Міністерство освіти і науки України
Львівський національний університет імені Івана Франка

Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Кваліфікаційна наукова праця
на правах рукопису

ПАВЛИШЕНКО БОГДАН МИХАЙЛОВИЧ

УДК 004.89:519.765

ДИСЕРТАЦІЯ

МЕТОДИ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ КОНСОЛІДОВАНИХ ДАНИХ ДЛЯ ПІДТРИМКИ ПРИЙНЯТТЯ РІШЕНЬ

05.13.23 – системи та засоби штучного інтелекту

Подається на здобуття наукового ступеня доктора технічних наук

Дисертація містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело
_____ Б. М. Павлишенко
Підпис

Науковий консультант – Дияк Іван Іванович,
доктор фізико-математичних наук, професор

Цей примірник дисертації ідентичний за змістом з іншими примірниками,
поданими до спеціалізованої вченої ради Д 64.052.01
Учений секретар спеціалізованої вченої ради Д 64.052.01 *Підпис* Є. І. Литвинова

Харків – 2021

АНОТАЦІЯ

Павлишенко Б.М. Методи інтелектуального аналізу консолідованих даних для підтримки прийняття рішень. – Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття наукового ступеня доктора технічних наук за спеціальністю 05.13.23 – системи та засоби штучного інтелекту. – Львівський національний університет імені Івана Франка, Харківський національний університет радіоелектроніки, Міністерство освіти і науки України, Львів, 2021.

Дисертаційну роботу присвячено розробленню методів моделювання, формування аналітичних ознак, інтелектуального аналізу табличних і текстових консолідованих даних для підвищення точності, достовірності та інформативності результатів аналізу, які використовуються для підтримки прийняття рішень в інформаційно-аналітичних системах. Об'єктом дослідження є процеси опрацювання та аналізу консолідованих даних із різною структурою та з різних джерел інформації. Предметом дослідження є моделі та методи інтелектуального аналізу консолідованих даних табличного та текстового типу. Методами дослідження є: теорія та алгоритми машинного та глибокого навчання для створення прогнозних моделей та їх ансамблів; теорія машинного навчання з підкріпленням для побудови моделей інтелектуальних агентів в алгоримах оптимізації послідовності прийняття рішень; теорія ймовірності та математична статистика для формування частотних семантичних характеристик текстових лексем та для створення ймовірнісних прогнозних моделей інтелектуального аналізу даних; теорія множин для створення теоретико-множинних моделей семантичних та тематичних полів; теорія частих множин та асоціативних правил, а також теорія аналізу формальних концептів для розробки підходів в аналітиці текстових потоків даних. Унаслідок проведених аналітичних та експериментальних досліджень отримано такі нові результати: розроблено метод оптимізації прогнозної аналітики часових рядів з використанням стекінгового об'єднання та відбору різнотипних моделей на основі лінійної регресії LASSO та байєсівської регресії, що

забезпечує підвищення точності прогнозування та формування оптимального прогнозного ансамблю моделей; розроблено метод виявлення технічних відмов, який, за рахунок поєднання байєсівської, лінійної та машинно-навчальної логістичних регресій, забезпечує підвищення достовірності результатів, що дозволяє побудувати ефективні диверсифіковані процеси прийняття рішень; отримали подальший розвиток методи оптимізації послідовності дій інтелектуального агента в задачах аналітики попиту з використанням глибокого Q-навчання та імітаційного моделювання середовища взаємодії на основі параметричної моделі та з використанням історичних даних, що забезпечує підвищення ефективності прийняття бізнес рішень; розроблено метод векторного представлення текстових даних, який, за рахунок використання теорії семантичних та тематичних полів, дозволяє представляти текстові документи у низькорозмірному просторі семантичних ознак та забезпечує зменшення складності розрахунків і підвищення достовірності результатів в аналізі текстових даних; розроблено метод аналізу текстових даних на основі алгоритмів машинного навчання з використанням кількісних ознак семантичних і тематичних полів та метод генетичної оптимізації набору цих ознак, що забезпечує підвищення достовірності результатів інтелектуального аналізу текстових масивів. удосконалено метод класифікаційного та регресійного аналізу різнотипних консолідованих даних на основі поєднання LSTM нейромережі з вхідними текстовими даними та нейромережі з повністю з'єднаними шарами з вхідними кількісними ознаками, що забезпечує підвищення точності та достовірності результатів; розроблено метод виявлення додаткових аналітичних ознак на основі лексемних поєднань у семантичних структурах текстових масивів, який, за рахунок використання теорії частих множин та асоціативних правил, розширює інформаційну основу для підтримки прийняття рішень в аналітиці консолідованих даних; розроблено модель семантичних концептів текстових масивів на основі теорії формальних концептів, що дозволяє виявляти ефективні аналітичні ознаки з урахуванням семантичної структури текстових масивів. Одержані у дисертаційному дослідженні результати та розроблені методи є складовою технологією для підтримки прийняття рішень у комплексних інформаційних системах і

забезпечують підвищення інформативності та надійності інтелектуального аналізу даних у прогнозній аналітиці різнотипних консолідованих даних. Одержані результати дають можливість: підвищити точність у задачах прогнозування та зменшити кількість моделей у стекінговому ансамблі на 30% для певного класу задач за рахунок розроблених методів стекінгового об'єднання різнотипних моделей у прогнозні ансамблі; оцінити невизначеність та прогнозні ризики складових моделей при прийнятті експертних рішень щодо формування прогнозного ансамблю моделей за рахунок розробленого методу використання байєсівської регресії для стекінгу прогнозних моделей; підвищити точність та інформативність результатів у задачах аналізу динаміки попиту та в аналітиці фінансових часових рядів за рахунок розроблених методів застосування лінійних, ймовірнісних та машинно-навчальних прогнозних моделей з урахуванням аналітичних ознак консолідованих даних заданої предметної області інтелектуального аналізу; оптимізувати набір прогнозних ознак та підвищити точність прогнозування за рахунок розроблених методів у прогнозуванні технічних відмов на лініях збірки на виробництві з використанням стекінгового об'єднання моделей; зменшити кількість аналітичних семантичних ознак текстових даних у 3-10 разів у порівнянні з набором лексемних частотних ознак для заданих характеристик інтелектуального аналізу текстових даних за рахунок розроблених методів використання теорії семантичних та тематичних полів; кількісно аналізувати семантичну складову авторського ідіолекта в текстових масивах за рахунок розробленого методу аналізу текстів із використанням теорії семантичних та тематичних полів; сформувати додаткові семантичні ознаки для прогнозних моделей та підвищити якість інформаційно-аналітичних систем за рахунок розроблених методів інтелектуального аналізу текстових потоків соціальної мережі Твіттер з використанням теорії частих множин і асоціативних правил та теорії формальних концептів.

У першому розділі наведено літературний огляд основних моделей, методів та підходів, які використовуються в інтелектуальному аналізі даних. Розглянуто методи аналізу даних табличного та текстового типів. Наведено основні положення лексемної семантики та теорії семантичних полів. Розглянуто семантичне структурування лексемного словника на основі

семантичної системи WordNet. На основі наведених літературних даних зроблено висновки та сформульовано невирішені питання.

У другому розділі розглянуто моделювання, формування ознак та інтелектуальний аналіз даних табличного типу. Для експериментального аналізу розглянуто історичні дані часових рядів, які описують динаміку продажів у роздрібній мережі. Розроблено комплексний підхід у прогнозній аналітиці табличних даних на основі параметричних та машинно-навчальних моделей, який дає змогу утворювати оптимальний набір аналітичних ознак та формувати ефективний підхід у побудові прогнозних моделей. Розглянуто об'єднання моделей різних типів у прогнозний ансамбль на основі стекінгового підходу, в якому прогнозні значення різних моделей, які було отримано на валідаційній вибірці, використано як прогнозні ознаки для моделей другого рівня. Розглянуто використанням LASSO регресії як стекінгової моделі другого рівня. У роботі проведено дослідження застосування байєсівської регресії, яка дає можливість оцінити невизначеність складових факторів аналізу і прогнозні ризики, а також врахувати екстремальні значення при використанні негаусових розподілів із "товстими хвостами" для цільової змінної. Досліджено реалізацію стекінгу різних прогнозних моделей за допомогою байєсівської регресії, яку було використано на другому стекінговому рівні, що дозволяє отримати розподіли для регресійних коефіцієнтів моделей першого рівня прогнозного ансамблю і оцінити невизначеність, внесену кожною моделлю в результат стекінгу. Як приклад застосування запропонованих підходів розглянуто лінійну модель для ціни біткоїна, яка використовує регресійні ознаки, що базуються на статистиці біткоїна, характеристиках процесів видобутку біткоїна, трендах пошукових запитів Google, візитах на сторінки Вікіпедії, а також змінній, яка описує експертну корекцію. Досліджено різні підходи для логістичної регресії на прикладі проблеми виявлення відмов на виробничих лініях. Розглянуто використання моделей глибокого Q-навчання у задачах часових рядів продажів. Розглянуто безмодельний підхід Q-навчання для аналізу задачі оптимальних стратегій ціноутворення та задачі попиту та постачання.

У третьому розділі розглянуто концепції семантичних та тематичних лексикографічних полів із точки зору їхнього використання в алгоритмах

інтелектуального аналізу текстових масивів. Під семантичними полями розглядають множини лексем, об'єднані деякою парадигмою. На основі концепцій семантичних полів створено теоретико-множинну модель, яка об'єднує поняття семантичного та тематичного лексемних полів і дає можливість представляти текстові дані у просторі семантичних ознак з метою інтелектуального аналізу заданого семантичного спектру текстових даних. Розглянуто векторну модель текстових документів у семантичному просторі, базис якого утворено частотно-дистрибутивними характеристиками семантичних та тематичних полів. Використання концепції семантичних полів є ефективним у векторній моделі текстових документів унаслідок зменшення розмірності простору семантичних ознак для векторного представлення текстових документів. Запропоновано модель некорельованих вторинних семантичних полів, які формуються на основі методу головних компонент шляхом визначення ортонормованого базису семантичного простору.

У четвертому розділі проаналізовано використання концепції семантичних полів в аналітиці текстових даних на основі методів машинного навчання. Розглянуто текстові вибірки різних типів, зокрема, масив авторських текстів англомовної художньої прози, повідомлення груп новин та текстові повідомлення соціальної мережі Твіттер. Як семантичні ознаки, розглянуто частотні характеристики семантичних та тематичних полів, а також компонент тематик латентного розміщення Діріхле. Розроблено метод кластеризації текстових документів у семантичному просторі, який дає можливість отримувати новий структурний поділ документів за семантичними ознаками. Розроблено метод класифікації текстових даних за експертно сформованими семантичними ознаками, зокрема, квантитативними ознаками семантичних та тематичних полів, що дозволяє проводити інтелектуальний аналіз текстових масивів із відповідними семантичними акцентами, які відображають семантичну сторону предметної області аналізу. Для реалізації алгоритмів класифікації вибрано різні алгоритми машинного навчання з учителем, зокрема, алгоритми Random Forest, XGBoost, нейронні мережі прямого поширення. Розроблено метод використання семантичних ознак у комбінованих нейромережах із

використанням рекурентних підмереж для текстових даних та підмереж із повністю з'єднаними шарами для кількісних ознак, що диверсифікує простір прогнозних ознак в алгоритмах глибокого навчання та покращує якість інтелектуального аналізу консолідованих даних. Розроблено метод використання генетичних алгоритмів для оптимізації набору семантичних полів, які утворюють векторний простір документів в алгоритмах інтелектуального аналізу текстових даних, що дозволяє формувати ефективні низькорозмірні простори семантичних ознак у задачах інтелектуального аналізу текстових даних. Як цільову функцію для генетичної оптимізації використано точність класифікаційного алгоритму машинного навчання.

Розглянуто квантовий алгоритм пошуку ключових семантичних образів у масивах текстових об'єктів. Реалізація цього алгоритму здійснюється на основі квантових логічних елементів, зокрема, з використанням вентиля Тоффолі. Ітерація Гровера використовується для підсилення амплітуд квантових станів, які описують семантичні вектори текстових об'єктів. Показано, що реалізація квантових алгоритмів аналізу семантичних образів текстових об'єктів для певного класу задач дає можливість поліноміально зменшити час виконання алгоритму у порівнянні з класичними алгоритмами внаслідок реалізації квантового паралелізму.

У п'ятому розділі розглянуто використання теорії частих множин та асоціативних правил в аналітиці текстових повідомлень соціальних мереж, зокрема Твіттера, що дає можливість сформувати тематичне семантичне поле, яке в подальшому можна використовувати для пошуку асоціативних правил. На основі відібраних частих множин семантичних ознак можна побудувати асоціативні правила, які будуть відображати семантичні зв'язки змісту повідомлень мікроблогів. Розглянуто використання теорії графів для аналізу повідомлень мережі Твіттер, зокрема, для аналізу зв'язків між користувачами та виявлення різних спільнот. Показано, що у потоках твітів, у яких обговорюються очікувані події, можна виявити ознаки на основі частих множин, які мають прогнозний потенціал стосовно цих подій. Використовуючи алгоритми аналізу графів, а також теорію частих множин та асоціативних правил, проведено інтелектуальний аналіз повідомлень

мережі Твіттер, пов'язаних з пандемією COVID-19.

У шостому розділі на основі теорії аналізу формальних концептів запропоновано модель семантичного контексту, яка відображає структурну семантичну організацію текстових масивів. У семантичному контексті формується частково впорядкована множина семантичних концептів, формальний зміст яких визначається семантичними полями, а формальний об'єм – текстовими документами. Розроблено метод використання моделі семантичного контексту в аналітиці текстових повідомлень соціальних мереж. Побудова ґратки семантичних концептів дає можливість описувати ієрархічну семантичну структуру в масиві документів та виявляти групи текстових документів, які об'єднані спільною групою семантичних ознак. Запропоновано застосування теорії аналізу формальних концептів в інтелектуальній обробці повідомлень Твіттера.

За результатами досліджень опубліковано 52 наукові праці, серед яких 30 статей у наукових фахових журналах і 22 публікації у матеріалах конференцій. Серед публікацій 7 статей опубліковано у наукових журналах із списку Scopus, а також 5 статей опубліковано у матеріалах конференцій, які реферуються у Scopus.

На основі проведених досліджень вирішено актуальну науково-прикладну проблему вибору, поєднання та оптимізації методів інтелектуального аналізу консолідованих даних шляхом розроблення методів моделювання, формування інформативних аналітичних ознак та інтелектуального аналізу табличних та текстових даних із врахуванням предметної області аналізу, що дозволило створювати ефективні прогнозні багаторівневі моделі, розширити інформативність інтелектуального аналізу різнотипних даних та вдосконалити підтримку прийняття рішень для комплексних інформаційно-аналітичних системах.

Ключові слова: інтелектуальний аналіз даних, методи машинного навчання, ознаки даних, часові ряди, семантичні поля,часті множини, асоціативні правила, аналіз формальних концептів.

ABSTRACT

Pavlyshenko B. M. Methods of intellectual analysis of consolidated data for decision-making support. – Qualification scientific work as a manuscript.

The thesis for the degree of Doctor of technical sciences specialty 05.13.23 – systems and means of artificial intelligence. – Ivan Franko National University of Lviv, Kharkiv National University of Radio Electronics, Ministry of Education and Science of Ukraine, Kharkiv, 2021.

The thesis is focused on the development of methods of modeling, formation of analytical features, intelligent analysis of tabular and textual consolidated data to improve the accuracy, reliability and self-descriptiveness of the analysis results, which are used to support decision-making in information and analytical systems. The object of the research is the processing and analysis of consolidated data with different structures and from different sources of information. The subject of the research is models and methods of the intellectual analysis of consolidated data of tabular and textual type. The methods of the research are: the theory and algorithms of machine and deep learning for creating predictive models and their ensembles; the theory of machine learning with reinforcement for building models of intelligent agents in the algorithms for optimizing the sequence of decision-making; the probability theory and mathematical statistics for the formation of frequency semantic characteristics of textual lexemes and for the creation of probabilistic predictive models of intellectual data analysis; the set theory for creating set-theoretic models of semantic and thematic fields; the theory of frequent sets and association rules, as well as the theory of analysis of formal concepts for the development of approaches in the analytics of text data streams. As a result of theoretical and experimental studies, the following scientific results were obtained: a method for optimizing the predictive analytics of time series using stacking combination and a selection of different types of models based on linear regression LASSO and Bayesian regression has been developed, providing an increase in forecasting accuracy as well as the formation of an optimal predictive ensemble of models; a method for detecting technical failures has been developed, which, due to a combination of Bayesian, linear and machine-learning logistic regression, provides an increase in the reliability of results, making it possible to

build effective diversified decision-making processes; the methods for optimizing the sequence of actions of an intelligent agent in the tasks of demand analytics using deep Q-learning and simulation modeling of the interaction environment based on a parametric model and using historical data were further developed, providing an increase in the efficiency of business decision-making; a method of vector representation of textual data has been developed, which, through the theory of semantic and thematic fields, makes it possible to present text documents in a low-dimensional space of semantic features, reduces the complexity of calculations and increases the reliability of results in the analysis of textual data; a method for analyzing textual data based on machine learning algorithms using quantitative features of semantic and thematic fields as well as a method for genetic optimization of a set of these features have been developed, providing an increase in the reliability of the results of the mining of text arrays; the method of classification and regression analysis of different types of consolidated data based on the combination of LSTM neural network with input text data and neural network with fully connected layers with input quantitative features has been improved, providing an increased reliability of the results; a method for identifying additional analytical features based on lexeme combinations in the semantic structures of text arrays has been developed, which, through the use of the theory of frequent sets and association rules, expands the information basis to support decision-making in the analytics of consolidated data; a model of semantic concepts of text based on the theory of formal concepts has been developed, which allows identifying effective analytical features taking into account the semantic structure of text arrays. The results obtained in the thesis research and the developed methods are a composite technology for decision-making support in complex information systems and provide an increase of self-descriptiveness and reliability of intellectual data analysis in predictive analytics of different types of consolidated data. The obtained results make it possible to: increase the accuracy in forecasting tasks and reduce the number of models in a stacking ensemble by 30% for a certain class of tasks due to the developed methods of stacking combination of different types of models into predictive ensembles; assess the uncertainty and predictive risks of the constituent models when making expert decisions on the formation of a predictive ensemble of models due to the developed method of using Bayesian regression for

stacking predictive models; increase the accuracy and self-descriptiveness of the results in the analyses of demand dynamics and in the analytics of financial time series due to the developed methods of applying linear, probabilistic and machine-learning predictive models based on analytical features of the consolidated data of a given subject area of intellectual analysis; optimize the set of predictive features and improve the forecasting accuracy due to the developed methods in predicting technical failures on assembly lines in production using a stacking combination of models; reduce the number of analytical semantic features of textual data by 3-10 times compared to a set of lexeme frequency features for the given characteristics of the intellectual textual data analysis due to the developed methods of using the theory of semantic and thematic fields; quantitatively analyze the semantic component of the author's idiolect in text arrays due to the developed method of text analysis using the theory of semantic and thematic fields; form additional semantic features for predictive models and improve the quality of information and analytical systems through the developed methods of intellectual analysis of text streams of Twitter using the theory of frequent sets and association rules as well as the theory of formal concepts.

The first section provides a literature overview of the main models, methods, and approaches used in intellectual data analysis. The methods for analyzing data of tabular and textual types are considered. The main points of lexeme semantics and the theory of semantic fields are given. The semantic structuring of the lexeme dictionary based on the WordNet semantic system is considered. On the basis of the given literature data, the conclusions are drawn and unresolved issues are formulated.

The second section deals with modeling, feature formation and intellectual analysis of tabular data. For experimental analysis, historical data of time series describing the dynamics of sales in retail network have been considered. An integrated approach has been developed in predictive analytics of tabular data based on parametric and machine-learning models, which allow one to create an optimal set of analytical features and form an effective approach in building predictive models. The combinations of models of various types into a predictive ensemble based on the stacking approach are considered, in which the predicted values of various models, obtained on the validation set, are used as predictive

features for second-level models. The use of LASSO regression as a second level stacking model has been analyzed. The paper studies the application of Bayesian regression, which makes it possible to assess the uncertainty of the constituent factors of the analysis and predictive risks, as well as take into account extreme values when using non-Gaussian distributions with "fat tails" for the target variable. The implementation of stacking of various predictive models using Bayesian regression, used at the second stacking level, has been investigated making it possible to obtain distributions for the regression coefficients of the models of the first level of the predictive ensemble and estimate the uncertainty introduced by each model into the stacking result. As an example of the application of the proposed approaches, a linear model for the price of bitcoin is considered, which uses regression features based on bitcoin statistics, characteristics of bitcoin mining processes, trends in Google search queries, visits to Wikipedia pages, and a variable describing the expert correction. Various approaches for logistic regression have been studied on the example of the problem of identifying failures on production lines. The use of deep Q-learning models in sales time series problems is considered. A non-model Q-learning approach for the analysis of the problem of optimal pricing strategies and the problem of demand and supply is considered.

The third section considers the concepts of semantic and thematic lexicographic fields in terms of their use in algorithms for the intellectual analysis of text arrays. Semantic fields are regarded as sets of lexemes united by a certain paradigm. Based on the concepts of semantic fields, a set-theoretical model has been created, which combines the concept of semantic and thematic lexeme fields and makes it possible to represent textual data in the space of semantic features for the purpose of intellectual analysis of a given semantic spectrum of textual data. A vector model of text documents in a semantic space, the basis of which is formed by the frequency-distribution characteristics of semantic and thematic fields, is considered. The use of the concept of semantic fields is effective in the vector model of text documents due to a decrease in the dimension of the space of semantic features for the vector representation of text documents. A model of uncorrelated secondary semantic fields, which are formed on the basis of the principal components method by determining the orthonormal basis of semantic space, is proposed.

The fourth section analyzes the use of the concept of semantic fields in text data analytics based on machine learning methods. The text samples of various types have been considered, particularly an array of author's texts of English fiction, newsgroup messages and text messages from Twitter. The frequency characteristics of the semantic and thematic fields, as well as the component of the themes of the latent Dirichlet placement are considered as semantic features. A method for clustering text documents in the semantic space has been developed, making it possible to obtain a new structural division of documents by semantic features. A method has been developed for the classification of text data by expertly formed semantic features, in particular quantitative features of semantic and thematic fields, which allows for intellectual analysis of text arrays with appropriate semantic accents that reflect the semantic side of the subject area of analysis. Various algorithms have been chosen to implement the classification algorithms of supervised machine learning, in particular, Random Forest, XGBoost, feedforward neural networks. A method for using semantic features in combined recurrent neural subnetworks for text data and subnetworks with fully connected layers for quantitative features has been developed, which diversifies the space of predictive features in deep learning algorithms and improves the quality of the intellectual analysis of consolidated data. A method has been developed for using genetic algorithms to optimize a set of semantic fields that form a vector space of documents in algorithms for the intellectual analysis of text data, it allows forming effective low-dimensional spaces of semantic features in the tasks of the intellectual analysis of text data. The accuracy of the classification algorithm of machine learning is used as a target function for genetic optimization. A quantum algorithm for searching for key semantic patterns in arrays of text objects is considered. The implementation of this algorithm is based on quantum logic gates, in particular, using the Toffoli gate. Grover's iteration is used to amplify the amplitudes of quantum states that describe the semantic vectors of text objects. It is shown that the implementation of quantum algorithms for the analysis of semantic patterns of text objects for a certain class of problems makes it possible to reduce polynomially the execution time of the algorithm in comparison with classical methods due to the implementation of quantum parallelism.

The fifth section considers the use of a frequent set theory and associative

rules in the analysis of text messages of social networks, Twitter in particular, making it possible to form a thematic semantic field, which can later be used to search for associative rules. Based on selected frequent sets of semantic features, one can build associative rules that will reflect the semantic relationships of the content of microblog posts. The use of graph theory for the analysis of Twitter messages is considered, particularly for the analysis of connections between users and the identification of different communities. It has been shown that in tweets that discuss expected events, it is possible to identify the features based on frequent sets that have predictive potential for these events. Using graph analysis algorithms, as well as the theory of frequent sets and associative rules, an intellectual analysis of Twitter messages related to the COVID-19 pandemic was performed.

The sixth section, basing on the theory of formal concepts analysis, proposes a model of the semantic context, which reflects the structural semantic organization of text arrays. In the semantic context, a partially ordered set of semantic concepts is formed, the formal intent of which is determined by semantic fields, and the formal extent is defined by text documents. A method of using the semantic context model in the analysis of text messages of social networks has been developed. Building a lattice of semantic concepts makes it possible to describe a hierarchical semantic structure in an array of documents as well as identify the groups of text documents that are united by a common group of semantic features. The application of the theory of analysis of formal concepts in the intellectual processing of Twitter messages is proposed.

According to the research results, 52 scientific works have been published, including 30 articles in scientific journals and 22 publications in conference proceedings. Among the publications, 7 articles were published in scientific journals from the Scopus list and 5 articles were published in conference proceedings which are reviewed in Scopus.

The conducted studies have solved the relevant scientific and applied problem of a choice, combination and optimization of methods of the intellectual analysis of consolidated data by developing methods of modeling, formation of informative analytical features and intellectual analysis of tabular and textual data, taking into account the subject area of analysis, which made it possible

to create effective predictive multilevel models, expand the self-descriptiveness of intellectual analysis of various types of data and improve the decision support for complex information-analytical systems.

Keywords: intellectual analysis, machine learning methods, data features, time series, semantic fields, frequent sets, associative rules, formal concepts analysis.

Публікації результатів дисертаційної роботи

Список публікацій здобувача, в яких опубліковано основні наукові результати дисертації:

1. Pavlyshenko B. M. Machine-learning models for sales time series forecasting // Data. 2019. Vol. 4, № 1. P. 15. (Входить до міжнародних наукометричних баз Web of Science та Scopus)

2. Pavlyshenko B. Genetic Optimization of Keyword Subsets in the Classification Analysis of Authorship of Texts // Journal of Quantitative Linguistics. 2014. Vol. 21, № 4. P. 341–349. (Входить до міжнародних наукометричних баз Web of Science та Scopus)

3. Pavlyshenko B. Clustering of Authors' Texts of English Fiction in the Vector Space of Semantic Fields // Cybernetics and Information Technologies. 2014. Vol. 14, № 3. P. 25–36. (Входить до міжнародних наукометричних баз Web of Science та Scopus)

4. Pavlyshenko B. Classification analysis of authorship fiction texts in the space of semantic fields // Journal of Quantitative Linguistics. 2013. Vol. 20, № 3. P. 218–226. (Входить до міжнародних наукометричних баз Web of Science та Scopus)

5. Pavlyshenko B. The Distribution of Semantic Fields in Author's Texts // Cybernetics and Information Technologies. 2016. Vol. 16, № 3. P. 195–204. (Входить до міжнародних наукометричних баз Web of Science та Scopus)

6. Павлишенко Б. Квантовий алґоритм еволюційного аналізу одновимірних кліткових автоматів // Журнал фізичних досліджень. 2011. Т. 15, № 3. С. 1–6. (Входить до міжнародної наукометричної бази Scopus)

7. Pavlyshenko B. M. Sales Time Series Analytics Using Deep Q-learning // International Journal of Computing. 2020. Sep. Vol. 19, № 3. P. 434–441. (Входить до міжнародної наукометричної бази Scopus)

8. Павлишенко Б. М. Модель семантичного контексту в алгоритмах інтелектуального аналізу текстів // Комп'ютинг. 2011. Т. 10, № 3. С. 216–222.

9. Павлишенко Б. Семантична кластеризація текстових документів методом k–середніх // Комп'ютерні науки та інформаційні технології. 2011. № 710. С. 215–218.

10. Павлишенко Б. М. Групування тегів користувачів мікроблогів на основі ґратки семантичних концептів // Комп'ютерні системи та мережі. 2011. № 717. С. 120–124.

11. Павлишенко Б. М. Пошук частих множин семантичних ознак та асоціативних правил в повідомленнях мікроблогів // Нові технології. 2011. № 3(33). С. 82–86.

12. Павлишенко Б. М. Моделювання нечітких семантичних полів у масивах текстових документів // Системи обробки інформації. 2011. № 8. С. 175–178.

13. Павлишенко Б. М. Квантовий алгоритм пошуку ключових слів у масивах текстових даних // Біоніка інтелекту. 2011. № 3(77). С. 157–161.

14. Павлишенко Б. Числове моделювання алгоритму Гровера для квантового пошуку даних // Теоретична електротехніка. 2010. № 61. С. 49–59.

15. Павлишенко Б. М. Сингулярна декомпозиція матриці семантичних ознак в алгоритмі ієрархічної кластеризації текстових масивів // Математичні машини і системи. 2012. Т. 1, № 1. С. 69–76.

16. Павлишенко Б. М. Групування текстових даних на основі моделі семантичного контексту // Східно-Європейский журнал передових технологій. 2011. № 5 (2). С. 39–42.

17. Павлишенко Б. М. Модель решітки семантичних концептів для інтелектуального аналізу мікроблогів // Штучний інтелект. 2012. № 1. С. 103–111.

18. Павлишенко Б. М. Часова залежність квантитативних характеристик ключових тегів у RSS каналах // Системи обробки інформації. 2012. № 3 (2). С. 199–202.

19. Павлишенко Б. Ймовірнісна класифікація текстових документів у просторі семантичних полів // Електроніка та інформаційні технології. 2012. № 2. С. 164–172.

20. Павлишенко Б. М. Кластерний аналіз повідомлень груп новин у просторі семантичних ознак // Комп'ютерні системи та мережі. 2012. № 745. С. 148–155.

21. Павлишенко Б. Класифікація повідомлень груп новин у векторному просторі семантичних полів // Комп'ютерні науки та інформаційні технології. 2012. № 744. С. 294–302.

22. Павлишенко Б. М. Аналіз семантичних образів у масивах текстових об'єктів за допомогою квантових обчислень // Математичні машини і системи. 2013. № 1. С. 34–43.

23. Павлишенко Б. М. Формування базису семантичного простору текстових документів за допомогою генетичних алгоритмів // Математичні машини і системи. 2013. № 2. С. 96–104.

24. Павлишенко Б. М. Використання лексемних полів у інтелектуальному аналізі текстових масивів // Штучний інтелект. 2013. № 1. С. 98–109.

25. Павлишенко Б. М. Модель вторинних некорельованих семантичних полів для аналізу текстових даних // Системні дослідження та інформаційні технології. 2014. № 3. С. 130–138.

26. Pavlyshenko B. M. Forecasting of Events by Tweets Data Mining // Electronics and information technologies. 2018. № 10. P. 71–85.

27. Pavlyshenko B. M. Can Twitter Predict Royal Baby's Name? // Electronics and information technologies. 2019. № 11. P. 52–60.

28. Pavlyshenko B. M. Detection of Technical Failures on Production Lines Using Machine Learning, Linear and Bayesian Models of Logistic Regression // Electronics and information technologies. 2019. № 12. P. 3–19.

29. Павлишенко Б. М. Використання методів машинного навчання та семантичних ознак в інтелектуальному аналізі текстових даних // Електроніка та інформаційні технології. 2020. № 13. С. 3–18.

30. Pavlyshenko B. M. Modeling COVID-19 Spread and Its Impact on Stock Market Using Different Types of Data // Electronics and information technologies. 2020. № 14. P. 3–21.

Публікації, які засвідчують апробацію матеріалів дисертації:

31. Павлишенко Б. М. Використання квантових алгоритмів в системах розпізнавання образів // Друга Всеукраїнська науково–практична конференція "Проблеми електроніки та інформаційні технології", 02–05 вересня 2010 р. – Львів–Чинадієво. 2010. С. А11.

32. Павлишенко Б. М. Алгоритми семантичної векторизації та кластеризації текстових масивів // Друга Всеукраїнська науково–практична конференція "Проблеми електроніки та інформаційні технології", 02–05 вересня 2010 р. – Львів–Чинадієво. 2010. С. А12.

33. Павлишенко Б. М. Кластерний аналіз текстових документів в просторі семантичних концептів // Збірник доповідей науково–практичної

конференції з міжнародною участю "Системи підтримки прийняття рішень. Теорія і практика", 6 червня 2011 р. – Київ. 2011. С. 146–149.

34. Павлишенко Б. М. Алгоритми семантичного групування текстових документів // III науково–практична конференція "Електроніка та інформаційні технології (ЕлІТ–2011)": тези доповідей, 01–04 вересня 2011 р. – Львів–Чинадієво. 2011. С. 22–23.

35. Павлишенко Б. М. Модель формального семантичного контексту в алгоритмах обробки текстових документів // III науково–практична конференція "Електроніка та інформаційні технології (ЕлІТ–2011)": тези доповідей, 01–04 вересня 2011 р. – Львів–Чинадієво. 2011. С. 24–27.

36. Павлишенко Б. М. Інтелектуальний аналіз мікроблогів за допомогою решітки семантичних концептів // 5-а міжнародна науково–технічна конференція ACSN–2011 "Сучасні комп'ютерні системи та мережі: розробка та використання": тези доповідей, 29 вересня – 1 жовтня 2011 р. – Львів. 2011. С. 85–87.

37. Павлишенко Б. М. Аналіз формальних семантичних понять в алгоритмах обробки даних // XVII Всеукраїнська наукова конференція "Сучасні проблеми прикладної математики та інформатики": тези доповідей, 6–7 жовтня 2011 р. – Львів. 2011. С. 80.

38. Павлишенко Б. М. Векторна модель текстових документів у семантичному ортонормованому базисі // XVIII Всеукраїнська наукова конференція "Сучасні проблеми прикладної математики та інформатики": тези доповідей, 4–5 жовтня 2012 р. – Львів. 2012. С. 127.

39. Павлишенко Б. М. Модель нечітких семантичних полів для інтелектуального аналізу текстових масивів // IV науково–практична конференція "Електроніка та інформаційні технології (ЕлІТ–2012)": тези доповідей, 30 серпня – 2 вересня 2012 р. – Львів–Чинадієво. 2012. С. 98.

40. Павлишенко Б. М. Аналіз семантичних асоціацій у веб–блогах за допомогою гратки формальних понять // Міжнародна науково–технічна конференція "Штучний інтелект. Інтелектуальні системи" (ШІ–2012): матеріали конференції, 1–5 жовтня, 2012 р. – Кацивелі, АР Крим. 2012. С. 118–122.

41. Павлишенко Б. М. Аналіз мікроблогів користувачів на основі гратки семантичних концептів // Збірник доповідей науково–практичної конференції з міжнародню участю "Системи підтримки прийняття рішень. Теорія і практика", 6 червня 2012 р. – Київ. 2012. С. 115–118.

42. Павлишенко Б. М. Прогнозування подій на основі інтелектуального аналізу повідомлень мікроблогів Twitter // XIII міжнародна наукова конференція імені Т. А. Таран "Інтелектуальний аналіз інформації" (ІАІ–2013): збірка праць, 15–17 травня 2013 р. – КПИ, Київ. 2013. С. 199–205.

43. Павлишенко Б. М. Чи може Твіттер передбачити ім'я британського принца? // XIX Всеукраїнська наукова конференція "Сучасні проблеми прикладної математики та інформатики": тези доповідей, 3–4 жовтня 2013 р. – Львів. 2013. С. 108.

44. Павлишенко Б. М. Використання інтелектуального аналізу повідомлень Twitter у прогнозуванні фінансових ринків // Матеріали 2–ї Міжнародної конференції "Інформація, комунікація, суспільство 2013" (ІКС–2013), 16–19 травня, 2013 р. – Львів–Славське. 2013. С. 86–87.

45. Павлишенко Б. М. Аналіз курсу акцій на основі твітів інформагенств // V науково–практична конференція "Електроніка та інформаційні технології" (ЕлІТ–2013): тези доповідей, 29 серпня–1 вересня 2013 р. – Львів–Чинадієво. 2013. С. 60.

46. Pavlyshenko B. M. Linear, machine learning and probabilistic approaches for time series analysis // Data Stream Mining & Processing (DSMP), IEEE First International Conference. 2016. P. 377–381. (Входить до міжнародної наукометричної бази Scopus)

47. Pavlyshenko B. Machine learning, linear and Bayesian models for logistic regression in failure detection problems // Big Data (Big Data), 2016 IEEE International Conference on, IEEE, Washington D.C. 2016. P. 2046–2050. (Входить до міжнародної наукометричної бази Scopus)

48. Pavlyshenko B. Using Stacking Approaches for Machine Learning Models // 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP). 2018. P. 255–258. (Входить до міжнародної наукометричної бази Scopus)

49. Pavlyshenko B. Predictive Analytics for Sales Time Series // Xth International Scientific and Practical Conference "Electronics and Information Technologies" (ELIT-2018) August 30 - September 2, 2018, Lviv, Karpaty village, Issue 10. 2018. P. 85–87.

50. Pavlyshenko B. M. Regression Approaches For Sales Time Series Forecasting // Матеріали XXIV Всеукраїнської наукової конференції "Сучасні проблеми прикладної математики та інформатики", APAMCS-2018 26-28 вересня 2018 року, Львів. 2018. С. 121–123.

51. Pavlyshenko B. Bitcoin Price Predictive Modeling Using Expert Correction // 2019 XIth International Scientific and Practical Conference on Electronics and Information Technologies (ELIT), September 16 – 18, 2019 Lviv, Ukraine. 2019. P. 163–167. (Входить до міжнародної наукометричної бази Scopus)

52. Pavlyshenko B. Using Bayesian Regression for Stacking Time Series Predictive Models // 2020 IEEE Third International Conference on Data Stream Mining & Processing (DSMP). 2020. P. 305–309. (Входить до міжнародної наукометричної бази Scopus)

ЗМІСТ

ВСТУП

Актуальність теми. Інтелектуальний аналіз даних є одним з важливих та перспективних напрямків сучасних інформаційних технологій. Суть такого аналізу полягає у пошуку складних закономірностей та виявленні патернів у масивах даних. Дані відображають різноманітні явища та процеси у бізнесі, суспільстві, соціальних мережах, технічних пристроях тощо. У сучасній інформаційній епосі існує багато різноманітних джерел даних різної структури, які містять кількісні та якісні величини різноманітних ознак. Актуальним є об'єднання усіх даних, дотичних до аналізованої задачі, в єдиній аналітичній моделі. Процес формування ознак даних, які відображають їхні характеристики та зведення масивів цих даних до реляційного вигляду, який часто використовується в алгоритмах інтелектуального аналізу даних, є складною нетривіальною проблемою, яку важко формалізувати. Складність полягає у тому, що різні процеси характеризуються даними з різною структурою, наприклад, частина даних може мати табличну структуру, а частина – текстову. Виявлення та формування ефективних аналітичних ознак цих процесів є різним у різних предметних областях і в основному базується на експертному досвіді. Важливим етапом в аналізі даних є їхня консолідація, під якою розуміють об'єднання масивів даних із різних джерел та з різною структурою для вирішення певної аналітичної проблеми. Це узагальнений етап аналізу даних, який може відрізнятись у різних предметних областях. Актуальним є створення узагальнених моделей та методів у аналізі даних, консолідації досліджуваних даних різних типів із різних джерел та різних предметних областей, виявлення та створення аналітичних ознак даних та їхнього узагальнення для підтримки прийняття рішень у заданому класі проблем.

В інтелектуальному аналізі даних використовують параметричні моделі та алгоритмічні моделі машинного навчання. Параметричні моделі, зокрема моделі лінійної регресії, дають можливість аналізувати вплив зовнішніх факторів на цільову змінну, однак вони не дозволяють враховувати складну взаємодію між факторами впливу. Методи машинного навчання дають можливість виявляти складні патерни у даних і здійснювати

прогнозування цільових змінних на основі натренованих на історичних даних моделях. Однак такі точні прогнози можливі у випадку достатньо великої вибірки історичних даних, які мають стаціонарний розподіл. Алгоритмічні моделі машинного навчання є певною інформацією, яка генерується алгоритмом машинного навчання на основі тренувальної вибірки даних і використовується цим алгоритмом для прогнозування цільової змінної. В алгоритмічному моделюванні використовуються дані, які відображають аналізовані процеси. Розроблення ефективних методів інтелектуального аналізу різноструктурованих консолідованих даних із використанням алгоритмічних моделей можливе шляхом аналізу різнотипних задач із різних предметних областей. На різних етапах такого аналізу стає очевидною доцільність розроблення нових методів та підходів, зокрема, шляхом поєднання наявних методів та алгоритмів. Аналізованим процесам властива деяка міра невизначеності, тому важливо враховувати та аналізувати невизначеність факторів впливу та цільової змінної для того, щоб оцінити ризики, пов'язані з неточністю прогнозування.

Прикладом слабоструктурованих типів даних можуть бути текстові масиви. Стимулом розвитку методів інтелектуального аналізу текстів є значний ріст слабоструктурованої інформації текстового типу, зокрема, у мережі Інтернет. Сучасний аналіз текстової інформації поряд із традиційними статистичними методами вимагає розвитку нових ефективних методів семантичного аналізу із заглибленням у зміст інформації, використовуючи методи машинного навчання. На сьогодні розроблено велику кількість алгоритмів та систем обробки природньої мови, які базуються на математичних статистичних методах і мають емпіричний характер. Виникає необхідність розвитку нових методик та алгоритмів, які би базувалися на глибоких теоретичних засадах лінгвістичної науки, зокрема, на результатах, отриманих у дослідженнях комп'ютерної лінгвістики.

Теоретичні питання та практичне застосування інтелектуального аналізу даних розглядають у своїх працях L. Breiman, J. H. Friedman, C. M. Bishop, D. H. Wolpert (машинне навчання), A. Gelman, J. Kruschke (байєсівські методи аналізу), G. E. Hinton, G. Cybenko, I. Goodfellow, Y. Bengio, A. Courville (глибоке навчання, нейромережі), R.S. Sutton, A. G. Barto,

V. Mnih (машинне навчання із підкріпленням), P. D. Turney, P. Pantel, D. M. Blei, T. Mikolov, F. Sebastiani (аналіз текстових даних), U. Priss, B. Ganter, G. Stumme, R. Wille (аналіз формальних концептів), R.J. Hyndman, G. Athanasopoulos, R.S. Tsay (аналіз часових рядів), D. E. Goldberg, J. H. Holland (генетичні алгоритми), A. Gliozzo, C. Strapparava, C. Fellbaum, G. A. Miller (семантика текстових даних), а також вітчизняні вчені – О.Г. Івахненко, Є.В. Бодянський, О.А. Винокурова, Д.Д. Пелешко (нейронні мережі), П.І. Бідюк (методи прогнозування), С.О. Субботін (системи штучного інтелекту), В.А. Широков (аналіз текстових даних), В.В. Пасічник, В.В. Литвин, Н.Б. Шаховська (інтелектуальні системи, аналіз консолідованих даних).

Практика інтелектуального аналізу показує, що сучасні бізнес процеси настільки складні, що важко виробити єдиний для всіх задач підхід у прогнозній аналітиці. Підбір, об'єднання прогнозних моделей та формування аналітичних ознак є об'єднаною комплексною проблемою інтелектуального аналізу, розв'язок якої базується як на сучасних методах аналізу даних, так і на знаннях у предметній області, до якої належать аналізовані процеси. Виникає потреба в удосконаленні наявних та розробці нових методів та підходів інтелектуального аналізу для підтримки прийняття рішень з урахуванням особливостей структури даних та предметної області. Актуальним є розгляд типових задач такого аналізу з різних предметних областей та узагальнення методів і алгоритмів розв'язку прикладних задач, беручи до уваги особливості заданої предметної області знань.

Отже, актуальною науково-прикладною проблемою є розроблення, вибір, поєднання та оптимізація моделей та методів інтелектуального аналізу різнотипних консолідованих даних з метою підвищення інформативності, точності та достовірності результатів для підтримки прийняття рішень в інформаційно-аналітичних системах.

Зв'язок роботи з науковими програмами, планами, темами. Тема дисертаційної роботи відповідає науковим напрямам факультету електроніки та комп'ютерних технологій Львівського національного університету імені Івана Франка, зокрема темі "Аналіз даних засобами машинного навчання" (номер держреєстрації 0119U002409).

Мета і задачі дослідження. Мета дисертаційної роботи полягає у розробленні методів моделювання, формування аналітичних ознак, інтелектуального аналізу табличних і текстових консолідованих даних для підвищення точності, достовірності та інформативності результатів аналізу, які використовуються для підтримки прийняття рішень в інформаційно-аналітичних системах.

Для реалізації мети дисертаційної роботи потрібно розв'язати такі задачі:

- проаналізувати наявні методи опрацювання та інтелектуального аналізу даних і сформулювати актуальні завдання для дисертаційного дослідження;
- розробити метод застосування машинно-навчальних та ймовірнісних моделей для покращення точності та якості інтелектуального аналізу даних табличного типу;
- розробити методи стекінгового об'єднання різнотипних моделей у прогнозні ансамблі на основі лінійної регресії LASSO та байєсівської регресії;
- удосконалити метод використання машинного навчання з підкріпленням в аналітиці табличних даних з імітаційним моделюванням середовища взаємодії інтелектуального агента;
- розробити метод використання теорії семантичних та тематичних полів у інтелектуальному аналізі даних з метою формування квантитативних семантичних ознак текстових даних;
- розробити метод інтелектуального аналізу текстових даних на основі машинного навчання з використанням семантичних ознак;
- удосконалити метод класифікаційного та регресійного аналізу з використанням нейромережі з вхідними текстовими даними та кількісними ознаками;

– розробити метод використання теорії частих множин та асоціативних правил для формування семантичних ознак в інтелектуальному аналізі текстових даних;

– розробити метод використання теорії формальних концептів в аналітиці текстових потоків соціальних мереж для аналізу семантичної структури текстових даних та формування аналітичних ознак;

– створити засоби для апробації розроблених у роботі методів інтелектуального аналізу табличних та текстових даних.

Об'єктом дослідження є процеси опрацювання та аналізу консолідованих даних із різною структурою та з різних джерел інформації.

Предметом дослідження є моделі та методи інтелектуального аналізу консолідованих даних табличного та текстового типу.

Методами дослідження є:

– теорія та алгоритми машинного та глибокого навчання для створення прогнозних моделей та їх ансамблів;

– теорія машинного навчання з підкріпленням для побудови моделей інтелектуальних агентів в алгоритмах оптимізації послідовності прийняття рішень;

– теорія ймовірності та математична статистика для формування частотних семантичних характеристик текстових лексем та для створення ймовірнісних прогнозних моделей інтелектуального аналізу даних;

– теорія множин для створення теоретико-множинних моделей семантичних та тематичних полів;

– теорія частих множин та асоціативних правил і теорія аналізу формальних концептів для розробки підходів в аналітиці текстових потоків даних;

Наукова новизна одержаних результатів. Унаслідок проведених теоретичних та експериментальних досліджень отримано такі наукові результати:

вперше:

- розроблено метод оптимізації прогнозної аналітики часових рядів з використанням стекінгового об'єднання та відбору різнотипних моделей на основі лінійної регресії LASSO та байєсівської регресії, що забезпечує підвищення точності прогнозування та формування оптимального прогнозного ансамблю моделей;
- розроблено метод виявлення технічних відмов, який, за рахунок поєднання байєсівської, лінійної та машино-навчальної логістичних регресій, забезпечує підвищення точності та достовірності результатів, що дозволяє побудувати ефективні диверсифіковані процеси прийняття рішень;
- розроблено метод векторного представлення текстових даних, який, за рахунок використання теорії семантичних та тематичних полів, дозволяє представляти текстові документи у низькорозмірному просторі семантичних ознак та забезпечує зменшення складності розрахунків і підвищення достовірності результатів в аналізі текстових даних;
- розроблено метод аналізу текстових даних на основі алгоритмів машинного навчання з використанням кількісних ознак семантичних і тематичних полів, а також метод генетичної оптимізації набору цих ознак, що забезпечує підвищення достовірності результатів інтелектуального аналізу текстових масивів;
- розроблено метод виявлення додаткових аналітичних ознак на основі лексемних поєднань у семантичних структурах текстових масивів, який, за рахунок використання теорії частих множин та асоціативних правил, розширює інформаційну основу для підтримки прийняття рішень в аналітиці консолідованих даних;

– розроблено модель семантичних концептів текстових масивів на основі теорії формальних концептів, що дозволяє виявляти ефективні аналітичні ознаки з урахуванням семантичної структури текстових масивів;

отримали подальший розвиток:

– методи оптимізації послідовності дій інтелектуального агента в задачах аналітики попиту з використанням глибокого Q-навчання та імітаційного моделювання середовища взаємодії на основі параметричної моделі та з використанням історичних даних, що забезпечує підвищення ефективності прийняття бізнес рішень;

удосконалено:

– метод класифікаційного та регресійного аналізу різнотипних консолідованих даних на основі поєднання LSTM нейромережі з вхідними текстовими даними та нейромережі з повністю з'єднаними шарами з вхідними кількісними ознаками, що забезпечує підвищення точності та достовірності результатів.

Практичне значення одержаних результатів. Одержані у дисертаційному дослідженні результати та розроблені методи є складовою технологією для підтримки прийняття рішень у комплексних інформаційних системах і забезпечують підвищення інформативності та надійності інтелектуального аналізу даних у прогнозній аналітиці різнотипних консолідованих даних. Одержані результати дають можливість:

– підвищити точність прогнозування та зменшити кількість моделей у стекінговому ансамблі на 30% для певного класу задач за рахунок розроблених методів стекінгового об'єднання різнотипних моделей у прогнозні ансамблі;

– оцінити невизначеність та прогнозні ризики складових моделей при прийнятті експертних рішень щодо формування прогнозного ансамблю моделей за рахунок розробленого методу використання байєсівської регресії для стекінгу прогнозних моделей;

– підвищити точність та інформативність результатів у задачах аналізу динаміки попиту та в аналітиці фінансових часових рядів за рахунок розроблених методів застосування лінійних, ймовірнісних та машинно-навчальних прогнозних моделей з урахуванням аналітичних ознак консолідованих даних заданої предметної області інтелектуального аналізу;

– оптимізувати набір прогнозних ознак та підвищити точність прогнозування за рахунок розроблених методів у прогнозуванні технічних відмов на лініях збірки на виробництві з використанням стекінгового об'єднання моделей;

– зменшити кількість аналітичних семантичних ознак текстових даних у 3-10 разів у порівнянні з набором лексемних частотних ознак для заданих характеристик інтелектуального аналізу текстових даних за рахунок розроблених методів використання теорії семантичних та тематичних полів;

– кількісно аналізувати семантичну складову авторського ідіолекта в текстових масивах за рахунок розробленого методу аналізу текстів із використанням теорії семантичних та тематичних полів;

– сформувати додаткові семантичні ознаки для прогнозних моделей та підвищити якість інформаційно-аналітичних систем за рахунок розроблених методів інтелектуального аналізу текстових потоків соціальної мережі Твіттер з використанням теорії частих множин і асоціативних правил та теорії формальних концептів.

Отримані у роботі результати використовуються у компанії Soft-Serve Inc. для розробки програмного забезпечення у задачах аналізу даних, а також впроваджені у відповідні навчальні курси у Львівському національному університеті імені Івана Франка.

Особистий внесок здобувача. Усі наукові результати, які виносяться на захист дисертаційної роботи, отримані автором самостійно. Усі наукові праці опубліковано одноосібно.

Апробація результатів дисертаційного дослідження. Основні результати роботи було представлено на таких наукових конференціях: Друга Всеукраїнська науково-практична конференція "Проблеми електроніки та інформаційні технології", 02–05 вересня 2010 р., Львів-Чинадієво; "Системи підтримки прийняття рішень. Теорія і практика", 6 червня 2011 р. – Київ; III науково-практична конференція "Електроніка та інформаційні технології (ЕлІТ–2011)": тези доповідей, 01–04 вересня 2011 р. – Львів-Чинадієво; 5-а міжнародна науково-технічна конференція ACSN–2011 "Сучасні комп'ютерні системи та мережі: розробка та використання", 29 вересня – 1 жовтня 2011 р. – Львів; XVII Всеукраїнська наукова конференція "Сучасні проблеми прикладної математики та інформатики", 6–7 жовтня 2011 р. – Львів; "Системи підтримки прийняття рішень. Теорія і практика", 6 червня 2012 р. – Київ; XVIII Всеукраїнська наукова конференція "Сучасні проблеми прикладної математики та інформатики", 4–5 жовтня 2012 р. – Львів; IV науково-практична конференція "Електроніка та інформаційні технології (ЕлІТ–2012)", 30 серпня–2 вересня 2012р. – Львів-Чинадієво; Міжнародна науково–технічна конференція Штучний інтелект. Інтелектуальні системи" (ШІ–2012), 1–5 жовтня, 2012 р. – Кацивелі, АР Крим; XIII міжнародна наукова конференція імені Т. А. Таран "Інтелектуальний аналіз інформації" (ІАІ–2013), 15–17 травня 2013 р. – КПИ, Київ; 2-а Міжнародна конференція "Інформація, комунікація, суспільство 2013" (ІКС–2013), 16–19 травня, 2013 р. – Львів-Славське; V науково-практична конференція "Електроніка та інформаційні технології" (ЕлІТ–2013), 29 серпня – 1 вересня 2013 р. – Львів-Чинадієво; XIX Всеукраїнська наукова конференція "Сучасні проблеми прикладної математики та інформатики", 3–4 жовтня 2013 р. – Львів; Data Stream Mining & Processing (DSMP), IEEE First International Conference 2016, Lviv; Big Data (Big Data), 2016 IEEE International Conference on, IEEE, Washington D.C.; 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP); Xth International Scientific and Practical Conference "Electronics and Information Technologies" (ELIT-2018) August 30 - September 2, 2018, Lviv; XXIV Всеукраїнська наукова конференція "Сучасні проблеми прикладної математики та інформатики", APAMCS-2018 26-28 вересня

2018 р., Львів; 2019 XIth International Scientific and Practical Conference on Electronics and Information Technologies (ELIT), September 16–18, 2019, Lviv, Ukraine; 2020 IEEE Third International Conference on Data Stream Mining & Processing (DSMP), August 21–25, 2020, Lviv, Ukraine.

Також, результати роботи було представлено на практичних конференціях для фахівців з аналізу даних Predictive Analytics World (London, 2018), Predictive Analytics World (Munich, 2019).

Публікації. За результатами досліджень опубліковано 52 наукові праці, серед яких 30 статей у наукових фахових журналах і 22 публікації у матеріалах конференцій. Серед публікацій 7 статей опубліковано у наукових журналах зі списку Scopus, а також 5 статей опубліковано у матеріалах конференцій, які реферуються у Scopus.

Структура та обсяг дисертаційної роботи. Дисертаційна робота складається зі вступу, шести розділів, висновків, списку літератури з 361 джерела та додатків, загальним обсягом 407 сторінки друкованого тексту, з яких 314 сторінок основного тексту.

1 АЛГОРИТМИ ТА МЕТОДИ В ІНТЕЛЕКТУАЛЬНОМУ АНАЛІЗІ ДАНИХ

1.1 Сучасні підходи в аналізі даних

Інтелектуальний аналіз даних є важливою складовою сучасних інформаційних технологій. Під консолідованими даними розглядають дані, які надходять із різних джерел. Актуальною задачею є об'єднання таких даних для використання у прогнозній аналітичній моделі та відповідних алгоритмах аналізу даних. На даний час можна спостерігати спроби створення універсальних підходів у аналітиці даних із різних предметних областей та об'єднання їх в одній аналітичній системі на основі методів машинного навчання. На нашу думку, створення узагальнених підходів дає результати лише для простих та стандартизованих задач. Для складних комплексних задач необхідно враховувати особливості предметної області на експертному рівні. Це в першу чергу стосується аналітики нестаціонарних процесів, зокрема, у випадку різних розподілів ознак на тренувальній та тестовій вибірках. Під тестовою вибіркою розуміють дані, для яких здійснюються прогнозування цільової змінної, яке в подальшому буде використане при прийнятті бізнес та технічних рішень. Наприклад, нестаціонарність може бути зумовлена трендом цільової змінної. У такому випадку прогнозна модель на основі методів машинного навчання може вносити суттєві зміщення у результат. Нестаціонарність також може бути зумовлена новими значеннями категоріальних змінних у тестовій вибірці, які не зустрічаються у тренувальній вибірці. Для формування оптимального набору ознак у таких випадках необхідно застосувати експертний підхід у формуванні стаціонарних або наближених до стаціонарних ознак. Тому для вироблення ефективних підходів важливо розглянути не узагальнений підхід до інтелектуального аналізу даних, а типові задачі певної предметної області. Проаналізуємо сучасні методи та підходи, які використовують в інтелектуальному аналізі даних. Дані можна поділити на структуровані, які можна описати деякою структурною схемою, та неструктуровані або напівструктуровані, які не мають чіткої структурної схеми даних. Прикладом структурованих даних можуть бути дані табличного типу, де

стрічки позначають зразки даних, а стовпці – ознаки даних. Ознаки даних можна поділити на числові та категоріальні. Числові дані можуть бути цілочисельного типу або типу із плаваючою комою. Категоріальні дані набувають деяких дискретних значень із заданої множини. Такі значення можуть бути як числові, так і стрічкові, прикладом такого типу даних може бути день тижня, місяць, назва міста, назва компанії тощо. Для того, щоб мати можливість аналізувати категоріальні дані, необхідно їх перетворити в числовий тип. Є різні методи таких перетворень. Один із широковживаних методів – це метод фіктивних змінних, або one-hot encoding. Суть методу полягає у заміні категоріальної змінної сукупністю бінарних числових змінних, назви яких відповідають можливим унікальним значенням категоріальної змінної. Якщо, наприклад, категоріальна змінна позначає день тижня, цю змінну можна замінити на сім бінарних змінних, які відповідають кожному дню тижня. Значення змінної дорівнює 1, якщо категоріальна змінна даних має значення, яке відповідає заданій бінарній змінній, в іншому випадку значення бінарної змінної дорівнює нулю. Прикладом слабоструктурованих даних можуть слугувати дані текстового типу. Формально текст можна представити як послідовність деяких елементів, які належать до множини словника. Елементи таких послідовностей можуть бути як стрічковими і відображати слова, так і числовими, де числа є закодованими словами. Числове представлення слів часто використовують в аналізі текстових даних за допомогою нейронних мереж. Розглянемо існуючі методи аналітики для даних табличного та текстового типів. У сучасній аналітиці даних табличного типу часто використовують методи машинного навчання з учителем, які полягають у знаходженні значень деякої цільової змінної за ознаками даних. Для прогнозування цільової змінної необхідно здійснити навчання прогнозної моделі на навчальній вибірці даних. Такі методи можна поділити на регресійні та класифікаційні. У випадку, коли цільова змінна є числовою, отримуємо регресію, якщо категоріальна – класифікацію. Якщо значення цільової змінної може бути лише двох типів, тоді отримуємо логістичну регресію або бінарну класифікацію. У прогнозній аналітиці використовують параметричні моделі та моделі машинного навчання. У параметричних

моделях зв'язок між цільовою змінною та ознаками описують певним виразом, у якому є декілька параметрів. Прикладом параметричної моделі може бути лінійна регресія, у якій зв'язок між цільовою змінною та ознаками описується лінійною залежністю, в якій кількісна ознака входить із деяким ваговим коефіцієнтом. Прикладами моделей машинного навчання можуть бути моделі із використання дерев рішень або нейронних мереж. Насправді такі моделі також мають параметри, однак їх кількість може бути великою за рахунок врахування складного взаємозв'язку між ознаками і прослідкувати зв'язок між цільовою змінною та ознаками дуже складно. У задачах класифікаційного та регресійного аналізу виникає проблема вибору оптимальної моделі із набором ефективних ознак. Одним із критеріїв може бути набір кількісних характеристик похибок. Однак, якщо похибки розраховувати на тій самій вибірці даних, які використовувались для вивчення моделі, може виникнути ефект перенавчання (*overfitting*). Тому часто вибірку даних розбивають на два набори даних. На одному наборі даних здійснюють навчання моделі, а на іншому здійснюють тестування моделі. Широко поширеним підходом є крос-валідація, при якій набір даних розбивають на n підмножин. На кожному кроці вибирають одну із підмножин для тестування, а решту підмножин використовують для тренування. Такий процес повторюють для кожної з підмножин. В кінці такого алгоритму отримаємо результати тестування для всіх підмножин, за умови, що дані цих підмножин не використовувались для навчання моделі. Коректний валідаційний підхід є дуже важливим у виборі ефективної моделі. Якщо здійснюється аналіз часових рядів, класичний кросвалідаційний підхід може бути неефективним у разі нестаціонарних процесів. Можна отримати значні похибки в оцінці моделі, якщо при тестуванні даних певного часового проміжку для навчання моделі були використані дані з наступних часових періодів. У таких задачах при побудові та аналізі прогнозної моделі дані розбивають на дві послідовні за часом множини. На першій за часом множині здійснюють тренування моделі, а на наступній множині здійснюють тестування та валідацію моделі.

В інтелектуальному аналізі даних часто використовують алгоритмічні прогнозні моделі [1, 2]. Алгоритмічну модель можна розглядати як

певну інформацію, яка генерується алгоритмом машинного навчання з використанням тренувальної вибірки даних і в подальшому використовується цим алгоритмом для прогнозування цільової змінної на вибірці нових даних, які не використовувалися в процесі тренування. Під алгоритмічним моделюванням можна розглядати процес створення алгоритмічної моделі, аналіз цієї моделі на валідаційній вибірці даних та підбір оптимальних параметрів алгоритму машинного навчання.

Розглянемо окремо методи та підходи, які використовують в аналітиці даних табличного та текстового типів. У роботі [3] розглянуто проблеми пошуку ефективних ознак у масивах даних із великим об'ємом нерелевантної інформації та методи виділення найбільш релевантних зразків даних для аналізу методами машинного навчання. В [4] розглянуто методи та підходи в побудові алгебраїчних моделей ознак. В [5, 6] проаналізовано підходи в аналітиці напівструктурованих даних. В [7] аналізуються алгоритми виділення ознак для інтелектуального аналізу даних. Розглянуті методи дають покращення результатів класифікаційного аналізу. В [8] розглядається методи та підходи в роботі із структурованими, слабоструктурованими та неструктурованими даними. В [9, 10, 11] аналізуються неструктуровані дані. Напівструктуровані дані розглянуті в [12, 13]. Створення реляційних даних на основі неструктурованих даних розглянуто в [14]. В [15] описано алгоритми відбору ознак для задач класифікації та кластеризації даних. В [16] розглянуто підходи до зменшення розмірності простору ознак, зокрема, за допомогою алгоритмів SVD, PCA. Особливістю неструктурованих та слабоструктурованих даних є відсутність чіткої структурної схеми даних. Загальні підходи в аналітиці таких даних розглянуті у роботах [13, 17, 18, 19]. В [20] розглянуто деякі підходи в отриманні слабоструктурованої інформації із Веб-сервісів мережі Інтернет. В [21] розглянуто підходи в отриманні інформації з слабоструктурованих даних, зокрема текстів. В [22] розглянуто підходи в аналітиці слабоструктурованої бази часових рядів. В [23] розглянуто різні концепції та підходи в аналітиці великих даних. В [24] розглянуто методи та підходи в аналітиці текстів у фінансовій області. Актуальним напрямком є використання методів прогнозної аналітики у дослідженні інформаційних систем [25]. Одним із напрямків ефективного

використання прогнозної аналітики є управління мережами поставок [26, 27]. Підходи та методи прогнозної аналітики часто використовують в області управління людськими ресурсами [28]. Прогнозна аналітика широко використовується в аналізі інформації Вебу та соціальних медіа [29, 30, 31, 32, 33, 34]. Методи прогнозної аналітики є ефективними в аналізі фінансових ринків. Інформація масмедіа та соціальних мереж відображає настрої та наміри інвесторів і, відповідно, аналітику текстових повідомлень інформаційних потоків інформагенств та соціальних мереж можна використати в прогнозуванні економічних та фінансових процесів [35, 36, 37, 38, 39, 40, 41, 42, 43]. У [44] розглянуто бази знань інтелектуальних систем підтримки прийняття рішень, ядром яких є онтології предметних галузей та онтології задач. В [45] проаналізовано проблеми опрацювання даних з різнотипних джерел, побудовано формальну модель простору даних та уведено операції над ним, визначено особливості інтеграції даних із різнотипних джерел, побудовано схему інтеграції даних та засоби обміну даними. У [46] розглянуто основні принципи побудови та функціонування сховищ даних, показано коло задач, для яких необхідно використовувати сховища даних, уведено формальну модель простору даних як нову абстракцію керування даними. У [47] формалізовано характеристики якості консолідованих даних у просторах даних, уведено поняття корисності даних, розроблено архітектуру системи оцінювання якості різнотипних даних, уведено метамову для формування запитів користувачів до різнотипних джерел, розроблено структури даних для опису інформаційних продуктів та схеми метаданих. У [48] введено поняття корисності даних з джерел даних та концептуальне визначення якості консолідованих даних простору даних, розроблено метамови опису джерел даних та встановлення відповідності між їхніми структурами даних, розроблено архітектури підсистеми оцінки якості консолідованих даних. У [49] розглянуто задачу адаптивного навчання еволюційної нейронної мережі з ядерними функціями активації, що базуються на об'єднанні різних архетипів навчання. У [50] аналізуються нейронні мережі. У [51] розглянуто архітектуру та застосування нейронних мереж. У [50] розглянуто використання нео-фаззі нейрона як основного компонента нейронної мережі, показано архітектуру глибокої нео-фаззі

нейронної мережі а також алгоритм зворотнього поширення похибки для цієї архітектури з трикутною функцією приналежності для нео-фаззі нейрона, наведено основні переваги щодо застосування нео-фаззі нейрона, як основного компонента нейронної мережі. У [52] розглянуто використання байєсівських мереж у технологіях інтелектуального аналізу даних. У [53] викладено концептуальний підхід до моделювання фінансової стійкості підприємства, що полягає в оцінюванні фінансового стану компанії шляхом прогнозування ймовірного банкрутства на основі аналогій між показниками діяльності цієї компанії, підприємств-банкрутів і фінансово стабільних компаній. У [54] розглянуто побудову функцій прогнозування для стаціонарних процесів авторегресії та авторегресії з ковзним середнім, процесів з детермінованими та стохастичними трендами, гетероскедастичних та коінтегрованих процесів. У [55] розглянуто можливість ідентифікації часових рядів на основі ймовірнісних нейронних мереж, розглянуто модифіковані версії ймовірнісних нейронних мереж та особливості їх застосування. У [56] наведено систематизований виклад математичних основ і методів опису, побудови та застосування моделей знань у системах штучного інтелекту та підтримки прийняття рішень, розглянуто семантичні, фреймові та нейро-нечіткі мережі, продукційні та логічні моделі.

1.2 Інтелектуальний аналіз даних табличного типу

Розглянемо основні методи та підходи, які використовують в аналітиці даних табличного типу, коли дані можна представити у вигляді таблиці. У такій таблиці рядки представляють зразки даних, а стовпці - ознаки даних.

1.2.1 Лінійні стохастичні регресійні моделі

Розглянемо узагальнену регресійну модель [57]. Цільову змінну представимо у вигляді комбінації базисних функцій:

$$y(x, w) = w_0 + \sum_{j=1}^{M-1} w_j \Phi_j(x), \tag{1.1}$$

де $\Phi_j(x)$ - базисні функції, M - загальна кількість параметрів моделі. Параметр w_0 визначає зміщення. У найпростішому вигляді базисні функції визначені значеннями вхідних змінних x_i, однак, також можуть бути використані нелінійні за вхідними ознаками функції. У загальному, деякий процес можна розглядати як стохастичний. У цьому випадку цільову функцію можна представити як суму детермінованої функції $y(x, w)$ та гаусового шуму ε

$$t = y(x, w) + \varepsilon. \tag{1.2}$$

Запишемо умовну ймовірність для цільової функції

$$p(t \mid x, w, \beta) = N(t, y(x, w), \beta^{-1}) \tag{1.3}$$

у випадку гаусового шуму, математичне очікування для $p(tx, w, \beta)$ буде рівне $y(x, w)$
Відповідно до [57], визначимо коефіцієнти регресії

$$w = (\mathbf{\Phi}^{-1}\mathbf{\Phi}^T)\mathbf{\Phi}^T t, \tag{1.4}$$

де

$$\mathbf{\Phi} = \begin{bmatrix} \Phi_0(x_1) & \Phi_1(x_1) & \dots & \Phi_{M-1}(x_1) \\ \Phi_0(x_2) & \Phi_1(x_2) & \dots & \Phi_{M-1}(x_2) \\ \dots & \dots & \dots & \dots \\ \Phi_0(x_N) & \Phi_1(x_N) & \dots & \Phi_{M-1}(x_N) \end{bmatrix}. \tag{1.5}$$

Розв'язок у вигляді (1.4) для великих масивів може вимагати багато обчислювальних витрат. Одними із алгоритмів, які є зручними з точки зору об'ємів обчислень, є *On-Line* алгоритми, зокрема, алгоритми стохастичного градієнтного спуску. На кожному ітераційному кроці оновлюється значення вектора коефіцієнтів:

$$w^{(t+1)} = w^{(t)} - \eta \nabla E_n, \tag{1.6}$$

де E_n - функція похибок, η - параметр темпу навчання. Розглянемо випадок, коли функція похибок визначається сумою квадратів похибок

$$E_n(w) = \frac{1}{2}\sum_{n=1}^{N}(t - w^T\Phi(x_n))^2. \tag{1.7}$$

Тоді ітераційних процес визначення вектора коефіцієнтів регресії можна описати так [57] :

$$w^{(t+1)} = w^{(t)} - \eta(t_n - w^{(t)T}\Phi_n)\Phi_n. \tag{1.8}$$

Важливою проблемою у задачах регресійного аналізу є перенавчання (*overfitting*). Одним із шляхів зменшення такого ефекту є регуляризація, яку можна реалізувати додаванням регуляризаційного доданку у функцію похибок:

$$E = E_0(w) + \lambda E_w(w), \tag{1.9}$$

де λ - регуляризаційний коефіцієнт. Одним із варіантів регуляризаційного доданку може бути сума квадратів коефіцієнтів регресії [57]

$$E_w(w) = \frac{1}{2}w^T w. \tag{1.10}$$

У випадку функції похибок у вигляді суми квадратів похибок отримаємо

$$E_n(w) = \frac{1}{2}\sum_{n=1}^{N}(t - w^T\Phi(x_n))^2 + \frac{\lambda}{2}w^T w. \tag{1.11}$$

Прирівнюючи похідні $E_n(w)$ по коефіцієнтах w_i до нуля, отримаємо

$$w = (\lambda I + \mathbf{\Phi}^{-1}\mathbf{\Phi}^T)\mathbf{\Phi}^T t. \tag{1.12}$$

Часто використовують регуляризацію Lasso [57], для якої функцію похибок можна розглянути у вигляді

$$E_n(w) = \frac{1}{2}\sum_{n=1}^{N}(t - w^T\Phi(x_n))^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|. \tag{1.13}$$

Параметр регуляризації λ вибирають експериментально, мінімізуючи похибку на валідаційній вибірці.

1.2.2 Байєсівська лінійна регресія

Байєсівський підхід в аналітиці даних базується на теоремі Байєса і дає можливість отримати функції розподілів для параметрів моделі та для цільової змінної [58, 59, 60]. Розглянемо випадок гаусової статистики. Розподіл параметрів розглянемо у вигляді

$$p(w) = \mathcal{N}(w|m_0, S_0), \tag{1.14}$$

де m_0 – середні значення відповідних параметрів, S_0 - коваріації. Розглянемо теорему Байєса у вигляді

$$p(w|D) = \frac{p(D|w)p(w)}{p(D)}, \tag{1.15}$$

де $D = \{t_1, t_2...t_N\}$ – набір історичних даних аналізу, w – вектор параметрів моделі. Величину $p(D|w)$ називають функцією правдоподібності і вона визначає розподіл отриманих даних як функцію від вектора параметрів моделі. $p(w)$ – апріорна фунція розподілу параметрів моделі. Знаменник у виразі (1.15) можна розглянути у вигляді інтегралу

$$p(D) = \int p(D|w)p(w)dw. \tag{1.16}$$

В аналізі стохастичних процесів розглядають два підходи – частотний (*frequentists*) та байєсівський. У частотному підході параметри моделі є сталими, а похибки моделі визначаються на вибірках валідаційних даних. У байєсівському підході невизначеність виражена через розподіл параметрів моделі. Можна показати [57], що максимізація функції суми квадратів похибок є еквівалентною мінімізації функції суми квадратів похибок із квадратичним регуляризаційним доданком. Часто в реальних задачах потрібно враховувати складні розподіли історичних даних, зокрема, розподіли з "товстими хвостами", тобто такі розподіли, у яких рідко,

але зустрічаються випадки величин із екстремальними значеннями. Опис таких процесів за допомогою гаусової статистики не дає адекватну оцінку математичного сподівання та відповідних квантилів. Зокрема, важливо враховувати екстремальні значення при оцінці ризиків. Наприклад, величина VaR (Value at Risk) для процесів із екстремальними значеннями, яку визначають як 1% або 5% перцентиль, може суттєво відрізнятися для гаусової апроксимації розподілу даних та для апроксимації за допомогою розподілів із "товстими хвостами". Прикладом роподілу із "товстими хвостами" може бути розподіл Стьюдента (t-розподіл). Для аналізу процесів, які описуються негаусовою статистикою, часто використовують чисельні методи семплювання, зокрема, методи Монте-Карло. Одна з проблем в отриманні репрезентативного набору значень полягає у складнощах семплювання для розподілів із великими розмірностями. Одним із ефективних підходів у такому класі задач є методи Монте-Карло для марківських ланцюгів (MCMC алгоритми). Одним із варіантів цих методів є алгоритм Metropolis [61]. Простою та ефективною реалізацією алгоритмів Монте-Карло є семплювання Гібса [62, 63, 64, 65]. На кожному кроці ітерації семплювання Гібса здійснюється заміна однієї із змінних значенням змінної, отриманої з розподілу, за умови, що інші змінні є сталими. На наступній ітерації здійснюється заміна наступної змінної. Ітерації повторюються поки не здійсниться заміна всіх змінних. Наступний цикл ітерацій починається знову із заміни першої змінної. Схему алгоритму Гібса можна записати так:

$$
\begin{aligned}
&1. init\ \{z_i \mid i = 1, 2, ..., M\} \\
&2. for\ t = 1, 2, ..., T \\
&\quad z_1^{(t+1)} \sim p(z_1 \mid z_2^{(t)}, z_3^{(t)}, ..., z_M^{(t)}) \\
&\quad z_2^{(t+1)} \sim p(z_2 \mid z_1^{(t)}, z_3^{(t)}, ..., z_M^{(t)}) \\
&\quad z_j^{(t+1)} \sim p(z_j \mid z_1^{(t)}, ...z_{j-1}^{(t)}, z_{j+1}^{(t)}, ..., z_M^{(t)}) \\
&\quad z_M^{(t+1)} \sim p(z_M \mid z_1^{(t)}, z_2^{(t)}, ..., z_{M-1}^{(t)})
\end{aligned}
\tag{1.17}
$$

Одним із обмежень алгоритмів MCMC є те, що вони базуються на ідеї броунівського руху і відстань, яку можна подолати у просторі станів, визначається квадратним коренем із кількості кроків. Збільшення кількості кроків веде до збільшення відмов прийняття кандидатів векторів параметрів

до репрезентативної вибірки. Вирішити таку проблему можна за допомогою гібридних алгоритмів Монте-Карло, які базуються на рівняннях Гамільтона, які описують поведінку фізичних систем [66, 67, 68]. Розглянемо опис алгоритму семплювання на основі гамільтонової динаміки згідно з [57]. У класичній механіці динаміка описується другим законом Ньютона, який можна представити у вигляді диференційних рівнянь другого порядку. Таке рівняння можна представити у вигляді двох рівнянь уводячи додаткову змінну, яка визначає швидкість зміни вектора z

$$r_i = \frac{dz_i}{dt}. \tag{1.18}$$

Ця змінна є аналогом моменту руху у механіці. Отже, вектор z можна розглядати в об'єднаному фазовому просторі значень та моментів. Широкий клас функцій розподілу можна розглядати у вигляді

$$p(z) = \frac{1}{z_p} exp\left(-E(z)\right). \tag{1.19}$$

$E(z)$ – це аналог потенційної енергії, яка визначає стан z .Прискорення визначається зовнішньою силою, яку можна розглядати як від'ємний градієнт енергії

$$\frac{\partial r_i}{\partial t} = -\frac{\partial E(z)}{\partial z_i}. \tag{1.20}$$

Кінетичну енергію розглянемо як

$$K(r) = \frac{1}{2}\sum_i r_i^2. \tag{1.21}$$

Сумарну енергію запишемо як

$$H(z, r) = E(z) + K(r), \tag{1.22}$$

де $H(z,r)$ - гамільтоніан. Динаміку системи можна описати рівняннями Гамільтона

$$\begin{aligned} \frac{\partial z_i}{\partial t} &= -\frac{\partial H}{\partial r_i}, \\ \frac{\partial r_i}{\partial t} &= \frac{\partial H'}{\partial z_i}. \end{aligned} \tag{1.23}$$

Протягом еволюції системи значення гамільтоніана не змінюються. Згідно із теоремою Ліувілля, також залишається сталим об'єм у фазовому просторі змінних. Інтегруючи гамільтоніан протягом скінченного часового періоду, можна отримати значення зразків z без реалізації процесу випадкового блукання. На відміну від алгоритму *Metropolis*, гамільтоновий підхід дає можливість враховувати як розподіл параметрів, так і градієнт логарифму функції розподілу ймовірності. Алгоритм семплювання Hibrid Monte Carlo комбінує гамільтонову динаміку та алгоритм *Metropolis* і, таким чином, дає можливість уникнути похибок, зумовлених дискретизацією. Для реалізації гамільтонового підходу у семплюванні використовують схему дискретизації *leapfrog* [57] :

$$\begin{aligned} r_i(t+\epsilon/2) &= r_i(t) - \frac{\epsilon}{2}\frac{\partial E}{\partial z_i}\left(z(t)\right), \\ z_i(t+\epsilon) &= z_i(t) + \epsilon r_i(t+\epsilon/2), \\ r_i(t+\epsilon) &= r_i(t+\epsilon/2) - \frac{\epsilon}{2}\frac{\partial E}{\partial z_i}\left(z(t+\epsilon)\right). \end{aligned} \tag{1.24}$$

Похибка дискретизації буде прямувати до нуля зі зменшенням величини ϵ до нуля. Згідно із алгоритмом *Metropolis*, якщо (z,r) і (z',r') – стан-кандидат після leapfrog інтегрування, тоді цей стан приймається з імовірністю

$$min\left(1, exp\{(H(z,r) - H(z',r')\}\right).$$

Якщо алгоритм *leapfrog* буде працювати ефективно, тоді майже кожний кандидат буде прийматись, оскільки гамільтоніан є незмінним. Відмінності обчислених гамільтоніанів можуть бути зумовлені лише похибками чисельних методів апроксимації алгоритму *leapfrog*. Гамільтонівський підхід

реалізовано, зокрема, у системі ймовірнісного моделювання *Stan* [59, 68].

1.2.3 Алгоритми машинного навчання на основі дерев рішень

В аналітиці даних часто використовують алгоритми машинного навчання на основі дерев рішень [69], наприклад Random Forest [70], Gradiend Boosting Machine [71, 72]. Одна із особливостей алгоритмів на основі дерев рішень полягає в їхній нечутливості до монотонних перетворень ознак. Відіграє роль порядок значень, які набувають ознаки, тобто, важливим є те, що деяке значення є більшим за інше і не важливо наскільки. Така особливість є важливою, коли прогнозні ознаки мають різну природу. Як приклад алгоритмів на деревах розглянемо алгоритм градієнтного бустінгу. Суть такого алгоритму полягає в об'єднанні слабких моделей в одну сильну прогнозну модель. На кожному ітераційному процесі навчання покращуються прогнозні результати, які отримані на попередньому кроці. Розглянемо алгоритм бустингу відповідно до [72] . Нехай існує деякий вектор незалежних змінних

$$\boldsymbol{x} = \{x_1, x_2, ..., x_n\}, \tag{1.25}$$

якому відповідає значення цільової змінної y. Набір зразків

$$\{y_i, \boldsymbol{x_i}\}_1^N$$

описує наявні дані. Необхідно знайти функцію $F^*(x)$, яка описує відображення x на множину y, так, щоб математичне очікування деякої функції втрат $\Psi(y, F(x))$ було мінімізоване.

$$F^*(x) = argmin_{F(\boldsymbol{x})} = E_{y,\boldsymbol{x}}\Psi(y, F(x)). \tag{1.26}$$

Алгоритм бустингу апроксимує $F^*(\boldsymbol{x})$ так:

$$F(x) = \sum_{m=0}^{M} \beta_m \left(h(\boldsymbol{x}, \boldsymbol{a}_m)\right), \tag{1.27}$$

де $h(\boldsymbol{x}, \boldsymbol{a}_m)$ – базові навчальні моделі, які є простими функціями від $\boldsymbol{x}$ із параметрами $\boldsymbol{a} = \{a_1, a_2, ...\}$.

Параметри β_m, a_m визначають так:

$$(\beta_m, a_m) = argmin_{\beta,a} \sum_{i=1}^{N} \Psi(y_i, F_{m-1}(x_i) + \beta h(\boldsymbol{x}, \boldsymbol{a}_m)). \tag{1.28}$$

Для функції $F_m(x)$ запишемо

$$F_m(x) = F_{m-1} + \beta_m h(x, a_m). \tag{1.29}$$

Наближений розв'язок рівняння для довільної функції $\Psi(y, F(x)$ можна здійснити за два кроки. На першому кроці

$$a_m = argmin_{\boldsymbol{a},\rho} \sum_{i=1}^{N} \left[\tilde{y}_{im} - \rho h(\boldsymbol{x}_i, \boldsymbol{a})\right]^2, \tag{1.30}$$

$$\tilde{y}_{im} = - \left[\frac{\partial \Psi(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(X)=F_{m-1}(x)}. \tag{1.31}$$

Беручи $h(\boldsymbol{x}, \boldsymbol{a}_m)$, оптимальне значення коефіцієнта β_m визначають як

$$\beta_m = argmin_{\beta} \sum_{i=1}^{N} \Psi(y_i, F_{m-1}(\overline{x_i})) + \beta h(\boldsymbol{x}_i, \boldsymbol{a}_m). \tag{1.32}$$

У випадку використання дерев рішень як базових класифікаторів, $h(\boldsymbol{x}, \boldsymbol{a}) \in L$ – нодове регресійне дерево. На кожній ітерації регресійне дерево розділяє x – простір у L – роз'єднаних зон $\{R_{lm}\}_{l=1}^{L}$ і прогнозує значення у кожній із цих зон

$$\begin{aligned} h(x, \{R_{lm}\}_1^L) &= \sum_{l=1}^{L} \tilde{y}_{lm} 1(x \in R_{lm}), \\ \overline{y}_{lm} &= mean_{x_i \in R_{lm}}(\tilde{y}_{im}). \end{aligned} \tag{1.33}$$

Величина $\overline{y}_{lm}$ визначає середнє значення $\tilde{y}_{im}$ з (1.31) для кожної зони R_{lm}. Оскільки дерево рішень (1.33) прогнозує значення констант $\overline{y}_{lm}$ на кожному

листку R_{lm}, розв'язок (1.32) зводиться до

$$\gamma = argmin_{\gamma} \sum_{x_i \in R_{lm}} \Psi(y_i, F_{m-1}(x_i) + \gamma). \tag{1.34}$$

Тоді на кожній ітерації отримаємо основне значення для функції $F_m(x)$ у відповідній зоні

$$F_m(x) = F_{m-1}(x) + \nu\gamma_{lm}1(x \in R_{lm}). \tag{1.35}$$

Параметр $0 < \nu \leqslant 1$ визначає темп навчання класифікатора. Найкраще значення ν визначають емпіричним шляхом.

Дерева рішень лежать,зокрема, в основі алгоритмів машинного навчання XGBoost [73] та LightGBM [74], які широко використовуються у класифікаційних та регресійних задачах інтелектуального аналізу даних.

1.2.4 Методи побудови ансамблів прогнозних моделей

Прогнозні моделі часто об'єднують в ансамблі для того, щоб отримати вищу точність та стабільність у прогнозуванні. Використання ансамблевих методів у класифікаційних задачах розглянуто у [75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86]. Технології стекінгу часто використовують для отримання вищої генералізованої точності [75, 76, 84]. Ідея стекінгу полягає в об'єднанні прогнозних моделей у багаторівневий ансамбль. На першому рівні отримують прогнози за допомогою моделей машинного навчання, другий рівень – мета рівень, на якому за допомогою деякої моделі об'єднують результати першого рівня у кінцевий результат. Для отримання тренувальної вибірки даних для мета рівня часто використовують кросвалідаційний підхід, при якому тренувальну вибірку розбивають випадковим чином на декілька підвибірок. Далі беруть одну підвибірку для прогнозування цільової змінної, а решту – для тренування прогнозної моделі. Таку процедуру повторюють для прогнозування кожної підвибірки. В результаті отримують прогнозовані значення для всієї вибірки за умови, що зразки даних, для яких було знайдено прогнозоване значення, не були використані для тренування моделі на тому ж кроці, на якому прогнозувалась цільова змінна для підвибірки цих зразків. У випадку нестаціонарних даних валідаційна підвибірка для прогнозування

вибирається за часовим поділом так, щоб дані для валідації знаходились на часовій осі після зразків даних для тренування. Отримані прогнозовані дані на валідаційній підвибірці використовують як незалежні змінні для прогнозної моделі другого мета рівня стекінгового ансамблю. Цільова змінна на мета рівні дорівнює цільовій змінній зразків даних валідаційної підвибірки. Кросвалідаційний підхід для отримання навчальної вибірки на мета рівні стекінгового ансамблю використовують з метою уникнення проблем, які можуть бути зумовлені ефектом перенавчання. Цей ефект виникає коли у навчанні моделі використовують підвибірки зразків даних, для яких здійснюється прогнозування. У такому випадку точність прогнозування для таких підвибірок може бути завищеною, а для нових даних - меншою. Стекінгові підходи часто використовують у прогнозних моделях на різних змаганнях спеціалістів із прогнозної аналітики, зокрема на платформі Kaggle [87]. Такі підходи дають можливість покращити результати прогнозування на заданій вибірці даних. Ефективність стекінгових ансамблів також визначається набором параметрів мета-моделі. Як мета моделі часто використовують лінійні регресійні моделі та моделі на основі машинного навчання, зокрема, нейронні мережі.

1.2.5 Штучні нейронні мережі

Штучні нейронні мережі часто використовують у прогнозній аналітиці. Базові принципи нейронних мереж розглянуто у працях [88, 89, 90, 91, 92, 93, 94, 95] Нейронну мережу можна розглядати як апроксимаційну функцію, яка відображає вхідні дані $\boldsymbol{x}$ на множину значень $\boldsymbol{y}$ [94]. Множина $\boldsymbol{y}$ може бути категоріальною або числовою. У випадку категоріальних значень ми отримаємо класифікацію значень, а у випадку числових – регресію. У загальному випадку нейронну мережу можна описати формулою

$$\boldsymbol{y} = f(\boldsymbol{x}, \boldsymbol{\Theta}), \tag{1.36}$$

де $\boldsymbol{y}$ – вихідні значення, $\boldsymbol{x}$ – вхідні значення, $\boldsymbol{\Theta}$ – параметри моделі, які відповідають найкращій апроксимації даних. Базовими мережами є мережі прямого поширення, у яких інформація поширюється від початкового

шару до вихідного без зворотніх зв'язків. За наявності зворотніх зв'язків така нейронна мережа називається рекурентною. Нейромережі прямого поширення можна зобразити за допомогою орієнтованого ациклічного графу. Нейромережу можна розглядати у вигляді ланцюга функцій

$$f(\boldsymbol{x}) = f^{(n)}(f^{(n-1)}(...(f^{(2)}(f^{(1)}(\boldsymbol{x}))...)). \tag{1.37}$$

Функцію $f^{(1)}(\boldsymbol{x})$ називають першим шаром, який є вхідним шаром, $f^{(2)}(\boldsymbol{x})$ – другим шаром, $f^{(n)}(\boldsymbol{x})$ – n-м шаром. Якщо $f^{(n)}(\boldsymbol{x})$ – кінцевий шар, то його також називають вихідним шаром. Шари, які описуються функціями $f^{(2)}(\boldsymbol{x}), ..., f^{(n-1)}(\boldsymbol{x})$ називають прихованими шарами. Важливим ефектом нейронних мереж є генералізація, яка полягає у ефективному відображенні вхідних $\boldsymbol{x}$ даних на множину вихідних значень $\boldsymbol{y}$ для зразків даних, які не були використані у процесі оптимізації нейромережі та знаходженні оптимальних нейромережевих параметрів $\boldsymbol{\Theta}$, які описані у формулі (1.36). Для того, щоб враховувати нелінійні зв'язки між вхідними змінними, необхідно, щоб функції $f^{(i)}(\boldsymbol{x})$ також були нелінійними. Нелінійність нейронних мереж зумовлює ітеративний процес пошуку оптимальних параметрів, який часто базується на градієнтних методах, зокрема: на стохастичному градієнтному спуску. Для пошуку оптимальних параметрів нейронної мережі необхідно визначити цільову функцію для пошуку глобального екстремуму, зокрема мінімуму. Модель визначає деякий розподіл $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\Theta})$ і для пошуку оптимальних параметрів можна використати пошук максимуму ймовірності. Часто, як цільову функцію розглядають крос-ентропію між історичними даними та прогнозами на основі моделі. У загальному, виходячи з принципу максимуму правдоподібності цільову функцію можна визначити як [94]

$$J(\boldsymbol{\Theta}) = -\mathbb{E}_{x,y\sim\hat{p}_{data}} log(p_{model}(\boldsymbol{y}|\boldsymbol{x})). \tag{1.38}$$

У випадку нормального розподілу

$$p_{model}(\boldsymbol{y}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{y}; f(\boldsymbol{x}, \boldsymbol{\Theta})), \tag{1.39}$$

$$J(\boldsymbol{\Theta}) = \frac{1}{2}\mathbb{E}_{x,y\sim\hat{p}_{data}}\|\boldsymbol{y} - f(\boldsymbol{x}, \boldsymbol{\Theta})\|^2 + const. \tag{1.40}$$

Вибір цільової функції визначається типом вихідних величин. У багатьох випадках використовують крос-ентропію між розподілами даних та моделі. Вихідні та приховані елементи нейромережі можуть бути однакового типу, тобто, здійснювати такі ж перетворення вхідних даних. Елементи вихідного шару можуть також відрізнятися від елементів прихованих шарів і здійснювати фінальні перетворення даних, які поступають від прихованих шарів. Розглянемо прихований шар нейромережі, який визначається функцією:

$$h = f(\boldsymbol{x}, \boldsymbol{\Theta}). \tag{1.41}$$

Розглянемо функцію вихідного шару у вигляді лінійної функції [94]

$$\boldsymbol{y} = \boldsymbol{W^T h} + \boldsymbol{b}. \tag{1.42}$$

У випадку наближення гаусового розподілу для умовної ймовірності $p(\boldsymbol{y}, \boldsymbol{x})$, максимізація логарифмічної функції правдоподібності еквівалентна мінімізації середньо-квадратичної похибки. Елементи прихованих шарів можуть бути описані деякою лінійною трансформацією вхідних даних, наприклад, $z = \boldsymbol{W^T x} + \boldsymbol{b}$, на яку накладається деяка нелінійна активаційна функція $g(z)$. Часто як активаційну функцію використовують елемент $ReLU$ (*Rectified Linear Unit*) [96], який можна описати як

$$g(z) = max\{0, z\}. \tag{1.43}$$

Тоді для елемента прихованого шару можна записати

$$\boldsymbol{h} = g\left(\boldsymbol{W^T x} + \boldsymbol{b}\right). \tag{1.44}$$

Нейронну мережу часто розглядають у вигляді послідовного об'єднання шарів елементів (нейронів), де елементи різних сусідніх шарів з'єднані між

собою. Модель такої мережі можна розглянути у вигляді

$$\begin{aligned}
&\boldsymbol{h}^{(1)} = g^{(1)}\left(\boldsymbol{W}^{(1)T}\boldsymbol{x} + \boldsymbol{b}^{(1)}\right),\\
&\boldsymbol{h}^{(2)} = g^{(1)}\left(\boldsymbol{W}^{(2)T}\boldsymbol{h}^{(1)} + \boldsymbol{b}^{(2)}\right),\\
&...\\
&\boldsymbol{h}^{(i)} = g^{(1)}\left(\boldsymbol{W}^{(i)T}\boldsymbol{h}^{(i-1)} + \boldsymbol{b}^{(i)}\right),\\
&...\\
&\boldsymbol{h}^{(n)} = g^{(1)}\left(\boldsymbol{W}^{(n)T}\boldsymbol{h}^{(n-1)} + \boldsymbol{b}^{(n)}\right).
\end{aligned} \tag{1.45}$$

Оптимальну кількість шарів та елементів у них можна визначити експериментально, виходячи з оптимізації параметрів моделі на валідаційному сеті даних. Теоретичні основи нейромереж базуються, зокрема, на універсальній апроксимаційній теоремі [97, 98], яка говорить, що маючи лінійний вихідний шар та прихований шар із нелінійністю, зокрема, із сигмоїдною активаційною функцією, можна здійснити апроксимацію довільної функції, визначеної в скінченно вимірному просторі $\mathbb{R}^n$. Ці результати отримані для випадку функцій сигмоїдного типу. Універсальну апроксимаційну теорему також розглянуто для активаційних функцій типу $ReLU$ [96], які широко використовуються у сучасних архітектурах нейронних мереж. Для оптимізації параметрів нейромереж використовують алгоритм зворотнього поширення [99]. Складні обчислення можна представити у вигляді обчислювального графу. Граф складається із змінних і з базових операцій. Вхідними даними операцій є одна або більше змінних, а вихідним результатом є одна змінна. Змінні можуть бути векторного типу. Змінні позначають вершинами, які з'єднані орієнтованими ребрами. Ребра, які сходяться до вершини відображають деяку операцію. Нехай є функції

$$\boldsymbol{y} = f(\boldsymbol{x}), \boldsymbol{z} = f(\boldsymbol{y}). \tag{1.46}$$

Для похідної від $\boldsymbol{z}$ по $\boldsymbol{x}$ можна записати

$$\nabla_{\boldsymbol{x}}\boldsymbol{z} = \left(\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\right)^T \nabla_{\boldsymbol{y}}\boldsymbol{z}, \tag{1.47}$$

де $\dfrac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}$ – Якобіан. Формулу (1.47) можна узагальнити для випадку тензорів. Градієнт для довільних нодів можна розглянути як

$$\frac{\partial U^{(n)}}{\partial U^{(j)}} = \sum_{i,j} \frac{\partial U^{(n)}}{\partial U^{(j)}} \frac{\partial U^{(n)}}{\partial U^{(j)}}. \tag{1.48}$$

Алгоритм backprop розраховує на ребрі між вершинами $U^{(j)}$ $iU^{(i)}$ похідну $\frac{\partial U^{(j)}}{\partial U^{(i)}}$. Ефективність використання активаційних функцій типу ReLU розглянуто у [100, 101]. Одна із важливих проблем у машинному навчанні полягає в отриманні моделі, яка добре прогнозує дані на валідаційній та тестовій вибірках, які не були використані у процесі навчання моделі. Для цього використовують різні регуляризаційні стратегії. У загальному регуляризаційну функцію можна розглянути у вигляді

$$\boldsymbol{J}\left(\boldsymbol{\Theta}, \boldsymbol{X}, \boldsymbol{y}\right) = \boldsymbol{j}\left(\boldsymbol{\Theta}, \boldsymbol{X}, \boldsymbol{y}\right) + \alpha\Omega(\boldsymbol{\Theta}), \tag{1.49}$$

де α – деякий регуляризаційний параметр. Для регуляризації часто використовують $L1$ $L2$ регуляризації. Як регуляризаційний підхід також часто використовують ранню зупинку оптимізаційного ітераційного процесу. Рання зупинка полягає у тому, що процес навчання зупиняється на деякому етапі, коли на певній кількості послідовнизх кроків не покращується цільова функція на валідаційному сеті. На деякому етапі навчання похибка на валідаційному сеті може почати збільшуватись, такий ефект називають перенавчанням (overfitting). Як регуляризхацію також використовують метод Dropout [102, 103], який полягає у видаленні елементів у прихованих шарах чи у видаленні зв'язків між елементами різних шарів.

Цільову функцію в алгоритмі навчання можна розглянути як [94]

$$J(\boldsymbol{\Theta}) = \mathbb{E}_{x,y\sim\hat{p}_{data}} L\left(f(\boldsymbol{x}, \boldsymbol{\Theta}), \boldsymbol{y}\right), \tag{1.50}$$

де L - функція втрат, $f(\boldsymbol{x}, \boldsymbol{\Theta})$ - прогнозований результат, $\hat{p}_{data}$ - емпіричний

розподіл історичних даних. У випадку емпіричних історичних даних

$$\mathbb{E}_{x,y\sim\hat{p}_{data}} = \frac{1}{m}\sum_{i=1}^{m} L\left(f(\boldsymbol{x}^{(i)}, \boldsymbol{\Theta}), y^{(i)}\right). \tag{1.51}$$

Використання стратегій, які базуються лише на оптимізації функції (1.51) може привести до перенавчання, тому додатково необхідно використовувати різні регуляризаційні підходи. В залежності від кількості історичних даних, які використовують у навчанні моделі, розрізняють детерміністичні, або пакетні (*batch*) методи, стохастичні, або *on-line* методи та мініпакетні (*minibatch*) стохастичні методи [94]. Пакетні методи використовують усі історичні дані для навчання моделі. *On-line* методи на кожному кроці використовують лише один зразок даних. Мініпакетні або стохастичні методи використовують на кожному кроці набір даних заданого розміру. Розмір мініпакету є важливим параметром і його можна оптимізувати на валідаційному сеті. Часто, коли говорять про стохастичні методи оптимізації алгоритму машинного навчання, мають на увазі мініпакетні стохастичні методи. Малий розмір пакетів може давати кращий ефект генералізації [104]. Важливим моментом для мініпакетних методів є те, що дані у мініпакетах мають бути вибрані випадково із вибірки історичних даних. Стохастичні методи градієнтного спуску мінімізують генералізовану похибку.

1.2.6 Методи глибокого Q-навчання

Машинне навчання з учителем можна розглядати як різновид пасивного навчання з використанням історичних даних. З іншої сторони, навчання з підкріпленням (Reinforcement Learning) дозволяє знаходити послідовності оптимізованих дій безпосередньо без історичних даних. У такому підході є середовище та агент навчання, який взаємодіє з середовищем. У результаті кожної взаємодії навчальний агент отримує винагороду. Таке навчання можна розглядати як різновид активного навчання. Мета навчання з підкріпленням полягає у пошуку такої послідовності дій, яка дозволить досягти максимальної середньої сукупної винагороди за епізодами взаємодій агент-середовище. Існують підходи,

засновані на стратегії (policy) та вільні від стратегії. Стратегія може бути описана параметризованою функцією розподілу для станів та дій. Параметри цих розподілів можна знайти, використовуючи градієнтні методи, де на кожній ітерації обчислюється градієнт цільової функції. Q-навчання можна розглядати як підхід без стратегії, яке базується на рівнянні Беллмана [105, 106, 107]. На кожній ітерації здійснюється оновлення Q-таблиці, де рядки представляють стани, а стовпці - дії. У випадку безперервної дії просторову Q-таблицю можна апроксимувати за допомогою нейронної мережі з використанням підходу DQN [106, 107]. Основні принципи навчання із підкріпленням можна знайти в [105]. В [106, 107] вивчено використання глибокого (глибинного) навчання у Q-навчанні. У [108] розглянуто динамічне ціноутворення в режимі реального часу в нестаціонарному стані середовища. У статті розглянуто проблему встановлення цінової політики, яка максимізує дохід від продажу певного товарного запасу у встановлений термін. В [109, 110, 111] розглянуто різні підходи на основі навчання підкріплення для динамічного ціноутворення. У статті [112] запропоновано адаптивні моделі управління запасами для ланцюга поставок, що складається з одного постачальника та декількох роздрібних продавців. У статті [113] описано використання методів навчання з підкріпленням у задачі визначення динамічних цін в електронних продажах. У статтях [114, 115] розглядаються питання використання навчання з підкріпленням для фінансової аналітики. У статті [116] розглядається підхід до оптимізації структури нейронної мережі та підвищення точності моделювання. У статті [115] розглянуто безмодельний підхід навчання з підкріпленням з використанням глибокого навчання для розв'язку задач управління портфельними інвестиціями. У [117], розглянуто глибоке навчання з підкріпленням у великих дискретних просторах дій інтелектуального агента. Метою Q-навчання є максимізація майбутньої сукупної винагороди [105, 106, 107]. Для навчання мережі Q-навчання часто використовують алгоритм градієнтного спуску. Щоб усунути вплив між послідовностями у даних та нестаціонарним розподілом, може бути використаний механізм reply [118]. Цей підхід полягає у випадковій вибірці попередніх даних, які представляють стани та дії. Це дає можливість усереднити розподіл даних, які описують попередню поведінку агента. Мета

агента полягає у виборі стратегії послідовних дій, яка максимізує майбутні винагороди [107]. Оптимальну функцією дія-значення (action-value) можна розглянути так:

$$Q^*(s,a) = \mathbb{E}_{s'\sim\varepsilon}\left[r + \gamma \max_{a'} Q(s',a') \mid s,a\right], \tag{1.52}$$

де r – винагорода, s – стан, a – дія, s', a' – можливі стани та дії на наступному часовому кроці. Для апроксимаційної оцінки функції $Q^*(s,a)$ можна використати ітераціний процес із рівнянням Беллмана:

$$Q_{i+1}(s,a) = \mathbb{E}\left[r + \gamma \max_{a'} Q_i(s',a') \mid s,a\right]. \tag{1.53}$$

Основна проблема такого підходу полягає в тому, що не відбувається узагальнення виявлених патернів взаємодії агент-середовище, оскільки $Q(s,a)$ оцінюється на кожному окремому кроці. Для покращення узагальнення можна використовувати функцію наближення для $Q^*(s,a)$. Для цього можна використовувати нейромережу. Параметри такої глибокої Q-мережі можна знайти за допомогою градієнтних методів мінімізації функції втрат

$$L_i(\theta_i) = \mathbb{E}_{s,a\sim\rho}\left[(y_i - Q(s,a;\theta_i))^2\right], \tag{1.54}$$

$$y_i = \mathbb{E}_{s'\sim\varepsilon}\left[r + \gamma \max_{a'} Q(s',a',\theta_{i-1}) \mid s,a\right], \tag{1.55}$$

де ρ – поведінковий розподіл, θ – вагові коефіцієнти Q-мережі. Підхід на основі відтворення досвіду (experience reply) [118] ефективно використовується у глибоких Q-мережах [107, 106]. У такому підході дії агента та стани зберігаються у пам'яті відтворення досвіду на кожному кроці у вигляді кортежів $e_t = (s_t, a_t, r_t, s_{t+1})$. Кортежі e_t зберігаються у наборі даних $D_t = \{e_1, ...e_t\}$. На кожному кроці оновлень алгоритм Q-навчання отримує зразки e_t з пам'яті відтворення за допомогою рівномірної випадкової вибірки $e_t \sim U(D)$ [107]. На наступному кроці відтворення досвіду алгоритм Q-навчання оновлює ваги θ. Далі на наступному кроці агент вибирає оптимальну дію, використовуючи ε -стратегію (ε - greedy policy) [107]. Міні-пакети даних формуються на кожній ітерації для

оновлення ваг Q-мережі. Міні-пакети вибирають випадковим чином. Такий підхід забезпечує генералізацію апроксимації на основі даних по взаємодії агент-середовище. Завдяки підходу на основі відтворення досвіду, розподіл поведінки усереднюється за багатьма попередніми станами та діями агента, що забезпечує збіжність процесу ітерації. Один підходів, які широко використовуються для Q-мережі полягає у розгляді станів агента як вхідних параметрів для аналізу, а результатами є Q-значення для кожної окремої дії агента [107].

1.2.7 Генетичні алгоритми

Генетичні алгоритми використовують у широкому класі оптимізаційних задач, які полягають у пошуку набору вхідних параметрів, що мінімізують деяку цільову функцію. Наприклад, генетичний алгоритм може бути використано в оптимізації метапараметрів та набору ознак у класифікаційних задачах на основі методів машинного навчання. Опис генетичних алгоритмів можна знайти у роботах [119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129]. Ідея генетичних алгоритмів полягає у використанні основних положень еволюційної теорії Дарвіна, зокрема, принципу природнього відбору та спадкової мінливості у розв'язку оптимізаційних задач. Розглянемо основні положення генетичних алгоритмів у контексті задачі пошуку оптимального семантичного базису для інтелектуального аналізу текстових документів, зокрема, на основі класифіканційних алгоритмів. Набір із вхідних параметрів називають хромосомою або особою. У простому випадку особа утворена на основі однієї хромосоми. Сукупність хромосом утворює популяцію. У генетичних алгоритмах базовими є оператор кросоверу для рекомбінації хромосом, оператор мутації, оператор відбору хромосом у наступне покоління. Оператор кросоверу може бути одноточковим, N-точковим та розсіяним. У одноточковому кросовері у послідовності генів вибирають точку розриву. Вибір точки розриву визначають випадковим чином із заданою функцією розподілу. Далі обмінюють ділянки генів у батьківських генах. У N-точковому кросовері існує N точок поділу хромосоми на ділянки генів, якими обмінюються батьківські хромосоми. У результаті обміну ділянок двох батьківских хромосом утворюють дві дочірні хромосоми нової популяції.

В операторі розсіяного або однорідного кросоверу (scatered crossover, uniform crossover) використовують бінарний вектор, який відіграє роль маски обміну генами. Розподіл бінарних значень у такій масці визначають заданим розподілом, зокрема, вибирають рівномірний розподіл. Розмір такого бінарного вектора дорівнює розміру батьківських хромосом. У дочірню хромосому попадає ген першої батьківської хромосоми, якщо значення складової бінарного вектора-маски, яке знаходиться на тому ж порядковому місці дорівнює «0». Якщо значення відповідної складової бінарного вектора дорівнює «1», тоді дочірній хромосомі надають відповідний цьому місцю ген другої батьківської хромосоми. Оператор мутації використовують для зміни окремих генів у новостворених хромосомах. Ці зміни можуть відбуватися в одній або декількох заданих точках хромосоми. Ймовірність мутації задають деякою функцією розподілу, яка визначена характером та умовами задачі. Використання оператора мутації зумовлене необхідністю виведення популяції з локального мінімуму цільової функції для задач з існуванням локальних та глобальних мінімумів. Для формування нової популяції використовують оператори відбору хромосом. При використанні відбору відсіканням формують нову популяцію з батьківських та дочірніх хромосом, які випадково відбирають з імовірністю, що визначається значенням цільової функції. Причому у відборі беруть участь ті хромосоми, у яких значення цільової функції менше за визначений поріг. Вибір здійснюють до тих пір, поки не отримають нову популяцію з такою ж кількістю хромосом як і у попередньої популяції. Очевидно, що деякі хромосоми можуть увійти у нову популяцію декілька разів. Також визначають деяку кількість хромосом із значенням цільової функції менше порогу, які можуть увійти у нову популяцію. При елітарному відборі, задають процент батьківських та дочірних хромосом із найвищим значенням цільової функції, які увійдуть у нову популяцію без генетичних змін. При використанні такого підходу у кожній популяції буде знаходитися сукупність елітарних хромосом, які є найкращими на заданий момент розв'язку. Коли будуть знайдені кращі хромосоми у наступних популяціях, тоді вже вони стануть елітарними, а попередні елітарні хромосоми стануть звичайними. Часто різні методи відбору хромосом об'єднують у комбінований оператор відбору. У

деяких алгоритмах використовують правило репродукції Холанда за яким хромосоми із значенням цільової функції вище середнього значення копіюють у наступну популяцію, а хромосоми із значенням цільової функції, яке є меншим за середнє значення – видаляють [127]. Класичний генетичний алгоритм містить такі кроки:

1. Утворюють початкову популяцію із n хромосом.
2. Для кожної хромосоми визначають цільову функцію.
3. На основі заданого правила відбору вибирають дві батьківські хромосоми, на основі яких буде утворена нова дочірня хромосома для наступної популяції.
4. До відібраних батьківських пар застосовують оператор кросовера, за допомогою якого утворюють нову дочірню хромосому.
5. Здійснюють мутацію нащадків із деякою заданою ймовірністю.
6. Повторюють кроки 3-5, доки не буде згенерована нова популяція n хромосом.
7. Кроки 2-6 повторюють до тих пір, поки не будуть виконуватись умови зупинки алгоритму. Такою умовою може бути, наприклад, задане значення цільової функції, або максимальна кількість ітерацій.

У дискретній оптимізації за допомогою генетичних алгоритмів кількість кроків, необхідних для пошуку оптимальних наборів вхідних параметрів, часто є поліноміально меншою у порівнянні із перебором можливих варіантів. Це пов'язано з наявністю деяких ділянок у хромосомах, які чимось подібні поведінкою на гени і які сукупно вносять оптимізаційний вклад у цільову функцію. Тобто, вхідні параметри розглядають деякими групами (генами), якими обмінюються хромосоми за допомогою оператора кросовера, що суттєво зменшує кількість комбінацій параметрів в оптимізаційному аналізі. Актуальною задачею для генетичних алгоритмів є задача оптимізації метапараметрів та набору ознак у задачах класифікації даних, зокрема, текстових документів.

1.3 Інтелектуальний аналіз текстових даних

1.3.1 Семантичні концепції в аналізі текстових даних

Інтелектуальний аналіз текстових масивів та комп'ютерна обробка природньої мови (NLP, Natural Language Processing) є перспективними напрямками сучасних інформаційних технологій. Особливим стимулом розвитку методів інтелектуального аналізу текстів є значний ріст слабоструктурованої інформації текстового типу у мережі Інтернет. Сучасний аналіз текстової інформації поряд із традиційними статистичними методами вимагає розвитку нових ефективних методів семантичного аналізу із заглибленням у зміст інформації, використовуючи методи машинного навчання, предметні онтології та семантичні мережі. Одним із прикладів ієрархічно-організованої семантичної мережі можна розглядати систему WordNet, яку розроблено у Прінстонському університеті [130, 131, 132, 133]. Ця система побудована на основі експертного лексикографічного аналізу семантичних структурних зв'язків, які відображають денотативні та конотативні характеристики лексемного складу словника. Глибина зв'язків у такій системі визначається експертною оцінкою лексемних комбінацій в текстових масивах і обмежується науковим досвідом експертів та об'ємом проаналізованого матеріалу. Лексемний склад в цій системі організований у вигляді синсетів, під якими розуміють набори лексем синонімічного ряду, які є взаємозамінними у заданих контекстах. Бази даних WordNet створені експертами-лексикографами. Іменники та дієслова згруповані відповідно до семантичних полів [134]. Семантичні поля у мережі WordNet представлені лексикографічними файлами, назви яких відображають основні семантичні значення лексем, які входять у склад цих файлів. Семантичні поля іменників складаються із 26 лексикографічних файлів: *noun.Tops, noun.act, noun.animal, noun.artifact, noun.attribute, noun.body, noun.cognition, noun.communication, noun.event, noun.feeling, noun.food, noun.group, noun.location, noun.motive, noun.object, noun.person, noun.phenomenon, noun.plant, noun.possession, noun.process, noun.quantity, noun.relation, noun.shape, noun.state, noun.substance, noun.time, verb.body.* Семантичні поля дієслів містять 15 лексикографічних файлів: *verb.change,*

verb.cognition, verb.communication, verb.competition, verb.consumption, verb.contact, verb.creation, verb.emotion, verb.motion, verb.perception, verb.possession, verb.social, verb.stative, verb.weather. У роботі [135] розглянуто використання системи WordNet для обробки текстових масивів. В [136] розглянуто структуру WordNet як семантичну мережу і проаналізовано метрику для вимірювання подібності між синсетами. Подібність між запитом та документом розглядається як подібність між синсетами у запиті та синсетами у документі. Одна із основних ідей у визначенні змісту лексем полягає у тому, що множина лексем, які зустрічаються одночасно в одному контексті, буде визначати відповідне значення для кожної з лексем, навіть якщо окремі лексеми є багатозначними. У дистрибутивному аналізі розглядають лексеми у їхньому співвідношенні з контекстуальними властивостями [137]. Припускають [137], що іменники, для яких розподіли в тексті подібні, будуть мати подібні значення. У роботі [137] розглянуто ймовірнісну формалізацію таксономії іменників у WordNet. Семантична структурна огранізація лексемного складу словника може бути використана у відповідних алгоритмах класифікації та кластеризації текстових об'єктів з точки зору зменшення розмірності задач аналізу та виявлення нових семантичних зв'язків в онтології предметної області, до якої відносять аналізований масив текстів.

Розглянемо лексикографічні концепції лексемних полів, які використовують у лінгвістиці. Семантичні групування слів відображають системність лексики. В основі визначення семантичних полів лежить лексико-семантична парадигма, під якою розуміють множину лексем, які об'єднані сукупністю семантичних ознак. Відмінність лексем у межах однієї парадигми визначається уточнювальними диференційними ознаками. Парадигми можуть бути одно- та багаторанговими. Ранги парадигми визначають структуру ієрархії лексемного об'єднання. Ядро семантичного поля утворюють лексеми, домінуюче значення яких визначають основними ознаками семантичного поля. Периферію семантичного поля утворюють лексеми, які містять основні поняття семантичного поля опосередковано через низку диференційних ознак, які мають відношення до основного поняття, яке утворює семантичне поле [138]. Одні і ті ж множини лексем

називають лексико-семантичними групами, семантичними полями та синонімічними рядами [139]. Під семантичним полем розуміють таку множину лексем, які об'єднані певним спільним поняттям [140, 138]. Прикладом семантичних полів може бути поле руху, поле комунікації, поле сприйняття тощо. Характерною особливістю семантичних лексемних полів є те, що деякі з багатозначних лексем входять у ці поля за основним значення, однак інші значення можуть суттєво відрізнятися від семантичного поняття, яке утворює це поле. У лінгвістиці вводять поняття семантичного простору, який інтегрує та об'єднує семантичні поля [141]. На вершині семантичної організації знаходиться поняття семантичного простору, далі поняття семантичного поля, лексико-семантичної групи, а на нижньому рівні знаходиться поняття слова. У роботі [142] проаналізовано закономірності розподілу лексемних полів дієслова в англомовній літературі. У роботі [143] проаналізовано семантичні сітки, семантичну структуру та ієрархію лексичних одиниць. У [144] проведено аналіз семантичних одиниць, уведено поняття семантичних станів мовних одиниць, які розглянуто як формальні репрезентативні стани та розглянуті моделі на основі теорії нечітких множин. Лексемний склад семантичних полів визначають різними способами. Один із способів полягає у виділенні загального поняття, на основі якого формують лексико-семантичне поле. Інший спосіб полягає у виділенні слова чи групи слів до яких підбирають синонімічні ряди. Також виділяють семантичні поля на основі експертного аналізу спільних появ лексем у заданих контекстах. У літературі розглядаються такі лексемні класи як семантичні поля, понятійні поля, тематичні групи лексем, семантичні групи, синонімічні ряди, семантичні домени та інші. Більшість визначень семантичної класифікації класів лексем є спорідненими, близькими до класичного визначення семантичного поля і базуються на моделі «мішка слів». У цій моделі розглядається сукупність слів текстових документів без розгляду їхньої контекстуальної послідовності. Це просто лексемні словники текстових документів, які впорядковані за алфавітом або деякими квантитативними характеристиками, наприклад за текстовою частотою. В роботах [145, 146, 147] запропоновано концепцію семантичних доменів, яка розширює поняття семантичних полів. Визначення семантичних

доменів є найбільш близьким до методів комп'ютерного аналізу текстів природньої мови і базується на відповідних текстових колекціях, які належать до аналізованого домена і характеризують семантичні поняття, які виокремлюють аналізований домен. У роботі [145] уведено поняття семантичного домена, який описує деяку предметну область, наприклад, економіку, політику, фізику, програмування тощо. Семантичні домени розглядають як семантичні поля, що описуються множинами слів, які часто зустрічаються в аналізованих семантичних областях. Семантичні домени можуть бути охарактеризовані та визначені підібраними для кожного домена колекціями текстів [145, 146]. Також, вони можуть розглядатися як сукупність взаємопов'язаних лексем, які описують деяку предметну область. Характерною властивістю доменних лексем є їхня кореляція з типом текстів, тобто, вони мають подібні розподіли у текстах одного і того ж типу [145]. У роботі [143] проаналізовано семантичні сітки, семантичну структуру та ієрархію лексичних одиниць. В інтелектуальному аналізі текстових даних часто широко використовують концепцію векторного представлення слів, в якому семантично подібні слова представлені подібними векторами [148, 149, 150].

В алгоритмах інтелектуального аналізу текстів використовують векторну модель текстових документів, яка базується на представленні документів як векторів у деякому фазовому просторі. Основна ідея цієї моделі полягає у представленні кожного текстового документа у вигляді вектора у деякому векторному просторі [151]. Вважають, що точки, які є близькими між собою у цьому просторі, відображають семантично близькі документи, і навпаки – семантично близькі документи відображаються близькими точками у фазовому просторі. Базис такого простору часто утворюють за допомогою частотно-дистрибутивних характеристик лексем текстового словника. Одним із методів представлення масиву документів у векторному просторі є організація векторів документів у вигляді матриці текстових частот типу лексеми-документи. Рядки таких матриць відповідають лексемам, а стовпці є векторами відповідних документів. Враховуючи надлишковість текстів з точки зору інформації, базиси таких просторів, які утворені на основі лексикону, у загальному випадку не

будуть ортонормованими, тобто, будуть спостерігатися залежності між частотами окремих лексем. Лексемний склад, який утворює базис векторного простору у свою чергу може бути структурований за семантичними класами, квантитативні характеристики яких у свою чергу можуть утворювати нові семантичні простори чи підпростори. Таким чином, масив документів може бути також охарактеризований семантичною структурою базисів векторних просторів текстових документів. Такі базиси формують на основі семантичних класів. Розглянемо низку гіпотез, які лежать в основі векторної моделі текстів [151]. Статистична семантична гіпотеза припускає, що статистичні характеристики вживання лексем можна використаовувати для визначення змісту сказаного. Якщо деякі частини тексту мають подібні вектори в частотних матрицях, тоді ці частини мають подібні значення. Гіпотеза сукупності лексем говорить про те, що частоти лексем в документі відображають зв'язок між документом та запитом. Запит можна розглядати як псевдокумент. Якщо документ та псевдодокумент мають подібні стовці в частотній матриці, тоді вони мають подібні значення. Дистрибутивна гіпотеза припускає, що лексеми, які зустрічаються в подібних контекстах мають подібні значення. Гіпотеза латентних відношень припускає, що пари лексем, які зустрічаються у подібних лексемних шаблонах, мають подібні семантичні зв'язки.

На основі проаналізованого матеріалу можна зробити висновок про доцільність розробки комплексної структурної багаторівневої класифікаційної моделі лексемного складу словників текстових масивів, яка б об'єднувала такі дистрибутивні лексемні відображення характеристик текстових масивів як семантика документа, тематика масиву документів, авторська стилістика текстового масиву. За допомогою квантитативних характеристик семантичних лексемних угрупувань можна утворити додаткові виміри у семантичному просторі представлення текстових документів. Уведення цих додаткових вимірів може бути ефективним у задачах інтелектуального аналізу текстів, зокрема, у класифікаційних задачах та задачах кластерного аналізу. Велика розмірність векторного простору, утвореного частототними характеристиками лексем словника текстового масиву, є значною проблемою класифікаційних алгоритмів

внаслідок значного обсягу обчислень. Тому актульними є методи зменшення розмірності базису такого векторного простору. Структурування словника, зокрема, на основі лінгвістичних семантичних концепцій може дати суттєве зменшення розмірності базису векторного простору текстових документів внаслідок використання квантитативних ознак семантичних полів.

1.3.2 Латентне розміщення Діріхле

Латентне розміщення Діріхле (Latent Dirichlet allocation, LDA) представляє текстові документи у вигляді сумішей прихованих тематик [152]. Тематики можуть характеризуватись певним набором слів. Текстовий документ можна представити у деякому багатовимірному просторі тематик. Ознаки документів на основі кількісних характеристик тематик можуть використовуватись у задачах кластеризації та класифікації текстових документів. У загальному LDA може бути застосоване до вибірки з великої кількості текстових документів. Розглянемо основні ідеї прихованого розміщення Діріхле, беручи за основу роботу [152]. Кожна тематика характеризується розподілом слів. Латентне розміщення Діріхле розглядають як генеративну ймовірнісну модель текстового корпусу. Словник слів текстового корпусу містить V слів. Документ розглядається як послідовність N слів:

$$\boldsymbol{w} = \{w_1, w_2, ..., w_N\}, \tag{1.56}$$

де w_n – n-е слово у послідовності слів текстового документа. Текстовий корпус розглядається як сукупність M документів

$$D = \{\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_M\}. \tag{1.57}$$

У теорії LDA розглядається такий генеративний процес [152]:

1. Вибрати $N \sim Poisson(\xi)$
2. Вибрати $\Theta \sim Dir(\alpha)$
3. Для кожного із N слів w_n :

(a) Вибрати тематику $z_n \sim Multinomial(\Theta)$

(b) Вибрати слово w_n із $p(w_n|z_n, \beta)$

Припускається, що розмірність k розподілу Діріхле $Dir(\alpha)$ є відомою. Ймовірності слів параметризовані матрицею β розмірності $k \times V$

$$\beta_{ij} = p(w^j = 1|z^j = 1). \tag{1.58}$$

Матриця β описує ймовірності приналежності заданого слова до заданої тематики. Розподіл Діріхле для деякої випадкової змінної Θ можна записати так

$$p(\Theta \mid \alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \Theta_1^{\alpha_1 - 1} ... \Theta_k^{\alpha_k - 1}. \tag{1.59}$$

де параметр α є k - вектор із компонентами $\alpha_i > 0$, Γ - гамма-функція. Розподіл суміші тематик розглядається як

$$p(\Theta, \boldsymbol{z}, \boldsymbol{w} \mid \alpha, \beta) = p(\Theta|\alpha) \prod_{n=1}^{N} p(z_n|\Theta) p(w_n|z_n, \beta). \tag{1.60}$$

$\Theta \sim Dir(\alpha)$, $z_n \sim Multinomial(\Theta)$ Інтегруючи по Θ і сумуючи по z, отримаємо розподіл слів у документі:

$$p(\boldsymbol{w} \mid \alpha, \beta) = \int p(\Theta|\alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\Theta) p(w_n|z_n, \beta) \right) d\Theta. \tag{1.61}$$

Для корпусу документів можна отримати:

$$p(D \mid \alpha, \beta) = \prod_{d=1}^{M} \int p(\Theta_d|\alpha) \left(\prod_{n=1}^{N} \sum_{z_{dn}} p(z_{dn}|\Theta_d) p(w_{dn}|z_{dn}, \beta) \right) d\Theta_d. \tag{1.62}$$

Розглядають три рівні представлення LDA моделей. Параметри α, β є параметрами рівня текстового корпусу і вибираються для текстової вибірки. Змінні Θ_d є змінними на рівні документів і вибираються для кожного документа. Змінні z_{dn}, w_{dn} є змінними на рівні слів і вибираються для кожного слова у кожному документі. Тематики в LDA можна

характеризувати за допомогою рангованого за частотою списку лексем, які належать до заданої тематики [153]. Одна з проблем полягає у тому, що серед лексем із найбільшими частотами є широковживані лексеми, що ускладнює семантичне диференціювання та інтерпретацію тематик. У [154] запропоновано розглядати поряд із частотою терму деякої тематики також і його унікальність для даної тематики. Аналогічний підхід, який полягає в аналізі релевантності лексем, розглянуто в [153], де релевантість до тематики визначається як

$$r(w, k, \mid \lambda) = \lambda log(\phi_{kw}) + (1 - \lambda) log\left(\frac{\phi_{kw}}{p_w}\right), \tag{1.63}$$

де ϕ_{kw} – ймовірність приналежності терму w до тематики k, p_w – ймовірність появи терму w у текстовій вибірці, λ – деякий заданий ваговий параметр. Тематики в LDA можна роглядати як багатомірні розподіли лексем, які належать до деякого словника. Формування словника для заданої задачі може розглядатися на експертному рівні. Важливим є вибір такого словника, який відображає предметну область аналізованої задачі. Різні методи візуалізації тематик у моделях на основі прихованого розміщення Діріхле розглянуто у роботах [153, 155, 156, 157, 158]. Один з ефективних підходів у візуалізації моделей LDA описано в [153] і реалізований в пакеті `LDAvis` для мови програмування Python. Кількісні характеристики розподілу прихованих тематик можна розглядати як ознаки в інтелектуальному аналізі даних на основі алгоритмів машинного навчання.

1.3.3 Кластерний аналіз

Завдання кластеризації полягає у побудові відображення множини вхідних даних на множину кластерів [159, 160, 161, 162, 163, 164, 165, 166]. Важливим етапом кластеризації є формування векторного простору текстових документів, у якому аналізують кластери. Алгоритми кластеризації та класифікації часто використовують у інтелектуальному аналізі текстів [167, 168]. Врахування семантики тексту у задачах кластеризації текстових документів дає можливість отримати більшу точність кластеризації [169]. ‘У [170] розглянуто алгоритми кластеризації

текстів та проаналізовано їхня ефективність. У поширеній векторній моделі документи відображають як вектори у багатомірному просторі, кожний вимір якого відповідає квантитативній характеристиці лексеми із словників текстових масивів. Текстовий масив можна представити у вигляді матриці ознак слів (термів) та документів. Такими ознаками можуть бути текстові частоти лексем. У матриці ознак колонки визначають документи, а рядки – частоти лексем у цих документах. Розглянемо групування документів за семантичними ознаками за допомогою алгоритму ієрархічної кластеризації. Нехай є множина текстових документів

$$D = \{d_i | m = 0, 1, 2..., N_d\} \tag{1.64}$$

та множина кластерів

$$C = \{c_m | m = 0, 1, 2..., N_c\}. \tag{1.65}$$

Необхідно побудувати відображення множини документів на множину кластерів :

$$U_{DC} : D \to C. \tag{1.66}$$

Відображення U_{DC} задає модель даних, яка є розв'язком задачі кластеризації [159, 160, 166, 164, 165]. Кожний елемент c_m множини кластерів C складається з підмножини текстових документів, які подібні між собою відповідно до деякої кількісної міри подібності r

$$c_m = \{d_i, d_j | d_i \in D, d_j \in D,\ r(d_i, d_j) < \epsilon\}, \tag{1.67}$$

де ε – визначає деякий поріг для включення документів в кластер. Величина $r(d_i, d_j)$ є відстанню між елементами d_i та d_j. Якщо виконується умова

$$r(d_i, d_j) < \varepsilon, \tag{1.68}$$

то елементи вибірки вважають подібними і приналежними до спільного кластера. В іншому випадку елементи знаходяться у різних кластерах.

Матриця

$$R = \{r_{ij} = r(d_i, d_j)\} \tag{1.69}$$

є матрицею відмінностей в алгоритмі кластеризації. Оскільки на множині текстових документів введено поняття відстані, то кожен документ представляється у вигляді точки в N_s–мірному просторі R^{N_s} семантичних полів. Є декілька методів обрахунку мір близькості точок в N_s–мірному просторі, зокрема, евклідова відстань обраховується так:

$$r_e(d_i, d_j) = \sqrt{\sum_{k=1}^{N_s} \left(p_{ki}^{sd} - p_{kj}^{sd}\right)^2}. \tag{1.70}$$

Подібність між двома текстовими документами в N_s–мірному просторі також визначається кутом між векторами цих документів і за кількісну міру можна взяти косинус цього кута. Розглянемо ієрархічний метод агломеративної кластеризації. На першому кроці вся множина текстових документів розглядається як множина кластерів:

$$c_1 = \{d_1\}, c_1 = \{d_1\}, ...c_{Nd} = \{d_{Nd}\} \tag{1.71}$$

На наступному кроці два близьких один до одного документи (наприклад d_p і d_q) об'єднуються в один спільний кластер, нова множина на цьому кроці вже складається із $N_d - 1$ кластерів і має вигляд

$$c_1 = \{d_1\}, c_2 = \{d_2\}, ...c_p = \{d_p, d_q\} ...c_{Nd-1} = \{d_{Nd-1}\}. \tag{1.72}$$

Повторюючи кроки, на яких будуть об'єднуватися кластери, отримаємо множину із N_c кластерів. Процес об'єднання кластерів завершується на тому кроці алгоритму, коли жодна пара кластерів не відповідає порогу об'єднання для міри близькості елементів. На кожній ітерації алгоритму необхідно робити перерахунок між кластерами. Враховуючи те, що кластери можуть складатися з декількох об'єктів, існують різні методи формування та об'єднання кластерів на основі відстаней між об'єктами всередині кластера. Метод найближчого сусіда полягає у виборі найменшої відстані між двома

кластерами. У методі дальніх сусідів вибирається відстань між двома найдальшими сусідами. У центроїдному методі розраховується евклідова відстань між центрами двох кластерів. У методі Варда обраховують квадрати евклідових відстаней від окремих документів до центру кожного кластера. Далі ці відстані сумують. У новий кластер об'єднуються ті кластери, при об'єднанні яких виходить найменший приріст суми квадратів відстаней. Графічним зображенням результату ієрархічної кластеризації є дендрограма, на якій відображається процес агломеративного об'єднання кластерів. По осі абсцис відкладають номери кластерів, а по осі ординат – відстані між кластерами. При певних значеннях відстаней починається об'єднання кластерів. З ростом міжкластерної відстані кластери об'єднуються аж до повного злиття кластерів в один. Для отримання інформативної кластерної структури вибирається деякий поріг міжкластерної відстані, при якому утворюється оптимальна з точки зору аналізу текстових масивів кластерна структура. Наприклад, при дослідженні можливості кластеризації текстових документів за авторами доцільно взяти таке порогове значення міжкластерної відстані, при якому утворюється кількість кластерів, що дорівнює кількості аналізованих авторів. Розглянемо алгоритм кластеризації документів методом k-середніх (k-means). На початковому кроці вибираємо k центрів кластеризації, це можуть бути випадкові точки семантичного простору. Для кожного центру формують групу текстових документів, які є найближчими за евклідовою відстанню у векторному просторі до цього центру. Утворені групи текстових документів формують проміжні кластери. Далі визначають центри мас цих кластерів. Координати вектору центрів мас розраховують як середні значення за координатами векторів текстових документів в утворених кластерах. Отримані центри мас беруть як центри кластеризації на наступній ітерації, у якій відбувається новий перерозподіл текстових документів за найменшою відстанню до центрів кластеризації. Процес кластеризації завершується на ітерації, при якій не відбувається нового перерозподілу текстових документів. Суть кластеризації полягає в

мінімізації дисперсії σ на точках кластерів у векторному просторі

$$\sigma = \sum_{m=1}^{N_c} \sum_{d_j \in c_j} r(d_j, \mu_m), \tag{1.73}$$

де μ_m – центр мас для векторів документів d_j, які належать до кластера c_j .

1.3.4 Часті множини та асоціативні правила

В аналізі слабоструктурованих даних, зокрема, текстових масивів часто використовують алгоритми пошуку частих множин ознак та асоціативних правил, за допомогою яких можна виявити взаємозв'язок між підмножинами даних [171, 172, 173, 174, 175, 176, 177, 178, 166]. Одне із завдань пошуку асоціативних правил полягає у виявленні певних сукупностей об'єктів, які часто зустрічаються у великих масивах. Розглянемо основні положення теорії частих множин та асоціативних правил з точки зору застосування в аналізі текстових масивів, використовуючи основні положення робіт [171, 172, 166]. Уведемо загальну множину об'єктів

$$I = i_1, i_2, ..., i_n, \tag{1.74}$$

де I_j - об'єкти, n – загальна кількість об'єктів. Під об'єктом можна розуміти, наприклад, текстовий файл, повідомлення мікроблогу. Сукупності об'єктів, які аналізуються з точки зору виявлення закономірностей, називають транзакціями. Транзакцію можна описати як підмножину множини I :

$$T = \{i_j | i_j \in I\}. \tag{1.75}$$

Набір транзакцій, які аналізуються розглянемо як множину

$$D = \{T_1, T_2, ...T_m\}, \tag{1.76}$$

де m — кількість транзакцій в аналізі. Розглянемо множину транзакцій, у яку входить об'єкт i_j

$$D(i_j) = \{T_r | i_j \in T_r; j = 1...n;\ r = 1, ...m,\ D(i_j) \in D\}. \tag{1.77}$$

Нехай існує деякий довільний набір об'єктів

$$F = \{i_j | i_j \in I\}. \tag{1.78}$$

Множину транзакцій, у яку входить набір F позначимо

$$D_F = \{T_r \mid F \in T_r; r = 1, ...m\}. \tag{1.79}$$

Відношення кількості транзакцій, у які входить множина F до загальної кількості транзакцій називають підтримкою набору F і позначають :

$$Supp(F) = \frac{|D_F|}{|D|}. \tag{1.80}$$

Набір називають частим, якщо значення його підтримки більше мінімальної підтримки, яка задається користувачем

$$Supp(F) > Supp_{\min}. \tag{1.81}$$

Враховуючи умову (1.81), знаходимо сукупність частих множин

$$L = \{F_j \mid Supp(F_j) > Supp_{\min}\}. \tag{1.82}$$

Для виявлення частих множин переважно використовують алгоритм Apriori [171]. В його основі лежить принцип, який полягає в тому, що підтримка деякої частої множини не перевищує підтримки будь-якої з її підмножин. Асоціативні правил розглядають у вигляді

$$X \to Y. \tag{1.83}$$

де X - це умова, в яку входять об'єкти, пов'язані з сукупністю об'єктів наслідку Y. Об'єкти умови та наслідку є підмножинами частої множини ознак. В задачах пошуку асоціативних правил виділяють два основні етапи: пошук всіх частих множин об'єктів та генерація асоціативних правил на основі знайдених частих множин. Асоціативні правила можна поділити на такі типи: Тривіальні правила – містять зрозумілий та очевидний взаємозв'язок об'єктів. Такі правила можуть бути використані для перевірки алгоритмів інтелектуального аналізу. Корисні правила – це правила, що містять інформацію, яка раніше не була відома, але має логічне пояснення. Незрозумілі правила – це правила, які містять інформацію, що не може бути логічно обґрунтована на основі відомих знань. Незрозумілі правила можуть виникати в некоректних алгоритмах чи некоректних вхідних даних, а також можуть відображати об'єктивні, але не відомі в даний час закономірності. Асоціативні правила типу (1.83) будують на основі частих множин F, для яких

$$X \cup Y = F. \tag{1.84}$$

На основі одного набору можна побудувати велику кількість асоціативних правил, які будуть визначатися всеможливими комбінаціями ознак. Значна частина таких правил є тривіальною або не несе корисної інформації. Для оцінки та відбору корисних правил вводять ряд кількісних характеристик, зокрема, підтримку, достовірність, покращення. Підтримка асоціативного правила показує, яка частка транзакцій містить це правило. Оскільки правило будується на основі частої множини ознак, то правило $X \to Y$ має підтримку, що дорівнює підтримці множини $F : X \in F, Y \in F$. Різні правила, побудовані на основі одного набору, мають одні і ті ж величини підтримки. Підтримку розраховують за допомогою формули (1.80). Достовірність асоціативного правила показує ймовірність того, що з наявності в транзакції підмножини ознак X випливає наявність підмножини ознак Y. Під достовірністю розуміють відношення числа транзакцій, які містять підмножини ознак X та Y до числа транзакцій, які містять лише

підмножину ознак X:

$$Conf_{X\to Y} = \frac{|D_{X\cup Y}|}{|D_X|} = \frac{Supp_{X\cup Y}}{Supp_X}. \tag{1.85}$$

Важливою особливістю є те, що в різних асоціативних правилах одного і того ж набору достовірність буде різною. Однак, достовірність не визначає корисність правила. Наприклад, якщо частка наявності ознак Y в транзакціях із наявними ознаками X є меншою, ніж частка безумовної наявності Y, тоді ймовірність вгадати наслідок Y є більшою, ніж ймовірність виявити Y на основі правила $X \to Y$. Для визначення корисності правила розраховують характеристику покращення (improvement):

$$Impr_{X\to Y} = \frac{Supp_{X\cup Y}}{Supp_X \cdot Supp_Y}. \tag{1.86}$$

Величина $Impr_{X\to Y}$ показує чи дане правило є корисніше випадкового вгадування. Якщо $Impr_{X\to Y} > 1$, то це означає, що за допомогою даного правила передбачити наслідок Y є більш ймовірним, ніж випадково його вгадати.

1.3.5 Теорія формальних концептів

Підхід на основі теорії формальних концептів (Formal Concept Analysis, FCA) дає можливість аналізувати структурні патерни в даних на основі їх атрибутів [179, 180, 181, 182, 183, 184, 180, 185, 186, 187, 188]. Ця теорія також використовується в аналітиці корпусу текстів [180]. Її можна розглядати з точки зору нових підходів у задачах групування даних. Розглянемо основні положення цієї теорії базуючись на [179, 180]. Одним з основних є поняття контексту:

$$K = (G, M, I), \tag{1.87}$$

де G, M є множини, а I – бінарне відношення між G і M. Елементи множини G є об'єктами, а множини M – атрибутами об'єктів, I є відношенням

інцидентності. Для $A \subseteq G$, $B \subseteq M$ визначають оператори Галуа:

$$A' = \{m \in M \mid \forall g \in M : (g, m) \in I\}, \tag{1.88}$$

$$B' = \{g \in G \mid \forall m \in B : (g, m) \in I\}. \tag{1.89}$$

Формальним концептом називають пару (A, B), для якої виконується умова

$$A \subseteq G, B \subseteq M, A' = B, B' = A.$$

Тобто (A, B) розглядається як формальний концепт, якщо множина атрибутів об'єктів множини A дорівнює множині B, з іншої сторони множину A можна розглядати як множину об'єктів, які мають всі атрибути з множини B. A називають екстентом, а B - інтентом формального концепту (A, B). Формальні концепти заданого контексту є частково впорядкованою множиною для яких виконується:

$$(A_1, B_1) \leq (A_2, B_2) \Leftrightarrow A_1 \subseteq A_2, B_2 \subseteq B_1. \tag{1.90}$$

Множину концептів можна зобразити за допомогою діаграми Гассе. У подальшому доцільно розглянути семантичну структуру текстових даних із використанням теорії формальних концептів.

1.4 Висновки

1. Розглянуто методи та підходи в аналітиці даних табличного типу, зокрема, розглянуто лінійні моделі, ймовірнісні моделі на основі байєсівського виведення та моделі машинного навчання. Лінійні моделі дають можливість встановити лінійний зв'язок між цільовою змінною та ознаками, однак не дають можливості виявити складну взаємодію між ознаками. Ймовірнісні байєсівські моделі дають можливість отримати густину розподілу ймовірностей для цільової змінної, що є важливим в аналітиці ризиків та оцінці невизначеності. Методи машинного та глибинного навчання дають можливість виявити складні закономірності між ознаками, однак їх можна ефективно застосовувати лише у випадку стаціонарного розподілу ознак. Розглянуто методи

глибокого Q-навчання.

2. В аналітиці текстових даних широко використовуються концепції семантичних структур, зокрема, семантичних полів. Одним з ефективних структурних поділів текстового словника є поділ на основі семантичних одиниць, зокрема, реалізований у лінгвістичній системі WordNet. Розглянуто лінгвістичні концепції семантичних та тематичних лексикографічних полів із точки зору їх використання в алгоритмах інтелектуального аналізу текстових масивів. Під семантичними полями розглядають множини лексем, які об'єднані деякою парадигмою. Під парадигмою можна розуміти, наприклад, спектр семантичних або тематичних понять, які відображені у структурі лексикографічних значень лексем.

3. Проведений аналіз виявив необхідність подальшого дослідження комбінування різних моделей, зокрема, за допомогою ансамблів моделей для того, щоб ефективно використати переваги кожного типу моделей в інтелектуальному аналізі даних у заданій предметній області. Структурна та інформаційна складність наборів даних, які описують різні явища та процеси у різних предметних областях, не дають можливості знайти загальний універсальний підхід в інтелектуальному аналізі даних. Тому виникає необхідність розробки методів та підходів у формуванні ознак даних та прогнозних моделей на прикладах реальних типових задач заданої предметної області. Необхідно реалізувати такі завдання: розробити метод застосування машинно-навчальних та ймовірнісних моделей для покращення точності та якості інтелектуального аналізу даних табличного типу; розробити методи стекінгового об'єднання різнотипних моделей у прогнозні ансамблі на основі лінійної регресії LASSO та байєсівського регресії; удосконалити метод використання навчання з підкріпленням в аналітиці табличних даних з імітаційним моделюванням середовища взаємодії інтелектуального агента; розробити метод використання теорії семантичних та тематичних полів у інтелектуальному аналізі даних з метою формування квантитативних семантичних ознак

текстових даних; розробити метод інтелектуального аналізу текстових даних на основі машинного навчання з використанням семантичних ознак; удосконалити метод класифікаційного та регресійного аналізу з використанням нейромережі з вхідними текстовими даними та кількісними ознаками; розробити метод використання теорії частих множин та асоціативних правил для формування семантичних ознак в інтелектуальному аналізі текстових даних; розробити метод використання теорії формальних концептів в аналітиці текстових потоків соціальних мереж для аналізу семантичної структури текстових даних та формування аналітичних ознак на основі виявленої семантичної структури даних; створити засоби для апробації розроблених у роботі методів інтелектуального аналізу табличних та текстових даних.

2 МЕТОДИ МОДЕЛЮВАННЯ, ФОРМУВАННЯ ОЗНАК ТА СТЕКІНГ МОДЕЛЕЙ В ІНТЕЛЕКТУАЛЬНОМУ АНАЛІЗІ ДАНИХ ТАБЛИЧНОГО ТИПУ

Проаналізуємо машинно-навчальні, ймовірнісні та лінійні моделі у прогнозній аналітиці даних табличного типу. Розглянемо побудову стекінгових ансамблів прогнозних моделей. Проаналізуємо використання методів Q-навчання та їх комбінації в аналізі даних табличного типу.

2.1 Реляційна модель даних та формування ознак для інтелектуального аналізу

Розглянемо узагальнену модель даних, використовуючи теорію реляційної алгебри [189, 190, 191]. Певний процес або об'єкт можна описати як елемент, або зразок даних за допомогою набору даних, наприклад так:

$$t = \{(F_1, x_1), (F_2, x_2), ..., (F_n, x_n)\}. \tag{2.1}$$

Кожна ознака F_i набуває значення x_i і визначається доменом D_i. Множині ознак $\{F_1, F_2, ..., F_n\}$ відповідає множина доменів $\{D_1, D_2, ..., D_n\}$. Домени описують тип даних та визначають допустиму множину значень ознак. Елемент даних можна описати у вигляді відношення з одним кортежем даних $R(F_1, F_2, ..., F_n)$. Сукупність елементів даних описує сукупність аналізованих об'єктів або процесів і може бути представлена відношенням із багатьма кортежами, в якому кожний кортеж описує окремий об'єкт або, наприклад, деякий часовий момент аналізованого процесу. Таке відношення описує табличне представлення даних, в якому стовпці описують ознаки даних, а рядки - зразки даних. Множина даних для кожного кортежу описує схему даних. Кортежі з однаковими множинами доменів та атрибутів можуть бути об'єднані. Набір даних складається із сукупності кортежів t_i

$$\begin{aligned} DataSet =& \{t_i\}, \\ DataSet =& \{(F_{11}, x_{11}), (F_{12}, x_{12}), ..., (F_{ij}, x_{ij})..., \\ & (F_n, x_n) | i = 1, ..., N_F, j = 1, 2, ..., N_D\} \end{aligned} \tag{2.2}$$

Нехай існує декілька різних джерел даних

$$DataSources = \{DataSource_i | i = 1, 2, ...N_{DSources}\}, \tag{2.3}$$

тоді консолідована множина даних буде складатися з сукупності наборів даних з різних джерел

$$ConsDataSet = \{DataSet_i | i = 1, 2, ...N_{DSources}\}. \tag{2.4}$$

Для кожного джерела даних є своя множина ознак F^k_{set}, якій відповідає своя множина доменів D^k_{set}. Загальна множина ознак консолідованих даних буде

$$F_{cons} = \{F : F \in F^1_{set} \vee F \in F^2_{set} \vee ...F \in F^i_{set} \vee ... | i = 1, 2, ...N_{DSources}\}. \tag{2.5}$$

Для об'єднаної множини ознак можна знайти відповідну множину доменів. Різні кортежі з однаковими схемами можуть бути об'єднані у відношення, навіть якщо вони походять із різних джерел консолідованого масиву даних. На основі цих відношень можна створити нові відношення, використовуючи оператори реляційної алгебри з новими атрибутами та схемами даних. Процес формування нових відношень визначається цілями інтелектуального аналізу і важко піддається формалізації. Формування нових відношень є важливою складовою інтелектуального аналізу. Підходи та методи у формуванні нових відношень можна дослідити на прикладах реальних задач із різних предметних областей. Проаналізуємо формування ознак на прикладі аналітики продажів товарів. Розглянемо дані продажів у вигляді алгебраїчного відношення, яке описує фрейм даних

$$SalesDF(Date, Promo, Store, Sales) \tag{2.6}$$

Розділимо $SalesDF$ на тренінговий, валідаційний та тестовий сет за часовим інтервалом у вигляді

$$\begin{aligned} t^{train} &= [t^{train}_{min}, t^{train}_{max}], \\ t^{val} &= [t^{val}_{min}, t^{val}_{max}], \\ t^{test} &= [t^{test}_{min}, t^{test}_{max}]. \end{aligned} \tag{2.7}$$

Фрейм даних для тренінгу, валідації та тесту визначимо як

$$\begin{aligned} SalesDF_{train} &= \sigma_{(SalesDF.Date \in t^{train})}(SalesDF), \\ SalesDF_{val} &= \sigma_{(SalesDF.Date \in t^{val})}(SalesDF), \\ SalesDF_{test} &= \sigma_{(SalesDF.Date \in t^{test})}(SalesDF). \end{aligned} \tag{2.8}$$

Ознака $Date$ містить додаткові ознаки, такі як рік, місяць, день місяця, день тижня. Функцію визначення ознак дати позначимо як $GetDateFeatures$. Ця функція буде утворювати відношення $DateFeatures$

$$\begin{aligned} DateFeatures(Date, Year, Month, MonthDay, WeekDay) = \\ GetDateFeatures(\Pi_{Date}(SalesDF)) \end{aligned} \tag{2.9}$$

Здійснимо ліве об'єднання відношення $SalesDF$ з відношенням $DateFeatures$ і замінимо відношення $SalesDF$ результатом цього об'єднання:

$$SalesDF \leftarrow SalesDF \rtimes_{(SalesDF.Date=DateFeatures.Date)} DateFeatures. \tag{2.10}$$

Для формування прогнозної моделі утворимо нові ознаки на основі агрегованих функцій. Розглянемо місячні об'єми продаж кожного магазину

$$StoreSales^{month}(Store, Month, Sales) =_{Store,Month} G_{Sum(Sales)}(SalesDF) \tag{2.11}$$

Знайдемо усереднені за місяцями продажі

$$StoreSales^{month}_{avg} =_{Store} G_{Avg(Sales)}(SalesDF^{month}). \tag{2.12}$$

У результаті лівого об'єднання отримаємо

$$SalesDF \leftarrow SalesDF \ltimes_{(SalesDF.Store=StoreSales_{avg}^{month}.Store)}(SalesDF) \quad (2.13)$$

Розділимо ознаки та цільову змінну для тренінгових та тестових фреймів:

$$\begin{aligned} &X_{train} = \Pi_{Date,Promo,Store}(SalesDF_{train}), \\ &y_{train} = \Pi_{Sales}(SalesDF_{train}) \\ &X_{test} = \Pi_{Date,Promo,Store}(SalesDF_{test}), \\ &y_{test} = \Pi_{Sales}(SalesDF_{test}). \end{aligned} \quad (2.14)$$

Ознаки, які характеризують дати продаж, зокрема, рік, місяць, день місяця, день тижня, є категоріальними. Для інтеграції категоріальних ознак у прогнозну модель використовують бінарні змінні [57]. Один із поширених методів такої бінаризації є *One Hot Encoding*. У цьому методі категоріальна ознака замінюється на декілька бінарних змінних, кількість яких визначається розміром множини значень категоріальної змінної. Назви цих бінарних змінних відповідають елементам множини значень категоріальної змінної. Бінарна змінна набуває значення 1, якщо значення категоріальної змінної в даному кортежі відповідає назві цієї бінарної змінної, в інших випадках бінарна змінна дорівнює 0. Утворення відношення бінарних змінних для категоріальної змінної можна записати так

$$\begin{aligned} &BinFeat_{CategFeat}(CategFeat, BinFeat_1, BinFeat_2, ...) \leftarrow \\ &OneHotEnc(SalesDF, CategFeat). \end{aligned} \quad (2.15)$$

Для подальшого аналізу здійснимо ліве об'єднання утвореного відношення бінарних значень $BinFeat_{feat}$ із відношенням $SalesDF$

$$\begin{aligned} &SalesDF \leftarrow \\ &SalesDF \ltimes_{(SalesDF.CategFeat=BinFeat_{CategFeat}.CategFeat)}(SalesDF) \end{aligned} \quad (2.16)$$

Відношення та операції, описані формулами (2.1)-(2.16), можна розглядати

як реляційну модель даних аналізованої задачі із заданої предметної області. Утворений фрейм $SalesDF$ в результаті проведених операцій містить нові інформативні ознаки, які можна використати для інтелектуального аналізу. Отже, консолідовані дані з різною структурою та з різних джерел можуть бути представлені у вигляді реляційної моделі. За допомогою операцій реляційної алгебри можна виділити та утворити нові ознаки аналізованої задачі, які в подальшому будуть використані для інтелектуального аналізу.

2.2 Методи машинного навчання у прогнозуванні часових рядів

Часові ряди даних можна розглядати у табличному представленні. Прикладом часових рядів можуть буди дані про продажі товарів. Прогнозування продажів є важливою частиною сучасної бізнес-аналітики [192, 193, 194]. На сьогодні існують різні моделі часових рядів, наприклад, Holt-Winters, ARIMA, SARIMA, SARIMAX, GARCH тощо. Різні підходи з використанням часових рядів описані у [195, 196, 197, 198, 199, 200, 201, 202, 203, 204, 205, 206]. У [207] автори досліджують прогнозованість часових рядів та вивчають різні методи прогнозування. В [208] розглядаються та порівнюються різні багатокрокові підходи до прогнозування часових рядів. У [209] досліджуються комбінування методів прогнозування. Показано, що у випадку, коли різні моделі базуються на різних алгоритмах та даних, можна отримати суттєво точніші результати. Покращення точності є важливим у випадках з великою невизначеністю. У [75, 76, 77, 78, 79, 80], розглядаються різні ансамблеві методи класифікаційних задач. У [210] показано, що за допомогою комбінування прогнозів, створених за допомогою різних алгоритмів, можна вдосконалити точність прогнозування. У роботі було розглянуто різні підходи ефективного поєднання прогнозів. У [211] автори розглянули вибір лагової змінної, оптимізацію гіперпараметрів, порівняння між класичними алгоритмами та алгоритмами на основі машинного навчання для часових рядів. На сетах даних температурних часових рядів автори показали, що класичні алгоритми та алгоритми на основі машинного навчання можна використовувати з однаковою ефективністю. Існують певні обмеження

використання підходів часових рядів до прогнозування часових рядів, зокрема тих, які описують продаж товарів. Деякі з них:

– Потрібно мати історичні дані за тривалий період, щоб визначити сезонність. Але часто немає історичних даних для цільової змінної, наприклад, коли запускається новий продукт у продаж. У той же час є часові ряди продажів аналогічного продукту і можна очікувати, що новий продукт матиме схожий патерн продажу.

– Дані про продажі можуть мати багато викидів (аномальних значень), а потрібні дані можуть бути відсутніми. Потрібно видалити викиди та інтерполювати дані перед використанням підходу на основі часового ряду.

– Потрібно враховувати багато зовнішніх факторів, які впливають на динаміку часових рядів, зокрема, продажі.

У прогнозуванні деяких типів часових рядів, зокрема продажів, використання регресійних підходів може дати кращі результати порівняно з статистичними методами часових рядів, такими як Holt-Winters, ARIMA. Алгоритми машинного навчання дозволяють знаходити закономірності у часових рядах. Можна знайти складні патерни в динаміці продажів, використовуючи методи машинного навчання з учителем. Популярними є алгоритми машинного навчання на основі дерев рішень [69], наприклад Random Forest [70], Gradiend Boosting Machine [71, 72]. Одне з головних припущень регресійних методів полягає у тому, що патерни попередніх даних будуть повторюватися в майбутньому. У [212] досліджувалися лінійні моделі, машинне навчання та ймовірнісні моделі для моделювання часових рядів. Для імовірнісного моделювання розглянуто підходи на основі використання копул та байєсівського висновування (виведення). У [213] розглянуто підходи на основі стекінгу моделей для часових рядів та логістичної регресії з сильно незбалансованими даними. У даних про продажі ми можемо спостерігати кілька типів моделей та ефектів, зокрема, тренд, сезонність, автокореляція, патерни, спричинені впливом таких зовнішніх факторів, як промоакції, ціноутворення, поведінка конкурентів. Також спостерігається

шум, викликаний факторами, які не враховуються у побудованій моделі. У даних про продажі також можна спостерігати екстремальні значення – викиди. Якщо потрібно провести оцінку ризику, тоді необхідно враховувати шум та екстремальні значення. Викиди можуть бути спричинені певними специфічними факторами, наприклад промо-подіями, зниженням цін, погодніми умовами тощо. Якщо такі події повторюються періодично, можна додати нову ознаку, яка вказує на ці особливі події та описує екстремальні значення цільової змінної. Розглянемо використання моделей машинного навчання для прогнозування часових рядів продажів. Дослідимо одинарну модель, ефект генералізації моделі машинного навчання та стекінговий ансамбль прогнозних моделей.

2.2.1 Прогнозні моделі на основі машинного навчання з учителем

Розглянемо приклад використання моделі на основі машинного навчання для прогнозування одного часового ряду. Для аналізу виберемо часовий ряд продажів товарів. Дані для проведеного аналізу базувалися на даних про продажі магазинів "Rossmann Store Sales" зі змагання по аналізу даних на платформі Kaggle [214]. Ці дані описують продажі в магазинах мережі Rossmann. Розрахунки проведено у середовищі Python, з використанням основних пакетів *pandas* [215, 216], *sklearn* [217], *numpy* [218], *keras* [219], *matplotlib* [220], *seaborn* [221]. Аналіз проведено у середовищі *Jupyter Notebook* На рис. 2.1 показано типові часові ряди продажів, величини продажів є нормованими на деяку довільно вибрану величину. Спочатку проведено описову аналітику, зокрема, дослідження розподілу продажів та візуалізацію даних за допомогою попарної візуалізації змінних. Така аналітика допомагає знайти відповідні характеристики, зокрема, фактори впливу на продажі. Як ознаки було вибрано категоріальні змінні – $weekday$ – день тижня, $StoreType$ – тип магазину, $Assortment$ – тип асортименту, $Promo$ - наявність промо акції, та числові ознаки – $CompetitionDistance$ - відстань до конкуруючих магазинів, $logMonthSales$ – натуральний логарифм середніх місячних продажів, $logSales$ - натуральний логарифм денних продажів, який розраховується за формулою $logSales = log(Sales + 1)$.

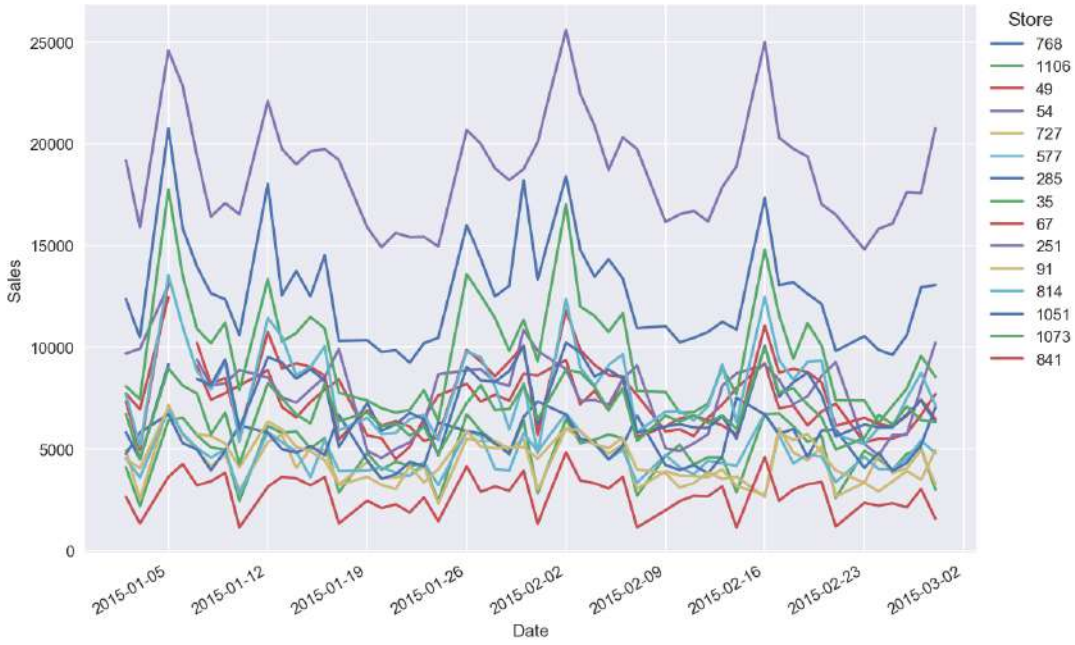

Рисунок 2.1 – Часові ряди продажів

Розглянемо результати проведеного аналізу citepavlyshenko2019machine, pavlyshenko2016linear, pavlyshenko2018using, pavlyshenko2018predictive, pavlyshenko2018regression. На рис. 2.2 наведено коробкові графіки розподілу продажів відносно дня тижня, на рис. 2.3 наведено агреговані об'єми продажів за різними категоріальними ознаками, на рис. 2.4 показано парні залежності для змінних $logMonthSales, logSales, CompetitionDistance$. Отримані залежності характеризують аналізовані процеси і можуть бути корисними при формуванні аналітичних ознак прогнозної моделі.

Розглянемо підхід на основі машинного навчання з учителем із використанням історичних даних часових рядів продажів. Для дослідження використано алгоритм машинного навчання Random Forest [70]. У випадку прогнозування часових рядів, які мають незначну нестаціонарність, наприклад, тренд, використання звичайного підходу кросвалідації на тренувальній вибірці може зумовити суттєве зміщення в оцінці прогнозованої змінної. У такому випадку доцільно розділити сет історичних даних на тренінговий та валідаційний сети за часовим поділом так, щоб тренувальні дані знаходились у першому часовому проміжку, а валідаційні – у наступному. У якості незалежних регресійних змінних було використано категоріальні ознаки, такі як промо, день тижня, день

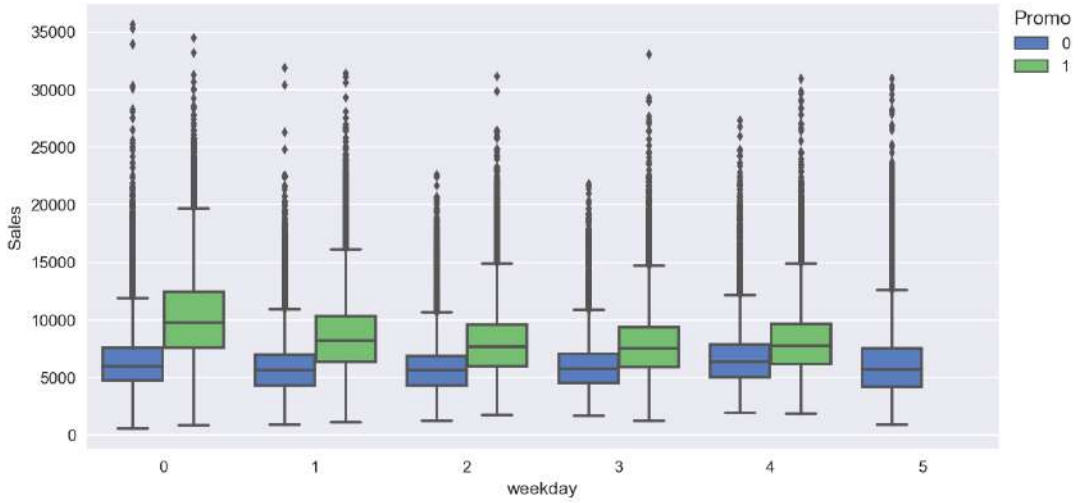

Рисунок 2.2 – Коробкові графіки розподілу продажів відносно дня тижня

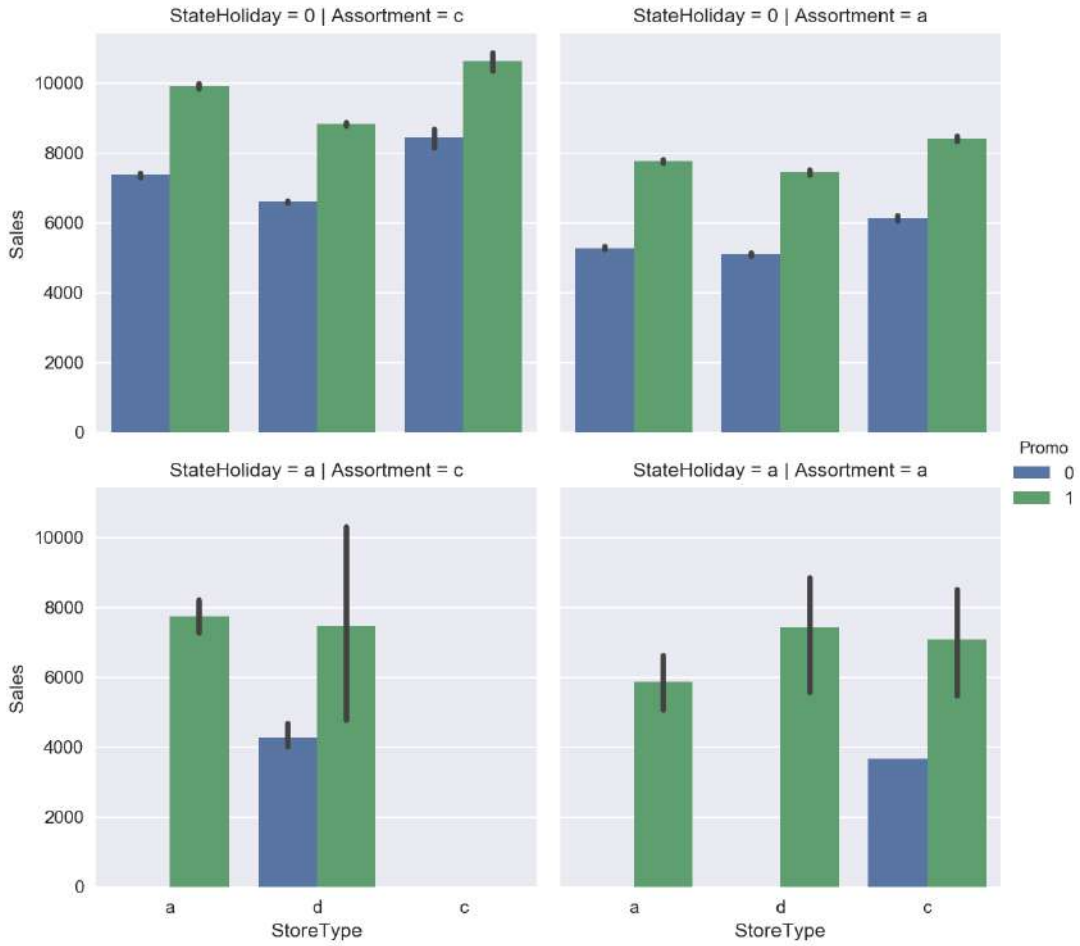

Рисунок 2.3 – Агреговані об'єми продажів за різними категоріальними ознаками

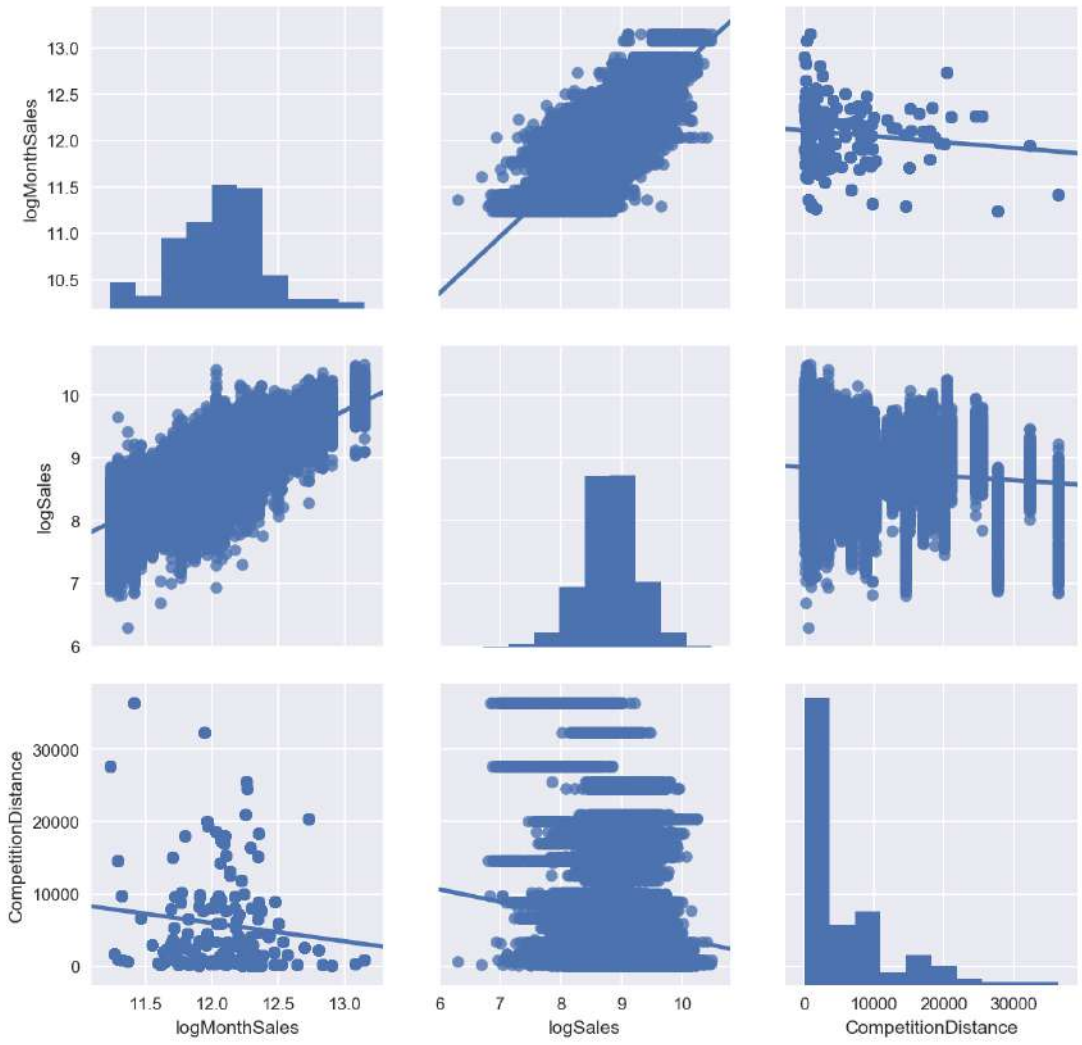

Рисунок 2.4 – Парні залежності для змінних $logMonthSales, logSales, CompetitionDistance$

місяця, місяць. Для категоріальних ознак застосовано one-hot кодування з використанням фіктивних змінних, коли одна категоріальна змінна була замінена на n бінарних змінних, де n – кількість унікальних значень категоріальних змінних. Розглянемо отримані результати прогнозування часового ряду на основі алгоритмічної моделі машинного навчання. На рис. 2.5 показано реальні та прогнозні значення для аналізованого часового ряду продажів. Вертикальна пунктрирна лінія розділяє часовий проміжок тренувальної та валідаційної вибірок даних. На рис. 2.6 показано розраховану характеристику важливості ознак, на основі яких побудована прогнозна модель. Для оцінки похибок використано відносне середнє значення абсолютної похибки (MAE), яка обчислюється як $error = MAE/mean(Sales) \cdot 100\%$. На рис. 2.7 показано залишки прогнозу, на , на рис. 2.8 – біжуче середнє залишків, а на рис. 2.9 – стандартне відхилення залишків прогнозу на тренувальній та валідаційній вибірках. Вертикальна лінія на рисунках розділяє часові проміжки для тренувальної та валідаційної вибірок. В отриманих прогнозних результатах спостерігається зміщення величини залишків на валідаційній вибірці, яке зумовлене застосуванням методу машинного навчання до нестаціонарних часових рядів. Можна провести корекцію такого зміщення, використовуючи лінійну регресію на валідаційній вибірці. Похибка на тренувальні вибірці у порівнянні з валідаційною вибіркою є суттєво меншою. Похибка на валідаційній вибірці є важливим показником для вибору оптимальної кількості ітерацій алгоритмів машинного навчання.

У Додатках наведено аналіз взаємного впливу часових рядів на прикладі взаємного впливу наявності товарів із подібними споживчими ознаками в аналітиці продажів.

2.2.2 Ефект генералізації моделей машинного навчання

Ефект генералізації у методах машинного навчання полягає у тому, що алгоритм регресії знаходить патерни, властиві для цілої вибірки даних. Якщо продажі мають виражені закономірності, то генералізація дозволяє отримати точніші результати, на які не впливають випадкові відхилення значень ознак. Для випадку дослідження генералізації моделі машинного

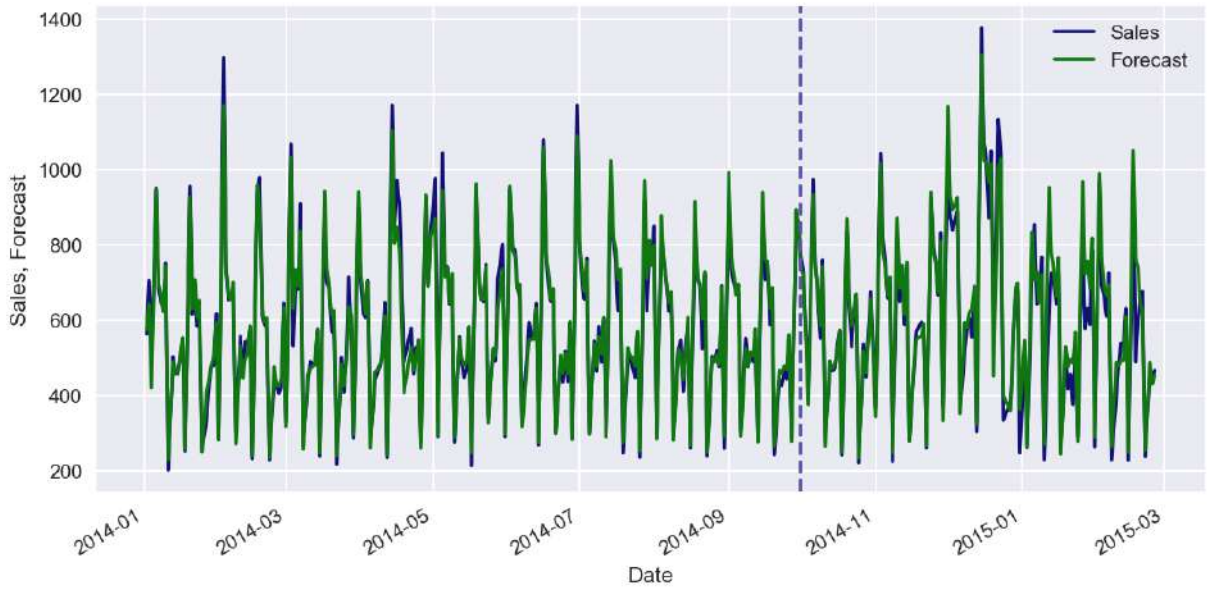

Рисунок 2.5 – Прогноз часових рядів продажів (похибка на тренувальній вибірці:3.9%, похибка на валідаційній вибірці: 11.6%)

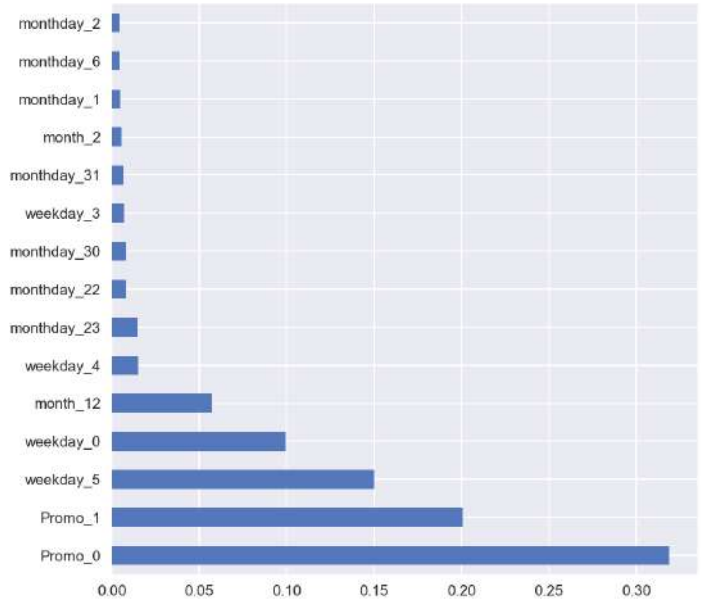

Рисунок 2.6 – Важливість ознак прогнозної моделі

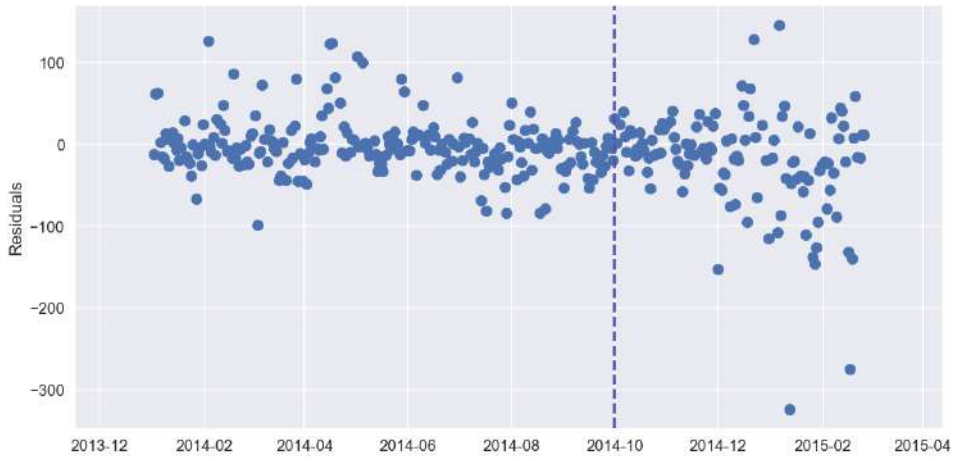

Рисунок 2.7 – Залишки прогнозування на тренувальній та валідаційній вибірках

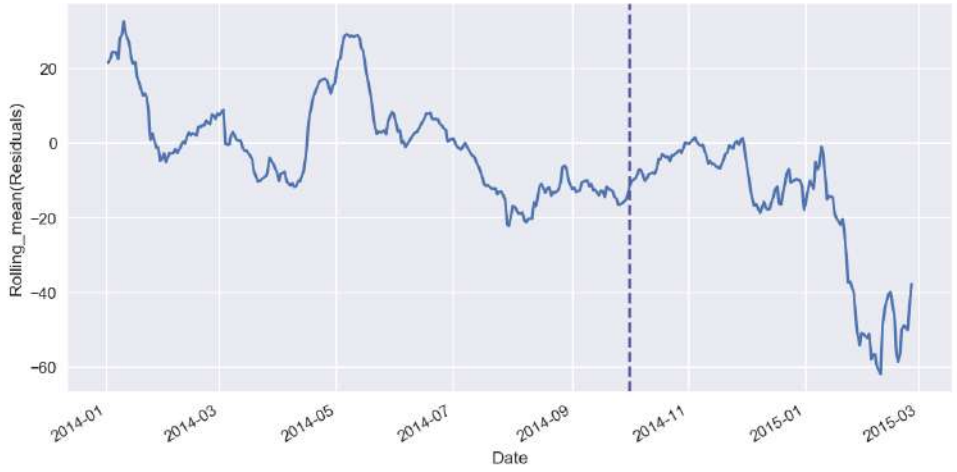

Рисунок 2.8 – Біжуче середнє для залишків прогнозування на тренувальній та валідаційній вибірках

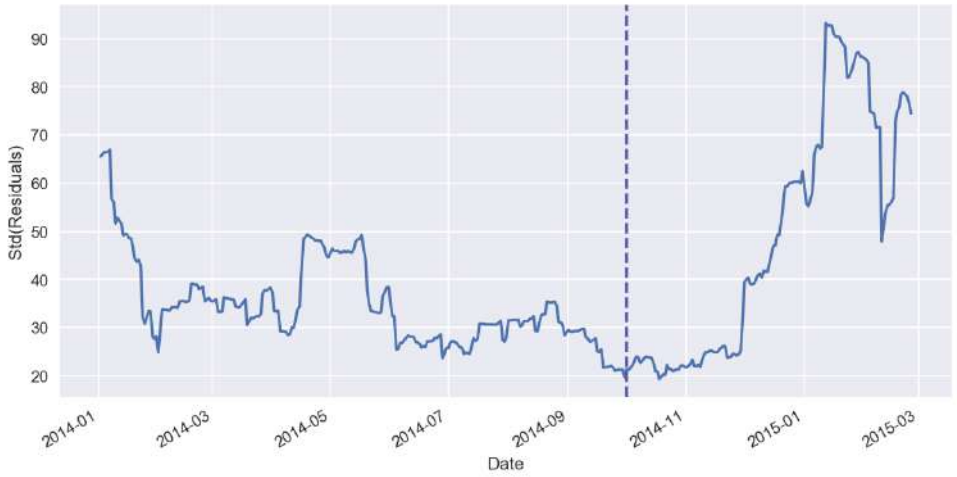

Рисунок 2.9 – Стандартне відхилення для залишків прогнозування на тренувальній та валідаційній вибірках

навчання використано такі додаткові ознаки у порівнянні із попередніми розглянутими випадками: середнє значення продажів за визначений часовий період історичних даних, змінні, які позначають державні вихідні та шкільні канікули, відстань від магазину аналізованої мережі до магазину конкурента, тип асортименту магазину. Алгоритм машинного навчання застосовано до вибірки даних, яка складається одночасно із різних часових рядів об'ємів продажів у різних магазинах [222]. На рис. 2.10 показано прогноз продажів на валідаційній вибірці для випадку історичних даних тренувальної вибірки за тривалий період часу (2 роки) для конкретного магазину, а на рис. 2.11 – прогноз для випадку тренувальних історичних даних за короткий період часу (3 дні) для одного і того ж заданого магазину. Як показують результати, у випадку короткого часового проміжку для тренувальних даних заданого часового ряду можна отримати навіть ще точніші результати. Ефект генералізації машинного навчання дозволяє здійснювати прогнози у випадку дуже малої кількості історичних даних продажів, що є важливим, наприклад, коли запускається у продаж новий продукт або відкривається новий магазин. Якщо необхідно передбачити продажі нових товарів, можна зробити експертну корекцію, помноживши отриманий часовий ряд прогнозу на залежний від часу деякий коефіцієнт для того, щоб врахувати перехідні процеси, наприклад, процес канібалізації продукту, коли попит на нові продукти заміщує попит на інші продукти.

2.2.3 Метод стекінгу моделей машинного навчання

Маючи різні моделі прогнозування з різними наборами ознак, доцільно поєднати ці моделі у прогнозний ансамбль моделей. Розглянемо підхід на основі стекінгу [75, 76, 77, 78, 79, 80] до побудови ансамблю прогнозних моделей. При такому підході результати прогнозів на валідаційній вибірці трактуються як вхідні регресійні змінні для моделей наступного рівня. Як модель наступного рівня, можна використати лінійну модель або модель на основі машинного навчання. Розглянемо двохрівневий стекінговий ансамбль моделей [222, 213]. На першому рівні ансамблю використано статистичну модель ARIMA, параметричну лінійну модель регресії LASSO, машинно-навчальні моделі на основі дерев рішень Random Forest та ExtraTree і модель

Рисунок 2.10 – Прогноз продажів для випадку історичних даних за тривалий період часу (2 роки), похибка=7.1%

Рисунок 2.11 – Прогноз продажів для випадку історичних даних за короткий період часу (3 дні), похибка=5.3%

Таблиця 2.1 – Похибки прогнозування різних моделей

Модель	**Валідаційна похибка**	**Позавибіркова похибка**
ExtraTree	14.6%	13.9%
ARIMA	13.8%	11.4%
RandomForest	13.6%	11.9%
LASSO	13.4%	11.5%
Neural Network	13.6%	11.3%
Stacking	12.6%	10.2%

глибокого навчання на основі нейронної мережі із повністтю з'єднаними шарами. На другому стекінговому рівні ансамблю використано модель на основі регуляризованої регресії LASSO [57], для якої функція похибок для регресії LASSO описується формулою (1.13). На рис. 2.12 показано прогнози часових рядів на валідаційних сетах, отриманих за допомогою використання моделей ансамблю. Вертикальна пунктирна лінія на рис. 2.12 розділяє валідаційний сет та позавибірковий набір даних, який не використовується в навчанні моделі та валідації (out-of-sample set). На такому позавибірковому наборі даних можна вирахувати похибки стекінгового ансамблю моделей. Розглянемо регуляризовану лінійну регресію LASSO як стекінгову модель на другому рівні ансамблю прогнозних моделей. Прогнози на валідаційних сетах моделей першого рівня розглядаються як регресійні змінні для лінійної моделі з регуляризацією LASSO. На рис. 2.13 показано значення стекінгових коефіцієнтів моделей першого рівня, отримані для стекінгової лінійної моделі на основі регресії LASSO. Лише три моделі першого рівня (ExtraTree, LASSO, Neural Network) мають ненульові коефіцієнти у регресійній моделі. Для інших випадків історичних даних часових рядів продажів результати можуть бути іншими, коли інші складові моделі можуть відігравати також важливу роль у прогнозуванні. У таблиці 2.1 показано похибки на валідаційній вибірці та на вибірці даних, які не входять у тренувальний та валідаційни сети (out-of-sample sets). Ці результати показують, що стекінговий підхід може підвищити точність на валідаційній вибірці та на позавибіркових даних, які не входять у тренувальний та валідаційний сети [222]. Як випливає із отриманих результатів, за рахунок розроблених методів стекінгового об'єднання різнотипних моделей у прогнозні ансамблі можна підвищити

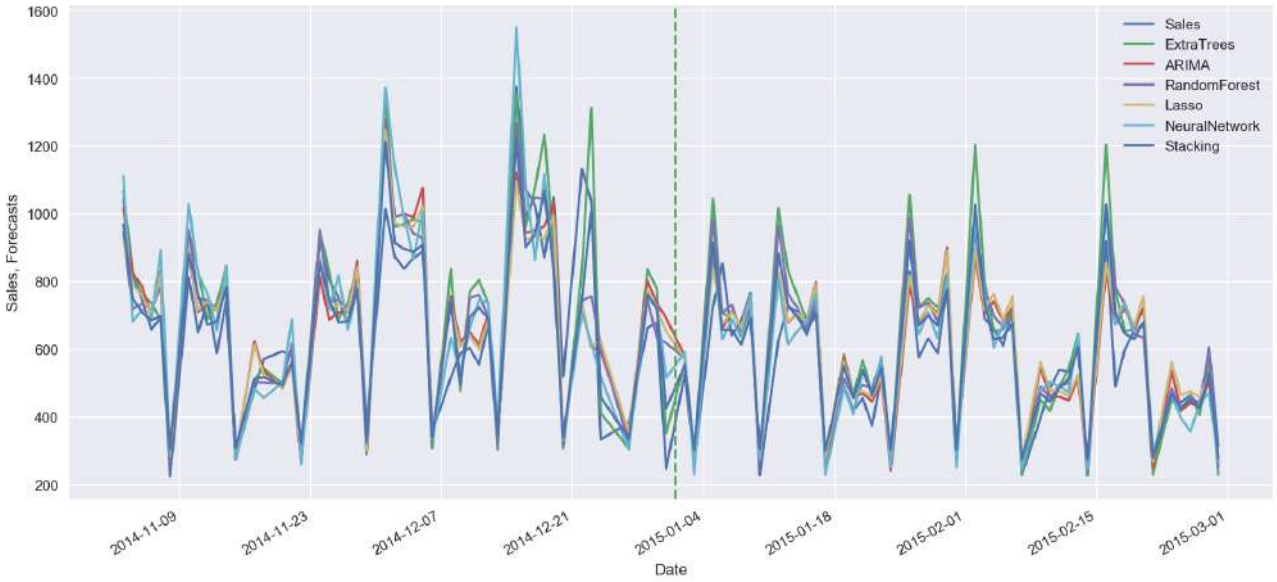

Рисунок 2.12 – Прогнозування часових рядів на валідаційній вибірці на основі різних моделей

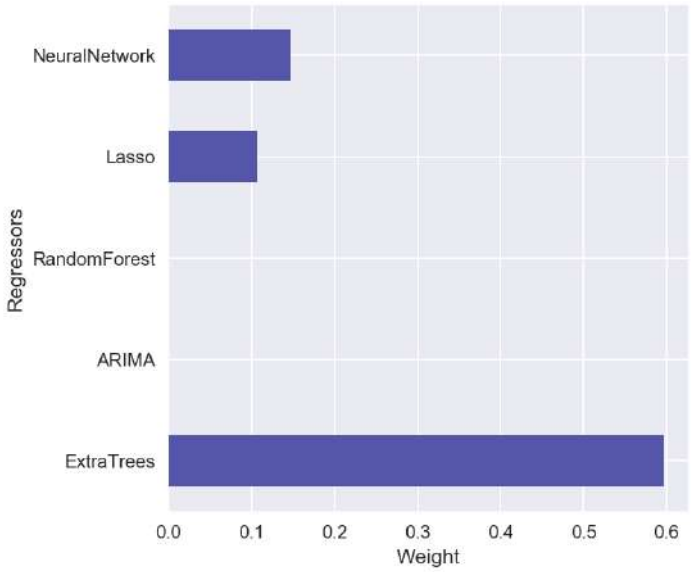

Рисунок 2.13 – Стекінгові коефіцієнти для прогнозних моделей

точність у задачах прогнозування за певних умов на 1-10% та зменшити кількість моделей у стекінговому ансамблі на 30% для певного класу задач.

Для пошуку нових підходів деякі компанії пропонують свої аналітичні задачі для змагань з інтелектуального аналізу даних, наприклад, на платформі Kaggle [87]. Одним із таких змагань було Gruppo Bimbo Inventory Demand [223]. Завданням цього змагання було передбачити попит на товари. Я був учасником міжнародної команди "The Slippery Appraisals", яка посіла перше місце на цьому змаганні. Деталі нашого переможного рішення розміщені на сайті Kaggle [224]. Рішення базувалось на трирівневій моделі, зображеній на рис. 2.14. На першому рівні використано багато одинарних моделей, більшість з яких базувалися на алгоритмі машинного навчання XGBoost [73]. Для другого стекінгового рівня використано дві моделі з пакету scikit-learn для мови програмування Python – модель ExtraTree, лінійну модель, а також нейронну мережу прямого поширення із повністю з'єднаними шарами. На третьому рівні моделі використано зважену суму результатів другого рівня, яку розрахували як $Sales = w_1 ET + w_2 LM + w_3 NN$, ET – результат прогнозування моделі ExtraTree, LM-результат прогнозування лінійної моделі, NN – результат прогнозування моделі на основі нейронної мрережі, w_1, w_2, w_3 – експертні вагові коефіцієнти, які вибирались на основі характеристик точності моделі на валідаційному наборі даних. Було побудовано багато нових ознак, найважливіші з яких базувалися на агрегуванні цільової змінної та її лагах з групуванням за різними факторами. Більше деталей нашого підходу можна знайти у [224]. Простий скрипт на мові R з одинарною моделлю машинного навчання для цієї задачі наведено у [225].

Отже, розглянуто різні методи прогнозування часових рядів. Використання регресійних підходів до прогнозування продажів часто можуть дати кращі результати порівняно з статистичними методами часових рядів. Одним з головних припущень методів регресії на основі машинного навчання є те, що патерни в історичних даних повторяться в майбутньому. Точність на валідаційного сеті даних є важливим показником для вибору оптимальної кількості ітерацій алгоритмів машинного навчання. Ефект генералізації моделі машинного навчання полягає у виявленні патернів у цілій вибірці

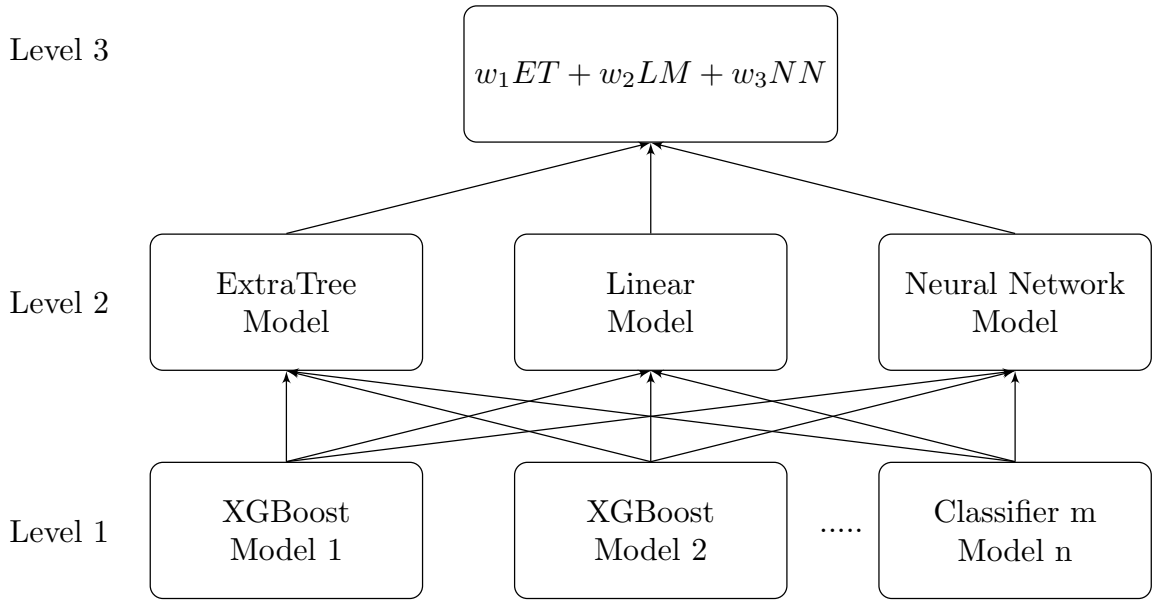

Рисунок 2.14 – Багаторівнева модель машинного навчання для прогнозування часових рядів продажів

даних. Цей ефект можна використовувати для прогнозування продажів, коли є невелика кількість історичних даних для заданих часових рядів продажів у випадку, коли запускається у продаж новий товар або відкривається новий магазин. У стекінговому підході, результати прогнозів багаторівневих моделей на валідаційних вибірках розглядаються як вхідні регресійні змінні для моделей наступного рівня. Як модель наступного рівня, використано регресію LASSO. Використання стекінгу дозволяє врахувати відмінності у результатах для різних моделей з різними наборами параметрів та ознак і підвищити точність на валідаційних, а також на позавибіркових даних, які не входять у тренінгові та валідаційні вибірки [222, 213, 226]. Використання регресії LASSO як стекінгової моделі дає можливість відібрати ефективні моделі першого рівня з ненульовими коефіцієнтами стекінгової регресії.

У Додатках розглянуто стохастичних патернів у часових рядах. Такі патерни можуть бути зумовлені факторами, не врахованими у прогнозній моделі.

2.2.4 Використання байєсівської регресії у прогнозній аналітиці часових рядів

Імовірнісні регресійні моделі можуть базуватися на байєсівській теоремі [59, 58, 227]. Імовірнісний підхід на основі байєсівського висновування у прогнозній аналітиці дає можливість отримувати щільність розподілу ймовірностей для цільової змінної [58, 59, 60]. Маючи таку функцію, можна зробити оцінку різних ризиків та обчислити величину VaR, яка дорівнює 5 %-му персентилю. Такий підхід дозволяє отримати апостеріорний розподіл параметрів моделі за допомогою умовної ймовірності та апріорного розподілу параметрів. Імовірнісний підхід є більш природним для таких стохастичних змінних, як часові ряди продажів. Різниця між байєсівським підходом і звичайним методом найменших квадратів полягає у тому, що в баєсівському підході невизначеність зумовлена параметрами моделі, на відміну від методу найменших квадратів, де параметри є постійні, а невизначеність зумовлена даними. У байєсівському виведенні можна використовувати інформаційні апріорні розподіли, які можуть бути задані експертом. Отже, результат можна розглядати як компроміс між історичними даними та експертним міркуванням. Це важливо у тих випадках, коли є мала кількість історичних даних. У байєсівській моделі можна розглядати цільову змінну з негаусовим розподілом, наприклад t-розподілом Стьюдента. Імовірнісний підхід дає можливість отримувати щільність розподілу ймовірностей для цільової змінної. Маючи таку функцію, можна оцінити ризики та обчислити величину ризику (VaR), яка дорівнює 5 %-му персентилю. Для розв'язання байєсівських моделей використовуються чисельні методи Монте-Карло. Семплювання Гіббса та Гамільтона є популярними методами пошуку апріорних розподілів для параметрів байєсівської моделі [59, 58, 227]. Байєсівське виведення (висновування) дозволяє реалізувати нелінійну регресію. Наприклад, у випадку трендів часових рядів із насиченістю можна використати модель логістичної кривої і знайти її параметри, використовуючи байєсівське висновування. Для розв'язання байєсівських моделей використовуються чисельні методи Монте-Карло. Розглянемо випадок нелінійної регресії для часового ряду із трендом, який виходить на

насичення [228]. Для моделювання розглянемо часові ряди продажів, які аналізувалися раніше. Таку модель можна описати так:

$$\begin{aligned} &log(Sales) \sim \mathcal{N}(\mu_{Sales},\, \sigma^2) \\ &\mu_{Sales} = \frac{a}{1 + exp(bt + c)} + \beta_{Promo} Promo + \\ &\beta_{Time} Time + \sum_j \beta_j^{wd} WeekDay_j \end{aligned} \tag{2.17}$$

Для аналізу було змодельовано часовий ряд, до якого було додано мультиплікативний тренд із насиченням. Чисельний аналіз моделі (2.17) було здійснено у системі *Stan* [59, 68, 227], використовуючи пакет *pystan* для мови *Python*. Результати моделювання наведено на рис. 2.15, де показано середні значення для розподілів прогнозованої величини у логарифмічному масштабі, а також розраховано характеристики ризику *VaR*, які визначалися як 5% перцентиль розподілу. На рис. 2.16 наведено функцію густини розподілу для коефіцієнта бінарного фактору Promo, на рис. 2.17 наведено коробкові графіки розподілів коефіцієнтів сезонності. Також можна використовувати байєсівське висновування для аналізу ієрархічних моделей, у яких деякі параметри моделі є загальними для всіх даних, а інші – різними для різних груп у вибірці даних. Наприклад, параметри сезонності можуть бути загальними для усіх товарів в аналітиці продажів, а вільний член у регресійній моделі може бути різним для різних магазинів. Ієрархічну модель можна описати у вигляді

$$\begin{aligned} &Sales \sim \mathcal{N}(\mu_{Sales},\, \sigma^2), \\ &\mu_{Sales} = \alpha(Store) + \beta_{Promo} Promo + \\ &\beta_{Time} Time + \sum_j \beta_j^{wd} WeekDay_j, \end{aligned} \tag{2.18}$$

де $\alpha(Store)$ – вільний член, для різних магазинів. Для числового аналізу розглянуто випадок із п'ятьма різними магазинами. Вважаємо тренди, промо-ефект та сезонність однаковими для всіх магазинів, а вплив конкретного магазину на продажі описується вільним членом регресії, тому цей параметр буде різним для різних магазинів. На рис. 2.18 показано

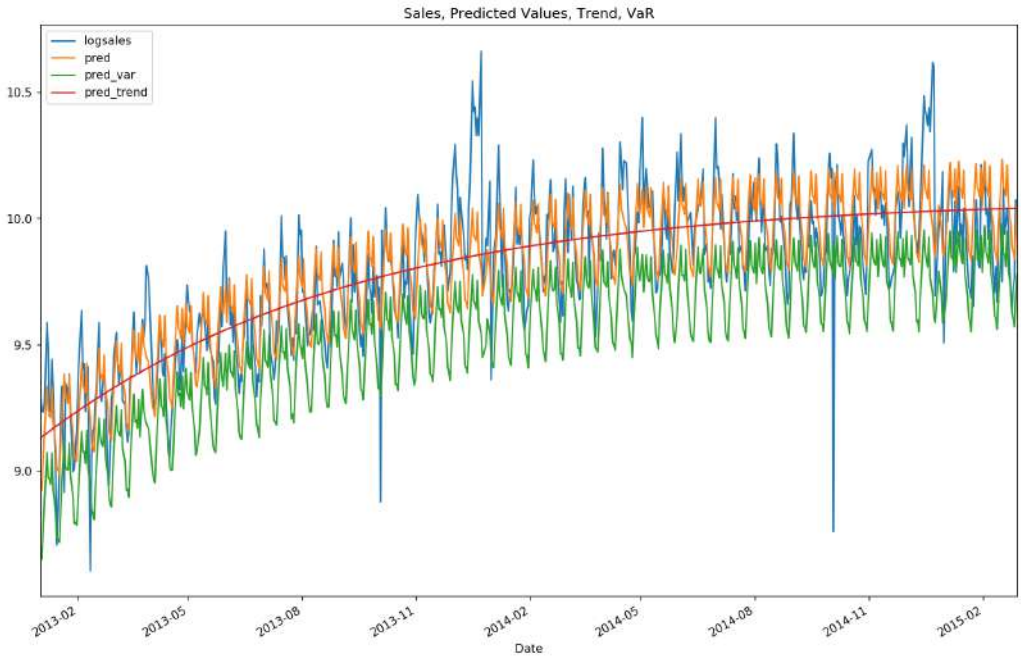

Рисунок 2.15 – Прогнозування часового ряду та нелінійного тренду

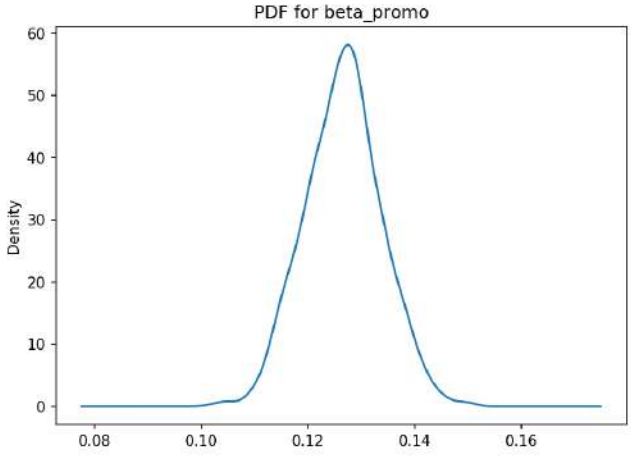

Рисунок 2.16 – Функція густини розподілу для коефіцієнта бінарного фактору Promo

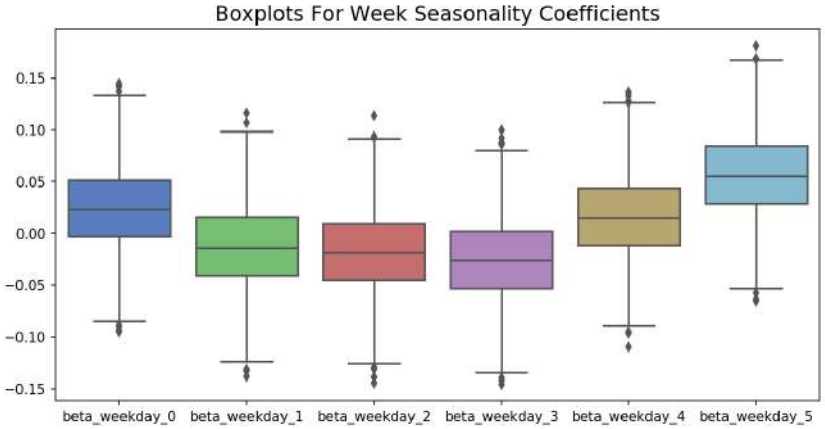

Рисунок 2.17 – Коробкові графіки розподілів коефіцієнтів сезонності

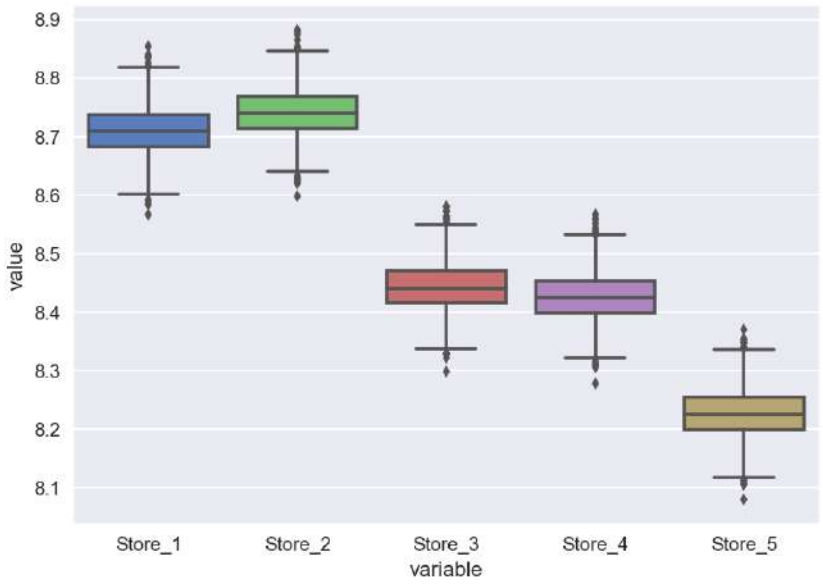

Рисунок 2.18 – Коробкові графіки розподілів для вільних членів різних часових рядів ієрархічної моделі

коробкові графіки для розподілу вільних членів. Дисперсія цих розподілів описує невизначеність впливу фактору заданого магазину на продаж. Ієрархічний підхід дозволяє використовувати таку модель при наявності короткострокових історичних даних для конкретних магазинів, наприклад, у випадку нових магазинів. На рис. 2.19 показано результати прогнозування у випадках використання дворічних історичних даних та лише п'ятиденних даних. Результати показують, що такі короткі історичні дані дозволяють правильно оцінити динаміку продажів. На рис. 2.20 показано коробкові графіки розподілів для вільних членів різних часових рядів ієрархічної моделі у випадку обмежених історичних даних заданого часового ряду. Отримані результати показують, що дисперсія для аналізованого магазину з короткими історичними даними стає більшою через невизначеність, викликану дуже короткими історичними даними для цього магазину. Отже, ієрархічна байєсівська модель дає можливість знаходити прогнозні значення цільової змінної у випадку коротких часових рядів за рахунок використання параметрів ієрархічної моделі, сформованих на основі інших подібних часових рядів, які належать до аналізованої вибірки. Нейронні мережі з шарами LSTM широко використовуються для прогнозування часових рядів. У процесі прийняття рішень важливо провести оцінку невизначеності результатів прогнозування. Для цього потрібно отримати функцію щільності

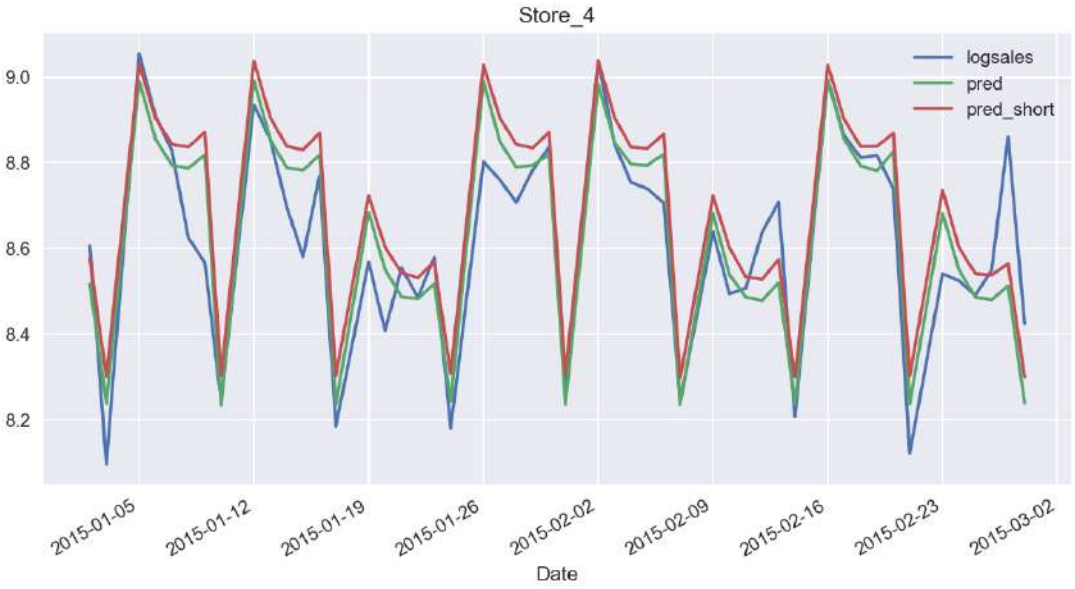

Рисунок 2.19 – Прогнозування на валідаційному сеті для заданого часового ряду із різним розміром історичних даних

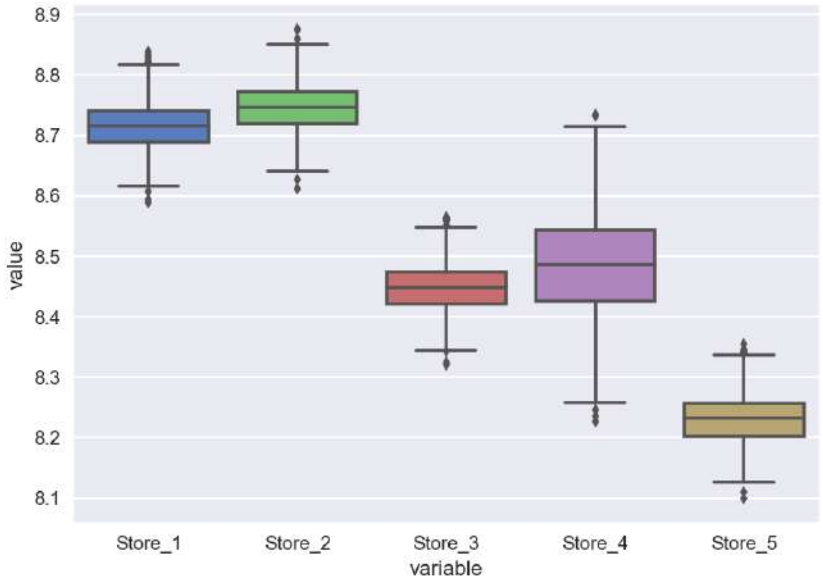

Рисунок 2.20 – Коробкові графіки розподілів для вільних членів різних часових рядів ієрархічної моделі у випадку обмежених історичних даних заданого часового ряду

розподілу ймовірностей (probability density function, PDF) цільових змінних. Враховуючи цю функцію, можна обчислити кількісні характеристики ризику, зокрема VaR, та отримати довірчий інтервал для прогнозування. Щільність розподілу ймовірностей для прогнозованої цільової змінної можна отримати за допомогою байєсівського підходу. У [91, 229, 230] показано, що шар Dropout, який використовується в процесі тренування нейронної мережі для зменшення ефекту перенавчання, можна використовувати для апроксимації байєсівського висновування.

2.2.5 Метод стекінгу прогнозних моделей часових рядів на основі байєсівської регресії

Прогнозні моделі можна об'єднувати в ансамблеву модель, використовуючи стекінг [75, 76, 77, 78, 79, 80]. У цьому підході результати прогнозування прогнозних моделей на валідаційній вибірці розглядаються як незалежні змінні регресійної стекінгової моделі. Прогнозні моделі розглядаються як перший рівень моделі прогнозного ансамблю. Стекінгова модель утворює другий рівень модельного ансамблю. Розглянемо використання байєсівської регресії для стекінгу прогнозних моделей часових рядів [231]. Використання байєсівського висновування для стекінгової регресії дає можливість отримати розподіли для коефіцієнтів регресії. За допомогою таких розподілів можна оцінити невизначеність прогнозних моделей першого рівня. Як прогнозні моделі для першого рівня ансамблю ми використовували такі моделі: *ARIMA, ExtraTree, RandomForest, LASSO, NeuralNetowrk.* Використання цих моделей для стекінгу за допомогою регресії LASSO було описано в [222]. Для стекінгу ми вибрали регресію з t-розподілом Стьюдента для цільової змінної як

$$\begin{aligned} y &\sim Student_t(\nu, \mu, \sigma), \\ \mu &= \alpha + \sum_i \beta_i x_i, \end{aligned} \tag{2.19}$$

де ν – параметр розподілу, який називається степінь вільності, i – індекс прогнозної моделі у стекінговій регресії, $i \in \{$ 'ARIMA', 'ExtraTree', 'RandomForest', 'LASSO', 'NeuralNetowrk' $\}$. Дані для чисельного

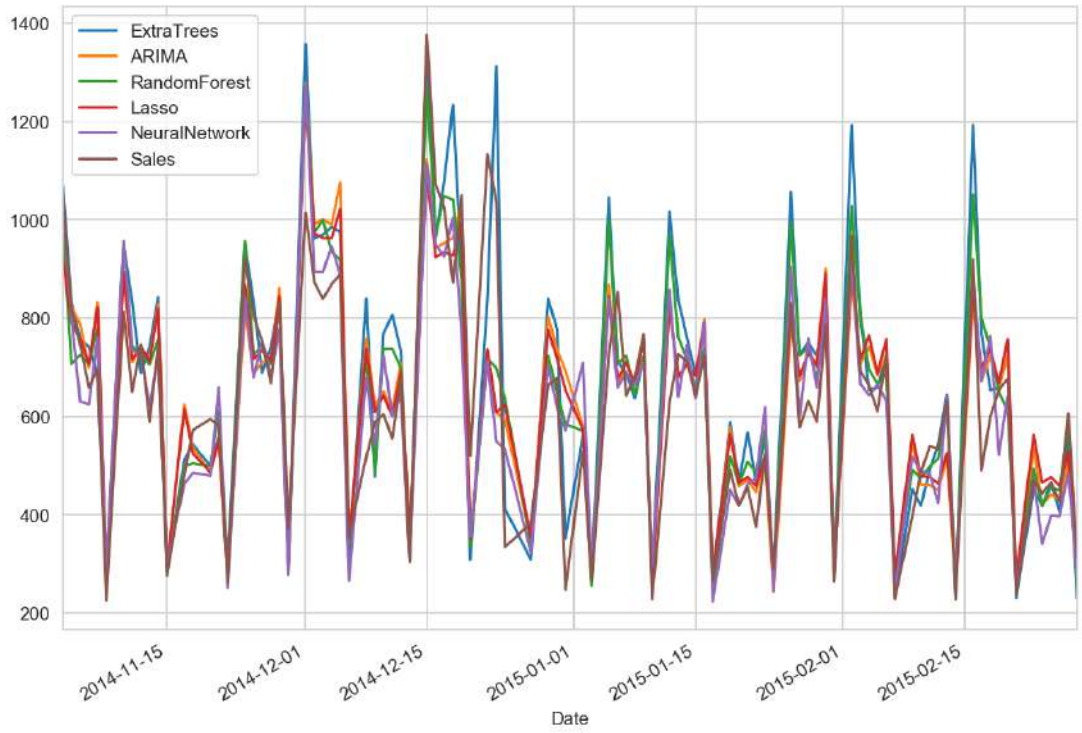

Рисунок 2.21 – Прогнозування із використанням різних моделей на валідаційному сеті

моделювання ймовірнісного стекінгу моделей сформовано на основі на історичних даних про продажі магазинів, взятих від Kaggle змагання 'Rossmann Store Sales' [214]. Для байєсівської регресії ми використовували платформу для статистичного моделювання Stan [227]. Аналіз проводився в середовищі Jupyter Notebook, використовуючи мову програмування Python та такі основні пакети Python *pandas* [215, 216], *sklearn* [217], *numpy* [218], *scipy* [232], *statsmodels* [233], *pystan* [227], *matplotlib* [220], *seaborn* [221]. Для проведення аналізу було середовище *Jupyter Notebook*. Було створено різні моделі прогнозування та розраховано прогнози на валідаційному сеті даних. Модель ARIMA було розраховано за допомогою пакета *statsmodels*, для моделювання нейронної мережі використовувався пакет *keras*, Random Forest і Extra Tree розраховувались за допомогою пакета *sklearn*. Підходи, які використані у цих розрахунках описано в [222]. На рис. 2.21 показані прогнози часових рядів на валідаційних сетах, отриманих за допомогою різних моделей. Результати прогнозування цих моделей на валідаційних сетах даних розглядають як незалежні змінні для регресії на другому рівні стекінгу ансамблю моделей. Для стекінгу прогнозних моделей ми розділили валідаційний сет на навчальний та тестовий. Для стекінгової регресії було

нормалізовано незалежні і цільові змінні за допомогою z-перетворення:

$$z_i = \frac{x_i - \mu_i}{\sigma_i}, \tag{2.20}$$

де μ_i – середнє значення, σ_i – стандартне відхилення. Апріорні розподіли для параметрів α, β у байєсівській регресійній моделі (2.19) вважаються гауссовими із середніми значеннями, рівними 0, і стандартним відхиленням, рівним 1. Ми розділили валідаційний сет на навчальний та тестовий за допогою часового фактору. Параметри апріорного розподілу можна коригувати, використовуючи оцінки прогнозування на тестових наборах або використовуючи експертний підхід у випадку малої кількості історичних даних. Для оцінки невизначеності коефіцієнтів регресії використано коефіцієнт варіації який визначається як співвідношення між стандартним відхиленням та середнім значенням розподілу коефіцієнтів моделі:

$$v_i = \frac{\sigma_i}{\mu_i}, \tag{2.21}$$

де v_i – коефіцієнт варіації, σ_i – стандартне відхилення, μ_i – середні значення для розподілу коефіцієнтів регресії i-ї моделі . Враховуючи, що μ_i може бути від'ємним, проаналізовано абсолютне значення коефіцієнта варіації $|v_i|$. Для оцінки результатів використано відносну середню абсолютну похибку (RMAE) та середньоквадратичну помилку (RMSE). Відносна середня абсолютна похибка (RMAE) розглядалася як відношення між середньою абсолютною похибкою (MAE) та середніми значеннями цільової змінної:

$$RMAE = \frac{E(|y_{pred} - y|)}{E(y)} 100\%. \tag{2.22}$$

Середньоквадратична похибка (RMSE) розраховувалась як:

$$RMSE = \sqrt{\frac{\sum_i^n (y_{pred} - y)^2}{n}}. \tag{2.23}$$

Дані з прогнозами різних моделей на валідаційному сеті було розділено на навчальну вибірку (48 зразків) та вибірку тестування (50 зразків)

за датою. Було використано байєсівську регресію з t-розподілом Стьюдента для цільової змінної. У результаті розрахунків отримано такі оцінки: RMAE(train)=12.4%, RMAE(test)=9.8%, RMSE(train)=113.7, RMSE(test)=74.7. На рис. 2.22 показано реальні значення та середні значення прогнозованої цільової змінної на валідаційних та тестових вибірках. Вертикальна пунктирна лінія розділяє навчальні та тестові вибірки даних. На рис. 2.23 показано щільність розподілу ймовірностей (PDF) для вільного члена регресії. Спостерігається додатнє зміщення цього розподілу. Це викликано тим, що алгоритми машинного навчання застосовано до нестаціонарних часових рядів. Якщо нестаціонарний тренд невеликий, його можна компенсувати за допомогою лінійної регресії на валідаційному сеті. На рис. 2.24 показано коробкові графіки для розподілу значень параметрів моделі. На рис. 2.25 показано абсолютні значення коефіцієнтів варіації для регресійних параметрів різних складових моделей. Ми також розглянули випадок з обмеженнями для регресійних коефіцієнтів, такими, щоб ці коефіцієнти були лише додатніми. Отримано подібні результати: RMAE(train)=12.9%, RMAE(test)=9.7%, RMSE(train)=117.3, RMSE(test)=76.1. На рис. 2.26 показано коробкові графіки регресійних коефіцієнтів для цього випадку. Усі моделі мають подібні середні значення та коефіцієнти варіації. Можна спостерігати, що характеристики похибок RMAE та RMSE на тестовому сеті можуть бути подібними до таких похибок на навчальному сеті даних. Це говорить про той факт, що байєсівська регресія не перенавчається на навчальному сеті, порівняно з алгоритмами машинного навчання, для яких є характерним істотне перенавчання на навчальній вибірці, особливо у випадках невеликих обсягів історичних даних. Ми вибрали найкращу стекінгову модель ExtraTree і провели байєсівську регресію лише з цією моделлю. Отримано такі оцінки: RMAE(train)=12.9%, RMAE(test)=11.1%, RMSE(train)=117.1, RMSE(test)=84.7. Ми також спробували виключити модель ExtraTree з стекінгової регресії та провели байєсівську регресію з рештою моделей без ExtraTree. У цьому випадку отримано такі оцінки: RMAE (train)=14.1%, RMAE (test)=10.2%, RMSE(train)=139.1, RMSE(test)=75.3. На рис. 2.27 показано коробкові графіки для модельних коефіцієнтів регресії, рис. 2.28 показує

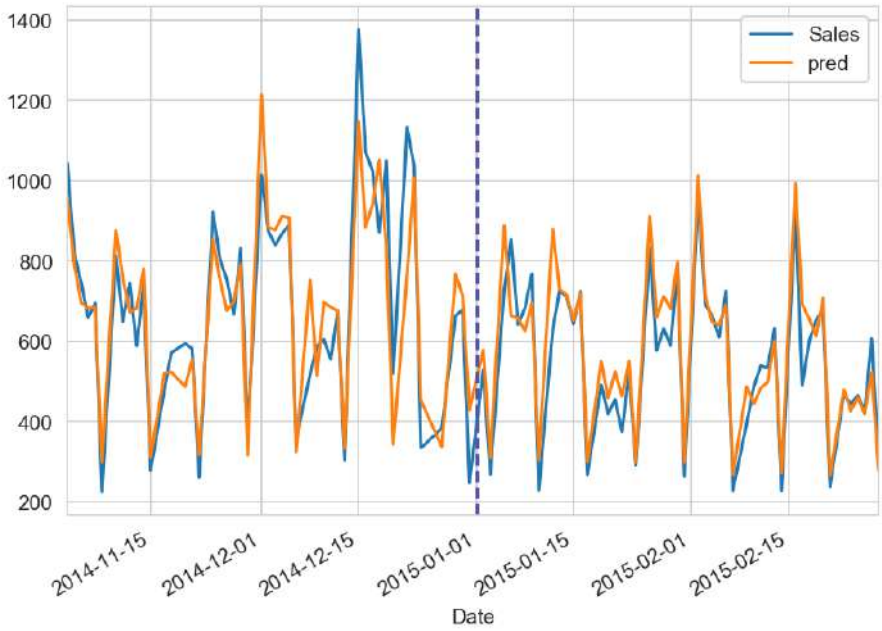

Рисунок 2.22 – Середні значення для реальних та прогнозованих продажів на валідаційному та тестовому сетах даних

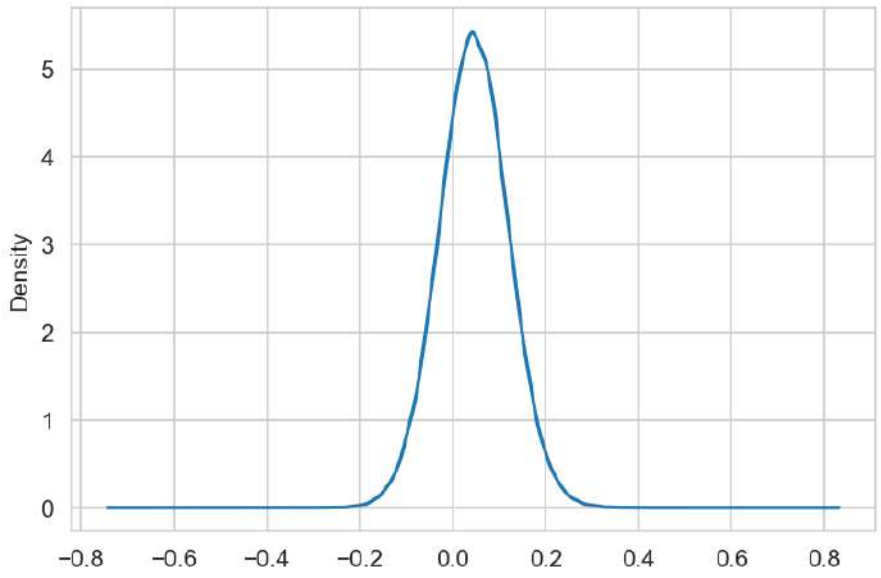

Рисунок 2.23 – Щільність розподілу ймовірностей вільного члена стекінгової регресії

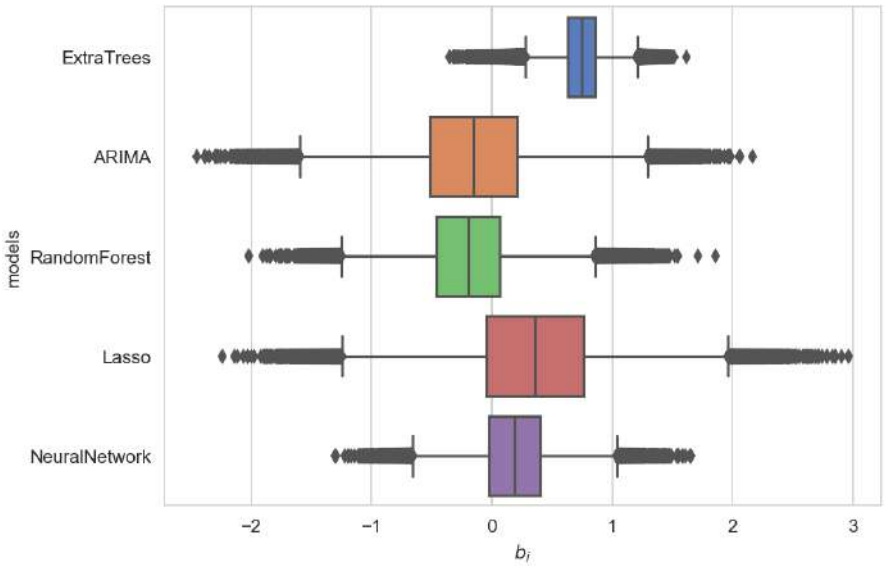

Рисунок 2.24 – Коробкові графіки для розподілу значень параметрів моделі

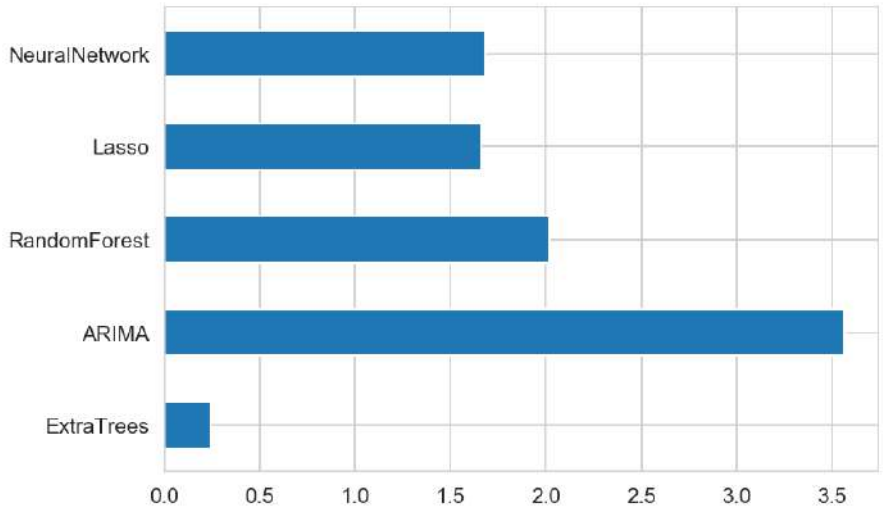

Рисунок 2.25 – Абсолютні значення коефіцієнтів варіації для регресійних коефіцієнтів різних складових моделей

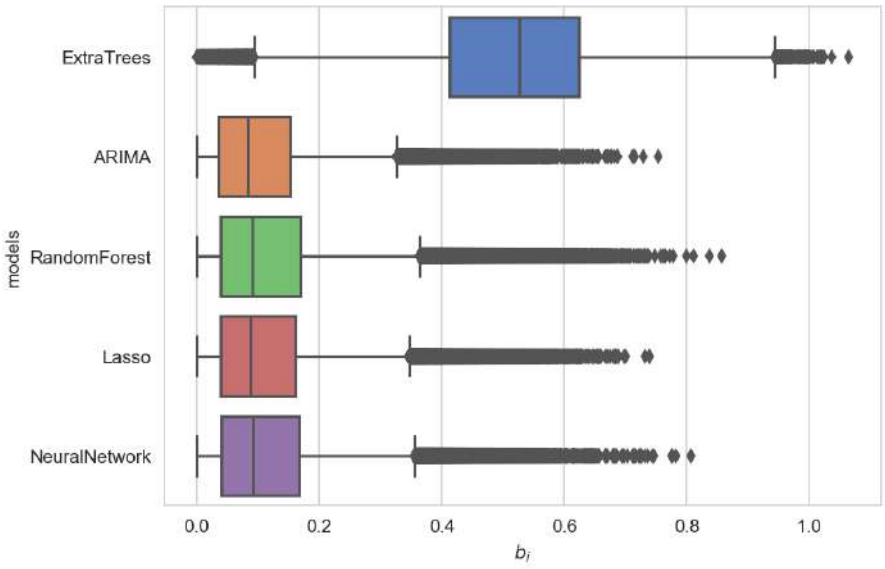

Рисунок 2.26 – Коробкові графіки для розподілу значень регресійних коефіцієнтів моделі

абсолютні значення коефіцієнтів варіації параметрів регресії стекінгових моделей для даного випадку. У результаті отримано гірші результати на тестовому наборі. У той же час ці моделі мають подібний вплив, і тому вони потенційно можуть забезпечити стабільніші результати в майбутньому через можливу зміну якості ознак. Моделі із стохастичним шумом можуть знижувати точність на великих наборах даних тренувань, у той же час вони вносять свій вклад у випадку малих сетів історичних даних. Ми розглянули випадок із невеликою кількістю історичних даних для навчання – 12 зразків. Щоб отримати стабільні результати, для параметру ν t-розподілу Стьюдента в байєсівській регресійній моделі (2.19) встановлено значення, що дорівнює 10. У результаті отримано такі оцінки: RMAE(train)=5.0%, RMAE(test)=14.2%, RMSE(train)=37.5, RMSE(test)=121.3. На рис. 2.29 показано середні значення часових рядів для реальних та прогнозованих продажів на валідаційних та тестових сетах. На рис. 2.30 показано коробкові графіки регресійних коефіцієнтів у випадку малого тренінгового сету. На рис. 2.31 показано абсолютне значення коефіцієнта варіації для складових моделей стекінгу. У цьому випадку можна побачити, що інші моделі починають грати важливу роль у порівнянні з попередніми випадками, а модель ExtraTree не відіграє вирішальну роль. Також здійснено розрахунки із різними параметрами інформативних апріорних розподілів для параметрів моделі, змінено параметр σ на 0.15 для t-

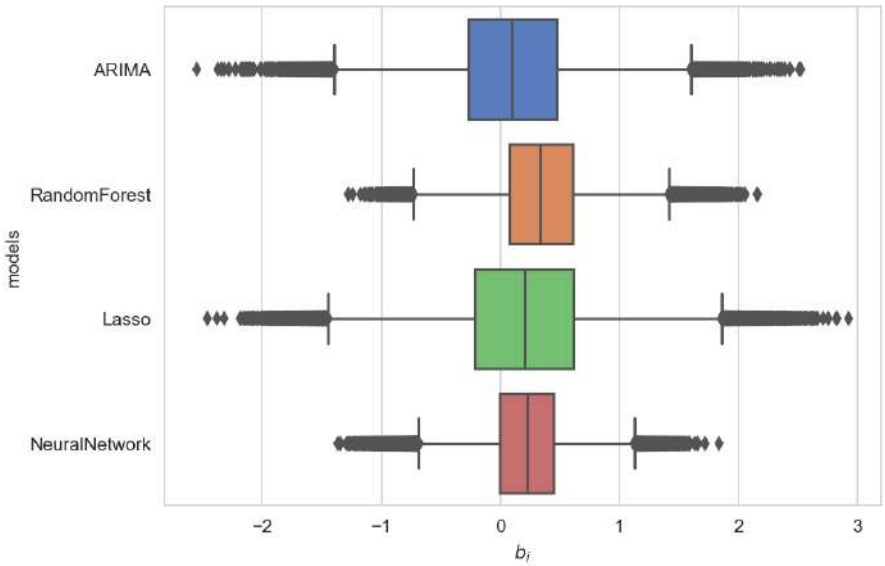

Рисунок 2.27 – Коробкові графіки для модельних коефіцієнтів регресії

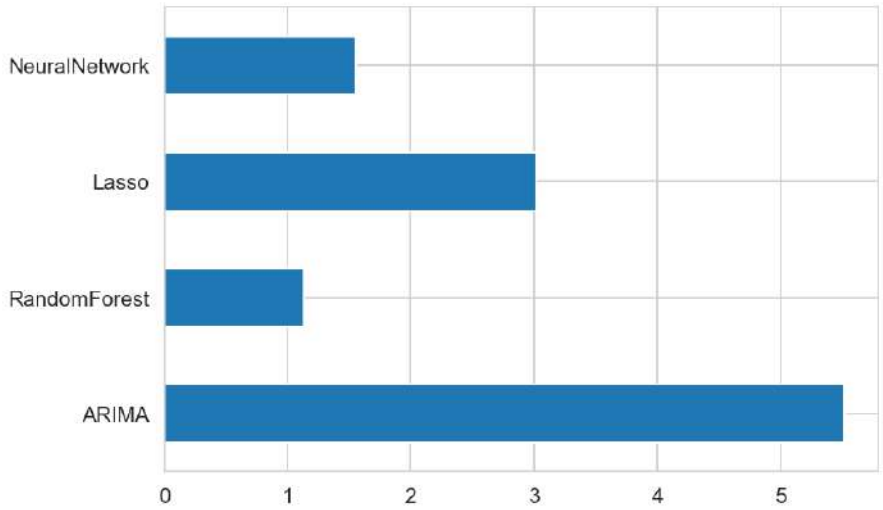

Рисунок 2.28 – Абсолютні значення коефіцієнтів варіації параметрів регресії стекінгових моделей

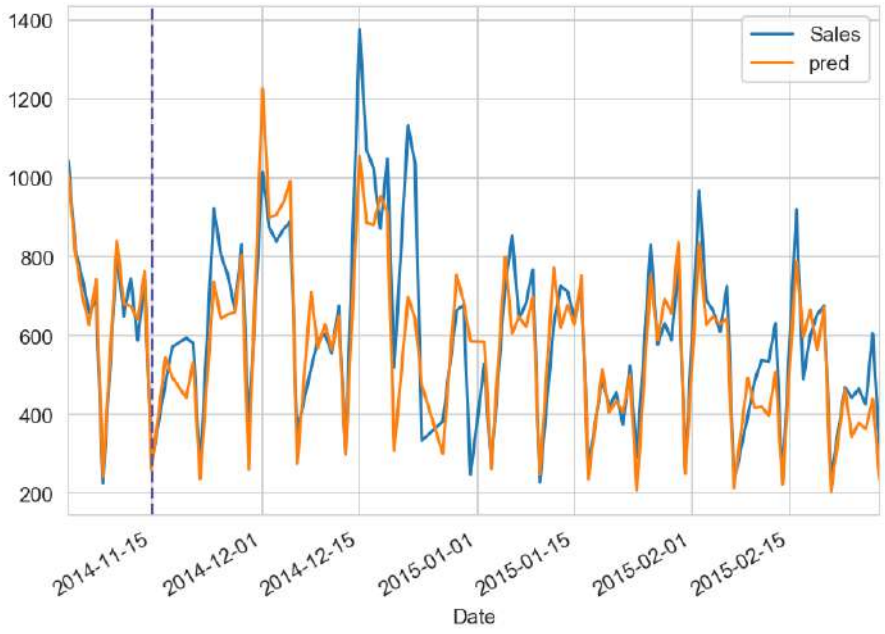

Рисунок 2.29 – Середні значення часових рядів для реальних та прогнозованих продажів на валідаційних та тестових сетах

розподілу Стьюдента цільової змінної. У результаті отримано покращені оцінки на тестовому наборі: RMAE(train)=7.0%, RMAE(test)=12.3%, RMSE(train)=54.3, RMSE(test)=109.9.

Отже, розглянуто дворівневий ансамбль прогнозних моделей для часових рядів. Для прогнозування на першому рівні ансамблю моделей було використано такі моделі як ARIMA, Neural Network, Random Forest, Extra Tree. На другому стекінговому рівні ансамблевої моделі було здійснено байєсівську регресію результатів прогнозування моделей на валідаційному сеті. Цей підхід дає можливість отримати розподіли для регресійних коефіцієнтів моделей першого рівня прогнозного ансамблю і оцінити невизначеність, внесену кожною моделлю у результат стекінгу. Інформація про ці розподіли дозволяє вибрати оптимальний набір моделей стекінгу, враховуючи знання із предметної області, у якій проводиться прогнозна аналітика. Імовірнісний підхід для стекінгу прогнозних моделей дозволяє зробити оцінку ризиків та невизначеності для прогнозів, що є важливим у процесі прийняття рішень. Моделі із зашумленими ознаками можуть знижувати точність на великих тренувальних сетах даних, у той же час вони сприяють задовільним результатам у випадку малих за об'ємом тренувальних сетів. Використання байєсівського висновування

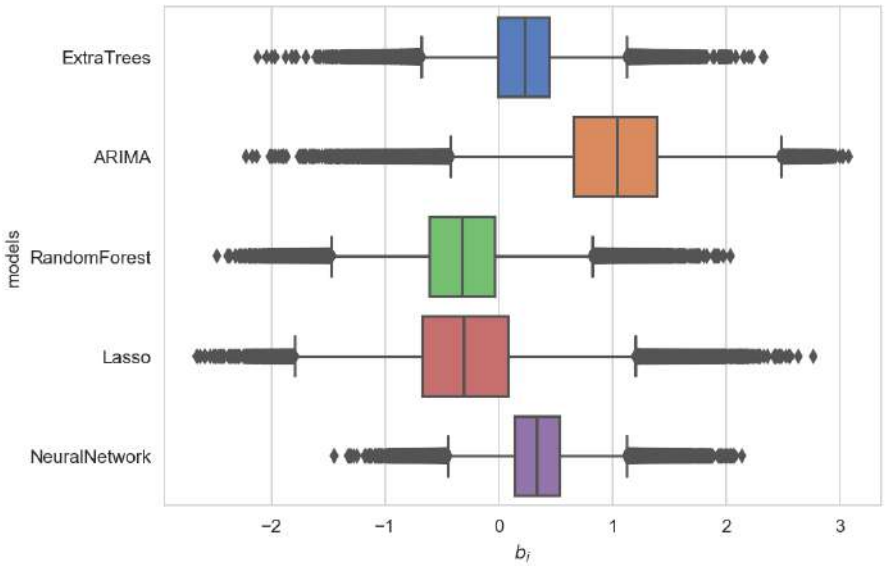

Рисунок 2.30 – Коробкові графіки регресійних коефіцієнтів у випадку малого тренінгового сету

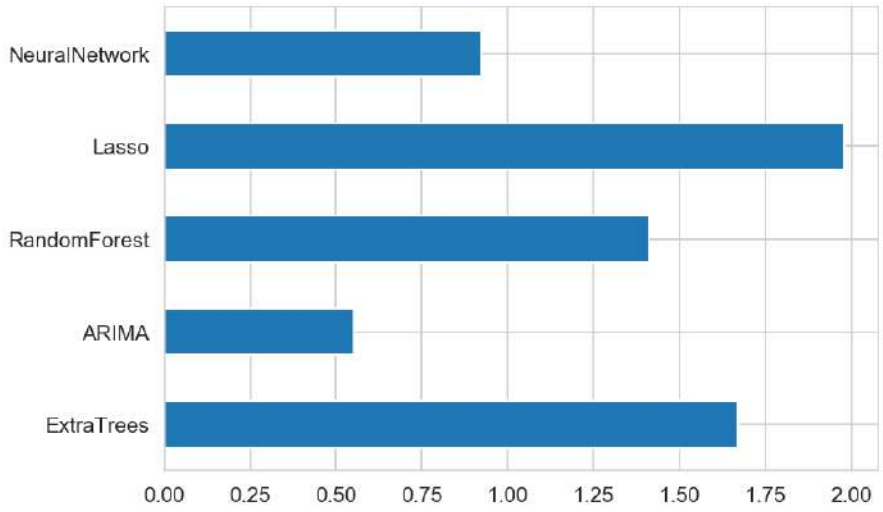

Рисунок 2.31 – Абсолютне значення коефіцієнта варіації для складових моделей стекінгу

для стекінгової регресії може бути корисним у випадках невеликих сетів даних та допоможе експертам вибрати набір моделей для стекінгу і оцінити різного типу ризики та невизначеності у прогнозуванні. Вибір кінцевих моделей для стекінгу може здійснюватися експертом, який враховує різні фактори, такі як невизначеність кожної моделі на рівні стекінгової регресії, кількість даних для навчання та тестування, стабільність моделей. При байєсівській регресії можна отримати кількісний показник невизначеності, який може бути корисною інформацією при виборі моделі для побудови стекінгового ансамблю моделей. Експерт також може задати апріорні інформативні розподіли для коефіцієнтів стекінгової регресії, з урахуванням закономірностей та знань із предметної області. У результаті проведених досліджень показано, що байєсівський підхід до стекінгової регресії може дати інформацію про невизначеність прогнозних моделей. Використовуючи цю інформацію та предметні знання, експерт може вибрати моделі для отримання стійкого стекінгового ансамблю прогнозних моделей.

2.3 Методи аналізу фінансових часових рядів на основі різнотипних консолідованих даних

Розглянемо аналіз фінансових часових рядів на прикладі моделювання ціни біткоїна та на прикладі впливу кризи, зумовленої пандемією COVID-19, на фондовий ринок.

2.3.1 Моделювання ціни біткоїна з використанням експертної корекції

Однією з головних цілей в аналітиці біткоїна є прогнозування цін. Існує багато факторів, які впливають на цінову динаміку. Найважливіші з них це: взаємодія між попитом і пропозицією, привабливість для інвесторів, фінансові та макроекономічні показники, технічні показники, такі як складність, кількість створених недавно блоків тощо. Дуже важливий вплив на ціну криптовалюти мають тренди у соціальних мережах та пошукових системах. Використовуючи ці фактори, можна створити регресійну прогнозну модель для ціни на біткойн на основі історичних даних.

У статті [234] показано, що перегляди сторінок, пов'язаних із біткоїнами у Вікіпедії, корелюють із динамікою ціни біткоїна та відображають інтереси потенційних інвесторів до криптовалюти. Тренди Google пошуку ключових слів, пов'язаних з біткоїнами, показують різні ефекти – зацікавленість інвесторів, активність спекулянтів тощо. У [235] було проаналізовано ціну на біткоїн. У статті [236] досліджено різні драйвери ціни біткоїна. У [237] показано значну кореляцію між ціною біткоїна та пошуковими трендами у соціальних мережах та Web. У [238] досліджено, що біткоїн має багато схожого з золотом і доларом. У [239] досліджено використання байєсівської регресії для аналізу цін на біткоїн. Вплив новин на поведінку інвесторів вивчається у [240]. Економіка біткоїна, його поведінка і видобуток розглядаються у [241, 242]. Поведінка ринку біткоїнів, особливо динаміка цін є предметом різних досліджень [243, 234]. У роботі [243] аналізуються різні фактори, що впливають на ціну біткойна. Особливістю біткоїна є те, що ця криптовалюта не випускається і не контролюється ні фінансовими, ні політичними установами, такими як Центральний банк чи уряд. Біткоїн видобувається без впливу економічних факторів. У [243] проаналізовано економіку формування ціни біткоїна. На динаміку цін найбільше впливає спекулятивна поведінка інвесторів. Один з головних драйверів біткойна – новини в Інтернеті. Відповідно до теорії ефективного ринку, фондові та фінансові ринки не передбачувані, оскільки вся наявна інформація вже відображена у ціні акцій. Але сьогодні домінування ефективної теорії ринку не таке очевидне. Деякі впливові вчені стверджують, що ринок може бути частково передбачуваним [244]. Для прогнозування ринку часто використовують поведінкові та психологічні теорії. Деякі економісти вважають, що історичні ціни, новини, активність соціальних мереж містять патерни, які дозволяють частково передбачити фінансовий ринок. Такі теорії та підходи розглянуто в огляді [244].

Розглянемо підхід до побудови регресійної моделі прогнозування ціни на біткоїн за допомогою експертної корекції шляхом додавання деякої коригувальної змінної [245]. У такому підході передбачається, що досвідчений експерт може скоригувати модель, опираючись на свій досвід. Регресійними незалежними змінними в моделі є історичні дані, які описують

статистику валюти біткоїн, процеси видобутку, тренди у пошуках в Google, динаміку відвідування сторінок Вікіпедії. Як статистичну характеристику біткоїна було обрано $total\_bitcoins$ – загальну кількість вже видобутих біткоїнів, $price$ – середню ціну в доларах США серед основних бірж біткоїнів, $volume$ – загальний об'єм долара на торгах на основних біржах біткоїна. Як інформацію про видобуток було обрано $difficulty$, що є відносним показником того, наскільки важко знайти новий блок. За мережеву активність було обрано $n\_unique\_addresses$, що є загальною кількістю унікальних адрес, які використовуються в блокчейні біткоїна. Часові ряди згаданих вище змінних було взято з сайту *Bitcoin.info*. Як фактори формування цін розглянуто Google тренди ключового слова "bitcoin" та кількість відвідувань сторінки Вікіпедії 'cryptocurrency'. Часові ряди обраних ознак показано на рис. 2.32. Цільова змінна $price$ та всі регресійні змінні розглядаються в логарифмічній шкалі, тобто $price' = ln(price + 1)$, $x'_i = ln(x_i + 1)$. Після логарифмічної трансформації всі регресійні змінні були нормалізовані шляхом віднімання середніх значень, та ділення на стандартне відхилення. Регресійну модель можна зобразити у такому вигляді:

$$\begin{aligned} price' = \alpha + \beta_{gtrend} \cdot gtrend' + \\ \beta_{wiki\_cryptocurrency} \cdot wiki\_cryptocurrency' + \\ \beta_{difficulty} \cdot difficulty + \\ \beta_{n\_unique\_addresses} \cdot n\_unique\_addresses' + \\ \beta_{total\_bitcoins} \cdot total\_bitcoins' + \\ \beta_{volume} \cdot volume' \end{aligned} \tag{2.24}$$

Для чисельного моделювання було використано Python з пакетами *pandas* [215, 216], *sklearn* [217], *numpy* [218], *scipy* [232], *pystan* [227], *matplotlib* [220], *seaborn* [221], *quandl*. Для проведення аналізу було обрано середовище *Jupyter Notebook*. Щоб отримати часові ряди відвідувань сторінок Вікіпедії, було використано пакет Python *mwviews*. Для знаходження коефіцієнтів α, β_i, було використано лінійну регресію з регуляризацією LASSO з пакету python scikit-learn. На рис. 2.33 показано дійсну динаміку ціни на біткоїн і ціну, передбачену регресійною моделлю 2.24. На рис. 2.34

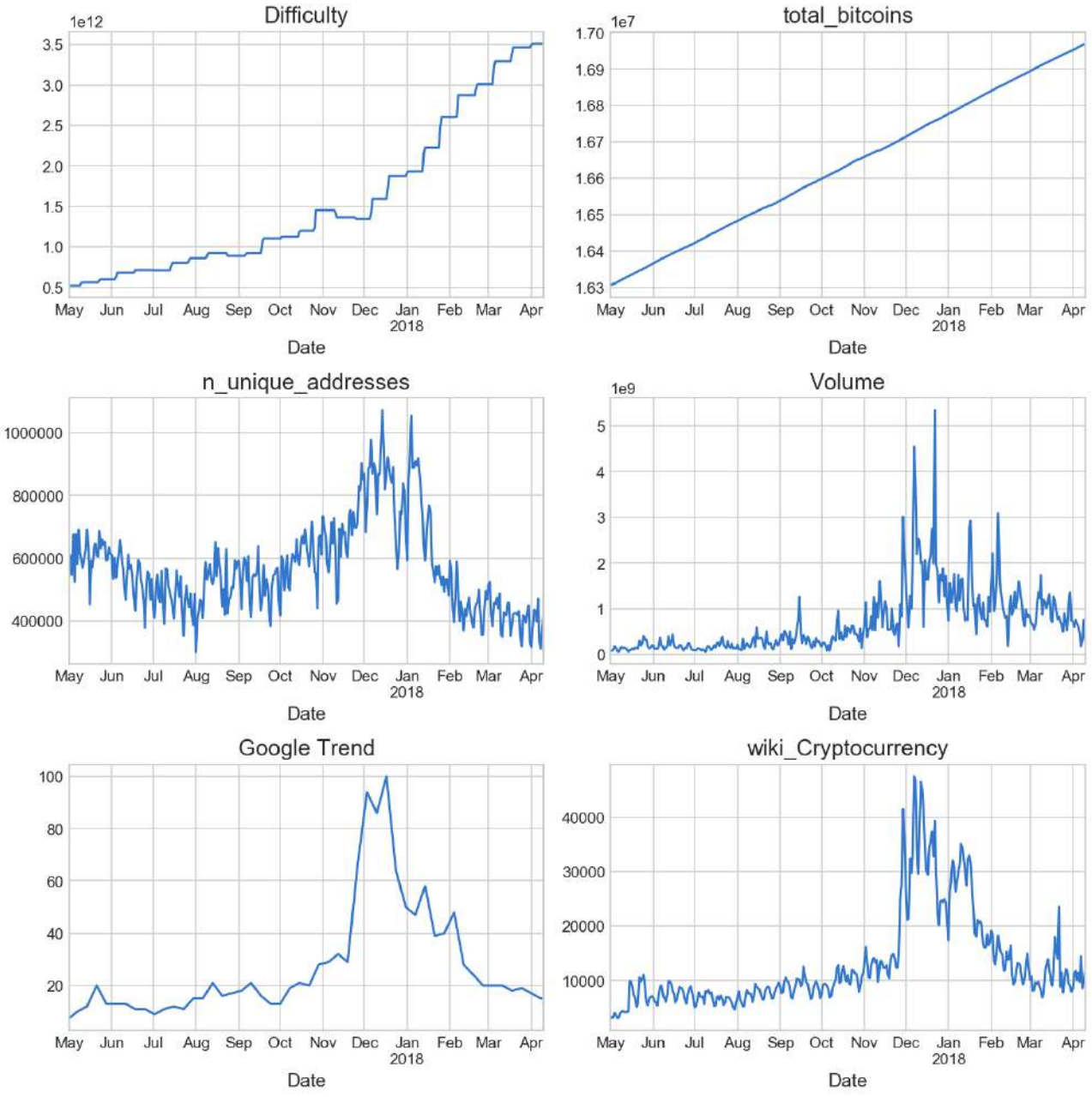

Рисунок 2.32 – Часові ряди ознак

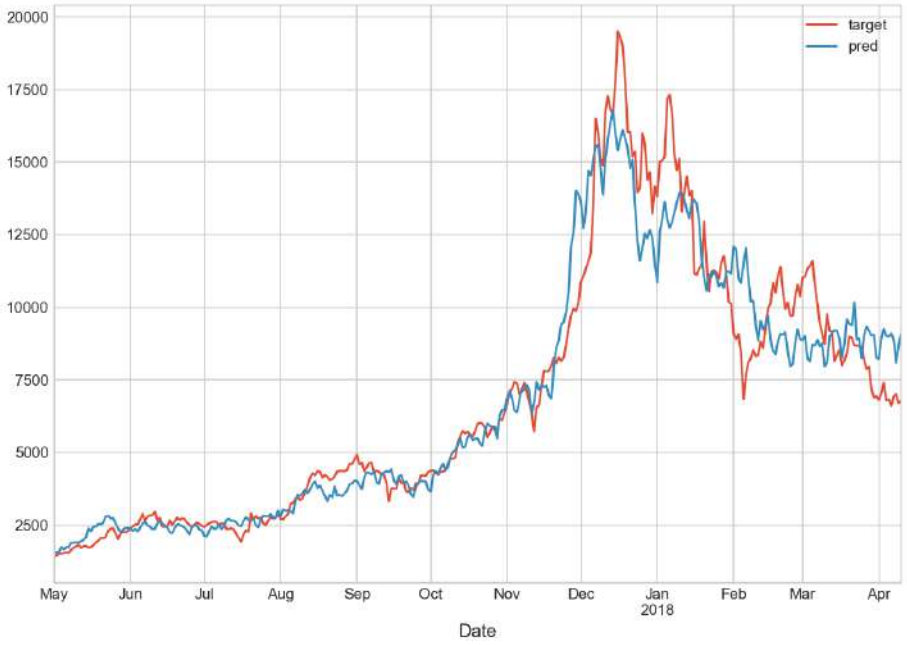

Рисунок 2.33 – Часова динаміка реальної та прогнозної ціни біткоїна

показано значення регресійних коефіцієнтів ознак в аналізованій моделі. Результати вказують на важливість ознаки, яка описує тренд Google для пошуку ключового слова Bitcoin та важливість ознаки, яка характеризує перегляди сторінок Вікіпедії про криптовалюту. Для цієї моделі було отримано значення похибки RMSE = 1277,8. На рис. 2.33 показано, що в окремих часових періодах прогнозована ціна є вищою порівняно з реальною ціною, а у деяких – нижчою. На рис. 2.35 показано співвідношення реальної та прогнозованої цін. Результати показують, що це співвідношення має періодичні коливання, які описують відхилення прогнозів від реальних значень. Це можна пояснити наявністю певних факторів, які впливають на ціну і які не включені у модель (2.24). Це можуть бути фактори складної поведінки інвесторів. Припустимо, що досвідчений експерт розуміє такий тип поведінки. У результаті він може пояснити динаміку відхилення прогнозування регресійної моделі (2.24) від реальних значень. Експертну корекцію в моделі можна застосувати за допомогою уведення додаткової регресійної незалежної змінної у регресійній моделі (2.24). Ця змінна описує описує динаміку відхилення моделі. Для опису такої динаміки відхилення,

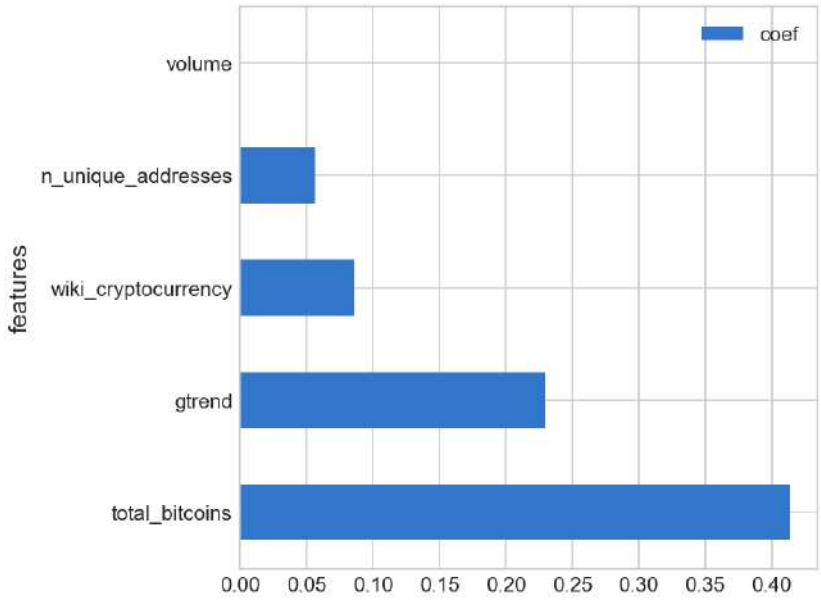

Рисунок 2.34 – Значення регресійних коефіцієнтів ознак

експерт повинен визначити локальні екстремуми у часових рядах відхилень, які є поворотними точками для трендів відхилень. Ми припускаємо, що експерт може визначити такі моменти правильно, спираючись на власний досвід. Часовий ряд такої можливої експертної корекції показано на рис. 2.36. Результати розрахунку регресійної моделі (2.24) з доданою змінною експертного корегування показано на рис. 2.37. У цьому випадку ми отримали значення похибки RMSE=856,4. На рис. 2.38 показано коефіцієнти регресії у випадку включення змінної експертної корекції в регресійну модель. Отримані результати показують, що додавання змінної експертної корекції покращує точність регресійної моделі.

Для ймовірнісного підходу, який дає можливість отримати оцінки ризику, можна використовувати підхід на основі байєсівського висновування. Байєсівська регресія – це метод, який має переваги, коли потрібно враховувати негаусівську статистику аналізованих процесів [59]. Для байєсівського моделювання було використано програмне забезпечення *Stan* [227] з пакетом Python *pystan* [227]. Щоб зробити модель стійкою до викидів та аномальних значень, можна розглядати t-розподіл Стьюдента для ціни біткоїна. Розподіл Стьюдента подібний до нормального розподілу, але має важчі "хвости". Розглянемо ціну біткоїна як стохастичну змінну, яка

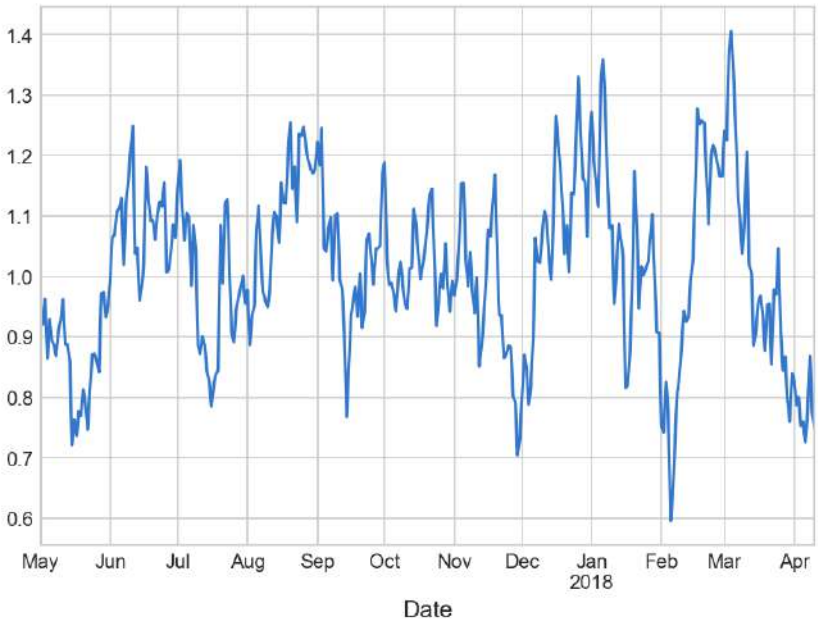

Рисунок 2.35 – Відношення реальної ціни на біткоїн до прогнозованої ціни

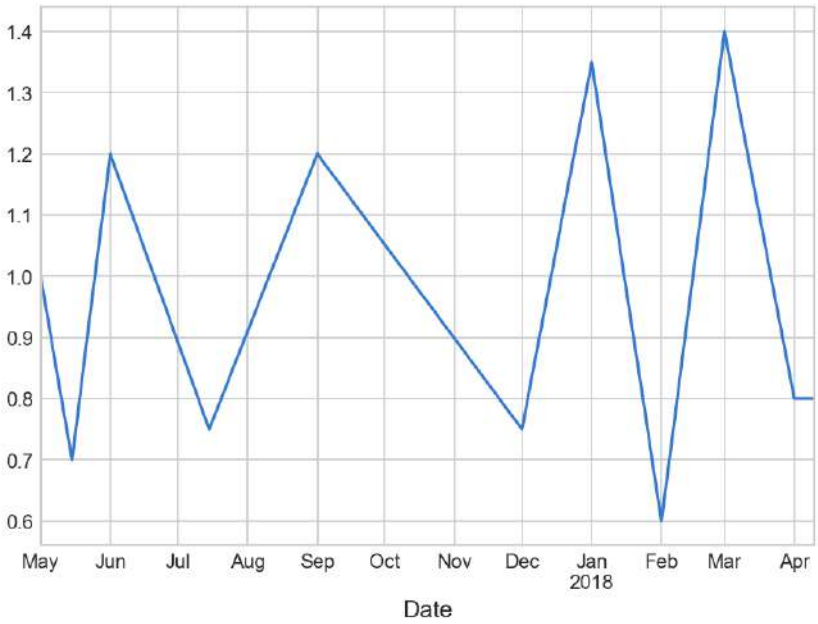

Рисунок 2.36 – Часовий ряд змінної, яка описує експертну корекцію

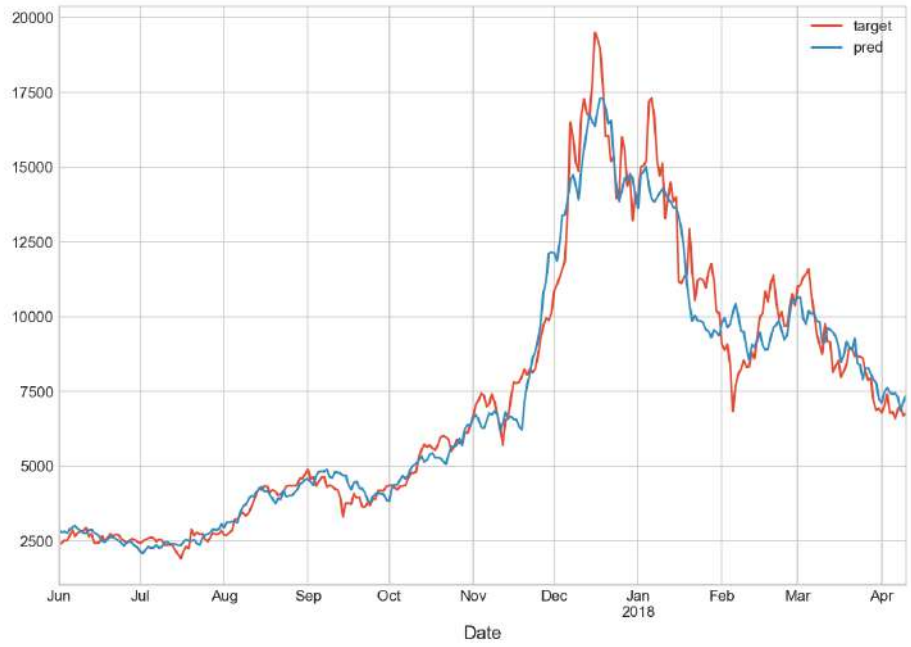

Рисунок 2.37 – Динаміка та прогнозування ціни біткоїна з експертною корекцією

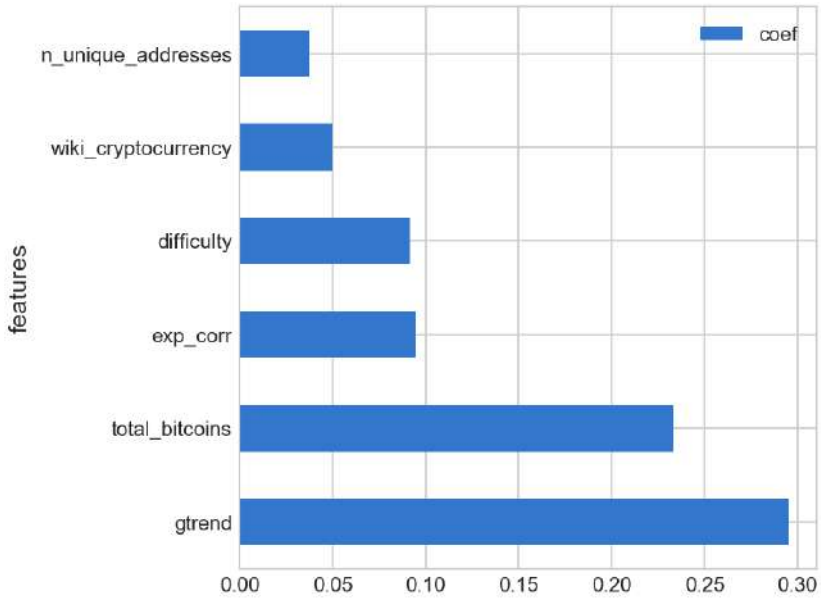

Рисунок 2.38 – Коефіцієнти ознак регресійної моделі з експертною корекцією

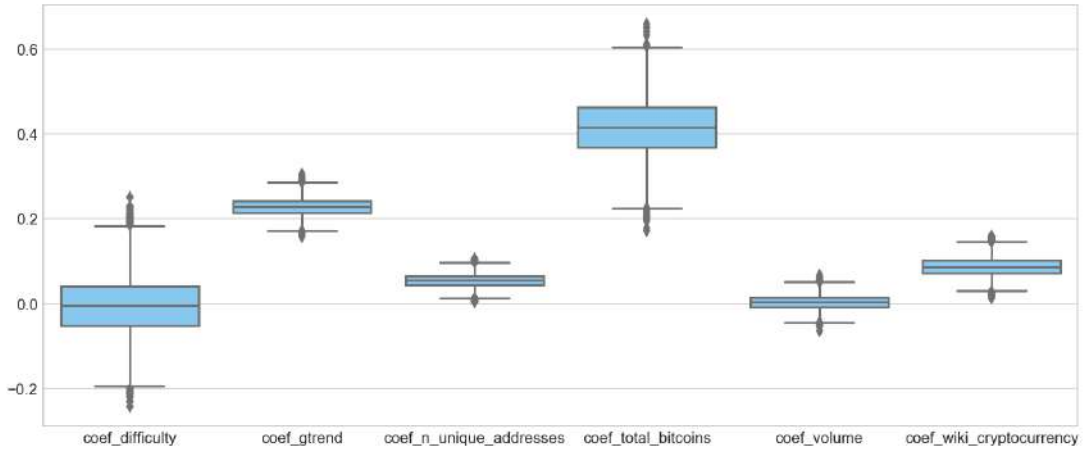

Рисунок 2.39 – Коробкові графіки коефіцієнтів регресійної моделі

розподіляється за допомогою t-розподілу Стьюдента

$$price'_s \sim Student_t(\mu, \sigma, \nu),$$

де ν, μ, σ – відповідні параметри t-розподілу Стьюдента. У баєсівському підході, вважаємо, що $\mu = price'$, де $price'$ визначається регресійною моделлю (2.24). На рис. 2.39 показано коробкові графіки для отриманих параметрів регресійної моделі. На рис. 2.40 показано коробкові графіки для параметрів у разі додавання змінної експертної корекції. Для параметра масштабу σ, значення 0.14 було отримано у разі відсутності змінної експертної корекції та значення 0.1 було отримано у разі додавання такої змінної. Це показує, що додавання змінної експертної корекції у байєсівську регресійну модель зменшує ширину кривої функції густини ймовірнісного розподілу для цільової змінної.

Отже, розглянуто лінійну модель для ціни біткоїна, яка включає в себе регресійні ознаки, які базуються на статистиці біткоїна, характеристиках процесів видобутку біткоїна, трендах пошукових запитів Google, візитах на сторінок Вікіпедії. Патерни відхилення регресійної моделі прогнозування від реальних цін біткоїна є простішими у порівнянні із часовими рядами ціни біткоїна. Припускаємо, що досвідчений експерт може передбачити цю закономірність. Поєднуючи регресійну модель та експертну корекцію, можна отримати кращі результати, ніж за допомогою регресійної моделі без змінної, яка описує експертну корекцію або лише на основі експертних міркувань.

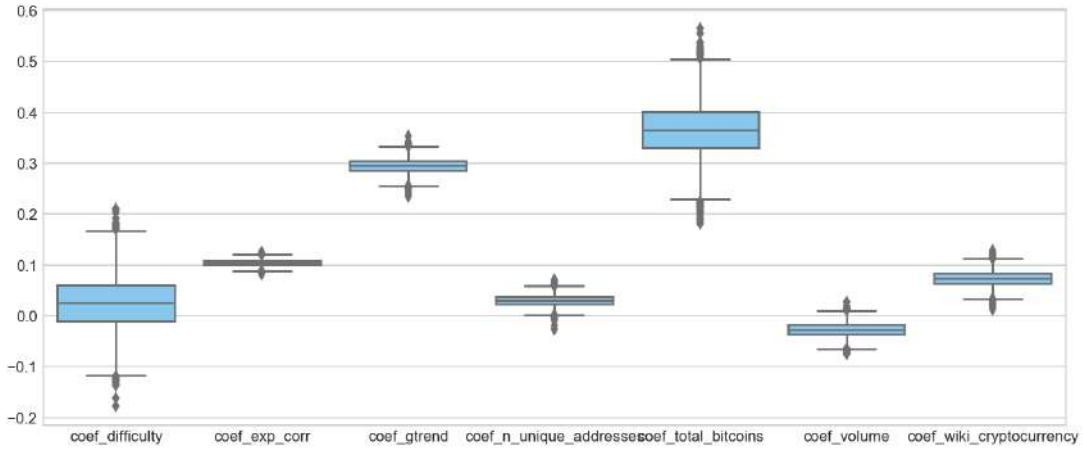

Рисунок 2.40 – Коробкові графіки для параметрів моделі із змінною експертної корекції

Припускається, що експерт може розуміти комплексні патерни часових рядів, які базуються на поведінковій теорії, економіці та політиці. Ці патерни неможливо вловити за допомогою даних історичних часових рядів, оскільки вони існують протягом короткого проміжку часу. Після опублікування чи обговорення цих патернів між інвесторами, вони швидко зникають згідно з положеннями теорії про ефективний ринок, оскільки починають враховуватися широким колом аналітиків та інвесторів. Отримані результати показують, що коректне експертне визначення часових поворотних точок у функції експертної корекції для відхилень регресійної моделі від реальних значень може суттєво покращити точність прогнозу ціни біткоїна. У запропонованому підході експерт повинен визначити часові поворотні точки, що описують відхилення регресійної моделі на основі порівняння історичних даних з часовим рядом реальних цін. За допомогою байєсівського висновування можна використовувати ймовірнісний підхід, використовуючи розподіли з "товстими хвостами" та враховувати викиди і аномальні значення у часовому ряді ціни біткоїна [245]. Маючи функції густини розподілу ймовірності для ціни біткоїна та для параметрів моделі, можна здійснити оцінку ризиків, розраховуючи кількісну характеристику ризику VaR (Value at risk).

Рисунок 2.41 – Часові ряди індексів фінансового ринку та цін на акції

2.3.2 Вплив кризи, зумовленої пандемією COVID-19, на часові ряди фондового ринку

Вплив різних факторів можна описати за допомогою альтернативних даних, таких як характеристики трендів пошуку, активності користувачів у соціальних мережах тощо. Розглянемо вплив поширення COVID-19 на зміни на фінансовому ринку [246]. На рис. 2.41 показано часові ряди індексів фінансового ринку та деяких цін на акції в період найбільшого впливу COVID-19 на фінансовий ринок. Як альтернативні дані, розглянуто часовий ряд відвідувань сторінок Вікіпедії, пов'язаних із COVID-19. Такі

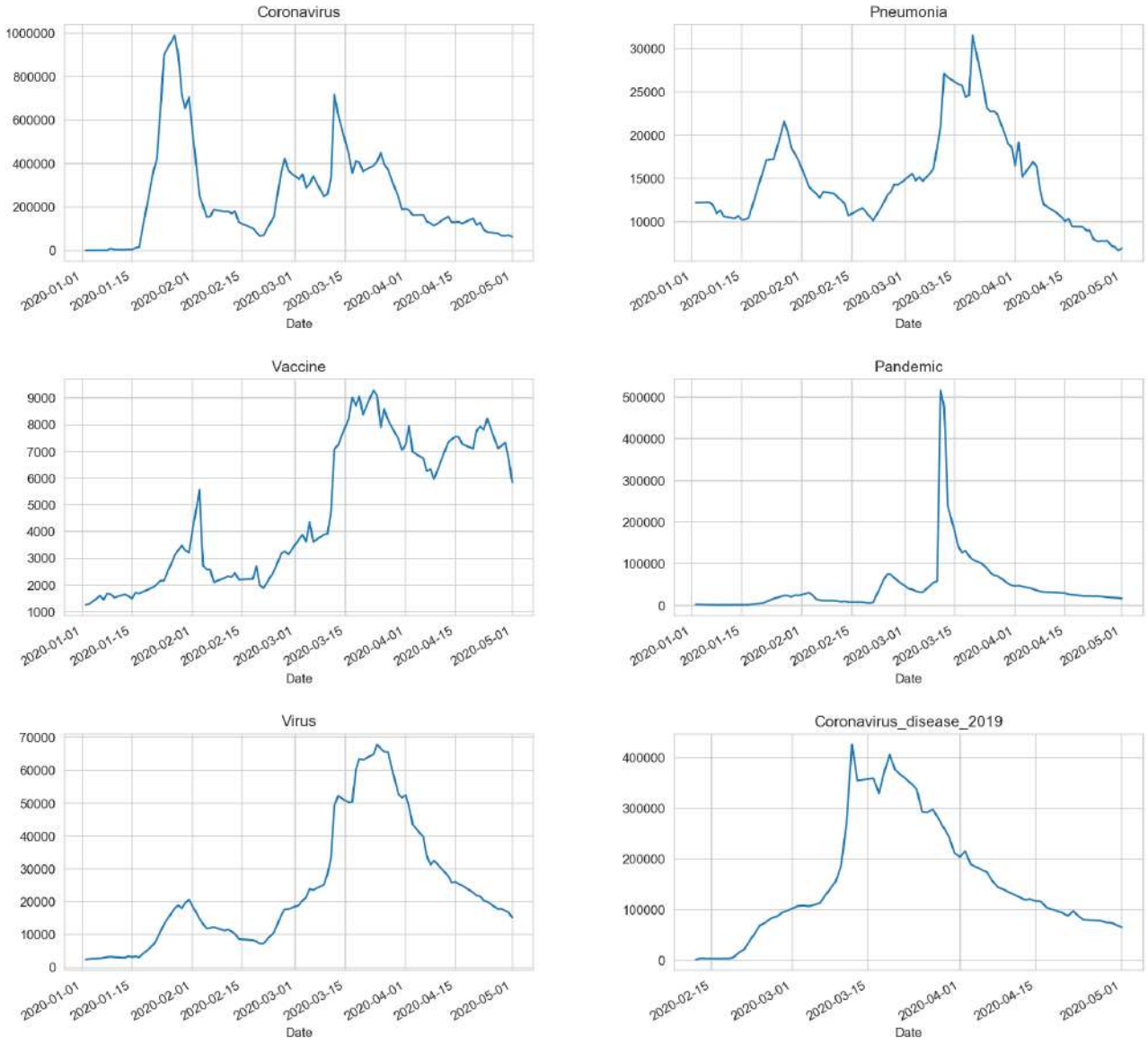

Рисунок 2.42 – Часові ряди відвідувань сторінок Вікіпедії пов'язаних із COVID-19

часові ряди показано на рис. 2.42. Баєсівський регресійний підхід у прогнозній аналітиці дозволяє отримати функцію щільності розподілу ймовірностей (PDF) для параметрів моделі і таким чином дозволяє зробити оцінку невизначеності факторів, які мають вплив на цільову змінну. Для аналізу розглянуто часовий період ['2020-02-15', '2020-05-01']. Для байєсівської регресії використано платформу Stan [227] для статистичного моделювання. Як регресійні ознаки використано z-оцінки часових рядів кількості відвідувань сторінок Вікіпедії. Як цільову змінну використано z-оцінки індексу S&P-500. На коефіцієнти лінійної регресії було накладено обмеження, щоб вони не були більші за 0. На рис. 2.43 показано розраховані

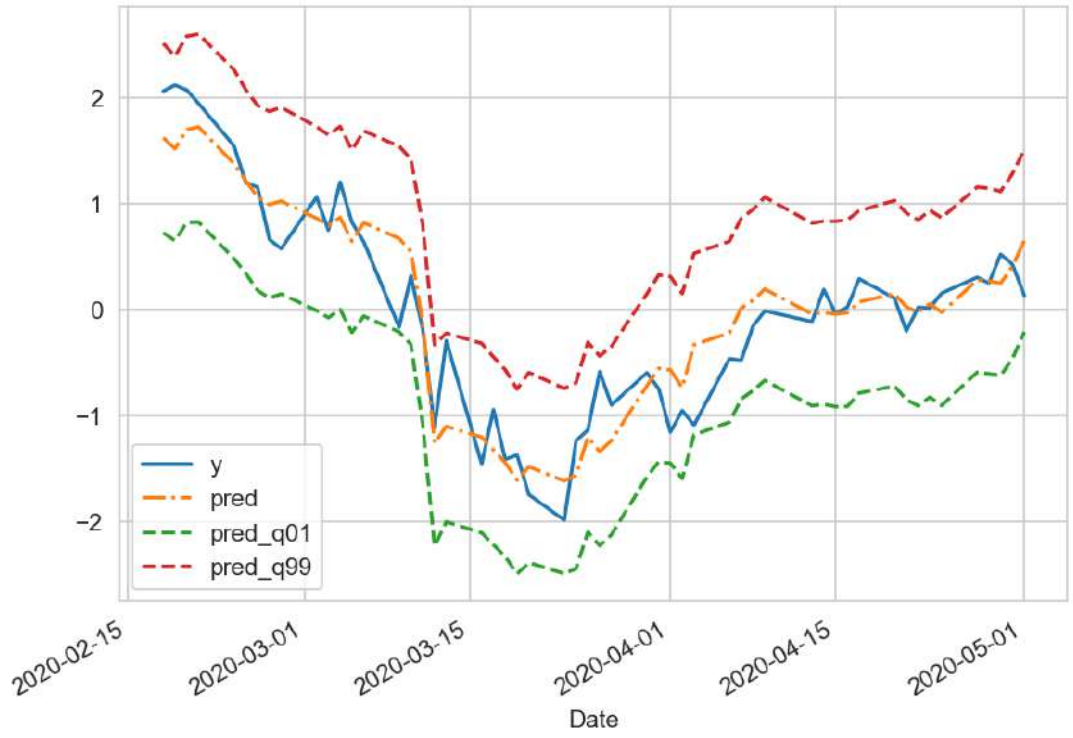

Рисунок 2.43 – Середні значення та 0.01, 0.99 квантилі розподілу значень прогнозів для індексу S&P-500

середні значення та 0.01, 0.99 квантилі розподілу значень прогнозів для індексу S&P-500. На рис. 2.44 показано коробкові графіки для розподілу значень коефіцієнтів лінійної регресії. На рис. 2.45 показано коефіцієнт варіації для аналізованих ознак, який дорівнює відношенню між стандартним відхиленням та абсолютними середніми значеннями коефіцієнтів регресії. Ці коефіцієнти описують невизначеність ознак регресії. Отримані результати показують, що різні ознаки мають різний вплив та різну невизначеність щодо цільової змінної. Найефективнішою та найменш варіативною серед розглянутих ознак була ознака на основі кількості відвідувань сторінки Вікіпедії про вакцину.

Моделювання динаміки поширення COVID-19 та порівняльний аналіз впливу кризи, зумовленої COVID-19 та інших криз на фондовий ринок, наведено у Додатках.

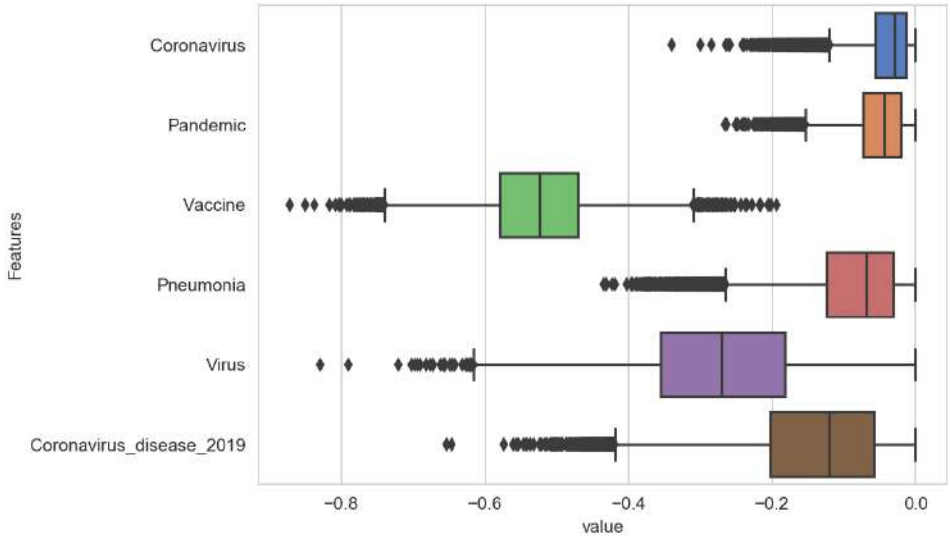

Рисунок 2.44 – Коробкові графіки для розподілу значень коефіцієнтів лінійної регресії

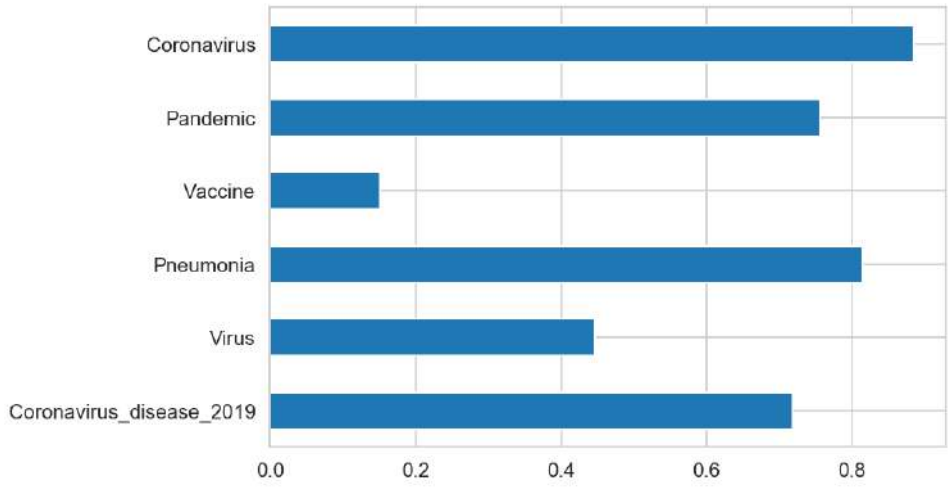

Рисунок 2.45 – Коефіцієнт варіації для аналізованих ознак

2.4 Методи машинного навчання, лінійної та байєсівської регресії у задачах виявлення технічних відмов

Розглянемо використання логістичної регресії на прикладі задачі виявлення відмов на виробничих лініях в компанії Bosch. У змаганні "Bosch Production Line Performance" [247] на платформі Kaggle компанія Bosch пропонує учасникам передбачити збої на лініях збірки, використовуючи численні ознаки, які моніторяться в процесі виробництва на лініях збірки. Особливістю запропонованої вибірки даних для цієї задачі є висока незбалансованість класів цільової змінної. Завдання прогнозування збоїв на лінії збірки можна розглядати як проблему логістичної регресії. Класифікація оцінюється за допомогою коефіцієнта кореляції Метюса (Matthews correlation coefficient, MCC) між прогнозованими та спостережуваними результатами. MCC визначається за формулою

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}, \tag{2.25}$$

де TP – кількість дійсних позитивних результатів, TN – кількість дійсних негативних результатів, FP – кількість помилкових позитивних результатів, а FN – кількість помилкових негативних результатів. Дані представляють реультати вимірювань під час переміщення деталей по виробничих лініях компанії Bosch. Кожна деталь має унікальний номер. Мета полягає у тому, щоб передбачити, які деталі механізмів не зможуть пройти контроль якості. В таких випадках значення бінарної цільової змінної дорівнює 1, в інших випадках – 0. Набір даних містить велику кількість анонімізованих ознак. Розглянемо розв'язок такої задачі, використовуючи методи машинного навчання, лінійної регресії та підхід на основі байєсівського висновування [248, 249] .

Для машинного навчання використовувався підхід на основі градієнтного бустінгу, реалізований у класифікаторі XGBoost. Для проведення аналізу було використано пакет R "xgboost" (скорочений термін від eXtreme Gradient Boosting) [73, 71, 72]. Було об'єднано фрейми даних із числовими та категоріальними ознаками. Також було

введено нову ознаку, що позначає час, протягом якого аналізована деталь знаходилася на виробничій лінії. Беручи до уваги великий об'єм даних, ми використовували підхід формування підвибірки для логістичної регресії з високо незбалансованими класами. Зразки даних з позитивною відповіддю 1 зберігались без змін, а кількість зразків із відповіддю 0 суттєво зменшувалась за допомогою випадкового відбору. Для категоріальних ознак було використано пряме кодування. До набору даних, отриманих таким чином, ми застосували класифікатор XGBoost. Загальна кількість ознак становила понад 4000. Ми визначили найважливіші ознаки, застосувавши класифікатор XGBoost до малої підмножини всього навчального набору. Для наступного кроку було взято 500 найважливіших ознак. Для валідаційної вибірки було взято 25% зразків даних. Результатами класифікації були ймовірності позитивних відповідей. На рис. 2.46 показано значимі ознаки. На рис. 2.47 показано криву ROC для результатів класифікації. Обчислена характеристика AUC становить 0.753. Для отримання бінарних значень цільової змінної потрібно застосувати деякий поріг для отриманих ймовірностей. Якщо ймовірність більша за поріг, то бінарне значення дорівнює 1, в інших випадках значення дорівнює 0. На рис. 2.48 показано коефіцієнт кореляції Метьюса для логістичної регресії для різних значень порогу. На рис. 2.49 показано коефіцієнт кореляції Метьюса для підмножини зразків, отриманих для сформованої підвибірки даних.

Учасники змагань на платформі *Kaggle.Com*, які працювали над цією задачею, також розглядали так звані 'магічні ознаки', які базувались на ідентифікації деталей [250, 251, 252]. Наявність 'магічних ознак' зумовлена характером формування набору даних. У загальному випадку такі ознаки можуть бути відсутні. Ці ознаки суттєво покращили показник логістичної регресії. Ми також проаналізували 'магічні ознаки', представлені у публікації форуму змагань [252], обчислили криву ROC (рис. 2.50) та коефіцієнт кореляції Метьюса (рис. 2.51) для різних наборів ознак. Набір ознак 2 – це набір ознак 1 з доданими 'магічними ознаками'. Для набору ознак 2 отримано коефіцієнт AUC = 0,91.

Логістичну регресію можна розглядати як окремий випадок узагальненої лінійної регресії. Такий тип логістичної регресії

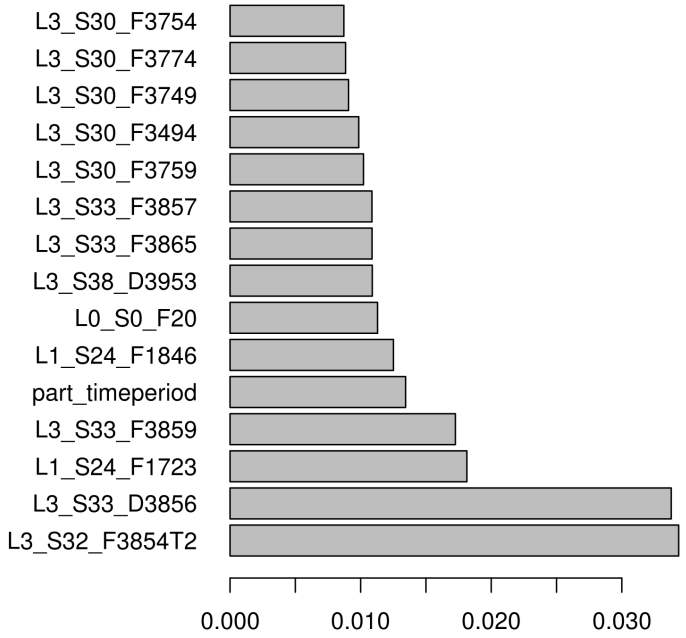

Рисунок 2.46 – Найбільш важливі ознаки

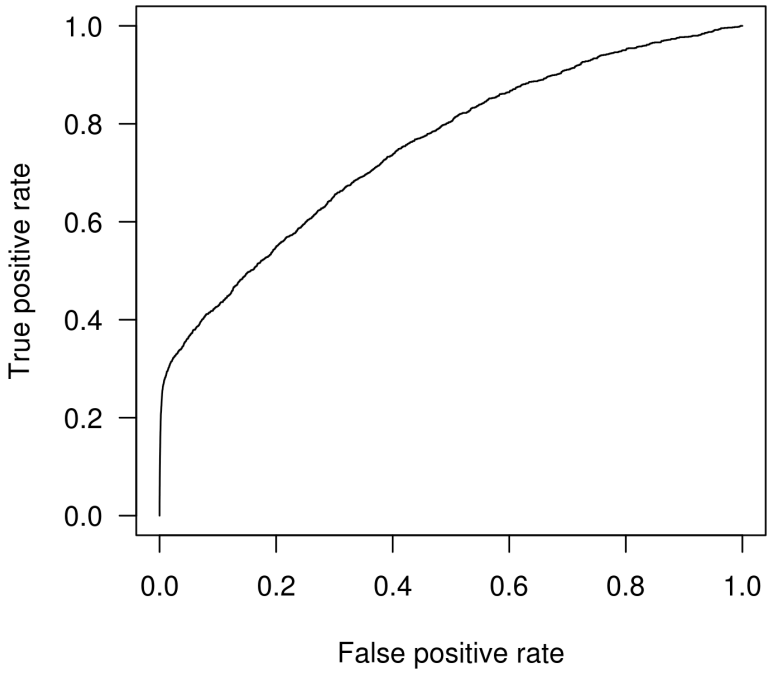

Рисунок 2.47 – ROC крива для тестової бінарної класифікації

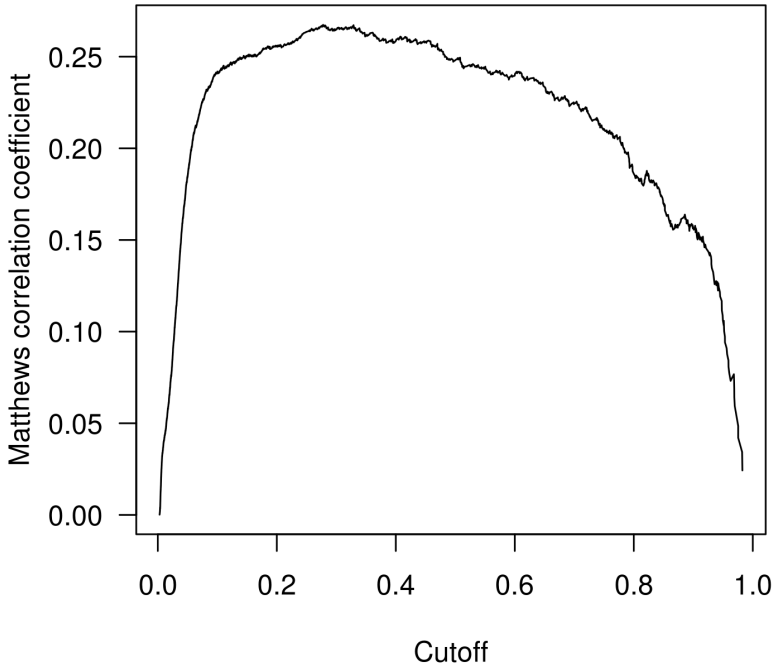

Рисунок 2.48 – Кореляційний коефіцієнт Метьюса для логістичної регресії при різних порогових значеннях класифікаційної ймовірності

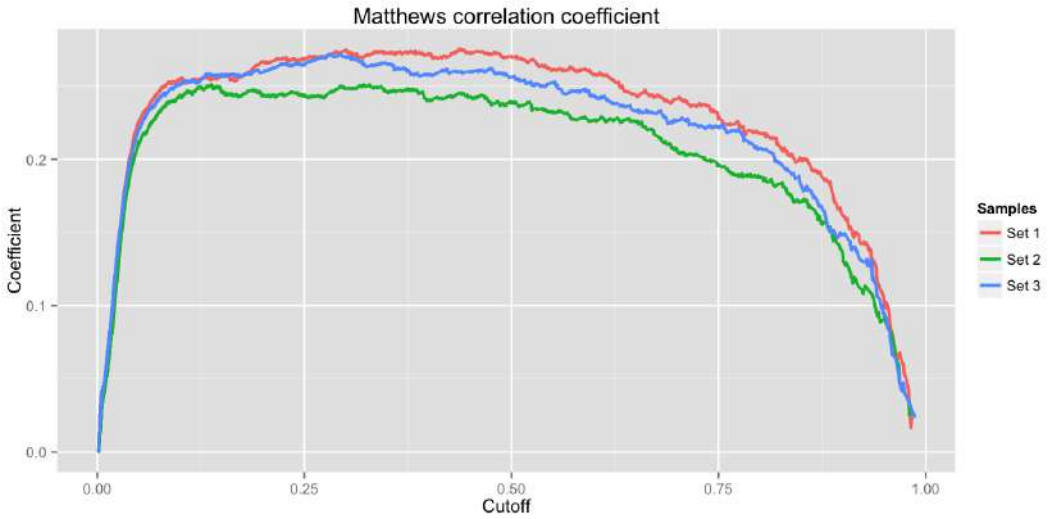

Рисунок 2.49 – Кореляційний коефіцієнт Метьюса для логістичної регресії для різних наборів даних

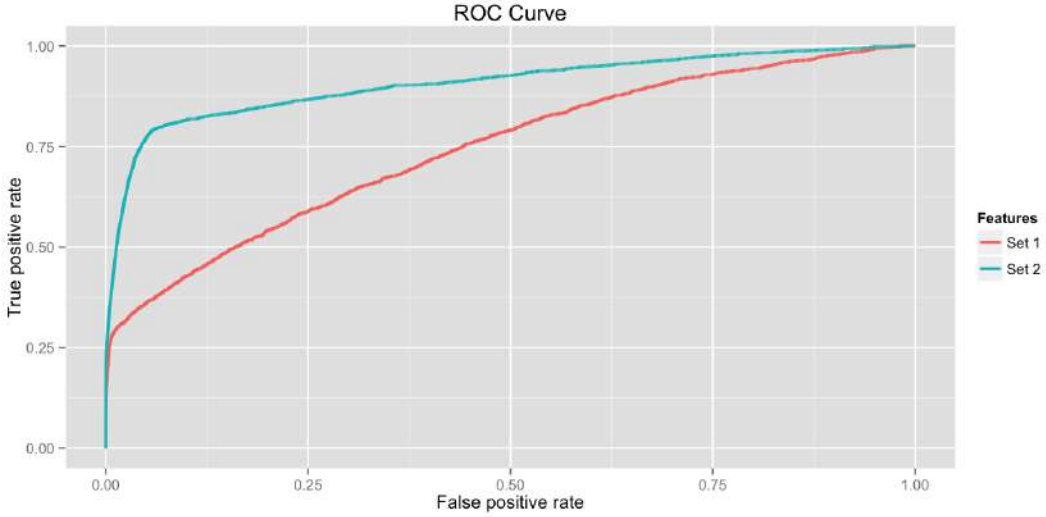

Рисунок 2.50 – ROC крива для різних наборів ознак

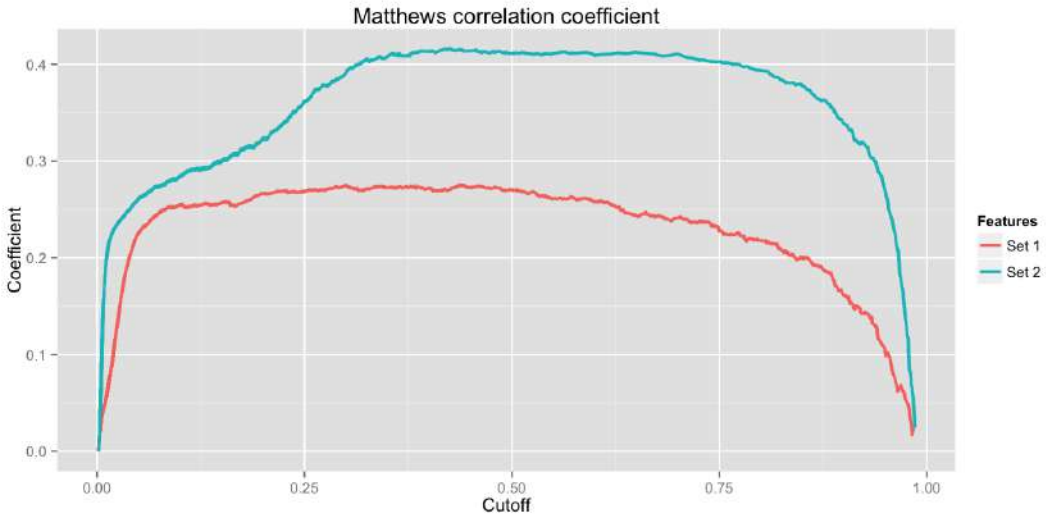

Рисунок 2.51 – Кореляційний коефіцієнт Метьюса для логістичної регресії для різних наборів ознак

дозволяє дослідити вплив числових факторів на бінарну цільову змінну. Використовуючи цю модель, можна отримати коефіцієнти для досліджуваних ознак. Набір даних, що розглядається, має велику кількість недоступних значень вимірювань (NA) для кожного зразка. Це не є великою проблемою для сучасних алгоритмів машинного навчання, побудованих на деревах рішень, але у випадку параметричної логістичної регресії, потрібно попередньо обробити дані, щоб уникнути значень NA у вибірці даних. Причина такої великої кількості значень NA може бути в тому, що в наборі даних є деталі різних типів. Для одного типу деталей застосовується один комплекс заходів для контролю якості, для іншого типу деталей застосовується інший набір заходів. Враховуючи те, що всі ознаки вимірювань у вибірці даних представлені у вигляді стовпців, показники, які не застосовуються до певної частини, матимуть значення NA. Отже, спочатку в лінійному аналізі потрібно згрупувати деталі за їх типами. Групування може виконуватися за допомогою кластеризації. Оскільки ознаки у вибірці даних є анонімізовані, було використано лише числові ознаки для вивчення можливості застосування узагальненої лінійної логістичної регресії для такої вибірки даних. Для кластеризації було використано алгоритм k-means. Для цільової змінної було встановлено значення 0 для недоступних значень ознак у вибірках даних та значення 1 для ознак із числовими значеннями. Для обчислення було вибрано 850 випадкових ознак. На рис. 2.50 показано залежність загальної суми квадратів у межах кластерів від кількості кластерів. Отримані результати показують, що оптимальна кількість груп деталей становить приблизно 20-30. Було вибрано 25 кластерів. Для лінійної регресії було обрано деталі лише з одного довільного кластера. Стовпці, де було більше 10% значень NA і рядки з більш, ніж 10% значення NA було видалено. Решта значень NA було замінено на середні значення, обчислені для відповідних числових ознак. Для лінійної логістичної регресії було використано пакет *glmnet* для мови статистичного програмування R[253, 254, 255, 256]. Розглянемо підхід на основі логістичної регресії з регуляризацією LASSO. Позначимо ймовірність того, що значення цільової змінної дорівнює 1 як p_1, тоді ймовірність того, що значення цільової змінної дорівнює 0, буде $p_0 = 1 - p_1$.

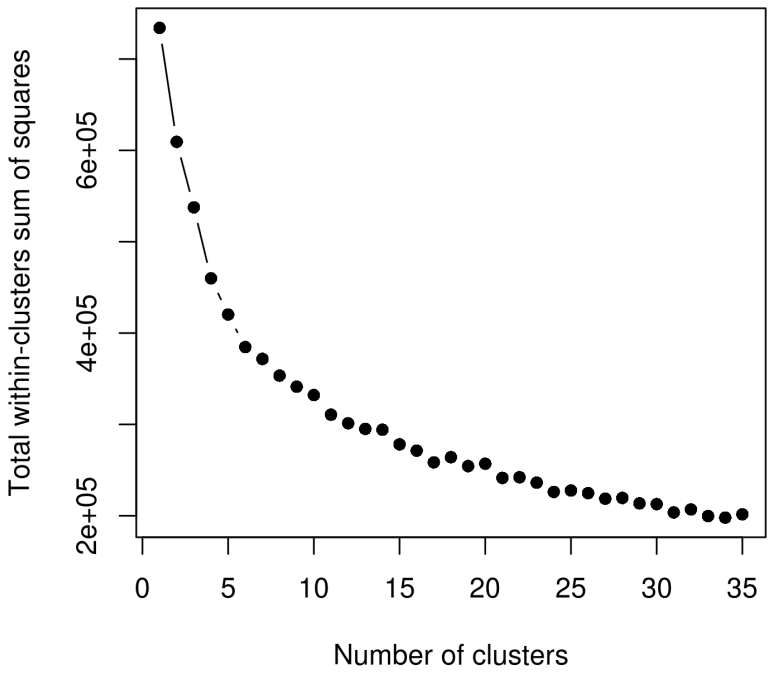

Рисунок 2.52 – Загальна внутрішньокластерна сума квадратів відстаней в залежності від кількості кластерів

Для логістичної регресії можна записати:

$$\begin{aligned} log\left(\frac{p_1}{p_0}\right) &= \beta_0 + \beta^T x, \\ p_1 &= \frac{exp(\beta_0 + \beta^T x)}{1 + exp(\beta_0 + \beta^T x)}. \end{aligned} \tag{2.26}$$

Коефіцієнти β_0, β^T можна знайти за допомогою мінімізації відповідної цільової функції. Щоб знайти оптимальне значення для параметра регуляризації *Lambda*, було використано підхід на основі кросвалідації. На рис. 2.53 показано залежність значення AUC від логарифмічного значення *Lambda*. Верхня вісь на рисунку вказує на кількість ненульових коефіцієнтів для *Lambda*. Маючи цю залежність, ми можемо знайти оптимальне значення параметра регуляризації, за яким ми можемо отримати найкращу точність. Використовуючи $Lambda = 0,03$, обчислено ненульові коефіцієнти узагальненої лінійної моделі для логістичної регресії, які показано на рис. 2.54. На рис. 2.55 показано гістограми, коефіцієнти кореляції, парні діаграми розкиду для ознак. Для байєсівської моделі ми взяли ознаки, які було знайдено в узагальненій лінійній моделі за допомогою регуляризації LASSO. Аналіз проводився за допомогою програмного забезпечення для ймовірнісного програмування *JAGS*, використовуючи пакет *rjags* для мови статистичного програмування R [59, 257]. Для моделювання було використано логістичну регресію. Значення цільової змінної 0 і 1 розподіляються за законом Бернуллі [59]. Логістичну регресію на основі байєсівського висновування можна описати так:

$$\begin{aligned} &p = Logistic(b_0 + b_1 x_1 + b_2 x_2 + ... + b_n x_n), \\ &y \backsim Bernully(p), \end{aligned} \tag{2.27}$$

де

$$Logistic(x) = \frac{1}{1 + exp(-x))}. \tag{2.28}$$

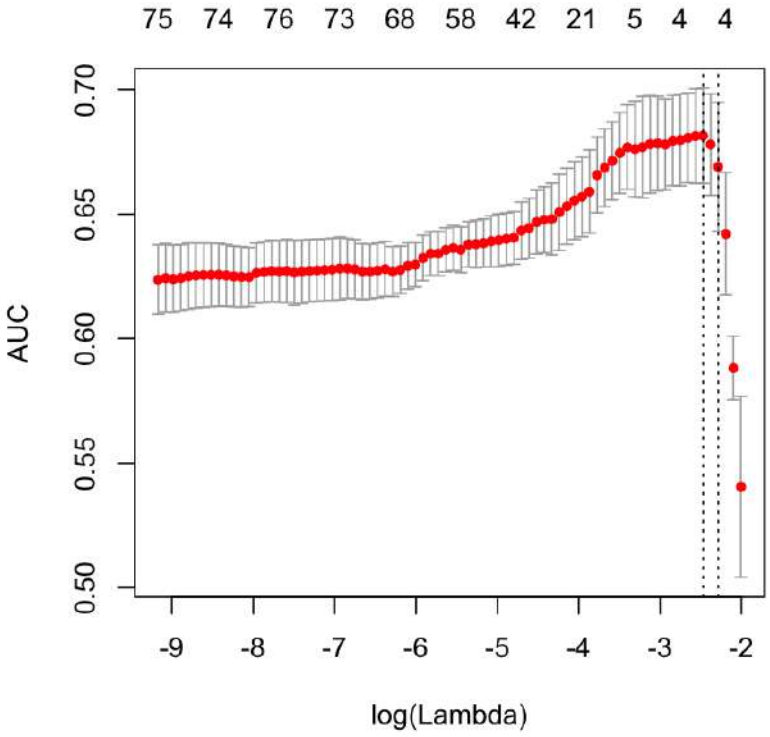

Рисунок 2.53 – Залежність AUC від λ

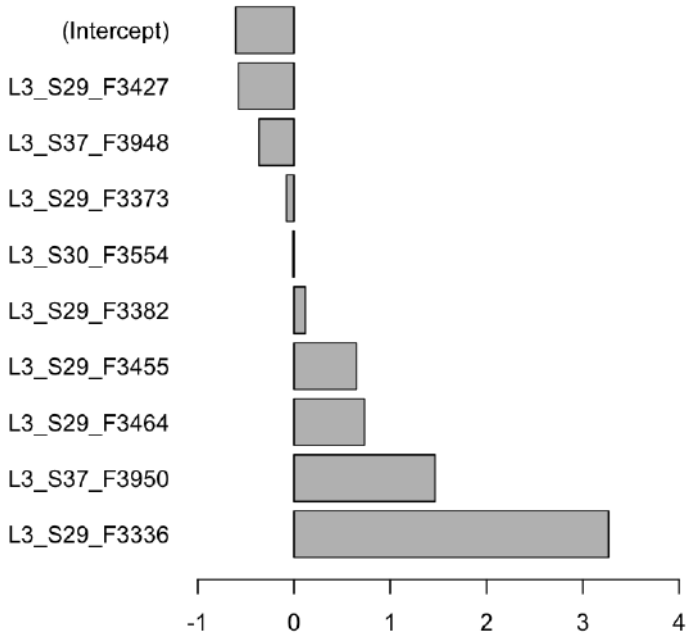

Рисунок 2.54 – Коефіцієнти узагальненої лінійної моделі для логістичної регресії

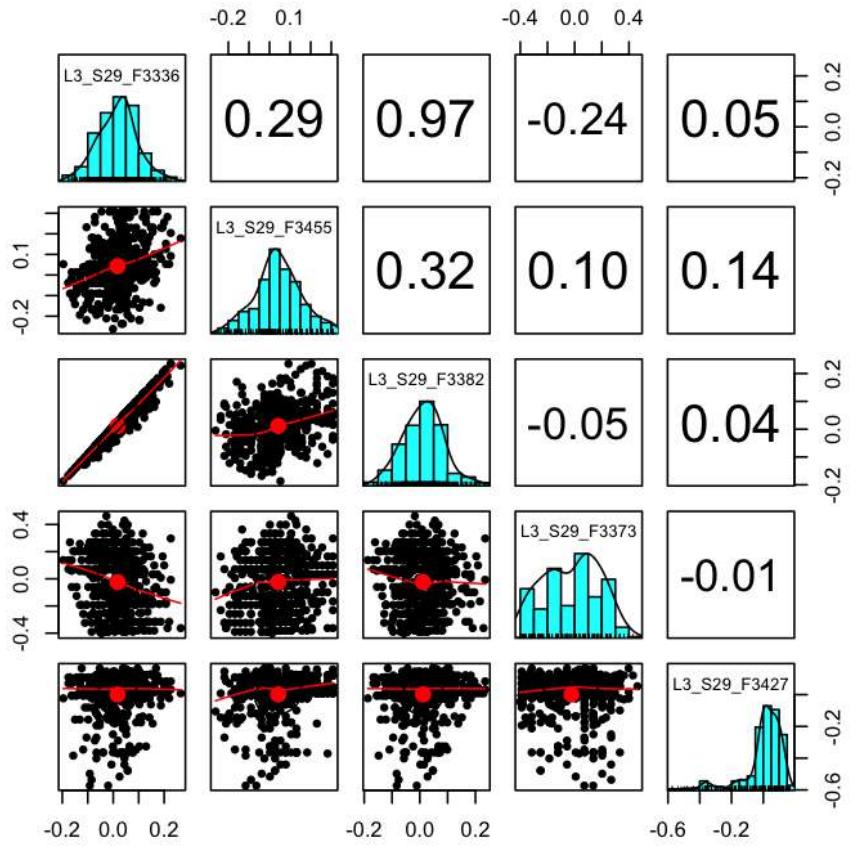

Рисунок 2.55 – Гістограми, коефіцієнти кореляції та парні діаграми розкиду значень ознак

Проста імовірнісна модель для логістичної регресії з використанням синтаксису *BUGS* має такий код:

```
model{
  for (i in 1:n) {
    y[i] ~ dbern(p[i])
    logit(p[i]) <- b0+inprod(b[ ],x[i,])
  }
  b0 ~ dnorm(0,0.0001)
  for (j in 1:nfeat) {
    b[j] ~ dnorm(0,0.0001)
  }
}
```

Візуалізація трасування змінних моделі є корисною при оцінці конвергенції процесу семплювання змінних моделі. На рис. 2.56 показано візуалізацію трасування для параметра b_0 логістичної моделі. Отриманий графік трасування показує стаціонарний процес. На рис. 2.57 показано щільність розподілу ймовірностей для параметра b_0. На рис. 2.58 показано приклади коробкових діаграм для деяких коефіцієнтів логістичної регресії. Точки на рисунку позначають значення коефіцієнтів, обчислених за допомогою узагальненої лінійної моделі без регуляризації ($Lambda = 0$), яку було застосовано до тих же ознак, що і у випадку байєсівської моделі. Отримані результати показують, що використання байєсівського підходу дозволяє моделювати стохастичні залежності між різними факторами та отримувати розподіли параметрів аналізованої моделі. Такий підхід може бути корисним для оцінки різних ризиків, пов'язаних із проблемами контролю якості.

Розглянемо використання комбінування прогнозних моделей за допомогою стекінгу. На рис. 2.59 показано схему багаторівневого ансамблю моделей із використанням стекінгу. На першому рівні ансамблю є моделі XGBoost з різними наборами ознак і підвибірками зразків даних, на другому рівні прогнозовані за допомогою моделей першого рівня ймовірності використовуються як вхідні ознаки для лінійної та байєсівської регресії. Для чисельного моделювання використовувались різні набори параметрів

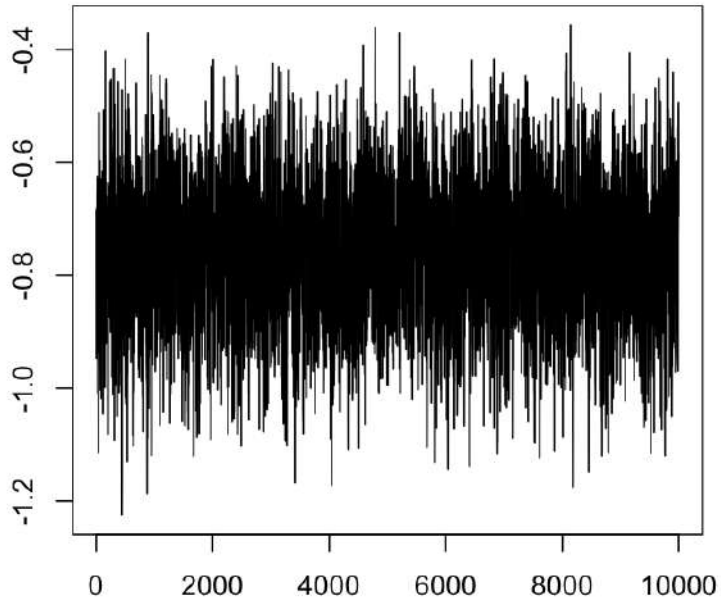

Рисунок 2.56 – Трасування для коефіцієнта вільного члена узагальненої лінійної моделі логістичної регресії

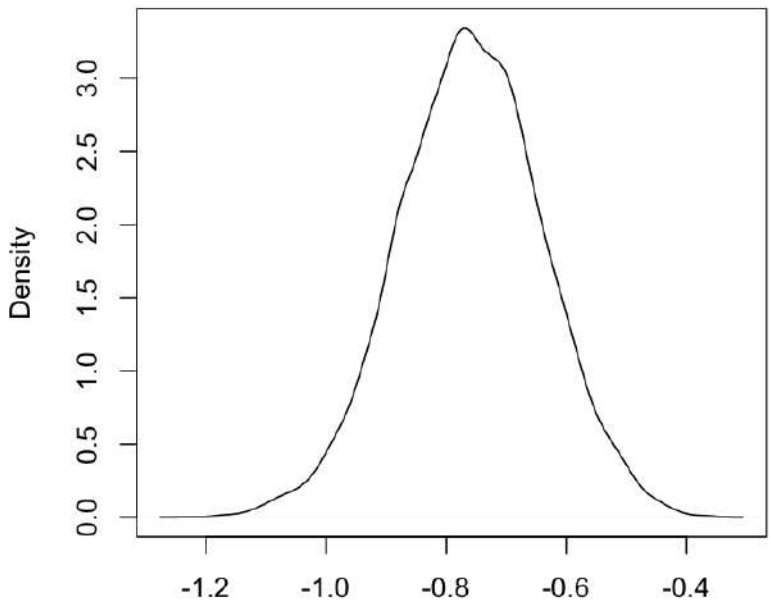

Рисунок 2.57 – Щільність розподілу ймовірностей для коефіцієнта вільного члена узагальненої лінійної моделі логістичної регресії

для 3 моделей XGBoost: набір 1: max.depth=15, colsample_bytree=0.7; набір 2: max.depth=5, colsample_bytree=0.7; набір 3: max.depth=15, colsample_bytree=0.3. Для цих трьох моделей було використано однакову підмножину зразків даних. На рис. 2.60, 2.61 показано залежність коефіцієнта корекції Метьюса від порогу ймовірності для різних підмножин ознак, де набір ознак 2 – це набір 1 з чотирма доданими 'магічними ознаками', згаданими вище. Для байєсівських моделей було використано 3 однакові підмножини параметрів з різними підмножинами зразків даних. Як було показано вище, для різних підмножин зразків ми отримали дещо різні результати для коефіцієнта корекції Метьюса (рис. 2.51). Як незалежні змінні для байєсівської моделі було використано прогнозовані ймовірності, отримані з використанням моделей XGBoost та трьох різних підмножин зразків даних. На рис. 2.62 показано коробкові графіки для розподілів коефіцієнтів для різних XGBoost моделей.

Отже, досліджено різні підходи в логістичній регресії у проблемі

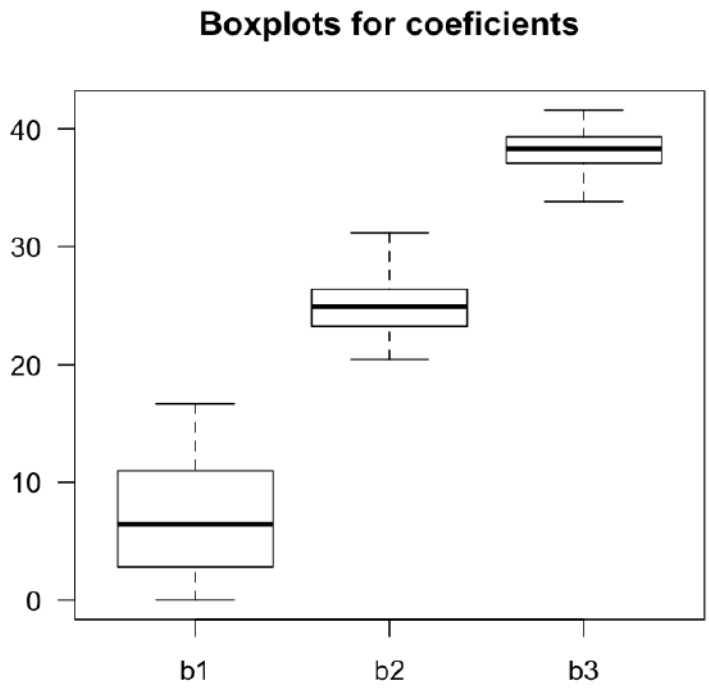

Рисунок 2.58 – Коробкові графіки для розподілів значень коефіцієнтів логістичної регресії

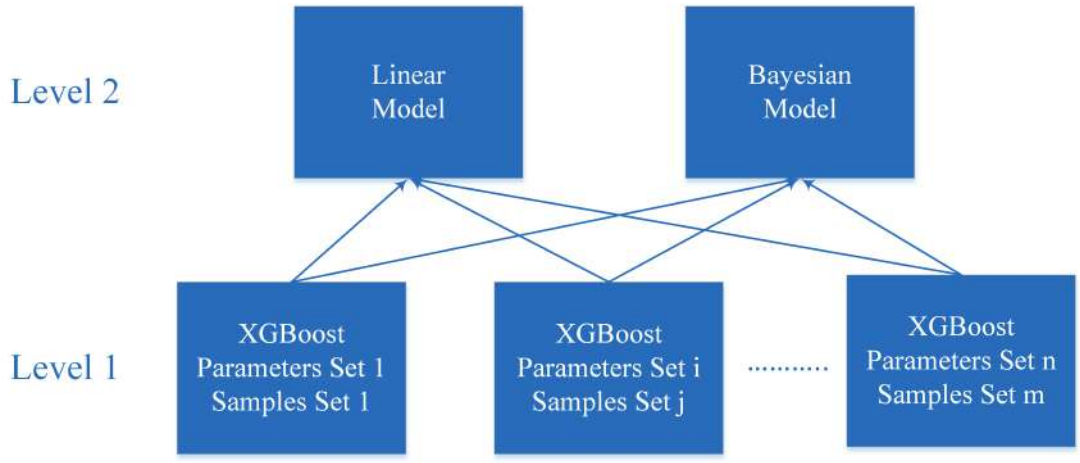

Рисунок 2.59 – Логістична регресія з моделями машинного навчання на першому рівні та лінійної і байєсівською регресією на другому рівні

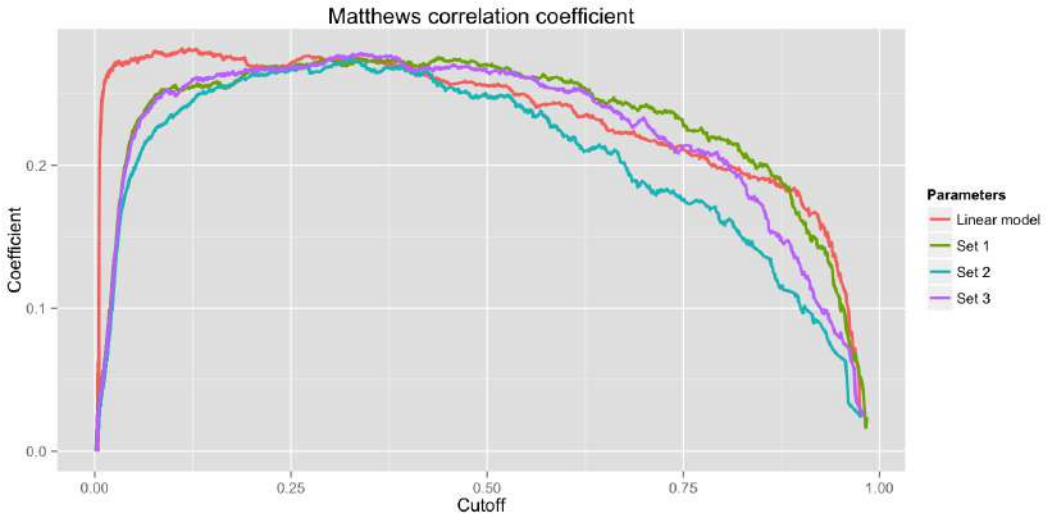

Рисунок 2.60 – Кореляційний коефіцієнт Метьюса для різних наборів параметрів XGBoost моделі (набір ознак 1) для різних наборів ознак

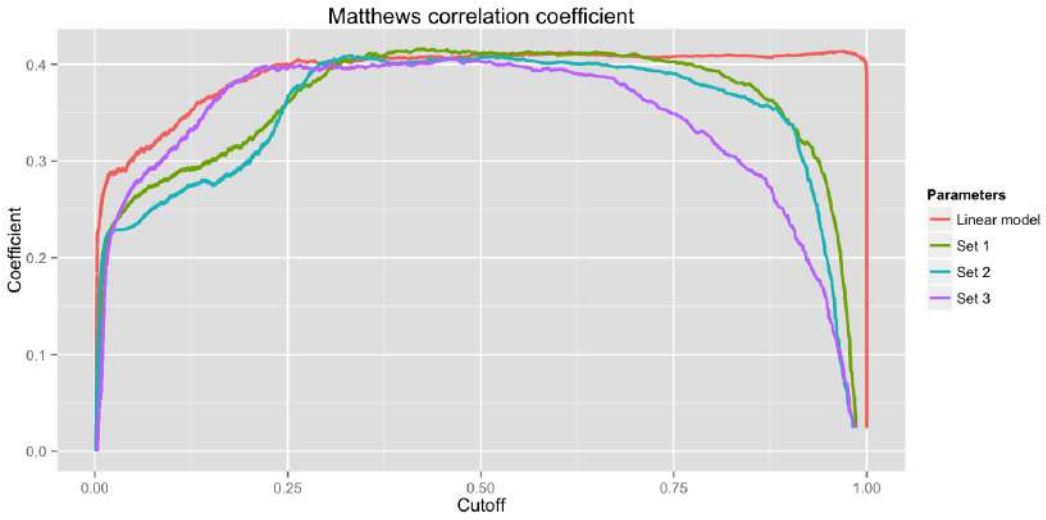

Рисунок 2.61 – Кореляційний коефіцієнт Метьюса для різних наборів параметрів XGBoost моделі (набір ознак 2) для різних наборів ознак

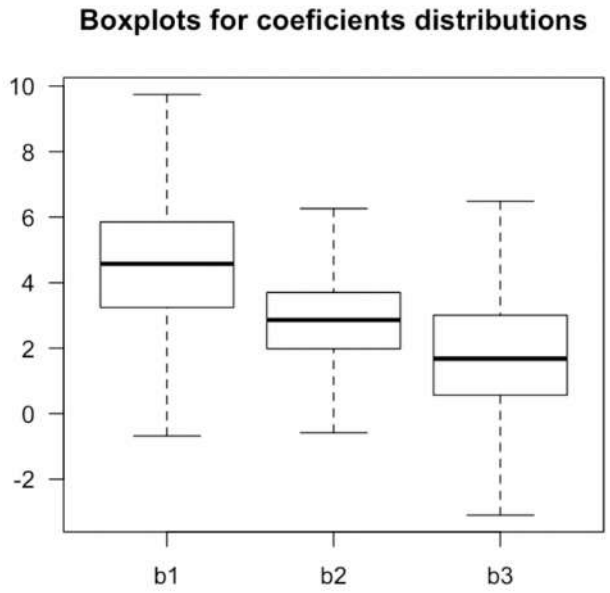

Рисунок 2.62 – Розподіли для коефіцієнтів різних XGBoost моделей

виявлення відмов на виробничих лініях [248, 249]. Машинне навчання дає можливість виявити складні патерни виникнення відмов у виробничих процесах. Узагальнена лінійна модель для логістичної регресії дає можливість досліджувати фактори впливу на виникненя відмов. Використовуючи байєсівську модель, можна отримати статистичні розподіли для параметрів моделі, які можна використовувати в аналізі ризиків. За рахунок розроблених методів у прогнозуванні технічних відмов на лініях збірки на виробництві з використанням стекінгового об'єднання моделей можна оптимізувати набір прогнозних ознак та підвищити точність прогнозування на 1-10%. Використання байєсівської моделі на другому рівні модельного ансамблю з незалежними змінними, які відображають ймовірності відмов, отриманих на основі прогнозних моделей першого рівня, можна проаналізувати невизначеності моделей першого рівня та оцінити ризики, які виникають при похибках прогнозування ансамблю прогнозних моделей.

2.5 Методи інтелектуального аналізу даних з використанням глибинного Q-навчання

Розглянемо два випадки використання Q-навчання в аналітиці часових рядів продажів [258]. Одним випадком є аналіз оптимальної цінової стратегії, другим - аналіз задачі попиту та постачання, яка часто виникає у сфері рітейлу. У чисельних експериментах використано алгоритми, які ґрунтувалися на модельній архітектурі, описаній у [106, 107], та підходи до реалізації агента Q-навчання з [259, 260, 261]. Для моделювання середовища взаємодії у випадку задачі оптимізації цін ми використовували параметричні моделі. У випадку оптимізації дій агента у задачі попиту та постачання використано історичні дані для часових рядів попиту. Для аналізу проблеми попиту та постачання ми використали історичні дані продажів магазинів зі змагання Kaggle 'Rossmann Store Sales' [214]. Ці дані описують продажі в магазинах Rossmann. Розрахунки проводилися в середовищі Python із використанням таких основних пакетів *pandas* [215, 216], *sklearn* [217], *numpy* [218], *keras* [219], *matplotlib* [220], *seaborn* [221]. Для проведення аналізу було використано середовище *Jupyter Notebook*.

2.5.1 Q-навчання інтелектуального агента з використанням моделі середовища

Розглянемо простий випадок стратегії ціноутворення. Питання полягає у тому, яку стратегію можна застосувати для максимізації прибутку за певний проміжок часу. Стратегія може складатися з багатьох факторів та засобів. У найпростішій моделі вона може складатися з дискретних додаткових значень ціни, які додаються до собівартості. Залежність між величиною продажу та додатковою ціною можна вважати, наприклад, такою:

$$F_{Sales} = \frac{a}{(1 + b \cdot exp(c \cdot (Price_m \cdot (1 + Price_e) - d)))}, \tag{2.29}$$

де $Price_m$ – мінімальна ціна продажу, яка рівна собівартості продукту, $Price_e$ – додаткова ціна у відносних частинах ціни $Price_m$, a, b, c, d – параметри для функції F_{Sales}. Функція F_{Sales} (2.29) описує відносне зменшення продажів

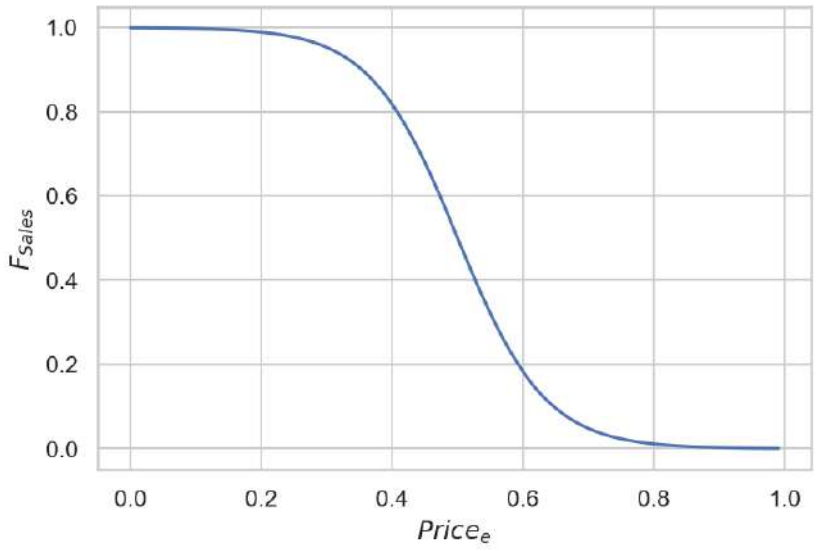

Рисунок 2.63 – Залежність продажів від додаткової ціни

із збільшенням додаткової ціни. Функція винагороди Q-навчання може розглядатися як

$$Reward = Demand \cdot F_{Sales} \cdot Price_e. \tag{2.30}$$

На рис. 2.63 показано залежність F_{Sales} від доданої ціни ($Price_m$=1, a=1, b=1, c=15, d=1.5). Можна спостерігати, що високій додатковій ціні відповідають невеликі продажі. Реальні параметри логістичної кривої можна знайти градієнтним методом на основі історичних даних. На рис. 2.64 показано залежність прибутку від ціни. Можна побачити, що прибуток є мінімальним при великих і малих значеннях доданої ціни, яка додається до собівартості товару. Завдання полягає в пошуку оптимальної додаткової ціни. Можна скласти таблицю цільової функції за допомогою змодельованих часових рядів попиту та знайти оптимальне значення для додаткової ціни. У реальних випадках залежність прибутку від додаткових цін може мати складну функціональну залежність, включно з залежністю від багатьох якісних факторів, які визначаються багаторівневою ціновою стратегією. Для формування оптимальної стратегії ціноутворення використаємо підхід на основі Q-навчання. У випадку практичного використання цю просту модель можна використовувати для холодного старту навчання агента перед взаємодією з реальним бізнес-середовищем.

Розглянемо параметри чисельного моделювання. Для обрахунку

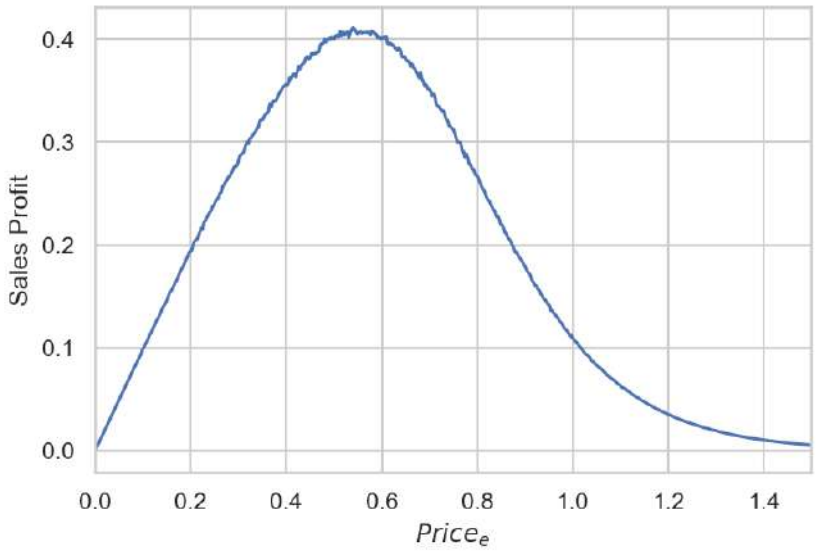

Рисунок 2.64 – Залежність прибутку від додаткової ціни

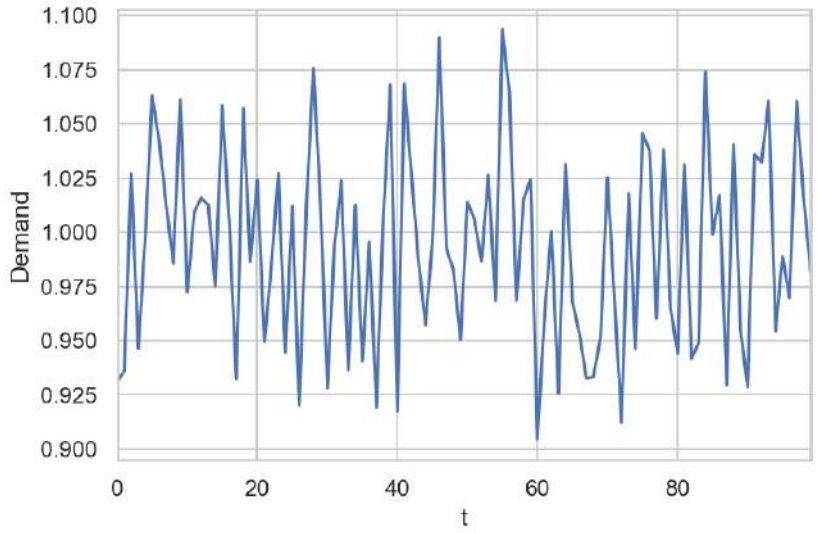

Рисунок 2.65 – Часовий ряд змодельованого попиту

Q-значень використано нейронну мережу прямого поширення з двома прихованими шарами із 32 нейронами в кожному шарі. Розмір вихідного шару нейронної мережі дорівнює кількості можливих дій агента і був вибраний рівним 8. Для значень, які описують дії агента, ми взяли список наступних значень для нормалізованої додаткової ціни [0, 0.15, 0.25, 0.5, 0.75, 0.85, 1, 1.5]. Додаткова ціна розглядається у відносних частинах ціни собівартості. Як часові кроки ми розглядаємо дні. Кількість етапів часу в кожному епізоді дорівнює 7, розмір пакету даних для тренування моделі – 32, кількість ітерацій навчання – 50, оптимізатор – Adam, параметр темпу навчання нейронної мережі – 0.001, коефіцієнт затухання (epsilon decay) – 0.97. Для апроксимації функції (2.29) використано такі параметри: $Price_m$=1, a=1, b=1, c=7, d=1.7 . У цьому прикладі використано підхід DQN з ε-жадібним алгоритмом розвідки-використання (epsilon-greedy exploration-exploitation trade off). Величина ε описує ймовірність випадкової дії. З кожною ітерацією ε зменшується. На рис. 2.66 показано залежність ε від часу. Для числового експерименту було змодельовано попит з випадковим рівномірним розподілом відносних одиниць. У результаті чисельного експерименту ми отримали модель DQN, яка може визначати оптимальні дії агента, що дозволяють максимально збільшити сукупну винагороду. На рис. 2.65 показано змодельований часовий ряд попиту. На рис. 2.67 показана середня винагорода на епізодах взаємодії агента із середовищем. Можна побачити, що винагорода збільшується із ітераціями. Це означає, що навчальний агент покращує спосіб взаємодії з навколишнім середовищем. На рис. 2.68 показані частоти для дій агента. На цьому відображена одна домінуюча дія, яка відповідає оптимальній ціновій стратегії для цієї простої моделі. На рис. 2.69 показано дії агента в часі, для яких є характерним великий розкид на початку процесу взаємодії внаслідок переважання розвідувального типу взаємодій. На наступних кроках після періоду розвідки переважає одна дія, яка може розглядатися як знайдена оптимальна стратегія ціноутворення [258].

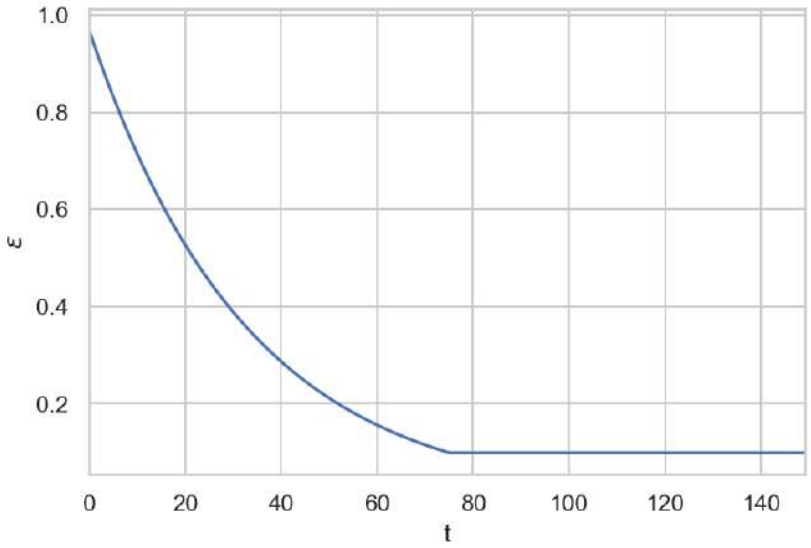

Рисунок 2.66 – Часова залежність ε

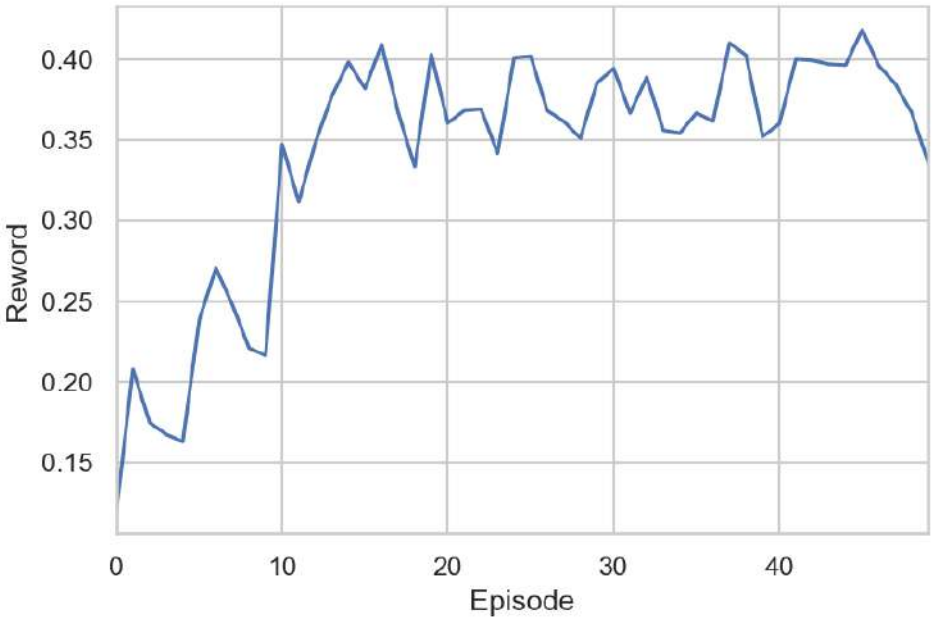

Рисунок 2.67 – Середня винагорода на епізодах взаємодії агента з середовищем

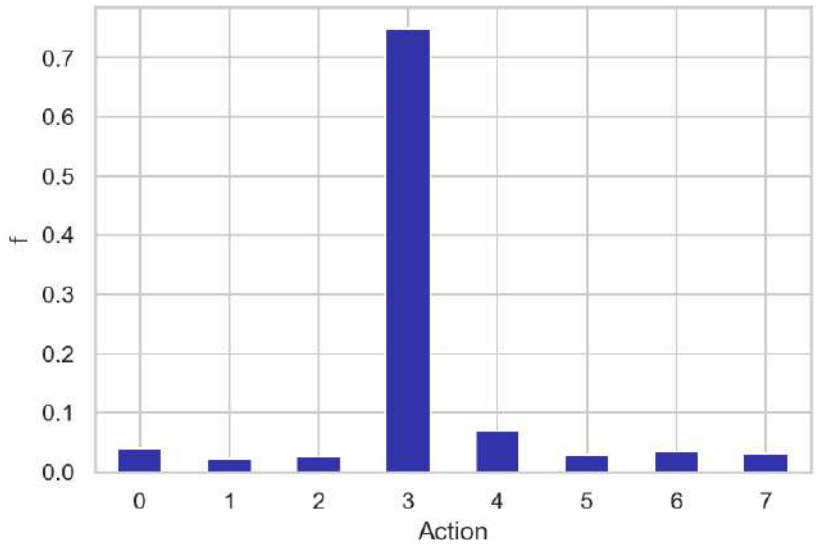

Рисунок 2.68 – Частоти дій агента

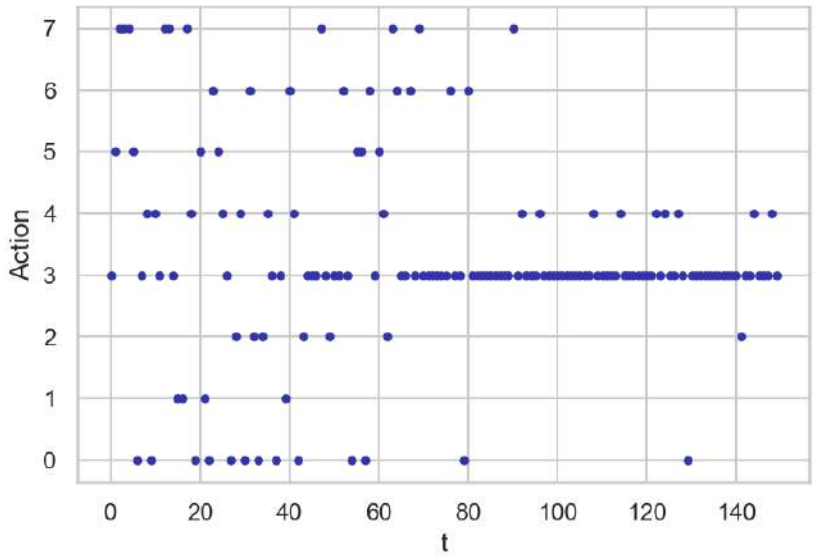

Рисунок 2.69 – Дії агента в часі

2.5.2 Q-навчання інтелектуального агента з використанням історичних даних

Розглянемо приклад використання Q-навчання для проблеми попиту та постачання в задачах оптимізації продажів. У цьому прикладі можна використовувати історичні дані для початкового запуску алгоритму Q-навчання разом з параметричним моделюванням середовища. Такий підхід дозволяє проводити холодний старт для агентів Q-навчання. У цьому випадку було взято нормовані часові ряди попиту з урахуванням сезонності та промо фактору. Завдання полягає у пошуку оптимальних дискретних дій агента у проблемі попиту та постачання. Продукти можна постачати партіями з дискретною кількістю. У моделі ми враховуємо витрати на переробку продукції, які пов'язані з логістичними, складськими та іншими витратами. Нагороду на кожному кроці можна вважати такою

$$Reward = SalesProfit - ProcessCost. \tag{2.31}$$

У цій задачі є більше можливостей для представлення ознак станів агента у порівнянні з задачею оптимального ціноутворення, яку ми розглянули раніше. Такими ознаками можуть бути попит на продукт напередодні, очікувані промо-акції, день тижня. Розглянемо параметри моделювання задачі попит-постачання. Для моделювання середовища використано історичний часовий ряд для попиту. Цей часовий ряд було змодельовано на основі часових рядів продажів які було взято на змаганні з аналізу даних 'Rossmann Store Sales' на платформі Kaggle [214]. Для зменшення ефекту перенавчання було застосовано випадковий лаг для часового ряду попиту, який випадково змінювався на кожному епізоді навчання. Лаг обчислювався за допомогою рівномірного випадкового розподілу цілих значень у проміжку від 0 до 25. Для кожного епізоду використано часовий проміжок 150 днів, кількість дій агента - 7. Ми використовували нейронну мережу прямого поширення з двома прихованими шарами з 64 нейронами у кожному шарі. Розмір пакету даних для ітерацій навчання - 32, параметр темпу навчання - 0.001, ε - затухання - 0.995, гамма-коефіцієнт для рівняння Беллмана - 0.3. Для підрахунку винагороди використано такі параметри: значення

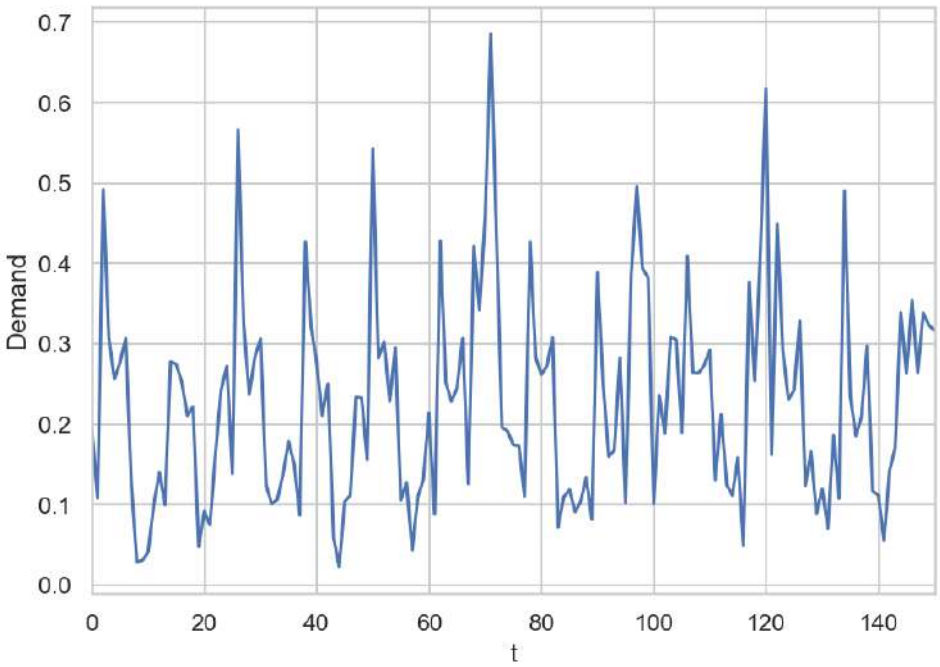

Рисунок 2.70 – Часовий ряд попиту

ціни прибутку на одиницю товару 1, роздрібна одиниця товару - 0.05, ціна обробки та логістичної підтримки товару - 0.5. Як ознаки стану агента використано бінарну ознаку промо акції, величини продажів за попередній день, внутрішньотижневу сезонність, яка була представлена у вигляді 7-ми бінарних ознак, отриманих на основі one-hot кодування ознаки дня тижня. Також встановлено бінарну змінну зупинки епізоду, коли винагорода стає нижчою за вказаний процес. Застосування змінної зупинки епізоду прискорює процес Q-навчання, оскільки при низьких значеннях винагороди процес навчання на даному епізоді закінчується і починається нова ітерація на новому епізоді. Дії агента розглядаються як дискретні значення для постачання товару. Для чисельного моделювання було обрано 7 дій із наступними значеннями [0, 2, 4, 6, 8, 10, 12], які є дискретною кількістю упаковок товарів. Кількість продукту в упаковці становила 0.025 відносних одиниць. На рис. 2.70 показано змодельований часовий ряд попиту для довільно обраного епізоду. Для чисельного експерименту попит моделюється у деяких відносних одиницях. На рис. 2.71 показано розраховані середні значення винагороди на епізодах. Результати показують, що навчальний агент оптимізує послідовності дій протягом епізодів [258]. На рис. 2.72 можна побачити, що домінують декілька дій, на відміну від попередньої

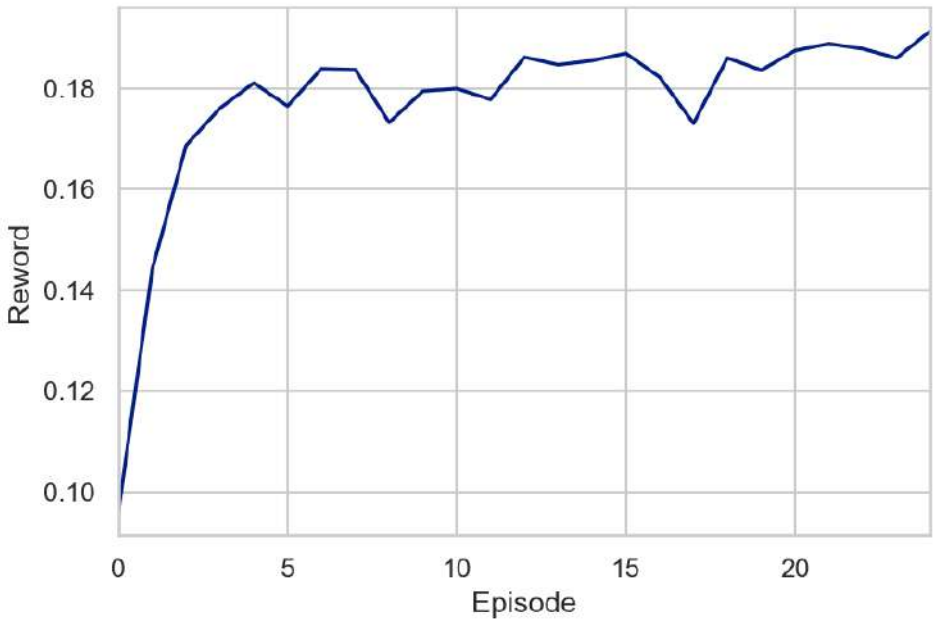

Рисунок 2.71 – Розраховані середні значення винагороди на епізодах

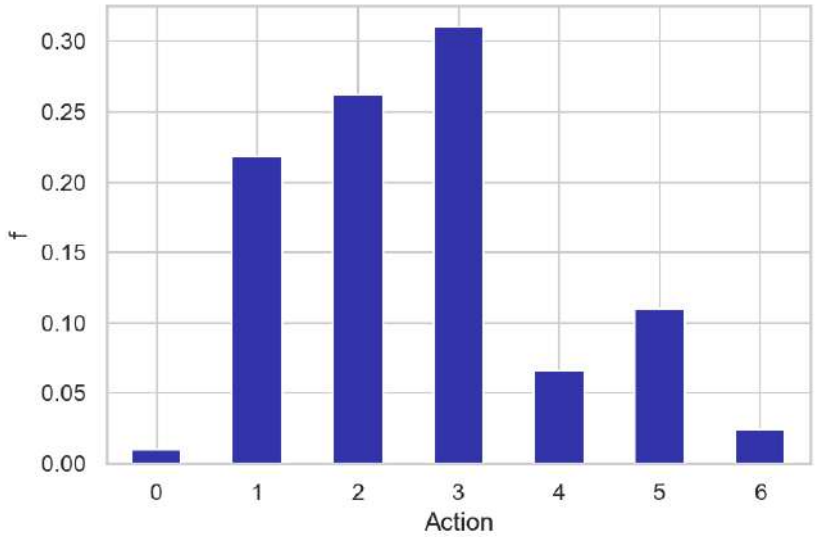

Рисунок 2.72 – Частоти дій у задачі попиту-постачання

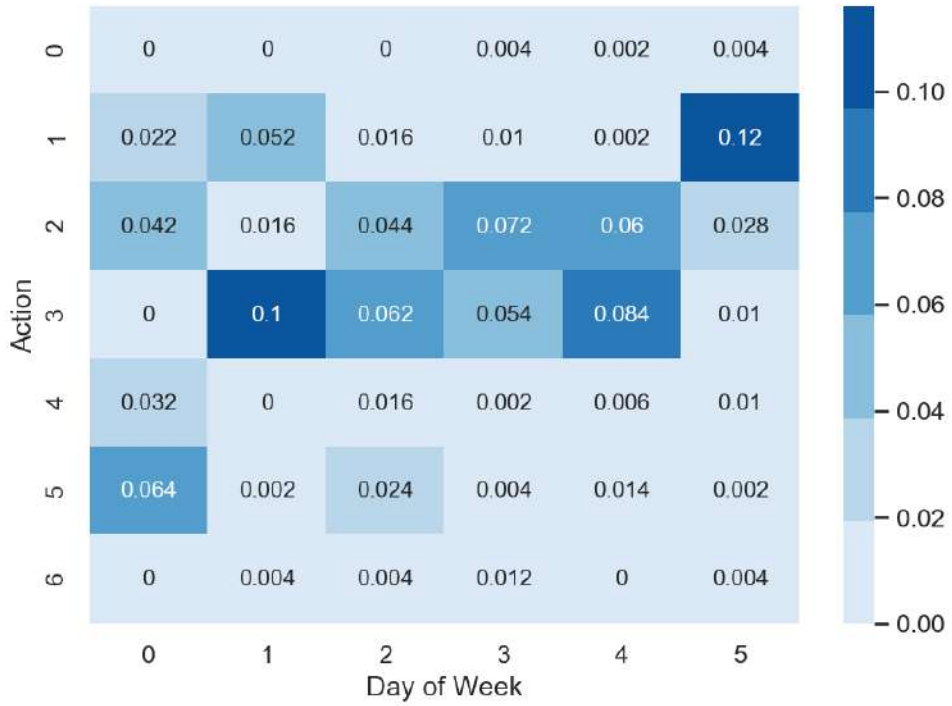

Рисунок 2.73 – Частоти дій агента для різних днів тижня

задачі дослідження оптимізації цінової стратегії де домінувала лише одна дія. У цьому випадку є більше ознак для представлення стану, тому алгоритм дає різні оптимальні дії для різних станів. На рис. 2.73 показано частоти дій агента для різних днів тижня. У різні дні переважають різні дії агента, залежно від внутрішньотижневої сезонності та промоакцій. На рис. 2.74 показано часовий ряд попиту, поставок, запасів та нестачі. Динаміка поставок, торговельних запасів та нестачі залежить від функції винагороди, яка може формуватися за рахунок прибутку, витрат на логістику, опрацювання продукції та недоотриманого прибутку.

Отже, розглянуто використання моделей глибокого Q-навчання у задачах часових рядів продажів [258]. На відміну від машинного навчання з учителем, яке можна розглядати як форму пасивного навчання із використанням історичних даних, Q-навчання – це вид активного навчання з метою виявлення оптимальної послідовності взаємодій інтелектуального агента з середовищем з метою досягнення максимальної винагороди. Розглянуто безмодельний підхід Q-навчання для аналізу задачі оптимальних стратегій ціноутворення та задачі попиту та постачання. У задачі ціноутворення середовище було змодельовано на основі випадково

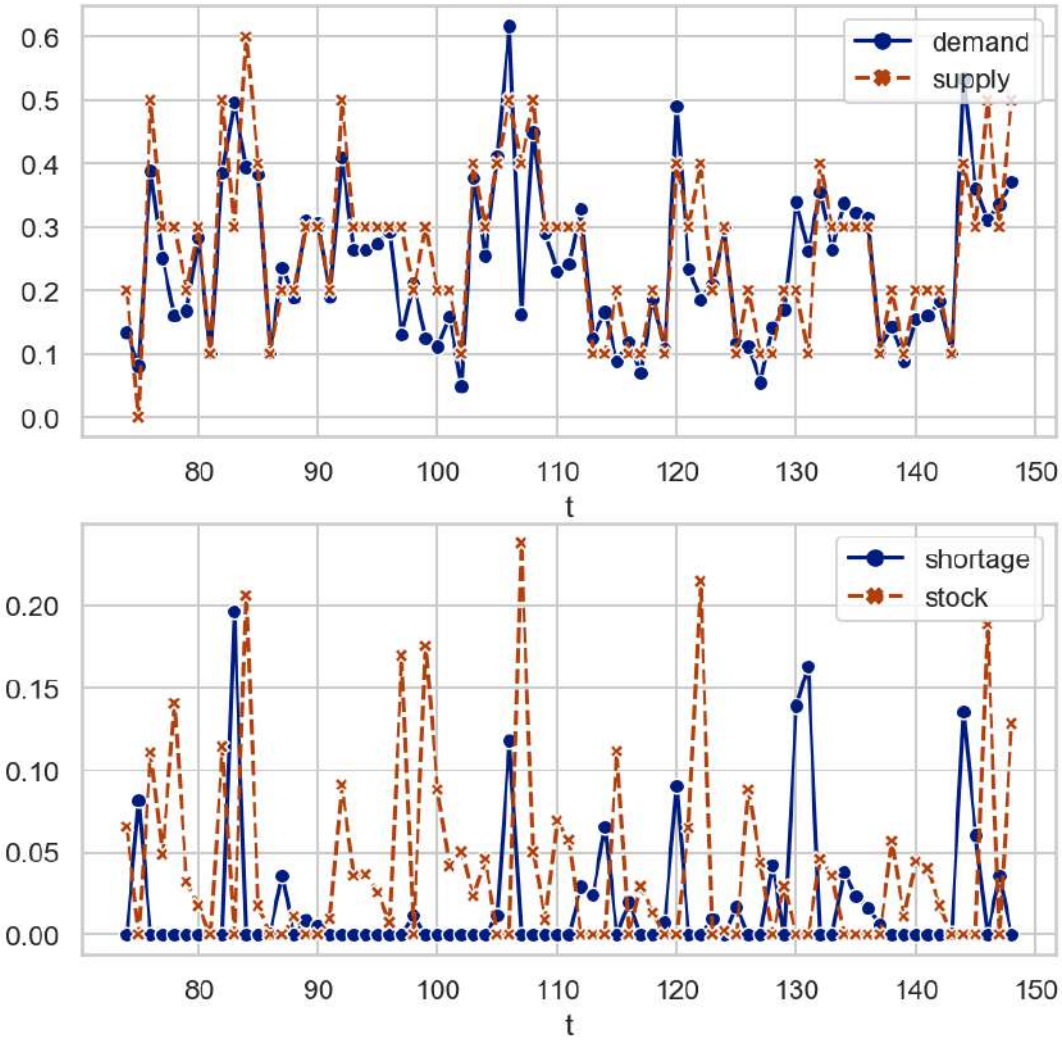

Рисунок 2.74 – Часові ряди для попиту, постачання та нестачі товарів

згенерованого попиту та залежності продажів від додаткової ціни. У задачі попиту та постачання запропоновано використовувати історичний часовий ряд попиту для моделювання середовища, ознаки станів агента було представлено промо-акціями, попередніми значеннями попиту та ознаками сезонних коливань. Отримані результати показують, що за допомогою глибокого Q-навчання можна формувати процес прийняття рішень для задач оптимізації цін та задач попиту та постачання. Моделювання середовища з використанням параметричних моделей та історичних даних може бути використано для холодного старту навчання інтелектуального агента. На наступних кроках, після холодного старту, агента можна використовувати в реальному бізнес-середовищі. Використовуючи Q-навчання, можна побудувати алгоритм прийняття оптимальних рішень. Цей алгоритм можна запустити з даними, змодельованими на основі параметричної експертної моделі або на основі моделі, яка базується на історичних даних, і далі цей алгоритм може працювати з реальним бізнес-середовищем та адаптуватися у процесі його використання до змін у бізнес-процесах.

2.6 Висновки

– Показано, що консолідовані дані з різною структурою та з різних джерел можуть бути представлені у вигляді реляційної моделі. За допомогою операцій реляційної алгебри можна виділити та утворити нові ознаки аналізованої задачі, які в подальшому будуть використані для інтелектуального аналізу.

– Розглянуто різні методи регресії на основі машинного навчання до прогнозування часових рядів. Використання регресійних методів для прогнозування часових рядів часто можуть дати кращі результати порівняно зі статистичними методами часових рядів. Одне з головних припущень методів регресії – це те, що патерни в історичних даних повторяться в майбутньому. Досліджено ефект генералізації моделі машинного навчання, який полягає у виявленні патернів у вибірці даних. Цей ефект можна використовувати для прогнозування, наприклад, продажів, коли є невелика кількість історичних даних для

заданих часових рядів.

– Розроблено метод оптимізації прогнозної аналітики часових рядів із використанням стекінгового об'єднання та відбору різнотипних моделей на основі лінійної регресії LASSO, що забезпечує підвищення точності прогнозування та формування оптимального прогнозного ансамблю моделей. Використання стекінгу дозволяє врахувати відмінності у результатах для різних моделей з різними наборами параметрів та ознак і підвищити точність на валідаційних даних, а також на позавибіркових даних, які не входять в тренінгові та валідаційні вибірки.

– За рахунок розроблених методів стекінгового об'єднання різнотипних моделей у прогнозні ансамблі можна підвищити точність у задачах прогнозування та зменшити кількість моделей у стекінговому ансамблі на 30% для певного класу задач;

– Показано, що використання ієрархічної байєсівської моделі дає можливість знаходити прогнозні значення цільової змінної у випадку коротких часових рядів за рахунок використання параметрів ієрархічної моделі, сформованих на основі інших подібних часових рядів, які належать до аналізованої вибірки.

– Запропоновано метод використання байєсівської регресії для стекінгу прогнозних моделей. Цей метод дає можливість отримати розподіли для регресійних коефіцієнтів моделей першого рівня прогнозного ансамблю і оцінити невизначеність, внесену кожною моделлю в результат стекінгу. Інформація про ці розподіли дозволяє вибрати оптимальний набір моделей стекінгу, враховуючи знання з предметної області, у якій проводиться прогнозна аналітика. Імовірнісний підхід для стекінгу прогнозних моделей дозволяє зробити оцінку ризиків для прогнозів, що є важливим у процесі прийняття рішень. За рахунок розробленого методу використання байєсівської регресії для стекінгу прогнозних моделей можна оцінити невизначенність та прогнозні ризики складових моделей при прийнятті експертних рішень щодо формування прогнозного ансамблю моделей.

– Розглянуто лінійну модель для ціни біткоїна, яка включає в себе регресійні ознаки, які базуються на статистиці біткоїна, характеристиках процесів видобутку біткоїна, трендах пошукових запитів Google, візитах на сторінки Вікіпедії. Патерни відхилення регресійної моделі прогнозування від реальних цін біткоїна є простішими у порівнянні з часовими рядами ціни біткоїна. Коректне експертне визначення часових поворотних точок у функції експертної корекції для відхилень регресійної моделі від реальних значень може суттєво покращити точність прогнозу ціни біткоїна.

– На прикладі аналізу впливу кризи, зумовленої пандемією COVID-19 на фінансовий ринок, показано, що кількісні характеристики активності користувачів у мережі Інтернет, зокрема, часові ряди відвідуваності сторінок Вікіпедії, які мають відношення до аналізованої тематики, володіють прогнозним потенціалом. Підхід на основі байєсівського виведення дозволяє аналізувати невизначеність впливу різних фінансових криз. Результати показують, що невизначеність коронавірусної кризи більша порівняно з іншими кризами. Розрахунок невизначеності дозволяє робити оцінку ризиків для інвестиційних портфелів та різних фінансових та бізнес-процесів.

– За рахунок розроблених методів застосування лінійних, ймовірнісних та машинно-навчальних прогнозних моделей з урахуванням аналітичних ознак консолідованих даних заданої предметної області інтелектуального аналізу можна підвищити точність та інформативність результатів у задачах аналізу динаміки попиту та в аналітиці фінансових часових рядів.

– Досліджено різні підходи у логістичній регресії на прикладі проблеми виявлення відмов на виробничих лініях. Машинне навчання дає можливість виявити складні патерни виникнення відмов у виробничих процесах. Узагальнена лінійна модель для логістичної регресії дає можливість дослідити фактори впливу на виникнення відмов. Використовуючи байєсівську модель, можна отримати статистичні розподіли для параметрів моделі, які можна використати при аналізі

ризиків. Використання дворівневого ансамблю дозволяє отримати стабільні диверсифіковані моделі. Різні моделі можуть бути об'єднані у дворівневі стекінгові ансамблі для підвищення точності та стабільності результатів прогнозування. За рахунок розроблених методів у прогнозуванні технічних відмов на лініях збірки на виробництві із використанням стекінгового об'єднання моделей можна оптимізувати набір прогнозних ознак та підвищити точність прогнозування на 1-10%;

– Отримав подальший розвиток метод оптимізації послідовності дій інтелектуального агента в задачах аналітики попиту із використання глибокого Q-навчання та імітаційного моделювання середовища взаємодії. У задачі ціноутворення середовище було змодельовано на основі випадково згенерованого попиту та залежності продажів від додаткової ціни. У задачі попиту та постачання запропоновано використовувати історичний часовий ряд попиту для моделювання середовища. Отримані результати показують, що за допомогою глибокого Q-навчання можна формувати процес прийняття рішень для задач оптимізації цін та задач попиту та постачання, що забезпечує підвищення ефективності прийняття бізнес рішень.

3 ВИКОРИСТАННЯ КОНЦЕПЦІЇ СЕМАНТИЧНОГО ПОЛЯ У ВЕКТОРНІЙ МОДЕЛІ ТЕКСТОВИХ ДОКУМЕНТІВ

Побудуємо теоретико-множинну модель лексемних полів, яка буде описувати як лексико-семантичні, так і тематичні поля у лексемній структурі словників. Розглянемо модель текстових документів у просторі лексемних полів. Розглянемо формування лексемного складу на основі лексемних відношень у текстових масивах. Проаналізуємо використання теоретико-множинної моделі семантичних полів у векторному представленні текстових документів.

3.1 Векторна модель текстових документів у базисі семантичних полів

Розглянемо формування базису лексемних семантичних та тематичних полів для векторного простору текстових документів [262]. Сукупність текстових документів опишемо такою множиною

$$D = \{d_j \mid j = 0, 1, 2, ..., N_d\}, \tag{3.1}$$

де N_d – кількість документів. Під документом з $j = 0$, будемо вважати документ з нейтральним текстом, який відповідає лінгвостилістичній нормі. Документ d_j з множини текстових документів D можна представити як упорядковану множину слів T_j^d, порядок елементів якої відповідає порядку слів у цьому документі:

$$T_j^D = \left\{t_{lj} \mid l = 0, 1, 2, ..., N_d^t\right\} \tag{3.2}$$

Упорядкований за алфавітом словник текстового документа d_j розглянемо як мультимножину W_j^d над множиною словника W

$$W_j^D = \left\{n_{lj}^{wd} \mid w_i \in d_j i = 1, 2, ..., N_w\right\}. \tag{3.3}$$

де n_{lj}^{wd} – кількість входжень лексеми w_i із словника W у множину лексем текстового документа d_j, яку можна визначити як

$$n_{ij}^{wd} = \sum_{l=1}^{N_j^t} f_{wd}(t_{ij}, w_i), \tag{3.4}$$

де

$$f_{wd}(t_{ij}, w_i) = \begin{cases} 1, t_{ij} = w_i \\ 0, w_{ij}^d \neq w_i \end{cases}. \tag{3.5}$$

Відображення лексемного складу словника W на множину семантичних полів S (3.27) задамо таблицею, яка визначена експертним лексикографічним аналізом. Лексемний склад семантичного поля s_k визначимо як

$$W_k^s = \left\{ w_i \mid w \xrightarrow{U_{ws}} s_k, i = 1, 2, ..., N_w \right\}. \tag{3.6}$$

Множину образів відображення U_{ws} розглянемо як мультимножину над множиною семантичних полів S

$$S_f = \left\{ n_k^s(s_k) \mid k = 1, 2, ...N_s \right\}, \tag{3.7}$$

де $n_k^s(s_k)$ – кількість лексем словника W, які відносять до семантичного поля s_k:

$$n_k^s = \sum_{i=1}^{N_w} f_s(w_i, s_k), \tag{3.8}$$

де

$$f_s(w_i, s_k) = \begin{cases} 1, \ w_i \in W_k^s, \\ 0, \ w_i \notin W_k^s. \end{cases}$$

Уведемо мультимножину образів відображення U_{ws} семантичних полів для окремого документа d_j

$$S_j^d = \left\{ n_{kj}^{sd}(s_k) \mid k = 1, 2..., N_s \right\}, \tag{3.9}$$

де n_{kj}^{sd} – кількість лексем семантичного поля в лексемному складі документа d_j

$$n_{kj}^{sd} = \sum_{i=1}^{N_j^t} f_s(t_{li}, s_k), \tag{3.10}$$

де

$$f_s(t_{lj}, s_k) = \begin{cases} 1, t_{lj} \in W_k^s, \\ 0, t_{lj} \notin W_k^s. \end{cases}$$

Уведемо оператор відображення лексемного словника W на множину квантитативних ознак у масиві документів

$$U_{wd} : w_i \to p_{ij}^{wd}, i = 1, 2, ..., N_w, = 1, 2..., N_d. \tag{3.11}$$

У загальному випадку величина p_{ij}^{wd} може представляти довільну квантитативну характеристику. У подальшому будемо розглядати цю величину як текстову частоту лексеми w_i у текстовому документі d_j, яка визначена такою функціональною залежністю

$$p_{ij}^{wd} = \frac{n_{ij}^{wd}}{N_j^t}. \tag{3.12}$$

Аналогічно уведемо оператор відображення семантичного складу S_j^d текстового документа d_j на множину квантитативних ознак:

$$U_{sd} : s_k \to p_{kj}^{sd}, k = 1, 2, ...N_s, j = 1, 2...N_d. \tag{3.13}$$

Величина p_{kj}^{sd} визначає структурну частоту лексем семантичного поля s_k у текстовому документі d_j. Визначимо p_{kj}^{sd} за такою формулою:

$$p_{kj}^{sd} = \sum_{i=1}^{N_w} p_{ij}^{wd} f_s(w_i, s_k), \tag{3.14}$$

де

$$f_s(w_i, s_k) = \begin{cases} 1, \ w_i \in W_k^s, \\ 0, \ w_i \notin W_k^s. \end{cases} .$$

Сукупність значень p_{ij}^{wd} утворює матрицю типу ознака-документ

$$M_{wd} = \left(p_{ij}^{wd}\right)_{i=1,j=1}^{N_w,N_d}. \tag{3.15}$$

У матриці M_{wd} роль ознаки відіграє текстова частота лексеми. Уведемо вектор

$$V_j^w = \left(p_{1j}^{wd}, p_{2j}^{wd}, ..., p_{N_wj}^{wd}\right) \tag{3.16}$$

Такий вектор відображає документ d_j в N_w-мірному просторі текстових документів. Сукупність значень p_{kj}^{sd} утворюють іншу матрицю ознака-документ, у якій ознаками виступають частоти семантичних полів у документах:

$$M_{sd} = \left(p_{kj}^{sd}\right)_{k=1,j=1}^{N_s,N_d} \tag{3.17}$$

Вектор

$$V_j^s = \left(p_{1j}^{sd}, p_{2j}^{sd}, ..., p_{N_sj}^{sd}\right) \tag{3.18}$$

відображає документ d_j в N_s-мірному просторі текстових документів [262, 263].

3.2 Векторна модель текстових документів у базисі тематичних полів

Уведемо поняття тематичного поля за аналогією з семантичним полем [262]. Вважаємо, що тематичне поле утворюють лексеми словника текстових масивів, які характеризують тематику деякої категорії текстових документів. Такі категорії можна визначати, наприклад, на основі дистрибутивних характеристик текстів, згрупованих за деякою визначеною тематикою, авторством текстів, джерелом походження тощо. Множину тематичних полів позначимо так:

$$Them = \{them_i \mid i = 1, 2, ...N_{them},\} \tag{3.19}$$

де $N_{them} = |Them|$ – розмір множини тематичних полів, який визначений кількістю тематичних категорій. Уведемо деякий коефіцієнт, який буде відображати у скільки разів деяку лексему вживають частіше у деякій

категорії у порівнянні із загальною вибіркою усіх категорій. Визначимо цей коефіцієнт як відношення частоти лексеми у документах заданої категорії до частоти цієї ж лексеми у загальній текстові вибірці

$$Kthem_{ij}^{wg} = \frac{p_{ij}^{wg}}{p_i^w}. \tag{3.20}$$

Назвемо $Kthem_{ij}^{wg}$ коефіцієнтом тематичної виразності. Визначимо тематичне поле $them_k$ деякої категорії текстових документів ctg_k як підмножину словника лексем, для яких коефіцієнт тематичної виразності більший за деяке наперед визначене значення:

$$W_k^{them} = \left\{ w_i \mid Kthem_{ij}^{wg}(w_i) > Kthem_i \right\}, \tag{3.21}$$

де $Kthem_i$ – деяке порогове значення коефіцієнта тематичної виразності. На основі визначення множини тематичного поля можна сформувати лексемний склад для кожного тематичного поля, заданого певною категорією текстових документів. Уведення простору семантичних та тематичних полів не тільки зменшує розмірність задачі аналізу текстів, а також вводить новий базис для текстових характеристик. У семантичному базисі можуть спостерігатися якісно нові групування текстових документів. Розгляд таких групувань може бути ефективним в алгоритмах комплексного аналізу текстів.

Розглянемо поняття тематичного поля як сукупності лексем, які в загальному випадку можуть належати різним частинам мови і повинні однозначно відображати понятійний спектр деякої категорії текстових документів. Аналогічно до частот семантичних полів визначимо частоти тематичних полів кожного документа як суми частот лексем, які належать до цього поля:

$$p_{kj}^{(them)d} = \sum_{i=1}^{N_w} p_{ij}^{wd} f_{them}(w_i, them_k), \tag{3.22}$$

де

$$f_{them}(w_i, them_k) = \begin{cases} 1, & w_i \in W_k^{them}, \\ 0, & w_i \notin W_k^{them}. \end{cases}$$

де $p_{kj}^{(them)d}$ – частота тематичного поля у текстовому документі d_j,

W_k^{them} – множина лексем тематичного поля $them_k$, визначена формулою (3.21). Розглянемо матрицю $M_{(them)d}$ типу *тематичні_поля-документи* за аналогією до матриці семантичних полів M_{sd}

$$M_{(them)d} = \left(p_{kj}^{(them)d}\right)_{k=1,j=1}^{N_{them},N_d}, \tag{3.23}$$

де $p_{kj}^{(them)d}$ – частоти тематичних полів, N_{them} – кількість тематичних полів, N_d – кількість текстових документів. Частоти тематичних полів утворюють координати текстових повідомлень у векторному семантичному просторі. Вектор

$$V_j^{them} = \left(p_{1j}^{(them)d}, p_{2j}^{(them)d}, ..., p_{N_{(them)j}}^{(them)d}\right) \tag{3.24}$$

відображає документ d_j в N_w-мірному просторі, базис якого утворений тематичними полями. Використання векторного представлення дає можливість пошуку подібних документів та псевдодокументів у векторному просторі з базисом, утвореним частотними характеристиками семантичних та тематичних полів. Цей базис має суттєво меншу розмірність у порівнянні із базисом, утвореним частотними характеристиками лексем словника текстових масивів [262]. Це дає можливість зменшити кількість необхідних обчислень в алгоритмах аналізу текстів.

3.3 Чисельний аналіз розподілу семантичних полів у текстових документах

Для чисельного аналізу розподілу різних типів семантичних полів у текстових документах ми вибрали два типи текстових документів – текстову вибірку художніх творів англомовної прози та текстові повідомлення груп новин із різних предметних областей [264]. Для аналізу семантичних полів було вибрано 41 семантичне поле іменників і дієслів, сформованих у лінгвістичній системі WordNet і описаних у розділі 1.3.1. Вибірку художніх творів, яка містила 368 творів десяти авторів англомовної прози, було завантажено з інтернет-ресурсу *Project Gutenberg* [265]. Розглянемо закономірності розподілів семантичних полів у авторських текстах. На рис. 3.1 наведено частотний розподіл лексем на основі матриці TF-IDF у

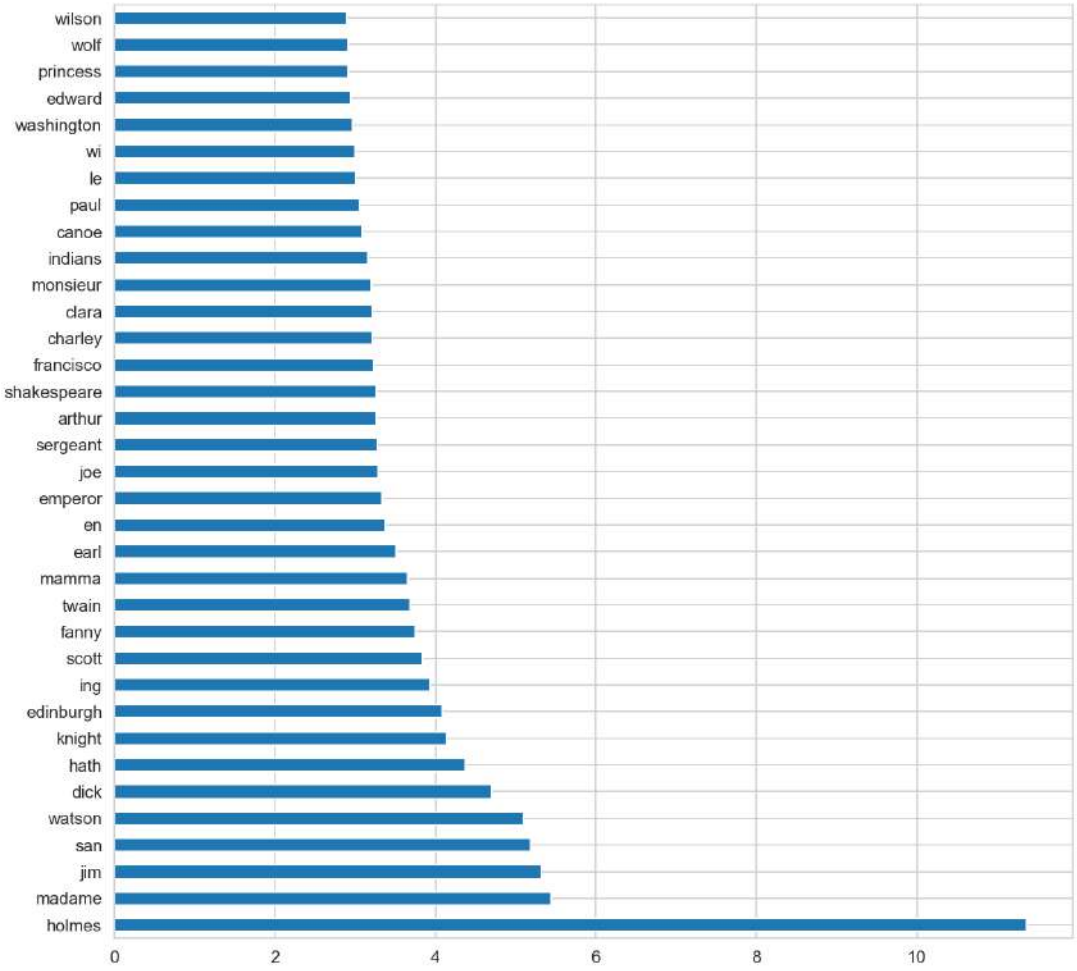

Рисунок 3.1 – Частотний розподіл лексем на основі матриці TF-IDF у текстовій вибірці художніх творів англомовної прози

текстовій вибірці художніх творів англомовної прози. На рис. 3.2, 3.3, 3.3 наведено коробкові графіки довільно вибраних семантичних полів у авторських текстах. На рис. 3.4 наведено функцію щільності розподілу ймовірностей частоти довільно вибраного семантичного поля *verb.communication* для різних класів документів у текстовій вибірці художніх творів англомовної прози. Розглянемо розподіл семантичних полів у просторі меншої розмірності. Для цього використаємо метод головних компонент (PCA) [266] для лінійного зменшення семантичного простору та метод t-SNE перетворення для нелінійного зменшення простору [267]. Розподіл семантичних полів у просторі двох перших компоннент PCA наведено на рис. 3.5, а у двовимірному просторі t-SNE на рис. 3.6.

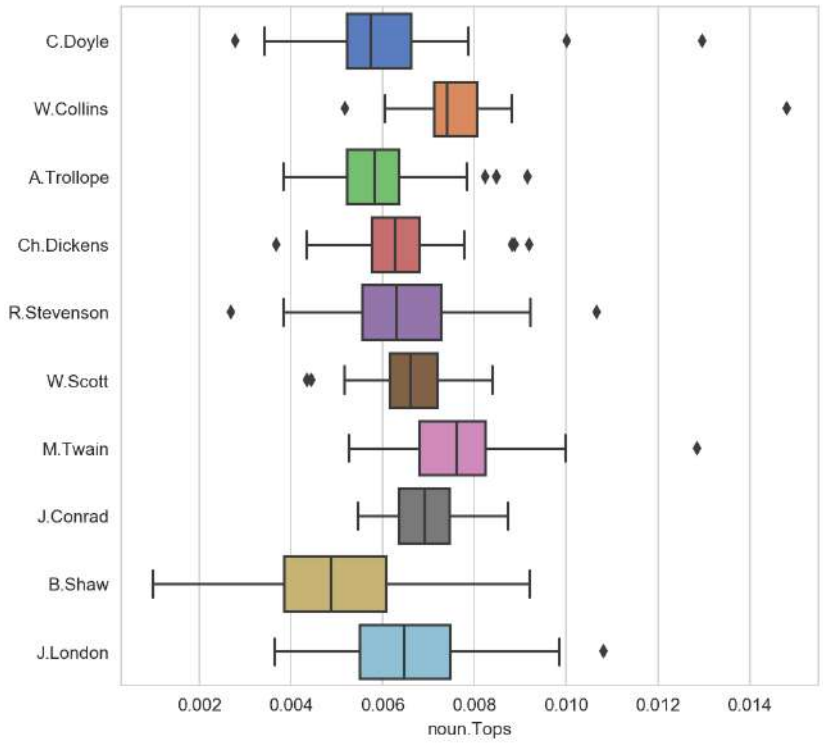

Рисунок 3.2 – Розподіл частот семантичного поля *noun.Tops* за класами документів у текстовій вибірці художніх творів англомовної прози

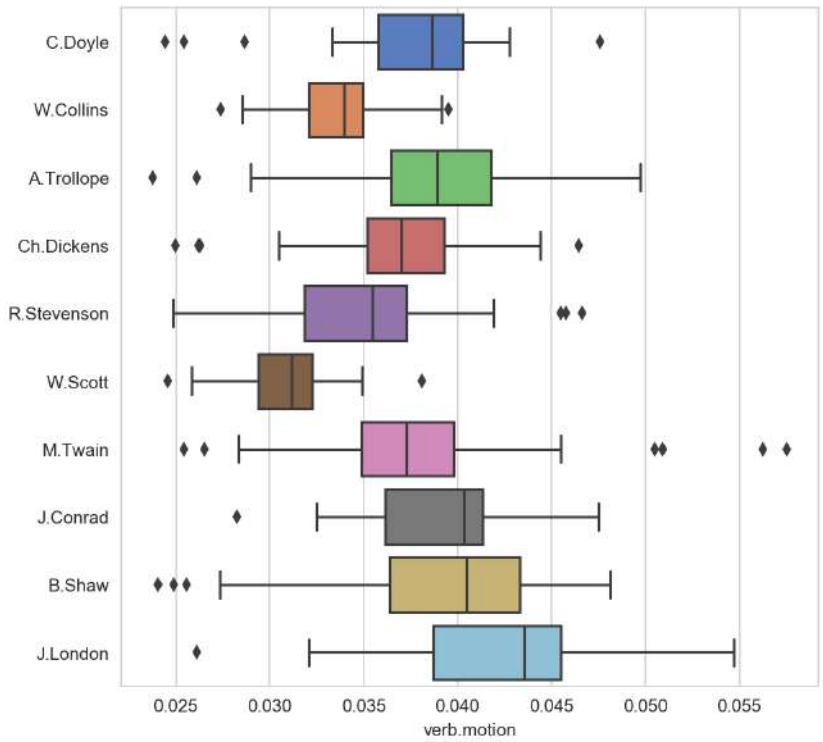

Рисунок 3.3 – Розподіл частот семантичного поля *verb.motion* за класами документів у текстовій вибірці художніх творів англомовної прози

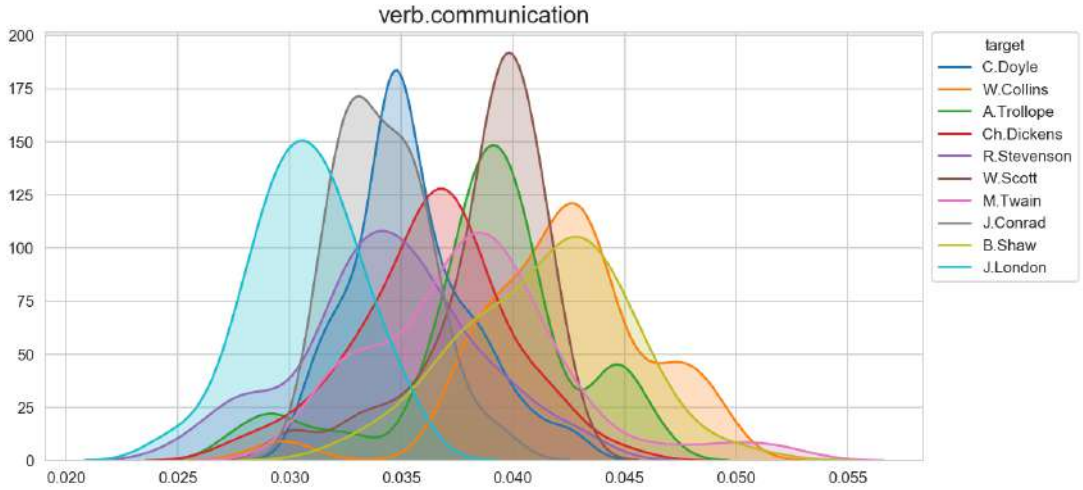

Рисунок 3.4 – Щільність розподілу ймовірностей частоти семантичного поля *verb.communication* для різних класів документів у текстовій вибірці художніх творів англомовної прози

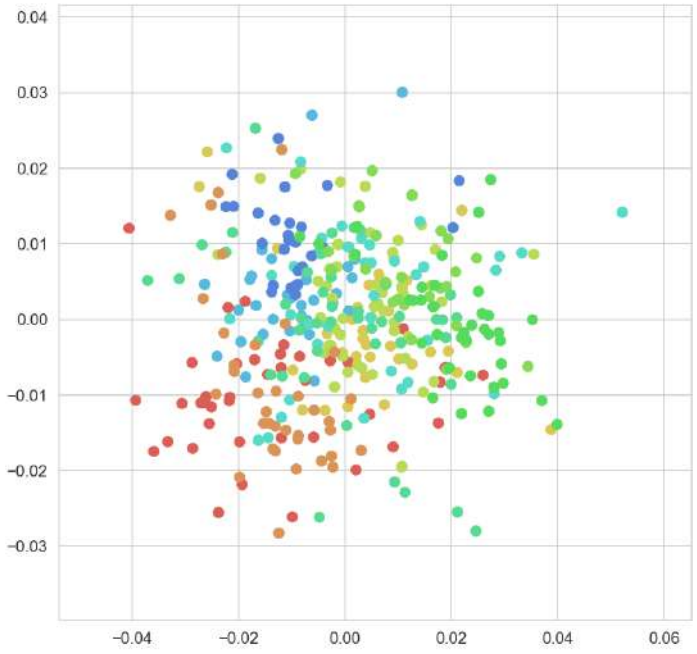

Рисунок 3.5 – Розподіл семантичних полів у двовимірному PCA просторі для вибірки художніх творів англомовної прози

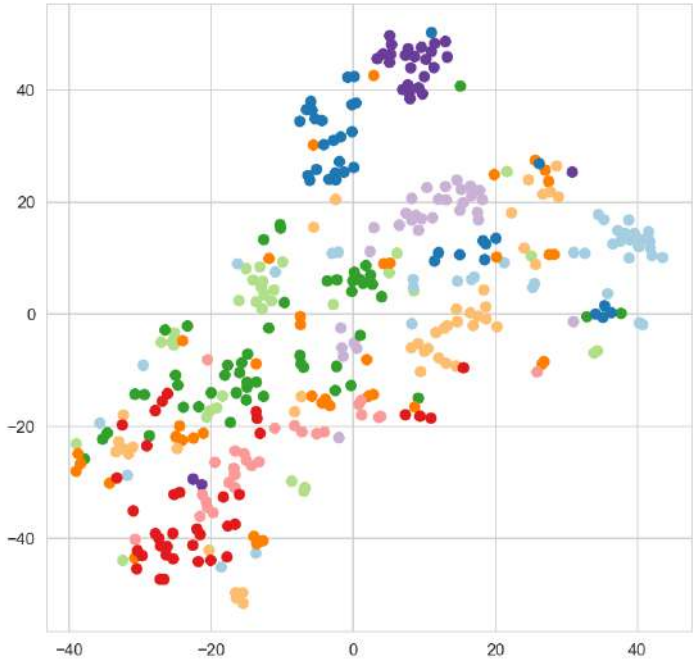

Рисунок 3.6 – Розподіл семантичних полів у двовимірному t-SNE просторі для вибірки художніх творів англомовної прози

Аналіз розподілу семантичних полів було також проведено у вибірці повідомлень груп новин 20 Newsgroups [268]. Ця вибірка містить близько 20000 текстових повідомлень, які рівномірно розподілені по 20 групах новин. Аналогічні розрахунки для розподілів семантичних полів у групах новин наведено на рис. 3.7-3.11

Як випливає із наведених графіків, розподіл семантичних полів за авторськими текстами має більшу дисперсію у порівнянні з таким розподілом по групах новин. Це свідчить про більший класифікаційних потенціал семантичних полів у масивах авторських текстів, а з іншої сторони говорить про більш диференційований авторський стиль у текстах художньої прози у порівнянні з текстами груп новин [264]. Інші розрахунки для семантичних полів наведено у Додатках.

Рисунок 3.7 – Частотний розподіл лексем на основі матриці TF-IDF у вибірці повідомлень груп новин

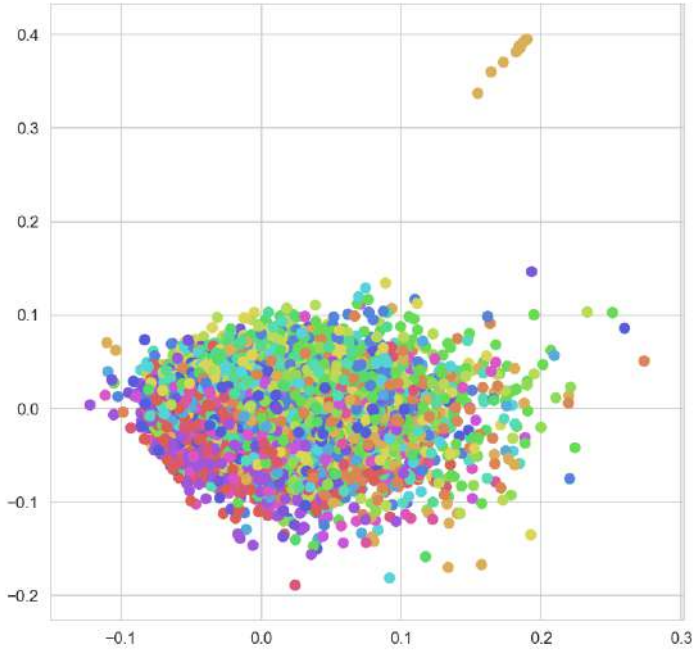

Рисунок 3.8 – Розподіл семантичних полів у двовимірному PCA просторі для вибірки повідомлень груп новин

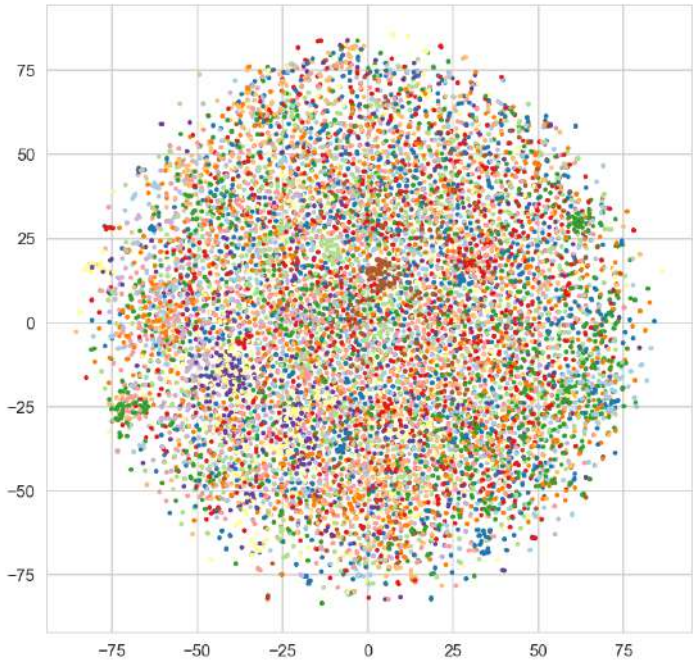

Рисунок 3.9 – Розподіл семантичних полів у двовимірному t-SNE просторі для вибірки повідомлень груп новин

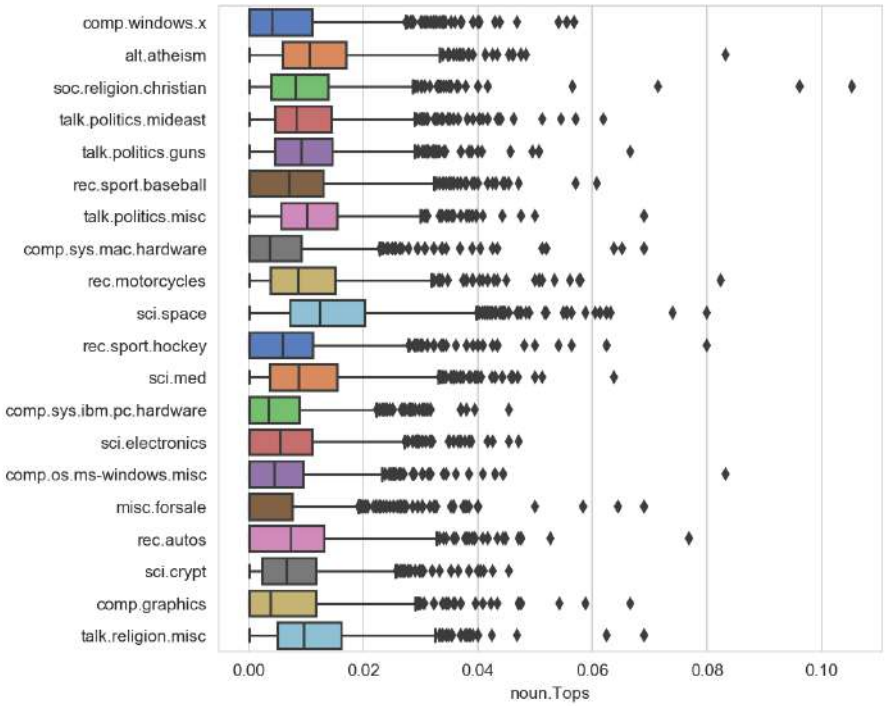

Рисунок 3.10 – Розподіл частот семантичного поля *noun.Tops* по класах документів у вибірці повідомлень груп новин

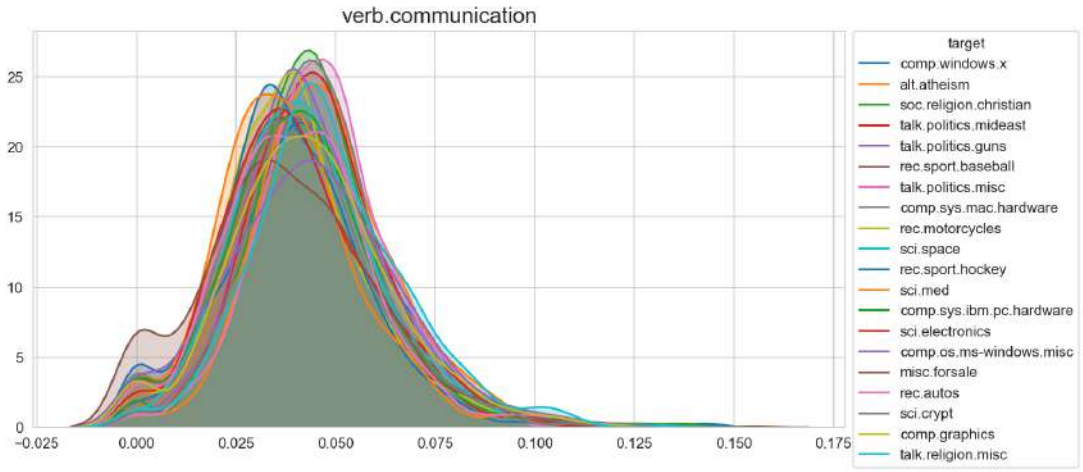

Рисунок 3.11 – Щільність розподілу ймовірностей частоти семантичного поля *verb.communication* для різних класів документів у вибірці повідомлень груп новин

3.4 Чисельний аналіз розподілу тематичних полів у текстових документах

Для чисельного аналізу розподілу різних типів тематичних полів у текстових документах ми вибрали такі ж, як і у випадку аналізу семантичних полів два типи текстових документів – текстову вибірку художніх творів англомовної прози та текстові повідомлення груп новин з різних предметних областей [264]. Для відбору лексем у тематичні поля було вибрано поріг для коефіцієнта тематичної виразності $Kthem_{ij}^{wg}$, який дорівнює 2. Тобто, до тематичних полів входили лексеми, які у текстах заданого класу мають текстову частоту у два рази більшу, ніж у текстах інших класів. Розглянемо результати чисельного аналізу розподілу тематичних полів у вибірці художніх творів англомовної прози. На рис. 3.12 зображено щільність розподілу ймовірностей частоти тематичного поля для різних класів документів. На рис. 3.23 зображено розподіл тематичних полів у двовимірному PCA просторі, на рис. 3.14 зображено розподіл тематичних полів у двовимірному t-SNE просторі. Також для аналізу була вибрана стандартизована колекція текстових документів груп новин 20 Newsgroups [268]. Результати розрахунків для розподілів тематичних полів у групах новин наведено на рис. 3.15–3.15. Додаткові розрахунки для інших тематичних полів наведено у Додатках. Як випливає із отриманих даних, тематичні поля мають значний класифікаційний потенціал. Це дає

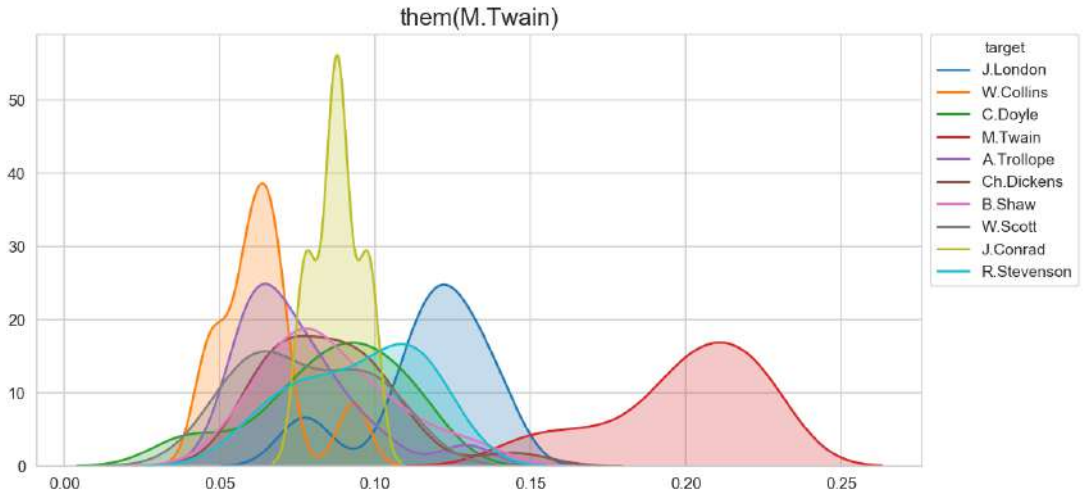

Рисунок 3.12 – Щільність розподілу ймовірностей частоти тематичного поля *them(M_ Twain)* для різних класів документів у вибірці художніх творів англомовної прози

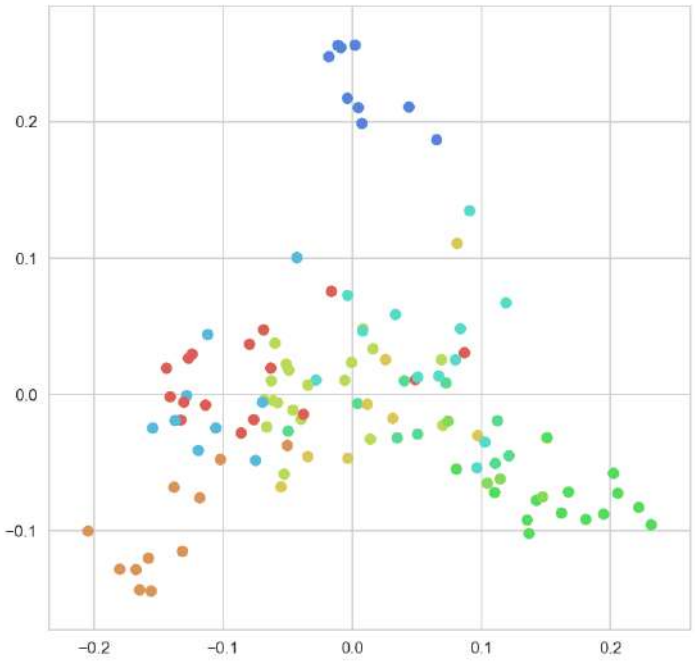

Рисунок 3.13 – Розподіл тематичних полів у двовимірному PCA просторі для вибірки художніх творів англомовної прози

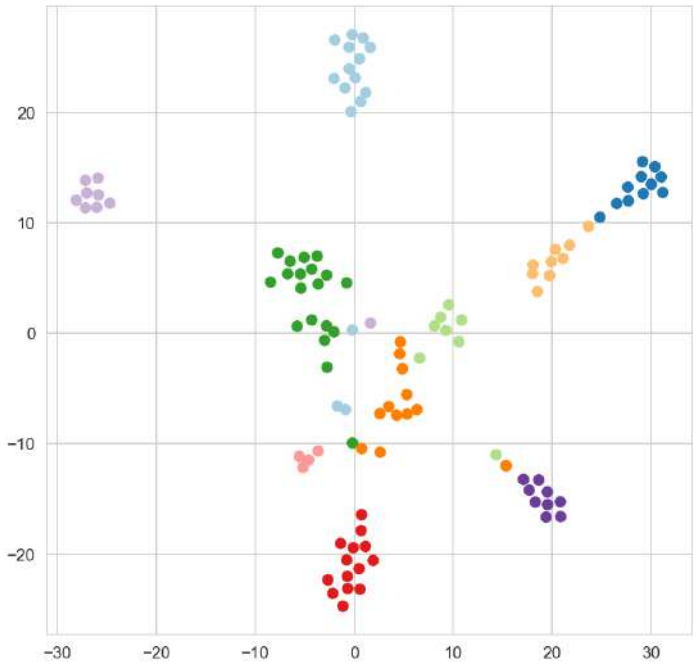

Рисунок 3.14 – Розподіл тематичних полів у двовимірному t-SNE просторі для вибірки художніх творів англомовної прози

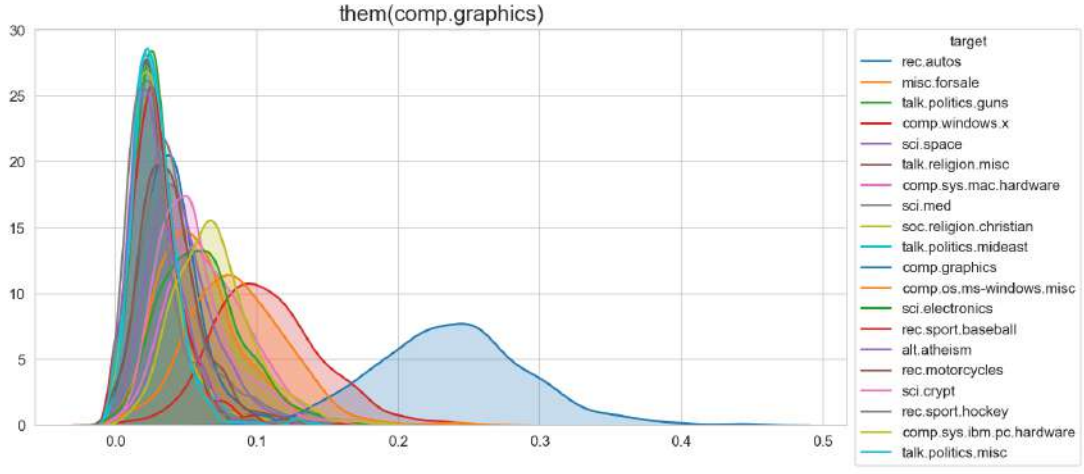

Рисунок 3.15 – Щільність розподілу ймовірностей частоти тематичного поля *them(comp_graphics)* для різних класів документів у вибірці повідомлень груп новин

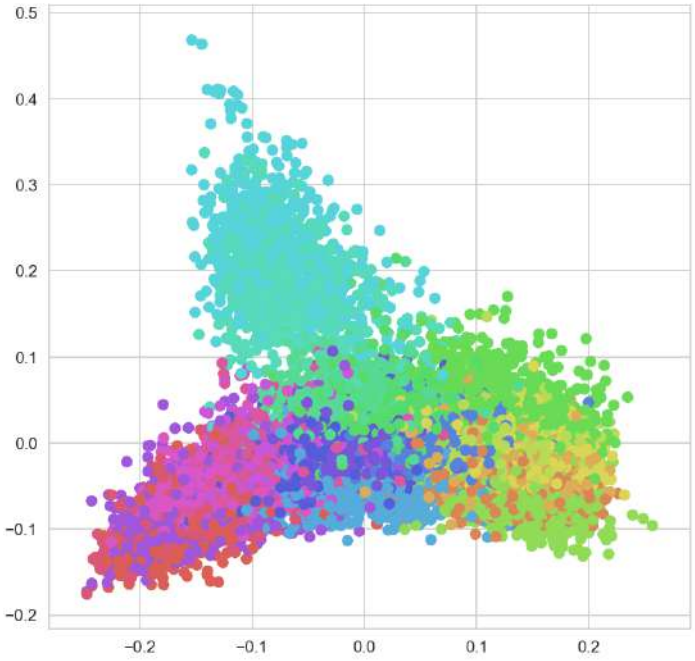

Рисунок 3.16 – Розподіл тематичних полів у двовимірному PCA просторі у вибірці повідомлень груп новин

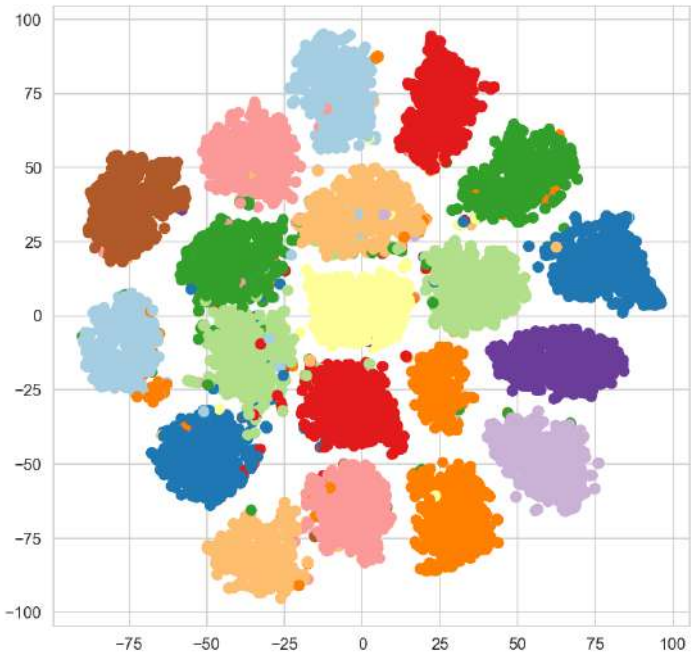

Рисунок 3.17 – Розподіл тематичних полів у двовимірному t-SNE просторі у вибірці повідомлень груп новин

можливість використовувати кількісні ознаки на основі тематичних полів у задачах інтелектуального аналізу текстових даних, зокрема, у класифікації текстових документів.

3.5 Теоретико-множинна модель лексемних полів

Розглянемо теоретико-множинну концепцію семантичного поля [263]. Нехай існує деякий словник лексем, які зустрічаються в аналізованих текстових масивах. Опишемо цей словник як впорядковану множину

$$W = \{w_i \mid i = 1, 2, ..., N_w\}, \tag{3.25}$$

де N_w – кількість лексем у словнику. Уведемо множину семантичних полів

$$S = \{S_k \mid k = 1, 2, ..., N_s\}, \tag{3.26}$$

де N_s – кількість семантичних полів. Семантичні ознаки лексем будемо характеризувати відображенням

$$U_{WS} : W \to S, w_i \to s_k, i = 1, 2, ..., N_w, k = 1, 2, ..., N_s. \tag{3.27}$$

Тобто у відповідність кожній лексемі ставлять деякий елемент множини S. Множина значень S може мати різну природу, наприклад, це може бути множина назв деяких семантичних класів. Шкала семантичних ознак є номінальною, якщо лексеми набувають деяких назв із множини S. Номінальна шкала володіє класифікаційним потенціалом, коли за допомогою відображення (3.27) можна утворити групування елементів множини W, які мають спільні назви з множиною S. У загальному, класифікацію лексем за семантичними полями будемо розглядати як відображення множини лексем на множину семантичних полів. Семантичну класифікацію розглянемо як деяку сукупність відображень лексем на множину дійсних чисел. Можливу квантифікацію лексемних відображень можна пов'язати з частотами лексем в текстових об'єктах. Розглянемо утворення семантичного поля на основі

відношення еквівалентності. Нехай існує деяке бінарне відношення

$$S_k^b \subseteq W \times W. \tag{3.28}$$

Розглянемо деяку квантитативну ознаку лексеми $x_k^s(w_i)$, яка кількісно характеризує лексемні відношення заданого типу у множині аналізованих текстових об'єктів. Наприклад, це може бути частота появи лексеми у заданому лексемному шаблоні. Пов'яжемо із ознакою $x_k^s(w_i)$ бінарне відношення

$$S_k^b = \{(w_i, w_j) \mid x_k^s(w_i) = x_k^s(w_j)\}. \tag{3.29}$$

Можна показати, що відношення S_k^b є рефлексивним, тобто

$$(w_i, w_j) \in S_k^b, \forall w_i, w_i \in W, \tag{3.30}$$

симетричним, тобто

$$(w_i, w_j) \in S_k^b \Rightarrow (w_j, w_i) \in S_k^b, \forall w_i, w_i \in W, \tag{3.31}$$

і транзитивним, тобто

$$(w_i, w_j) \in S_k^b, (w_j, w_l) \in S_k^b \Rightarrow (w_i, w_l) \in S_k^b. \forall w_i, w_i, w_l \in W. \tag{3.32}$$

Рефлексивне, симетричне і транзитивне відношення називають еквівалентністю [269]. Еквівалентність S_k^b повністю характеризує ознаку $x_k^s(w_i)$, яка його породжує і дає можливість визначити множину лексем, які не розрізняють за цією ознакою :

$$S_k^c = \{w_i \mid (w_i, w_j) \in S_k^b\}. \tag{3.33}$$

Якщо S_k^c є деяким семантичним відношенням, тоді множини, що не співпадають, утворюють розбиття лексемного словника W на семантичні класи

$$S_{sk} = \{S_k^c \mid k = 1, 2, ..., N_s\}. \tag{3.34}$$

Такі семантичні класи, враховуючи теорію лексико-семантичних полів, можна розглядати як лексемні поля. Бінарне відношення S_k^b може також породжуватися деяким логічним висловлюванням $Q(w_i, w_j)$

$$S_k^b = \{(w_i, w_j) \mid Q(w_i, w_j) = True\}, \tag{3.35}$$

де $Q(w_i, w_j)$ описує деяку умову, наприклад, одночасне використання в текстових шаблонах заданої структури. Умова породження бінарного відношення S_k^b може також описуватися деяким правилом підстановки у заданій схемі формальної граматики. Таке правило може бути сформоване деяким регулярним виразом. Розглянемо рангову ознаку $r_k^{rs}(w_i)$, яка утворює бінарне відношення

$$S_k^{rb} = \{(w_i, w_j) \mid x_k^s(w_i) \leq x_k^s(w_j)\}. \tag{3.36}$$

Можна показати, що таке бінарне відношення є рефлексивним, транзитивним та лінійним. Такі відношення називають лінійними квазіпорядками [269]. Квазіпорядок S_k^{rb} породжує рангову шкалу семантичного поля S_k^r. У випадку формування семантичного поля за допомогою рангових ознак, можна визначити внутрішню структуру поля, для якої можна сформувати внутрішній частковий порядок, виділивши структурні групи всередині семантичного поля. Такими групами можуть бути, наприклад, частотне ядро семантичного поля, основна частотна область, периферійна частотна область. Для кожної з цих груп можна визначити умови для семантичної ознаки, за якою лексеми всередині цих груп не розрізняють. Відношення еквівалентності та квазіпорядку визначають номінальні та рангові семантичні шкали для лексемного складу словника текстових масивів на основі лексемних відношень елементів різних класів семантичного розбиття.

У семантичному розбитті словника можна виявити відповідні структурні зв'язки [263]. Семантичним розбиттям із структурою назвемо пару (S^W, Z_{str}), де S^W – розбиття лексемного словника на відповідні

семантичні класи. Відношення

$$Z_{str} \subseteq \{1, 2, ...N_{sc}\} \times \{1, 2, ...N_{sc}\} \tag{3.37}$$

є відношенням відповідних зв'язків між семантичними класами, де N_{sc} - кількість семантичних класів розбиття. Умова $(i, j) \in Z_{str}$ означає, що існує зв'язок між лексемами семантичних класів S_i^W та S_j^W. Відношення Z_{str} можна представити за допомогою булевої матриці $M^{str} = \left\{m_{ij}^{str}\right\}$, вважаючи, що $m_{ij}^{str} = 1$ тоді, коли $(i, j) \in Z_{str}$. Аналізуючи матрицю M^{str}, можна виявити сегменти розбиття Z_{str}, тобто групи класів, які пов'язані між собою за допомогою розбиття. Під сегментом розбиття розуміють деяку множину, утворену об'єднанням класів розбиття. Здійснюючи перестановку рядків та стовпців цієї матриці, можна виявити підматриці, які заповнені одиницями. Індекси стовпців та рядків будуть описувати індекси семантичних класів, які можна віднести до єдиного спільного сегменту. Такі підматриці формують бінарні кластери, тобто групи рядків та стовпців, об'єднаних спільними властивостями. У цьому випадку елементи матриці, які відповідають цим рядкам та стовпцям дорівнюють одиниці. Визначимо бінарний кластер у структурному відношенні Z_{str}. Нехай X_{str}^Z є множиною рядків матриці Z_str, а Y_{str}^Z – множина стовпців. Матрицю Z_str позначимо так

$$Z_{str} = \left(X_{str}^Z, Y_{str}^Z\right). \tag{3.38}$$

Підмножини

$$I_{str}^Z \subseteq X_{str}^Z, J_{str}^Z \subseteq Y_{str}^Z, \tag{3.39}$$

утворюють підматрицю

$$M_{bc} = \left(I_{str}^Z, J_{str}^Z\right),$$

яка описує бінарний кластер за рядками та стовпцями підматриці, якщо виконується умова

$$M_{bc} = \left\{m_{ij}^{str} \mid m_{ij}^{str} = 1, i \in I_{str}^Z, j \in J_{str}^Z\right\}. \tag{3.40}$$

Враховуючи (3.38)–(3.40), семантичний сегмент можна визначити,

наприклад, так

$$S_k = \left\{ S_i^R \mid i \in \left(I_{str}^Z \cup J_{str}^Z \right) \right\}. \tag{3.41}$$

Виходячи із наведеного аналізу семантичним полем назвемо сегмент, який утворюється семантичними класами, об'єднаними бінарним кластером у структурному відношенні семантичного розбиття лексемного словника текстових масивів. Отже, за допомогою пари (S^W, Z_{str}), яка описує розбиття словника на семантичні класи із структурою, можна задавати семантичні поля лексемного словника [263].

3.6 Утворення семантичних полів на основі лексемних відношень

Розглянемо можливість утворення семантичних полів на основі дистрибутивної гіпотези та гіпотези латентних відношень [263]. Визначимо множину W_k^{sc} деяких базових лексем, що будуть виконувати роль ядра, яке утворює семантичне поле на основі деякого перетворення. Вміст семантичного поля буде утворюватися лексемами, які мають задане відношення до лексем множини W_k^{sc}. Такі відношення можуть бути, наприклад, зумовлені утворенням лексемних пар у тексті лексем семантичного поля з лексемами заданого ядра, або приналежністю лексеми семантичного поля і лексеми ядра, що утворює семантичні поля до спільних логічних сегментів текстів. Такими логічними сегментами можуть бути, наприклад, речення, абзаци, розділи. Розглянемо спочатку утворення семантичних полів на основі лексемних пар. Розглянемо текст як ланцюжок лексем

$$T = l_1 l_2 ... l_{N_t}. \tag{3.42}$$

Будемо вважати, що множина лексем в тексті впорядкована за номером лексеми в тексті. Розглянемо деяке відношення

$$I_k^s \subseteq W \cdot W. \tag{3.43}$$

Пара (w_i, w_j) належить відношенню I_k^s, якщо перша лексема w_i належить до ядра W_k^{sc}, що формує поле, а друга лексема w_j утворює з нею пару лексемного

сполучення у тексті, тобто, виконується умова

$$w_i \in T, w_j \in T, w_i \in W_k^{sc}, w_i \prec w_j. \tag{3.44}$$

Необхідно врахувати, що у великих текстових масивах можуть зустрічатися деякі лексемні пари, які не відображають стійких семантичних зв'язків. Тому необхідно визначити додаткову умову, за допомогою якої можна було б вилучати такі пари з розгляду. Аналогічно до теорії частих множин [171, 172, 174, 176, 177], де розглядається поняття підтримки частих множин, розглянемо поняття підтримки лексемних сполучень. Підтримкою лексемного сполучення для лексеми w_j будемо вважати відношення кількості лексемних сполучень цієї лексеми з лексемами ядра, що утворює поле, до загальної кількості появ цієї лексеми в аналізованому текстовому масиві

$$Supp(w_j) = \frac{n((w_i, w_j) \in I_k^s)}{n(w_j)}. \tag{3.45}$$

До подальшого розгляду будемо брати лексеми, для яких величина підтримки $Supp(w_j)$ буде перевищувати деяке наперед задане число

$$Supp(w_j) \geq Supp_{\min}. \tag{3.46}$$

Враховуючи (3.44)-(3.46), семантичне поле можна визначити як

$$S_k = \{w_j \mid \exists w_i \in W_k^{sc}, (w_i, w_j) \in I_k^s, Supp(w_j) \geq Supp_{\min}\}. \tag{3.47}$$

Аналогічно можна визначити семантичне поле на основі умови спільної появи лексеми із множини ядра, що утворює поле, та лексеми семантичного поля у деякому заданому логічному сегменті тексту, наприклад, реченні. Розглянемо текст як впорядковану множину логічних сегментів:

$$T = \left\{g_1, g_2, ...g_{N_g}\right\}. \tag{3.48}$$

У цьому випадку семантичне поле можна розглядати як таку множину

$$S_k = \{w_j \mid \exists w_i \in W_k^{sc}, \exists g_m \in T, (w_i, w_j) \in g_m, Supp_{sg}(w_j) \geq Supp_{min}\}. \tag{3.49}$$

Підтримка $Supp_{sg}(w_j)$ має аналогічний зміст до підтримки лексемних сполучень (3.46), замість кількості лексемних сполучень у тексті розглядається кількість логічних текстових сегментів. Розглянуті лексемні відношення можна узагальнити для випадку лексемних відношень у складному текстовому шаблоні з заданою множиною додаткових умов.

Отже, семантичні класи утворюються як відношення еквівалентності. Семантичне поле визначається як сегмент, який утворюється семантичними класами, об'єднаними бінарним кластером у структурному відношенні семантичного розбиття лексемного словника текстових масивів. Розглянуто відношення, яке описує розбиття словника на семантичні класи із структурою, яка визначає семантичні поля лексемного словника. Проаналізовано утворення семантичних полів на основі лексемних відношень, зокрема, таких як сполучення у тексті лексем семантичного поля та лексем множини, яка утворює семантичне поле. Використання концепції семантичних полів є ефективним у векторній моделі текстових документів унаслідок зменшення розмірності фазового простору представлення документів [263]. Це дає можливість зменшити кількість необхідних обчислень в алгоритмах аналізу текстових даних.

3.7 Моделювання нечітких семантичних полів у масивах текстових документів

Враховуючи денотативні закономірності лексикографічного складу словника, можна зауважити, що одні і ті ж лексеми можуть знаходитись у різних семантичних полях. Не існує чіткого розмежування семантичних полів. Таку особливість доцільно враховувати в алгоритмах аналізу текстів. Перспективним є створення моделі семантичних полів на основі теорії нечітких множин [270, 271]. Моделювання та аналіз нечіткої структури семантичних полів є актуальною задачею при створенні алгоритмів аналізу даних із використанням векторного представлення текстових об'єктів. Використовуючи теоретико-множинний підхід, створимо модель нечітких семантичних полів [272]. Уведемо клас семантичних полів як клас нечітких множин α-рівня. Використаємо поняття лінгвістичної змінної для відображення структурних нечітких зв'язків між елементами семантичних

полів. Розглянемо словник W як універсальну множину із теорії нечітких множин. При моделюванні семантичних полів на основі звичайних множин розглянемо характеристичну функцію множини семантичного поля S_k:

$$\mu_k^c(w_i) = \begin{cases} 0, w_i \notin S_k, \\ 1, w_i \in S_k. \end{cases} \tag{3.50}$$

Функція $\mu_k^c(w_i)$ набуває лише двох значень – одиниці, у випадку приналежності лексеми w_i семантичному полю S_k, і нуля, у випадку, якщо w_i не належить цій множині. Якщо допустити, що функція $\mu_k^c(w_i)$ може набувати проміжних значень в інтервалі [0,1], тоді множина, яка буде описуватися такою функцією, буде називатися нечіткою множиною [270, 271]. Нечітким семантичним полем $\tilde{S}_k$ назвемо пару $(W, \mu_k^s(w_i))$, де W – універсальна множина словника лексем, $\mu_k^s(w_i)$ - функція, визначена на множині W, яка приймає значення на відрізку [0,1]. Таку функцію $\mu_k^s(w_i)$ назвемо функцією приналежності лексеми w_i нечіткому семантичному полю $\tilde{S}_k$. Загальну форму запису нечіткого семантичного поля будемо розглядати у вигляді

$$\tilde{S}_k = \sum_{i=1}^{N_w} \frac{\mu_k^s(w_i)}{w_i}, w_i \in W. \tag{3.51}$$

Точками переходу нечіткого семантичного поля назвемо лексеми, для яких $\mu_k^s(w_i) = 0.5$. Нечітке семантичне поле можна зобразити за допомогою діаграми Заде [270, 271], яка є графіком функції $\mu_k^s(w_i)$. Також можна зобразити у вигляді сингельтона – пари $(W, \mu_k^s(w_i))$, де на першому місці знаходиться назва лексеми, а на другому – величина її приналежності семантичному полю $\tilde{S}_k$. Сингельтон називають чітким, якщо $\mu_k^s(w_i) = 1$. Очевидно, що розглядаючи деяке семантичне поле, можна знайти ненульові значення приналежності цьому полю для більшості лексем словника. Тому доцільно визначити деякі додаткові критерії формування нечіткого семантичного поля. Такі критерії визначимо, використовуючи поняття нечітких множин α-рівня [270, 271]. Нечітким семантичним полем α-рівня $\tilde{S}_k^\alpha$ назвемо множину лексем ($w_i \in W$), для яких виконується умова $\mu_k^s(w_i) > \alpha$.

Отже,

$$\tilde{S}_k^{\alpha} = \{w_i \mid \mu_k^s(w_i) > \alpha\}. \tag{3.52}$$

Можна показати, що нечітке семантичне поле $\tilde{S}_k$ можна розкласти за полями всіх α-рівнів у вигляді

$$\tilde{S}_k = \sum_{\alpha} \alpha \tilde{S}_k^{\alpha}. \tag{3.53}$$

Складним завданням у теорії нечітких множин є побудова функції приналежності $\mu_k^s(w_i)$. Основними методами побудови таких функцій є методи експертних оцінок [270, 271]. Можна виділити прямі та опосередковані методи. У прямому методі експерти напряму задають значення функції приналежності для кожної лексеми семантичного поля, наприклад, на основі лексикографічного аналізу. В опосередкованому методі здійснюють попарні порівняння елементів нечіткої множини. Такі попарні порівняння можуть здійснюватися на основі порівняння текстових частот лексемних відношень, які характерні для заданих лексемних шаблонів аналізованого семантичного поля. В результаті отримаємо квадратну матрицю попарних порівнянь лексем семантичного поля з одиничними діагональними елементами $a = \{a_{ij}\}$. Можна припустити, що

$$a_{ij} = \frac{\mu_k^s(w_i)}{\mu_k^s(w_j)}. \tag{3.54}$$

На основі утвореної матриці A розглянемо рівняння

$$A\vec{\mu_k^s} = \lambda\vec{\mu_k^s}. \tag{3.55}$$

Із розв'язку рівняння (3.55) виберемо вектор $\vec{\mu_k^s}$, який відповідає найбільшому власному значенню матриці A. Елементи вектора $\vec{\mu_k^s}$ можна розглядати як наближення значень приналежності лексем нечіткого семантичного поля $\tilde{S}_k$. Звичайне семантичне поле можна розглядати як частковий випадок нечіткого семантичного поля. Зв'язок між звичайним та нечітким семантичним полем можна визначити через наближену апроксимацію функції приналежності

характеристичною функцією

$$\mu_k^c(w_i) = \begin{cases} 1, \mu_k^s(w_i) \geq 0.5, \\ 0, \mu_k^s(w_i) < 0.5. \end{cases} \tag{3.56}$$

Звичайним семантичним полем, найближчим до нечіткого семантичного поля, назвемо поле із характеристичною функцією $\mu_k^c(w_i)$, яка визначається виразом (3.56). Вимір нечіткості семантичного поля можна визначити як відстань від його множини до множини найближчого до нього звичайного семантичного поля у заданій метриці. В евклідовій метриці індекс нечіткості семантичного поля можна обрахувати так

$$Ind_f^e(\tilde{S}) = \frac{2}{s\sqrt{N_w}\sqrt{\sum_{i=1}^{N_w}(\mu_k^s(w_i) - \mu_k^c(w_i))^2}}. \tag{3.57}$$

У семантичному полі можна виділити структурні семантичні групи, зокрема, синонімічні ряди, які можна розглядати як підмножини семантичного поля. Підмножиною нечіткої множини семантичного поля $\{W, \mu_k^s(w_i)\}$ будемо називати нечітку множину $\{W, \mu_\alpha(w_i)\}$, для якої виконується нерівність

$$\mu_a(w_i) < \mu_k^s(w_i). \tag{3.58}$$

Аналогічно до визначення нечіткого семантичного поля можна визначити поняття семантично нечіткої лексеми $\tilde{w}$. Такою лексемою будемо називати пару

$$\tilde{w_i} = \left(S, \mu_k^w(s_k)\right), \tag{3.59}$$

де S – універсальна множина семантичних полів, $\mu_k^w(s_k)$ – функція приналежності, визначена на множині S, яка приймає значення на відрізку [0,1]. За допомогою такої функції приналежності можна характеризувати спектр семантичних полів лексеми. Цю функцію можна визначати як за допомогою експертного аналізу, так і на основі статистичних характеристик текстового розподілу даної лексеми у заданих шаблонах, які відповідають певним семантичним полям. Розглянемо опис нечітких лексем $\tilde{w_i}$ за допомогою лінгвістичної змінної [270, 271] з урахуванням особливостей

семантичних полів. Лінгвістичною змінною лексеми $\tilde{w}_i$ назвемо набір

$$L_s(\tilde{w}) = \{w_i, T_w(w_i), S, G_{synt}, M_{sem}\}, \tag{3.60}$$

де w_i – назва змінної, $T_w(w_i)$ – терм-множина імен значень змінної, S – універсальна множина семантичних полів, G_{synt} – синтаксичне правило утворення імен значень змінної w_i, M_{sem} – семантичне правило, яке ставить у відповідність нечітку підмножину універсальної множини семантичних полів S кожному елементу терм-множини $T_w(w_i)$. Синтаксичне правило розглядають як алгоритмічну процедуру породження елементів множини $T_w(w_i)$, а семантичне правило як процедуру обрахунку функцій приналежності нечітких підмножин універсальної множини S. Уведення поняття лінгвістичної змінної в опис семантично нечітких лексем є ефективним тоді, коли синтаксична процедура полягає в утворенні терм-множини на основі конкатенації назви лінгвістичної змінної з назвами елементів семантичних полів інших частин мови. Зв'язок лінвістичних змінних елементів одних семантичних полів із елементами інших семантичних полів через синтаксичну процедуру формування терм-множин відображає внутрішню ієрархічну структуру нечітких семантичних полів лексемного словника текстових масивів. Окремим питанням є семантична процедура побудови функцій приналежності нечітких підмножин універсальної множини лексемного словника S, імена яких є елементами терм-множини лінгвістичної змінної. Одним із можливих шляхів розв'язку цього завдання є аналіз статистичних характеристик розподілу лексемних сполучень терм-множини у текстових шаблонах, які відповідають аналізованому семантичному полю. На основі знайдених текстових частот лексемних сполучень можна побудувати табличну апроксимацію функції приналежності для кожного лексемного сполучення, яке належить терм-множині аналізованої лінгвістичної змінної. Аналізуючи статистичні розподіли лексемних сполучень, можна об'єднати синтаксичне та семантичне правила формування лінгвістичної змінної. Нехай, наприклад, існує деякий елемент w_i множини семантичного поля іменників S_1 і деяка множина

прикметників семантичного поля S_2. Розглянемо бінарне відношення

$$I_b(w_j) \subseteq S_2 \times \{ w_j \}. \tag{3.61}$$

Будемо розглядати пари $(w_i, w_j) \in S_2 \times \{ w_j \}$ які утворюють лексемні сполучення $w_i w_j$ в тексті. На основі статистичного аналізу текстових масивів можна виявити текстові частоти $p(w_i, w_j)$ для цих сполучень. Виберемо деяке мінімальне значення текстових частот, за яким будемо формувати бінарне відношення $I_b(w_j)$. Будемо вважати, що

$$I_b(w_j) = \begin{cases} (w_i, w_j), w_i \in S_2, \\ (w_i, w_j) \in S_2 \times \{ w_j \}, \\ p(w_i w_j) > p_{\min}. \end{cases} \tag{3.62}$$

Правило (3.62) формування бінарного відношення $I_b(w_j)$ будемо розглядати як статистичне відображення синтаксичного правила формування лінгвістичної змінної, а набір частот $p(w_i w_j)$ із умовою $p(w_i w_j) > p_{\min}$ як деякі коефіцієнти для функції приналежності. Табличну табуляцію функції приналежності з врахуванням цих коефіцієнтів можна здійснювати на основі експертного аналізу, враховуючи як денотативні лексикографічні значення лексем так і їх конотативні значення в тексті. На основі правила (3.62) формування бінарного відношення $I_b(w_j)$ можна побудувати терм-множину для лексеми w_j

$$T_w(w_j) = \{ w_i w_j \mid (w_i, w_j) \in I_b(w_j) \}. \tag{3.63}$$

Запис $w_i w_j$ означає словосполучення лексем внаслідок конкатенації w_i та w_j. В загальному випадку словосполучення $w_i w_j$ слід розглядати не просто як з'єднання двох лексем, а як входження цих лексем в один із наперед визначених шаблонів. В результаті ми отримаємо визначення статистичної лінгвістичної змінної для нечіткої лексеми $\tilde{w}_i$ як

$$L(\tilde{w}_i) = \{ w_i, T_w(w_i), St(w_i) \}, \tag{3.64}$$

де w_i – назва статистичної лінгвістичної змінної, $T_w(w_i)$ - терм-множина, $St(w_i)$ – статистична процедура утворення терм-множини. Отже, на основі теорії нечітких множин створено модель нечіткого семантичного поля лексемного складу текстових масивів. Визначено характеристики для нечіткого семантичного поля - функцію приналежності, найближче звичайне семантичне поле, міру нечіткості семантичного поля, семантичне поле α-рівня. Поряд із поняттям нечіткого семантичного поля уведено поняття семантично нечіткої лексеми, для якої визначено лінгвістичну змінну. Формування синтаксичних та семантичних правил для лінгвістичних змінних нечітких лексем дає можливість визначити ієрархічну структуру нечітких семантичних полів. Визначено статистичну лінгвістичну змінну семантично нечіткої лексеми. Для такої змінної терм-множина формується на основі статистичних характеристик розподілу лексемних сполучень. Запропоновані в роботі характеристики семантичних полів та лексем дають можливість відобразити нечіткість семантичної структури словника в алгоритмах інтелектуального аналізу текстових масивів [272, 273].

3.8 Модель вторинних семантичних полів в ортонормованому базисі

Семантичні поля розглядають як групи лексем, об'єднаних спільним поняттям. Такі групи лексем утворюють нові характеристики текстових даних, використання яких є ефективним у задачах кластеризації та класифікації текстових документів. Формування додаткових семантичних ознак на основі концепції семантичних полів утворює новий семантичний простір, що збільшує можливості аналізу векторного простору текстових документів. Основним методом формування семантичних полів є експертний метод лексикографічного аналізу. У такому методі неможливо сформувати структуру семантичних полів так, щоб вони були не зв'язані між собою і не корелювали у статистичних розподілах алгоритмів аналізу текстових даних. Однак, можна припустити, що внаслідок лінійної комбінації частотних характеристик семантичних полів можна утворити нові семантичні поля, частотні характеристики яких будуть не корельовані. Такі поля назвемо некорельованими вторинними

семантичними полями. Утворення нових некорельованих вторинних семантичних полів оптимізує задачі аналізу текстових даних та зменшує розмірність семантичного простору текстових документів. Задача зводиться до представлення текстових документів у новому семантичному ортонормованому базисі. Проаналізуємо коваріаційну матрицю для частотних характеристик семантичних полів. Проаналізуємо зв'язок між методом головних компонент та сингулярним розкладом у задачі формування ортонормованого семантичного простору. Проаналізуємо сингулярні числа для тестового масиву текстових документів. Розглянемо представлення текстових документів у новому семантичному ортонормованому базисі, у якому коефіцієнти коваріації між різними семантичними частотними складовими текстових документів будуть дорівнювати нулю [274]. Тобто, задача зводиться до реалізації перетворення до нового базису, який буде описуватися діагональною коваріаційною матрицею. Такий базис може бути утворений за допомогою перетворення Карунена-Лоева, яке лежить в основі методу головних компонент [275, 276, 166]. Розглянемо це перетворення для просторового базису утвореного частотними характеристиками семантичних полів. Коваріаційну матрицю розглянемо у вигляді

$$\begin{aligned} &Cov_s = [cov_{ij}^s], \\ &cov_{ij}^s = cov(p_i^{sd}, p_j^{sd}) = \mathbb{E}[((p_i^{sd} - \mathbb{E}(p_i^{sd}))((p_j^{sd} - \mathbb{E}(p_j^{sd}))], \end{aligned} \tag{3.65}$$

де $\mathbb{E}$ – математичне сподівання. Враховуючи вибірку текстових документів, запишемо

$$cov_{ij}^s = \frac{1}{|D|-1}\sum_{i=1}^{|D|}(p_{il}^{sd} - \overline{p}_l^{sd})(p_{jl}^{sd} - \overline{p}_l^{sd}), \overline{p}_l^{sd} = E(p_l^{sd}). \tag{3.66}$$

Знайдемо множину вторинних семантичних полів

$$S' = \{s'_k \mid k = 1, 2, ...|S'|\}, \tag{3.67}$$

які описують текстові документи d_j за допомогою нових частотних векторів

$$V_j'^s = \left(p_{1j}'^{sd}, p_{2j}'^{sd}, ..., p_{|S'|j}'^{sd}\right). \tag{3.68}$$

Для складових частотних векторів $V_j'^s$ має виконуватись умова

$$cov(p_{ik}'^{sd}, p_{jk}'^{sd}), i \neq j. \tag{3.69}$$

Знайти семантичні вектори, для яких виконується умова (3.69), можна за допомогою методу головних компонент [275, 276, 166]. Розглянемо основні положення цього методу для випадку семантичного простору текстових документів. Нехай є відомою матриця базисних частотних семантичних векторів A_s, яка описує зв'язок між векторами первинних та вторинних семантичних полів. Будемо вважати цю матрицю ортогональною, для якої виконується умова

$$A_s^{-1} = A_s^T. \tag{3.70}$$

Тоді вектори первинних та вторинних семантичних полів зв'язані такими співвідношеннями

$$V_j = A_s V_j', V_j' = A_s^T V_j. \tag{3.71}$$

Складові векторів V_j' називають головними компонентами. Для матриць M_{sd} (3.17) можна записати аналогічні співвідношення

$$M_{sd} = A_s M_{sd}', M_{sd}' = A_s^T M_{sd}. \tag{3.72}$$

Здійснимо центрування семантичних векторів та матриць

$$\begin{aligned} \dot{V}_j &= V_j - \mathbb{E}\left(V_j\right), \\ \dot{M}_{sd} &= M_{sd} - \mathbb{E}((M_{sd}). \end{aligned} \tag{3.73}$$

Розглянемо таку коваріаційну матрицю

$$Cov_s' = \dot{M}_{sd}' \left(\dot{M}_{sd}'\right)^T \tag{3.74}$$

Враховуючи (3.72), отримаємо

$$\begin{aligned} Cov'_s &= A_s^T \dot{M}_{sd} \left(\dot{M}_{sd}\right)^T, \\ A_s &= A_s^T Cov_s A_s, \\ Cov_s &= \dot{M}_{sd} \left(\dot{M}_{sd}\right)^T. \end{aligned} \tag{3.75}$$

Нехай матриця A складається із власних векторів матриці Cov_s, тоді Cov'_s буде діагональною матрицею із власними значеннями матриці Cov_s.

$$Cov'_s = diag(\lambda_1, \lambda_2, ..., \lambda_{|S'|}), \tag{3.76}$$

де $\lambda_1, \lambda_2, ..., \lambda_{|S'|}$ – власні значення матриці Cov_s у порядку спадання їх величин. Задача знаходження матриці A_s, яка описує зв'язок між векторами первинних та вторинних семантичних полів зводиться до знаходження власних векторів та значень коваріаційної матриці Cov_s первинних семантичних полів. Визначивши матрицю A_s, частотні семантичні вектори V_j можна розкласти по частотних векторах V'_j вторинних семантичних полів. Характерною властивістю базисних векторів вторинного семантичного простору є їх ортонормованість. Якщо множину вторинних семантичних полів впорядкувати за величиною власних чисел базисних векторів, тоді можна відкинути крайні в цьому ряді вторинні поля як несуттєві для аналізу. У результаті отримаємо

$$p_{ij}^{sd} = \sum_{l=1}^{|\tilde{S}'|} a_{il} p_{lj}'^{sd}, |\tilde{S}'| < |S'|. \tag{3.77}$$

Тобто, для подальшого аналізу беруть підпростір простору вторинних семантичних полів. Складові семантичних векторів $\tilde{p}_{ij}^{sd}$ є проекціями складових $p_{ij}'^{sd}$ на цей підпростір. Якщо базисні ортонормовані вектори розмістити у порядку спадання власних значень коваріаційної матриці, тоді

оцінити похибку такої апроксимації можна за формулою

$$\varepsilon = \sum_{i=|\tilde{S}'|}^{|S'|} \lambda_j. \tag{3.78}$$

Тобто, похибка визначається сумою власних значень базисних векторів, які не вносять вклад у апроксимацію. Звідси випливає, що для зменшення похибки при апроксимації необхідно взяти базисні вектори, для яких власні значення є максимальними. Виникає питання, яка розмірність простору вторинних полів є достатньою для векторного представлення текстових документів. Одним із простих методів відбору головних компонент є правило Кайзера, згідно з яким залишають ті компоненти, для яких виконується умова

$$\lambda_j > \frac{1}{|S'|} tr\left(Cov'_s\right). \tag{3.79}$$

Умова (3.79) визначає ті головні компоненти, для яких власне значення коваріаційної матриці є більшим за середнє всіх власних значень. У загальному випадку метод головних компонент можна розглядати як спектральний розклад коваріаційної матриці частотних характеристик семантичних полів. Задачу про спектральний розклад коваріаційної матриці Cov_s можна звести до задачі сингулярного розкладу матриці (singular value decomposition, SVD) *частоти_семантичних_полів-документи* M_{sd}. Сингулярний розклад матриці *ключові_слова-документи* лежить в основі латентно-семантичного аналізу текстів(Latent Semanti Analysis,LSA) [277, 278, 279]. Нехай існує матриця типу *"частоти_семантичних_полів–документи"*, яка описується формулою (3.17). Вектор V_j відображає документ d_j в Ns-мірному просторі текстових документів. Добуток двох векторів $(V_p)^T V_q$ визначає кількісну міру близькості цих векторів у N_s–мірному семантичному просторі текстових документів. Відповідно добуток матриць $(M_{sd})^T M_{sd}$ містить скалярні добутки векторів $(V_p)^T V_q$ всіх документів і відображає їхні кореляції у просторі семантичних векторів. Нехай існує сингулярна декомпозиція

матриці M_{sd}

$$M_{sd} = U_{sd}\Sigma_{sd}Y_{sd}^T. \tag{3.80}$$

Тоді добуток матриць $(M_{sd})^T M_{sd}$ можна розглянути у вигляді

$$(M_{sd})^T M_{sd} = \left(U_{sd}\Sigma_{sd}Y_{sd}^T\right)^T \left(U_{sd}\Sigma_{sd}Y_{sd}^T\right) = Y_{sd}\Sigma_{sd}^T\Sigma_{sd}Y_{sd}^T. \tag{3.81}$$

У відповідності до теорії сингулярного розкладу матриць [277, 278] діагональна матриця Σ_{sd} містить сингулярні числа в порядку їх спадання. Якщо взяти K найбільших сингулярних чисел матриці Σ_{sd} і відповідно K сингулярних векторів матриць U_{sd} і Y_{sd}, то отримаємо K-рангову апроксимацію матриці M_{sd} :

$$(M_{sd})_K = (U_{sd})_K(\Sigma_{sd})_K(Y_{sd})_K^T \tag{3.82}$$

Матриця $(Y_{sd})_K$ відображає зв'язок між векторами документів $\tilde{V}_j$ у новому комбінованому K-мірному семантичному просторі з ортонормованим семантичним базисом. Зв'язок між вектором V_j документу в первинному семантичному просторі та вектором $\tilde{V}_j$ у просторі вторинних семантичних полів можна описати так

$$\begin{aligned} V_j &= (U_{sd})_K(\Sigma_{sd})_K\tilde{V}_j, \\ \tilde{V}_j &= (\Sigma_{sd})_K^{-1}(U_{sd})_K^T V_j \end{aligned} \tag{3.83}$$

Для чисельного аналізу розподілу вторинних семантичних полів на основі SVD компонент у текстових документах ми вибрали такі ж, як і у випадку аналізу семантичних полів два типи текстових документів - текстову вибірку художніх творів англомовної прози та текстові повідомлення груп новин із різних предметних областей. Наведемо результати, отримані для вибірки художніх творів англомовної прози. На рис. 3.18 показано функцію щільності розподілу ймовірностей значень SVD компонент для різних класів документів, а на рис. 3.19 показано розподіл SVD компонент у двовимірному t-SNE просторі. Аналогічні розрахунки для розподілів семантичних полів у групах новин наведено у Додатках.

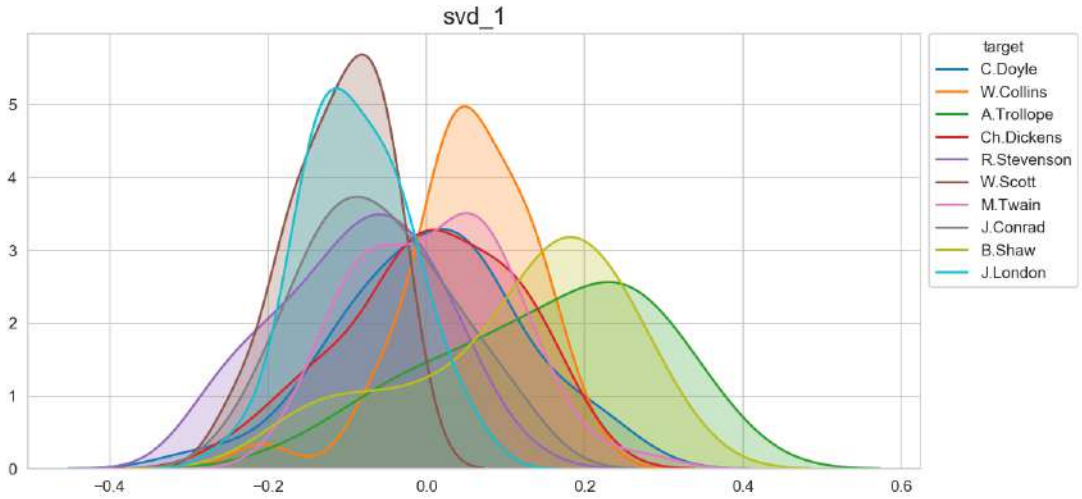

Рисунок 3.18 – Щільність розподілу ймовірностей значень SVD компоненти для різних класів документів у вибірці художніх творів англомовної прози

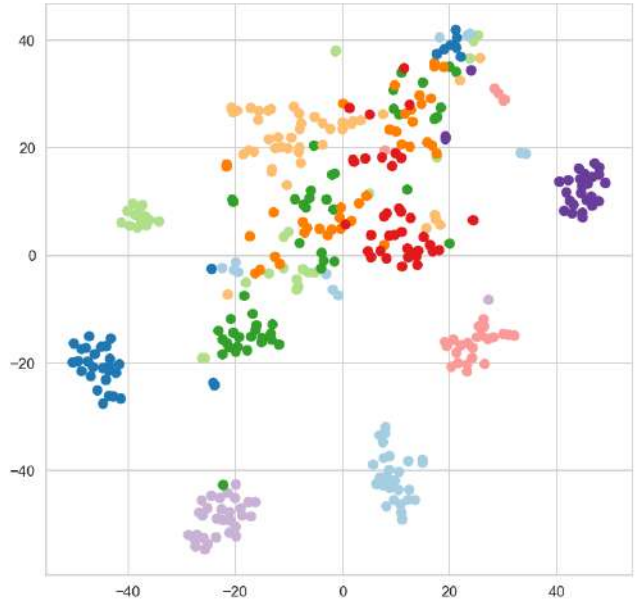

Рисунок 3.19 – Розподіл SVD компонент у двовимірному t-SNE просторі для вибірки художніх творів англомовної прози

Як випливає із отриманих даних, SVD семантичні компоненти володіють класифікаційним потенціалом і доповнюють текстові ознаки на основі семантичних та тематичних полів. Ранг апроксимації матриці M_{sd}, який визначається числом K, також визначає розмірність простору вторинних семантичних полів. Очевидно, що число K, може бути суттєво меншим за розмірність N_s початкового семантичного простору. Це зменшує розмірність задачі аналізу подібності текстових документів у семантичному векторному просторі.

Отже, запропоновано модель некорельованих вторинних семантичних полів, які формуються на основі методу головних компонент шляхом визначення ортонормованого базису семантичного простору, утвореного власними векторами коваріаційної матриці частотних семантичних векторів. Розмірність простору вторинних семантичних полів є суттєво меншою за розмірність простору первинних семантичних полів внаслідок заміни взаємопов'язаних складових некорельованими семантичними характеристиками. Ортонормований базис вторинних семантичних полів може бути також утворений за допомогою сингулярного розкладу матриць *частоти_семантичних_полів-документи.* Аналіз тестової вибірки текстових документів показав різке спадання значень сингулярних чисел. Це дає можливість брати до розгляду лише ті складові вторинних семантичних полів, які описуються першими сингулярними числами. Використання низькорозмірного ортонормованого базису вторинних семантичних полів може бути ефективним у задачах класифікації та кластеризації текстових даних [274, 280].

3.9 Суміш розподілів семантичних полів у текстовій вибірці

Введення простору семантичних полів не тільки зменшує розмірність задачі аналізу тексту, але й вводить новий базис опису текстових документів. Однією з можливих моделей, що пояснюють такий результат, може бути модель суміші нормальних розподілів [281, 282, 283, 284, 285]. Відповідно до цієї моделі розподіл частот розглядається як сума функцій нормальних розподілів семантичних полів із відповідними коефіцієнтами. Кожна така функція описує частотний розподіл семантичних полів у документах даної

категорії. Як категорію документів розглянуто авторство тексту у текстовій вибірці англомовної прози. Деякі розподіли, де нульову гіпотезу тесту про нормальний розподіл відхилено, можна розглядати як суміш нормальних розподілів для авторських підкатегорій за заданим семантичним полем. Враховуючи індивідуальний характер розподілу частот семантичних полів у текстах різних авторських категорій, можна побудувати ймовірнісну модель розподілу авторських стилів у документах текстового масиву. Семантичні поля у такій моделі можуть відігравати роль прихованих параметрів. Таку модель можна представити у вигляді ймовірності розподілу авторського стилю у документах текстового масиву

$$P(Style_j^a, d_j) = \sum_k^{N_s} f_k^s(Style_j^a \mid p_k^s) f_k^s(p_k^s \mid d_i), \tag{3.84}$$

де $f_k^s(p_k^s \mid d_i)$ – розподіл частот семантичних полів у аналізованому документі d_i, $f_k^s(p_k^s \mid d_i) = p_{ki}^{sd}$. Значення $f_k^s(Style_j^a \mid p_k^s)$ можна знайти на основі побудованих функцій розподілу частоти семантичного поля у документах заданої категорії. Семантичні поля, як приховані параметри, відіграють роль стилерозділювальних факторів у класифікаційному аналізі. Розподіли, де нульову гіпотезу про нормальний розподіл було відхилено, можна розглядати як суміш нормальних розподілів для авторських підкатегорій за заданим семантичним полем. Для обчислення параметрів розподілів використано реалізацію алгоритму EM пакету 'mixtools' для середовища R [286]. Розглянемо розподіл семантичних полів у множині текстів одного автора. Негаусовий частотний розподіл семантичних полів можна представити як суміш нормальних розподілів. На рис. 3.20 показано розрахований приклад гістограми та суміш нормального розподілу семантичного поля noun.animal для текстів А. Дойла. Модель суміші пояснює наявність текстових підгруп у аналізованій авторській вибірці текстів. Ці підгрупи визначаються розподілом семантичних полів. У [287] досліджено частотний розподіл семантичних полів іменників та дієслів у текстах англійської художньої літератури. Нуль гіпотеза тесту Шапіро-Вілка про нормальний розподіл частот семантичних полів у масиві текстів відкидається для деяких семантичних полів. Це

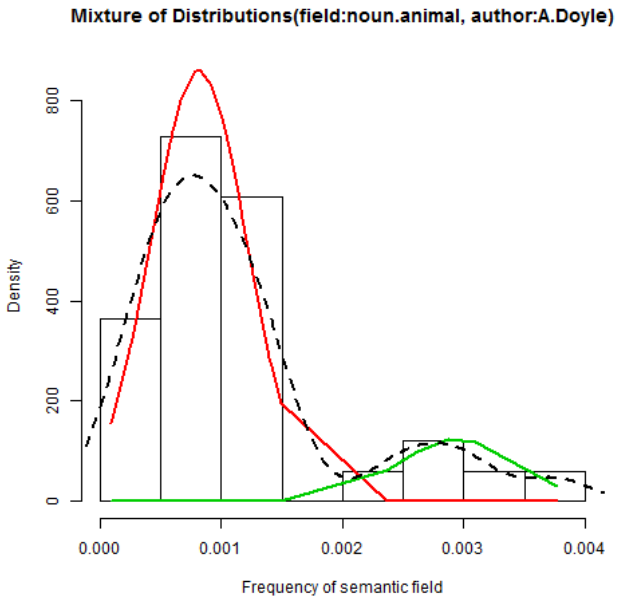

Рисунок 3.20 – Гістограма та суміш нормальних розподілів для вибраного семантичного поля у авторській текстовій вибірці

дає можливість розглядати частотний розподіл таких семантичних полів як категоризовану суміш нормальних розподілів. Як фактор категоризації ми вибрали авторство тексту. Категорії автора з відхиленою гіпотезою нормального розподілу поділено на підкатегорії з нормальним розподілом. Отже, аналіз отриманих результатів показав, що авторський ідіолект відображений у векторному просторі семантичних полів. Такий простір може бути використаний у задачах прогнозної аналітики авторського ідіолекту текстів. Деякі семантичні поля мають високий розділювальний потенціал для диференціювання авторського стилю. Як показують результати, розподіл семантичних полів може розглядатися як додатковий фактор структурного дослідження авторських текстів. Негаусові розподіли частот семантичних полів можна описати на основі моделі суміші категоризованих розподілів частот семантичних полів. Оскільки вибрано класифікацію текстів за категоріями авторів, то в деяких випадках розподіл частот семантичних полів у категоріях може бути негаусовим. У такому випадку авторську категорію текстів можна розділити на додаткові підкатегорії з гаусовим розподілом.

3.10 Розподіли компонент латентного розміщення Діріхле (LDA) в текстових документах

У підрозділі 1.3.2 наведено основні положення латентного розміщення Діріхле (Latent Dirichlet allocation, LDA) [152, 153, 154, 155, 156, 157, 158]. Розглянемо розподіл LDA компонент у вибірці творів англомовної прози, описаної раніше. Для формування LDA компонент ми вибрали 1000 лексем на основі впорядкованих за спаданням частот на основі TF-IDF матриці. Було вибрано 30 LDA тематик, сформованих із використанням пакету *gensim* [288]. Структуру утворених LDA тематик зображено на рис. 3.21. На лівій панелі тематики відображені як круги на двовимірній площині, відстань між ними характеризує їхню тематичну близькість. Методику розрахунку таких відстаней описано в [153]. Площа кругів описує поширеність тематики в аналізованому текстовому масиві. На правій панелі відображені найбільш вживані лексеми вибраної тематики та їх частоти у загальному словнику текстової вибірки та у словнику аналізованої тематики при заданому значенні параметра λ, який входить у формулу (1.56). Приклад розподілу LDA компоненти для різних класів документів наведено на рис. 3.22, розподіл LDA компонент у двовимірному PCA просторі наведено на рис. 3.23, а розподіл LDA компонент у двовимірному t-SNE просторі наведено на рис. 3.19. Використання семантичних ознак на основі LDA компонентів в інтелектуальному аналізі даних розглянуто в [264]. Аналогічні розрахунки для розподілів семантичних полів у групах новин наведено у Додатках.

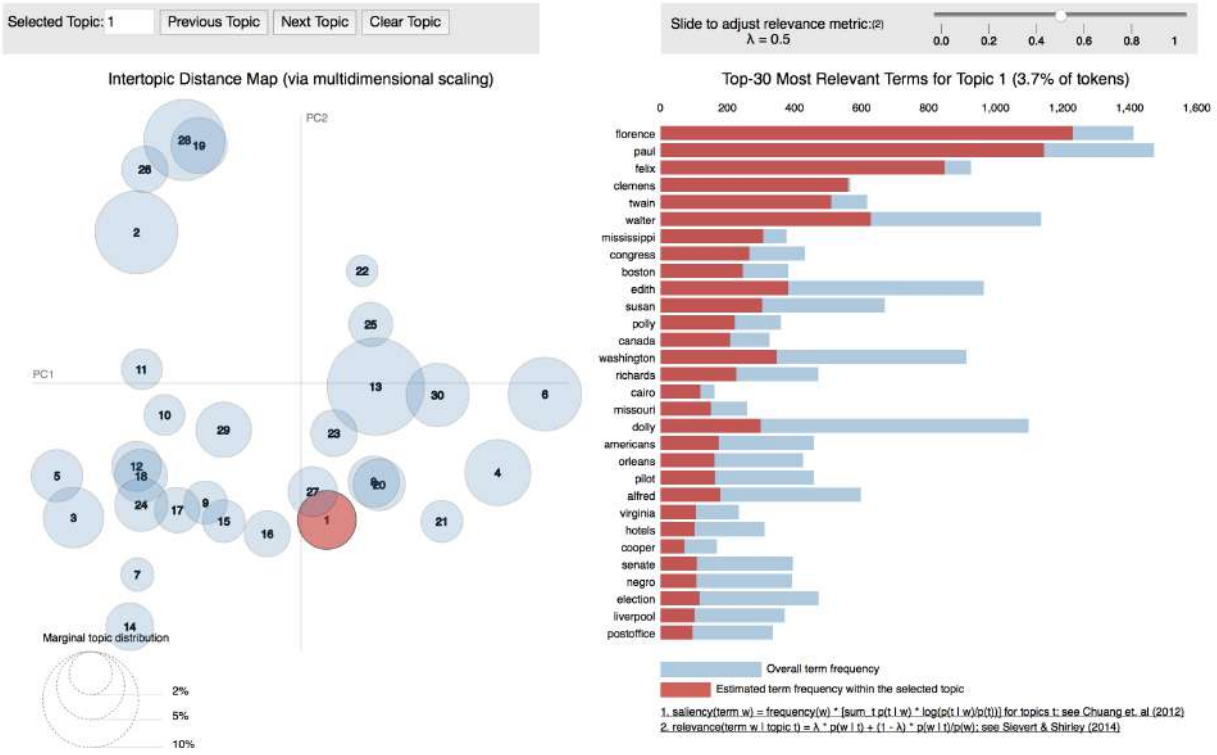

Рисунок 3.21 – Структура утворених LDA тематик у текстовій вибірці художніх творів англомовної прози

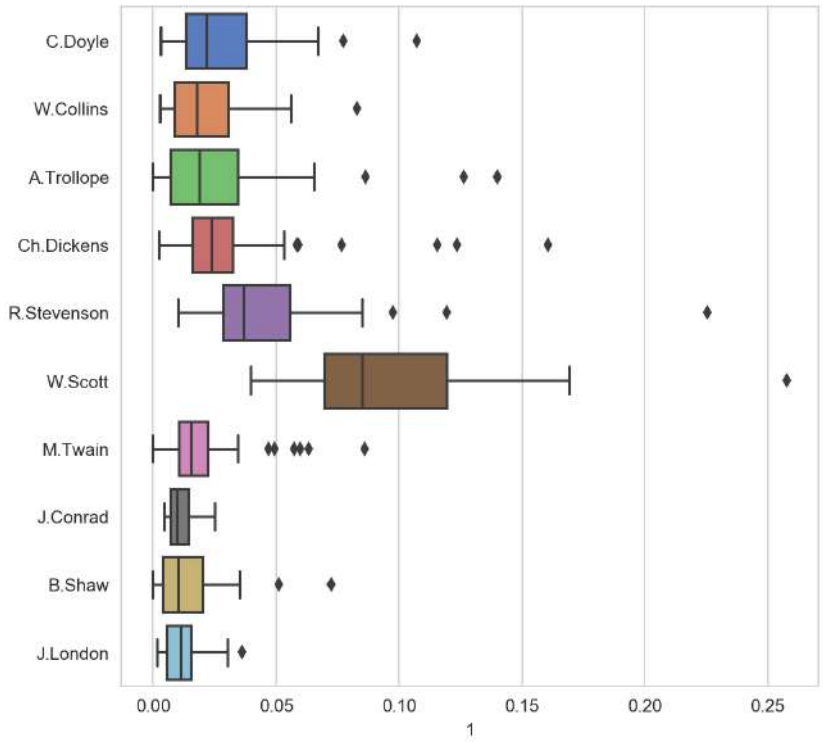

Рисунок 3.22 – Приклад розподілу LDA компоненти для різних класів документів у текстовій вибірці художніх творів англомовної прози

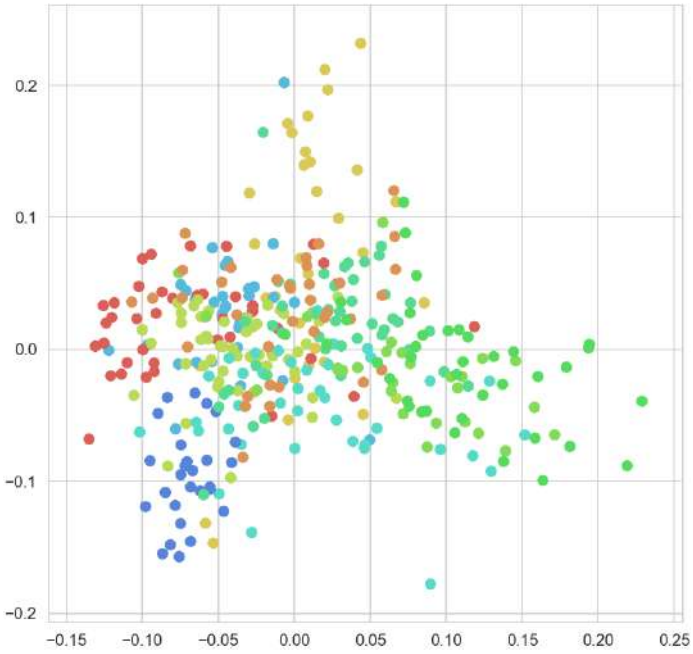

Рисунок 3.23 – Розподіл LDA компонент у двовимірному PCA просторі для текстової вибірки художніх творів англомовної прози

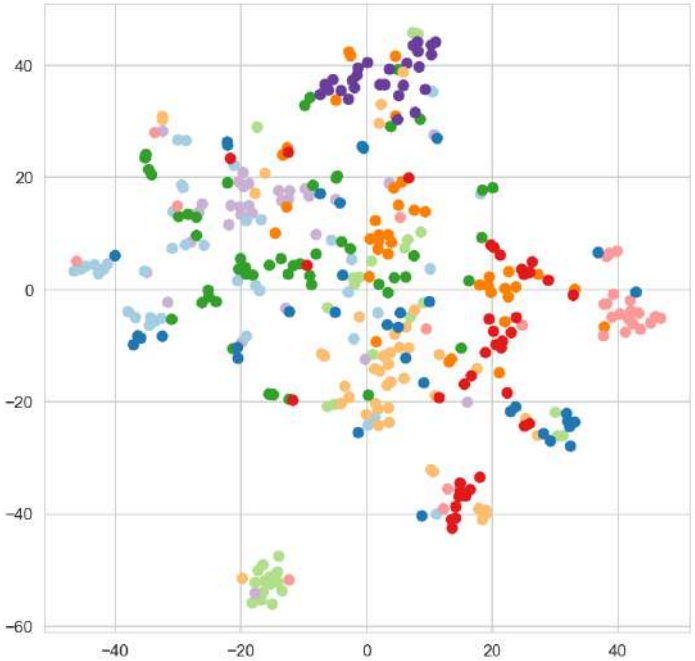

Рисунок 3.24 – Розподіл LDA компонент у двовимірному t-SNE просторі для текстової вибірки художніх творів англомовної прози

3.11 Висновки

– На основі концепцій семантичних полів створено теоретико-множинну модель, яка об'єднує поняття семантичного та тематичного лексемних полів. Лексикографічні семантичні поля та тематичні поля можна розглядати як підкласи об'єднувального класу лексемних полів. Лексикографічні поля утворено на основі експертного семантичного групування лексемного складу словника. Тематичні поля утворено на основі лексем, характерних для тематично категоризованих текстових документів, і визначаються на основі коефіцієнта тематичної виразності. Цей коефіцієнт показує у скільки разів лексеми тематичного поля зустрічаються частіше у текстах заданої тематичної категорії у порівнянні з текстами лінгвостилістичної норми.

– Розглянуто векторну модель текстових документів у семантичному просторі, базис якого утворено частотно-дистрибутивними характеристиками семантичних та тематичних полів. Базис лексикографічних семантичних полів є незалежним від вибірки, а базис тематичних полів є індивідуальним для кожної текстової вибірки.

– На основі теорії нечітких множин створено модель нечіткого семантичного поля лексемного складу текстових масивів. Визначено характеристики для нечіткого семантичного поля - функцію приналежності, найближче звичайне семантичне поле, міру нечіткості семантичного поля, семантичне поле α-рівня. Поряд із поняттям нечіткого семантичного поля уведено поняття семантично нечіткої лексеми, для якої визначено лінгвістичну змінну.

– Запропоновано модель некорельованих вторинних семантичних полів, які формуються на основі методу головних компонент шляхом визначення ортонормованого базису семантичного простору, утвореного власними векторами коваріаційної матриці частотних семантичних векторів. Розмірність простору вторинних семантичних полів є суттєво меншою за розмірність простору первинних семантичних

полів унаслідок заміни взаємопов'язаних складових некорельованими семантичними характеристиками. Ортонормований базис вторинних семантичних полів може бути також утворений за допомогою сингулярного розкладу матриць частоти_семантичних_полів-документи.

– Деякі семантичні поля мають високий розділювальний потенціал для диференціювання авторського стилю. Як показують результати, розподіл семантичних полів може розглядатися як додатковий фактор структурного дослідження авторського стилю.

– Розглянуто латентне розміщення Діріхле, компоненти якого відображають приховані тематики в текстових масивах. Ці компоненти можуть бути використані як додаткові семантичні ознаки текстових документів у задачах інтелектуального аналізу даних.

4 МЕТОДИ МАШИННОГО НАВЧАННЯ ДЛЯ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ТЕКСТОВИХ ДАНИХ ІЗ ВИКОРИСТАННЯМ СЕМАНТИЧНИХ ОЗНАК

Розглянемо особливості реалізації алгоритмів машинного навчання без учителя та з учителем, використовуючи текстові вибірки різних типів. Алгоритми навчання без учителя базуються на основі методів кластеризації, а алгоритми навчання з учителем – на основі методів класифікації з використанням навчальних та тестових вибірок. Для аналізу використаємо вибірку авторських текстів, яку сформовано на основі масиву текстів художньої прози, згрупованих за авторами та стандартизовану вибірку повідомлень груп новин.

4.1 Кластерний аналіз текстових документів

4.1.1 Кластеризація текстових документів у просторі семантичних та тематичних полів

Розвиток методів аналізу текстових документів є перспективним напрямком сучасних інформаційних технологій. Одним із таких методів є кластерний аналіз, який дає можливість виявити групи об'єктів, які подібні між собою за певними критеріями [159, 160, 162, 163, 164, 165, 166]. Кластерний аналіз є ефективним при вивченні структури текстових масивів [151, 162, 163]. Для представлення текстових документів часто використовують модель векторного простору [151]. У цій моделі кожний документ відображається як вектор у багатовимірному просторі, кожний вимір якого відповідає квантитативній характеристиці лексеми зі словників аналізованих текстових масивів. Текстовий масив можна представити у вигляді матриці слів та документів, у якій колонки визначають документи, а рядки – частоти лексем у цих документах. Тоді кожна колонка є вектором частот лексем для заданого документа, який задається номером колонки. Мірою відстані між двома документами може бути кут між векторами цих документів в утвореному векторному просторі. Такий підхід має також низку проблем, зокрема, розмірність аналізованого простору є великою,

оскільки вона зумовлена розміром словника. Документи також можуть бути квантитативно близькими не тільки за частотами окремих лексем, а також за характеристиками заданих лексемних об'єднань, наприклад, семантичних полів. Пошук комплексних характеристик текстових документів є важливим, зокрема, при аналізі авторства текстів, оскільки лексемний частотний спектр творів може бути подібним, але відрізнятися за характеристиками комбінованих лексемних груп. Також часто в алгоритмах інтелектуального аналізу використовують представлення текстових документів у якій кожна колонка представляє знаку, а рядок – окремий документ. Розмірність матриці ознак *"семантичні_поля–документи"* є суттєво меншою у порівнянні з матрицею ознак для лексем словника текстових масивів. Семантичні поля формуються на основі експертного аналізу, одні і ті ж лексеми можуть одночасно належати до різних семантичних полів. Перспективним, на нашу думку, є аналіз відображення у семантичній кластерній структурі документів класифікаційної структури документів за низкою ознак, зокрема, за авторством текстів. На основі даної моделі реалізуємо кластеризацію вибірки текстових документів [289, 290, 291]. Проведемо аналіз отриманих результатів з точки зору ефективності кластеризації у просторі семантичних полів для тестової вибірки документів. Частоти семантичних полів утворюють семантичний векторний простір, у якому кожний документ може бути представленим за допомогою вектора V_j^s, який описується виразом (3.18). Очевидно, що розмірність семантичного векторного простору є суттєво меншою за розмірність простору, утвореного лексемами частотного словника, оскільки $N_s < N_w$. Крім того, семантичний простір відображає додаткову комбіновану складову характеристику текстових документів, яка може утворювати іншу кластерну структуру у порівнянні зі структурою в лексемному просторі. Розглянемо групування документів за семантичними ознаками за допомогою алгоритму ієрархічної кластеризації. Ми обмежили кластерну структуру десятьма кластерами. Для аналізу ефективності розглянутих алгоритмів кластеризації взято текстову вибірку англомовної прози із 368 творів десяти відомих авторів, яку було завантажено з інтернет-ресурсу *Project Gutenberg* [265]. Для утворення семантичного простору сформовано 41 семантичне поле. Деталізація

літературних та лексикографічних характеристик вхідних даних не є суттєвою для аналізу можливості кластерного структурування даних, тому для подальшого аналізу будемо розглядати лише статистичні характеристики текстових документів. Для кожного документа було розраховано частотні словники, на основі яких розраховано частотні спектри семантичних полів документів. Отже, кожний документ розглядається як вектор в n-мірному семантичному просторі. Для кластеризації текстових документів у просторі семантичних полів використано пакет *scikit-learn* [217] для мови *Python*. Розглянемо результати чисельного моделювання кластерної структури масиву авторських текстів у просторах семантичних полів різних типів. На рис. 4.1 показано розподіл кількості документів за кластерами в алгоритмі агломеративної кластеризації у просторі семантичних полів. На рис. 4.2 показано дендрограму агломеративної кластеризації у просторі семантичних полів. На рис. 4.3 показано розподіл авторських документів за кластерами в алгоритмі агломеративної кластеризації у просторі семантичних полів. На рис. 4.4 показано розподіл документів у кластерах за авторами в алгоритмі агломеративної кластеризації у просторі семантичних полів.

На рис. 4.5 показано розподіл кількості документів за кластерами в алгоритмі агломеративної кластеризації у просторі тематичних полів. На рис. 4.6 показано дендрограму агломеративної кластеризації у просторі тематичних полів. На рис. 4.7 показано розподіл авторських документів за кластерами в алгоритмі агломеративної кластеризації у просторі тематичних полів. На рис. 4.8 показано розподіл документів у кластерах за авторами в алгоритмі агломеративної кластеризації у просторі тематичних полів. На прикладі вибірки авторських текстів англомовної прози реалізуємо ітераційну кластеризацію текстів методом k–середніх і проаналізуємо утворену в семантичному просторі кластерну структуру. Розглянемо використання алгоритмів кластеризації документів методом k-середніх у просторі семантичних полів. Для апробації семантичної кластеризації методом к-середніх виберемо текстовий масив та параметри семантичних полів такі самі, як у випадку ієрархічної кластеризаії, описаної вище.

Кожний документ розглянемо як вектор у семантичному просторі. Для кластеризації авторських текстів методом k-середніх у векторному

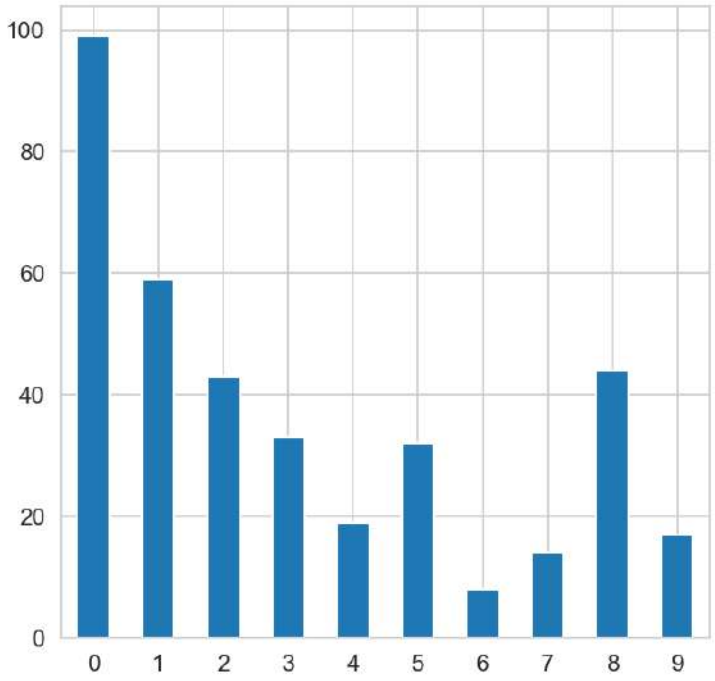

Рисунок 4.1 – Розподіл кількості документів за кластерами в алгоритмі агломеративної кластеризації у просторі семантичних полів

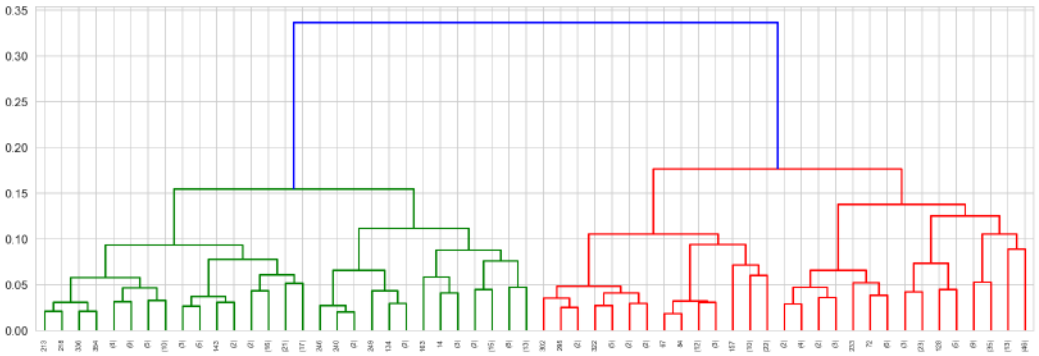

Рисунок 4.2 – Дендрограма агломеративної кластеризації у просторі семантичних полів

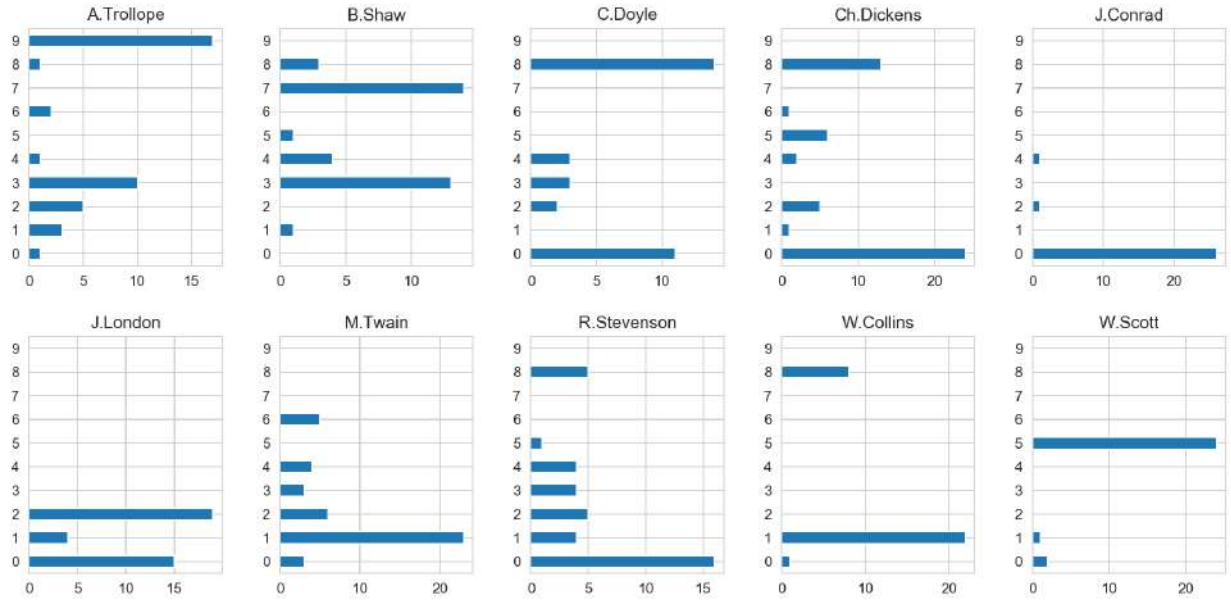

Рисунок 4.3 – Розподіл авторських документів за кластерами в алгоритмі агломеративної кластеризації у просторі семантичних полів

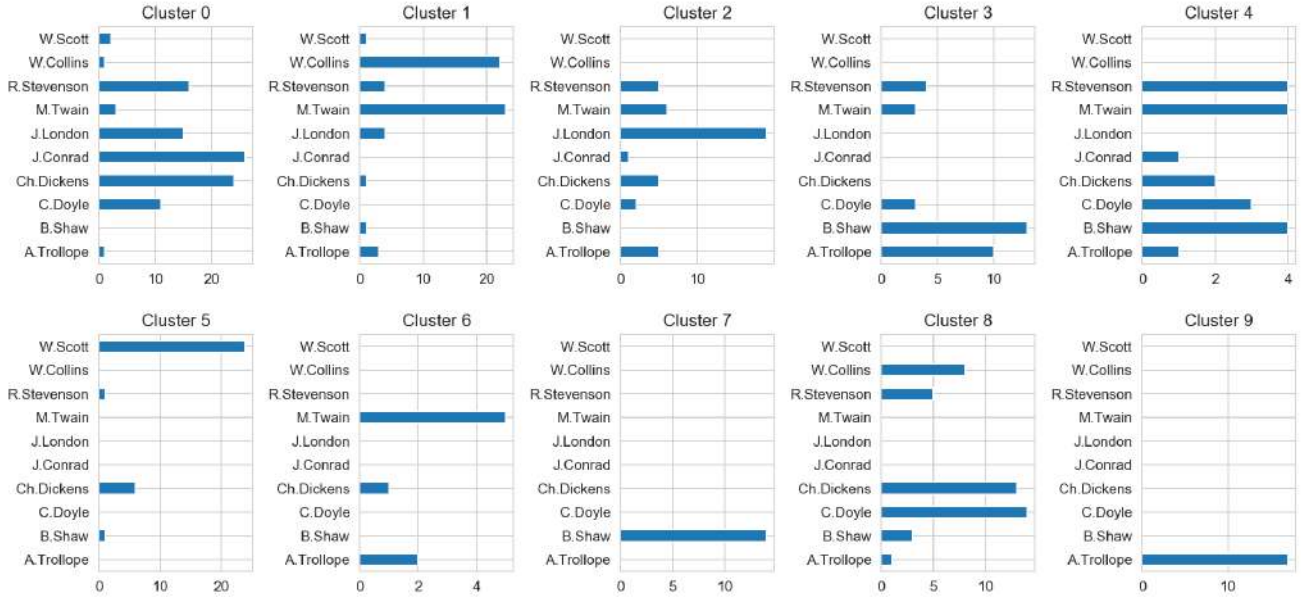

Рисунок 4.4 – Розподіл документів у кластерах за авторами в алгоритмі агломеративної кластеризації у просторі семантичних полів

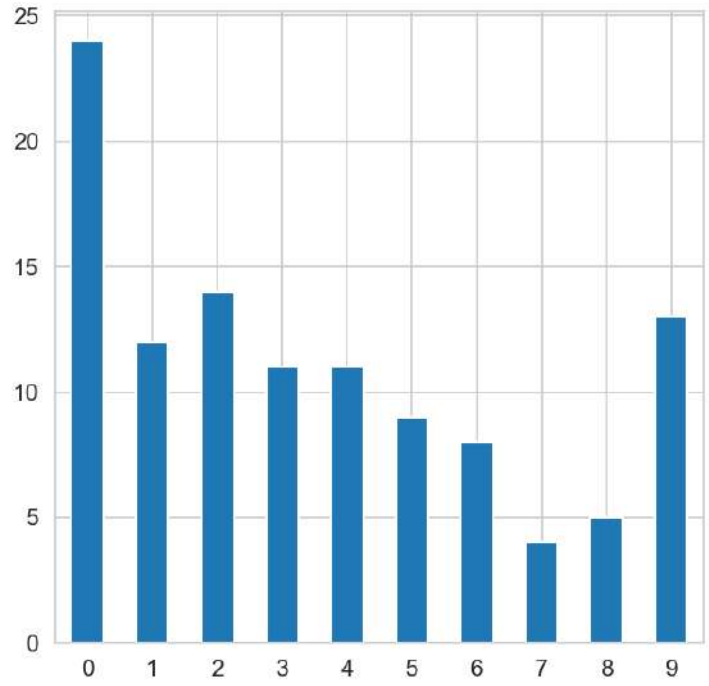

Рисунок 4.5 – Розподіл кількості документів по кластерах в алгоритмі агломеративної кластеризації у просторі тематичних полів

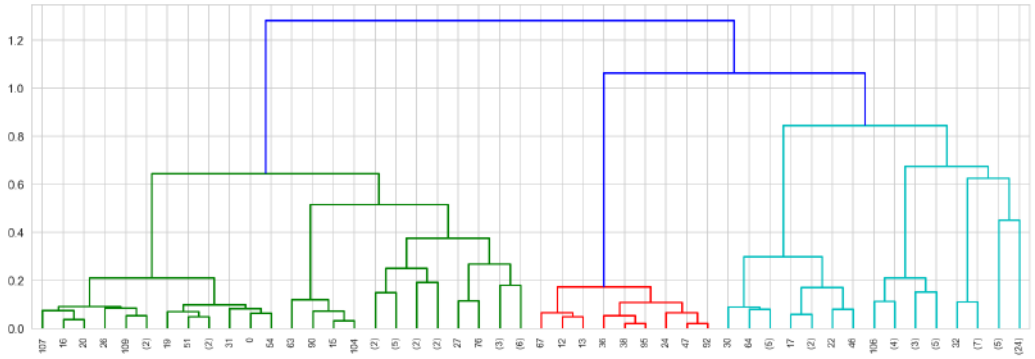

Рисунок 4.6 – Дендрограма агломеративної кластеризації у просторі тематичних полів

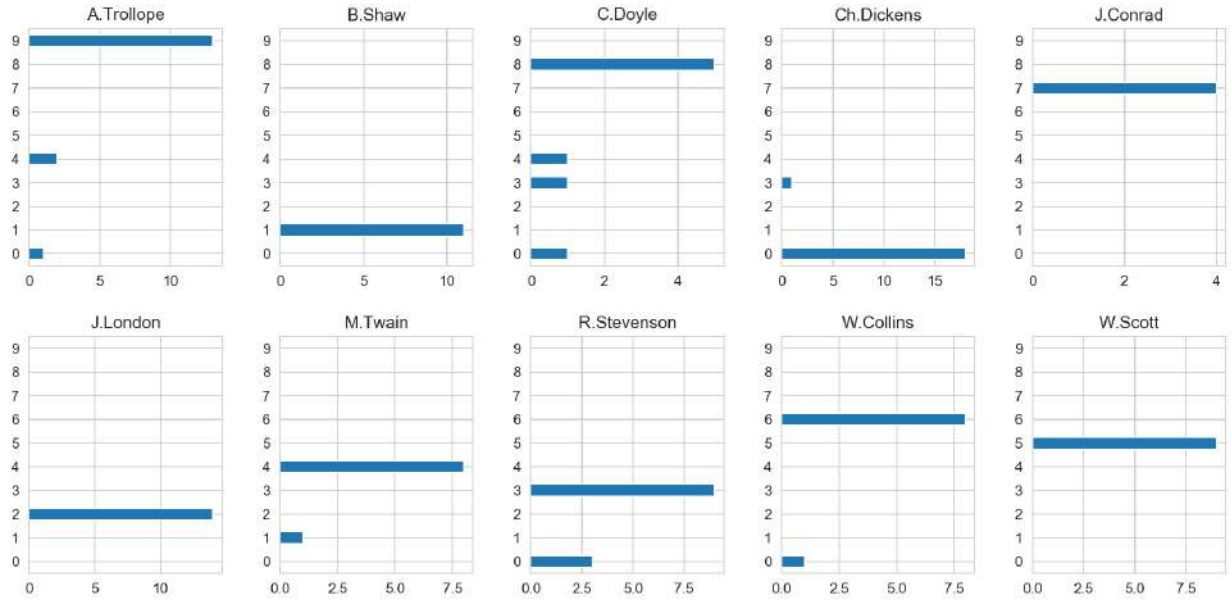

Рисунок 4.7 – Розподіл авторських документів по кластерах в алгоритмі агломеративної кластеризації у просторі тематичних полів

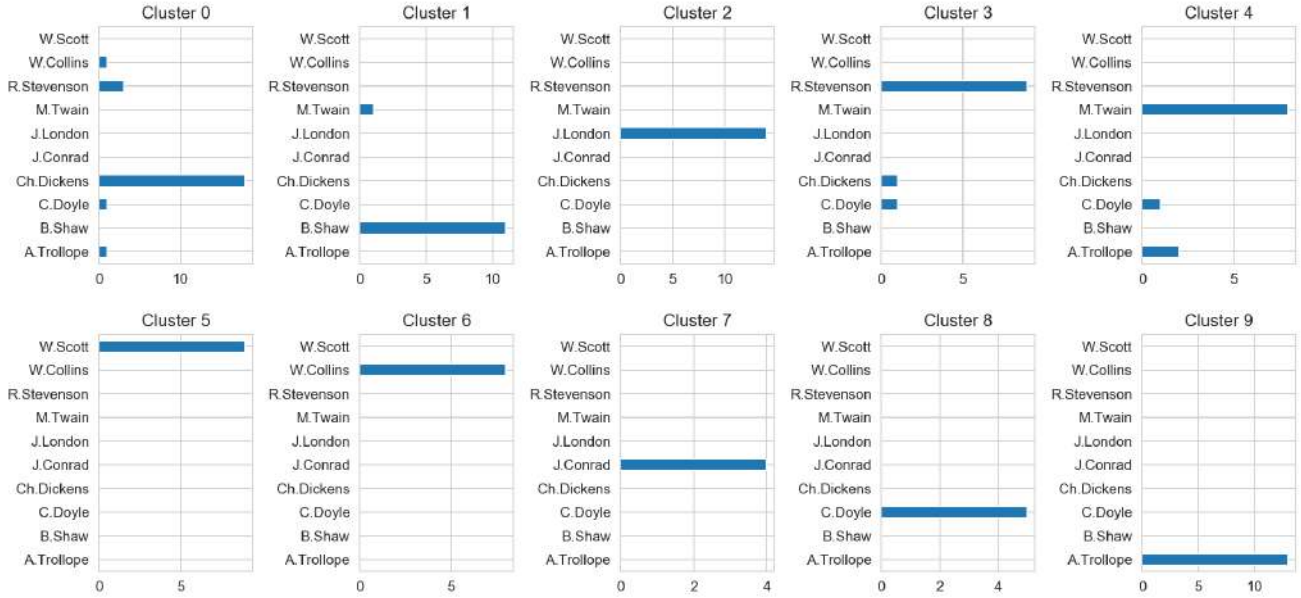

Рисунок 4.8 – Розподіл документів у кластерах по авторах в алгоритмі агломеративної кластеризації у просторі тематичних полів

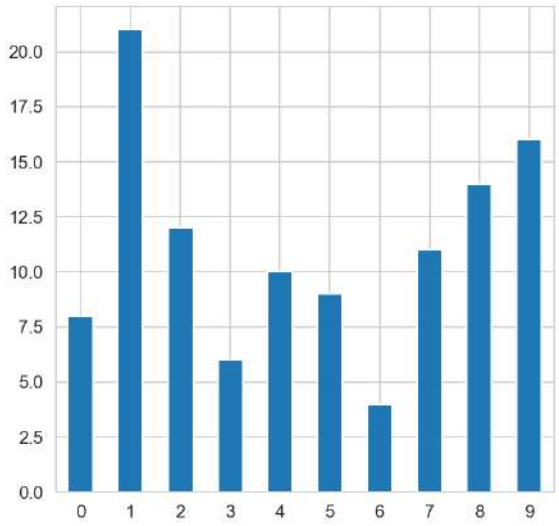

Рисунок 4.9 – Розподіл кількості документів за кластерами в алгоритмі k-means у просторі тематичних полів

семантичному просторі вибрано 10 центрів кластеризації як випадкові точки у семантичному просторі. У результаті реалізації алгоритму кластеризації отримано розподіл текстових документів за десятьма кластерами у семантичному просторі. Розглянемо результати чисельного кластерного аналізу авторських текстів у просторі тематичних полів за допомогою алгоритму k-means. На рис. 4.9 показано розподіл кількості документів за кластерами в алгоритмі k-means у просторі тематичних полів. На рис. 4.10 показано розподіл авторських документів за кластерами в алгоритмі k-means у просторі тематичних полів. На рис. 4.11 наведено розподіл документів у кластерах за авторами в алгоритмі k-means у просторі тематичних полів. Результати чисельного кластерного аналізу авторських текстів та текстів груп новин у просторі семантичних полів за допомогою алгоритму k-means наведено у додатках. Як випливає із наведених даних, деякі кластери містять тексти широкого семантичного спектру. Очевидно, що область цих кластерів у семантичному просторі є семантично однорідною і має низький семантично-диференціювальний потенціал. Однак, також спостерігаються кластери, у яких домінуюче положення займають тексти одного чи декількох авторів. Такі кластери характеризують авторський ідіолект окремих авторів. Семантичні просторові області

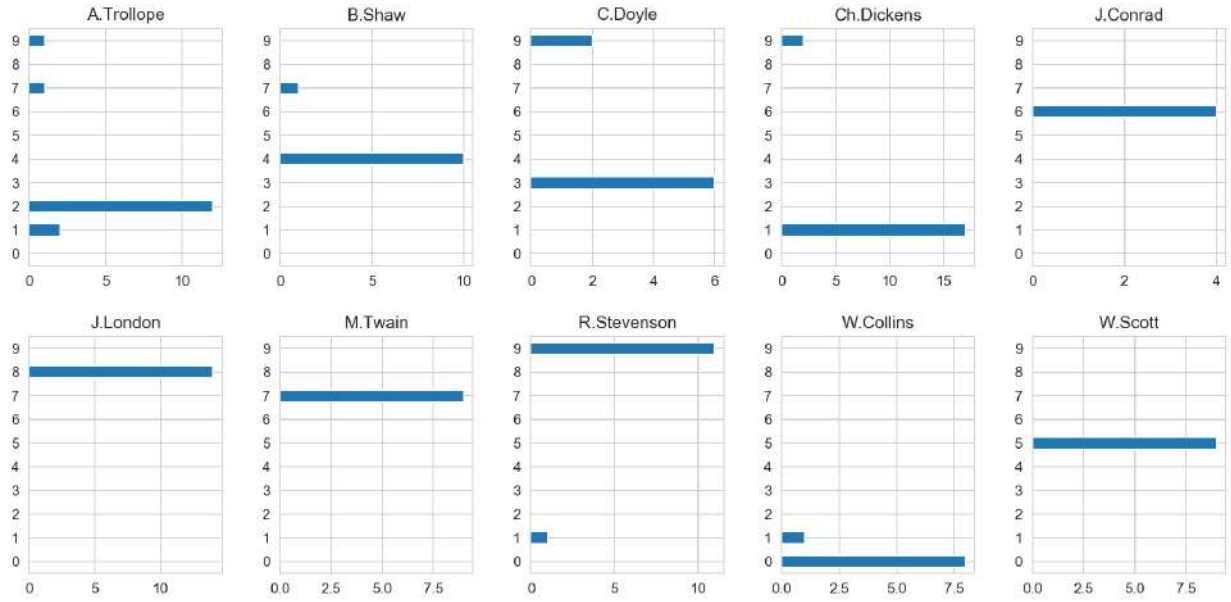

Рисунок 4.10 – Розподіл авторських документів за кластерами в алгоритмі k-means у просторі тематичних полів

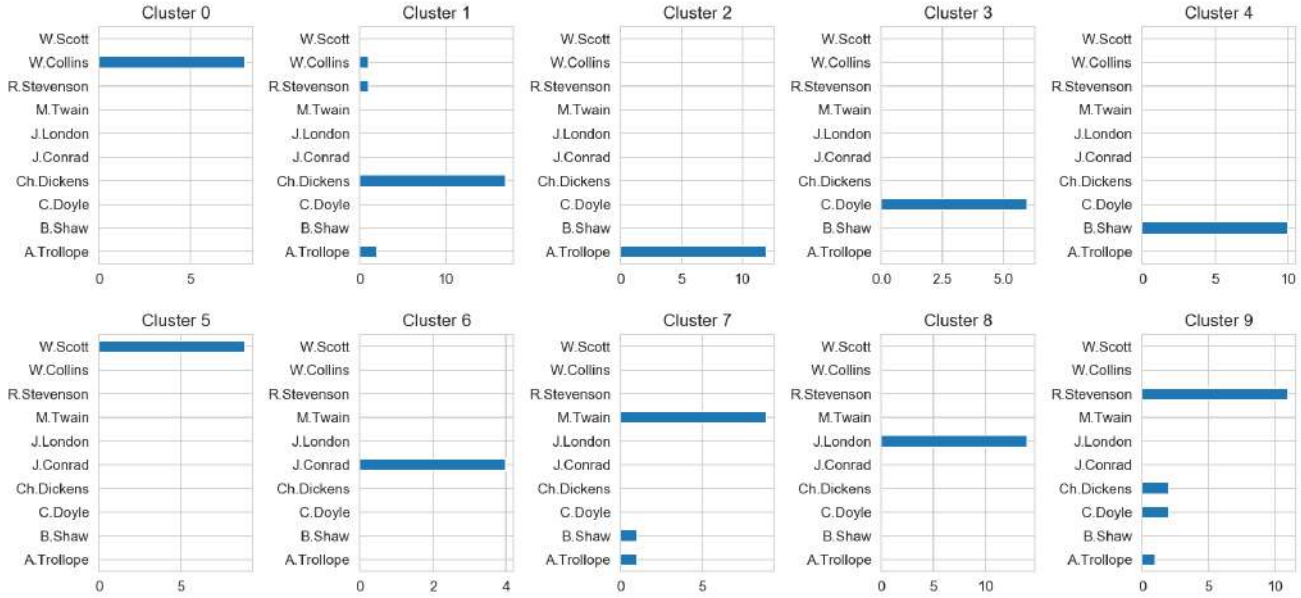

Рисунок 4.11 – Розподіл документів у кластерах за авторами в алгоритмі k-means у просторі тематичних полів

цих кластерів володіють диференціювальним потенціалом для авторських ідіолектів і можуть бути використані в аналізі авторських текстів як додатковий фактор аналізу авторського лексикону. Області семантичного простору, що відповідають кластерам, у яких домінують два або декілька авторів, можна розглядати як області семантичної спорідненості цих авторів. Такі кластери можна розглядати як області семантичного простору, які можуть бути використані для диференціювання авторського ідіолекту у завданнях аналізу авторського стилю та авторських текстів. Формування простору семантичних полів дає можливість отримувати новий структурний поділ документів за семантичними характеристиками. Кластеризація документів у такому просторі методом k–середніх відображає класифікаційну структуру документів за різними ознаками, зокрема, за авторством текстів. Кількість центрів кластеризації є вхідним параметром у методі k–середніх і може бути вибрана експериментальним шляхом, виходячи із наявності в кластерній структурі домінуючих кластерів для документів із спільними класифікаційними ознаками. Як випливає із отриманих результатів, авторські тексти містять індивідуальний стиль авторів, що відображається у кластерній структурі. Тексти деяких авторів домінують в окремих кластерах. Структурованість текстів за авторським ідіолектом спостерігається у просторах семантичних полів різних типів. Найбільш виражена структурованість спостерігається у просторі тематичних полів. Щодо методів кластеризації, то семантична структурованість текстів спостерігається при використанні різних методів кластеризації. Належність текстового документа до певного кластера у різних методах кластеризації у різних семантичних просторах можна розглядати як додаткову ознаку у класифікаційному та структурному аналізі текстових масивів. Як випливає із отриманих даних, для текстів груп новин також характерна структурованість документів у різних семантичних просторах. Проведений кластерний аналіз текстових вибірок різних типів показує, що дослідження текстів різними алгоритмами кластеризації у різних просторах семантичних ознак є ефективним методом структурного аналізу текстів та методом формування семантичних ознак на основі приналежності текстових документів до відповідних кластерів [289, 292, 290, 291, 280,

293, 294]. Запропонована модель кластеризації текстових документів у семантичному просторі дає можливість отримувати новий структурний поділ документів за семантичними ознаками у просторі суттєво меншої розмірності, ніж простір, утворений лексемним складом текстової вибірки. Такий структурний поділ відображає класифікацію документів за новими ознаками документів, зокрема, за авторством текстів. Текстові вибірки деяких авторів можуть мати свої чіткі області в семантичному просторі. Це дає можливість вивчати авторство текстових документів через аналіз приналежності семантичних векторів цих документів до заданих областей простору семантичних полів. Деякі кластери, в яких переважають тексти деяких авторів, є семантично інваріантними і не залежать від зміни базису семантичного простору та методу кластеризації. На основі кластерного аналізу можна диференціювати авторський ідіолект у векторному просторі семантичних ознак.

У Додатках наведено розгляд сингулярної декомпозиції семантичних ознак в алгоритмі ієрархічної кластеризації.

4.1.2 Кластеризація повідомлень груп новин у просторі семантичних ознак

Розглянемо особливості кластеризації текстових документів на прикладі стандартизованої вибірки текстових повідомлень групп новин 20 Newsgroups [268]. Для кластерного аналізу у просторі тематичних полів вибрано коефіцієнт тематичної виразності (3.20), який дорівнює 2. Ми розрахували частотні словники як для окремих документів, так і для масивів повідомлень кожної окремої групи. Для кожної групи було виявлено лексеми, для яких коефіцієнт тематичної виразності був більшим за 2. Ці лексеми утворюють тематичні поля, тематики яких задані кожною групою новин. На основі сформованих тематичних груп, розраховано частоти тематичних полів у кожному документі. Сукупність таких частот є складовими векторного представлення кожного повідомлення у семантичному просторі. Розглянемо результати чисельного кластерного аналізу текстів груп новин у просторі тематичних полів за допомогою алгоритму k-means. На рис. 4.12 показано розподіл кількості документів за кластерами в алгоритмі k-

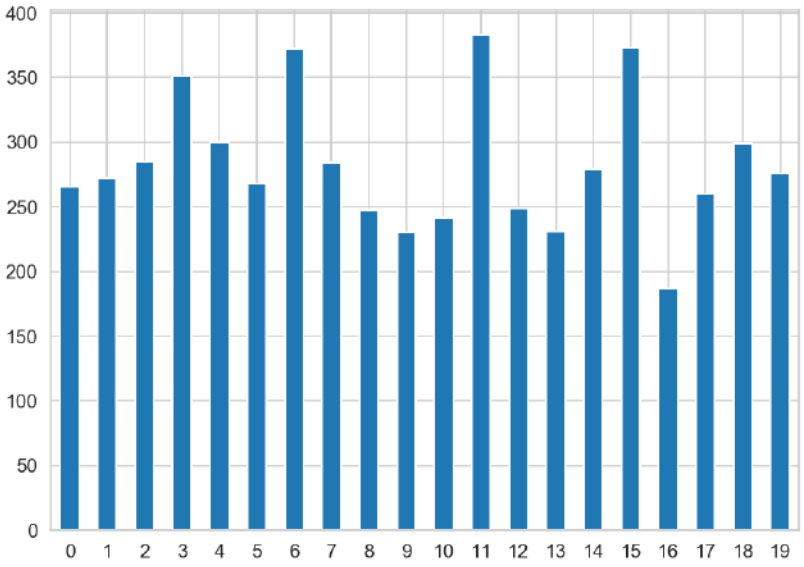

Рисунок 4.12 – Розподіл кількості документів за кластерами в алгоритмі k-means у просторі тематичних полів

means у просторі тематичних полів. На рис. 4.13 показано розподіл текстових документів груп новин за кластерами в алгоритмі k-means у просторі тематичних полів. На рис. 4.14 наведено розподіл документів у кластерах за групами новин в алгоритмі k-means у просторі тематичних полів. Результати чисельного кластерного аналізу текстів груп новин у просторах семантичних полів, SVD та LDA компонент різними методами кластеризації наведено у Додатках. На основі аналізу розподілу груп новин у кластерах можна зробити низку висновків. Кластери, у яких знаходяться повідомлення багатьох груп, характеризують у семантичному просторі області семантично нейтральних повідомлень, у яких відображені повідомлення з рівномірним семантичним розподілом лексем. Кластери, у яких домінують групи новин, характеризують області семантично виразних лексем. Кластери, які містять декілька домінуючих груп, можна розглядати як області семантичних зв'язків між цими групами. Порівнюючи кластерні розподіли у просторі семантичних та тематичних полів, можна виявити, що простір тематичних полів є більш семантично диференціювальним для аналізованого масиву документів у порівняння із простором семантичних полів. Однак, такий простір потребує додаткового формування тематичного

Рисунок 4.13 – Розподіл текстових документів груп новин за кластерами в алгоритмі k-means у просторі тематичних полів

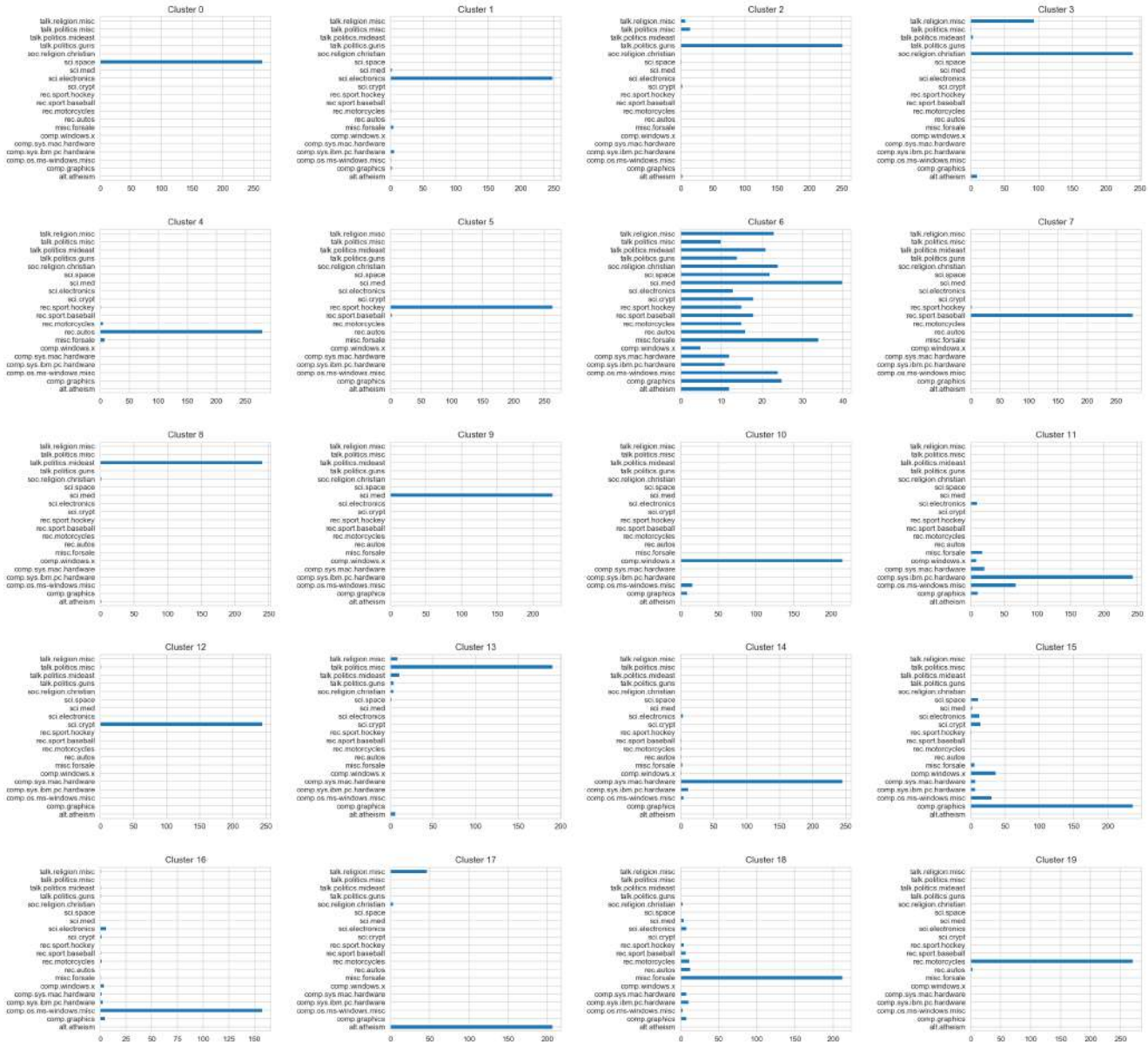

Рисунок 4.14 – Розподіл документів у кластерах за групами новин в алгоритмі k-means у просторі тематичних полів

базису векторного простору, який є ефективним для аналізованого масиву текстових документів. Крім того, формування базису тематичних полів потребує категоризованої вибірки документів, оскільки кожна категорія є основою формування заданого нею тематичного поля. Такий тип кластеризації можна вважати кластеризацією з навчальною вибіркою. Однак, на відміну від класифікаційного аналізу, у якому навчальна вибірка є необхідним елементом, сформовані тематичні поля можуть бути використані в аналізі некатегоризованих за тематичним базисом масивах документів. Таким чином, тематичний базис дає можливість виявити нові групування аналізованих текстів, які не проявлялися в інших базисах. У просторі тематичних полів спостерігається більша кількість кластерів, у яких домінує лише одна група новин. Це пояснюється тим, що один із вимірів тематичного простору утворений тематично виразними лексемами певної групи. Однак, також спостерігаються кластери, у яких одна і та ж група домінує у декількох кластерах. Це свідчить про можливість розбиття цієї групи на підгрупи, у тематику яких можуть входити тематичні складові інших груп.

Отже, використання моделі векторного простору із базисом семантичних ознак є ефективним у алгоритмах кластерного аналізу текстових повідомлень груп новин [291]. Як семантичні ознаки розглянуто частотні характеристики семантичних та тематичних полів. Тематичні групи новин утворюють тематичні поля на основі тематично виразних лексем. Аналіз розподілу груп новин у кластерній структурі показав наявність областей семантичного простору, в яких відображені окремі групи новин, та областей, які відображають семантичні зв'язки між масивами повідомлень окремих груп. Кластерна структура повідомлень у просторі тематичних полів є більш семантично диференційованою у порівнянні із кластерною структурою у просторі семантичних полів. Базис векторного простору на основі семантичних та тематичних полів є універсальним і не потребує експертного підбору ключових слів. Розмірність такого базису є суттєво меншою у порівнянні з методами кластеризації за ключовими словами.

4.2 Класифікація текстових даних у векторному просторі семантичних ознак

Розглянемо класифікацію текстових даних на прикладі текстових вибірок різних типів [295, 296, 297, 264]. Як текстові дані виберемо масив авторських текстів англомовної художньої прози, повідомлення груп новин та короткі повідомлення Твіттера. Для реалізації алгоритмів класифікації вибрано різні алгоритми машинного навчання з учителем, зокрема, алгоритми Random Forest, XGBoost, нейронні мережі прямого поширення. Для вивчення ефекту генералізації класифікаційних алгоритмів текстову вибірку було розділено на тренувальну та тестову вибірки даних. Ознаки на основі текстових частот семантичних полів можуть бути ефективно використані у алгоритмах машинного навчання з учителем, зокрема, у класифікаційних задачах. Використання таких ознак дає можливість суттєво зменшити розмірність аналізованого семантичного простору у порівнянні із використанням TF-IDF матриць. Класифікація текстових даних за ознаками семантичних полів дає можливість провести інтелектуальний аналіз текстів у експертному просторі з відповідними семантичними акцентами, які відображають семантичну сторону предметної області аналізу. Прогнозні моделі класифікаційного аналізу текстових даних за семантичними полями можуть бути складовими прогнозних багаторівневих ансамблів на основі стекінгового підходу. Ефективність використання класифікаторів із ознаками на основі семантичних полів зумовлена використанням семантичних підходів у формуванні прогнозних ознак, які відмінні від ознак в інших складових моделях ансамблю. Така відмінність зумовлює низьку кореляцію похибок класифікації з іншими моделями і, відповідно, зумовлює покращення прогнозних характеристик ансамблю у цілому. Використання ознак на основі експертно сформованих семантичних полів дає можливість будувати стабільні прогнозні моделі з урахуванням можливої зміни розподілу лексем тестових вибірок по відношенню до тренувальних вибірок моделей машинного навчання. Розглянемо класифікацію текстових документів у просторі семантичних полів. У [296] проаналізовано можливість використання наївного баєсівського

класифікатора (NB), а у [297] проаналізовано використання класифікатора за найближчими сусідами (kNN) у класифікаційному семантичному аналізі повідомлень груп новин. Отримані результати свідчать про ефективність реалізації NB та kNN класифікації у просторі семантичних полів і відображають сукупність характеристик розглянутих класифікаторів та текстової вибірки заданого типу повідомлень груп новин.

Авторські тексти художньої літератури відображають авторський ідіолект, який характеризує сукупність авторських засобів вираження, зокрема, характерний семантичний спектр авторського лексикону. Для аналізу було взято вибірку текстів англомовної прози, описану вище. Використання методів квантитативного інтелектуального аналізу дає можливість досліджувати особливості авторського стилю, аналізувати авторство невідомих текстів. Частотну характеристику семантичного поля розраховувалось як суму текстових частот слів, які входять у це поле, розділено на суму текстових частот лексем усіх семантичних полів, які розглядаються. Розрахунки тематичних полів здійснювалися на тренувальній вибірці даних, яка складала 70% від загального обсягу вибірки. Кількість тематичних полів дорівнює кількості класів у вибірці. У кожне тематичне поле входили слова, які зустрічаються у два рази частіше у документах відповідного класу, ніж у загальній вибірці. Кількісні характеристики текстових документів для кожного тематичного поля розраховувались як суми текстових частот слів, які входять у кожне тематичне поле. Такі розрахунки було здійснено для тренувальної та тестової вибірок на основі множин слів тематичних полів, виявлених на тренувальній вибірці. Семантичні ознаки для латентного семантичного аналізу знайдено за допомогою сингулярного розкладу (Singular Value Decomposition, SVD) матриці TF-IDF. Для реалізації SVD розкладу використано алгоритм, описаний у [298] і реалізований у пакеті *scikit-learn* [217]. Розрахунок LDA компонент здійснювався за допомогою пакету *gensim* [288]. Числове моделювання та візуалізація результатів здійснювались на мові *Python* у середовищі *Jupyter Notebook* із використанням відповідних пакетів, зокрема *pandas* [215, 216], *scikit-learn* [217], *numpy* [218], *keras* [219], *matplotlib* [220], *seaborn* [221]. Розглянемо можливість використання векторного простору

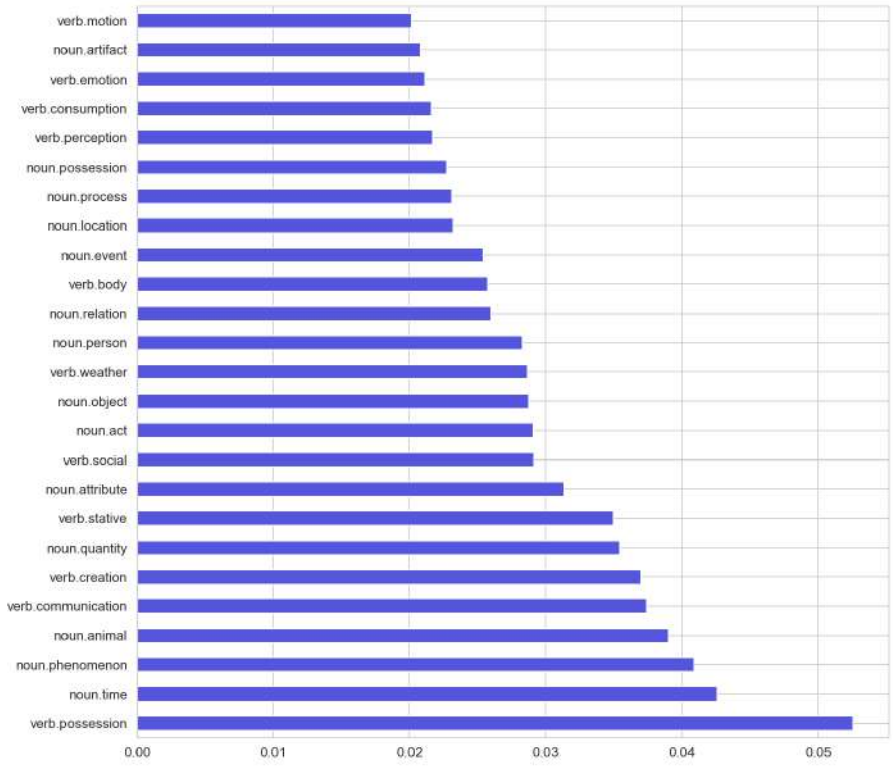

Рисунок 4.15 – Кількісна характеристика важливості семантичних полів

семантичних полів у класифікаційному аналізі авторських текстів художньої літератури [264]. Класифікаційний аналіз авторських текстів здійснювався за допомогою алгоритму *Random Forest* з пакету *scikit-learn* [217]. Для класифікації вибрано параметр кількості ітерацій, який дорівнює 300. На рис. 4.15 наведено кількісну характеристику важливості семантичних полів у класифікаційному аналізі. Для оцінки класифікації були використані такі характеристики як точність (precision), повнота (recall) та f1-оцінка (f1-score). На рис. 4.16, 4.17 наведено оцінки класифікації при використанні ознак на основі семантичних полів. На рис. 4.18 наведено кількісна характеристика важливості тематичних полів у класифікаційному аналізі. На рис. 4.19 наведено оцінки класифікації при використанні ознак на основі тематичних полів. Результати аналогічного класифікаційного аналізу у семантичному просторі SVD та LDA компонент наведено у Додатках. Ми розглянули випадок використання сукупного набору семантичних ознак різних типів за допомогою класифікатора на основі нейронної мережі. Також розглянуто випадок, у якому взято сукупні семантичні ознаки, які складались із семантичних та тематичних полів, складових компонент SVD

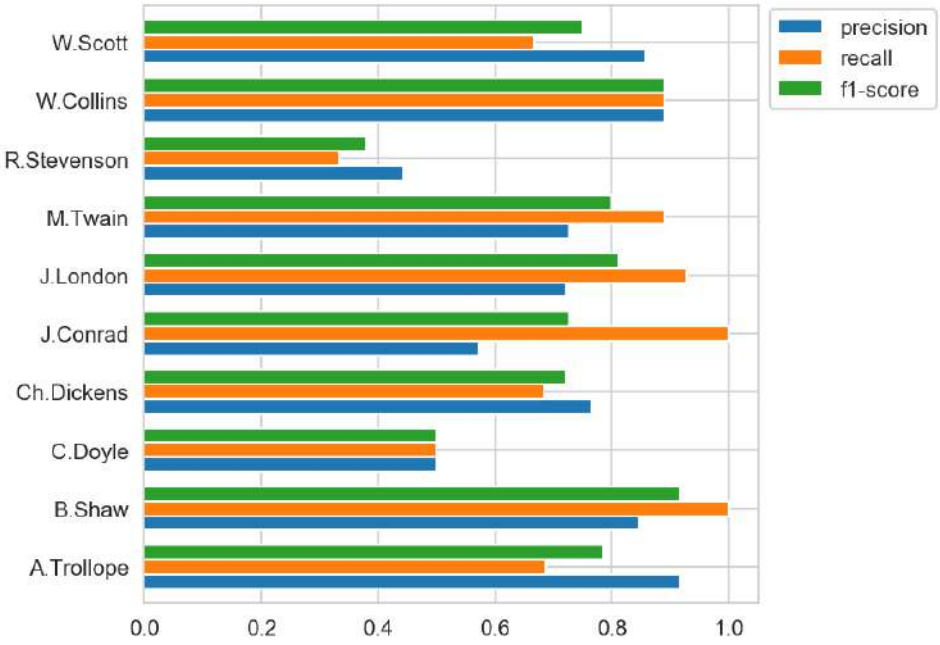

Рисунок 4.16 – Оцінки класифікації при використанні ознак на основі семантичних полів

	precision	recall	f1-score	support
A.Trollope	0.92	0.69	0.79	16
B.Shaw	0.85	1.00	0.92	11
C.Doyle	0.50	0.50	0.50	8
Ch.Dickens	0.76	0.68	0.72	19
J.Conrad	0.57	1.00	0.73	4
J.London	0.72	0.93	0.81	14
M.Twain	0.73	0.89	0.80	9
R.Stevenson	0.44	0.33	0.38	12
W.Collins	0.89	0.89	0.89	9
W.Scott	0.86	0.67	0.75	9
accuracy			0.74	111
macro avg	0.72	0.76	0.73	111
weighted avg	0.74	0.74	0.73	111

Рисунок 4.17 – Оцінки класифікації на валідаційному сеті

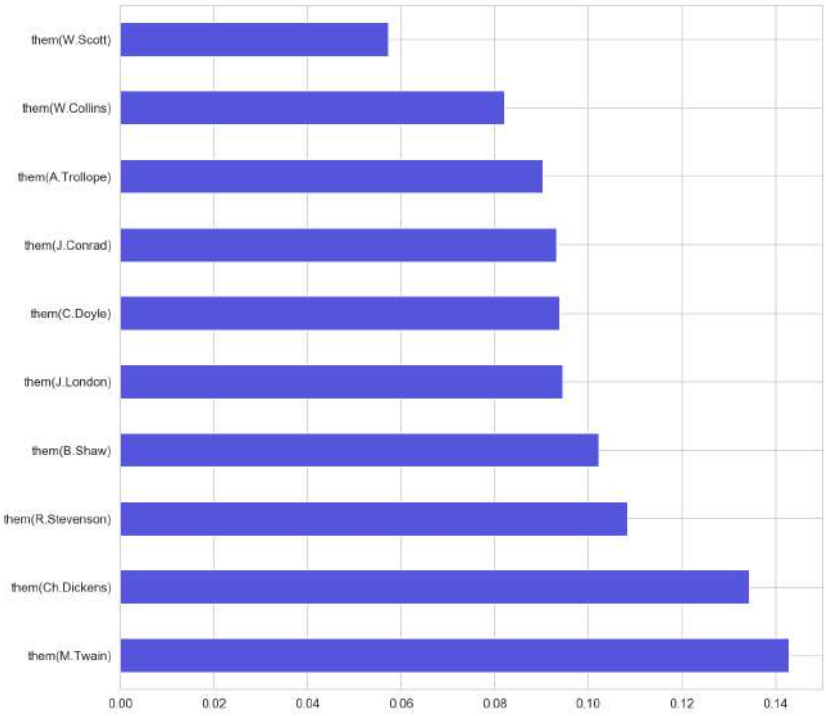

Рисунок 4.18 – Кількісна характеристика важливості тематичних полів

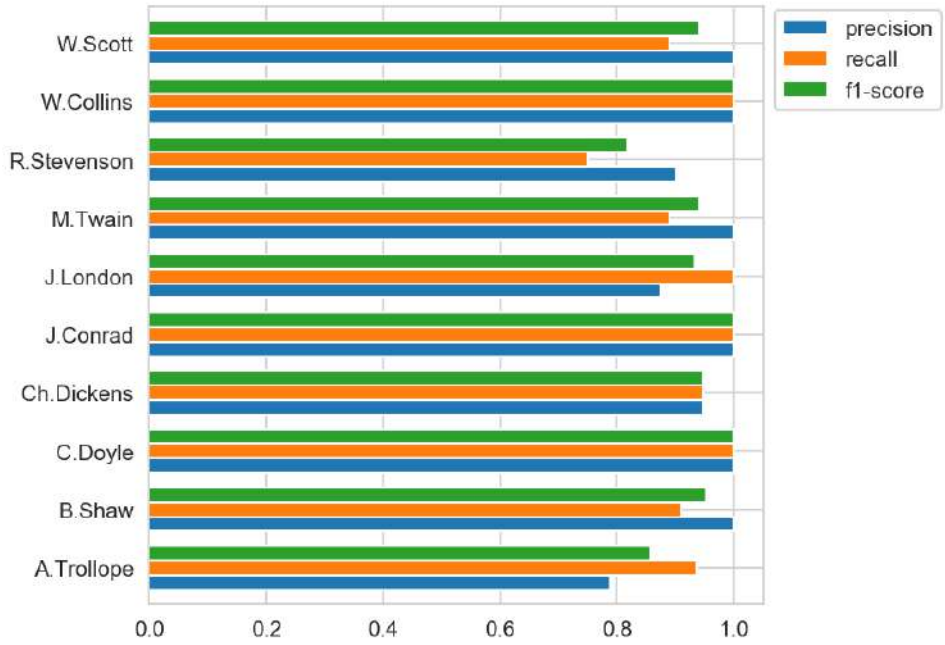

Рисунок 4.19 – Оцінки класифікації при використанні ознак на основі тематичних полів

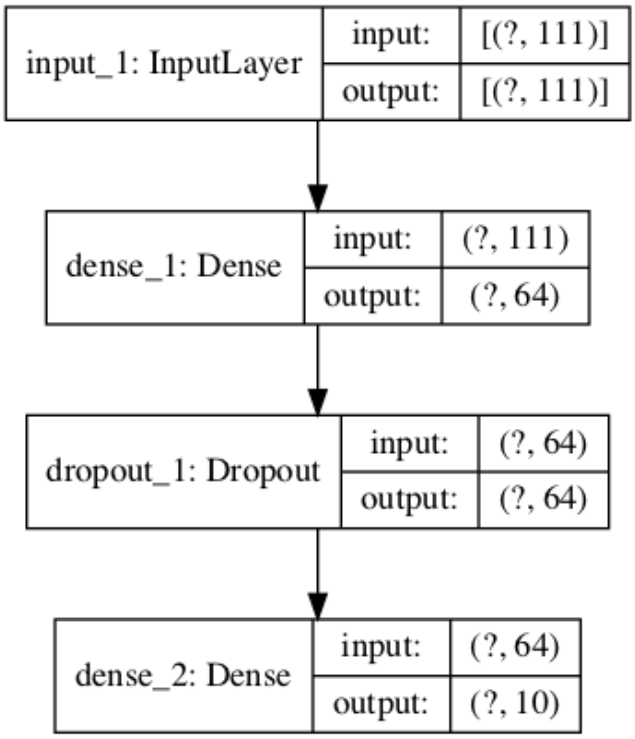

Рисунок 4.20 – Структура нейронної мережі із повністю з'єднаними шарами

розкладу та компонент LDA розміщення. Класифікацію здійснено на основі нейронної мережі із повністю з'єднаними шарами, структуру якої наведено на рис. 4.20. Нейронну мережу реалізовано за допомогою пакету *keras* [219] для мови *Python*. Для зменшення ефекту перенавчання мережі, між шарами було введено *Dropout* шари, які випадковим чином обривають заданий відсоток зв'язків між шарами. На рис. 4.21 наведено динаміку кількісних характеристик втрат та точності для тренувальної та валідаційної вибірок на ітераціях навчання нейронної мережі. На рис. 4.22 наведено величини точності (precision), повноти (recall) та f1-оцінки для класифікації текстових документів нейронною мережею із повністю з'єднаними шарами у випадку сукупних семантичних ознак різних типів.

Також було розглянуто алгоритм із використанням бустингу на деревах рішень XGBoost [73] із такими параметрами: {'colsample_bytree':0.15,'subsample':0.85, 'objective':'multi:softprob', 'n_estimators':30, 'learning_rate':0.01, 'max_depth':5}. Як ознаки класифікації було використано усі розглянуті семантичні ознаки – ознаки на основі семантичних і тематичних полів, компоненти сингулярного розкладу TF-IDF матриці та компоненти латентного розміщення Діріхле. На рис. 4.23 наведено

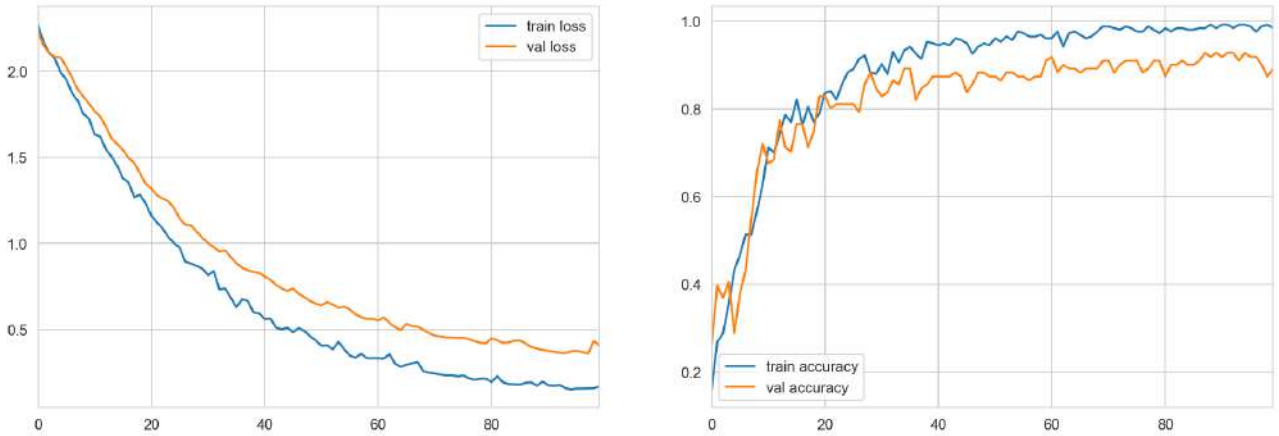

Рисунок 4.21 – Динаміка кількісних характеристик втрат та точності для тренувальної та валідаційної вибірок на ітераціях навчання нейронної мережі

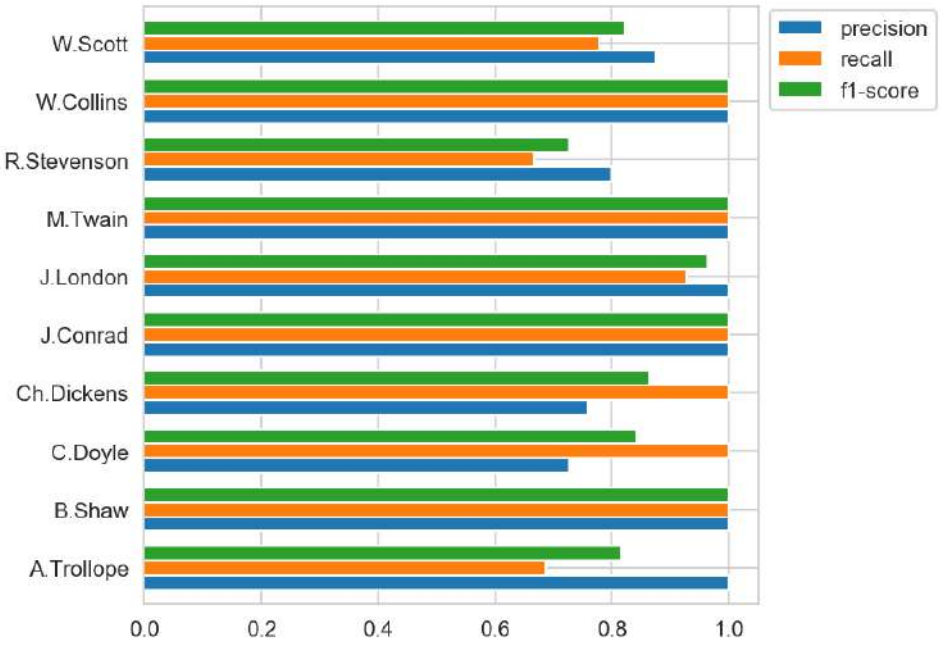

Рисунок 4.22 – Оцінки класифікації нейронної мережі при використанні сукупних ознак

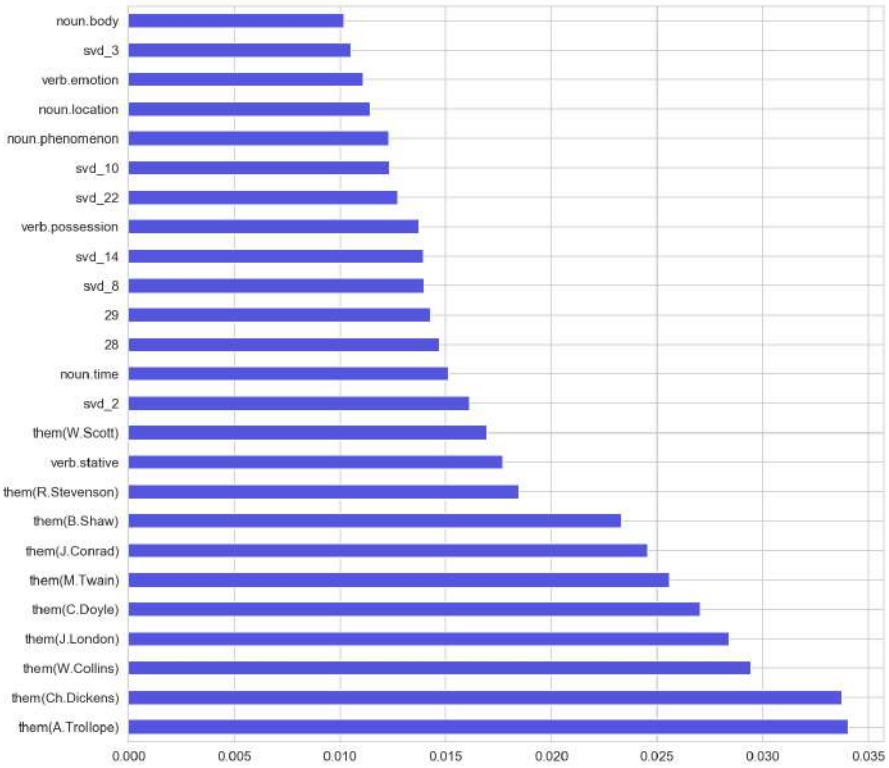

Рисунок 4.23 – Важливість ознак в алгоритмі XGBoost

важливість ознак в алгоритмі XGBoost. На рис. 4.24 наведено оцінки класифікації при використанні сукупних семантичних ознак в алгоритмі XGBoost. На рис. 4.25 наведено динаміку багатокласової класифікаційної похибки на тренувальному та валідаційному сетах при використанні сукупних семантичних ознак в алгоритмі XGBoost. На рис. 4.26 наведено приклад одного із дерев рішень алгоритму XGBoost. Висока точність класифікації авторських текстів у векторному просторі семантичних ознак свідчить про наявність у цьому просторі відокремлених областей авторського ідіолекта, які характеризують індивідуальний стиль авторів.

Також проведено класифікаційний аналіз для стандартизованої текстової вибірку груп новин 20 Newsgroups [268] з використанням алгоритму XGBoost при використанні сукупних семантичних ознак на основі семантичних і тематичних полів, компонент сингулярного розкладу TF-IDF матриці та компонент латентного розміщення Діріхле. Параметри алгоритму XGBoost були такими: 'colsample_bytree':0.15, 'subsample':0.85, 'objective':'multi:softprob', 'n_estimators':300, 'learning_rate':0.01, 'max_depth':3. Результати класифікаційного аналізу наведено на рис. 4.27.

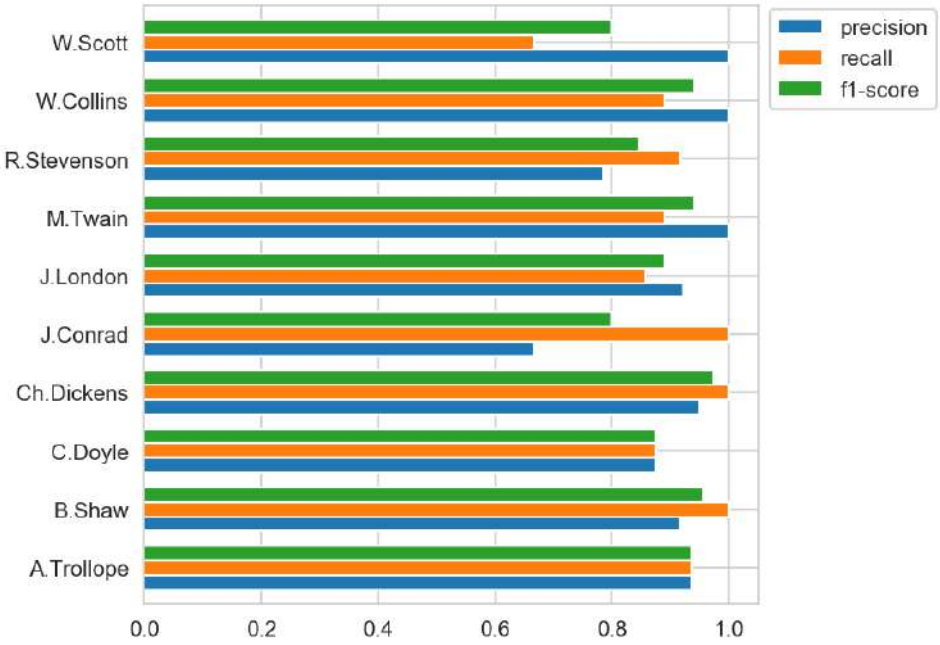

Рисунок 4.24 – Оцінки класифікації при використанні сукупних семантичних ознак в алгоритмі XGBoost

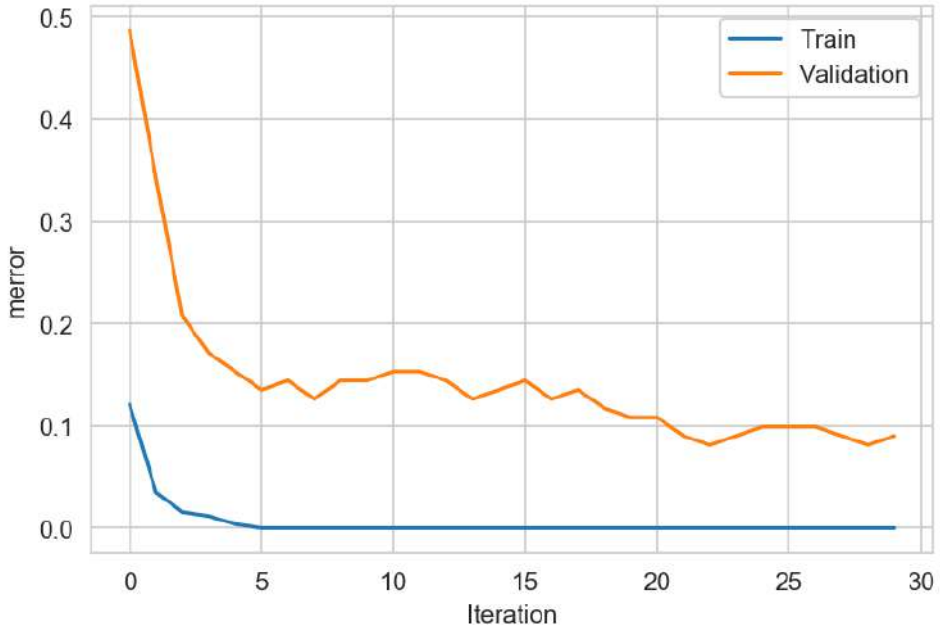

Рисунок 4.25 – Динаміка багатокласової класифікаційної похибки на тренувальному та валідаційному сетах при використанні сукупних семантичних ознак в алгоритмі XGBoost

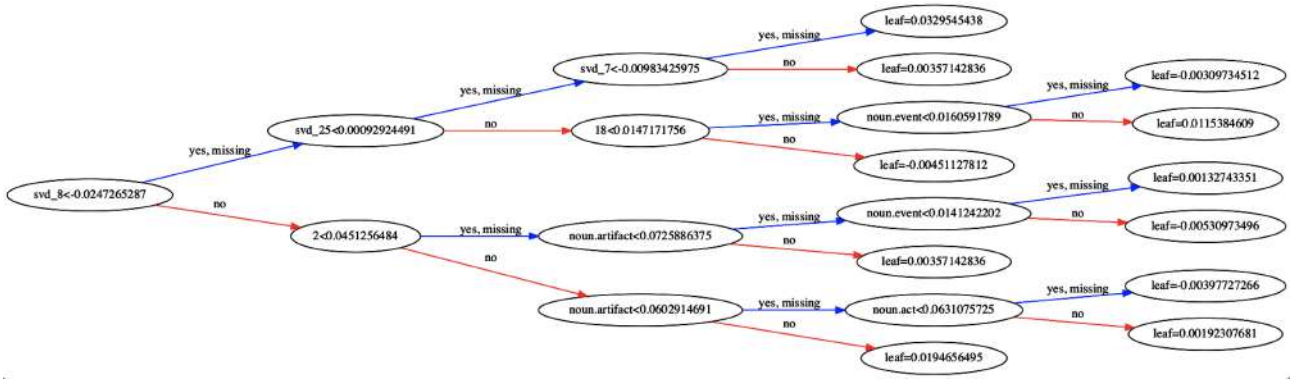

Рисунок 4.26 – Приклад одного із дерев рішень алгоритму XGBoost

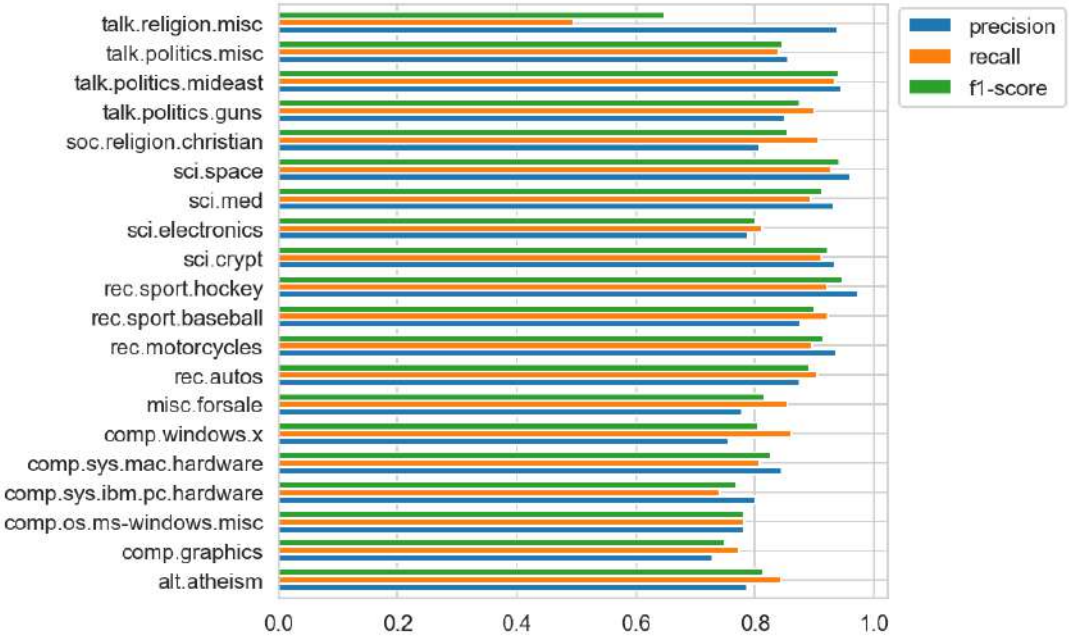

Рисунок 4.27 – Оцінки класифікації при використанні сукупних семантичних ознак в алгоритмі XGBoost

Отримані результати показують, що використання комбінованого набору класифікаційних ознак на основі семантичних та тематичних полів, компонент сингулярного розкладу матриці TF-IDF та компонент латентного розміщення Діріхле дає можливість зменшити кількість аналітичних ознак у 3-10 разів для певного класу задач інтелектуального аналізу текстів у порівнянні з набором ознак на основі частотних характеристик слів.

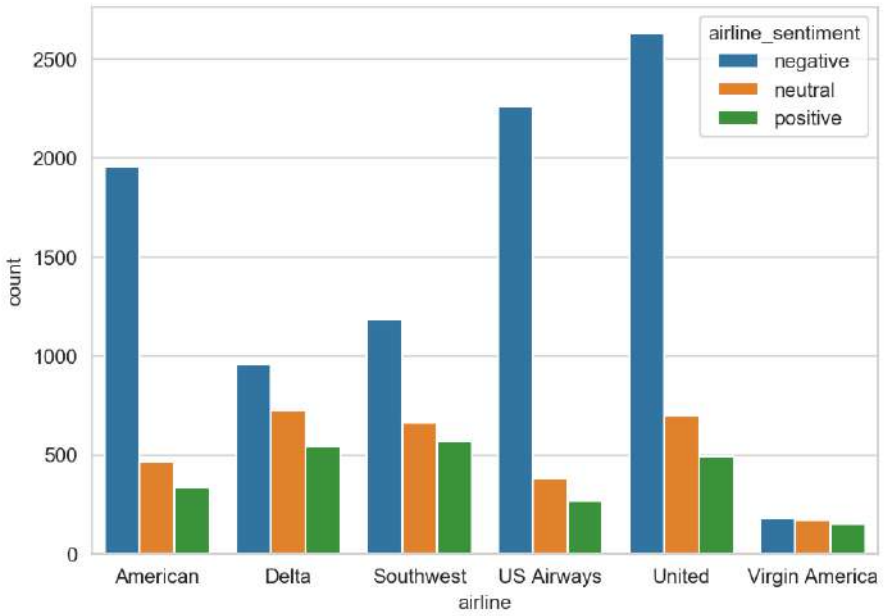

Рисунок 4.28 – Розподіл кількості твітів за класами цільової змінної для кожної авікомпанії

4.3 Використання рекурентних нейронних мереж та семантичних ознак в аналітиці текстових даних

Розглянемо формування семантичних ознак текстових даних, класифікаційний аналіз текстових документів та використання рекурентних нейронних мереж в аналітиці текстових даних [264]. Рекурентні нейронні мережі із шарами з довгою короткочасною пам'яттю (Long short-term memory, LSTM) [299, 300] часто використовують в аналітиці текстових даних. Розглянемо аналітику повідомлень соціальної мережі Твіттер за допомогою рекурентних нейронних мереж. Повідомлення соціальної мережі Твіттер можна розглядати як короткі текстові документи. Для аналізу твітів ми вибрали базу даних із повідомленнями Твітів, які стосуються сервісу декількох авіакомпаній [301]. Ця вибірка містить близько 14640 твітів. Твіти мають мітки 'negative', neutral', 'positive'. Розподіл класів за авіакомпаніями показано на рис.4.28. Для класифікаційного аналізу було вибрано рекурентну нейронну мережу із шаром LSTM. Структуру такої мережі наведено на рис. 4.29. Для чисельних розрахунків було вибрано такі параметри: максимальна довжина текстових стрічок – 50 слів, максимальна кількість ознак вбудованого шару – 300, розмірність вбудованого шару – 5, параметр швидкості навчання – 0.0003, розмір пакету даних на ітерації

навчання – 64. Набір твітів було розділено на тренувальну та валідаційну вибірки, розмір валідаційної вибірки 30%. На рис. 4.30 наведено динаміку кількісних характеристик втрат та точності для тренувальної та валідаційної вибірок на ітераціях навчання рекурентної нейронної мережі. На рис. 4.31 зображено кількісні величини точності, повноти та f1-оцінки класифікації твітів рекурентною мережею.

Розглянемо випадок класифікаційного аналізу за допомогою мережі, яка складається з двох підмереж. Одна підмережа містить двонаправлений LSTM шар, а друга складається із повністю з'єднаних шарів. На вхід першої мережі подаються текстові дані, а на вхід другої – числові дані, які характеризують текстові документи. Такими числовими даними можуть бути кількісні характеристики семантичних ознак. Для простоти розгляду вибрано один тип семантичних характеристик, який формується на основі сингулярного розкладу TF-IDF матриці. Для розрахунків взято 30 перших компонент такого розкладу. Структуру нейронної мережі із підмережами для текстових та числових даних зображено на рис. 4.29. На рис. 4.33 наведено оцінки класифікації твітів такою комбінованою нейронною мережею, а на рис. 4.34 - динаміку кількісних характеристик втрат та точності для тренувальної та валідаційної вибірок на ітераціях навчання такої нейронної мережі. Як випливає із отриманих результатів, використання нейронної мережі, яка складається із об'єднаних підмереж, дає дещо кращі результати на валідаційному сеті, криві втрат та точності прогресують швидше, ніж у випадку рекурентної мережі, яка складається лише із шарів для обробки текстової інформації. Розглянемо задачу числової регресії за наявності комбінованих даних текстового та числового типу. Для цього випадку було вибрано дані опису товарів та їх цін. Ці дані було завантажено із платформи Kaggle [302]. Для простоти аналізу вибрано лише одну категорію даних, для аналізу було взято 15000 зразків даних. Крім текстового опису дані містять також категоріальні змінні, які було об'єднано як стрічнові дані з текстовим описом. Як числові характеристики вибрано 30 перших компонент SVD розкладу TF-IDF матриці. Структуру нейронної мережі, яку було використано у цьому аналізі, зображено на рис. 4.35. Для аналізу було вибрано такі основні параметри: максимальна довжина текстових стрічок –

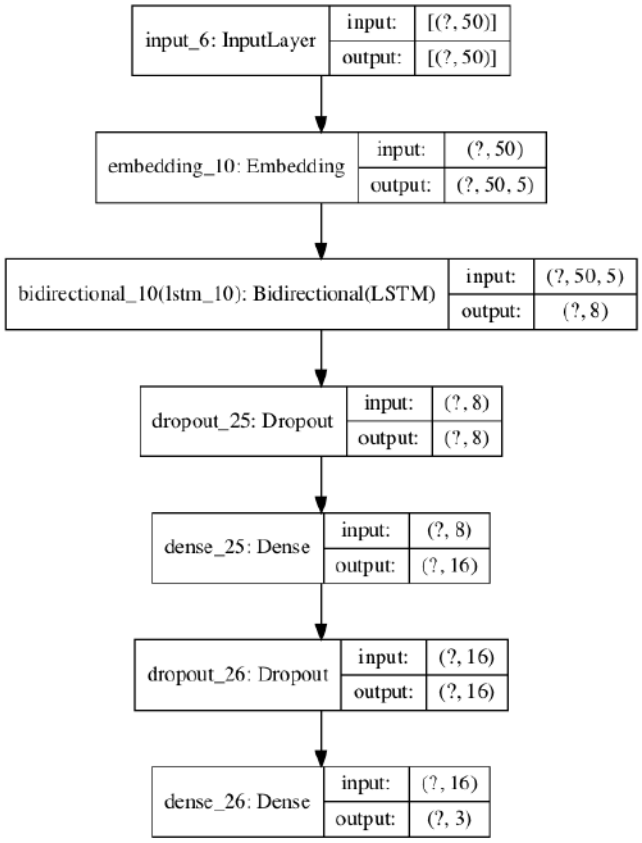

Рисунок 4.29 – Структура рекурентної нейронної мережі із двонаправленим LSTM шаром

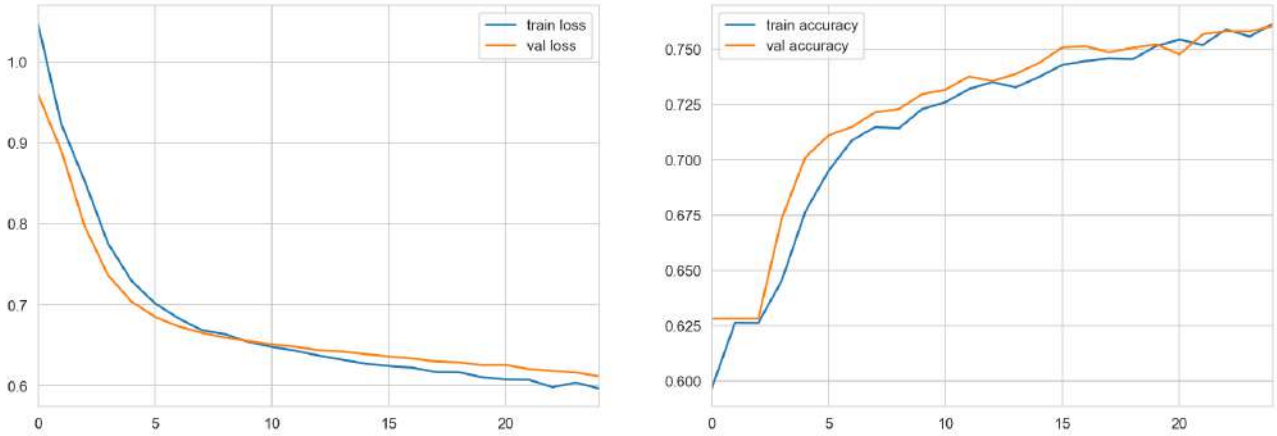

Рисунок 4.30 – Динаміка кількісних характеристик втрат та точності для тренувальної та валідаційної вибірок на ітераціях навчання рекурентної нейронної мережі

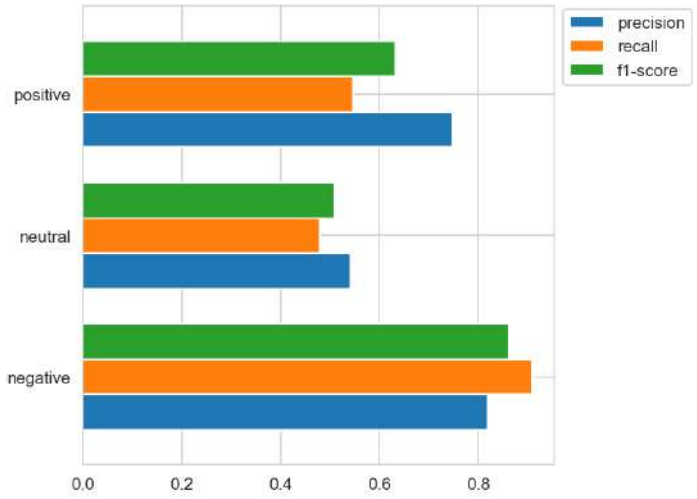

Рисунок 4.31 – Оцінки класифікації твітів рекурентною мережею

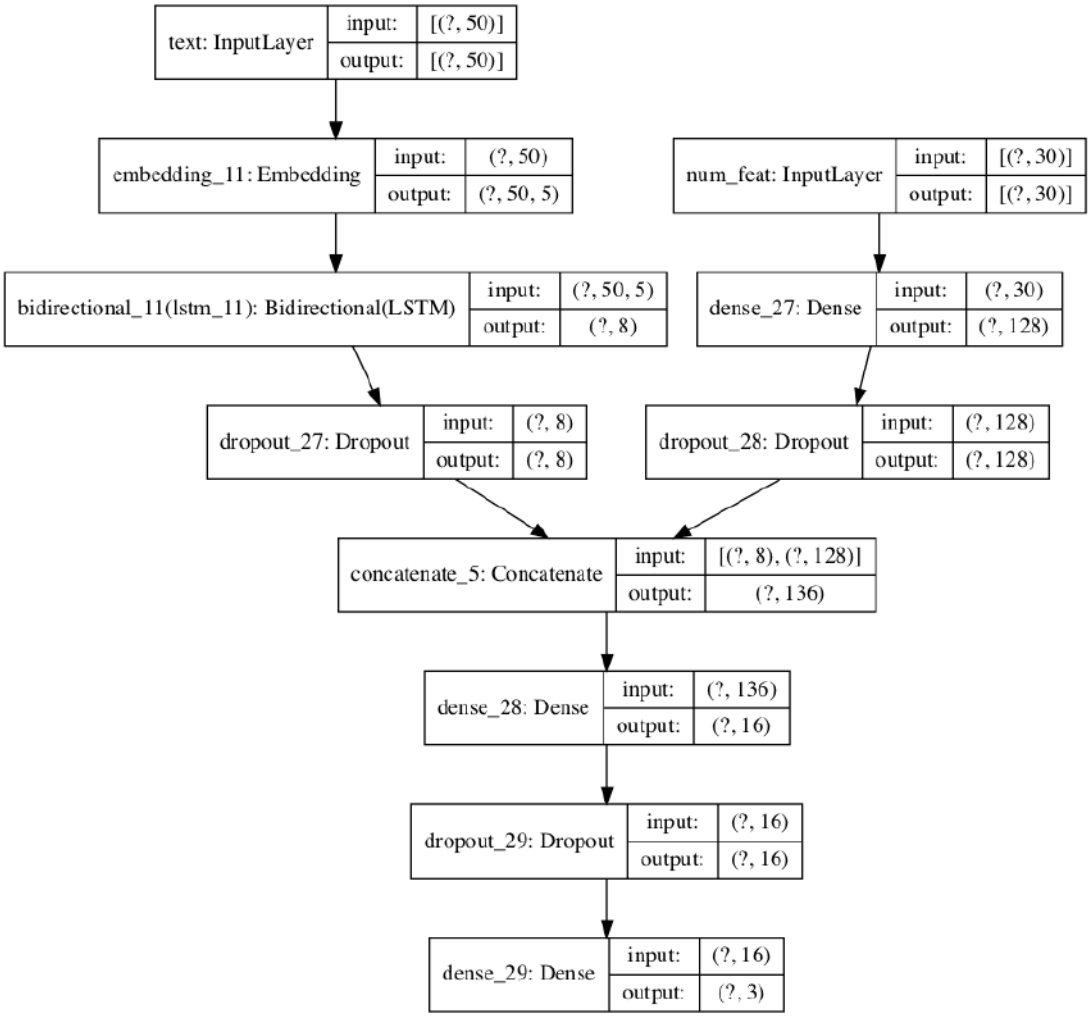

Рисунок 4.32 – Структура нейронної мережі із підмережами для текстових та числових даних

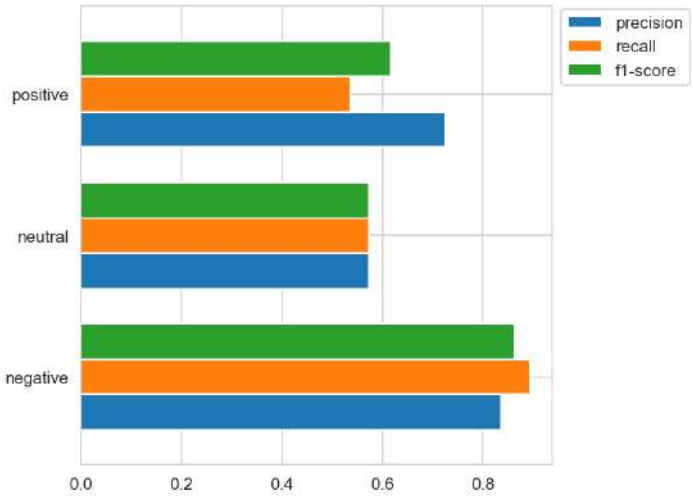

Рисунок 4.33 – Оцінки класифікації твітів комбінованою нейронною мережею

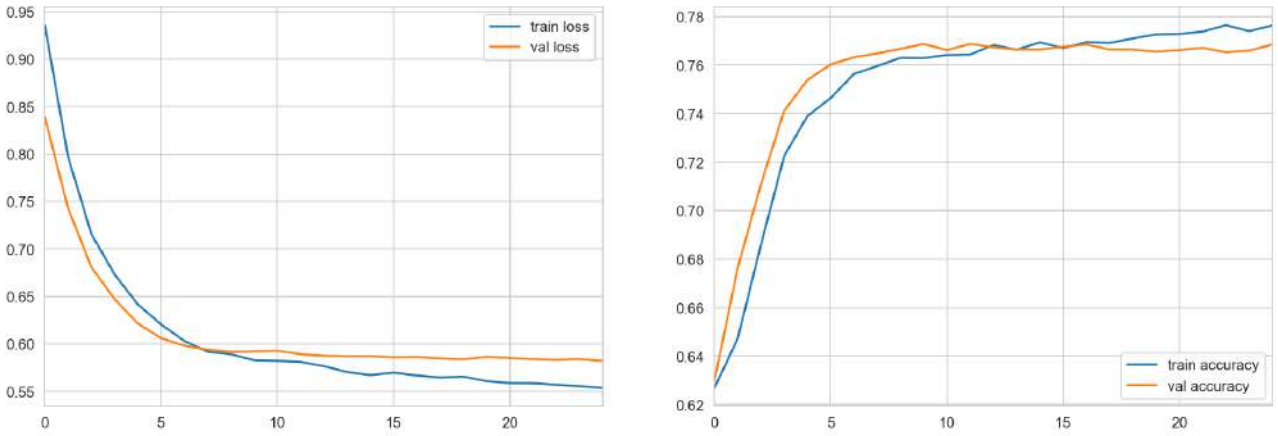

Рисунок 4.34 – Динаміка кількісних характеристик втрат та точності для тренувальної та валідаційної вибірок на ітераціях навчання комбінованої нейронної мережі.

250 слів, максимальна кількість ознак вбудованого шару – 3000, розмірність вбудованого шару – 10, параметр швидкості навчання – 0.0003, розмір пакету даних на ітерації навчання – 32. Як функція втрат розглядались характеристики RMSE, MAE. Цільову змінну, яка описує ціну товару, розглянуто в алгоритмі навчання у логарифмічному масштабі на основі перетворення $y = lg(x + 1)$. Після отримання прогнозованих значень було проведено зворотнє перетворення масштабу даних за функцією $y = 10^x - 1$. Оцінка похибки MAE, отримана на валідаційному наборі даних, дорівнює 4.1, відносна оцінка MAE, яка є відношенням абсолютної оцінки MAE до середнього значення ціни, дорівнює 0.3. На рис. 4.36 наведено динаміку функцій RMSE та MAE на ітераціях навчання нейронної мережі для тренувального та тестового наборів даних. Як випливає із отриманих даних, алгоритм глибокого навчання покращує точність прогнозування після реалізації деякої кількості ітерацій процесу навчання.

Отже, досліджено використання семантичних ознак в інтелектуальному аналізі текстових даних, зокрема, у класифікації текстових документів. Як семантичні ознаки розглянуто семантичні та тематичні поля, складові сингулярного розкладу матриці TF-IDF та складові латентного розміщення Діріхле. Класифікаційний аналіз здійснено за допомогою алгоритму Random Forest та алгоритмів глибокого навчання нейромереж із різною структурою з використанням двонаправлених шарів із довгою короткостроковою памяттю (LSTM). LSTM шари нейронної мережі дають можливість враховувати порядок та комбінації лексем. Розглянуто випадок використання комбінованої нейронної мережі, яка складається із рекурентної нейронної підмережі для аналізу текстових даних та підмережі для числових семантичних ознак текстових документів. Проаналізовано числову регресію, у якій як вхідні розглядались текстові дані для випадку аналізу цін за текстовим описом товарів. Для аналізу було вибрано аналогічну комбіновану нейронну мережу із LSTM підмережею для текстових даних і підмережею із повністю з'єднаними шарами для числових компонент SVD розкладу TF-IDF матриці. Результати показують, що у текстових даних опису товарів можна бути знайти відповідні патерни і, як наслідок, точність прогнозування ціни товару за текстовим описом покращується на ітераціях

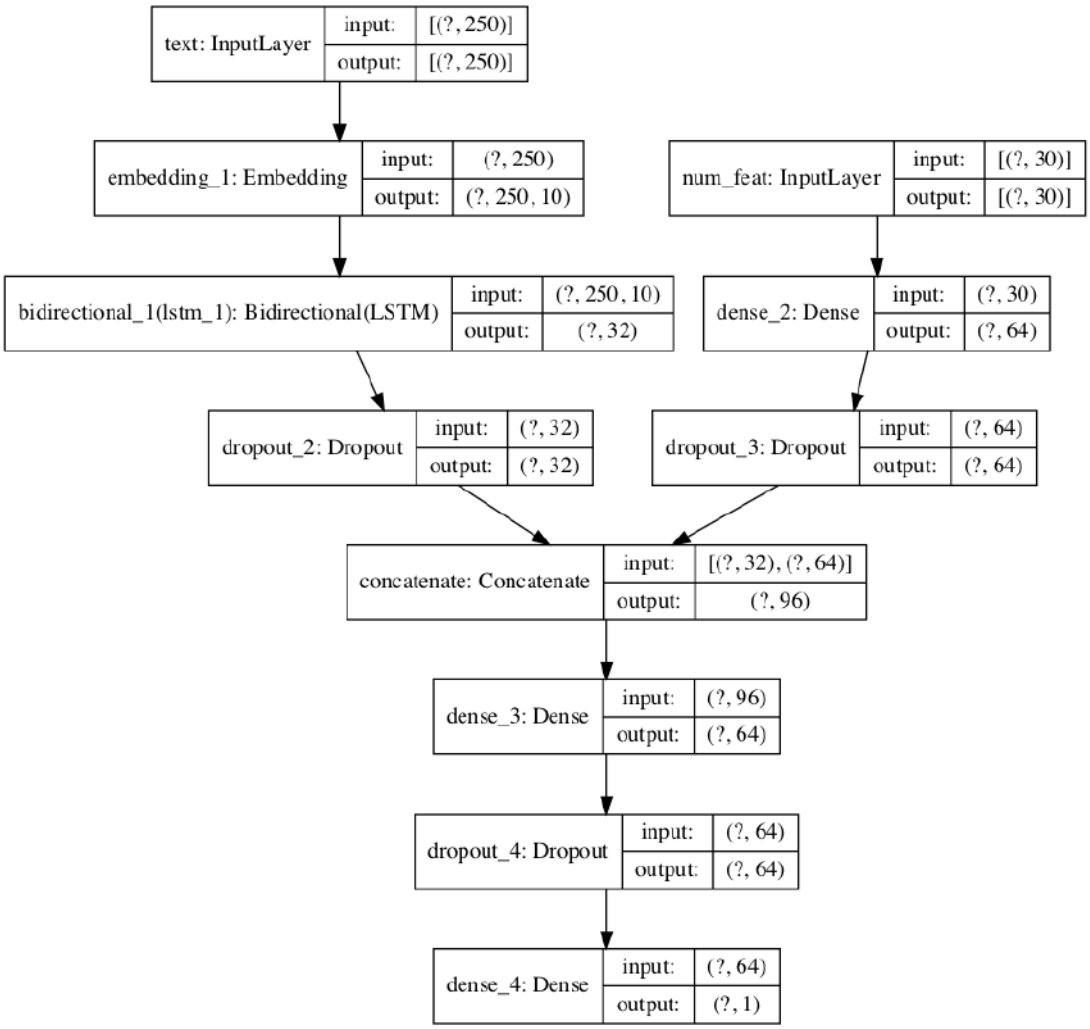

Рисунок 4.35 – Структура нейронної мережі з підмережами для текстових та числових даних

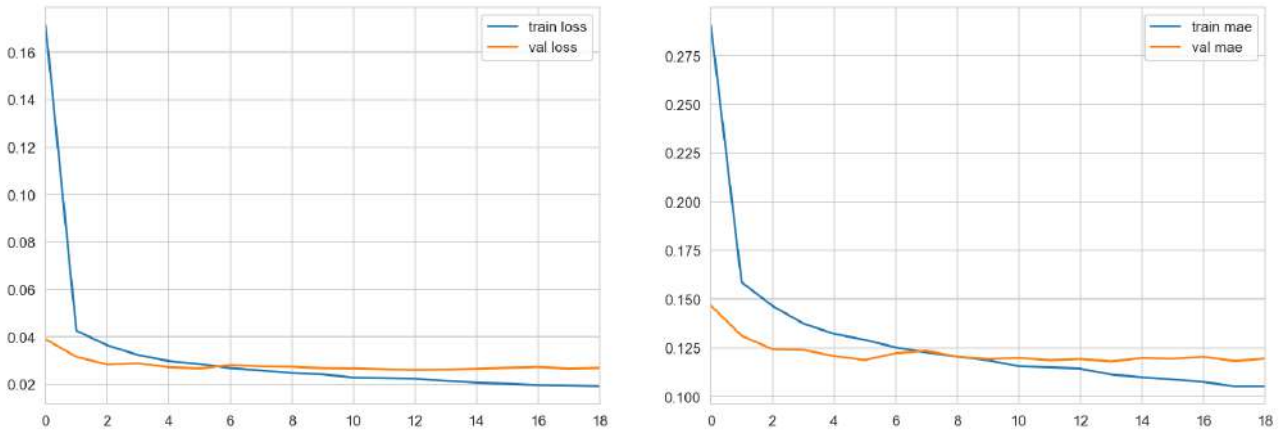

Рисунок 4.36 – Динаміка кількісних характеристик втрат та точності для тренувальної та валідаційної вибірок на ітераціях навчання рекурентої нейронної мережі

тренування такої нейронної мережі. Комбінації різних семантичних ознак дають можливість отримати вищу точність у задачах класифікацій текстових документів. Використання широкого класу семантичних ознак у задачах інтелектуального аналізу диверсифікує аналітичні підходи і збільшує простір ознак в аналітичних задачах, що є важливим при невеликій кількості даних та при аналізі нестаціонарних процесів, коли прогнозний потенціал різних ознак може змінюватися з часом [264]. Як випливає із отриманих результатів, компоненти векторів текстових документів у просторі семантичних ознак на основі кількісних характеристик семантичних та тематичних полів, компонент розміщення Діріхле та компонент сингулярного розкладу матриці TF-IDF ефективно відображають семантичну структуру цих документів. Розмірність такого простору є суттєво меншою за розмірність простору на основі частот лексем текстових документів. Розроблений метод класифікації текстових даних за експертно сформованими семантичними ознаками дозволяє проводити інтелектуальний аналіз текстових масивів із відповідними семантичними акцентами і дає можливість за певних умов зменшити кількість семантичних ознак в 3-10 разів у порівнянні з кількістю лексемних частотних ознак для заданих характеристик точності інтелектуального аналізу текстових даних.

4.4 Метод формування базису семантичного простору текстових документів за допомогою генетичних алгоритмів

Однією із актуальних задач є пошук оптимальних векторних підпросторів документів для класифікаційного та кластерного аналізу текстових документів. Зокрема, задача полягає у відборі семантичних полів, частотні характеристики яких можуть використовуватися як вхідні параметри текстових класифікаторів із задовільною точністю. Розв'язок такої задачі оптимізує необхідну кількість обрахунків та точність класифікатора в інтелектуальному аналізі текстів. Один із перспективних методів формування базису семантичного простору може бути побудований із використанням генетичних алгоритмів. Розглянемо можливість використання еволюційного програмування та генетичних алгоритмів

для формування оптимального простору семантичних полів у задачах інтелектуального аналізу текстів [303, 304]. Створимо векторну модель генетичного відбору семантичних полів у задачі класифікації текстових документів. Експериментально дослідимо формування семантичного підпростору у класифікаційному аналізі на прикладі класифікатора за найближчими k сусідами. Як основу для експериментального дослідження виберемо стандартизовану колекцію повідомлень груп новин 20-Newsgroups. Генетичні алгоритми використовують в аналітиці текстових даних [305, 126, 305]. Використання генетичних алгоритмів у формуванні аналітичних ознак розглянуто в [124, 125]. Зменшення аналітичного простору за допомогою генетичних алгоритмів розглянуто в [123]. Як цільову функцію можна розглядати похибку класифікатора або деяку кількісну характеристику кластерної структури текстових документів. Як вхідні параметри оптимізаційної задачі розглянемо набір семантичних полів, які утворюють базис векторного простору текстових документів.

Розглянемо теоретико-множинну модель генетичного відбору семантичних полів. Сукупність семантичних полів векторного базису у контексті генетичних алгоритмів назвемо хромосомою семантичних полів. У нашій задачі у ролі генів виступають індекси семантичних полів. Розглянемо теоретико-множинну модель генетичного алгоритму оптимізації відбору семантичних полів для утворення семантичного простору текстових документів. Еволюцію генетичної оптимізації розглянемо у вигляді впорядкованої множини популяцій

$$Ev^s = \{\, Pop_k^s | k = 1, 2, ... | Ev^s | \,\}. \tag{4.1}$$

Вважаємо, що одне покоління хромосом утворюється однією популяцією. Популяція складається із множини хромосом

$$Pop_k^s = \left\{ \chi_{jk}^{sp} | j = 1, 2, ... | Pop_k^s |; k = 1, 2, ... | Ev^s | \right\}. \tag{4.2}$$

У загальному випадку різні популяції можуть містити різну кількість хромосом. У спрощеному випадку вважаємо, що кількість хромосом є

однаковою у всіх популяціях, тобто

$$|Pop_k^s| = |Pop^s| = N_{pop}^{\chi}. \tag{4.3}$$

Кожну хромосому розглянемо як набір семантичних полів

$$\chi_{jk}^{sp} = \left\{ s_{ijk}^{f\chi p} | i = 1, 2...|\chi^s|; j = 1, 2, ...|Pop_k^s|; k = 1, 2, ...|Ev^s| \right\}, \tag{4.4}$$

де верхні індекси в $s_{ijk}^{f\chi p}$ позначають назви нижніх індексів: f – індекс семантичного поля; χ – індекс хромосоми; p – індекс популяції. Оператор одноточкового кросинговеру розглянемо у вигляді

$$\begin{array}{l} Crossover^p(\chi_{1k}^{sp}, \chi_{2k}^{sp}, m): \\ \left\{ \begin{array}{l} s_{11k}^{f\chi p} s_{21k}^{f\chi p} ... s_{m1k}^{f\chi p} ... s_{|\chi^s|1k}^{f\chi p} \\ s_{12k}^{f\chi p} s_{22k}^{f\chi p} ... s_{m2k}^{f\chi p} ... s_{|\chi^s|2k}^{f\chi p} \end{array} \right. \Rightarrow \left\{ \begin{array}{l} s_{11k}^{f\chi p} s_{21k}^{f\chi p} ... s_{m2k}^{f\chi p} ... s_{|\chi^s|2k}^{f\chi p} \\ s_{12k}^{f\chi p} s_{22k}^{f\chi p} ... s_{m1k}^{f\chi p} ... s_{|\chi^s|1k}^{f\chi p} \end{array} \right. . \end{array} \tag{4.5}$$

Індекс m позначає точку поділу хромосоми на дві частини семантичних полів, які обмінюють у батьківських хромосомах для утворення двох дочірніх хромосом наступної популяції. Оператор мутації розглянемо у вигляді

$$Mutation(\chi_{jk}^{sp}, m) : s_{1jk}^{f\chi p} s_{2jk}^{f\chi p} ... s_{mjk}^{f\chi p} ... s_{|\chi^s|jk}^{f\chi p} \Rightarrow s_{1jk}^{f\chi p} s_{2jk}^{f\chi p} ... \tilde{s}_{mjk}^{f\chi p} ... s_{|\chi^s|jk}^{f\chi p}. \tag{4.6}$$

Внаслідок дії оператора $Mutation(\chi_{jk}^{sp}, m)$ змінюється семантичне поле $s_{mjk}^{f\chi p}$ на семантичне поле $\tilde{s}_{mjk}^{f\chi p}$. Текстові документи представимо у семантичному просторі у вигляді вектора текстових частот семантичних полів p_{kj}^{sd}, які відображають частоту семантичного поля S_k у текстовому документі d_j. Значення частот p_{kj}^{sd} визначені як суми текстових частот лексем у аналізованому документі d_j, які належать заданому семантичному полю S_k. Сукупність значень p_{kj}^{sd} утворюють матрицю ознака-документ, у якій ознаками виступають частоти семантичних полів у документах:

$$M_{sd} = \left(p_{kj}^{sd}\right)_{k=1,j=1}^{N_s,N_d}. \tag{4.7}$$

Формування матриці частоти_семантичних_полів-документи розглянуто у роботах [263, 280, 292]. Значення частот p_{kj}^{sd} визначені як суми текстових

частот в аналізованому документі, які належать заданому семантичному полю. Вектор

$$V_j^s = \left(p_{1j}^{sd}, p_{2j}^{sd}, ..., p_{N_sj}^{sd}\right)^T. \tag{4.8}$$

відображає документ d_j в N_s-мірному просторі текстових документів із базисом, утвореним семантичними полями. Розглянемо використання генетичного алгоритму для оптимізації набору семантичних полів у задачі класифікації текстових документів. Як цільову функцію для еволюційної оптимізації набору семантичних полів базису семантичного простору розглянемо точність класифікатора. Нехай існують деякі категорії текстових документів. Ці категорії можуть мати різну природу, наприклад, можуть визначати авторський ідіолект, дискурс, характеризувати різні об'єкти, явища, події тощо. У нашому експериментальному аналізі такі категорії утворюють групи новин. Множину цих категорій позначимо

$$Categories = \{Ctg_m | m = 1, 2, ..., N_{ctg}\}, \tag{4.9}$$

де $N_{ctg} = |Categories|$ визначає розмір множини категорій. За даними категоріями розподілені текстові документи множини D. Завдання полягає у пошуку цільової функції, яку описує відображення

$$F_{d \to ctg} : Categories \times D \to \{0, 1\}. \tag{4.10}$$

Цільову функцію генетичної оптимізації визначимо так:

$$F_S^{ga} = 1 - Pr_{avg}, \tag{4.11}$$

де Pr_{avg} – усереднена за всіма категоріями точність класифікатора. Завдання генетичної оптимізації буде полягати у мінімізації цільової функції F_S^{ga}. Існують підходи, які базуються на мінімізації інших типів функцій. У роботі [119] розглянуто генетичні алгоритми, які базуються на мінімізації штрафних функцій. Штрафні функції обраховують на основі цільових функцій з урахуванням обмежень на мінімальні та максимальні значення вхідних параметрів. Такий підхід часто використовують при оптимізації цілочисельних параметрів.

Розглянемо експериментальні дослідження генетичної оптимізації базису простору семантичних полів. Як метод класифікації у дослідженні генетичної оптимізації розглянемо класифікацію за найближчими k сусідами, яку називають kNN класифікацією [306, 167, 166]. Цей метод відносять до векторних класифікаторів. В основі векторних методів класифікації лежить гіпотеза компактності. Згідно з цією гіпотезою, документи, які належать одному і тому ж класу, утворюють компактну область, а області, які належать різним класам, не перетинаються. Як міру близькості між документами виберемо евклідову відстань. У kNN класифікації границі категорій визначають локально. Деякий документ відносять до категорії, яка домінує у k його сусідів. У випадку $k = 1$ документу приписують категорію його найближчого сусіда. Згідно з гіпотезою компактності тестовий документ d має ту категорію, яку мають більшість документів навчальної вибірки у деякому просторовому локальному околі документа d. У генетичному відборі семантичних полів як вхідні параметри задачі оптимізації використаємо індекси масиву семантичних полів. Результатом генетичної оптимізації буде масив індексів, який визначає оптимальний набір семантичних полів. Для експериментального вивчення класифікації текстових документів у просторі семантичних полів ми вибрали стандартизовану текстову базу повідомлень груп новин 20 Newsgroups [268]. Для формування семантичного простору вибрано лексеми, згруповані за семантичними полями іменників та дієслів у семантичній мережі WordNet. Проведено початкову обробку текстового масиву, вилучено допоміжні символи та текстові елементи, які не несуть семантичної інформації. Для кожного документа та вибірки в цілому, обраховано частотні словники, на основі яких розраховано матрицю типу частота_семантичного_поля–документ. Навчальну та тестову вибірки було вибрано рівними загальному об'єму аналізованого текстового масиву повідомлень. Для обчислень були використані генетичні алгоритми пакету прикладних програм Matlab [307]. Для реалізації генетичної оптимізації складу семантичних полів у класифікаційному аналізі використано оператор розсіяного кросоверу, однорідну селекцію батьківських хромосом та наявність групи елітарних хромосом. Оптимальні значення були вибрані експериментальним шляхом. Аналіз проведено

при різних значеннях параметрів оптимізації. Розглянуто популяції розміром 30 хромосом. Кількість елітарних хромосом дорівнює 3. На рис. 4.37, 4.38, 4.39 наведено динаміку мінімального $F^{ga}_{s(\min)}$ та середнього значення $F^{ga}_{s(avg)}$ цільової функції $F^{ga}_{s(avg)}$ із різними значеннями частки хромосом, які утворені оператором кросовера та оператором мутації. Загальна кількість аналізованих семантичних полів дорівнює 41. Розмір семантичних хромосом був вибраний таким, що дорівнює 5. Тобто, здійснювалась генетична оптимізація набору із 5 семантичних полів, для яких класифікація здійснювалась із найменшою похибкою. На рис. 4.37 спостерігається динаміка швидкого спадання середнього значення цільової функції до мінімального значення. Це зумовлено відсутністю у популяції нових значень вхідних параметрів. Генетичний відбір здійснюється лише на основі тих вхідних індексів семантичних полів, які знаходились у початковій популяції і були згенеровані випадковим чином. На рис. 4.38 середнє значення цільової функції наближується до мінімального значення і коливається в околі деякого значення. Ці коливання зумовлені наявністю хромосом, утворених за допомогою оператора мутації. Мутації дають можливість отримувати хромосоми з новими значеннями вхідних параметрів, що є ефективним у випадку наявності локальних мінімумів цільової функції. Поява нових значень індексів семантичних полів дає можливість генетичному алгоритму вийти із області можливого локального мінімуму і продовжити пошук глобального мінімуму цільової функції. Динаміка середнього значення цільової функції на рис. 4.39 характеризується коливаннями на значній відстані від мінімального значення, що зумовлено малою фракцією хромосом, утворених оператором кросовера. Як випливає із отриманих даних, підбір параметрів генетичної оптимізації, зокрема частки кросовера, є важливим для ефективного пошуку глобального мінімуму цільової функції. В аналізованих дослідженнях оптимальне значення кросовера дорівнює 0.8. Також досліджувався вплив елітарних хромосом. При зменшенні кількості елітарних хромосом із 3 до 1 суттєво збільшувався розкид середніх значень цільової функції у послідовних популяціях. У результаті генетичної оптимізації отримано мінімальне значення цільової функції, яке відповідає набору індексів, які визначає такі семантичні поля у класифікації WordNet:

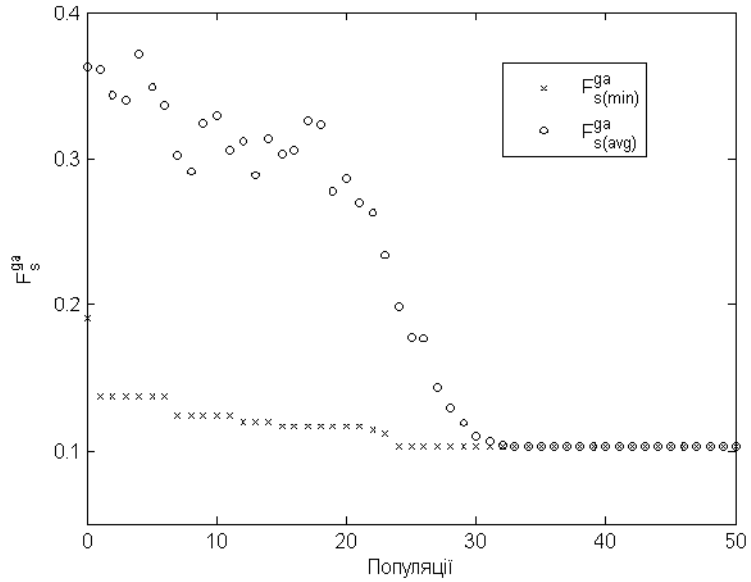

Рисунок 4.37 – Динаміка мінімального та середнього значення цільової функції при фракції кросовера, яка дорівнює 1

noun.event, noun.phenomenon, verb.competition, verb.possession, verb.weather.

У [303] розглянуто генетичну оптимізацію частотного лексемного поля у класифікаційному аналізі текстових даних. Проаналізовано генетичну оптимізацію ключових слів, частоти яких є складовими векторів документів і відіграють роль атрибутів у класифікаційному аналізі текстів. Генетична оптимізація здійснювалась на множині слів, яка є фракцією частотного словника із заданими частотними границям. Частотний словник утворено на основі аналізованого масиву текстовів творів англомовної прози. Експериментальний масив текстових документів складався із 503 текстів англомовної художньої прози, які було класифіковано за категоріями 17 авторів. Навчальна вибірка складалась із 350 випадково вибраних документів, а тестова вибірка складалась із 153 документів. Множину ключових слів для генетичної оптимізації утворено першими 1000 лексемами частотного словника, для яких текстова частота менша за 0.001. Ці лексеми утворюють частотний проміжок. Аналізувались популяції розміром 50 хромосом. Застосовано оператор розсіяного кросовера із розміром фракції 0.8. У кожній популяції було вибрано 5 елітних хромосом. Аналізувались хромосоми із розміром 30 та 10 лексем. Як класифікатор було вибрано класифікатор за k наближчими сусідами. Як цільову функцію, яку

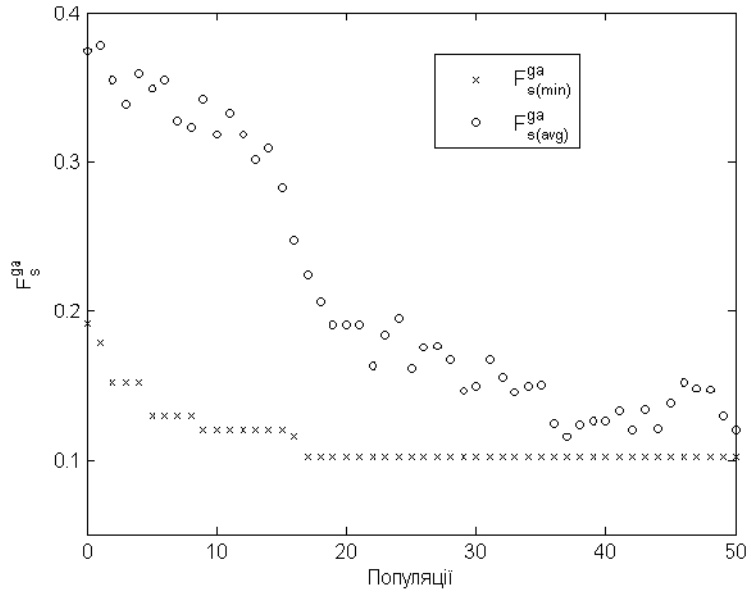

Рисунок 4.38 – Динаміка мінімального та середнього значення цільової функції при фракції кросовера, яка дорівнює 0.8

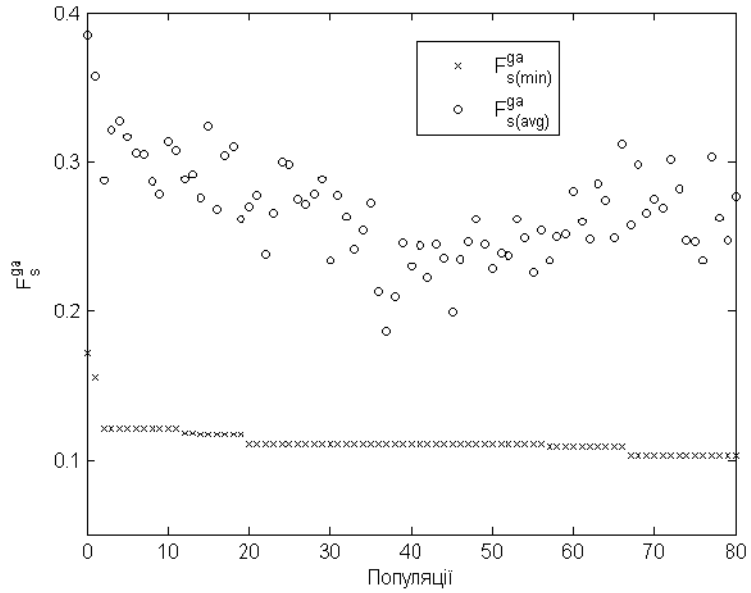

Рисунок 4.39 – Динаміка мінімального та середнього значення цільової функції при фракції кросовера рівній 0.5

мінімізує генетичний алгоритм, було використано похибку класифікатора за найближчими к сусідами. Отримані результати показують високу точність та повноту класифікації текстів за категоріями авторства на основі множини атрибутів ключових слів, які відібрано генетичним алгоритмом з частотного словника.

Отже, за допомогою генетичних алгоритмів можна оптимізувати набір семантичних полів, які утворюють векторний простір документів в алгоритмах інтелектуального аналізу текстових даних. Як цільову функцію для генетичної оптимізації використано точність класифікатора за найближчими k сусідами. Проведені експериментальні дослідження на тестовій вибірці текстових повідомлень груп новин показують ефективність використання генетичних алгоритмів для оптимізації набору семантичних полів та лексем, які утворюють базис векторного простору документів у класифікаційному аналізі текстових документів [303, 304].

4.5 Аналіз семантичних образів у масивах текстових об'єктів за допомогою квантових обчислень

Квантові комп'ютери та алгоритми є новим перспективним напрямком сучасних інформаційних технологій. Вони дають можливість суттєво пришвидшити розв'язок деяких класів задач унаслідок реалізації квантового паралелізму та заплутаності квантових станів. Пошук нових алгоритмів аналізу слабоструктурованих даних, зокрема, текстового типу є одним з актуальних напрямів сучасних інформаційних технологій. Якісно нові підходи до такого аналізу можливі при використанні квантових обчислень, теорія і методи яких активно розвиваються [308, 309, 310, 311, 312, 313, 314, 315]. Одним із ефективних квантових алгоритмів є алгоритм Гровера [310, 311, 312], який дає можливість реалізувати пошук у невпорядкованій вибірці даних поліноміально швидше у порінянні із класичними алгоритмами внаслідок реалізації квантового паралелізму.

Розглянемо можливість використання квантових алгоритмів в інтелектуальному аналізі даних. У роботі [316] показано, що реалізація одновимірних кліткових автоматів на квантових логічних елементах дає можливість побудувати квантові алгоритми аналізу правил переходів

кліткових автоматів, які дають поліноміальне прискорення в порівнянні із класичними алгоритмами. У [317] розглянуто числове моделювання алгоритму Гровера для квантового пошуку даних. Актуальним є розгляд можливості використання квантових алгоритмів в інтелектуальному аналізі даних слабоструктурованого типу, зокрема, текстових даних. Розглянемо представлення семантичного вектора текстового об'єкта у квантовій пам'яті. Складові семантичного вектора V_j^s (3.18) відображення текстових документів у просторі семантичних полів є дійсними величинами, які можуть набувати значень у проміжку [0,1]. Однак можна знайти відображення цього вектора на множину бінарних значень:

$$V_j^s \to V_j^{sb}, \tag{4.12}$$

де

$$V_j^{sb} = \left(p_{1j}^{stb}, p_{2j}^{stb}, ..., p_{N_s j}^{stb}\right), p_{ij}^{stb} \in \{0,1\}. \tag{4.13}$$

У найпростішому випадку відображення (4.12) може здійснюватися так:

$$p_{ij}^{stb} = \begin{cases} 0, p_{ij}^{st} < (p_i^{st})_{th}, \\ 1, p_{ij}^{st} \geq (p_i^{st})_{th}, \end{cases} \tag{4.14}$$

де $(p_i^{st})_{th}$ – деякі порогові значення частот семантичних полів, вибрані експериментальним шляхом. У загальному випадку, якщо потрібно враховувати різні значення частот, можна використати декілька двійкових розрядів для кодування. Наприклад, якщо збільшити розмір бінарного вектора V_j^{sb} у два рази, тоді можна буде кодувати кожну частоту двома двійковими розрядами, що буде відповідати чотирьом інтервалам значень частот. Нехай існує деякий масив текстових об'єктів. Кожне семантичне поле можна закодувати його номером за допомогою відображення

$$U_s : S_i \to i. \tag{4.15}$$

. Для простоти розгляду будемо вважати, що розмір множини масиву текстових об'єктів дорівнює $N_t = 2^{(nt)}$, а розмір множини семантичних полів дорівнює $N_w = 2^{(ns)}$. Тоді, для кодування положення об'єкта в

масиві потрібно nt двійкових елементів, а для кодування семантичного образу потрібно ns двійкових елементів. Нехай квантовий регістр складається з двох частин $|qt\rangle$ і $|qs\rangle$:

$$|qt\rangle \otimes |qs\rangle, |qt\rangle = |q_1^t, q_1^t, ..., q_{nt}^t\rangle, |qs\rangle = |q_1^s, q_1^s, ..., q_{ns}^s\rangle. \tag{4.16}$$

Регістр $|qt\rangle$ описує положення текстового об'єкта у масиві, а регістр $|qt\rangle$ описує семантичний образ, який складається з набору семантичних полів. Уведемо додатковий кубіт $|f\rangle$, який буде відображати наявність семантичного поля із заданим номером у заданому текстовому об'єкті. Якщо дане поле присутнє в заданому текстовому об'єкті, тобто, якщо його бінарна частота p_{ij}^{stb} (4.14) дорівнює 1, то значення цього кубіта дорівнює $|1\rangle$, в іншому випадку дорівнює $|0\rangle$. Тоді весь масив текстових об'єктів можна буде записати у вигляді такої системи кубітів:

$$|qr\rangle = |q_1^t, q_1^t, ..., q_{nt}^t\rangle \otimes |q_1^s, q_1^s, ..., q_{ns}^s\rangle \otimes |f\rangle. \tag{4.17}$$

Для запису у квантову пам'ять масиву текстових об'єктів розміром N_t, який містить набір семантичних полів розміром N_s, достатньо $nt + ns + 1 = \log_2(2N_tN_s)$ кубітів. Така експоненційна економія квантової пам'яті у порівнянні із класичною пам'яттю можлива внаслідок реалізації квантового паралелізму. Наприклад, якщо набір семантичних полів містить 2^5 елементів, а масив текстових об'єктів – 10^{15} елементів, то для запису такої інформації необхідно лише 5+15+1=21 кубітів. Запис масиву текстових об'єктів у квантову пам'ять розглянемо феноменологічно за допомогою квантового оракула. У теорії квантових обчислень показано, що на основі однокубітних та двокубітних квантових унітарних вентилів можна побудувати еквівалентні алгоритми класичної машини Тюрінга. Під оракулом будемо розуміти деяке формалізоване унітарне перетворення, за допомогою якого реалізуються наперед задані обчислення. Елементи масиву текстових об'єктів визначаються квантовими станами складеного регістру кубітів (4.17). Суперпозиція цих станів утворює вектор у комплексному Гільбертовому просторі. Цей вектор є квантовим еквівалентним відображенням текстових об'єктів. Розглянемо послідовність

квантового запису масиву текстових об'єктів. Кубіт $|f\rangle$ візьмемо в початковому стані $|0\rangle$, а регістри $|qt\rangle$, $|qs\rangle$ – в початкових станах $|0\rangle^{\otimes(nt)}$ та $|0\rangle^{\otimes(ns)}$ відповідно. Застосуємо однокубітні унітарні перетворення Адамара до регістрів $|qt\rangle \otimes |qs\rangle \otimes |f\rangle$:

$$|\psi\rangle = H^{\otimes(nt)} \otimes H^{\otimes(nw)} \otimes I\left(|qt\rangle \otimes |qs\rangle \otimes |f\rangle\right). \tag{4.18}$$

У результаті отримаємо

$$|\psi\rangle = \frac{1}{\sqrt{2^{nt+ns}}} \sum_{i=0,j=0}^{N_t,N_s} |i\rangle \otimes |j\rangle \otimes |0\rangle. \tag{4.19}$$

Суперпозиція $|\psi\rangle$ містить базисні ортогональні стани, кожен з яких відповідає одному запису семантичного вектора текстового об'єкта. У процесі вимірювання відбувається редукція суперпозиції до одного базового стану, який відповідає одному текстовому об'єкту. Отже, та кількість пам'яті, яка в класичному випадку необхідна для запису семантичного вектора одного текстового об'єкта, у квантовому випадку є достатньою для запису цілого масиву текстового об'єкта. Нехай наявність заданого семантичного вектора текстового об'єкта описується функцією $f_q(qt, qs)$, де індекс qt описує текстовий об'єкт, а індекс qs описує спектр семантичних полів текстового об'єкта. У випадку наявності заданого спектра семантичних полів у заданому текстовому об'єкті функція набуває значення 1, інакше 0. Процес запису тексту у квантову пам'ять опишемо унітарним перетворенням U_F, яке визначається квантовим оракулом:

$$U_F : |qt\rangle \otimes |qs\rangle \otimes |f\rangle \to |qt\rangle \otimes |qs\rangle \otimes |f\rangle \otimes f_q(qt, qs)\rangle, \tag{4.20}$$

де $\otimes$ означає сумування за модулем 2. Враховуючи, що кубіт $|f\rangle$ є в початковому стані $|0\rangle$, отримаємо

$$U_F : |qt\rangle \otimes |qs\rangle \otimes |0\rangle \to |qt\rangle \otimes |qs\rangle \otimes |f_q(qt, qs)\rangle, \tag{4.21}$$

Розглянемо пошук заданих семантичних образів текстових об'єктів у

квантовій базі даних. Завдання полягає у пошуку деякого ключового семантичного образу, який може бути закодований у вигляді квантового стану:

$$|qk\rangle = |q_1^k, q_1^k, ..., q_{ns}^k\rangle. \tag{4.22}$$

Використаємо вентиль Тоффолі для знаходження квантових станів, у яких закодовано ключові семантичні образи. Уведемо в систему кубітів (4.17) додатковий кубіт $|z\rangle$ – анцилу, отримаємо

$$|qr\rangle = |qt\rangle \otimes |qs\rangle \otimes |f_q(qt, qs)\rangle \otimes |z\rangle. \tag{4.23}$$

Подіємо на кубіт $|z\rangle$ у стані $|0\rangle$ оператором Адамара:

$$|z\rangle = H|1\rangle = \frac{1}{\sqrt{2}}\left(|0\rangle - \|1\rangle\right). \tag{4.24}$$

Допоміжний кубіт $|z\rangle$ буде керуватись за допомогою $ns+1$-мірного елемента Тоффолі T_{ns+1}, у якому керуючими кубітами виступають ns кубітів регістру $|qs\rangle$ та кубіт $|f(qt, qs)\rangle$. Значення кубіта $|z\rangle$ у квантовому стані може змінитися під впливом вентиля Тоффолі у випадку, коли всі кубіти регістру $|qs\rangle$ та кубіт $|f(qt, qs)\rangle$ будуть дорівнювати одиниці. Щоб перевести значення кубітів у значення $|1\rangle$ для квантових станів, які описують ключовий семантичний образ $|qk\rangle$, розглянемо унітарний оператор, який є тензорним добутком однокубітних операторів:

$$S_T^q = I^{\otimes(nt)} \otimes (\otimes_{i=1}^{ns} S_i^q) \otimes I \otimes I, S_i^q = \begin{cases} I, q_i^k = 1, \\ X, q_i^k = 0. \end{cases} \tag{4.25}$$

Оператор S_T^q переводить квантові базисні стани регістру $|qr\rangle$ в стани, у яких значення регістру $|qs\rangle$ дорівнюють 1, якщо співпадають кодування семантичного вектора текстового об'єкта у тексті та ключового семантичного образу $|qs\rangle = |qk\rangle$, тобто

$$S_T|qr\rangle = |qt\rangle \otimes |1, 1, ..., \rangle_{ns} \otimes |f_q(qt, qs)\rangle \otimes |z\rangle, \tag{4.26}$$

якщо

$$|qs\rangle = |qk\rangle.$$

Розглянемо оператор

$$U_T = (S_T^q) \cdot (I^{\otimes(nt)} \otimes T_{ns+1}) \cdot (S_T^q). \tag{4.27}$$

. Подіємо цим оператором на систему регістрів кубітів $|qr\rangle$. Перша група операторів справа, виділених дужками, переводить регістр $|qs\rangle$ у значення $|1,1,...,1\rangle_{nw}$ для станів, які відповідають шуканим ключовим семантичним образам, друга група операторів реалізує інверсію керованого кубіта $|z\rangle$ для шуканих станів семантичних образів, третя група повертає змінені першою групою стани у стан перед застосуванням оператора U_T. У результаті дії оператора U_T отримаємо

$$\begin{array}{l} |\psi\rangle_T = U_T \left(\frac{1}{\sqrt{2^{nt+ns}}} \sum\limits_{x=1}^{x=2^{nt+ns}} |x\rangle \otimes |f(x)\rangle \otimes |z\rangle \right) = \\ \frac{1}{\sqrt{2^{nt+ns}}} \left(\sum\limits_{x \notin X_k} |x\rangle \otimes |f(x)\rangle - \sum\limits_{x \in X_k} |x\rangle \otimes |f(x)\rangle \right) \otimes |z\rangle, \\ |x\rangle = |qt\rangle \otimes |qs\rangle, \end{array} \tag{4.28}$$

де X_k – множина станів суперпозиції, які відповідають кодуванню шуканих ключових семантичних образів. Допоміжний керований кубіт $|z\rangle$ не змінив свого значення під дією оператора U_T і знаходиться у стані (4.24), в якому він знаходився перед дією оператора U_T, тому його можна винести за дужки та вилучити із подальшого розгляду. Це зумовлено тим, що кубіт $|z\rangle$ було переведено у новий базисний стан (4.24) за допомогою оператора Адамара. Інверсія стану у цьому базисі рівнозначна інверсії знаку амплітуди підсистеми квантових станів, у яких закодовані ключові слова. Роль цього кубіта полягала в тому, щоб під дією вентиля Тофолі він змінював свій знак на протилежний у квантових станах, які відповідають умові пошуку. Умова рівності регістрів семантичного вектора текстового об'єкта та регістру ключового семантичного образу враховується в операторі S_T^q . Наступним кроком алгоритму є переведення додаткового кубіта $|f\rangle$ у

початковий базисний стан $|0\rangle$ та вилучення його із подальшого розгляду. Це можна зробити за допомогою унітарного оператора U_F^{-1}, який є оберненим до оператора U_F (4.21). У результаті отримаємо

$$|\psi\rangle_{F^{-1}} = U_F^{-1}|\psi\rangle_T = \frac{1}{\sqrt{2^{nt+ns}}}\left(\sum_{x\notin X_k}|x\rangle - \sum_{x\in X_k}|x\rangle\right), |x\rangle = |qt\rangle\otimes|qs\rangle. \quad (4.29)$$

Отже, в результаті дії складеного оператора $U_F^{-1}U_TU_F$ отримаємо суперпозицію базисних рівноймовірнісних станів (4.19), у якій стани, що кодують наявні ключові семантичні образи, мають від'ємну амплітуду. Для того, щоб при вимірюванні виявити квантові стани, у яких закодовано ключові семантичні образи, необхідно підсилити амплітуди цих станів. Таке підсилення амплітуд станів, які мають задані властивості, можна здійснити за допомогою оператора інверсії відносно середнього, який використовується в алгоритмі Гровера для пошуку в неструктурованій квантовій базі даних [310, 312]. Оператор інверсії розглянемо у вигляді

$$U_G = 2|\psi_c\rangle\langle\psi_c| - I, \quad (4.30)$$

де

$$|\psi_c\rangle = H^{\otimes n}|0\rangle^{\otimes n} = \frac{1}{\sqrt{2^n}}\sum_{i=0}^{i=2^n-1}|i\rangle. \quad (4.31)$$

. У геометричній інтерпретації оператор U_G здійснює в Гільбертовому просторі дзеркальне відображення деякого вектора відносно осі, яка визначається вектором $|\psi_c\rangle$. Оператор інверсії можна представити сукупністю однокубітних операторів Адамара та операторів інверсії стану кубіта відносно базисного вектора $|0\rangle$:

$$U_G = H^{\otimes n}(2|0\rangle\langle 0| - I)^{\otimes n}H^{\otimes n}. \quad (4.32)$$

Підсилення амплітуд станів з інверсними знаками амплітуд відбувається внаслідок дії оператора інверсії U_G аналогічно до механізму, описаного в алгоритмі Гровера [310, 312]. Враховуючи визначення операторів (4.20), (4.27), (4.30) розглянемо ітерацію алгоритму Гровера у загальному вигляді

складеного оператора:

$$U_I = U_G U_F^{-1} U_T U_F. \tag{4.33}$$

Можна показати, що внаслідок реалізації ітерації U_I можна підсилити амплітуди заданих станів у 3 рази. Якщо шуканий ключовий семантичний образ зустрічається в масиві текстових об'єктів лише один раз, то, аналогічно алгоритму Гровера [310, 312], оптимальна кількість ітерацій U_I буде

$$k_U \approx \frac{\pi}{4}\sqrt{N}, N = 2^{nt+ns}, \tag{4.34}$$

де N – кількість квантових станів рівна $N = N_t N_s$. Отже, складність алгоритму в даному випадку буде $O(\sqrt{N})$. Якщо кількість шуканих об'єктів із ключовим семантичним образом дорівнює l, тоді, згідно з [310, 312], кількість необхідних ітерацій буде дорівнювати

$$k_U \approx \frac{\pi}{4}\sqrt{\frac{N}{l}}. \tag{4.35}$$

Однак, наперед невідомо, скільки текстових об'єктів відповідають ключовому семантичному образу. В такому випадку можна провести серію реалізацій алгоритму Гровера з кількостями ітерацій, які утворюють прогресію

$$k_U = 1, 2, 4, 8, ...l_U. \tag{4.36}$$

де l_U – деяке максимальне значення кількості ітерацій U_I. Тобто, спочатку реалізується алгоритм з однією ітерацією, потім з двома і т.д. Якщо в цій серії реалізацій алгоритму при вимірюванні виявлено шукані квантові стани, тоді можна прийняти рішення про наявність шуканого семантичного образу у текстових об'єктах аналізованого масиву, а також оцінити кількість текстових об'єктів із заданим семантичним образом у цьому масиві. Можна показати, що складність алгоритму в такому випадку є також $O(\sqrt{N})$. У класичному алгоритмі складність пошуку семантичних образів у неструктурованому масиві текстових об'єктів буде $O(N)$.

На даний час існує можливість реалізації квантових обчислень на реальних квантових комп'ютерах, зокрема, на комп'ютерах IBM Q через

відповідний хмарковий сервіс з використанням пакету Qiskit для мови програмування Python [318, 319, 320, 321, 322]. Опис деяких квантових обчислень наведено у Додатках.

Отже, розглянуто модель квантового представлення семантичних векторів масиву текстових об'єктів, яка дає можливість експоненційно зменшити об'єм необхідної квантової пам'яті у порівнянні із класичним записом. Запропоновано квантовий алгоритм пошуку ключових семантичних образів у масивах текстових об'єктів. Реалізація цього алгоритму здійснюється на основі квантових логічних елементів, зокрема, з використанням вентиля Тоффолі. Ітерація Гровера використовується для підсилення амплітуд квантових станів, які описують семантичні вектори текстових об'єктів. Показано, що реалізація квантових алгоритмів аналізу семантичних образів текстових об'єктів для деякого класу задач дає можливість поліноміально зменшити час виконання алгоритму у порівнянні з класичними алгоритмами внаслідок реалізації квантового паралелізму [323, 324].

4.6 Висновки

– Запропоновано метод кластеризації текстових документів у семантичному просторі дає можливість отримувати новий структурний поділ документів за семантичними ознаками у просторі суттєво меншої розмірності, ніж простір, утворений лексемним складом текстової вибірки.

– Проведений кластерний аналіз текстових вибірок різних типів показує, що дослідження текстів різними алгоритмами кластеризації у різних просторах семантичних ознак є ефективним методом структурного аналізу текстів та методом формування семантичних ознак на основі приналежності текстових документів до відповідних кластерів. Належність текстового документа до певного кластера у різних методах кластеризації у різних семантичних просторах можна розглядати як додаткову ознаку у класифікаційному та структурному аналізі текстових масивів.

– Авторські тексти містять індивідуальний стиль авторів, що відображається у кластерній структурі. Тексти деяких авторів домінують в окремих кластерах. Структурованість текстів за авторським ідіолектом спостерігається у просторах семантичних полів різних типів. Найбільш виражена структурованість спостерігається у просторі тематичних полів. Семантичні просторові області кластерів із домінуванням текстів окремих авторів володіють диференціювальним потенціалом для авторських ідіолектів і можуть бути використані в аналізі авторських текстів як додаткові фактори аналізу авторського лексикону. Області семантичного простору, що відповідають кластерам, у яких домінують декілька авторів, можна розглядати як області семантичної спорідненості цих авторів.

– Досліджено використання семантичних ознак у класифікації текстових документів. Як семантичні ознаки розглянуто семантичні та тематичні поля, складові компоненти сингулярного розкладу TF-IDF матриці та складові латентного розміщення Діріхле. Класифікаційний аналіз здійснено за допомогою алгоритму Random Forest та алгоритмів глибокого навчання нейромереж із різною структурою з використанням двонаправлених шарів із довгою короткостроковою пам'яттю (LSTM). Розглянуто випадок використання комбінованої нейронної мережі, яка складається із рекурентної нейронної підмережі для аналізу текстових даних та підмережі для числових семантичних ознак текстових документів.

– Запропоновано підхід на основі комбінації різних семантичних ознак, зокрема, семантичних та тематичних полів, компонент сингулярного розкладу TF-IDF матриці та компонент латентного розміщення Діріхле, який дає можливість отримати вищу точність у задачах класифікацій текстових документів. Використання широкого класу семантичних ознак у задачах інтелектуального аналізу диверсифікує аналітичні підходи і збільшує простір ознак в аналітичних задачах, що є важливим при невеликій кількості даних та при аналізі нестаціонарних процесів, коли прогнозний потенціал різних ознак може змінюватися з

часом.

– Розроблений метод класифікації текстових даних за експертно сформованими семантичними ознаками дозволяє проводити інтелектуальний аналіз текстових масивів із відповідними семантичними акцентами і дає можливість за певних умов зменшити кількість семантичних ознак у 3-10 разів у порівнянні із набором лексемних частотних ознак для заданих характеристик точності інтелектуального аналізу текстових даних.

– Проаналізовано регресійну задачу, у якій, як вхідні, розглядалися текстові дані для випадку аналізу цін за текстовим описом товарів. Для аналізу було вибрано комбіновану нейронну мережу із LSTM підмережею для текстових даних і підмережею із повністю з'єднаними шарами для числових компонент сингулярного розкладу матриці TF-IDF. Показано, що у зразках даних із числовими та текстовими типами ознак та числовою цільовою змінною нейромережа може виявляти відповідні патерни, що проявляється у зростанні точності прогнозування на ітераціях навчання комбінованої нейронної мережі.

– Запропоновано та досліджено використання генетичних алгоритмів для оптимізації набору семантичних полів, які утворюють векторний простір документів в алгоритмах інтелектуального аналізу текстових даних. Як цільову функцію для генетичної оптимізації використано точність класифікатора.

– Розглянуто модель квантового представлення семантичних векторів масиву текстових об'єктів, яка дає можливість експоненційно зменшити об'єм необхідної квантової пам'яті у порівнянні з класичним записом. Запропоновано квантовий алгоритм пошуку ключових семантичних образів у масивах текстових об'єктів. Реалізація цього алгоритму здійснюється на основі квантових логічних елементів, зокрема, з використанням вентиля Тоффолі. Ітерація Гровера використовується для підсилення амплітуд квантових станів, які описують семантичні вектори текстових об'єктів. Показано, що реалізація квантових

алгоритмів аналізу семантичних образів текстових об'єктів для деякого класу задач дає можливість поліноміально зменшити час виконання алгоритму у порівнянні з класичними алгоритмами внаслідок реалізації квантового паралелізму.

5 ВИКОРИСТАННЯ ТЕОРІЇ ЧАСТИХ МНОЖИН ТА АСОЦІАТИВНИХ ПРАВИЛ У ІНТЕЛЕКТУАЛЬНОМУ АНАЛІЗІ ТЕКСТОВИХ ДАНИХ

5.1 Семантичний аналіз текстових даних з використанням частих множин та асоціативних правил

Теорію частих множин та асоціативних правил можна застосувати в аналітиці текстових даних для виявлення та аналізу певних сукупностей об'єктів, які часто зустрічаються у великих масивах і характеризуються деякими ознаками. Об'єктом може бути текстовий документ, а ознаками – домінуючі лексеми чи семантичні поля. Розглянемо алгоритми виявлення частих множин та асоціативних правил на прикладі обробки повідомлень мікроблогів Твіттера [325]. Об'єднання частих множин, які відображають суть тематики повідомлень мікроблогів, будемо розглядати як семантичні поля заданої тематики. Це дасть можливість звузити семантичний аналіз повідомлень до заданих тематичних рамок. На основі отриманих частих семантичних множин проаналізуємо можливі асоціативні правила, які відображають внутрішні семантичні зв'язки тематичних понять у повідомленнях. Використовуючи API системи Twitter, завантажено тестовий масив повідомлень, які містять ключове слово "software" а також хештег "#software" загальним об'ємом 10Мб. Тобто, відібрано повідомлення заданого тематичного напряму, пов'язаного з програмним забезпеченням. Далі повідомлення були згруповано за користувачами. На основі отриманих повідомлень було розраховано часті множини ключових лексем. На основі цих множин сформоване тематичне семантичне поле, у яке входили ключові лексеми заданого семантичного значення. Із розгляду було вилучено повідомлення, які не містили термінів із заданого семантичного поля. Загальна кількість повідомлень, взятих для аналізу, дорівнює 44720. Були відфільтровані всі лексеми, які зустрічаються в повідомленнях менше 10 раз. Після фільтрації словник повідомлень містив 4062 лексеми, в якому мінімальна кількість появи лексеми дорівнює 10, а максимальна – 2095. Максимальна кількість появ була в хештегу #software. Очевидно, що ключове слово *software* входило в кожне повідомлення за умовою пошуку,

тому з розгляду було виключене. Як обмеження підтримки частих множин було взято величину зустрічань частих множин термінів, яка дорівнює 5. Наведемо приклади отриманих частих множин ключових лексем:

{netscape, browser, internet}, {telemarketing, sales}, {earth, features, tracking, google}, {#linux, ubuntu, center}, {#opensource, #linux}, {installs, printer, drivers}, {phones, popular, android}, {phones, games, android}, {#hiring, #job, developer }, {#job, developer}, {#careers, #job }, {coding, programming, hardware}, {satellite, internet}, {explorer, firefox, internet}, {trial, antivirus, security}

Наведені часті множини ключових лексем відображають семантику тенденцій в повідомленнях мікроблогів у певний момент часу. Очевидно, що з часом масив частих множин ключових лексем із заданою підтримкою змінюється. Відслідковуючи зміну величини підтримки $Supp(F)$ (1.80), можна відслідковувати тенденції тематик обговорень у мікроблогах. На основі отриманих частих множин було сформовано тематичне семантичне поле, у яке ввійшли лексеми, що відображають задану тематику аналізу. Таке семантичне поле може використовуватись у фільтруванні повідомлень для відкидання лексем, які не є актуальними для семантичного аналізу із заданою тематикою. Семантичне фільтрування суттєво зменшує об'єм необхідних обчислень. На основі відфільтрованого масиву повідомлень було побудовано асоціативні правила, які відображають семантику зв'язків тематичних ключових лексем. В таблиці 5.1 наведено деякі побудовані асоціативні правила. Асоціативні правила характеризуються величиною підтримки $Supp(F)$ (1.80), та величиною достовірності $Conf_{X\to Y}$ (1.85). Асоціативні правила, для яких $Conf_{X\to Y=100\%}$, утворюють імплікації, тобто є справедливі завжди, коли виконується умова правила. Деякі правила є тривіальними та очевидними, деякі відображають сучасні тенденції в програмному забезпеченні. Наприклад, у наведених правилах відображається популярність операційної системи Android. Можна побачити актуальність пошуку програмістів для операційних систем Android та Linux.

Отже, пошук частих множин у повідомленнях мікроблогів дає можливість сформувати тематичне семантичне поле, яке можна у подальшому використовувати для пошуку асоціативних правил. На основі

Таблиця 5.1 – Асоціативні правила на основі частих множин семантичних ознак.

Умова X	Наслідок Y	$Supp(F)$	$Conf_{X \to Y}$
{location}	{phone}	1.62%	65.62%
{#hiring}	{#job}	1.22%	61.71%
{internet, popular}	{satellite}	0.77%	94.33%
{#opensource}	{#linux}	0.66%	66.15%
{#careers}	{#hiring}	0.43%	54.9%
{explorer}	{internet}	0.3%	52.63%
{antivirus, internet}	{security}	0.3%	100.0%
{#careers, #job}	{#hiring}	0.27%	94.73%
{#job, location}	{#hiring}	0.21%	100.0%
{netscape}	{internet}	0.17%	100.0%
{antivirus, internet}	{trial}	0.17%	55.0%
{telemarketing}	{sales}	0.15%	90.9%
{netscape}	{browser}	0.15%	90.9%
{browser, internet}	{netscape}	0.15%	90.9%
{browser, netscape}	{internet}	0.15%	100.0%
{greater, sales}	{leader, telemarketing}	0.13%	81.81%
{sales, telemarketing}	{greater, leader}	0.13%	89.99%
{leader, telemarketing}	{leader, sales}	0.13%	100.0%
{android, popular}	{greater, sales}	0.13%	100.0%
{android, popular}	{phones}	0.12%	80.0%
{#hiring, location}	{installs}	0.12%	80.0%
{#job, location}	{developer}	0.12%	57.14%
{drivers, installs}	{developer}	0.12%	57.14%
{#hiring, developer, location}	{android}	0.12%	100.0%
{installs, printer}	{printer}	0.12%	100.0%
{#job, developer, location}	{#job}	0.12%	100.0%
{#discounts, security}	{android}	0.1%	100.0%
{android, games}	{internet}	0.09%	67%
{firefox, internet}	{phone}	0.09%	100.0%
{#job, android}	{internet}	0.07%	100.0%
{#linux, ubuntu}	{developer}	0.06%	67%
{#job, telemarketing}	{#linux}	0.04%	100.0%
{#hiring, programming}	{android}	0.03%	100.0%
{#linux, leader}	{#job}	0.03%	100.0%
{android, browser}	{#opensource}	0.01%	50.0%

відібраних частих множин семантичних ознак можна побудувати асоціативні правила, які будуть відображати семантичні зв'язки змісту повідомлень мікроблогів.

5.2 Використання семантичної структури твітів у прогнозній аналітиці

Проаналізуємо наявність прогнозного потенціалу семантичної структури твітів. Для нашого прикладу ми завантажували твіти, пов'язані з компанією *Tesla*, протягом певного періоду часу. Якісна структура твітів може бути використана для агрегування різних кількісних характеристик часових рядів і створення на основі них нових ознак для прогнозної моделі, яку можна використовувати, наприклад, для прогнозування цін на акції. Розглянемо які функції можна отримати з твітів для прогнозної аналітики. Зв'язки користувачів можуть бути представлені графом, де вершини відображають користувачів, а ребра – їх з'єднання. Використовуючи алгоритми інтелектуального аналізу графів, можна виявити спільноти користувачів і знайти впорядковані списки користувачів за різними характеристиками, зокрема такими, як *Hub, Authority, PageRank, Betweenness*. Для виявлення спільнот користувачів ми використовували алгоритм *Community Walktrap Algorithm*, який реалізовано в пакеті igraph [326] для середовища мови програмування R. Для візуалізації ми використовували Fruchterman-Reingold алгоритм з цього пакету. Алгоритм *Community Walktrap* здійснює пошук зв'язаних підграграфів, які також називаються спільнотами, шляхом випадкового блукання. Ідея полягає у тому, що короткі шляхи блукання, як правило, залишаються в одній спільноті [327]. Граф, який відображає зв'язки між користувачами, можна зобразити за допомогою алгоритму Fruchterman-Reingold [328]. Для оцінки графу користувачів можна використовувати такі оцінки:

– *Hub* – оцінки вершин визначаються як головний власний вектор $A \cdot A^T$, де A - матриця суміжності графу;

– *Authority* – оцінки вершин визначаються як головний власний вектор $A^T \cdot A$, де A - матриця суміжності графу;

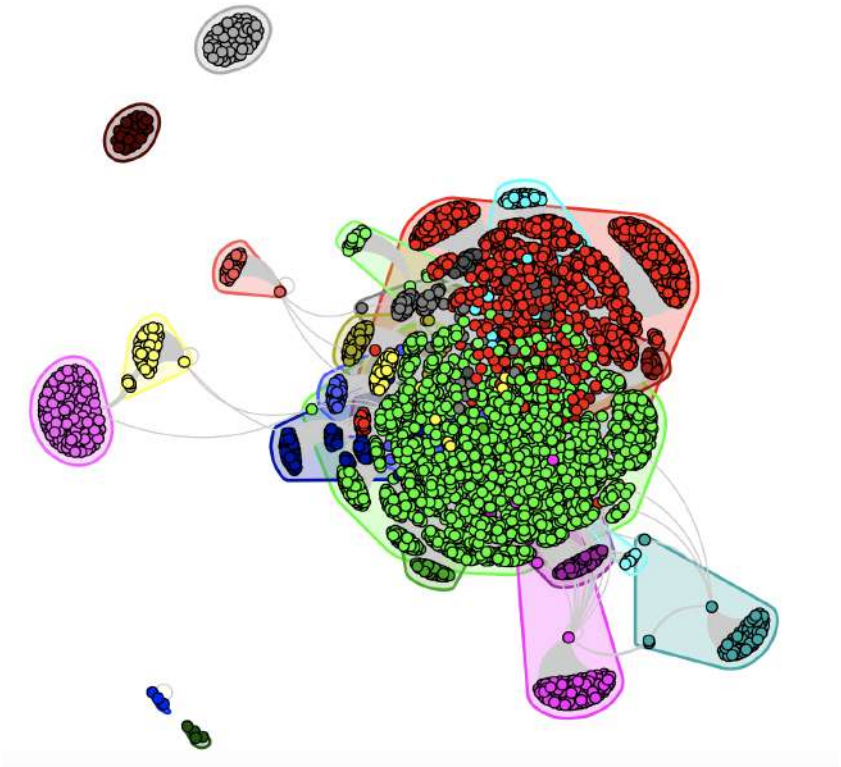

Рисунок 5.1 – Структура груп користувачів Твіттера

– PageRank – обчислює Google PageRank для вказаних вершин;

– *Betweenness* – визначається кількістю геодезичних (найкоротших шляхів), що проходять через вершину чи ребро.

На рис. 5.1 показано виявлені спільноти користувачів для аналізованої підмножини твітів. На рис. 5.2 наведено підграф користувачів високоізольованих спільнот. На рис. 5.3 наведено гістограму користувачів за характеристикою Authority. На рис. 5.4 показано топові ключові слова для твітів, які містять хештег 'teslasolarissues'. Цей хештег пов'язаний із комплексом проблем, які виникли із сонячними батареями, виготовленими компанією *Tesla*. Актуальним є розгляд відображення трендів, пов'язаних із даною тематикою на різні процеси, зокрема, на динаміку цін акцій компанії на фінансовому ринку. Для аналізу було виділено задане тематичне поле лексем, частотний розподіл якого наведено на рис. 5.5. Використовуючи теорію частих множин та асоційованих правил, можна знайти семантичну структуру у визначених семантичних полях лексем. На рис. 5.6, 5.7 показано семантичні часті множини для різних заданих тем, пов'язаних з

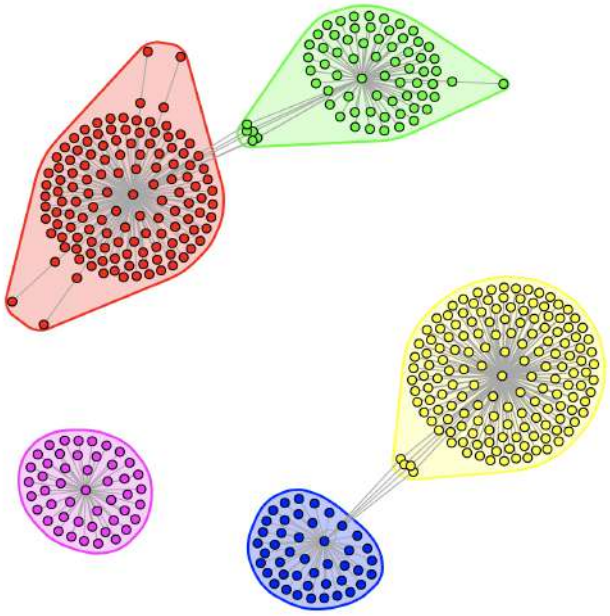

Рисунок 5.2 – Структура користувачів високоізольованих спільнот

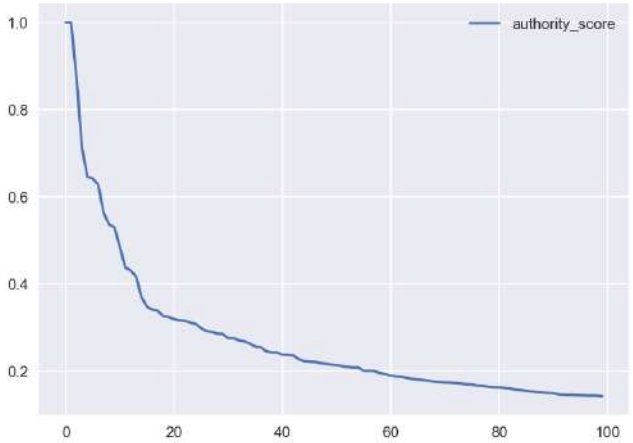

Рисунок 5.3 – Гістограма користувачів за характеристикою Authority

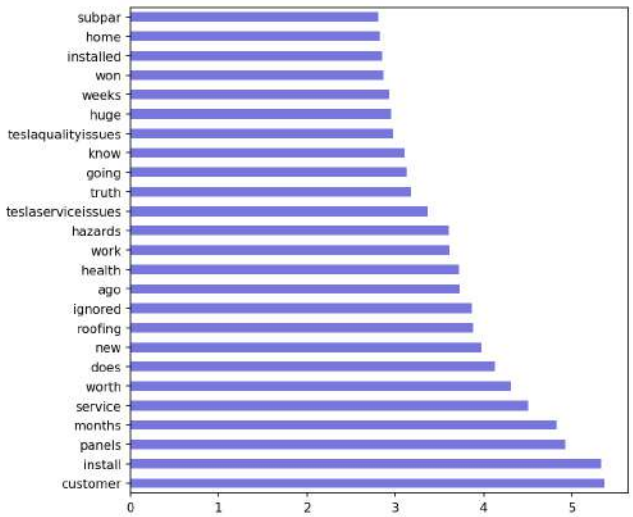

Рисунок 5.4 – Топові ключові слова для твітів, які містять хештег 'teslasolarissues'

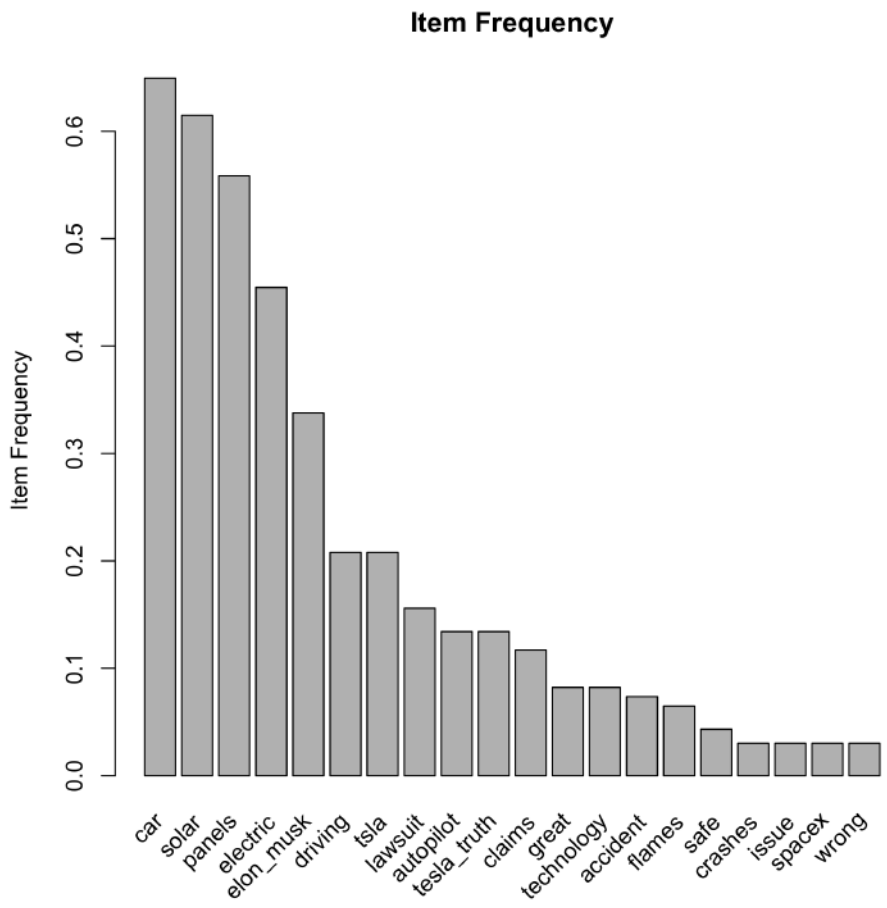

Рисунок 5.5 – Частотний розподіл лексем заданого тематичного поля

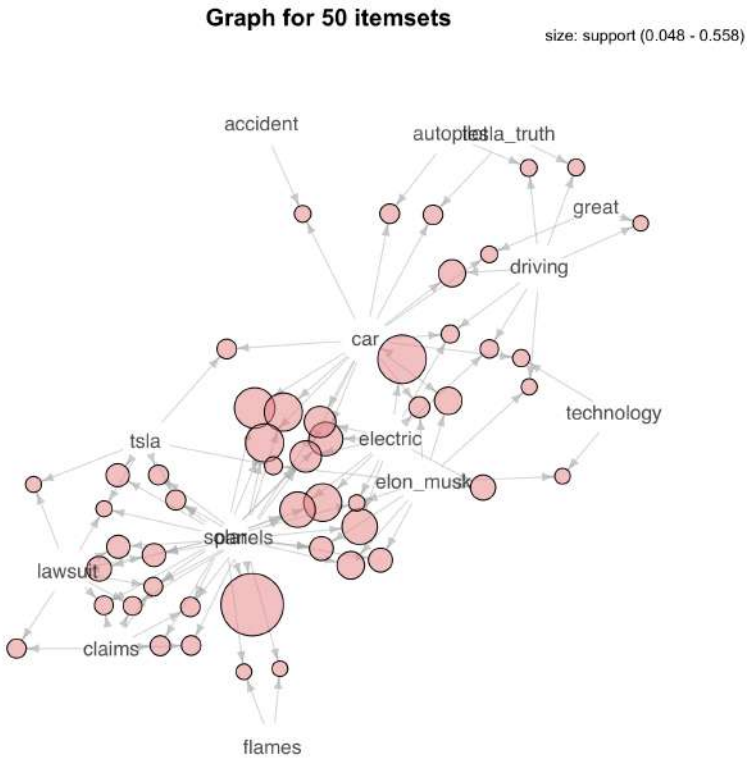

Рисунок 5.6 – Семантичні часті множини

компанією *Tesla*. На рис. 5.8, 5.9 показано асоціативні правила, відображені за допомогою графу та згрупованої матриці. Використовуючи структуру графів, семантичну структуру та ключові слова, пов'язані з темами та хештегами, що обговорюються користувачами, можна отримувати часовий ряд ключових слів для підрахунків твітів за день. Ці часові ряди можна розглядати як ознаки у прогнозних моделях. На рис. 5.10 показано аналітичні характеристики настрою та особистості, отримані за допомогою аналітичного хмаркового сервісу IBM Watson Personality Insights [329]. На рис. 5.11 показано часові ряди для ключових слів та хештегів. На рис. 5.12 показано нормалізовані часові ряди ключових слів. Можна побачити, що в часові моменти, коли стався інцидент із сонячними панелями компанії *Tesla*, активність твітів зростає за часовими рядами деяких ключових слів. Проаналізуємо як цей інцидент впливає на ціну акцій *Tesla*. На рис. 5.13 показано динаміку ціни акцій компанії *Tesla* (тікер TSLA) на ринку акцій. Соціальні мережі впливають на формування інвестиційних настроїв потенційних учасників фондового ринку.

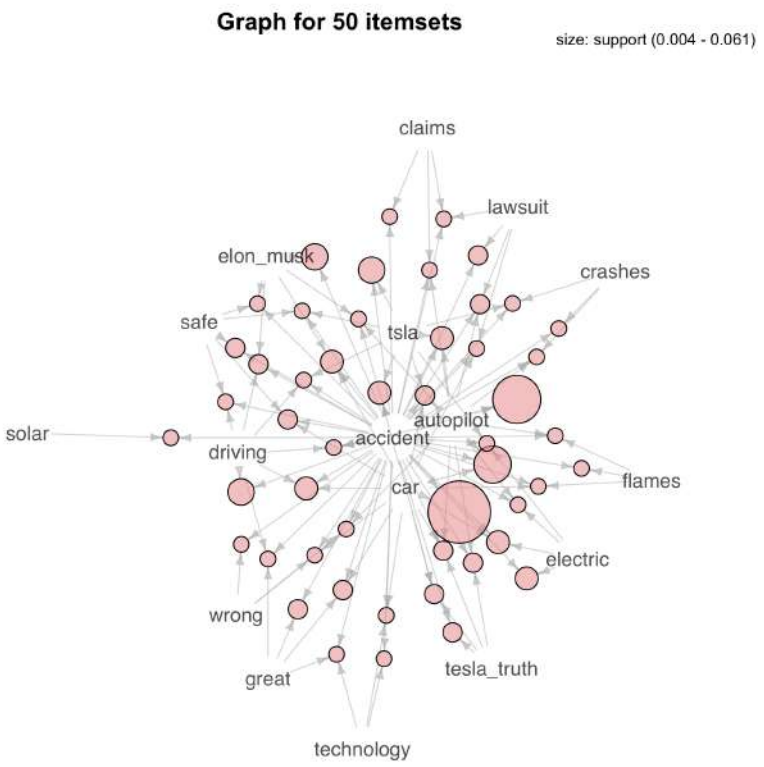

Рисунок 5.7 – Семантичні часті множини

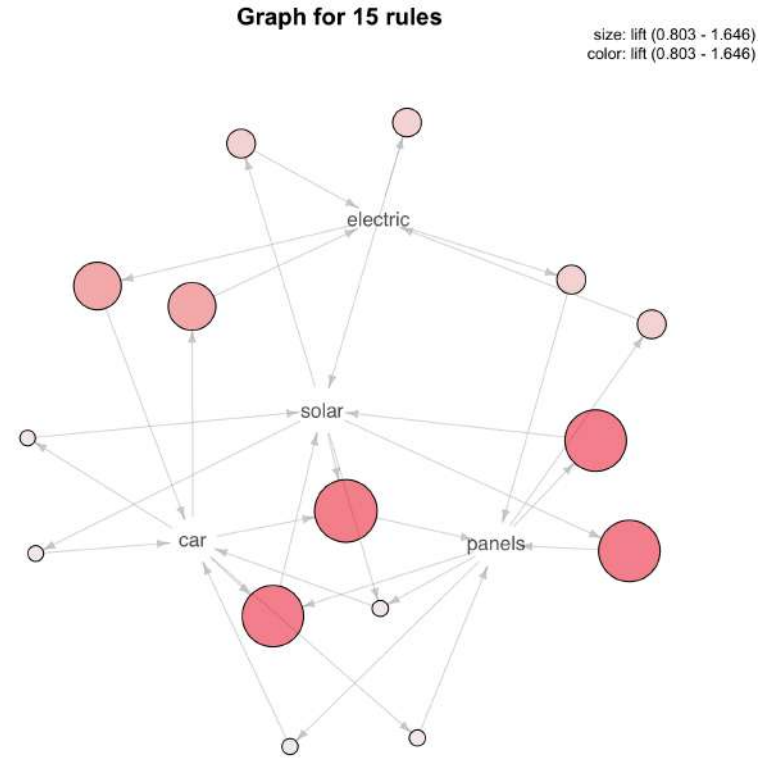

Рисунок 5.8 – Асоціативні правила, відображені за допомогою графу

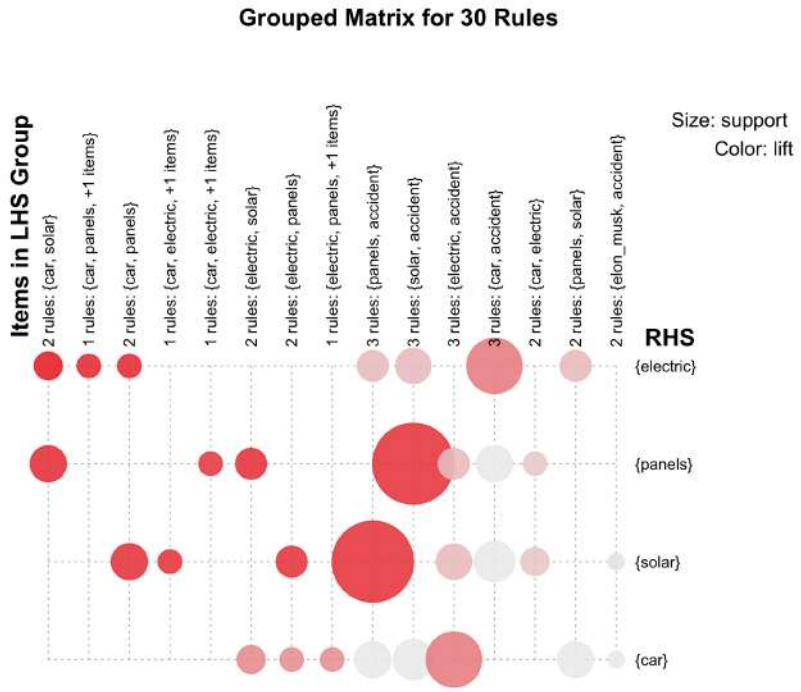

Рисунок 5.9 – Асоціативні правила, відображені за допомогою згрупованої матриці

Розглянемо динаміку акцій компанії *Тесла* в часовий період інциденту із сонячними батареями. Створено лінійну модель, де часові ряди ключових слів та їх зміщені в часі значення (лаги) розглядалися як незалежні регресійні змінні. В якості цільової змінної ми розглядали часовий ряд відносної зміни ціни протягом дня (price return). Використовуючи регресію LASSO, знайдено вагові коефіцієнти для аналізованих ознак. На рис. 5.14 показано реальну відносну зміну ціни акції та її прогнозовані значення. На рис. 5.15 показано регресійні коефіцієнти аналізованих ознак прогнозної моделі. Також проведено регресію, використовуючи підхід на основі байєсівського виведення. Такий підхід дозволяє розрахувати розподіли для параметрів моделі та цільової змінної, що є важливим у задачах оцінки ризиків. За допомогою такого підходу можна враховувати негаусовий розподіл цільової змінної, що є характерним для багатьох часових рядів у фінансовій сфері. На рис. 5.16 показано коробкові графіки для коефіцієнтів назалежних змінних у байесівській регресійній моделі. Актуальним є аналіз можливості використання Q-навчання для пошуку оптимальної стратегії на

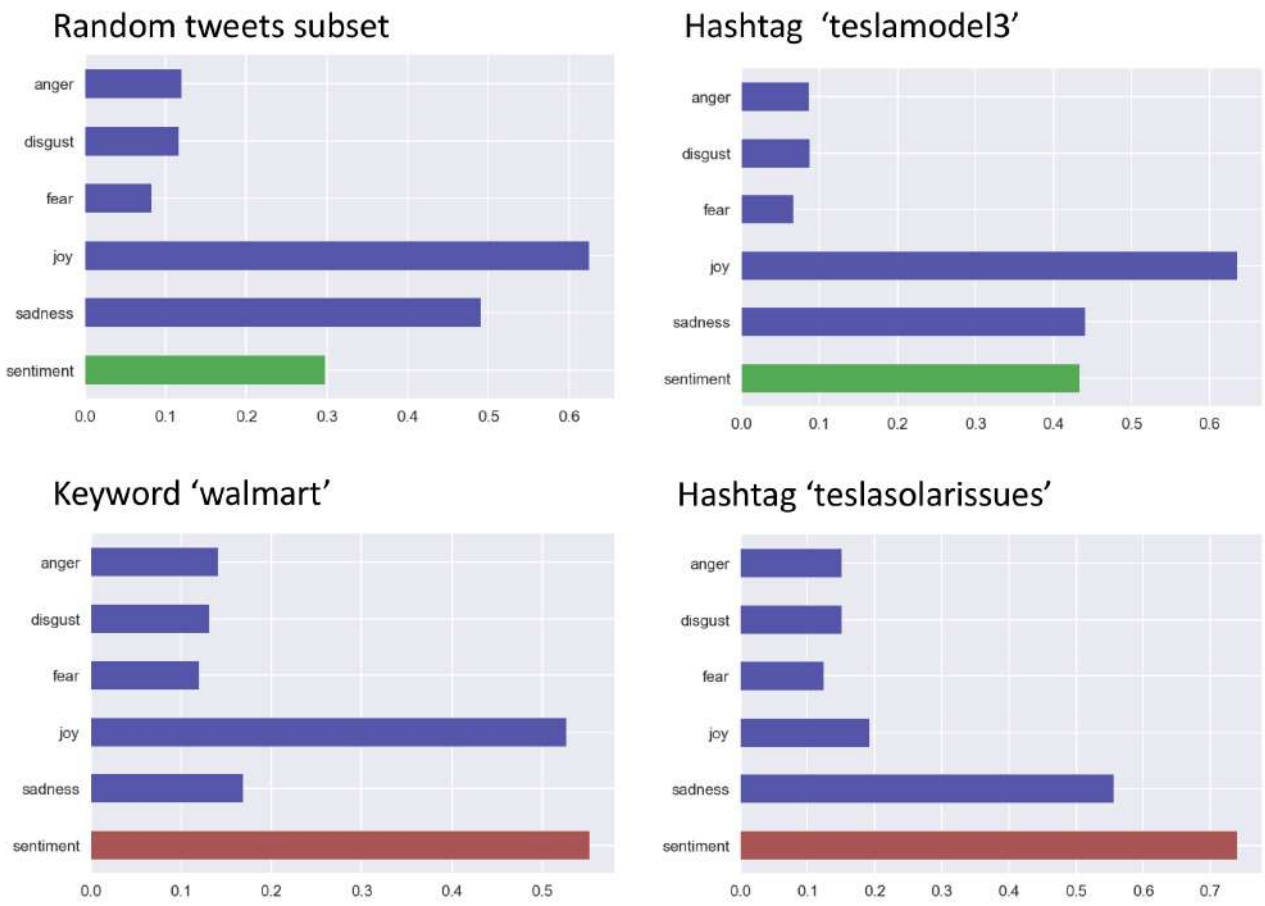

Рисунок 5.10 – Аналітичні характеристики настрою (sentiment) та особистості (personality)

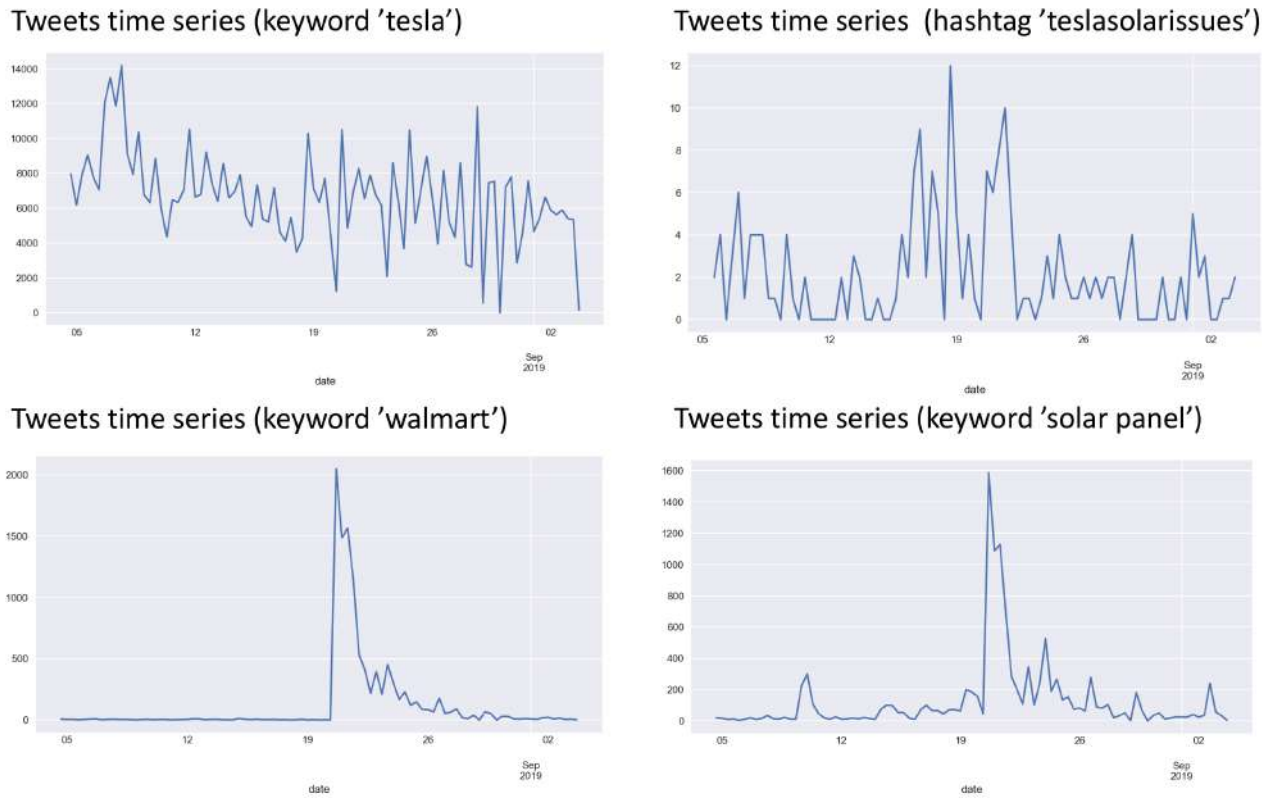

Рисунок 5.11 – Часові ряди для ключових слів та хештегів

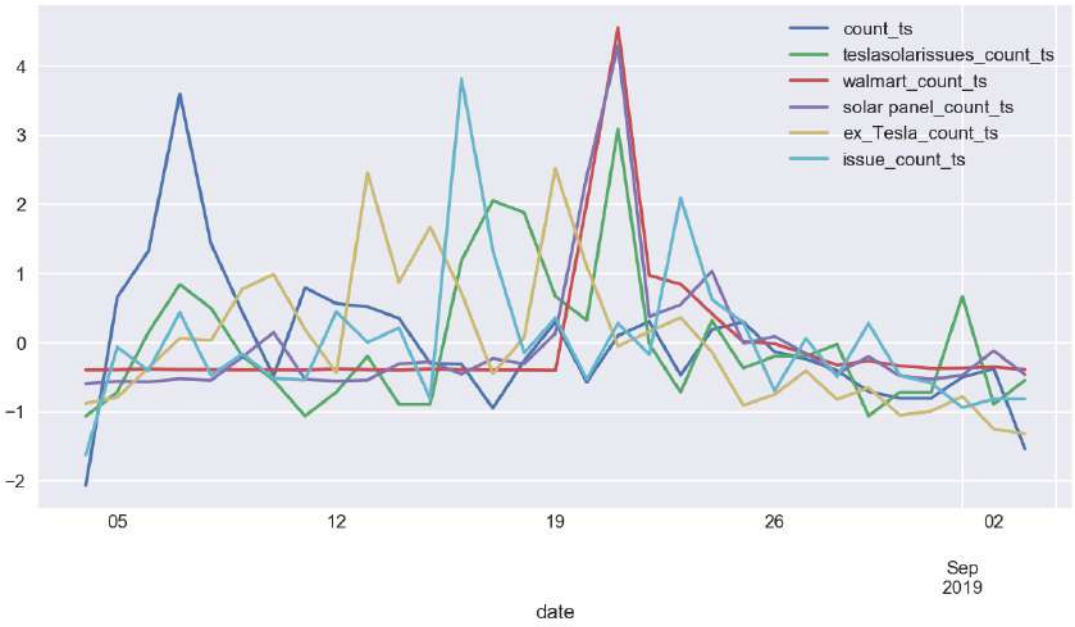

Рисунок 5.12 – Нормалізовані часові ряди ключових слів

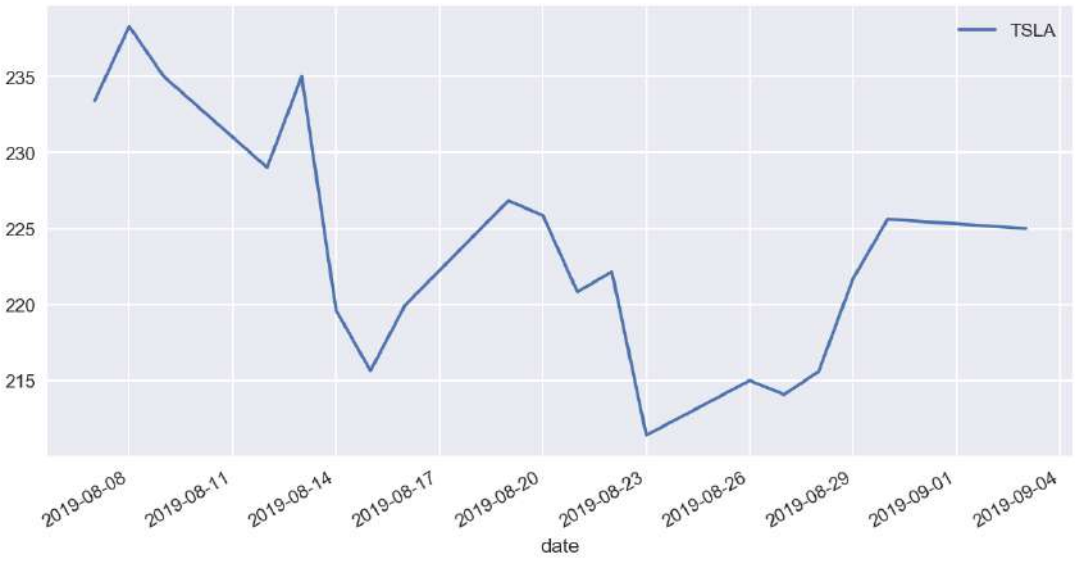

Рисунок 5.13 – Динаміка ціни акцій компанії Тесла (тікер TSLA)

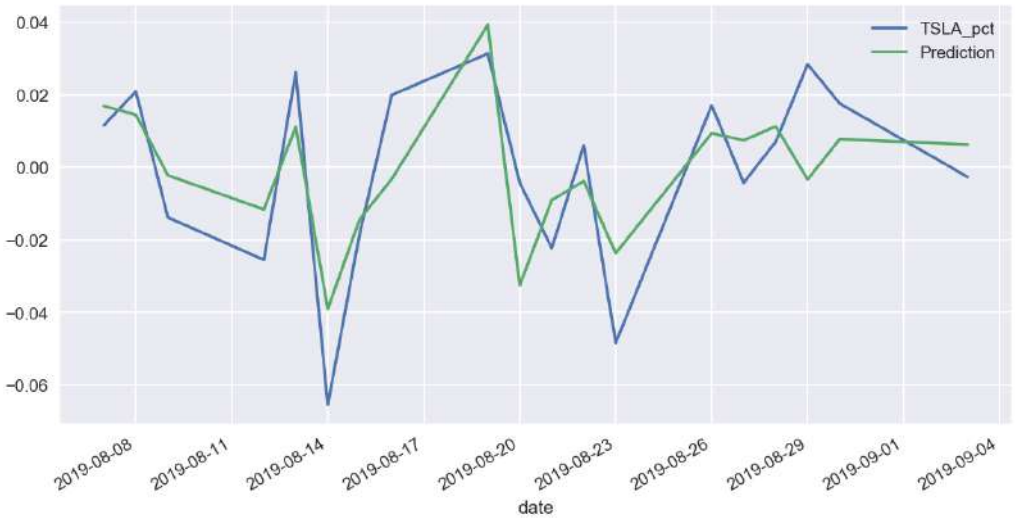

Рисунок 5.14 – Відносна зміна ціни акції та її прогнозовані значення

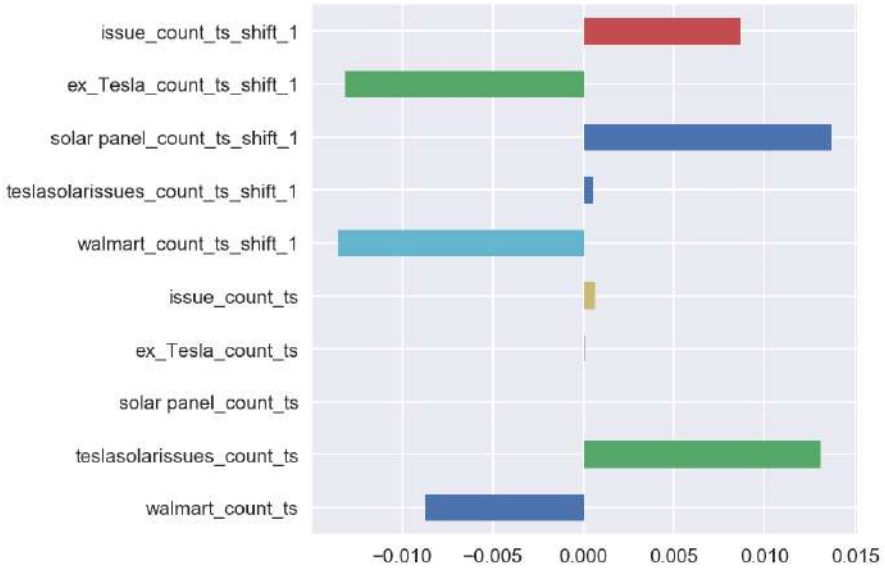

Рисунок 5.15 – Регресійні коефіцієнти аналізованих ознак прогнозної моделі

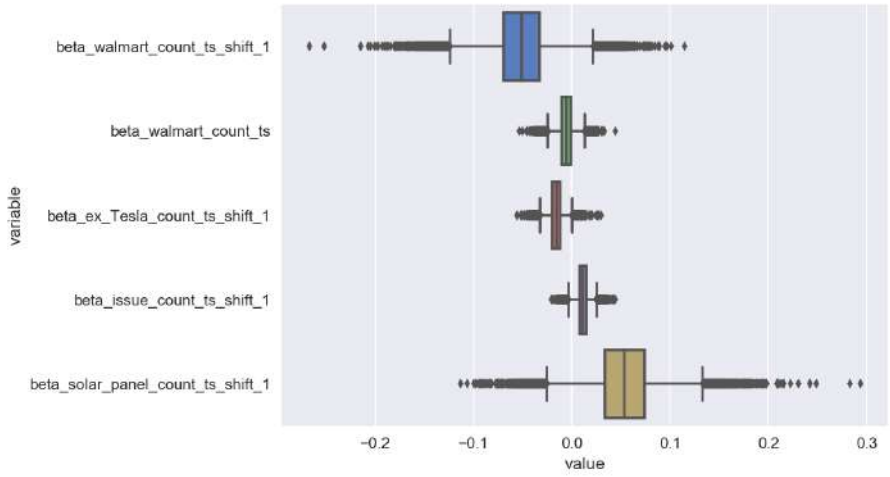

Рисунок 5.16 – Коробкові графіки для коефіцієнтів назалежних змінних у байесівській регресійній моделі

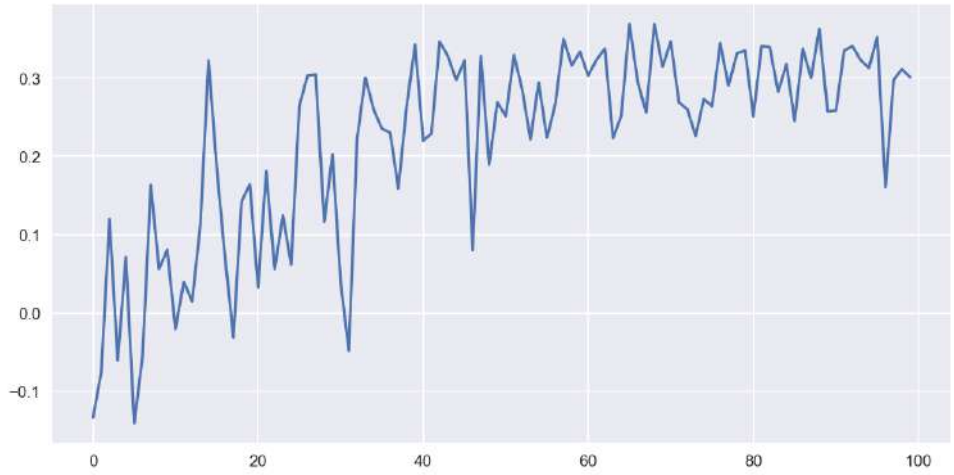

Рисунок 5.17 – Винагорода на ітераційних епізодах взаємодії інтелектуального агента із середовищем

ринку акцій. У найпростішому випадку використання глибокого Q-навчання (Deep Q-Learning, DQN) можна застосувати три дії "купити", "продати", "утримувати". Як ознаки стану інтелектуального агента використовувались часові ряди ключових слів у твітах, як винагороду використано відносну зміну ціни акції. Середовище інтелектуального агента моделювалось за допомогою історичних даних ключових слів та відносної зміни ціни акції. На рис. 5.17 показано відносну зміну ціни акції на ітераційних епізодах взаємодії інтелектуального агента із середовищем. Отримані результати показують, що інтелектуальний агент може знайти оптимальну прибуткову стратегію. Звичайно, це дуже спрощений випадок аналізу, у якому може виникнути ефект перенавчання, тому такий підхід потребує подальшого дослідження. Основна ціль – показати, що, використовуючи навчання із підкріпленням та модель середовища на основі історичних фінансових даних та кількісних характеристик текстових повідомлень Твіттера, можна побудувати таку модель, у якій інтелектуальний агент зможе знайти оптимальну стратегію, яка оптимізує функцію винагороди на епізодах інтерактивної взаємодії агента із середовищем.

5.3 Методи прогнозування подій на основі інтелектуального аналізу повідомлень мікроблогів Твіттера

Система мікроблогів Твіттер (Twitter) є одним із популярних засобів взаємодії користувачів за допомогою коротких повідомлень (не більше 140 символів). Для повідомлень Твіттера характерна висока густина тематично значимих ключових слів. Ця особливість зумовлює перспективність досліджень мікроблогів засобами інтелектуального аналізу та актуальність розвитку методів інтелектуального аналізу текстових повідомлень для виявлення семантичних зв'язків між основними поняттями та тематиками обговорень у мікроблогах. Перспективним є аналіз прогностичної ефективності часових залежностей квантитативних характеристик ключових тегів та тематичних концептів у повідомленнях мікроблогів твіттер. Особливості соціальних мереж та поведінки користувачів досліджуються у багатьох роботах. Низку робіт присвячено аналізу можливості передбачати події на основі аналізу повідомлень у мікроблогах. Повідомлення Твіттера

характеризуються високою щільністю контекстно-значущих ключових слів. Це дає можливість вивчати мікроблоги за допомогою інтелектуального аналізу даних з метою виявлення семантичних зв'язків між основними поняттями та предметами обговорення в мікроблогах. Особливості соціальних мереж та поведінки користувачів розглядаються у багатьох дослідженнях. У [330, 331] досліджено явища мікроблогінгу. У роботі [332] показано, що соціальні мережі структурно відрізняються від інших типів мереж. Вплив користувачів у мережі Twitter вивчався у [333]. Поведінка користувачів у соціальних мережах проаналізована в [334]. В [335] проаналізовано методи виявлення позицій користувачів у масиві повідомлень Твіттера. Є також дослідження можливого прогнозування подій шляхом аналізу повідомлень у мікроблогах. У [35] досліджено кореляції настроїв користувачів із процесами на фондових ринках. У роботі [336] показано, що проста модель на основі динаміки твітів дає можливість будувати прогнози поведінки ринку акцій. У [337] проаналізовано динаміку продажу фільмів на основі аналізу дискусій у мікроблогах. У роботі [338] досліджувалася активність мікроблогів Твіттера під час медійних подій. Перспективним є аналіз часових залежностей ключових кількісних характеристик тематичних понять у повідомленнях мікроблогів Твіттера.

Проведемо інтелектуальний аналіз повідомлень Твіттера на прикладі обговорень подій різних типів, зокрема прогнозування переможців на пісенному конкурсі Євробачення 2013 та обговорення імені новонародженого британського принца у 2013 році. Дослідження проведемо, використовуючи теорію частих множин та асоціативних правил. Проаналізуємо динаміку підтримки та достовірності у виявлених частих множинах та асоціативних правилах.

Розглянемо випадок прогнозування Євробачення 2013. У шведському місті Малмо, 18 травня 2013 року пройшов конкурс Євробачення 2013. Прогнозування результатів голосування з визначення переможця є цікавим тому, що виконавці пісень широко обговорювались у соціальних медіа, зокрема, у Твіттері. Крім того, можна припустити, що активні користувачі Твіттера, які брали участь у обговоренні, також будуть активно голосувати відповідно до своїх поглядів, які вони виражають у своїх мікроблогах.

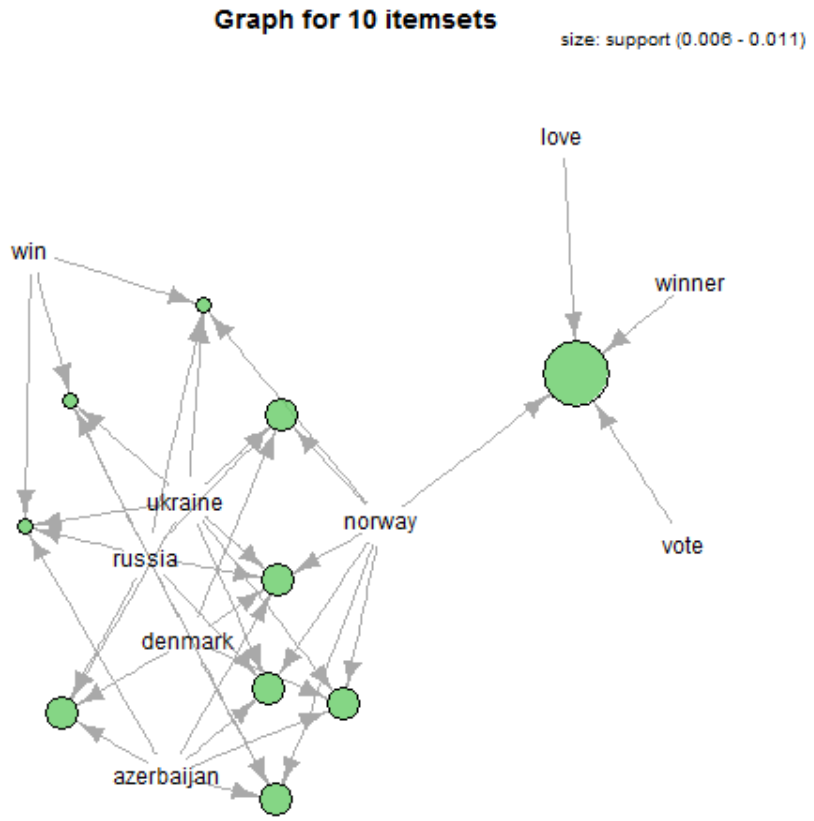

Рисунок 5.18 – Граф для частих множин із найбільшими значеннями підтримки.

Тому можна очікувати, що семантична структура твітів із обговоренням конкурсу Євробачення буде відображатись у результатах голосування. Для прогнозування фіналу конкурсу Євробачення 2013 за день до фінального конкурсу 17 травня ми завантажили твіти з ключовими словами *"eurovision win"*, *"eurovision winner"*. Всього 2400 твітів. Аналіз було проведено у середовищі статистичних розрахунків R. Твіти завантажувались із використанням пакету *"twitteR"* [339], аналіз частих множин та асоціативних правил проведено за допомогою пакетів *"arules"* [340], *"arulesViz"* [341]. Було відфільтровано малоінформативні стоп слова. Далі було утворено часті множини ключових слів, із яких були відібрані ті, що мають семантичне відношення до фіналу Євробачення. На рис. 5.18 наведено граф для частих множин із найбільшою підтримкою. На основі відібраних частих множин утворено асоціативні правила. Підтримку та достовірність асоціативних правил було визначено експериментальним шляхом так, щоб утворилась множина асоціативних правил, права і ліва частина яких відображала б

семантичні поняття, які мають відношення до аналізованої події. Зокрема, це назви країн, ключові слова *"win", "winner", "favorite", "vote", "love"* etc. Отримані асоціативні правила було відсортовано за величиною підтримки. У таблиці 5.2 наведено приклади отриманих асоціативних правил із найбільшими значеннями підтримки у лівій частині асоціативного правила. Розглянемо графічні представлення асоціативних правил. Для

Таблиця 5.2 – Асоціативні правила з двома ключовими словами у лівій частині правила.

	Antecedent	consequent	support	confidence	lift
1	denmark, norway	win	0.01461	0.900	1.314
2	denmark, favourites	win	0.01136	1.000	1.460
3	azerbaijan, norway	win	0.01136	0.875	1.277
4	denmark, ukraine	win	0.00812	0.833	1.216
5	azerbaijan, russia	win	0.00812	0.833	1.216
6	azerbaijan, denmark	win	0.00812	0.714	1.043
7	finland, sweden	win	0.00812	1.000	1.460
8	russia, ukraine	win	0.00649	0.800	1.168
9	azerbaijan, ukraine	win	0.00649	0.800	1.168
10	norway, ukraine	win	0.00649	0.8000	1.168

побудови графічного відображення асоціативних правил використано пакет *"arulesViz"* [341] для мови програмування R. На рис. 5.19 зображено асоціативні правила з найбільшими значеннями підтримки. На рис. 5.20 зображено утворення асоціативних правил на основі ключових слів. На рис. 5.21 наведено графічне зображення матриці згрупованих асоціативних правил. На основі отриманих даних можна зробити висновок, що лідером серед претендентів на перемогу була Данія. Серед фаворитів наступні три місця належали Україні, Росії та Ірландії. Реальні результати змагань такі: перше місце у Данії, друге місце в Азербайджану, третє місце – в України, четверте місце – у Норвегії, п'яте місце – у Росії. Як випливає із результатів дослідження та з реальних результатів фіналу Євробачення 2013, проведений інтелектуальний аналіз твітів за день до фіналу дав точний прогноз для переможця Євробачення та коректно визначив країни у верхній частині рейтингу за результатами голосувань телеглядачів країн учасниць пісенного конкурсу. Особливість проаналізованого типу подій на

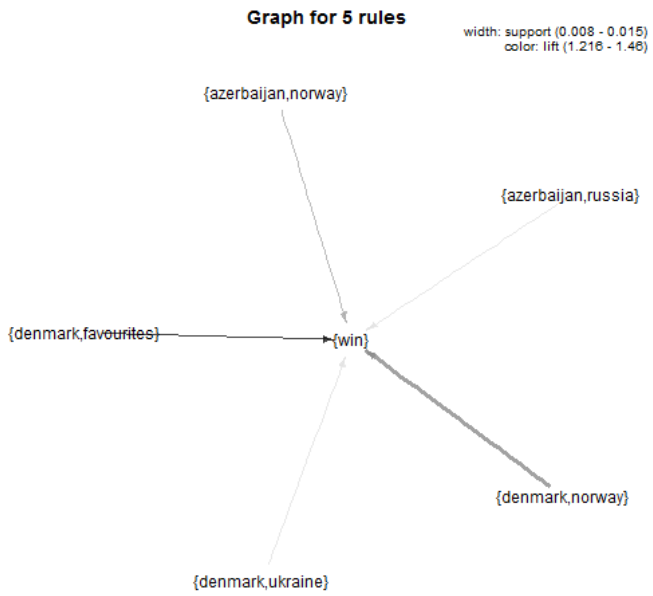

Рисунок 5.19 – Асоціативні правила з найбільшими значеннями підтримки.

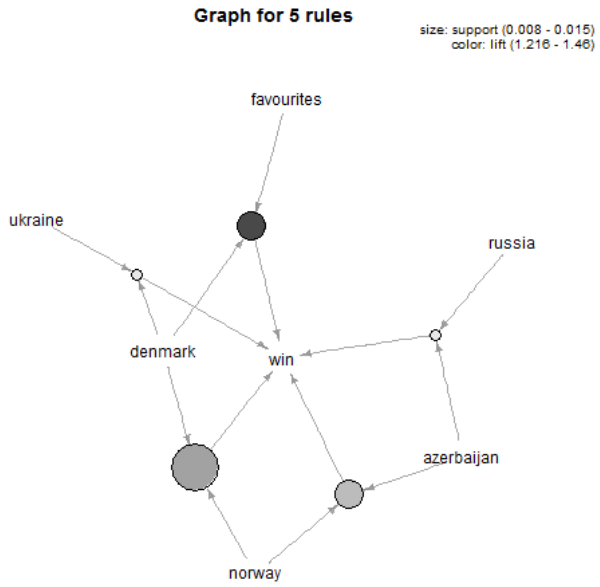

Рисунок 5.20 – Утворення асоціативних правил на основі ключових слів.

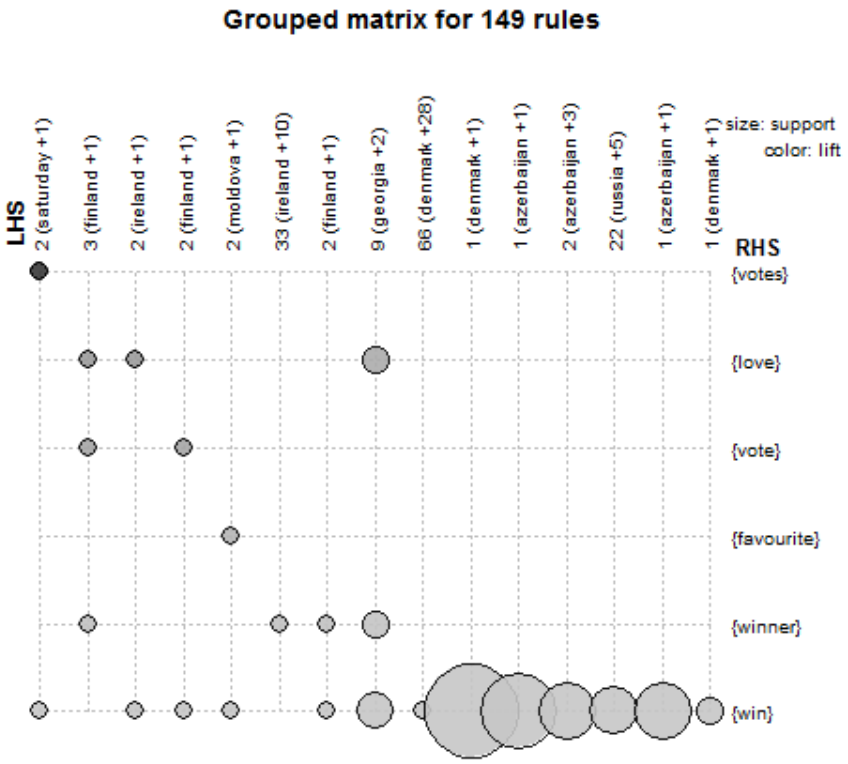

Рисунок 5.21 – Матриця згрупованих асоціативних правил.

прикладі пісенного конкурсу Євробачення зумовлена тим, що користувачі, які обговорюють учасників конкурсу, репрезентують групу тих, хто активно бере участь у телефонному голосуванні. Відмінність прогнозування цього типу подій від, наприклад, спортивних подій полягає у суттєво меншому впливу випадкових факторів, оскільки учасники конкурсу та їх пісні відомі наперед і в учасників блогосфери вже є власна сформована оцінка учасників конкурсу, яка малоймовірно зміниться, у той час як спортивний результат із суттєво більшою ймовірністю може бути відмінним від очікувань. Тому для цього типу подій характерна висока кореляція між результатами події та результатами попереднього інтелектуального аналізу твітів засобами інтелектуального аналізу текстів [342].

Розглянемо приклад можливої кореляції між суспільною думкою користувачів Твіттера і прийняттям рішень впливових у суспільстві осіб. Однією з подій літа 2013 року, яка широко обговорювалась у соціальних мережах, було народження британського принца. У соціальних мережах, зокрема, велись дискусії щодо можливого імені королівського немовляти. Це не є серйозною науковою проблемою, якщо її розглядати буквально. Головна мета цього дослідження – дослідити можливу кореляцію між висловлюваннями користувачів соціальної мережі та прийняттям рішень особами, які є впливовими у певних сферах суспільства. Проаналізуємо наявність можливого зв'язку між суспільною думкою користувачів Твіттера та прийняттям рішень особами, які мають вагу у суспільстві. Цей аналіз проведемо на прикладі обговорень можливого імені народженого у липні 2013 року британського принца. Розглянемо послідовність аналізу. Ми завантажували твіти з каналів в період з 19 по 25 липня 2013 року. Для виділення твітів, які мали відношення до аналізованої тематики імені новонародженого наслідного принца, ми використовували такі ключові слова як "#RoyalBaby", "#RoyalBabyWatch", "#Royals", "#RoyalBaby-Name", "#goodluckKate", "Kate Middleton". Для завантаження твітів ми використали Twiitter API та Python пакет python-twiter. Завантажені дані записувались у базу даних SQLite. Аналіз проведено у середовищі статистичних розрахунків R з використанням додаткових пакетів, зокрема, *xts, tm, arules, "arulesViz", RSQLite, igraph*. На основі отриманих даних ми

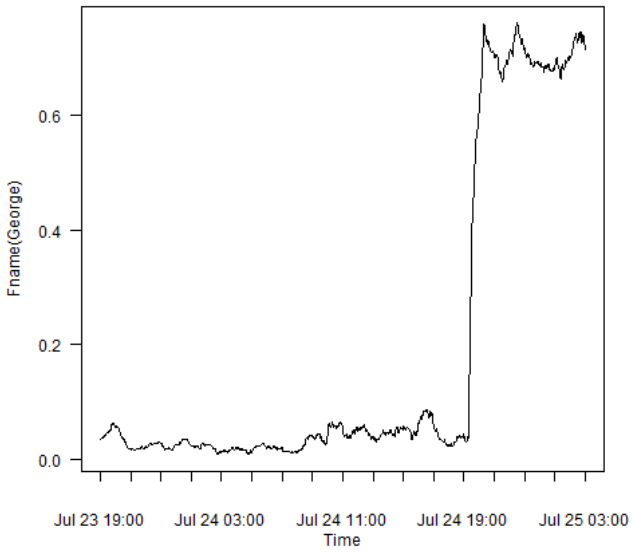

Рисунок 5.22 – Часова динаміка частоти твітів із іменем George

виділили масив твітів із ключовим словом "name". Далі утворили частотний словник, із якого було видалено стоп-слова. Аналізуючи частотний словник ми виділили лексеми, які позначають імена. Множина потенційно можливих імен має такий вигляд:

Snames ={"George", "Philip", "Spencer", "Stuart","Edward","John", "james", "arthur", "freddy", "henry", "alexander", "charles", "joffrey", "boris", "rudiger", "richard", "joseph", "harry","michael", "albert", "andrew", "louis"}

Відомо, що ім'я британського принца – Prince George of Cambridge. Повне ім'я наслідного принца – George Alexander Louis. Розглянемо часову динаміку імені George. Як кількісну характеристику розглянуто частоту твітів із заданим ключовим іменем, розрахунок якої описано в [343]. На рис. 5.22 наведено часову динаміку частоти твітів, які містять ім'я George. Час наведено у стандарті UTC. На цьому рисунку спостерігається різкий скачок, який очевидно відповідає часовому моменту оголошення імені новонародженого принца. Часовий скачок відбувається у часовий період з 19:20 по 19:30 24 липня 2013 року. Розглянемо масив твітів, який відповідає часовому періоду перед цим різким скачком. Для аналізу виберемо твіти, які було надіслано перед 18:00 24 липня, коли ім'я принца було ще невідомим. До розгляду було взято 22411 твітів, які було завантажено

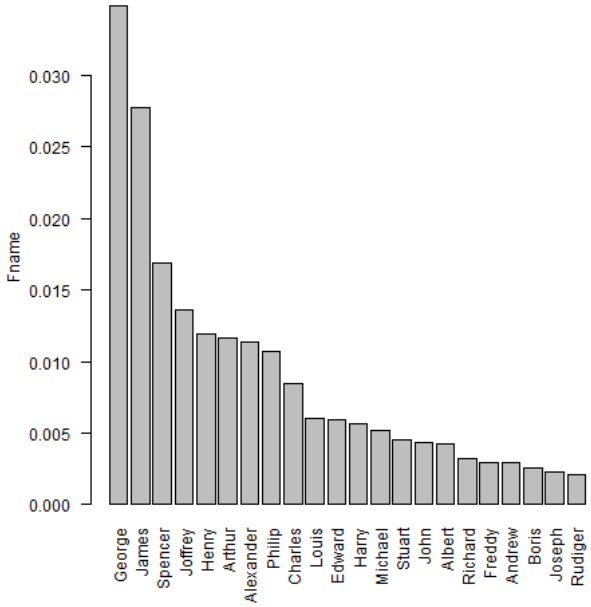

Рисунок 5.23 – Розподіл імен у твітах у період до оголошення імені принца

перед згаданим скачком, починаючи із 19 липня, і які містили хоча б одне із згаданих імен, а також ключове слово 'name'. На рис. 5.23 наведено частоти твітів із аналізованими іменами, які входять у множину аналізованих імен у період перед офіційним оголошенням імені принца. Імена наведено у порядку спадання частоти Fname. Як випливає із отриманого графіку, лідирує ім'я George. Проаналізуємо чи можна було б передбачити повне ім'я, яке має три складових імені. Розглянемо імена, які зустрічались у твітах одночасно. Для аналізу використаємо елементи теорії частих множин. Розглянемо в аналізованому масиві твіти, у яких існують часті множини із трьох таких ключових слів, які належать до множини аналізованих імен. Використовуючи алгоритм a priori [171, 344], було виявлено часті множини лексем, які позначають потенційні імена у масиві твітів. Часті множини з найбільшим значенням підтримки наведено у таблиці 5.3 Як випливає з отриманих даних, часта множина лексем, у яку входять три імені принца {George, Alexander, Louis}, є у перших п'ятьох частих множинах, які впорядковані за величиною підтримки. Якщо говорити про два складових імені принца, то вони входять у першу трійку списку частих множин. На рис. 5.24 наведено граф утворення перших десяти частих множин за величиною підтримки. Із п'яти імен, які входять

Таблиця 5.3 – Часті множини із найбільшим значенням підтримки

#	Часта множина	Підтримка
1	alexander, george, james	0.121
2	george, henry, james	0.107
3	george, james, louis	0.090
4	alexander, james, louis	0.085
5	alexander, george, louis	0.085
6	george, henry, louis	0.080
7	alexander, henry, james	0.077
8	alexander, george, henry	0.077
9	alexander, henry, louis	0.075
10	henry, james, louis	0.075
11	arthur, george, james	0.019
12	charles, james, spencer	0.016
13	george, james, spencer	0.015
14	charles, george, philip	0.014
15	george, james, richard	0.012

у перших 10 частих множин, три імені є складовими іменами принца. Як випливає із отриманих даних, використання теорії частих множин дає можливість отримати точніший прогноз для повного імені у порівнянні з отриманим частотним списком імен, який дає можливість спрогнозувати лише основне ім'я. Розглянемо структуру множини користувачів, які взяли участь у обговоренні імені принца. Для виявлення спільнот, які динамічно сформувалися в аналізованій дискусії ми використали швидкий жадібний алгоритм оптимізації модулярності(fast greedy modularity optimization algorithm) для знаходження спільнот, описаний у [345] . Для побудови графу було використано алгоритм Fruchterman-Reingold [328]. Цей алгоритм належить до силових алгоритмів, які ще називають пружинними. Характер графу зумовлений моделлю, яка використовується у силових алгоритмах. Особливістю моделі є те, що вершини розглядають як кульки між якими діють сили відштовхування, а ребра розглядають як моделі пружин, що притягують з'єднані цими ребрами вершини. У масиві твітів знайдено 6919 користувачів, які надіслали 37191 твітів. У цих твітах згадано 2645 користувачів. Суттєва частина згадок інших користувачів пов'язана із ретвітами. У наших дослідженнми ми не розділяємо ретвіти

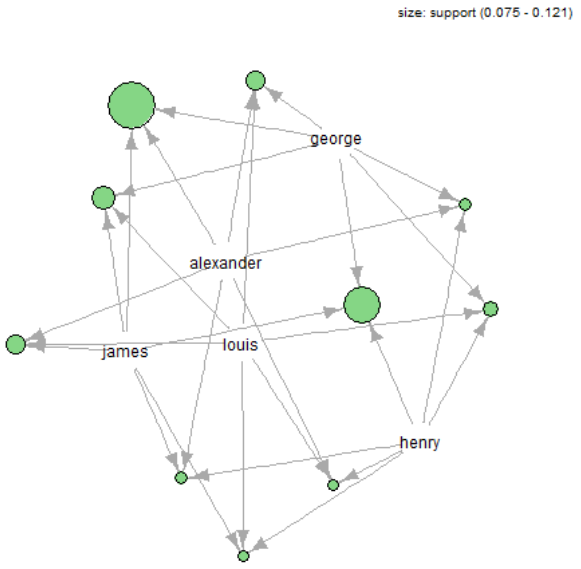

Рисунок 5.24 – Граф утворення перших десяти частих множин за величиною підтримки

від інших типів твітів, у яких згадуються користувачі. Для подальшого розгляду візьмемо активних користувачів, які в процесі обговорення надіслали більше одного твіта, або були згадані у твітах більше одного разу. Активних користувачів, які надіслали більше одного твіта, виявлено 2300, а користувачів, яких згадали у твітах більше одного разу, було 923. На рис. 5.25 наведено побудований граф зв'язків між користувачами, на якому відтінками кольорів виділено виявлені спільноти користувачів. На отриманому графі спостерігається декілька багаточисельних спільнот користувачів. Проведено аналіз із вилученням із розгляду найбільш популярних користувачів, які згадувались у твітах 100 і більше разів. Таких користувачів виявлено лише 6. Видаливши цих користувачів із розгляду, отримано граф із виявленими спільнотами, який зображено на рис. 5.26. Видалені користувачі складають близько 0.2% від згаданих у твітах користувачів. Як випливає із отриманих даних, видалення із розгляду лише цих найпопулярніших користувачів суттєво змінює структуру спільнот, залишаючи лише численні дрібні спільноти. Отже, у результаті проведених досліджень, показано, що інтелектуальний аналіз твітів дає можливість спрогнозувати ім'я наслідного принца. Показано, що основне ім'я новонародженого принца George домінувало у спектрі імен перед офіційним

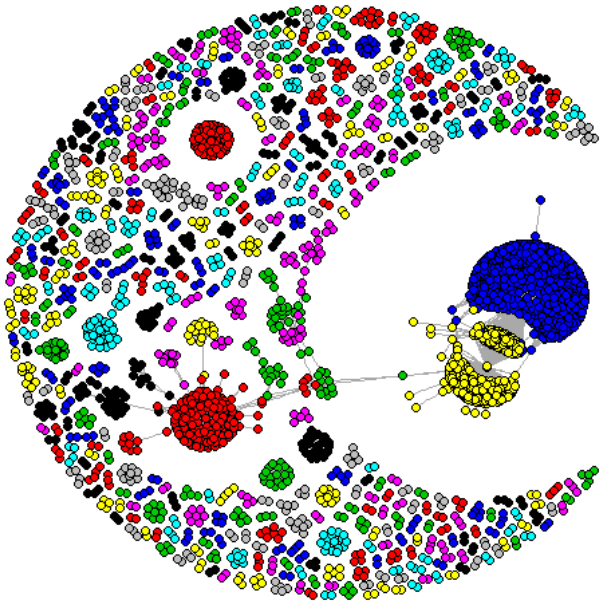

Рисунок 5.25 – Виявлені спільноти користувачів

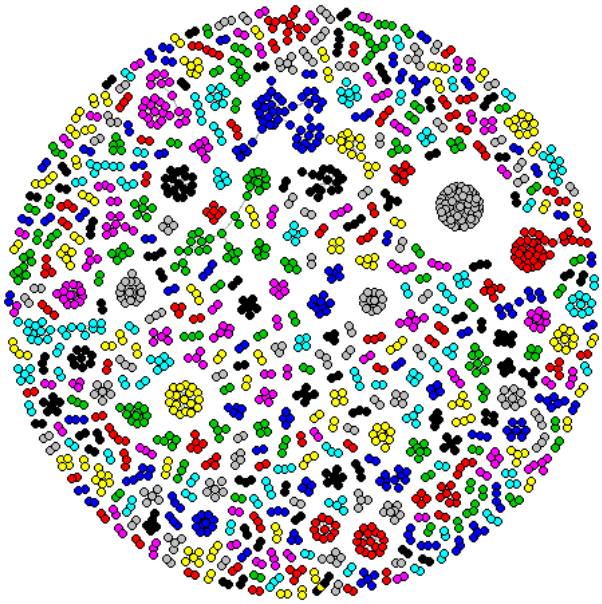

Рисунок 5.26 – Виявлені спільноти користувачів при вилученні шести найпопулярніших користувачів

оголошенням імені. Як випливає із отриманих даних, використання теорії частих множин дає можливість отримати точніший прогноз для повного імені у порівнянні з аналізом частотного ряду імен, який дає можливість спрогнозувати лише основне ім'я. Три складових імені принца George, Alexander, Louis утворюють часту множину лексем, яка входила у топ 5 частих множин за величиною підтримки. Показано, що структура динамічно утворених спільнот користувачів, які взяли участь у обговоренні, визначається, лише декількома лідерами, які мають суттєвий вплив на формування позиції інших користувачів. Проаналізована проблема не є серйозною, якщо її розглядати буквально. Основна ціль проведеного аналізу – дослідити чи існує кореляція між міркуваннями користувачів соціальної мережі та прийняттям рішень впливовими у певних сферах суспільства особами. Проведені дослідження показують, що така кореляція існує [343].

5.4 Аналіз повідомлень мережі Твіттер, пов'язаних із пандемією COVID-19

Проведемо інтелектуальний аналіз повідомлень мережі Твіттер, пов'язаних із COVID-19, використовуючи алгоритми аналізу графів, а також теорію частих множин та асоціативних правил [246]. COVID-19 активно обговорюється у соціальних мережах, тому характеристики повідомлень, зокрема у мережі Твіттер можуть мати прогнозні властивості. Розглянемо зв'язки між користувачами з точки зору теорії графів. Зв'язки користувачів можуть бути зображені графом, в якому вершини представляють користувачів, а ребра - їх зв'язки, зокрема на основі відповідей на повідомлення та ретвітів. Використовуючи алгоритми аналізу графів, можна виявляти спільноти користувачів та знаходити топових користувачів за різними оцінками вершин, зокрема такими оцінками як *Hub, Authority, PageRank, Betweenness.* Для виявлення спільнот використовувався алгоритм *Community Walktrap* з пакету *igraph* [326]. Для візуалізації використовувався алгоритм *Fruchterman-Reingold* із цього пакету. Для аналізу було взято твіти із ключовим словом *'coronavirus'* які завантажувались у січні 2020 року. На рис. 5.27 показано виявлені спільноти користувачів у повідомленнях Твіттера для заданої підмножини

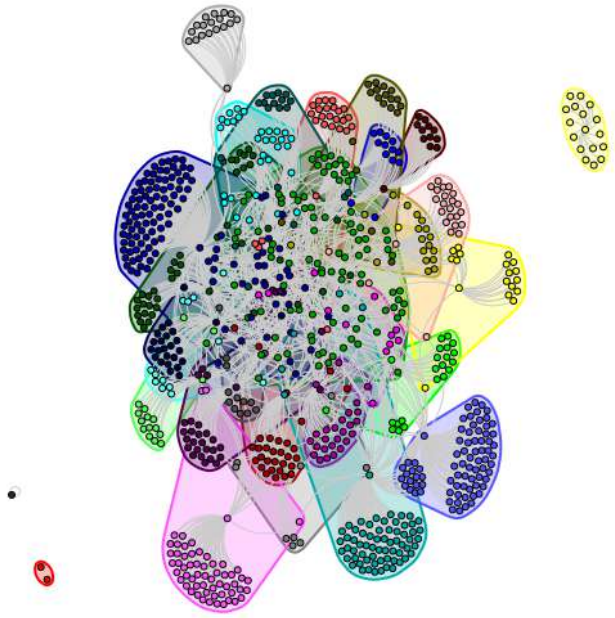

Рисунок 5.27 – Виявлені спільноти користувачів у повідомленнях Твіттера

твітів. Використовуючи теорію частих множин та асоціативних правил можна знайти семантичну структуру у масиві повідомлень в межах заданого семантичного поля. На рис. 5.28 показано частоти ключових слів. Деякі ключові слова, зокрема 'fear', позначають сумарну частоту семантично близьких слів. Сукупність ключових слів формує тематичне поле для інтелектуального аналізу твітів. На рис. 5.29-5.33 показано часті множини та асоціативні правила для різних семантичних структур.

Знайдені часті множини та асоціативні правила відображають семантичну структуру в масиві твітів, пов'язаних із COVID-19. Квантитативні характеристики таких частих множин, наприклад величина підтримки можуть бути використані як додаткові ознаки в задачах прогнозної аналітики.

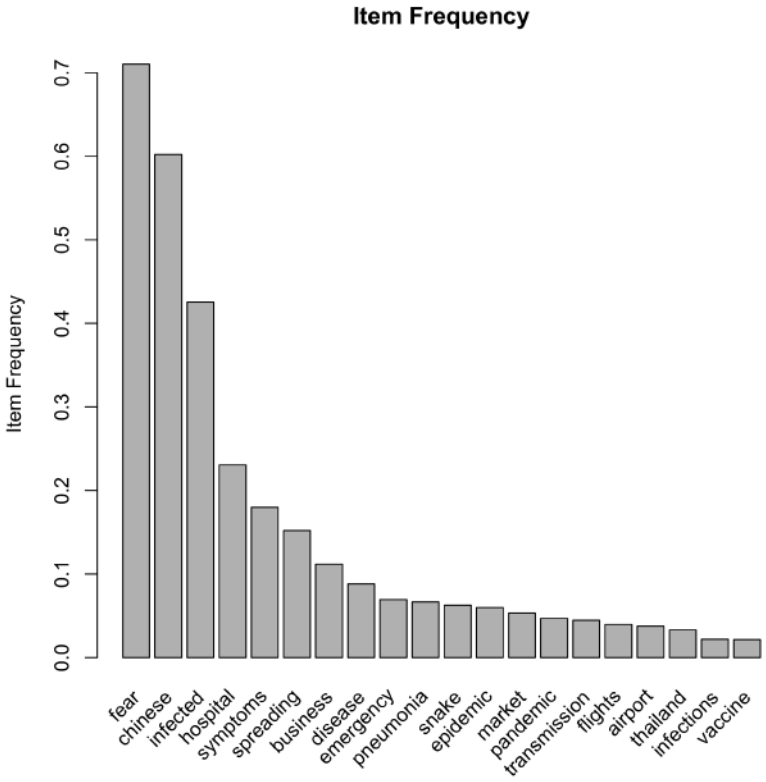

Рисунок 5.28 – Частоти ключових слів

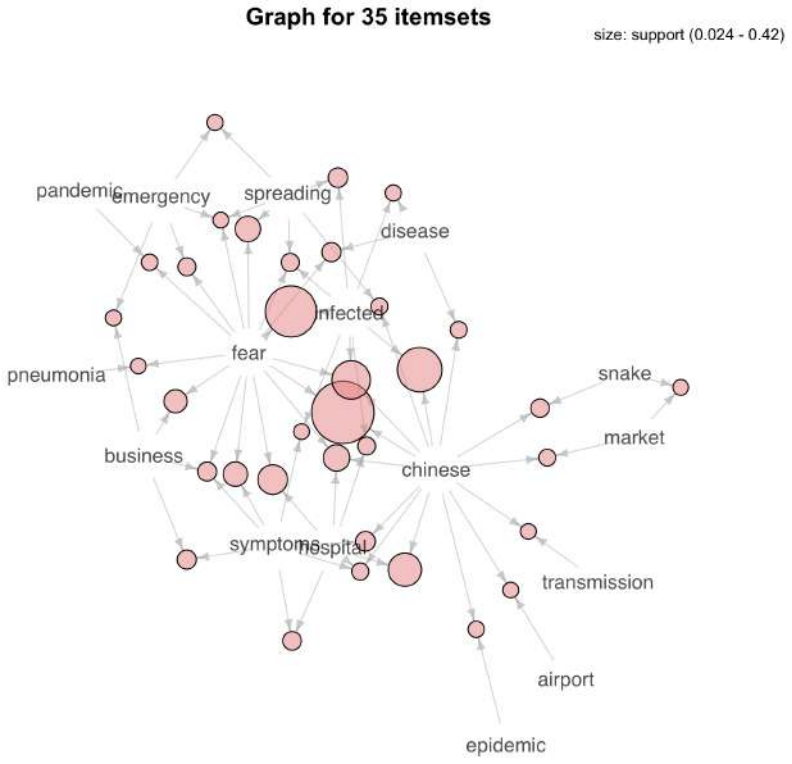

Рисунок 5.29 – Граф частих множин в заданому тематичному полі

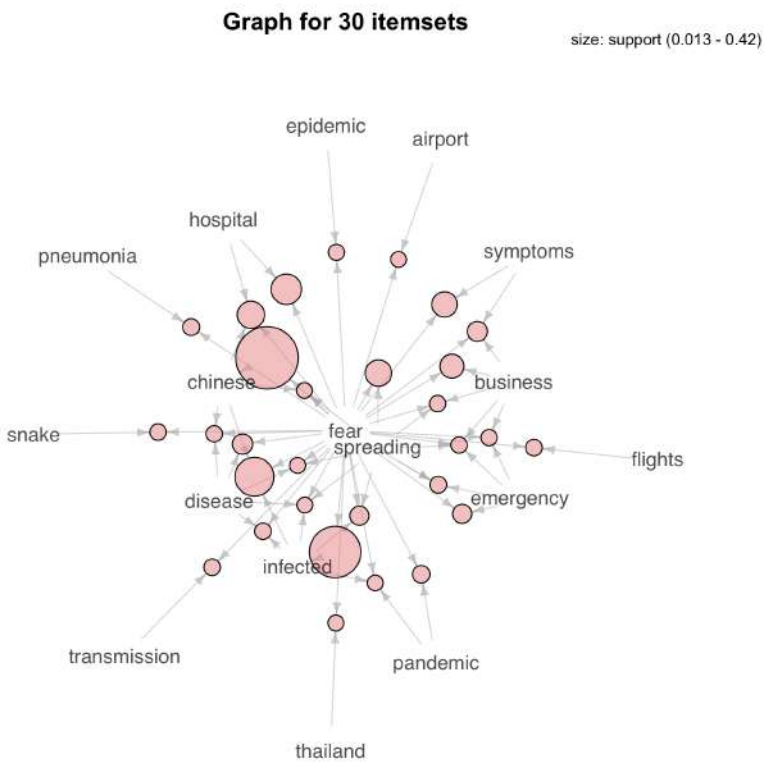

Рисунок 5.30 – Граф частих множин, які містять ключове слово *'fear'*

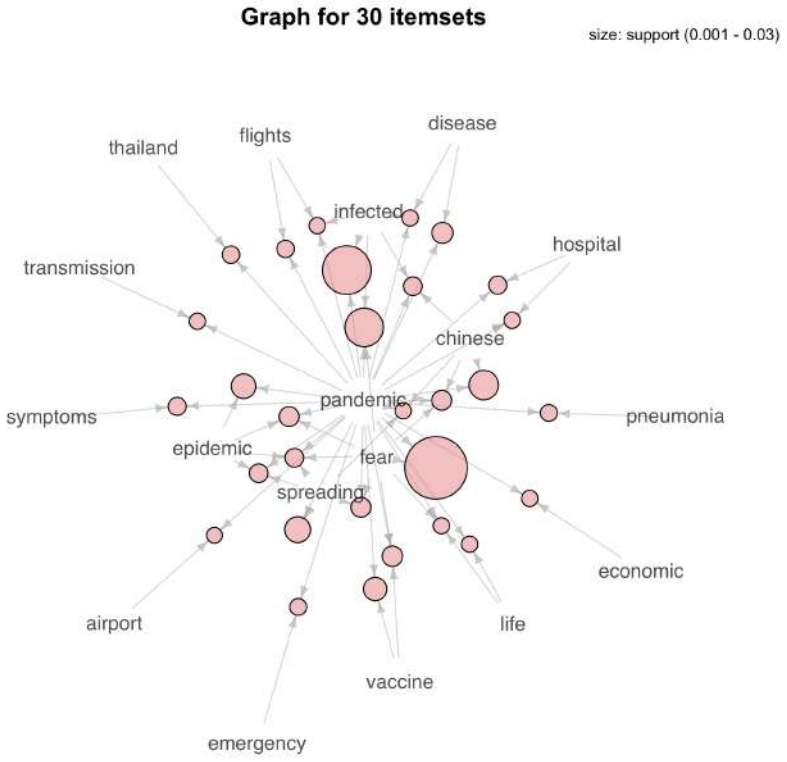

Рисунок 5.31 – Граф частих множин, які містять ключове слово *'pandemic'*

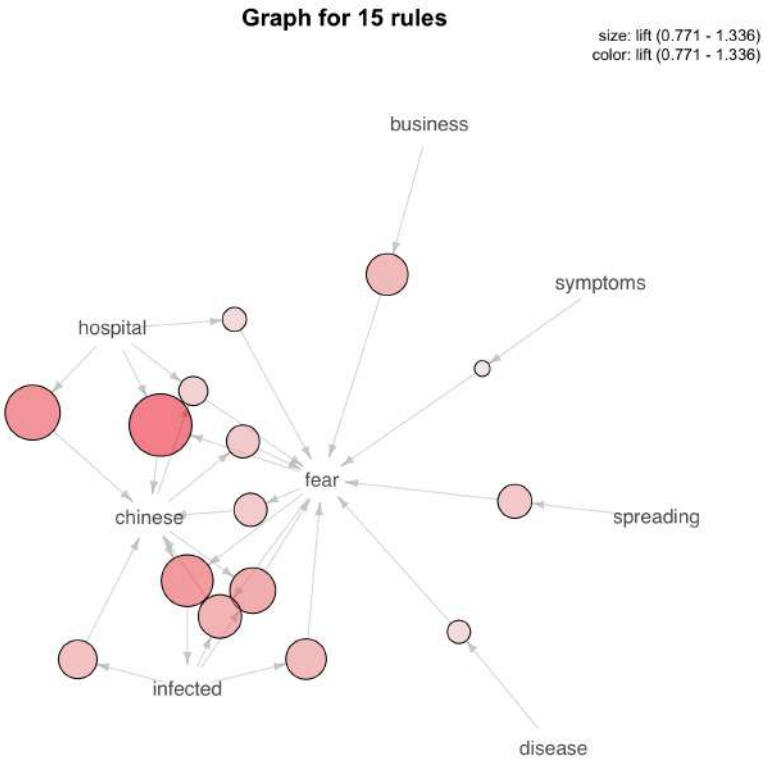

Рисунок 5.32 – Асоціативні правила, виявлені в частих множинах заданого тематичного поля

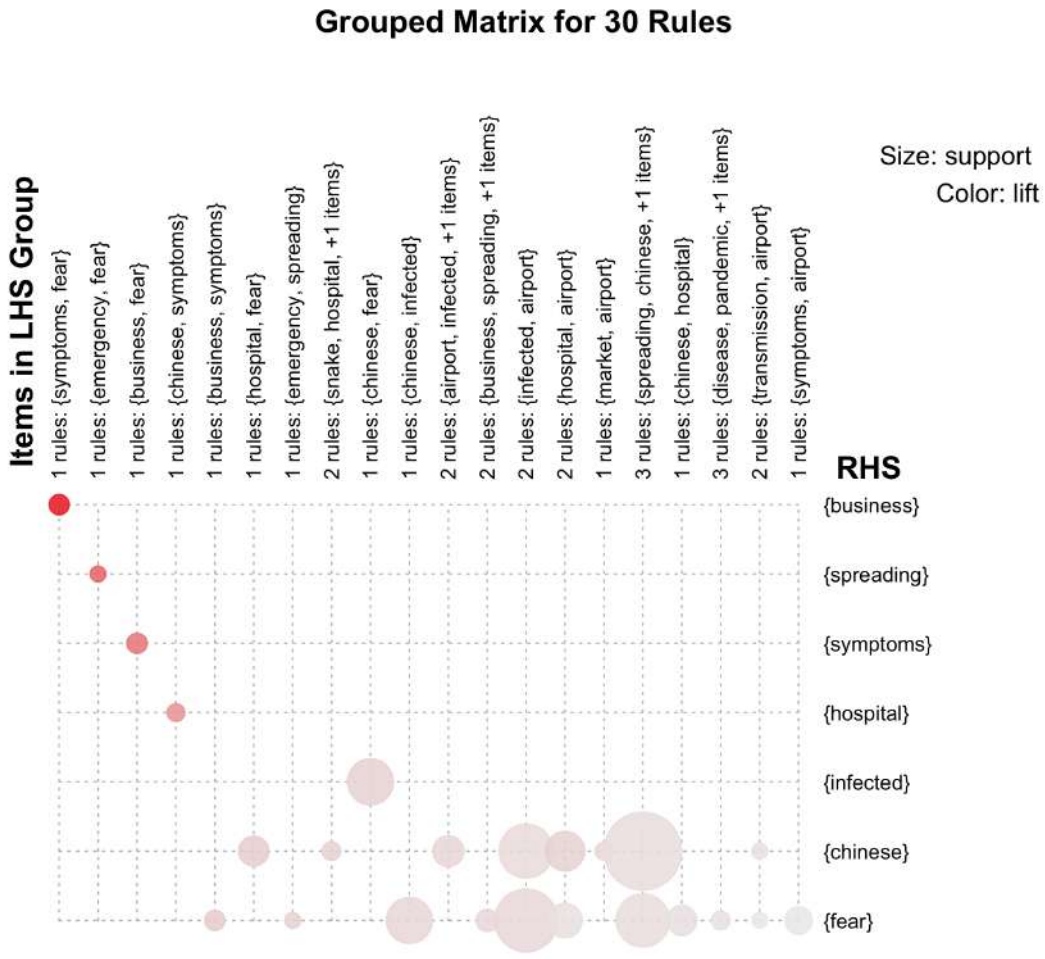

Рисунок 5.33 – Згруповані асоціативні правила

5.5 Висновки

– Розглянуто використання теорії частих множин та асоціативних правил в аналітиці текстових документів, зокрема, коротких повідомлень соціальної мережі Твіттер. Виявлені асоціативні правила характеризують семантичні зв'язки між концептами аналізованої тематики. Динаміка підтримки та достовірності деяких виявлених асоціативних правил відображає інформаційні тренди в обговоренні аналізованого процесу чи очікуваної події. Прогнозні часті множини ключових лексем можуть бути сформовані на основі тематичних полів. Під тематичним полем можна розглядати наперед задану множину лексем, яка характеризує аналізовану тематику. Часті множини лексем, які не входять у тематичне поле, можна відкинути як ситуативні та не характерні для аналізованої тематики.

– Досліджено, що на основі текстових повідомлень мережі Твіттер можна побудувати аналітичні моделі, які відображають вплив значущих подій, пов'язаних із аналізованими бізнес процесами, на динаміку відповідних фінансових часових рядів, зокрема ціни акції аналізованої компанії на фондовому ринку.

– Показано, що в потоках твітів, у яких обговорюються очікувані події, можна виявити ознаки на основі частих множин, які мають прогнозний потенціал стосовно цих подій.

– Показано, що актуальним є аналіз зв'язків між користувачами на основі теорії графів та виявлення впливових користувачів та спільнот. Різні спільноти можуть формувати різні інформаційні тренди. Належність користувача до спільноти може бути додатковою ознакою у прогнозній моделі.

– Квантитативні характеристики частих множин та асоціативних правил, виявлених у масиві твітів, можна використовувати в аналітиці як додаткові ознаки у прогнозних моделях.

6 АНАЛІЗ ФОРМАЛЬНИХ КОНЦЕПТІВ У МАСИВАХ ТЕКСТОВИХ ДАНИХ

6.1 Методи групування текстових даних на основі моделі семантичного контексту

Розглянемо аналіз текстових даних на основі об'єднання методів теорії аналізу формальних концептів. Зокрема, методи аналізу формальних концептів можна використовувати для формування семантичного базису векторного простору, в якому кластеризуються текстові документи. Пошук комплексних характеристик текстових документів є важливим, зокрема, при аналізі авторства текстів, оскільки лексемний частотний спектр творів може бути однаковим, але відрізнятися за характеристиками комбінованих лексемних груп. У теорії аналізу формальних понять (Formal Concept Analysis) [179, 180, 181, 182, 183, 184, 180, 185, 186, 187] аналізують ієрархії формальних понять, використовуючи математичний апарат теорії алгебраїчних ґраток (решіток). Однією із актуальних проблем є побудова моделі формального контексту для семантичних характеристик текстових даних на основі векторної моделі текстових документів та формального аналізу понять. Розглянемо модель формального семантичного контексту текстових масивів [346, 347, 348, 349, 350]. Проаналізуємо алгебраїчну ґратку семантичних концептів. На основі формальних змістів концептів, які відображають тематику аналізу, побудуємо тематичне семантичне поле. Лексемний склад цього поля використаємо як базис векторного простору, в якому можна реалізувати кластеризацію текстових документів. Розглянемо модель, яка описує сукупність текстових документів, лексемний склад та семантичні поля. Розглянемо матрицю семантичних ознак (3.17) типу *"частоти_семантичних_полів–документи"*

$$M_{sd} = \left(p_{kj}^{sd}\right)_{k=1,j=1}^{N_s,N_d}. \tag{6.1}$$

Вектор

$$V_j^s = \left(p_{1j}^{sd}, p_{2j}^{sd}, ..., p_{N_sj}^{sd}\right) \tag{6.2}$$

відображає документ d_j в N_s-мірному семантичному просторі текстових документів. Розглянемо бінарні семантичні характеристики текстового документа

$$p_{kj}^{bs} = \begin{cases} 1, p_{kj}^{sd} \geq \bar{p}_k^{sd}, \\ 0, p_{kj}^{sd} < \bar{p}_k^{sd}, \end{cases} \tag{6.3}$$

де $\bar{p}_k^{sd}$ – деяке порогове значення частоти семантичного поля s_k. Враховуючи (6.3), вектор бінарних семантичних характеристик можна записати у вигляді

$$V_j^{bs} = \left(p_{1j}^{bs}, p_{2j}^{bs}, ..., p_{N_s j}^{bs}\right). \tag{6.4}$$

Побудуємо модель ґратки семантичних концептів. Розглянемо модель семантичної структури текстових масивів використовуючи теорію аналізу формальних понять [179, 180, 188]. Визначимо семантичний контекст як трійку

$$K_s = (D, S, I), \tag{6.5}$$

де D – масив документів, S – множина семантичних полів, I – відношення належності семантичного поля до даного документу

$$I \subseteq D \cdot S, I = \{ (d_i, s_k) \}. \tag{6.6}$$

Пара (d_i, s_k) означає, що документ d_i характеризується семантичним полем s_k, тобто $p_{kj}^{bs} = 1$. Уведемо ґратку семантичних концептів. Для деяких $Ext \subseteq D$, $Int \subseteq S$ визначимо такі відображення

$$Ext' = \{s \in S \mid d \in Ext : (d, s) \in I\}, \tag{6.7}$$

$$Int' = \{d \in D \mid s \in Int : (d, s) \in I\}. \tag{6.8}$$

Множина Ext' описує семантичні поля, властиві документам множини Ext, а множина Int' описує документи, які володіють семантичними полями множини Int. Уведемо семантичний концепт як пару

$$Concept = (Ext, Int), \tag{6.9}$$

до якої належать лексеми з множини $Ext \subseteq D$ та семантичні поля з множини $Int \subseteq S$ з такими умовами

$$\begin{cases} Ext' = Int, \\ Int' = Ext. \end{cases} \tag{6.10}$$

Множину Ext назвемо формальним об'ємом, а Int – формальним змістом семантичного концепту $Concept$. У семантичному контексті утворюється частково-впорядкована множина семантичних концептів

$$\Psi(D, S, I) = \{\, Concept_m = (Ext_m,\, Int_m) \mid m = 1, 2, ...N_{ct} \,\}, \tag{6.11}$$

де N_{ct} – кількість виявлених семантичних концептів у формальному семантичному контексті масиву текстових документів. Семантичний концепт

$$Concept_1 = (Ext_1, Int_1) \tag{6.12}$$

є менш загальним за формальним об'ємом, ніж концепт

$$Concept_2 = (Ext_2, Int_2), \tag{6.13}$$

тобто, виконується умова

$$(Ext_1, Int_1) \leq (Ext_2, Int_2), \tag{6.14}$$

якщо

$$Ext_1 \subseteq Ext_2 \Leftrightarrow Int_1 \supseteq Int_2. \tag{6.15}$$

У цьому випадку концепт $Concept_2$ можна вважати узагальненням концепту $Concept_1$. Семантичний концепт можна розглядати як підматрицю семантичного контексту, яка повністю заповнена одиницями. Ґратку концептів часто відображають за допомогою діаграм Гассе. В аналізі семантичного контексту кожний елемент діаграми представляє семантичний концепт. На верхньому рівні діаграми концепт включає в себе всі текстові документи і нульову множину семантичних полів. На другому рівні в елементи діаграми входить одне семантичне поле, на третьому –

два семантичних поля і так до найнижчого рівня, який включає в себе всі семантичні поля та нульову множину текстових документів. Такі діаграми відображають внутрішню семантичну структурну організацію масивів текстових документів на основі теорії формального аналізу понять. Семантичні концепти об'єднують групи текстових документів та семантичні поля, властиві цим документам. У випадку тематичного аналізу текстових даних в ґратці семантичних концептів можна виявити підмножину формальних змістів концептів Int_j, які будуть відображати тематику аналізу. Тематичне семантичне поле розглянемо як об'єднання формальних змістів таких концептів:

$$S_t = \{ s_i \mid s_i \in Int_j \} . \tag{6.16}$$

Розглянемо векторний простір, базис якого утворюється на основі елементів множини S_t. У такому просторі кожен текстовий документ буде розглядатися як вектор

$$V_j^{td} = \left(p_{1j}^{ts}, p_{2j}^{ts}, ..., p_{N_{St}j}^{ts}\right) , \tag{6.17}$$

де N_{St} – кількість семантичних полів у тематичному полі S_t. На відміну від векторного представлення документів (6.2), у цьому векторному представленні присутні частоти лише деякої підмножини семантичних полів, які відображають тематику заданого аналізу. Векторне представлення документа (6.17) використаємо для групування документів за допомогою ієрархічної кластеризації, яка дасть можливість виявити групи документів, близькі за визначеною тематикою. Такий підхід ефективніший, ніж кластеризація за наперед визначеною множиною семантичних полів, оскільки тематично близькі документи можуть сильно відрізнятися за несуттєвими семантичними полями і в результаті не ввійдуть у спільний кластер.

Отже, запропонована модель семантичного контексту відображає структурну семантичну організацію текстових масивів. У семантичному контексті формується частково впорядкована множина семантичних концептів, формальний зміст яких визначається семантичними полями, а формальний об'єм – масивами текстових документів. Побудова ґратки семантичних концептів у текстових документах дає можливість описувати ієрархічну семантичну структуру в масиві документів та виявляти групи

текстових документів, об'єднані спільною групою семантичних ознак. На основі формальних змістів концептів, які відповідають заданій тематиці, можна сформувати базис семантичного простору текстових документів. Ієрархічна кластеризація документів у такому просторі дає можливість згрупувати у спільних кластерах тематично близькі документи та ігнорувати відмінності за несуттєвими для тематики семантичними полями.

6.2 Модель гратки семантичних концептів для інтелектуального аналізу текстових повідомлень Твіттера

У теорії аналізу формальних концептів [179, 180, 181, 182, 183, 184, 180, 185, 186, 187, 351] розглядають відношення об'єктів та їх атрибутів, на основі якого будують алгебраїчну ґратку формальних концептів. Кожен концепт об'єднує множину об'єктів та їх спільних атрибутів. На основі частих множин спільних атрибутів виявляють асоціативні правила, які відображають зв'язки між атрибутами на множині аналізованих об'єктів. Актуальним є створення моделі формальних концептів для аналізу мікроблогів, яка б враховувала семантичну структуру повідомлень. Для цього доцільно ввести поняття семантичного поля, яке б об'єднувало ключові лексеми тематики аналізу. Соціальна мережа Твіттер є одним із популярних засобів взаємодії користувачів за допомогою коротких повідомлень, які називають твітами. Формат таких повідомлень є простим і дозволяє згадувати в тексті інших користувачів (наприклад, *@username*) та тематичні групи за допомогою хештегів з позначкою # (наприклад, *#software*). Такий формат дає можливість за деяким ключовим словом виявляти повідомлення, які містять це слово, а також виявляти користувачів та групи, які стосуються тематики заданої цим ключовим словом. Такі повідомлення також несуть інформацію про взаємозв'язок між окремими користувачами та ключовими словами. Для твітів характерна висока густина тематично значущих ключових слів. Ця особливість зумовлює перспективність досліджень мікроблогів засобами інтелектуального аналізу та актуальність розвитку методів інтелектуального аналізу текстових повідомлень для виявлення семантичних зв'язків між основними поняттями та тематиками обговорень

в мікроблогах. Розглянемо можливість використання методів теорії аналізу формальних концептів в інтелектуальному аналізі низькорозмірних текстових документів. Розглянемо утворення семантичних концептів та асоціативних правил. На основі утвореної моделі проаналізуємо текстовий масив повідомлень соціальної мережі Твіттер. Розглянемо теоретико-множинну модель, яка описує повідомлення мікроблогів, згрупованих за користувачами. Проаналізуємо утворення семантичних концептів та асоціативних правил для груп хештегів та ключових слів користувачів. Нехай вибрано деяке ключове слово kw, яке задає тематику повідомлень і є наявним у всіх повідомленнях, наприклад kw=*"software"*. Визначимо множину повідомлень мікроблогів:

$$TW^{kw} = \left\{ tw_i^{(kw)} \mid kw \in tw_i \right\}. \tag{6.18}$$

Загальний словник аналізованого масиву повідомлень розглянемо як мультимножину

$$W_s^{tw(kw)} = \left\{ n_i^{st}(w_i) \mid w_i \in TW^{kw} \right\}, \tag{6.19}$$

де n_i^{st} – кількість зустрічань лексеми w_i у повідомленнях аналізованого масиву. Оскільки всі повідомлення містять наперед задане ключове слово (в наших дослідженнях це слово – *"software"*), то такий масив повідомлень буде охоплювати деякий наперед заданий семантичний спектр інформації. Множину користувачів позначимо

$$USR = \{usr_i\}. \tag{6.20}$$

Розглянемо об'єднання всіх повідомлень кожного окремого користувача usr_i як цілісні інформаційні об'єкти

$$tw_j^{usr(kw)} = \{tw_i \mid usr(tw_i) = usr_j\}. \tag{6.21}$$

Масив повідомлень розглянемо як об'єднання повідомлень окремих користувачів, тобто інформаційних об'єктів $tw_j^{usr(kw)}$

$$TW_s^{usr(kw)} = \left\{ tw_j^{usr(kw)} \right\}. \tag{6.22}$$

Уведемо узагальнене поняття семантичного поля. Під семантичнним полем будемо розуміти деяку підмножину словника, елементи якої об'єднані деяким спільним семантичним поняттям. У загальному випадку такі поняття можуть об'єднувати ключові слова, які належать до підрозділів аналізованої тематики. Уведемо множину семантичного поля, в яку входять ключові слова та хештеги назв тематичних груп

$$Keywords = \{keyword_i\}. \tag{6.23}$$

Множина $Keywords$, яка відображає задану тематику може бути сформованою на основі експертного аналізу, коли експерт формує масив ключових слів, які охоплюють напрям досліджень. Семантичне поле можна також утворити на основі знайдених частих множин лексем. Такі множини формуються з наборів лексем, які одночасно зустрічаються у повідомленнях із частотою, більшою за деякий заданий поріг. Очевидно, що деяка підмножина масиву частих множин ключових лексем буде відображати семантику напрямку досліджень мікроблогів. Використовуючи теорію аналізу формальних концептів [179, 180, 188, 351], розглянемо формальний контекст як трійку

$$K^{tw(kw)} = \left(TW_s^{(kw)}, Keywords, I_s\right), \tag{6.24}$$

де I_s – відношення $I_s \subseteq TW_s^{(kw)} \times Keywords$, яке описує зв'язки повідомлень із ключовими лексемами в цих повідомленнях. Вважаємо, що $(tw_i^{(kw)}, keyword_j) \in I_s$, якщо термін $keyword_j$ зустрічається у повідомленні $tw_i^{(kw)}$. Відношення I_s можна розглядати як множину

$$I_s = \left\{(tw_i, keyword_j) | keyword_j \in tw_i^{(kw)}\right\}. \tag{6.25}$$

Уведемо ґратку семантичних концептів. Для деяких $Ext \subseteq TW_s^{(kw)} Int \subseteq Keywords$ визначимо такі відображення:

$$Ext' = \{keyword_j \in Keywords \mid tw_i^{(kw)} \in Ext : (tw_i^{(kw)}, keyword_j) \in I_s\}, \tag{6.26}$$

$$Int' = \{tw_i^{(kw)} \in TW_s^{(kw)} \mid keyword_j \in Int : (tw_i^{(kw)}, keyword_j) \in I_s\}. \quad (6.27)$$

Множина Ext' описує ключові терміни, властиві документам множини Ext, а множина Int' описує повідомлення, які містять ключові терміни множини Int. Уведемо семантичний концепт як пару

$$Concept = (Ext, Int), \quad (6.28)$$

до якої належать повідомлення з множини $Ext \subseteq TW_s^{(kw)}$ та ключові терміни з множини $Int \subseteq Keywords$ з такими умовами

$$\begin{cases} Ext' = Int, \\ Int' = Ext. \end{cases} \quad (6.29)$$

Множину Ext назвемо формальним об'ємом, а Int – формальним змістом семантичного концепту $Concept$. У семантичному контексті $K^{tw(kw)}$ утворюється частково-впорядкована множина семантичних концептів

$$\Psi(TW_s^{(kw)}, Keywords, I_s) = \{Concept_m = (Ext_m, Int_m)\}. \quad (6.30)$$

В аналізі семантичного контексту $K^{tw(kw)}$ кожний елемент діаграми представляє семантичний концепт. Такі діаграми відображають внутрішню семантичну структурну організацію повідомлень користувачів та відповідних їм груп ключових слів.

Аналогічно розглянемо формальний контекст для повідомлень, згрупованих за користувачами

$$K_{usr}^{tw(kw)} = \left(TW_s^{usr(kw)}, Keywords, I_s^{usr}\right), \quad (6.31)$$

де I_s^{usr} – відношення $I_s^{usr} \subseteq TW_s^{usr(kw)} \times Keywords$, яке описує зв'язки повідомлень користувачів із ключовими словами у цих повідомленнях. Будемо вважати, що $(tw_i^{usr(kw)}, keyword_j) \in I_s^{usr}$, якщо ключове слово $keyword_j$ зустрічається у масиві повідомлень $tw_i^{usr(kw)}$ певну кількість разів

n_{ij}^{usr}. Відношення I_s^{usr} можна розглядати як множину

$$I_s^{usr} = \left\{(tw_i^{usr}, keyword_j) | keyword_j \in tw_i^{usr(kw)}, n_{ij}^{usr} > n_{th}^{usr}\right\}. \quad (6.32)$$

Введення порогового значення n_{th}^{usr} є необхідними для того, щоб включити у розгляд лише ключові слова понять, які активно обговорюються. Уведемо ґратку семантичних концептів. Для деяких

$$Ext^{usr} \subseteq TW_s^{usr(kw)}, Int^{usr} \subseteq Keywords,$$

визначимо такі відображення

$$(Ext^{usr})' = \left\{keyword_j \in Keywords | tw_i^{usr(kw)} \in Ext^{usr} : (tw_i^{usr(kw)}, keyword_j) \in I_s^{usr}\right\}, \quad (6.33)$$

$$(Int^{usr})' = \left\{ tw_i^{usr(kw)} \in TW_s^{usr(kw)} | keyword_j \in Int^{usr} : (tw_i^{usr(kw)}, keyword_j) \in I_s^{usr}\right\}. \quad (6.34)$$

Множина $(Ext^{usr})'$ описує ключові слова, властиві згрупованим за користувачами повідомленням множини Ext^{usr}, а множина $(Int^{usr})'$ описує повідомлення користувачів, які містять ключовими словами множини Int^{usr}. Аналогічно до (6.28) уведемо семантичний концепт як пару $Concept^{usr} = (Ext^{usr}, Int^{usr})$, до якої належать згруповані за користувачами повідомлення з множини $Ext^{usr} \subseteq TW_s^{usr(kw)}$ та ключові слова з множини $Int^{usr} \subseteq Keyword$ з умовами $(Ext^{usr})' = Int^{usr}, (Int^{usr})' = Ext^{usr}$, де Ext^{usr} назвемо формальним об'ємом, а Int^{usr} – формальним змістом семантичного концепту $Concept^{usr}$. Таким чином, об'єм семантичного концепту об'єднує користувачів, які часто вживають у своїх повідомленнях ключові слова, які утворюють множину формального змісту цього концепту. У семантичному контексті $K_{usr}^{tw(kw)}$ утворюється частково-впорядкована множина семантичних концептів повідомлень, згрупованих за користувачами

$$\Psi^{usr}(TW_s^{usr(kw)}, Keywords, I_s^{usr}) = \{Concept_m^{usr} = (Ext_m^{usr}, Int_m^{usr})\}. \quad (6.35)$$

Розглянемо поняття порядкового ідеалу та фільтра для деякої частково впорядкованої множини $(P, \leq)$. Порядковим ідеалом називають підмножину $J \subseteq P$, для якої

$$\forall x \in J, y \leq x \Rightarrow y \in J. \tag{6.36}$$

Порядковим фільтром називають підмножину $F \subseteq P$, для якої

$$\forall x \in F, y \geq x \Rightarrow y \in F. \tag{6.37}$$

Використання понять порядкового ідеала та фільтра може бути ефективним в аналізі ґратки семантичних концептів. Порядковим ідеалом деякого концепта будуть концепти, які пов'язані з ним на діаграмі Гассе і знаходяться нижче нього, включно з концептом, який відповідає інфімуму ґратки. Порядковим фільтром деякого концепту є множина пов'язаних із ним концептів, які знаходяться вище нього в ґратці, включно з концептом, який відповідає супремуму ґратки. Формальний зміст деякого концепту є підмножиною формальних змістів концептів, які належать до його порядкового ідеалу, а об'єднання формальних змістів концептів, які утворюють порядковий фільтр деякого концепту, утворює формальних зміст цього концепту. Інформативним для аналізу є також розгляд об'єднання порядкового фільтра та ідеала. Множина змістів такого об'єднання утворює деяке семантичне поле, яке відображає множину взаємопов'язаних понять. В одній ґратці може знаходитися декілька таких незалежних об'єднань порядкових ідеалів та фільтрів. Отже, одним із методів формування семантичних полів є пошук множини формальних змістів концептів деякого об'єднання ідеала та фільтра заданого формального контексту. На основі розрахованої ґратки семантичних концептів можна виявити асоціативні правила, які відображають семантичні структурні зв'язки між ключовими словами. Під асоціативним правилом деякого контексту

$$K^{tw(kw)} = \left(TW_s^{(kw)}, Keywords, I_s\right)$$

будемо розуміти вираз

$$A \to B,\ A, B \subseteq Keywords. \tag{6.38}$$

Підмножину A називають передумовою, а B – наслідком асоціативного правила $A \to B$. Важливими характеристиками асоціативних правил є підтримка (support) $Supp_{A\to B}$ та достовірність (confidence) $Conf_{A\to B}$, які можна обрахувати за такими виразами:

$$Supp_{A\to B} = \frac{|(A \cup B)'|}{\left|TW_s^{(kw)}\right|}, \tag{6.39}$$

$$Conf_{A\to B} = \frac{|(A \cup B)'|}{|A'|}. \tag{6.40}$$

У випадку, коли $Supp_{A\to B} = 1$, асоціативне правило (6.38) є імплікацією, тобто, виконується завжди, коли зустрічається передумова A. Значення $Supp_{A\to B}$ характеризує частку повідомлень $TW_s^{(kw)}$, яка містить ознаки $A \cup B$. Величина $Conf_{A\to B}$ характеризує частку повідомлень із ключовими словами множини A, яка також містить ключові слова множини B. Актуальними для аналізу є правила з деяким заданим мінімальним значенням достовірності та підтримки:

$$Supp_{A\to B} > Supp_{\min}, \tag{6.41}$$

$$Conf_{A\to B} > Conf_{\min}. \tag{6.42}$$

Асоціативні правила з умовами (6.41)–(6.44) називають частими та отримують із частої підмножини ключових лексем:

$$F \subseteq Keywords, |F'| > \theta, F = A \cup B, A \cap B = \varnothing, \tag{6.43}$$

де θ – деякий поріг частої множини.

Розглянемо експериментальний аналіз ґратки семантичних концептів. Використовуючи API соціальної мережі Твіттер, завантажено текстовий масив повідомлень, які містять ключове слово "software" а також хештег "#software". Тобто, відібрано повідомлення заданого тематичного напряму

пов'язаного з програмним забезпеченням. Твіти з ключовим словом "software" завантажувались в період з 06.08.11 по 11.08.11. В загальному завантажено 75977 твітів. Далі проведено фільтрацію твітів і взято до розгляду лише лексеми, які повторюються не менше 10 разів і не більше 4000 разів. Наведемо приклади високочастотних лексем у порядку спадання частоти зустрічань

> #software (3371), engineer (3186), download (2615), #jobs (2279), online (2098), business (1865), marketing (1758), windows (1751), development (1704), developer (1673), management (1525)

Отриманий частотний словник містить 6325 лексем. Було відфільтровано високочастотні стоп-слова, які не несуть семантичної інформації. Знайдені часті множини термінів із підтримкою більше 10. До розгляду було взято твіти, які містили не менше 5 лексем. Також розглядались часті множини із кількістю термінів від 2 до 5. Отримано список із 2879 частих множин, які відповідають наведеним вище умовам. Для зменшення кількості частих множин було збільшено мінімальну підтримку частих множин до 20. У цьому випадку кількість частих множин зменшилась до 1049. Наведемо деякі із них:

> {manager, #job}, {computer, windows}, {#job, developer}, {microsoft, windows}, {security, internet}, {looking, #job}, {player, traffic, video}, {script, #php}, {servers, hosting}, {servers, remote, desktop, hosting}, {salary, #hiring, location, #job}, {blackberry, android}

В аналізі розглядались ґратки формальних концептів для семантичних полів різного розміру та формального змісту. Розглянемо твіти, в яких присутні лексеми такого найпростішого семантичного поля *S1*

> {london, lndia, windows, microsoft, android, #mysql, scripts, #linux, #job, developer }

У це семантичне поле включено географічні назви, операційні системи, хештег *#job*. Ґратка семантичних концептів буде відображати взаємозв'язок цих понять у повідомленнях мікроблогів. Після фільтрації масиву вхідних повідомлень за наведеним семантичним полем отримано масив із 8920

твітів. Для розрахунку ґратки концептів та побудови діаграм Гассе було використано пакет програм *Lattice Miner* [352]. На рис. 6.1 наведено діаграму Гассе, яка відображає утворену ґратку семантичних концептів для семантичного поля *S1*. На цій діаграмі наведено формальний зміст концептів верхнього рівня. Формальні змісти концептів нижніх рівнів є комбінаціями наведених формальних змістів відповідно до зв'язків на діаграмі.

На рис. 6.2 виділено фільтр та ідеал для концепта *{android, developer}*. Для концептів наведено формальний зміст та об'єм у процентах. Концепт інфімуму не наведено, оскільки він містить нульовий об'єм. Таблиця 6.1 містить приклади асоціативних правил та їхні кількісні характеристики для наведеної на рис. 6.1 ґратки семантичних концептів.

Розглянемо інше об'ємніше семантичне поле *S2*, яке складається із таких лексем та хештегів:

{#linux, #opensource, android, browser, center, drivers, earth, features, google, greater, installs, internet, iphone, latest, leader, linux, netscape, phones, popular, powerful, printer, sales, telemarketing, tracking, ubuntu}

Для цього семантичного поля отримано відфільтрований контекст, який містить 4681 твітів. Розраховану ґратку семантичних концептів наведено на рис. 6.3. На рис. 6.4 показано зв'язки для концепта *{android}*, які відображають його порядковий фільтр та ідеал. Фільтр представлений лише концептом супремуму та самим концептом *{android}*. Для наведеного вище семантичного поля S2 розраховано асоціативні правила, приклади яких наведено у таблиці 6.2. Серед наведених у таблиці 6.2 правил можна виявити

Таблиця 6.1 – Асоціативні правила відфільтрованого за семантичним полем S1 масиву повідомлень

N	Передумова A	Наслідок B	$Supp_{A\to B}$	$Conf_{A\to B}$
1	{#php}	{#mysql}	0.33%	22.55%
2	{london}	{#job}	0.58%	20.47%
3	{#job, android}	{developer}	0.04%	50.0%
4	{#mysql, #php}	{script}	0.2%	60.0%
5	{linux, mysql}	{developer}	0.04%	30.76%
6	{android, london}	{developer}	0.01%	100.0%

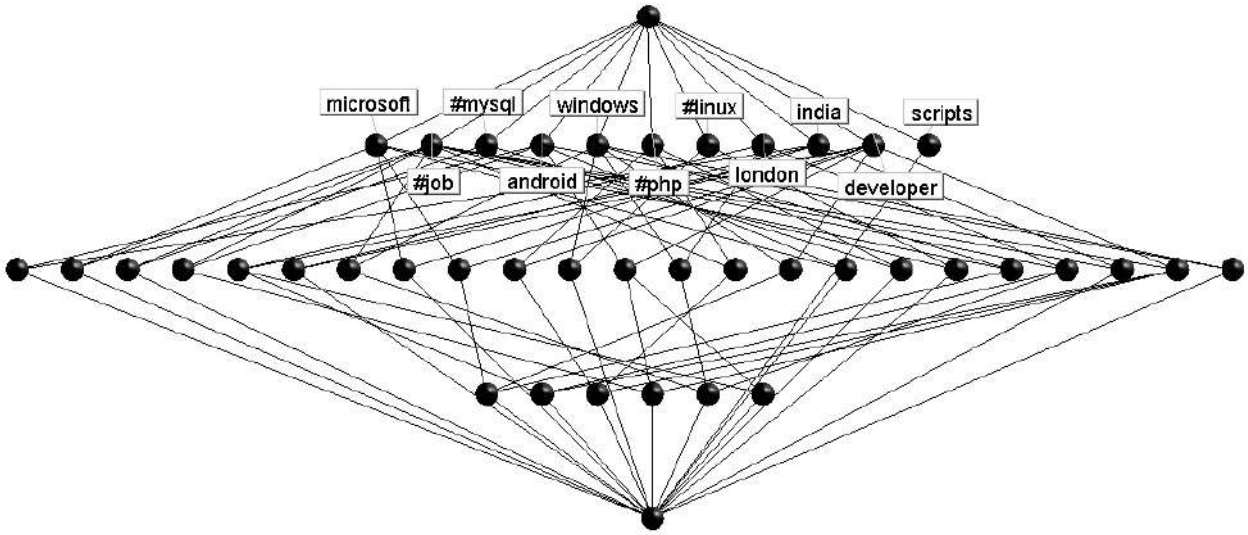

Рисунок 6.1 – Діаграма Гассе для ґратки семантичних концептів семантичного поля S1

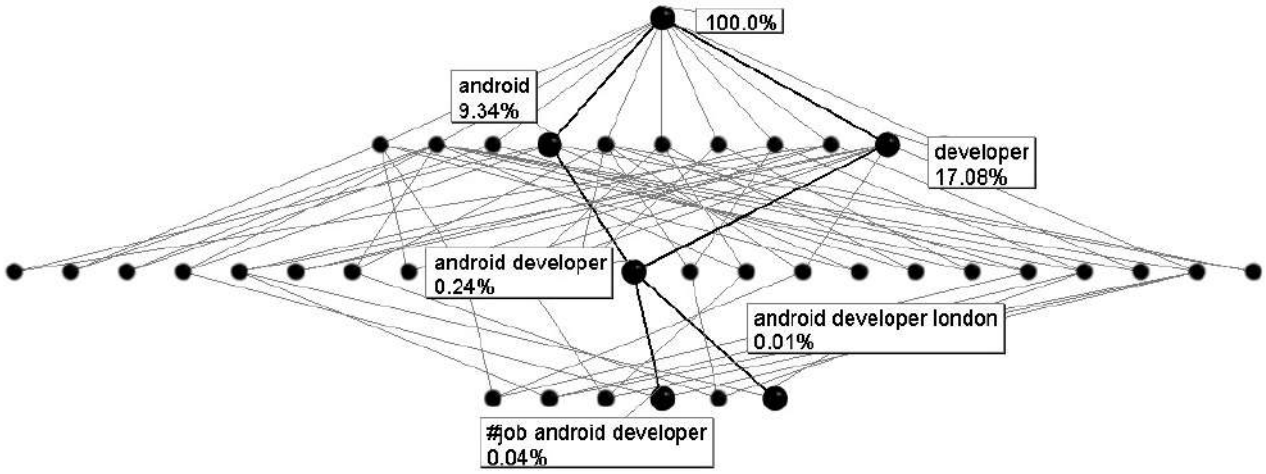

Рисунок 6.2 – Порядковий фільтр та ідеал для концепта *{android, developer}*

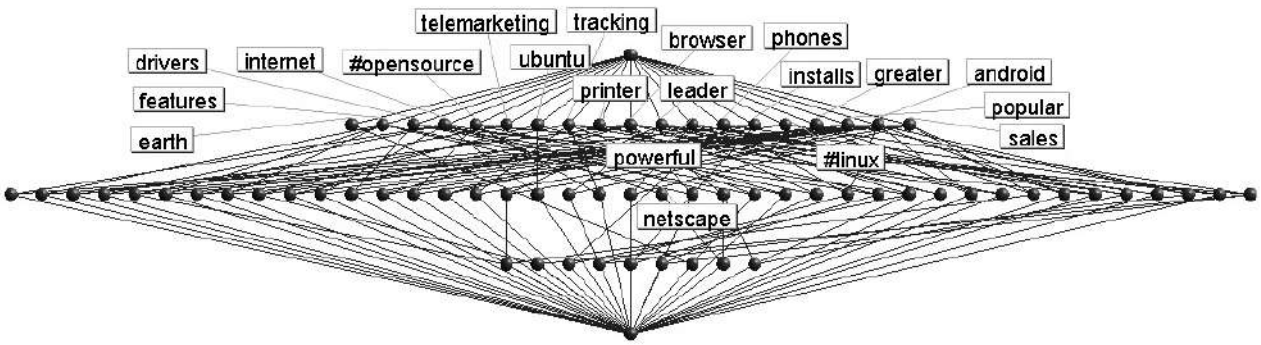

Рисунок 6.3 – Діаграма Гассе для ґратки семантичних концептів семантичного поля S2

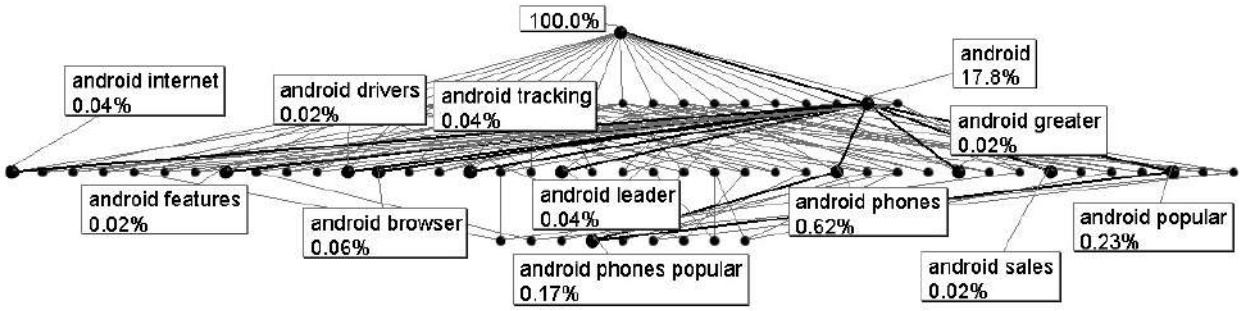

Рисунок 6.4 – Порядковий фільтр та ідеал для концепта {android}

Таблиця 6.2 – Асоціативні правила відфільтрованого за семантичним полем *S2* масиву повідомлень

N	Передумова A	Наслідок B	$Supp_{A\to B}$	$Conf_{A\to B}$
1	{#linux}	{#opensource}	1.62%	45.5%
2	{telemarketing}	{sales}	0.21%	83.33%
3	{browser, internet}	{netscape}	0.21%	55.55%
4	{#linux, ubuntu}	{#opensource}	0.14%	46.66%
5	{android, popular}	{phones}	0.17%	72.72%
6	{#opensource, ubuntu}	{#linux}	0.14%	100.0%
7	{phones, popular}	{android}	0.17%	100.0%

такі імплікації:

$$\{\#opensource, ubuntu\} \Rightarrow \{\#linux\}, \{phones, popular\} \Rightarrow \{android\} \quad (6.44)$$

Правила (6.44) є імплікаціями лише для відфільтрованого масиву твітів і в загальному випадку повідомлень мікроблогів можуть не бути імплікаціями. Отже, застосування теорії аналізу формальних концептів є ефективним в інтелектуальній обробці повідомлень мікроблогів. Використання моделі ґратки семантичних концептів дає можливість аналізувати семантично зв'язані множини лексем та будувати асоціативні правила. Формування семантичних полів на основі масиву виявлених частих множин дає можливість суттєво звузити пошук асоціативних правил та розмір ґратки семантичних концептів в алгоритмах інтелектуального аналізу текстів.

Розглянемо повідомлення, згруповані за користувачами. Для аналізу було взято таку саму базу повідомлень, як і в попередньому дослідженні. Далі було сформовано контекст, у якому рядки відображали об'єднані повідомлення кожного із дописувачів мікроблогу. Було відфільтровано стоп-слова. З розгляду було вилучено лексеми які користувач використовував менше 25 разів. Було сформовано контекст, який містив 112 об'єктів та 99 атрибутів. На основі цього контексту отримано асоціативні правила при підтримці більше 0.1%, приклади цих правил наведено у таблиці 6.3. Деякі асоціативні правила утворюють імплікації, тобто є справедливі для всіх

Таблиця 6.3 – Приклади виявлених асоціативних правил

N	Передумова A	Наслідок B	$Supp_{A\to B}$	$Conf_{A\to B}$
1	{android}	{phone, windows}	0.89%	50.0%
2	{windows}	{office}	0.89%	50.0%
3	{#linux}	{#opensource}	0.89%	50.0%
4	{#income}	{programmer}	0.89%	100.0%
5	{engineering}	{computer, hardware}	0.89%	100.0%
7	{#ecommerce}	{business, shopping}	0.89%	100.0%
8	{security}	{internet}	1.78%	100.0%
9	{freeware}	{download}	0.89%	100.0%
10	{stream, traffic}	{internet}	1.78%	100.0%
11	{developer, iphone}	{application, technologies}	0.89%	100.0%
12	{internet, online}	{#movies, #tv, satellite}	0.89%	100.0%

випадків появи передумови правила. Для наведених у таблиці 6.3 прикладів можна виявити, зокрема, такі імплікації:

{internet, online} => {#movies, #tv, satellite},
{developer, iphone}=>{application, technologies},
{freeware}=>{download},
{#ecommerce}=>{business, shopping}.

Фрагмент розрахованої діаграми Гассе для отриманої ґратки семантичних концептів із виділеним порядковим ідеалом та фільтром концепту *{internet}* наведено на рис. 6.5. Семантична структура концептів на рис. 6.5, відображає поняття та тематики, якими інтенсивно цікавиться

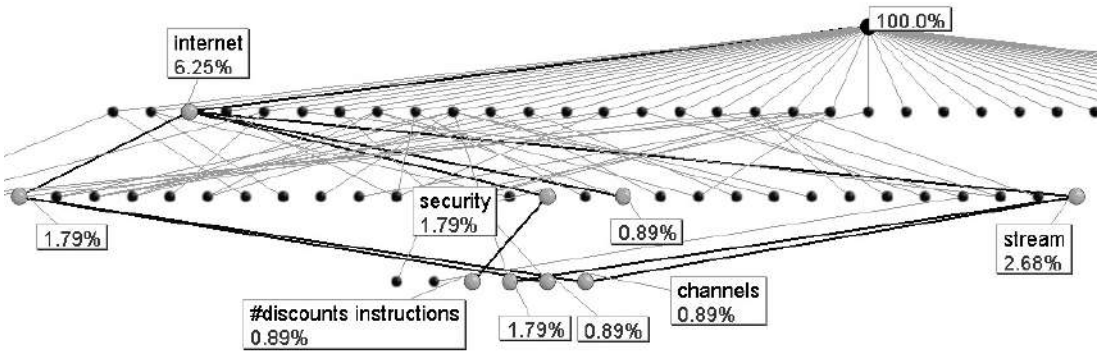

Рисунок 6.5 – Порядковий ідеал та фільтр концепту *{internet}*

певна група користувачів, кожен із яких згадує наведені поняття не менше 25 разів у своїх повідомленнях. Очевидно, що ця семантична структура може не відображатися в аналізі масиву твітів без групування за користувачами, оскільки деякі взаємопов'язані поняття можуть знаходитись в різних повідомленнях одного і того ж користувача, а, отже, не увійдуть у часті множини лексем масиву повідомлень без групування. Отже, розглянуто модель ґратки семантичних концептів для аналізу хештегів та ключових слів у повідомленнях, згрупованих за користувачами мікроблогів. Уведення поняття семантичного поля як множини тематично об'єднаних лексем зменшує обсяг необхідних обчислень унаслідок фільтрації масиву повідомлень. Фільтрація полягає у відкиданні лексем повідомлень, які не входять у задану семантичним полем тематику. Використання аналізу формальних концептів дає можливість утворити алгебраїчну ґратку семантичних концептів, характеристиками яких є формальним об'єм та зміст. Формальний об'єм семантичного концепту об'єднує користувачів, які часто вживають у своїх блогах ключові слова, що утворюють множину формального змісту цього концепту. Групування ключових слів користувачів здійснюється на основі множин формальних змістів семантичних концептів. Утворена ґратка семантичних концептів дає можливість виявляти асоціативні правила у множинах ключових слів, які є складовими формальних змістів семантичних концептів. Ці правила відображають зв'язки між ключовими словами, що характеризують семантичні поняття у повідомленнях користувачів [349, 347, 353, 350].

6.3 Методи прогнозування подій на основі інтелектуального аналізу повідомлень Твіттера з використанням ґратки семантичних концептів

Розглянемо можливість використання теорії формальних концептів на прикладі аналізу фіналу тенісного турніру на олімпіаді в Лондоні 2012 року [342]. Побудуємо алгебраїчну ґратку формальних концептів, яка відображає семантичні зв'язки ключових понять у повідомленнях мікроблогів. Проаналізуємо динаміку підтримки та достовірності у виявлених частих множинах та асоціативних правилах. Особливістю прогнозування спортивних подій є те, що твіти відображають очікування блогерів. Ці очікування не завжди корелюють із реальними результатами, на які часто впливають випадкові фактори, стан підготовки спортсменів. Для аналізу ми завантажили повідомлення мікроблогів Твіттера, які стосувались олімпіади 2012 у Лондоні. Завантаження здійснювалось з 26 липня по 15 серпня 2012 року. Завантажувались твіти, які містили такі ключові слова та хештеги: *"olympics", "#london2012", "#london2012 tennis"* та інші. Твіти завантажувалися в окремі файли для кожного вибраного набору ключових слів та хештегів. За час аналізу було завантажено близько 1Гб твітів. Розглянемо послідовність проведеного аналізу. Щоб отримати очевидні результати, які легко перевірити, виберемо вузький набір понять, які описують напівфінал та фінал олімпійського тенісного турніру. Побудуємо частотний словник файлу твітів із хештегами "#london2012". До розгляду візьмемо лексеми, які зустрічаються не менше 10 разів. Отримано відфільтрований масив повідомлень із вилученими високочастотними стоп-словами, та рідковживаними термами, до яких можна також віднести більшість нікнеймів користувачів. Як відомо, фінал тенісного турніру для жінок пройшов 4 серпня 2012 року і у ньому взяли участь M. Sharapova та S. Williams. А фінал для чоловіків пройшов 5 серпня і у ньому взяли участь A. Murrey і R. Federer. Формалізуємо семантичний фрейм аналізу, включивши до нього лексеми, які позначають етап турніру (final), стать учасників (men, women), тип (single, double), дату (Aug 4, Aug 5), результат (gold,silver) імена атлетів (Federer, Myrray, Sharapova, Williams).

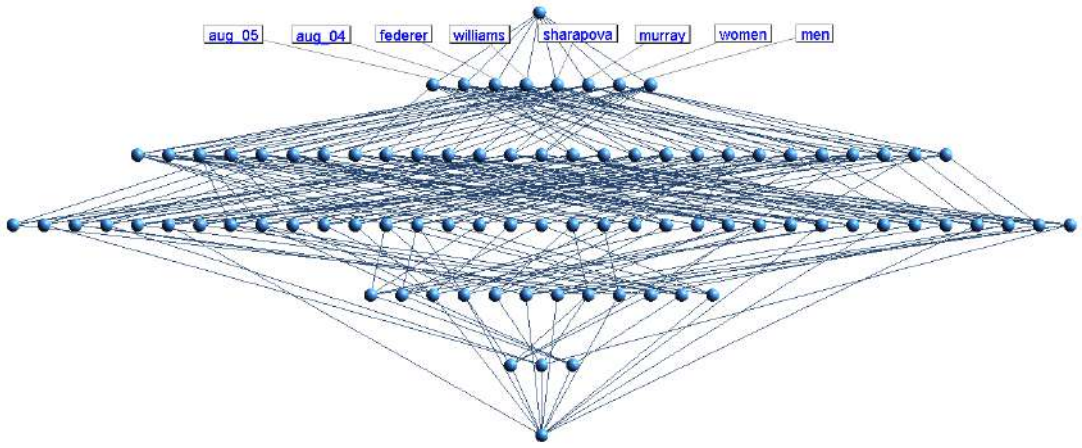

Рисунок 6.6 – Гратка семантичних концептів для тематичного поля масиву твітів із ключовими словами { #london2012, tennis}

Відфільтруємо множину повідомлень так, щоб вони містили лише лексеми заданого тематичного поля. Для побудови діаграми Гассе, яка відображає гратку семантичних концептів, використано програмне забезпечення *Lattice Miner* [352]. При побудові ґратки семантичних концептів фрейму, заданого створеним тематичним полем, утворено 1000 семантичних концептів, тому додатково розіб'ємо множину семантичних ознак на декілька підмножин. На рис. 6.6 наведено ґратку семантичних концептів, яка відображає такі концепти, як можливі дати проведення, стать, імена учасників. Очевидно, що у загальному випадку необхідно було б врахувати всіх можливих учасників, всі дати та види спорту. Але таку ґратку важко було б відобразити графічно. Тому для наочності взято мінімальну кількість концептів. На рис. 6.7 показано гратку концептів, на якій виділено порядковий ідеал та фільтр для концепту *{aug_5, federer, murrey, man}*, об'єм якого дорівнює 2% і є найбільшим для цього рівня ґратки. На основі цього результату можна встановити, що фінал тенісного турніру серед чоловіків пройшов 5 серпня 2012. Очевидно, що для вибору концепту, який відповідає реальній події, необхідно проаналізувати всі можливі концепти та вибрати концепти із максимальним значенням об'єму. Концепти із максимальним значенням об'єму будуть найбільш ймовірними у реальності. Розглянемо детальніше концептуальні зв'язки на основі ґратки із таким семантичним полем *{sharapova, williams, aug_05, aug_04, aug_01, final, wins, gold}*.

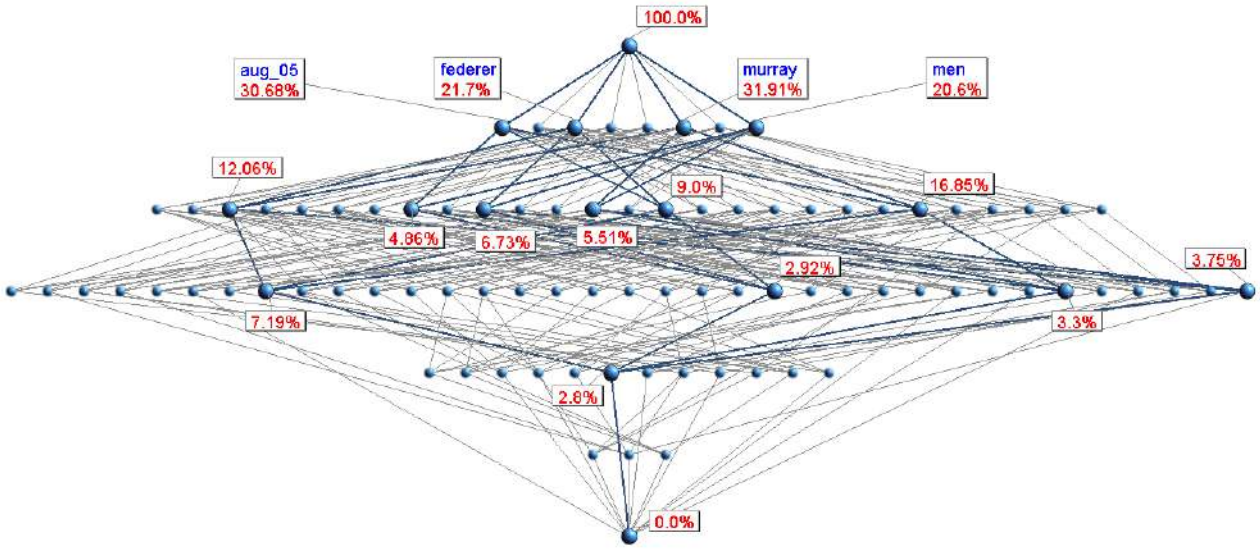

Рисунок 6.7 – Гратка семантичних концептів із виділеним порядковим ідеалом та фільтром для концепту *{aug_5, federer, murrey, man}*

У результаті аналізу отримано такі значення об'ємів для аналізованих концептів:

Ext(sharapova, aug_04, gold)= 3.0 %,
Ext(sharapova, aug_05, gold)= 0.07 %,
Ext(sharapova, aug_01, wins)= 0.04 %,
Ext(sharapova, aug_04, wins)= 1.81 %,
Ext(williams, aug_04, gold)=3.76 %,
Ext(williams, aug_05, gold)=0.79 %,
Ext(williams, aug_01, wins)=0.05 %,
Ext(williams, aug_04, wins)=1.97 %.

Як відомо, у фіналі 4 серпня 2012 року перемогу отримала S. Williams. На рис. 6.8 наведено фільтр та ідеал для концепту *{sharapova, aug_04, gold}*, а на рис. 6.9 наведено фільтр та ідеал для концепту *{williams, aug_04, gold}*. Як випливає з наведених даних

Ext(sharapova, aug_04, gold)< Ext(williams, aug_04, gold)

Ця нерівність відображає реальні результати змагань. Різниця між об'ємами цих двох концептів є незначною. Ще меншою є різниця для розрахованих об'ємів концептів *{sharapova, aug_04, wins }* та *{williams, aug_04, wins}*. Однак, для аналогічних концептів із датою *aug_5* різниця є більш суттєвою,

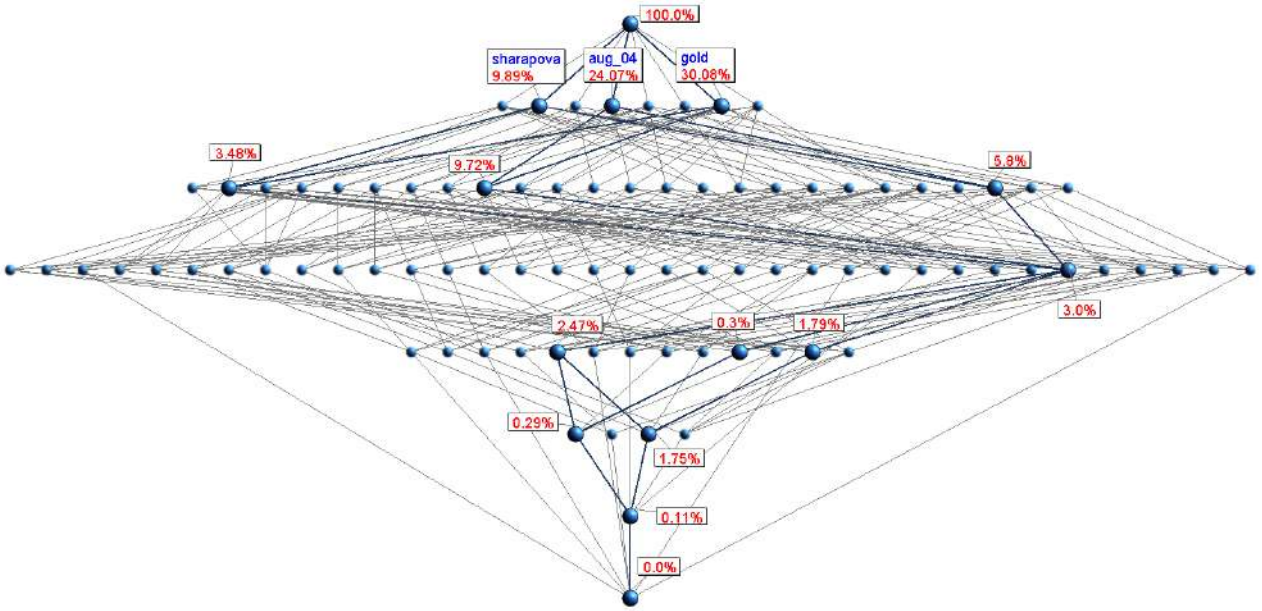

Рисунок 6.8 – Ґратка семантичних концептів із виділеним порядковим ідеалом та фільтром для концепту *{sharapova, aug_04, gold}*

що відображає результат фіналу, у той час як концепти із датами *aug_1, aug_4* відображають здебільшого очікування блогерів. На рис. 6.10 наведено порядковий ідеал та фільтр для концепту *{sharapova, aug_05, gold}*, а на рис. 6.11 наведено фільтр та ідеал для концепту *{williams, aug_05, gold}*.

Розглянемо динаміку кількісних характеристик асоціативних правил. Для аналізу вибрано масив із ключовими словами *#london2012, tennis*. На рис. 6.12 наведено динаміку підтримки для асоціативних правил $Gold \rightarrow Sharapova$, $Gold \rightarrow Williams$, а на рис. 6.13 наведено динаміку достовірності для цих правил. Для побудови графіків використовувався пакет *gnuplot* [354]. Характерними для отриманих кривих є максимуми у день змагань та наступного дня. Це говорить про те, що підтримка та достовірність асоціативних правил у день змагань відображають очікування, а у день після змагань – реальний результат.

Отже, розглянуту в роботі модель семантичних концептів в масиві повідомлень Твіттера апробовано в аналізі та прогнозуванні подій [342]. Як приклад, розглянуто спортивні події, у яких очікування користувачів можуть бути відкореговані випадковими факторами та реальним рівнем підготовки спортсменів, який може відрізнятися від очікувань болільників. Так, величини підтримки відповідних концептів

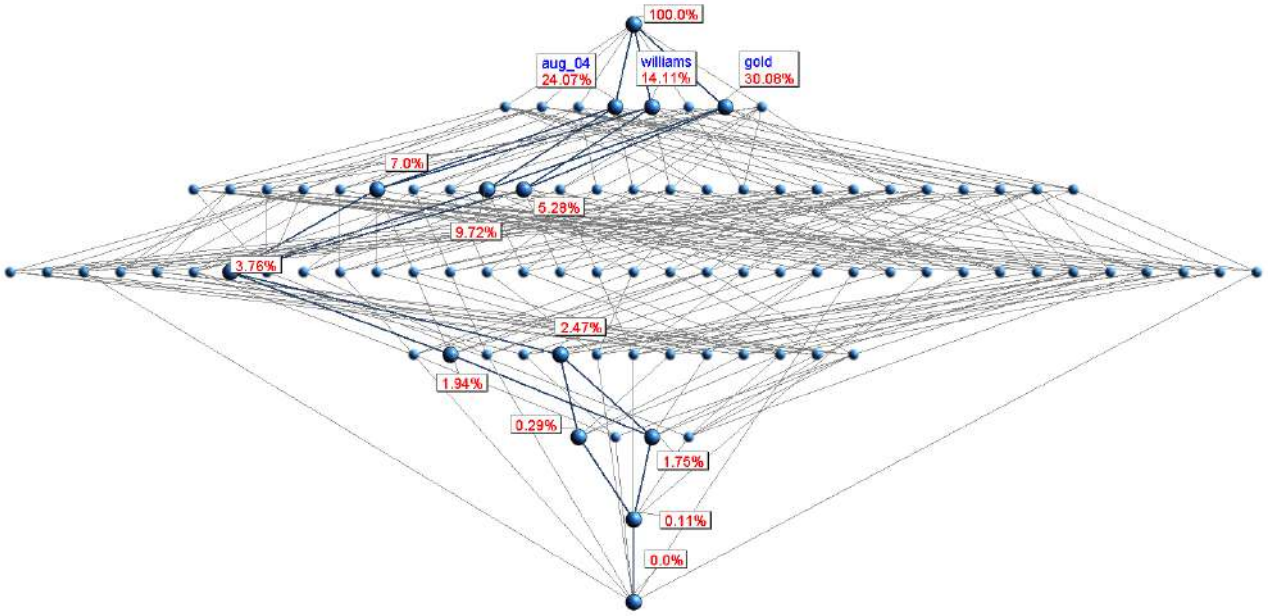

Рисунок 6.9 – Ґратка семантичних концептів із виділеним порядковим ідеалом та фільтром для концепту *{williams, aug_04, gold}*

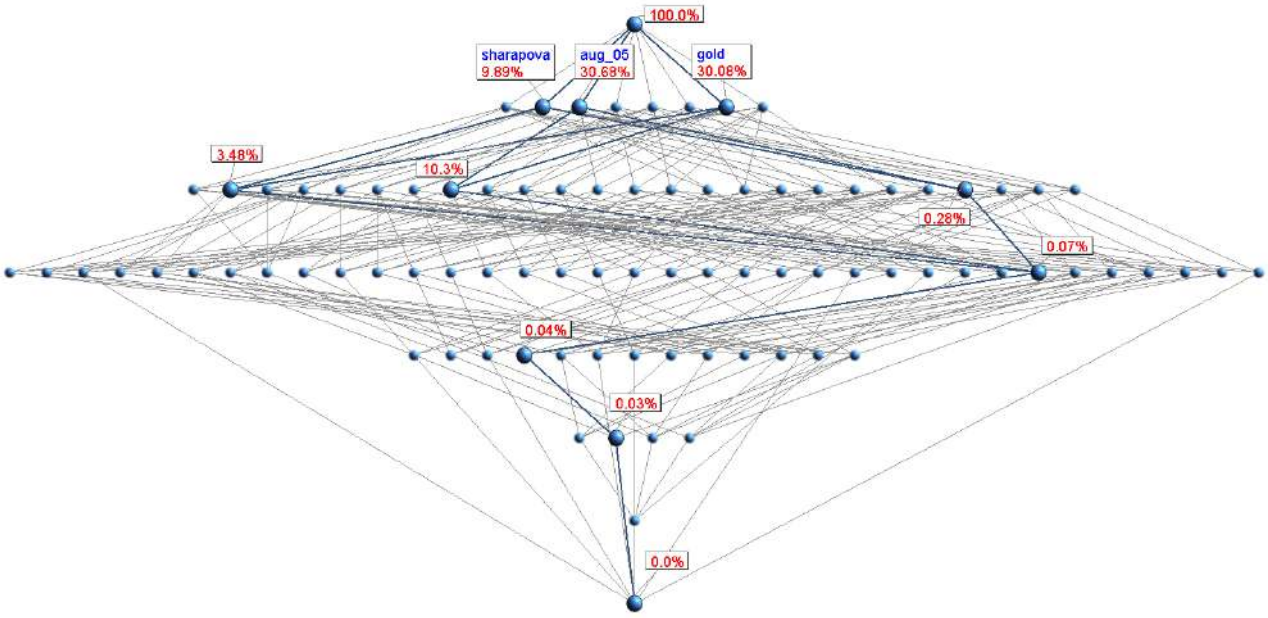

Рисунок 6.10 – Ґратка семантичних концептів із виділеним порядковим ідеалом та фільтром для концепту *{sharapova, aug_05, gold}*

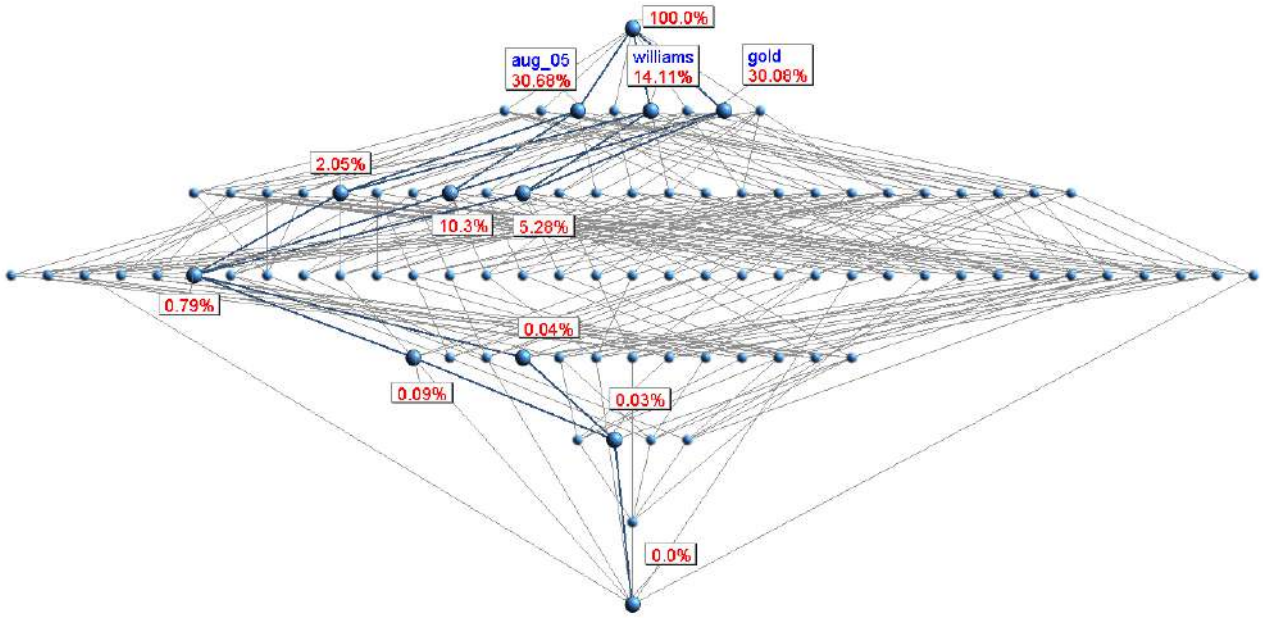

Рисунок 6.11 – Ґратка семантичних концептів із виділеним порядковим ідеалом та фільтром для концепту *{williams, aug_05, gold}*

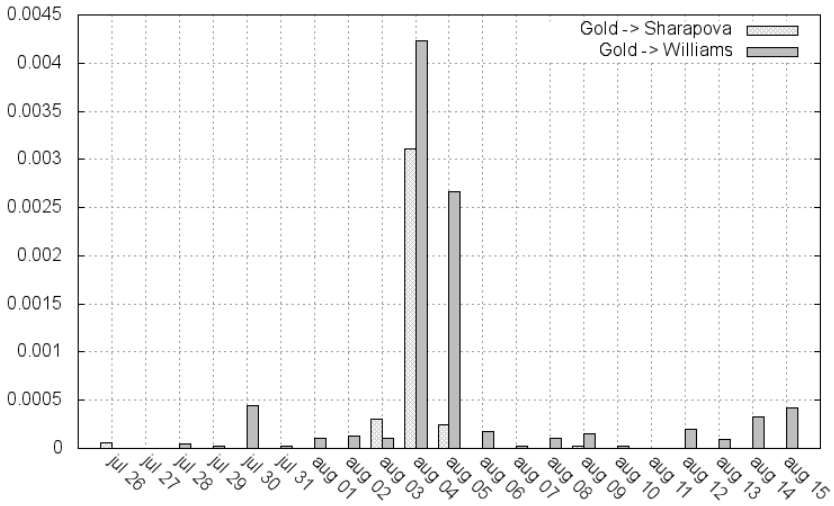

Рисунок 6.12 – Динаміка підтримки для асоціативних правил $Gold \to Sharapova$, $Gold \to Williams$

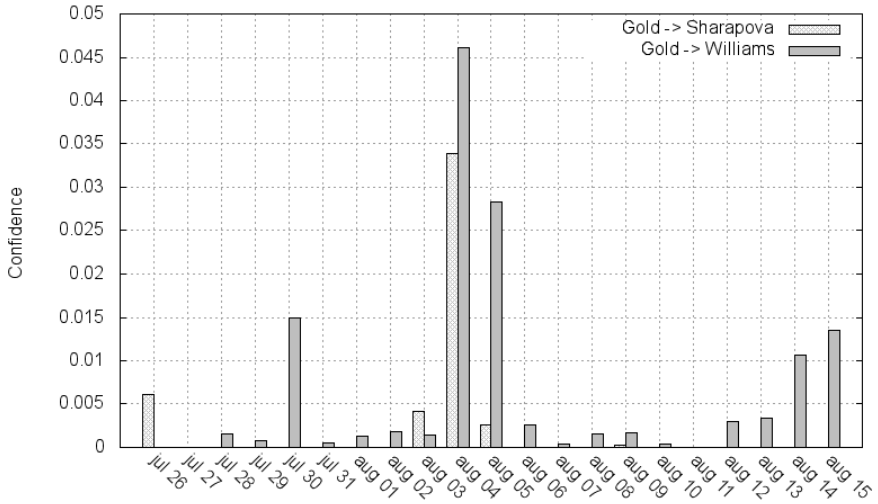

Рисунок 6.13 – Динаміка достовірності для асоціативних правил $Gold \rightarrow Sharapova$, $Gold \rightarrow Williams$

для фіналістів були майже однаковими перед початком турніру і суттєво відмінними після завершення турніру, що відповідає результатам фіналу. Використання моделі формальних концептів в аналізі повідомлень Твіттера дає можливість ефективно виявляти семантичні зв'язки між такими тематичними концептами спортивних подій, як час проведення змагань, стать учасників, вид спорту, імена учасників, результати змагань та імена переможців. Поряд із відображенням реальних фактів, побудована ґратка семантичних концептів відображає прогнози та очікування блогерів.

6.4 Висновки

– Запропонована модель семантичного контексту відображає структурну семантичну організацію текстових масивів. У семантичному контексті формується частково впорядкована множина семантичних концептів, формальний зміст яких визначається семантичними полями, а формальний об'єм – масивами текстових документів. Побудова решітки семантичних концептів у текстових документах дає можливість описувати ієрархічну семантичну структуру в масиві документів та виявляти групи текстових документів, об'єднаних спільною групою семантичних ознак. На основі змістів концептів, які відповідають заданій тематиці, можна сформувати базис семантичного простору для

векторного представлення текстових документів.

– Запропоновано застосування теорії аналізу формальних концептів в інтелектуальній обробці повідомлень мікроблогів. Використання моделі решітки семантичних концептів дає можливість аналізувати семантично зв'язані множини лексем та будувати асоціативні правила. Формування семантичних полів на основі масиву виявлених частих множин дає можливість суттєво звузити пошук асоціативних правил та розмір решітки семантичних концептів в алгоритмах інтелектуального аналізу текстів.

– Розглянуто модель ґратки семантичних концептів для аналізу тегів у повідомленнях, згрупованих за користувачами мікроблогів. Уведення поняття семантичного поля як множини тематично об'єднаних лексем зменшує обсяг необхідних обчислень унаслідок фільтрації масиву повідомлень. Використання аналізу формальних концептів дає можливість утворити алгебраїчну ґратку семантичних концептів, характеристиками яких є об'єм та зміст. Об'єм семантичного концепту об'єднує користувачів, які часто вживають у своїх блогах групи тегів, що утворюють множину змісту цього концепту. Групування тегів користувачів здійснюється на основі множин змістів семантичних концептів. Утворена решітка семантичних концептів дає можливість виявляти асоціативні правила у групах тегів, які є складовими змістів семантичних концептів. Ці правила відображають зв'язки між тегами, які характеризують семантичні поняття у повідомленнях користувачів.

ВИСНОВКИ

У дисертаційній роботі на основі проведених досліджень вирішено актуальну науково-прикладну проблему вибору, поєднання та оптимізації методів інтелектуального аналізу консолідованих даних шляхом розроблення методів моделювання, формування інформативних аналітичних ознак та інтелектуального аналізу табличних та текстових даних з урахуванням предметної області аналізу, що дозволило створювати ефективні прогнозні багаторівневі моделі, розширити інформативність інтелектуального аналізу різнотипних даних та вдосконалити підтримку прийняття рішень у комплексних інформаційно-аналітичних системах. Отримано такі основні результати:

1. Проаналізовано сучасний стан в області інтелектуального аналізу різнотипних даних, сформульовано актуальні питання, обґрунтовано тематику, напрям досліджень та необхідність розроблення нових методів інтелектуального аналізу консолідованих даних.

2. Розроблено комплексний підхід у прогнозній аналітиці табличних даних на основі параметричних та машинно-навчальних моделей, який дає змогу утворювати оптимальний набір аналітичних ознак та формувати ефективний підхід у побудові прогнозних моделей. Розроблено метод об'єднання різнотипних моделей в ансамблі на основі LASSO регресії, який покращує точність прогнозування та стабільність прогнозних результатів і дозволяє підвищити точність у задачах прогнозування, а також зменшити кількість моделей у стекінговому ансамблі на 30% для певного класу задач.

3. Показано ефективність байєсівської регресії для отримання ймовірнісних розподілів параметрів прогнозних моделей. Досліджено, що використання байєсівської регресії на стекінговому рівні дає можливість оцінити невизначеність, яку вносить кожна складова модель ансамблю, що дозволяє формувати оптимальний ансамбль прогнозних моделей.

4. Подальший розвиток отримали підходи в оптимізації послідовності прийняття рішень інтелектуальним агентом на основі глибокого Q-навчання із моделюванням середовища взаємодії параметричної моделі та на основі історичних даних, що дозволяє побудувати процес формування послідовності оптимальних рішень у складних інформаційних середовищах.

5. На основі теорії семантичних полів створено теоретико-множинну модель, яка об'єднує поняття семантичного та тематичного лексемних полів і дає можливість представляти текстові дані у просторі семантичних ознак з метою інтелектуального аналізу заданого семантичного спектру текстових даних. Розроблено метод використання концепції семантичного поля у векторній моделі текстових документів на основі частотно-дистрибутивних семантичних ознак.

6. Розроблено метод кластеризації текстових документів у семантичному просторі, який дає можливість отримувати новий структурний поділ документів за семантичними ознаками. Такий структурний поділ відображає групування документів за їх новими ознаками, зокрема, за авторством текстів.

7. Розроблено метод класифікації текстових даних за експертно сформованими семантичними ознаками, зокрема, квантитативними ознаками семантичних та тематичних полів, що дозволяє проводити інтелектуальний аналіз текстових масивів із відповідними семантичними акцентами та дає можливість за певних умов зменшити кількість семантичних ознак у 3-10 разів у порівнянні з набором лексемних частотних ознак для заданих характеристик точності інтелектуального аналізу текстових даних.

8. Розроблено метод використання семантичних ознак у комбінованих нейромережах із використанням рекурентних підмереж для текстових даних та підмереж із повністю з'єднаними шарами для кількісних ознак, що диверсифікує простір прогнозних ознак в алгоритмах

глибокого навчання та покращує якість інтелектуального аналізу консолідованих даних.

9. Розроблено метод використання генетичних алгоритмів для оптимізації набору семантичних полів, які утворюють векторний простір документів в алгоритмах інтелектуального аналізу текстових даних, що дозволяє формувати ефективні низькорозмірні простори семантичних ознак у задачах інтелектуального аналізу текстових даних.

10. Запропоновано квантовий алгоритм пошуку ключових семантичних образів у масивах текстових об'єктів. Реалізація цього алгоритму здійснюється на основі квантових логічних елементів, зокрема, з використанням вентиля Тоффолі. Ітерація Гровера використовується для підсилення амплітуд квантових станів, які описують семантичні вектори текстових об'єктів. Показано, що реалізація квантових алгоритмів аналізу семантичних образів текстових об'єктів для деякого класу задач дає можливість поліноміально зменшити об'єм обчислень у порівнянні з класичними алгоритмами внаслідок реалізації квантового паралелізму.

11. Розроблено метод використання теорії частих множин та асоціативних правил для формування інформативних ознак у задачах інтелектуального аналізу повідомлень мікроблогів, який дає можливість формувати аналітичні ознаки на основі поєднання лексем у текстах.

12. Розроблено модель семантичного контексту, яка відображає структурну семантичну організацію лексемного складу текстових масивів. У семантичному контексті формується частково впорядкована множина семантичних концептів, формальний зміст яких визначається семантичними полями, а формальний об'єм – текстовими документами. Розроблено метод використання моделі семантичного контексту в аналітиці текстових повідомлень соціальних мереж.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] Breiman L. et al. Statistical modeling: The two cultures (with comments and a rejoinder by the author) // Statistical science. 2001. Vol. 16, № 3. P. 199–231.

[2] Kuhn M., Johnson K. et al. Applied predictive modeling. Springer, 2013. Vol. 26.

[3] Blum A. L., Langley P. Selection of relevant features and examples in machine learning // Artificial intelligence. 1997. Vol. 97, № 1-2. P. 245–271.

[4] An algebra for features and feature composition / S. Apel, C. Lengauer, B. Möller et al. // International Conference on Algebraic Methodology and Software Technology / Springer. 2008. P. 36–50.

[5] Nestorov S., Abiteboul S., Motwani R. Extracting schema from semistructured data // Acm Sigmod Record / ACM. Vol. 27. 1998. P. 295–306.

[6] Chawathe S. S., Abiteboul S., Widom J. Managing historical semistructured data // Theory and practice of object systems. 1999. Vol. 5, № 3. P. 143–162.

[7] Jain A., Zongker D. Feature selection: Evaluation, application, and small sample performance // IEEE transactions on pattern analysis and machine intelligence. 1997. Vol. 19, № 2. P. 153–158.

[8] Losee R. M. Browsing mixed structured and unstructured data // Information processing & management. 2006. Vol. 42, № 2. P. 440–452.

[9] A query language and optimization techniques for unstructured data / P. Buneman, S. Davidson, G. Hillebrand et al. // ACM SIGMOD Record / ACM. Vol. 25. 1996. P. 505–516.

[10] Adding structure to unstructured data / P. Buneman, S. Davidson, M. Fernandez et al. // International Conference on Database Theory / Springer. 1997. P. 336–350.

[11] Mansuri I. R., Sarawagi S. Integrating unstructured data into relational databases // 22nd International Conference on Data Engineering (ICDE'06) / IEEE. 2006. P. 29–29.

[12] Kim J., Xue X., Croft W. B. A probabilistic retrieval model for semistructured data // European conference on information retrieval / Springer. 2009. P. 228–239.

[13] Buneman P. Semistructured data // Proceedings of the sixteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems. 1997. P. 117–121.

[14] Michelson M., Knoblock C. A. Creating relational data from unstructured and ungrammatical data sources // Journal of Artificial Intelligence Research. 2008. Vol. 31. P. 543–590.

[15] Liu H., Yu L. Toward integrating feature selection algorithms for classification and clustering // IEEE Transactions on Knowledge & Data Engineering. 2005. № 4. P. 491–502.

[16] Świniarski R. W. Rough sets methods in feature reduction and classification // International Journal of Applied Mathematics and Computer Science. 2001. Vol. 11. P. 565–582.

[17] Abiteboul S. Querying semi-structured data // International Conference on Database Theory / Springer. 1997. P. 1–18.

[18] Abiteboul S., Buneman P., Suciu D. Data on the Web: from relations to semistructured data and XML. Morgan Kaufmann, 2000.

[19] Efficient substructure discovery from large semi-structured data / T. Asai, K. Abe, S. Kawasoe et al. // IEICE TRANSACTIONS on Information and Systems. 2004. Vol. 87, № 12. P. 2754–2763.

[20] Extracting Semistructured Information from the Web.: Tech. Rep.: / J. Hammer, H. Garcia-Molina, J. Cho et al.: Stanford InfoLab, 1997.

[21] Soderland S. Learning information extraction rules for semi-structured and free text // Machine learning. 1999. Vol. 34, № 1-3. P. 233–272.

[22] Optimizing Data Analysis with a Semi-structured Time Series Database. / L. Bitincka, A. Ganapathi, S. Sorkin et al. // SLAML. 2010. Vol. 10. P. 1–9.

[23] Gandomi A., Haider M. Beyond the hype: Big data concepts, methods, and analytics // International journal of information management. 2015. Vol. 35, № 2. P. 137–144.

[24] Das S. R. et al. Text and context: Language analytics in finance // Foundations and Trends® in Finance. 2014. Vol. 8, № 3. P. 145–261.

[25] Shmueli G., Koppius O. R. Predictive analytics in information systems research // MIS quarterly. 2011. P. 553–572.

[26] Schoenherr T., Speier-Pero C. Data science, predictive analytics, and big data in supply chain management: Current state and future potential // Journal of Business Logistics. 2015. Vol. 36, № 1. P. 120–132.

[27] Big data and predictive analytics for supply chain sustainability: A theory-driven research agenda / B. T. Hazen, J. B. Skipper, J. D. Ezell et al. // Computers & Industrial Engineering. 2016. Vol. 101. P. 592–598.

[28] Fitz-Enz J., John Mattox I. Predictive analytics for human resources. John Wiley & Sons, 2014.

[29] Integrating predictive analytics and social media / Y. Lu, R. Krüger, D. Thom et al. // 2014 IEEE Conference on Visual Analytics Science and Technology (VAST) / IEEE. 2014. P. 193–202.

[30] Social Web mining and exploitation for serious applications: Technosocial Predictive Analytics and related technologies for public health, environmental and national security surveillance / M. N. K. Boulos, A. P. Sanfilippo, C. D. Corley et al. // Computer methods and programs in biomedicine. 2010. Vol. 100, № 1. P. 16–23.

[31] Gundecha P., Liu H. Mining social media: a brief introduction // New Directions in Informatics, Optimization, Logistics, and Production. Informs, 2012. P. 1–17.

[32] Reece A. G., Danforth C. M. Instagram photos reveal predictive markers of depression // EPJ Data Science. 2017. Vol. 6, № 1. P. 1–12.

[33] Siegel E. Predictive analytics: The power to predict who will click, buy, lie, or die. John Wiley & Sons, 2013.

[34] Understanding the predictive power of social media / D. Gayo-Avello, P. T. Metaxas, E. Mustafaraj et al. // Internet Research. 2013.

[35] Bollen J., Mao H., Zeng X. Twitter mood predicts the stock market // Journal of computational science. 2011. Vol. 2, № 1. P. 1–8.

[36] Predictive sentiment analysis of tweets: A stock market application / J. Smailović, M. Grčar, N. Lavrač et al. // International Workshop on Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data / Springer. 2013. P. 77–88.

[37] Schumaker R. P., Chen H. Textual analysis of stock market prediction using breaking financial news: The AZFin text system // ACM Transactions on Information Systems (TOIS). 2009. Vol. 27, № 2. P. 1–19.

[38] Oh Chong, Sheng Olivia. Investigating predictive power of stock micro blog sentiment in forecasting future stock price directional movement. 2011.

[39] Stream-based active learning for sentiment analysis in the financial domain / J. Smailović, M. Grčar, N. Lavrač et al. // Information sciences. 2014. Vol. 285. P. 181–203.

[40] Mao H., Counts S., Bollen J. Predicting financial markets: Comparing survey, news, twitter and search engine data // arXiv preprint arXiv:1112.1051. 2011.

[41] Paye B. S. 'Déjà vol': Predictive regressions for aggregate stock market volatility using macroeconomic variables // Journal of Financial Economics. 2012. Vol. 106, № 3. P. 527–546.

[42] Groß-Klußmann A., Hautsch N. When machines read the news: Using automated text analytics to quantify high frequency news-implied market reactions // Journal of Empirical Finance. 2011. Vol. 18, № 2. P. 321–340.

[43] Stock market prediction from WSJ: text mining via sparse matrix factorization / F. Ming, F. Wong, Z. Liu et al. // 2014 IEEE International Conference on Data Mining / IEEE. 2014. P. 430–439.

[44] Литвин В. В. Бази знань інтелектуальних систем підтримки прийняття рішень // Львів: Видавництво Львівської політехніки. 2011. С. 240.

[45] Шаховська Н. Б. Методи опрацювання консолідованих даних за допомогою просторів даних // Проблеми програмування / Національна академія наук України, Інститут програмних систем НАН України. 2011. № 4. С. 72–84.

[46] Шаховська Н. Б., Пасічник В. В. Сховища та простори даних. Львів: Видавництво Львівської політехніки, 2009. С. 244.

[47] Шаховська Н. Б., Пшеничний О. Ю., Чорней І. М. Проблеми якості консолідованих даних у просторах даних // Системи обробки інформації. 2011. № 3. С. 80–84.

[48] Шаховська Н. Б. Дослідження якості консолідованих даних у просторах даних // Математические машины и системы. 2012. Т. 1, № 1. С. 77–88.

[49] Бодянський Є. В., Тесленко Н. О., Дейнеко А. О. Еволюційна нейронна мережа з ядерними функціями активації й адаптивний алгоритм її навчання // Наукові праці [Чорноморського державного університету імені Петра Могили]. Сер.: Комп'ютерні технології. 2011. № 160, Вип. 148. С. 53–58.

[50] Руденко О. Г., Бодянський Є. В. Штучні нейронні мережі // Харків: Компанія СМІТ. 2006.

[51] Бодянский Е. В., Руденко О. Г. Искусственные нейронные сети: архитектуры, обучение, применения // Харьков: Телетех. 2004. С. 372.

[52] Бідюк П. І., Терентьєв О. М., Коновалюк М. М. Байєсівські мережі в технологіях інтелектуального аналізу даних // Наукові праці. Комп'ютерні технології. Т. 134, № 121. С. 6–16.

[53] Матвійчук А. Моделювання фінансової стійкості підприємств із застосуванням теорій нечіткої логіки, нейронних мереж і дискримінатного аналізу // Вісник НАН України. 2010. № 9. С. 24–46.

[54] Бідюк П. І. Системний підхід до прогнозування на основі моделей часових рядів // Системні дослідження та інформаційні технології. 2003. С. 88–110.

[55] Квєтний Р. Н., Кабачій В. В., Чумаченко О. О. Імовірнісні нейронні мережі в задачах ідентифікації часових рядів // Наукові праці ВНТУ. 2010. № 3. С. 2–6.

[56] Субботін С. О. Подання й обробка знань у системах штучного інтелекту та підтримки прийняття рішень. 2008.

[57] Bishop C. M. Pattern recognition and machine learning. springer, 2006.

[58] Bayesian data analysis / A. Gelman, J. B. Carlin, H. S. Stern et al. Chapman and Hall/CRC, 2013.

[59] Kruschke J. Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan. Academic Press, 2014.

[60] Box G. E., Tiao G. C. Bayesian inference in statistical analysis. John Wiley & Sons, 2011. Vol. 40.

[61] Equation of state calculations by fast computing machines / N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth et al. // The journal of chemical physics. 1953. Vol. 21, № 6. P. 1087–1092.

[62] Plummer M. et al. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling // Proceedings of the 3rd international workshop on distributed statistical computing / Vienna, Austria. Vol. 124. 2003. P. 1–10.

[63] Illustration of Bayesian inference in normal data models using Gibbs sampling / A. E. Gelfand, S. E. Hills, A. Racine-Poon et al. // Journal of the American Statistical Association. 1990. Vol. 85, № 412. P. 972–985.

[64] Geman S., Geman D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images // IEEE Transactions on pattern analysis and machine intelligence. 1984. № 6. P. 721–741.

[65] Gilks W. R., Thomas A., Spiegelhalter D. J. A language and program for complex Bayesian modelling // Journal of the Royal Statistical Society: Series D (The Statistician). 1994. Vol. 43, № 1. P. 169–177.

[66] Hoffman M. D., Gelman A. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. // Journal of Machine Learning Research. 2014. Vol. 15, № 1. P. 1593–1623.

[67] Chen T., Fox E., Guestrin C. Stochastic gradient hamiltonian monte carlo // International conference on machine learning. 2014. P. 1683–1691.

[68] Gelman A., Lee D., Guo J. Stan: A probabilistic programming language for Bayesian inference and optimization // Journal of Educational and Behavioral Statistics. 2015. Vol. 40, № 5. P. 530–543.

[69] An introduction to statistical learning / G. James, D. Witten, T. Hastie et al. Springer, 2013. Vol. 112.

[70] Breiman L. Random forests // Machine learning. 2001. Vol. 45, № 1. P. 5–32.

[71] Friedman J. H. Greedy function approximation: a gradient boosting machine // Annals of statistics. 2001. P. 1189–1232.

[72] Friedman J. H. Stochastic gradient boosting // Computational Statistics & Data Analysis. 2002. Vol. 38, № 4. P. 367–378.

[73] Chen T., Guestrin C. Xgboost: A scalable tree boosting system // Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining / ACM. 2016. P. 785–794.

[74] Lightgbm: A highly efficient gradient boosting decision tree / G. Ke, Q. Meng, T. Finley et al. // Advances in neural information processing systems. 2017. Vol. 30. P. 3146–3154.

[75] Wolpert D. H. Stacked generalization // Neural networks. 1992. Vol. 5, № 2. P. 241–259.

[76] Rokach L. Ensemble-based classifiers // Artificial Intelligence Review. 2010. Vol. 33, № 1-2. P. 1–39.

[77] Sagi O., Rokach L. Ensemble learning: A survey // Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2018. Vol. 8, № 4. P. e1249.

[78] A survey on ensemble learning for data stream classification / H. M. Gomes, J. P. Barddal, F. Enembreck et al. // ACM Computing Surveys (CSUR). 2017. Vol. 50, № 2. P. 23.

[79] Dietterich T. G. Ensemble methods in machine learning // International workshop on multiple classifier systems / Springer. 2000. P. 1–15.

[80] Rokach L. Ensemble methods for classifiers // Data mining and knowledge discovery handbook. Springer, 2005. P. 957–980.

[81] Džeroski S., Ženko B. Is combining classifiers with stacking better than selecting the best one? // Machine learning. 2004. Vol. 54, № 3. P. 255–273.

[82] Seni G., Elder J. Ensemble methods in data mining: improving accuracy through combining predictions. Morgan & Claypool Publishers, 2010.

[83] Zhang C., Ma Y. Ensemble machine learning: methods and applications. Springer, 2012.

[84] Zhou Z.-H. Ensemble methods: foundations and algorithms. CRC press, 2012.

[85] Smyth P., Wolpert D. Stacked density estimation // Advances in neural information processing systems. 1998. P. 668–674.

[86] Ting K. M., Witten I. H. Issues in stacked generalization // Journal of artificial intelligence research. 1999. Vol. 10. P. 271–289.

[87] Kaggle: Your Home for Data Science. URL: `http://kaggle.com`.

[88] Schmidhuber J. Deep learning in neural networks: An overview // Neural networks. 2015. Vol. 61. P. 85–117.

[89] Hinton G. E., Salakhutdinov R. R. Reducing the dimensionality of data with neural networks // science. 2006. Vol. 313, № 5786. P. 504–507.

[90] Rumelhart D. E., Hinton G. E., Williams R. J. Learning representations by back-propagating errors // nature. 1986. Vol. 323, № 6088. P. 533–536.

[91] Dropout: a simple way to prevent neural networks from overfitting / N. Srivastava, G. Hinton, A. Krizhevsky et al. // The journal of machine learning research. 2014. Vol. 15, № 1. P. 1929–1958.

[92] Ivakhnenko A. G. Polynomial theory of complex systems // IEEE transactions on Systems, Man, and Cybernetics. 1971. № 4. P. 364–378.

[93] Ivakhnenko A. G., Ivakhnenko G. A., Muller J. A. Self-organization of neural networks with active neurons // Pattern Recognition and Image Analysis. 1994. Vol. 4, № 2. P. 185–196.

[94] Goodfellow I., Bengio Y., Courville A. Deep learning. MIT press, 2016.

[95] LeCun Y., Bengio Y., Hinton G. Deep learning // nature. 2015. Vol. 521, № 7553. P. 436–444.

[96] Multilayer feedforward networks with a nonpolynomial activation function can approximate any function / M. Leshno, V. Y. Lin, A. Pinkus et al. // Neural networks. 1993. Vol. 6, № 6. P. 861–867.

[97] Hornik K., Stinchcombe M., White H. Multilayer feedforward networks are universal approximators // Neural networks. 1989. Vol. 2, № 5. P. 359–366.

[98] Cybenko G. Approximation by superpositions of a sigmoidal function // Mathematics of control, signals and systems. 1989. Vol. 2, № 4. P. 303–314.

[99] Learning representations by back-propagating errors / D. E. Rumelhart, G. E. Hinton, R. J. Williams et al. // Cognitive modeling. 1988. Vol. 5, № 3. P. 1.

[100] What is the best multi-stage architecture for object recognition? / K. Jarrett, K. Kavukcuoglu, Y. LeCun et al. // 2009 IEEE 12th international conference on computer vision / IEEE. 2009. P. 2146–2153.

[101] Glorot X., Bordes A., Bengio Y. Deep sparse rectifier neural networks // Proceedings of the fourteenth international conference on artificial intelligence and statistics. 2011. P. 315–323.

[102] Srivastava N. Improving neural networks with dropout // University of Toronto. 2013. Vol. 182. P. 566.

[103] Improving neural networks by preventing co-adaptation of feature detectors / G. E. Hinton, N. Srivastava, A. Krizhevsky et al. // arXiv preprint arXiv:1207.0580. 2012.

[104] Wilson D. R., Martinez T. R. The general inefficiency of batch training for gradient descent learning // Neural networks. 2003. Vol. 16, № 10. P. 1429–1451.

[105] Sutton R. S., Barto A. G. et al. Introduction to reinforcement learning. MIT press Cambridge, 1998. Vol. 2.

[106] Human-level control through deep reinforcement learning / V. Mnih, K. Kavukcuoglu, D. Silver et al. // Nature. 2015. Vol. 518, № 7540. P. 529.

[107] Playing atari with deep reinforcement learning / V. Mnih, K. Kavukcuoglu, D. Silver et al. // arXiv preprint arXiv:1312.5602. 2013.

[108] Rana R., Oliveira F. S. Real-time dynamic pricing in a non-stationary environment using model-free reinforcement learning // Omega. 2014. Vol. 47. P. 116–126.

[109] Reinforcement learning for fair dynamic pricing / R. Maestre, J. Duque, A. Rubio et al. // Proceedings of SAI Intelligent Systems Conference / Springer. 2018. P. 120–135.

[110] Vengerov D. A gradient-based reinforcement learning approach to dynamic pricing in partially-observable environments. 2007.

[111] den Boer A. V. Dynamic pricing and learning: historical origins, current research, and new directions // Surveys in operations research and management science. 2015. Vol. 20, № 1. P. 1–18.

[112] Adaptive inventory control models for supply chain management / C. O. Kim, J. Jun, J. Baek et al. // The International Journal of Advanced Manufacturing Technology. 2005. Vol. 26, № 9-10. P. 1184–1192.

[113] Raju C., Narahari Y., Ravikumar K. Reinforcement learning applications in dynamic pricing of retail markets // EEE International Conference on E-Commerce, 2003. CEC 2003. / IEEE. 2003. P. 339–346.

[114] Huang C. Y. Financial Trading as a Game: A Deep Reinforcement Learning Approach // arXiv preprint arXiv:1807.02787. 2018.

[115] Jiang Z., Xu D., Liang J. A deep reinforcement learning framework for the financial portfolio management problem // arXiv preprint arXiv:1706.10059. 2017.

[116] Liu F., Quek C., Ng G. S. Neural network model for time series prediction by reinforcement learning // Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005. / IEEE. Vol. 2. 2005. P. 809–814.

[117] Deep reinforcement learning in large discrete action spaces / G. Dulac-Arnold, R. Evans, H. van Hasselt et al. // arXiv preprint arXiv:1512.07679. 2015.

[118] Lin L.-J. Reinforcement learning for robots using neural networks: Tech. Rep.: : Carnegie-Mellon Univ Pittsburgh PA School of Computer Science, 1993.

[119] Deb K. An efficient constraint handling method for genetic algorithms // Computer methods in applied mechanics and engineering. 2000. Vol. 186, № 2-4. P. 311–338.

[120] Goldberg D. E., Holland J. H. Genetic algorithms and machine learning. 1988.

[121] Whitley D. A genetic algorithm tutorial // Statistics and computing. 1994. Vol. 4, № 2. P. 65–85.

[122] Booker L. B., Goldberg D. E., Holland J. H. Classifier systems and genetic algorithms // Artificial intelligence. 1989. Vol. 40, № 1-3. P. 235–282.

[123] Dimensionality reduction using genetic algorithms / M. L. Raymer, W. F. Punch, E. D. Goodman et al. // IEEE transactions on evolutionary computation. 2000. Vol. 4, № 2. P. 164–171.

[124] A genetic algorithm-based method for feature subset selection / F. Tan, X. Fu, Y. Zhang et al. // Soft Computing. 2008. Vol. 12, № 2. P. 111–120.

[125] Vafaie H., De Jong K. A. Genetic Algorithms as a Tool for Feature Selection in Machine Learning. // ICTAI. 1992. P. 200–203.

[126] Atkinson-Abutridy J., Mellish C., Aitken S. Combining information extraction with genetic algorithms for text mining // IEEE Intelligent Systems. 2004. Vol. 19, № 3. P. 22–30.

[127] Гладков Л. А., Курейчик В. М., Курейчик В. В. Генетические алгоритмы. 2006.

[128] Панченко Т. В. Генетические алгоритмы. 2007.

[129] Батищев Д. И., Неймарк Е. А., Старостин Н. В. Применение генетических алгоритмов к решению задач дискретной оптимизации. Изд-во Нижегор. госуниверситета Н. Новгород, 2006.

[130] Fellbaum C. editeur. Wordnet: An electronic lexical database // Language, Speech, and Communication. MIT Press, Cambrige, MA. 1998.

[131] Miller G. A. WordNet: An electronic lexical database. MIT press, 1998.

[132] Fellbaum C. A semantic network of English verbs // WordNet: An electronic lexical database. 1998. Vol. 3. P. 153–178.

[133] WordNet. URL: `http://wordnet.princeton.edu`.

[134] Tengi R. I. Design and implementation of the WordNet lexical database and searching software // WordNet: an electronic lexical database. 1998. P. 105–127.

[135] Voorhees E. M. Using WordNet for text retrieval // WordNet: an electronic lexical database. 1998. P. 285–303.

[136] Sussna M. Word sense disambiguation for free-text indexing using a massive semantic network // Proceedings of the second international conference on Information and knowledge management. 1993. P. 67–74.

[137] Resnik P. WordNet and class-based probabilities // WordNet: An electronic lexical database. 1998. P. 239–263.

[138] Вердиева З. Н. Семантические поля в современном английском языке: Учебное пособие. Высшая школа, 1986.

[139] Полевые структуры в системе языка. коллективная монография под.ред. проф. З.Д.Попова. Воронеж.: Изд-во Воронежского ун-та, 1989. – 197с., 1989.

[140] Левицкий В. В., Стернин И. А. Экспериментальные методы в семасиологии. Воронеж: Изд-во ВГУ, 192с., 1989.

[141] Русанівський В. М., Широков В. А. Інформаційно-лінгвістичні основи сучасної тлумачної лексикографії.

[142] Pavlyshenko O. The lexical-semantic fields of verbs in English texts // Glottometrics 25. 2013. P. 69.

[143] Скороходько Е. Ф. Сіткове моделювання лексики: лінгвістична інтерпретація параметрів семантичної складності // К.: Мовознавство. 1995. № 6. С. 19–28.

[144] А. Широков В. Семантичні стани мовних одиниць та їх застосування в когнітивній лексикографії // Мовознавство. 2005. Т. 3-4. С. 47–62.

[145] Gliozzo A., Strapparava C. Semantic domains in computational linguistics. Springer Science & Business Media, 2009.

[146] Gliozzo A., Strapparava C., Dagan I. Unsupervised and supervised exploitation of semantic domains in lexical disambiguation // Computer Speech & Language. 2004. Vol. 18, № 3. P. 275–299.

[147] The role of domain information in word sense disambiguation / B. Magnini, C. Strapparava, G. Pezzulo et al. // Natural Language Engineering. 2002. Vol. 8, № 4. P. 359–373.

[148] Distributed representations of words and phrases and their compositionality / T. Mikolov, I. Sutskever, K. Chen et al. // Advances in neural information processing systems. 2013. Vol. 26. P. 3111–3119.

[149] Efficient estimation of word representations in vector space / T. Mikolov, K. Chen, G. Corrado et al. // arXiv preprint arXiv:1301.3781. 2013.

[150] Le Q., Mikolov T. Distributed representations of sentences and documents // International conference on machine learning. 2014. P. 1188–1196.

[151] Turney P. D., Pantel P. From frequency to meaning: Vector space models of semantics // Journal of artificial intelligence research. 2010. Vol. 37. P. 141–188.

[152] Blei D. M., Ng A. Y., Jordan M. I. Latent dirichlet allocation // Journal of machine Learning research. 2003. Vol. 3, № Jan. P. 993–1022.

[153] Sievert C., Shirley K. LDAvis: A method for visualizing and interpreting topics // Proceedings of the workshop on interactive language learning, visualization, and interfaces. 2014. P. 63–70.

[154] Bischof J., Airoldi E. M. Summarizing topical content with word frequency and exclusivity // Proceedings of the 29th International Conference on Machine Learning (ICML-12). 2012. P. 201–208.

[155] The topic browser: An interactive tool for browsing topic models / M. J. Gardner, J. Lutes, J. Lund et al. // Nips workshop on challenges of data visualization / Whistler Canada. Vol. 2. 2010.

[156] Chaney A. J.-B., Blei D. M. Visualizing topic models // Sixth international AAAI conference on weblogs and social media. 2012.

[157] Chuang J., Manning C. D., Heer J. Termite: Visualization techniques for assessing textual topic models // Proceedings of the international working conference on advanced visual interfaces. 2012. P. 74–77.

[158] Topicnets: Visual analysis of large text corpora with topic modeling / B. Gretarsson, J. O'donovan, S. Bostandjiev et al. // ACM Transactions on Intelligent Systems and Technology (TIST). 2012. Vol. 3, № 2. P. 1–26.

[159] Jain A. K., Murty M. N., Flynn P. J. Data clustering: a review // ACM computing surveys (CSUR). 1999. Vol. 31, № 3. P. 264–323.

[160] Xu R., Wunsch D. Clustering. John Wiley & Sons, 2008. Vol. 10.

[161] Rokach L., Maimon O. Clustering methods // Data mining and knowledge discovery handbook. Springer, 2005. P. 321–352.

[162] Beil F., Ester M., Xu X. Frequent term-based text clustering // Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. 2002. P. 436–442.

[163] Aggarwal C. C., Zhai C. A survey of text clustering algorithms // Mining text data. Springer, 2012. P. 77–128.

[164] Ким Д. О., Мьюллер Ч. У., Р. Клекка У. Факторный, дискриминантный и кластерный анализ. М.: Финансы и статистика – 215 с., 1989.

[165] Жамбю М. Иерархический кластер-анализ и соответствия: Пер. с фр. Финансы и статистика – 342 с, 1988.

[166] Анализ данных и процессов: учеб. Пособие / А. А. Брасегян, М. С. Куприянов, И. И. Холод [и др.] // СПб.: БХВ–Петербург – 512 с. 2009.

[167] Sebastiani F. Machine learning in automated text categorization // ACM computing surveys (CSUR). 2002. Vol. 34, № 1. P. 1–47.

[168] Schütze H., Manning C. D., Raghavan P. Introduction to information retrieval. Cambridge University Press Cambridge, 2008. Vol. 39.

[169] Shehata S., Karray F., Kamel M. An efficient concept-based mining model for enhancing text clustering // IEEE Transactions on Knowledge and Data Engineering. 2009. Vol. 22, № 10. P. 1360–1371.

[170] Larsen B., Aone C. Fast and effective text mining using linear-time document clustering // Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining. 1999. P. 16–22.

[171] Agrawal R., Srikant R. et al. Fast algorithms for mining association rules // Proc. 20th int. conf. very large data bases, VLDB. Vol. 1215. 1994. P. 487–499.

[172] Fast discovery of association rules. / R. Agrawal, H. Mannila, R. Srikant et al. // Advances in knowledge discovery and data mining. 1996. Vol. 12, № 1. P. 307–328.

[173] Chui C.-K., Kao B., Hung E. Mining frequent itemsets from uncertain data // Pacific-Asia Conference on knowledge discovery and data mining / Springer. 2007. P. 47–58.

[174] Gouda K., Zaki M. J. Efficiently mining maximal frequent itemsets // Proceedings 2001 IEEE International Conference on Data Mining / IEEE. 2001. P. 163–170.

[175] Srikant R., Vu Q., Agrawal R. Mining association rules with item constraints. // Kdd. Vol. 97. 1997. P. 67–73.

[176] Finding interesting rules from large sets of discovered association rules / M. Klemettinen, H. Mannila, P. Ronkainen et al. // Proceedings of the third international conference on Information and knowledge management. 1994. P. 401–407.

[177] Discovering frequent closed itemsets for association rules / N. Pasquier, Y. Bastide, R. Taouil et al. // International Conference on Database Theory / Springer. 1999. P. 398–416.

[178] Brin S., Motwani R., Silverstein C. Beyond market baskets: Generalizing association rules to correlations // Proceedings of the 1997 ACM SIGMOD international conference on Management of data. 1997. P. 265–276.

[179] Ganter B., Wille R. Formal concept analysis: mathematical foundations. Springer Science & Business Media, 2012.

[180] Cimiano P., Hotho A., Staab S. Learning concept hierarchies from text corpora using formal concept analysis // Journal of artificial intelligence research. 2005. Vol. 24. P. 305–339.

[181] Ganter B., Stumme G., Wille R. Formal concept analysis: foundations and applications. springer, 2005. Vol. 3626.

[182] Formal concept analysis in knowledge processing: A survey on applications / J. Poelmans, D. I. Ignatov, S. O. Kuznetsov et al. // Expert systems with applications. 2013. Vol. 40, № 16. P. 6538–6560.

[183] Formal concept analysis in knowledge processing: A survey on models and techniques / J. Poelmans, S. O. Kuznetsov, D. I. Ignatov et al. // Expert systems with applications. 2013. Vol. 40, № 16. P. 6601–6623.

[184] Kaytoue M., Kuznetsov S. O., Napoli A. Revisiting numerical pattern mining with formal concept analysis // Twenty-Second International Joint Conference on Artificial Intelligence. 2011.

[185] Priss U. Formal concept analysis in information science // Annual review of information science and technology. 2006. Vol. 40, № 1. P. 521–543.

[186] Completing Description Logic Knowledge Bases Using Formal Concept Analysis. / F. Baader, B. Ganter, B. Sertkaya et al. // IJCAI. Vol. 7. 2007. P. 230–235.

[187] Poshyvanyk D., Marcus A. Combining formal concept analysis with information retrieval for concept location in source code // 15th IEEE International Conference on Program Comprehension (ICPC'07) / IEEE. 2007. P. 37–48.

[188] Kuznetsov S. O., Obiedkov S. A. Comparing performance of algorithms for generating concept lattices // Journal of Experimental & Theoretical Artificial Intelligence. 2002. Vol. 14, № 2-3. P. 189–216.

[189] Codd E. F. The relational model for database management: version 2. Addison-Wesley Longman Publishing Co., Inc., 1990.

[190] Date C. J. An introduction to database systems. Pearson Education India, 2004.

[191] Нікольський Ю. В., Пасічник В. В., Щербина Ю. М. Дискретна математика // К.: Видавнича група BHV. 2007. С. 368.

[192] Mentzer J. T., Moon M. A. Sales forecasting management: a demand management approach. Sage, 2004.

[193] Efendigil T., Önüt S., Kahraman C. A decision support system for demand forecasting with artificial neural networks and neuro-fuzzy models: A comparative analysis // Expert Systems with Applications. 2009. Vol. 36, № 3. P. 6697–6707.

[194] Zhang G. P. Neural networks in business forecasting. IGI Global, 2004.

[195] Chatfield C. Time-series forecasting. Chapman and Hall/CRC, 2000.

[196] Brockwell P. J., Davis R. A., Calder M. V. Introduction to time series and forecasting. Springer, 2002. Vol. 2.

[197] Time series analysis: forecasting and control / G. E. Box, G. M. Jenkins, G. C. Reinsel et al. John Wiley & Sons, 2015.

[198] Time series sales forecasting for short shelf-life food products based on artificial neural networks and evolutionary computing / P. Doganis, A. Alexandridis, P. Patrinos et al. // Journal of Food Engineering. 2006. Vol. 75, № 2. P. 196–204.

[199] Hyndman R. J., Athanasopoulos G. Forecasting: principles and practice. OTexts, 2018.

[200] Tsay R. S. Analysis of financial time series. John Wiley & Sons, 2005. P. 543.

[201] Wei W. W. Time series analysis // The Oxford Handbook of Quantitative Methods in Psychology: Vol. 2. 2006.

[202] Arbitrage of forecasting experts / V. Cerqueira, L. Torgo, F. Pinto et al. // Machine Learning. 2018. P. 1–32.

[203] Hyndman R. J., Khandakar Y. et al. Automatic time series for forecasting: the forecast package for R. № 6/07. Monash University, Department of Econometrics and Business Statistics, 2007.

[204] Papacharalampous G. A., Tyralis H., Koutsoyiannis D. Comparison of stochastic and machine learning methods for multi-step ahead forecasting of hydrological processes // Journal of Hydrology. 2017. Vol. 10.

[205] Tyralis H., Papacharalampous G. Variable selection in time series forecasting using random forests // Algorithms. 2017. Vol. 10, № 4. P. 114.

[206] Tyralis H., Papacharalampous G. A. Large-scale assessment of Prophet for multi-step ahead forecasting of monthly streamflow // Advances in Geosciences. 2018. Vol. 45. P. 147–153.

[207] Papacharalampous G., Tyralis H., Koutsoyiannis D. Predictability of monthly temperature and precipitation using automatic time series forecasting methods // Acta Geophysica. 2018. P. 1–25.

[208] A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition / S. B. Taieb, G. Bontempi, A. F. Atiya et al. // Expert systems with applications. 2012. Vol. 39, № 8. P. 7067–7083.

[209] Combining forecasts: An application to elections / A. Graefe, J. S. Armstrong, R. J. Jones Jr et al. // International Journal of Forecasting. 2014. Vol. 30, № 1. P. 43–54.

[210] Armstrong J. S. Combining forecasts: The end of the beginning or the beginning of the end? // International Journal of Forecasting. 1989. Vol. 5, № 4. P. 585–588.

[211] Papacharalampous G., Tyralis H., Koutsoyiannis D. Univariate time series forecasting of temperature and precipitation with a focus on machine learning algorithms: A multiple-case study from Greece // Water Resources Management. 2018. P. 1–33.

[212] Pavlyshenko B. M. Linear, machine learning and probabilistic approaches for time series analysis // Data Stream Mining & Processing (DSMP), IEEE First International Conference on / IEEE. 2016. P. 377–381.

[213] Pavlyshenko B. Using Stacking Approaches for Machine Learning Models // 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP) / IEEE. 2018. P. 255–258.

[214] 'Rossmann Store Sales', Kaggle.Com. URL: `http://www.kaggle.com/c/rossmann-store-sales`.

[215] McKinney W. et al. Data structures for statistical computing in python // Proceedings of the 9th Python in Science Conference / Austin, TX. Vol. 445. 2010. P. 51–56.

[216] McKinney W. Python for data analysis: Data wrangling with Pandas, NumPy, and IPython. " O'Reilly Media, Inc.", 2012.

[217] Scikit-learn: Machine Learning in Python / F. Pedregosa, G. Varoquaux, A. Gramfort et al. // Journal of Machine Learning Research. 2011. Vol. 12. P. 2825–2830.

[218] Oliphant T. E. A guide to NumPy. Trelgol Publishing USA, 2006. Vol. 1.

[219] Chollet F. et al. Keras. `https://keras.io`. 2015.

[220] Hunter J. D. Matplotlib: A 2D graphics environment // Computing in Science & Engineering. 2007. Vol. 9, № 3. P. 90–95.

[221] mwaskom/seaborn: v0. 8.1 (September 2017) / M. Waskom, O. Botvinnik, D. O'Kane et al. // Zenodo, doi. 2017. Vol. 10. URL: https://doi.org/10.5281/zenodo.883859.

[222] Pavlyshenko B. M. Machine-learning models for sales time series forecasting // Data. 2019. Vol. 4, № 1. P. 15.

[223] Kaggle competition 'Grupo Bimbo Inventory Demand '. URL: `https://www.kaggle.com/c/grupo-bimbo-inventory-demand`.

[224] Kaggle competition 'Grupo Bimbo Inventory Demand' #1 Place Solution of The Slippery Appraisals team. URL: `https://www.kaggle.com/c/grupo-bimbo-inventory-demand/discussion/23863`.

[225] Kaggle competition 'Grupo Bimbo Inventory Demand' Bimbo XGBoost R script LB:0.457. URL: `https://www.kaggle.com/bpavlyshenko/bimbo-xgboost-r-script-lb-0-457`.

[226] Pavlyshenko B. Predictive Analytics for Sales Time Series // Xth International Scientific and Practical Conference "Electronics and Information Technologies" (ELIT-2018) August 30 - September 2, 2018, Lviv, Karpaty village, Issue 10. 2018. P. 85–87.

[227] Stan: a probabilistic programming language. / B. Carpenter, A. Gelman, M. D. Hoffman et al. // Grantee Submission. 2017. Vol. 76, № 1. P. 1–32.

[228] Pavlyshenko B. Bayesian Regression Approach for Building and Stacking Predictive Models in Time Series Analytics // International Conference on Data Stream Mining and Processing / Springer. 2020. P. 486–500.

[229] Gal Y., Ghahramani Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning // international conference on machine learning. 2016. P. 1050–1059.

[230] Gal Y., Ghahramani Z. A theoretically grounded application of dropout in recurrent neural networks // Advances in neural information processing systems. 2016. P. 1019–1027.

[231] Pavlyshenko B. Using Bayesian Regression for Stacking Time Series Predictive Models // 2020 IEEE Third International Conference on Data Stream Mining & Processing (DSMP). 2020. P. 305–309.

[232] SciPy 1.0–Fundamental Algorithms for Scientific Computing in Python / P. Virtanen, R. Gommers, T. E. Oliphant et al. // arXiv e-prints. 2019. Jul. P. arXiv:1907.10121.

[233] Seabold S., Perktold J. statsmodels: Econometric and statistical modeling with python // 9th Python in Science Conference. 2010.

[234] Kristoufek L. BitCoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era // Scientific reports. 2013. Vol. 3. P. 3415.

[235] Bouoiyour J., Selmi R., Tiwari A. K. Is Bitcoin business income or speculative foolery? New ideas through an improved frequency domain analysis // Annals of Financial Economics. 2015. Vol. 10, № 01. P. 1550002.

[236] What drives Bitcoin price / J. Bouoiyour, R. Selmi, A. K. Tiwari et al. // Economics Bulletin. 2016. Vol. 36, № 2. P. 843–850.

[237] Matta M., Lunesu I., Marchesi M. Bitcoin Spread Prediction Using Social and Web Search Media. // UMAP Workshops. 2015.

[238] Dyhrberg A. H. Bitcoin, gold and the dollar–A GARCH volatility analysis // Finance Research Letters. 2016. Vol. 16. P. 85–92.

[239] Shah D., Zhang K. Bayesian regression and Bitcoin // Communication, Control, and Computing (Allerton), 2014 52nd Annual Allerton Conference on / IEEE. 2014. P. 409–414.

[240] Barber B. M., Odean T. All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors // The review of financial studies. 2007. Vol. 21, № 2. P. 785–818.

[241] Grinberg R. Bitcoin: An innovative alternative digital currency // Hastings Sci. & Tech. LJ. 2012. Vol. 4. P. 159.

[242] Kroll J. A., Davey I. C., Felten E. W. The economics of Bitcoin mining, or Bitcoin in the presence of adversaries // Proceedings of WEIS. Vol. 2013. 2013. P. 11.

[243] Ciaian P., Rajcaniova M., Kancs A. The economics of BitCoin price formation // Applied Economics. 2016. Vol. 48, № 19. P. 1799–1815.

[244] Malkiel B. G. The efficient market hypothesis and its critics // Journal of economic perspectives. 2003. Vol. 17, № 1. P. 59–82.

[245] Pavlyshenko B. M. Bitcoin Price Predictive Modeling Using Expert Correction // 2019 XIth International Scientific and Practical Conference on Electronics and Information Technologies (ELIT),September 16 – 18, 2019 Lviv, Ukraine. 2019. P. 163–167.

[246] Pavlyshenko B. M. Modeling COVID-19 Spread and Its Impact on Stock Market Using Different Types of Data // Electronics and information technologies. 2020. № 14. P. 3–21.

[247] 'Bosch Production Line Performance', Kaggle.Com. URL: `https://www.kaggle.com/c/bosch-production-line-performance`.

[248] Pavlyshenko B. Machine learning, linear and Bayesian models for logistic regression in failure detection problems // Big Data (Big Data), 2016 IEEE International Conference on / IEEE. 2016. P. 2046–2050.

[249] Pavlyshenko B. M. Detection of Technical Failures on Production Lines Using Machine Learning, Linear and Bayesian Models of Logistic Regression // Electronics and information technologies. 2019. № 12. P. 3–19.

[250] 'Kaggle competition "Bosch Production Line Performance". The Magical Feature : from LB 0.3- to 0.4+', Kaggle.Com. URL: https://www.kaggle.com/c/bosch-production-line-performance/forums/t/24065/the-magical-feature-from-lb-0-3-to-0-4.

[251] 'Kaggle competition "Bosch Production Line Performance". Road-2-0.4+, Kaggle.Com. URL: https://www.kaggle.com/mmueller/bosch-production-line-performance/road-2-0-4.

[252] 'Kaggle competition "Kaggle competition "Bosch Production Line Performance". Road-2-0.4+ -> FeatureSet++, Kaggle.Com. URL: https://www.kaggle.com/alexxanderlarko/bosch-production-line-performance/road-2-0-4-featureset.

[253] Friedman J., Hastie T., Tibshirani R. Regularization paths for generalized linear models via coordinate descent // Journal of statistical software. 2010. Vol. 33, № 1. P. 1.

[254] Regularization paths for Cox's proportional hazards model via coordinate descent / N. Simon, J. Friedman, T. Hastie et al. // Journal of statistical software. 2011. Vol. 39, № 5. P. 1.

[255] Strong rules for discarding predictors in lasso-type problems / R. Tibshirani, J. Bien, J. Friedman et al. // Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2012. Vol. 74, № 2. P. 245–266.

[256] Hastie T., Qian J. Glmnet vignette // Retrieve from https://web.stanford.edu/ hastie/glmnet/glmnet_alpha.html. 2014.

[257] Plummer Martyn. JAGS Version 4.3. 0 user manual // URL: http://www.stat.yale.edu/ jtc5/238/materials/jags_4.3.0_manual_with_dist 2012.

[258] Pavlyshenko B. M. Sales Time Series Analytics Using Deep Q-learning // International Journal of Computing. 2020. Sep. Vol. 19, № 3. P. 434–441. URL: https://computingonline.net/computing/article/view/1892.

[259] Github Repository. URL: `https://github.com/dennybritz/reinforcement-learning`. On-line resource, accessed 5 December 2019.

[260] Github Repository. URL: `https://github.com/keon/deep-q-learning`. On-line resource, accessed 5 December 2019.

[261] Github Repository. URL: `https://github.com/rlcode/reinforcement-learning`. On-line resource, accessed 5 December 2019.

[262] Павлишенко Б. М. Використання лексемних полів у інтелектуальному аналізі текстових масивів // Штучний інтелект. 2013. № 1. С. 98–109.

[263] Павлишенко Б. М. Використання концепції семантичного поля у векторній моделі текстових документів // Східно-Європейский журнал передових технологій. 2011. Т. 6, № 2. С. 7–11.

[264] Павлишенко Б. М. Використання методів машинного навчання та семантичних ознак в інтелектуальному аналізі текстових даних // Електроніка та інформаційні технології. 2020. № 13. С. 3–18.

[265] Project Gutenberg. URL: `https://www.gutenberg.org`.

[266] Abdi H., Williams L. J. Principal component analysis // Wiley interdisciplinary reviews: computational statistics. 2010. Vol. 2, № 4. P. 433–459.

[267] Maaten L. v. d., Hinton G. Visualizing data using t-SNE // Journal of machine learning research. 2008. Vol. 9, № Nov. P. 2579–2605.

[268] The 20 Newsgroups data set. URL: http://qwone.com/~jason/20Newsgroups/.

[269] Миркин Б. Г. Анализ качественных признаков и структур. М.: Статистика, 319с, 1980.

[270] Заде Лютфи, Ринго Н. И. Понятие лингвистической переменной и его применение к принятию приближенных решений. М.: Мир – 165 с., 1976.

[271] Конышева Л. К., Назаров Д. М. Основы теории нечетких множеств. 2011.

[272] Павлишенко Б. М. Моделювання нечітких семантичних полів у масивах текстових документів // Системи обробки інформації. 2011. № 8. С. 175–178.

[273] Павлишенко Б. М. Модель нечітких семантичних полів для інтелектуального аналізу текстових масивів // IV науково–практична конференція "Електроніка та інформаційні технології (ЕлІТ–2012)": тези доповідей, 30 серпня–2 вересня 2012 р. – Львів–Чинадієво. 2012. С. 98.

[274] Павлишенко Б. М. Модель вторинних некорельованих семантичних полів для аналізу текстових даних // Системні дослідження та інформаційні технології. 2014. № 3. С. 130–138.

[275] Jolliffe I. T., Cadima J. Principal component analysis: a review and recent developments // Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 2016. Vol. 374, № 2065. P. 20150202.

[276] Jolliffe I. T. Principal components in regression analysis // Principal component analysis. Springer, 1986. P. 129–155.

[277] Indexing by latent semantic analysis / S. Deerwester, S. T. Dumais, G. W. Furnas et al. // Journal of the American society for information science. 1990. Vol. 41, № 6. P. 391–407.

[278] Mirzal A. Clustering and latent semantic indexing aspects of the singular value decomposition // International Journal of Information and Decision Sciences. 2016. Vol. 8, № 1. P. 53–72.

[279] Landauer T. K., Foltz P. W., Laham D. An introduction to latent semantic analysis // Discourse processes. 1998. Vol. 25, № 2-3. P. 259–284.

[280] Павлишенко Б. М. Сингулярна декомпозиція матриці семантичних ознак в алгоритмі ієрархічної кластеризації текстових масивів // Математичні машини і системи. 2012. Т. 1, № 1. С. 69–76.

[281] Modeling text with generalizable Gaussian mixtures / L. K. Hansen, S. Sigurdsson, T. Kolenda et al. // 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. № 00CH37100) / IEEE. Vol. 6. 2000. P. 3494–3497.

[282] Hofmann T. Probabilistic latent semantic indexing // Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. 1999. P. 50–57.

[283] The author-topic model for authors and documents / M. Rosen-Zvi, T. Griffiths, M. Steyvers et al. // Proceedings of the 20th conference on Uncertainty in artificial intelligence. 2004. P. 487–494.

[284] Zhai ChengXiang, Velivelli Atulya, Yu Bei. A cross-collection mixture model for comparative text mining // Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. 2004. С. 743–748.

[285] Mei Q., Zhai C. A mixture model for contextual text mining // Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. 2006. P. 649–655.

[286] Tatiana B., David H., Derek Y. mixtools: An R package for analyzing finite mixture models // Journal of Statistical Software. 2009. Vol. 32, № 6. P. 1–29.

[287] Pavlyshenko B. The Distribution of Semantic Fields in Author's Texts // Cybernetics and Information Technologies. 2016. Vol. 16, № 3. P. 195–204.

[288] Rehurek R., Sojka P. Software framework for topic modelling with large corpora // In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks / Citeseer. 2010.

[289] Pavlyshenko B. Clustering of Authors' Texts of English Fiction in the Vector Space of Semantic Fields // Cybernetics and Information Technologies. 2014. Vol. 14, № 3. P. 25–36.

[290] Павлишенко Б. Семантична кластеризація текстових документів методом k–середніх // Комп'ютерні науки та інформаційні технології. 2011. № 710. С. 215–218.

[291] Павлишенко Б. М. Кластерний аналіз повідомлень груп новин у просторі семантичних ознак // Комп'ютерні системи та мережі. 2012. № 745. С. 148–155.

[292] Павлишенко Б. Ієрархічна кластеризація текстових документів у векторному просторі семантичних полів // Електроніка та інформаційні технології. 2011. № 1. С. 212–222.

[293] Павлишенко Б. М. Алгоритми семантичної векторизації та кластеризації текстових масивів // Друга Всеукраїнська науково–практична конференція "Проблеми електроніки та інформаційні технології", 02–05 вересня 2010 р. – Львів–Чинадієво. 2010. С. А12.

[294] Павлишенко Б. М. Кластерний аналіз текстових документів в просторі семантичних концептів // Збірник доповідей науково–практичної конференції з міжнародню участю "Системи підтримки прийняття рішень. Теорія і практика", 6 червня 2011 р. – Київ. 2011. С. 146–149.

[295] Pavlyshenko B. Classification analysis of authorship fiction texts in the space of semantic fields // Journal of Quantitative Linguistics. 2013. Vol. 20, № 3. P. 218–226.

[296] Павлишенко Б. Ймовірнісна класифікація текстових документів у просторі семантичних полів // Електроніка та інформаційні технології. 2012. № 2. С. 164–172.

[297] Павлишенко Б. Класифікація повідомлень груп новин у векторному просторі семантичних полів // Комп'ютерні науки та інформаційні технології. 2012. № 744. С. 294–302.

[298] Halko N., Martinsson P.-G., Tropp J. A. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions // SIAM review. 2011. Vol. 53, № 2. P. 217–288.

[299] Gers F. A., Schraudolph N. N., Schmidhuber J. Learning precise timing with LSTM recurrent networks // Journal of machine learning research. 2002. Vol. 3, № Aug. P. 115–143.

[300] Sundermeyer M., Schlüter R., Ney H. LSTM neural networks for language modeling // Thirteenth annual conference of the international speech communication association. 2012.

[301] Airline Twitter sentiment. URL: `https://www.figure-eight.com/data-for-everyone/,https://www.kaggle.com/crowdflower/twitter-airline-sentiment`.

[302] Mercari Price Suggestion Challenge, Kaggle.com. URL: `https://www.kaggle.com/c/mercari-price-suggestion-challenge`.

[303] Pavlyshenko B. Genetic Optimization of Keyword Subsets in the Classification Analysis of Authorship of Texts // Journal of Quantitative Linguistics. 2014. Vol. 21, № 4. P. 341–349.

[304] Павлишенко Б. М. Формування базису семантичного простору текстових документів за допомогою генетичних алгоритмів // Математичні машини і системи. 2013. № 2. С. 96–104.

[305] Automatic text summarization with genetic algorithm-based attribute selection / C. N. Silla, G. L. Pappa, A. A. Freitas et al. // Ibero-American Conference on Artificial Intelligence / Springer. 2004. P. 305–314.

[306] Manning C. D., Raghavan P., Schütze H. Introduction to information retrieval. Cambridge university press, 2008.

[307] MATLAB. URL: https://www.mathworks.com. 2010.

[308] Quantum computers / T. D. Ladd, F. Jelezko, R. Laflamme et al. // Nature. 2010. Vol. 464, № 7285. P. 45–53.

[309] Preskill J. Reliable quantum computers // Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences. 1998. Vol. 454, № 1969. P. 385–410.

[310] Grover L. K. Quantum mechanics helps in searching for a needle in a haystack // Physical review letters. 1997. Vol. 79, № 2. P. 325.

[311] Grover L. K. Quantum computers can search arbitrarily large databases by a single query // Physical review letters. 1997. Vol. 79, № 23. P. 4709.

[312] Zalka C. Grover's quantum searching algorithm is optimal // Physical Review A. 1999. Vol. 60, № 4. P. 2746.

[313] Крохмальський Т. Квантові комп'ютери: основи й алґоритми (короткий огляд) // Журнал фізичних досліджень. 2004. Т. 8. С. 1–15.

[314] Китаев А., Шень А., Вялый М. Классические и квантовые вычисления // М.: МЦНМО. 1999. Т. 192.

[315] Castelvecchi D. IBM's quantum cloud computer goes commercial // Nature News. 2017. Vol. 543, № 7644. P. 159.

[316] Павлишенко Б. Квантовий алґоритм еволюційного аналізу одновимірних кліткових автоматів // Журнал фізичних досліджень. 2011. Т. 15, № 3. С. 1–6.

[317] Павлишенко Б. Числове моделювання алгоритму Гровера для квантового пошуку даних // Теоретична електротехніка. 2010. № 61. С. 49–59.

[318] Qiskit: An open-source framework for quantum computing / G. Aleksandrowicz, T. Alexander, P. Barkoutsos et al. // Accessed on: Mar. 2019. Vol. 16.

[319] Cross A. The IBM Q experience and QISKit open-source quantum computing software // Bulletin of the American Physical Society. 2018. Vol. 63.

[320] Koch D., Wessing L., Alsing P. M. Introduction to Coding Quantum Algorithms: A Tutorial Series Using Qiskit // arXiv preprint arXiv:1903.04359. 2019.

[321] Qiskit backend specifications for OpenQASM and OpenPulse experiments / D. C. McKay, T. Alexander, L. Bello et al. // arXiv preprint arXiv:1809.03452. 2018.

[322] An efficient quantum circuits optimizing scheme compared with QISKit / X. Zhang, H. Xiang, T. Xiang et al. // arXiv preprint arXiv:1807.01703. 2018.

[323] Павлишенко Б. М. Аналіз семантичних образів у масивах текстових об'єктів за допомогою квантових обчислень // Математичні машини і системи. 2013. № 1. С. 34–43.

[324] Павлишенко Б. М. Квантовий алгоритм пошуку ключових слів у масивах текстових даних // Біоніка інтелекту. 2011. № 3(77). С. 157–161.

[325] Павлишенко Б. М. Пошук частих множин семантичних ознак та асоціативних правил в повідомленнях мікроблогів // Нові технології. 2011. № 3(33). С. 82–86.

[326] Csardi G., Nepusz T. et al. The igraph software package for complex network research // InterJournal, complex systems. 2006. Vol. 1695, № 5. P. 1–9.

[327] Pons P., Latapy M. Computing communities in large networks using random walks // International symposium on computer and information sciences / Springer. 2005. P. 284–293.

[328] Fruchterman T. M., Reingold E. M. Graph drawing by force-directed placement // Software: Practice and experience. 1991. Vol. 21, № 11. P. 1129–1164.

[329] Mahmud J. IBM Watson Personality Insights: The science behind the service: Tech. Rep.: : Technical report, IBM, 2016.

[330] Why we twitter: understanding microblogging usage and communities / A. Java, X. Song, T. Finin et al. // Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis. 2007. P. 56–65.

[331] What is Twitter, a social network or a news media? / H. Kwak, C. Lee, H. Park et al. // Proceedings of the 19th international conference on World wide web. 2010. P. 591–600.

[332] Newman M. E., Park J. Why social networks are different from other types of networks // Physical review E. 2003. Vol. 68, № 3. P. 036122.

[333] Measuring user influence in twitter: The million follower fallacy / M. Cha, H. Haddadi, F. Benevenuto et al. // fourth international AAAI conference on weblogs and social media. 2010.

[334] Characterizing user behavior in online social networks / F. Benevenuto, T. Rodrigues, M. Cha et al. // Proceedings of the 9th ACM SIGCOMM conference on Internet measurement. 2009. P. 49–62.

[335] Pak A., Paroubek P. Twitter as a corpus for sentiment analysis and opinion mining. // LREc. Vol. 10. 2010. P. 1320–1326.

[336] Asur S., Huberman B. A. Predicting the future with social media // 2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology / IEEE. Vol. 1. 2010. P. 492–499.

[337] Mishne G., Glance N. S. et al. Predicting movie sales from blogger sentiment. // AAAI spring symposium: computational approaches to analyzing weblogs. 2006. P. 155–158.

[338] Shamma D., Kennedy L., Churchill E. Tweetgeist: Can the twitter timeline reveal the structure of broadcast events // CSCW Horizons. 2010. P. 589–593.

[339] Gentry Jeff. Package 'twitteR'. URL: `https://cran.r-project.org/web/packages/twitteR/`.

[340] Hahsler M., Gruen B., Hornik K. arules – A Computational Environment for Mining Association Rules and Frequent Item Sets // Journal of Statistical Software. 2005. October. Vol. 14, № 15. P. 1–25.

[341] Hahsler M. arulesViz: Interactive Visualization of Association Rules with R // R Journal. 2017. Vol. 9, № 2. P. 163–175. URL: https://journal.r-project.org/archive/2017/RJ-2017-047/RJ-2017-047.pdf.

[342] Pavlyshenko B. M. Forecasting of Events by Tweets Data Mining // Electronics and information technologies. 2018. № 10. P. 71–85.

[343] Pavlyshenko B. M. Can Twitter Predict Royal Baby's Name? // Electronics and information technologies. 2019. № 11. P. 52–60.

[344] Mannila H., Toivonen H., Verkamo A. I. Efficient algorithms for discovering association rules // KDD-94: AAAI workshop on Knowledge Discovery in Databases / Citeseer. 1994. P. 181–192.

[345] Clauset A., Newman M. E., Moore C. Finding community structure in very large networks // Physical review E. 2004. Vol. 70, № 6. P. 066111.

[346] Павлишенко Б. М. Модель семантичного контексту в алгоритмах інтелектуального аналізу текстів // Комп'ютинг. 2011. Т. 10, № 3. С. 216–222.

[347] Павлишенко Б. М. Групування тегів користувачів мікроблогів на основі решітки семантичних концептів // Комп'ютерні системи та мережі. 2011. № 717. С. 120–124.

[348] Павлишенко Б. М. Групування текстових даних на основі моделі семантичного контексту // Східно-Європейский журнал передових технологій. 2011. № 5 (2). С. 39–42.

[349] Павлишенко Б. М. Модель решітки семантичних концептів для інтелектуального аналізу мікроблогів // Штучний інтелект. 2012. № 1. С. 103–111.

[350] Павлишенко Б. М. Аналіз мікроблогів користувачів на основі гратки семантичних концептів // Збірник доповідей науково–практичної конференції з міжнародню участю "Системи підтримки прийняття рішень. Теорія і практика", 6 червня 2012 р. – Київ. 2012. С. 115–118.

[351] El Qadi A., Aboutajdine D., Ennouary Y. Formal Concept Analysis for Information Retrieval // International Journal of Computer Science & Information Security. 2010. Vol. 7, № 2. P. 119–125.

[352] Lahcen B., Kwuida L. Lattice miner: a tool for concept lattice construction and exploration // Suplementary Proceeding of International Conference on Formal concept analysis (ICFCA'10). 2010.

[353] Павлишенко Б. М. Інтелектуальний аналіз мікроблогів за допомогою решітки семантичних концептів // 5 міжнародна науково–технічна конференція ACSN–2011 "Сучасні комп'ютерні системи та мережі: розробка та використання": тези доповідей, 29 вересня – 1 жовтня 2011 р. – Львів. 2011. С. 85–87.

[354] Gnuplot: an interactive plotting program. URL: `http://gnuplot.info/`.

[355] 'M5 Forecasting - Accuracy', Kaggle.Com. URL: `https://www.kaggle.com/c/m5-forecasting-accuracy`.

[356] COVID I., Murray C. J. et al. Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by US state in the next 4 months. URL: `https://www.medrxiv.org/content/10.1101/2020.03.27.20043752v1`.

[357] COVID19 Global Forecasting (Week 2). Kaggle.Com. URL: https://www.kaggle.com/c/covid19-global-forecasting-week-2.

[358] CSSE COVID-19 Dataset. GitHub.Com. URL: https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data.

[359] Coronavirus (Covid-19) Data in the United States. GitHub.Com. URL: https://github.com/nytimes/covid-19-data.

[360] COVID-19 reports. URL: https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/.

[361] COVID-19 Kaggle community contributions. Kaggle.Com. URL: https://www.kaggle.com/covid-19-contributions.

А ДОДАТКИ

А.1 Враховування стохастичних патернів у прогнозній аналітиці часових рядів

Стохастичність часових рядів частково зумовлена факторами впливу, які не враховуються як ознаки в прогнозних моделях. Це може бути, наприклад погода, непередбачувана поведінка конкурентів тощо. Для враховування таких патернів в якості прогнозних ознак можна використовувати зміщені в часі значення цільової змінної, або іншими словами – лаги цільової змінної. Використовуючи такий підхід, можна виявити нові додаткові патерни, які зумовлені невідомими ознаками. Важливим є вибір оптимального підходу для таких задач. Для нашого дослідження ми використали 1000 довільно вибраних часових рядів із набору даних із змагання по прогнозуванні на платформі Kaggle 'M5 Forecasting - Accuracy' [355]. Ці часові ряди описують продажі в магазинах мережі Walmart. Як цільові змінні використано значення продажів товарів у грошовому вираженні. На рис. А.1 представлені типові часові ряди продажів. Динаміка продажів деяких товарів є переривиста, що може бути спричинено стохастичною присутністю товарів на полицях. Як лагові ознаки, ми використовували змінну, визначену як 10-денне зміщення цільової змінної та 15-денне ковзне середнє для цієї визначеної лагової змінної. Розглянемо прогнозування на 10 днів вперед. Це дає можливість використовувати пряме прогнозування без рекурсії, оскільки лагові змінні, мають лаг 10 днів. Ми використали алгоритм машинного навчання LightGBM [74] для регресійного підходу в прогнозуванні часових рядів. На рис. А.2 показано значимість ознак в алгоритмі LightGBM для моделі без лагових ознак. RMSE для цього випадку рівне 8.03. На рис. А.3 показано значимість ознак в алгоритмі LightGBM для моделі із описаними вище лаговими ознаками. RMSE ля цього випадку рівне 7.70. Можна бачити, що використання лагових ознак дає покращення для валідаційної оцінки. Пряме прогнозування не можна використовувати у випадку, якщо горизонт прогнозування перевищує найменший лаг цільової змінної. Для таких випадків можна використовувати два підходи. Перший підхід базується на використанні рекурсії для розрахунку лагових ознак на основі прогнозованих значень цільової змінної. Другий підхід базується на прямому прогнозуванні з використанням різних моделей з різними лаговими ознаками для різних часових періодів прогнозування. Розглянемо застосування рекурсійного підходу. Щоб порівняти результати у випадках з прямим прогнозуванням та із застосуванням рекурсії, використано однаковий часовий період та однакові метапараметри алгоритму LightGBM в обох випадках. Розраховувався прогноз на один місяць вперед. У цьому випадку для прямого прогнозування без лагових ознак оцінка RMSE=8,14, у прогнозуванні з рекурсією RMSE=7,48. Отже, використання лагових ознак дозволяє отримати більш точний результат прогнозування за умови врахування стохастичних патернів. У наступному дослідженні ми додали додаткові лагові ознаки

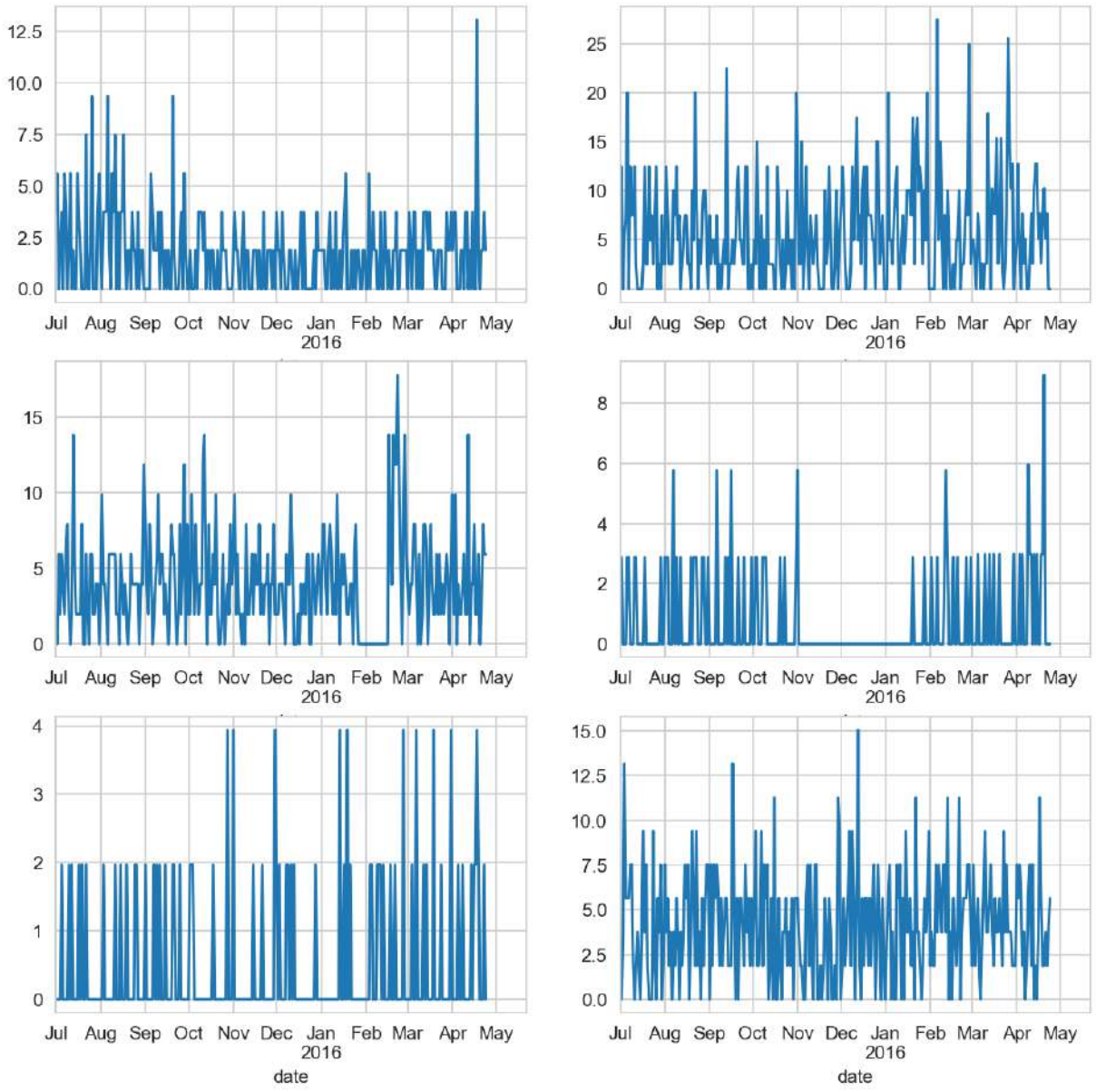

Рисунок А.1 – Типові часові ряди продажів

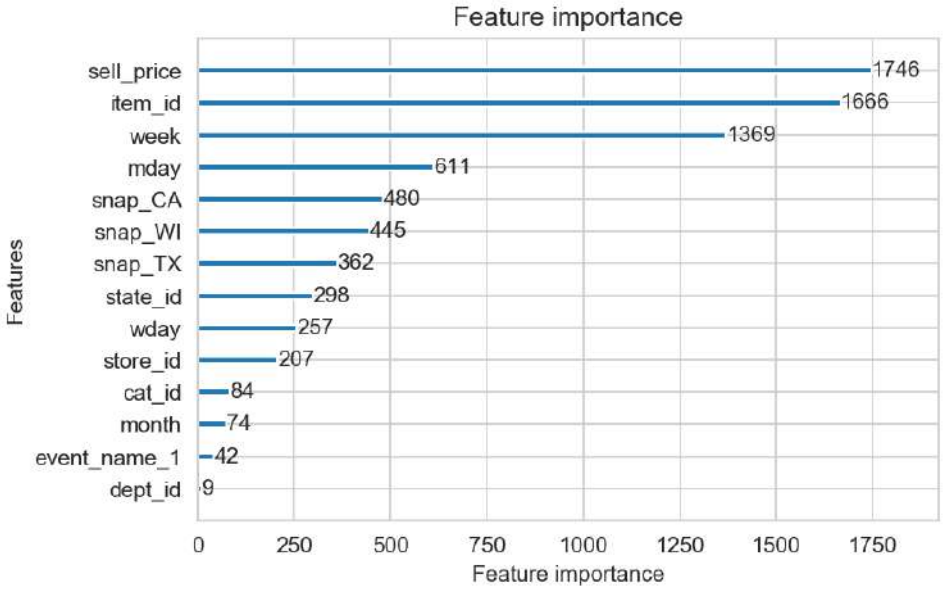

Рисунок А.2 – Значимість ознак в алгоритмі LightGBM для моделі без лагових ознак

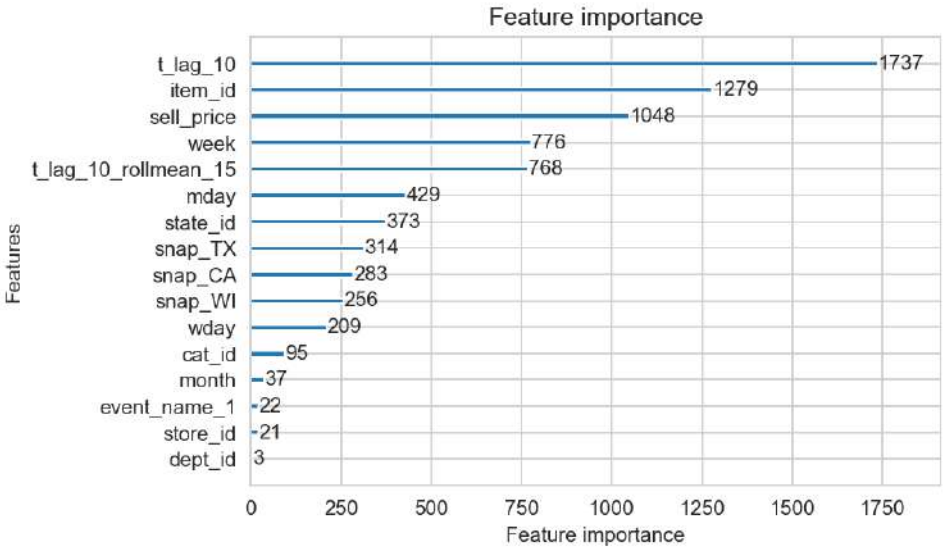

Рисунок А.3 – Значимість ознак в алгоритмі LightGBM для моделі з лаговими ознаками

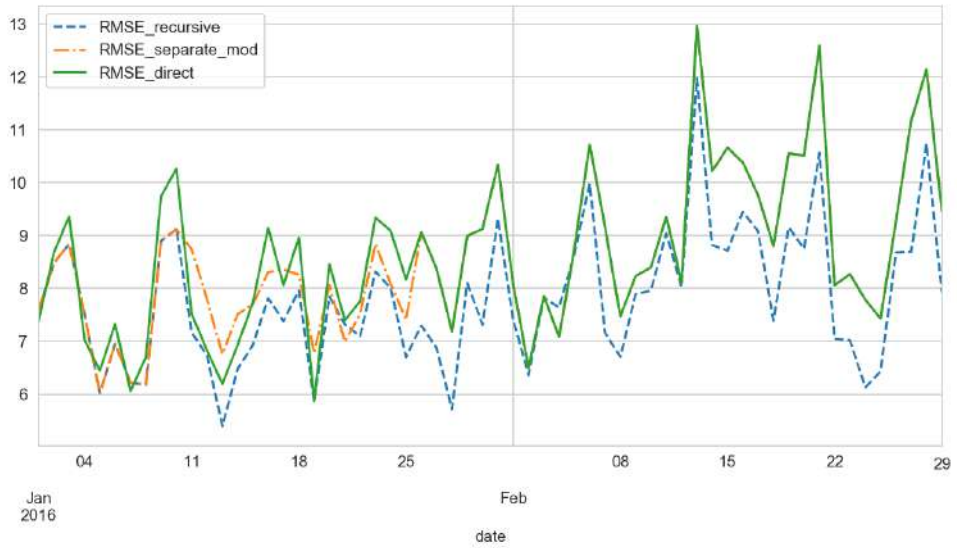

Рисунок А.4 – Динаміка оцінки RMSE для різних підходів прогнозування

– 25-денний лаг цільової змінної та 15-денне ковзне середнє для цієї лагової ознаки. Прогнозування здійснювалось для двохмісячного часового періоду. Поряд з прямими та рекурсивними підходами також було використано пряме прогнозування із різнимим моделями та різними ознаками для різних періодів прогнозування. Для перших 10 днів ми використовували функції з 10 та 25-денними лагами, на період від 11 до 25 днів, включно, використовувались ознаки із 25-денними лагами та для наступних днів використано пряме прогнозування без лагових ознак. На рис. А.4 показано динаміку оцінки RMSE для різних підходів прогнозування. Результати показують, що динаміка RMSE для рекурсивного підходу та підходу на основі окремих моделей для різних часових періодів прогнозування є подібна на періоді прогнозування протягом перших 10 днів, оскільки в цьому випадку використовуються ті ж самі ознаки. Також динаміка RMSE є подібною для підходу на основі окремих моделей та підходу на основі прямого прогнозування без лагових ознак на періоді прогнозування від 25 днів завдяки однаковим наборам ознак на цьому періоді часу для обох підходів. Загальна оцінка похибки RMSE за весь часовий період прогнозування для різних підходів така: RMSE(direct)=8.80, RMSE(recursive)=7.92, RMSE(separate models)=8.73. На рис. А.5 показано часову динаміку оцінки похибки RMSE для різних магазинів. Аналіз часової динаміки похибки RMSE дозволяє виявити моменти часу, коли похибка є дуже високою. Для цих точок знайдені алгоритмом машинного навчання узагальнені патерни не є характерними. Це важливо для аналізу можливих прихованих зовнішніх факторів, які викликають високу похибку прогнозування.

Розглянемо стекінг агрегованих результатів прогнозування різних моделей. Агреговані дані отримані сумування продажів для кожної дати аналізованого часового проміжку. Стекінгова модель тренувалась на агрегованих даних валідаційного періоду. Стекінг агрегованих результатів здійснювався за допомогою лінійної регресії. На рис. А.6 показано результати стекінгу агрегованих результатів прогнозування різних

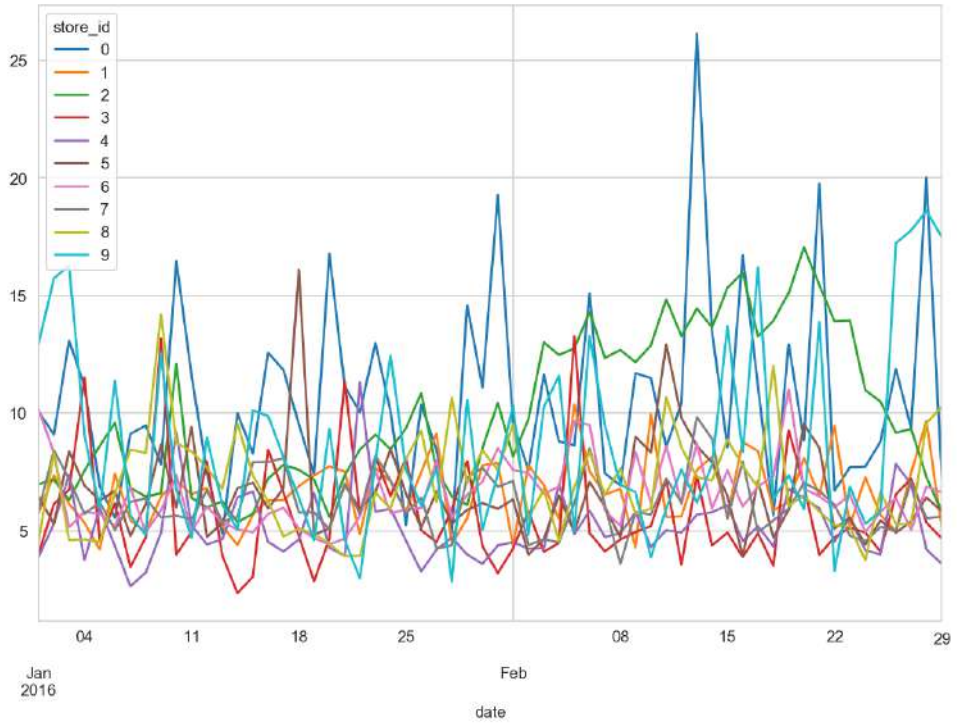

Рисунок А.5 – Часова динаміка похибки RMSE для різних магазинів

моделей. Вертикальною пунктирною лінією, розділено валідаційний та тестовий період прогнозування. Для рекурсивного підходу отримано найменший показник похибки на тестовому періоді – RMSE=463.2. Для стекінгового часового ряду на тестовому періоді похибка RMSE=429.4. Як випливає із отриманих результатів, стекінг агрегованих результатів прогнозування покращує точність прогнозування. Отже, стохастичні патерни, спричинені факторами, які не входять в набір ознак прогнозних моделей, можна врахувати за допомогою лагових ознак, які базуються на цільовій змінній. У таких випадках можна використовувати прямий підхід, рекурсивний підхід та підхід на основі окремих моделей із різними лаговими змінними для різних часових періодів. Оптимізовані результати прогнозування можна отримати, комбінуючи різні моделі в стекінговий ансамбль моделей.

А.2 Взаємний вплив наявності товарів в аналітиці продажів

Якщо різні товари знаходяться на полиці магазину одночасно, це може вплинути на продаж певного товару внаслідок, наприклад, ефекту канібалізації. Цей ефект полягає в часткові заміні продажу одного товару іншим товаром із аналогічними споживчими властивостями. Для проведення дослідження такого випадку ми взяли набір даних із змагання по прогнозуванні на платформі Kaggle 'M5 Forecasting - Accuracy' [355]. Ці дані представляють продажі товарів у магазинах Walmart. Для аналізу взято часові ряди для продажів товарів одного і того ж довільно вибраного магазину. На рис. А.7 наведено часовий ряд продажів у логарифмічному масштабі для довільно вибраного товару. Для

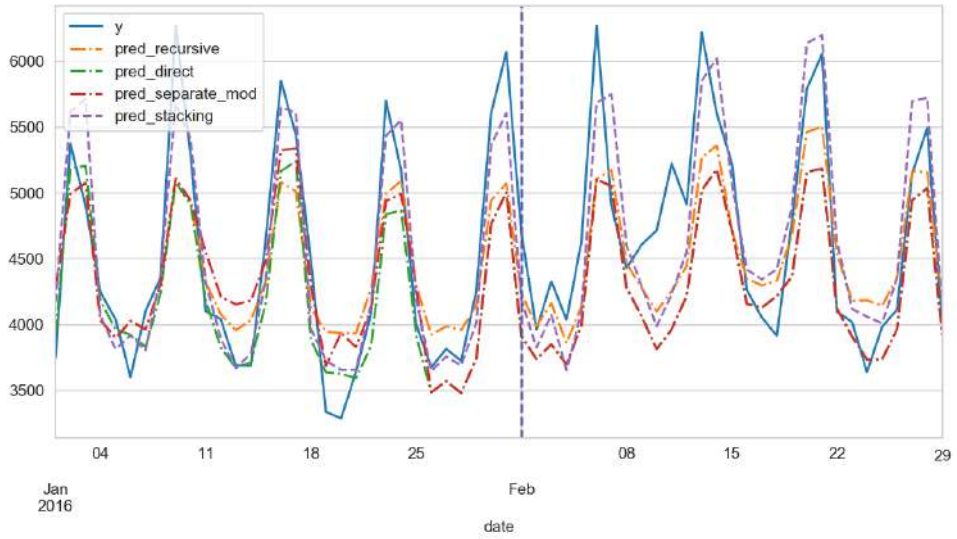

Рисунок А.6 – Стекінг агрегованих результатів прогнозування різних моделей

прогнозування використано регресію LASSO. На першому кроці аналізу як ознаки для зразків даних часових рядів використано ознаки сезонності на основі дати, такі як день тижня, день місяця, місяць, кількість днів від початку спостереження. На рис. А.8 показано результати прогнозування на цьому кроці. Вертикальні лінії розділяють послідовні частові періоди навчання, валідації та тестування. На валідаційному наборі даних проведено лінійну регресію для корекції результатів. Для цього кроку отримано RMSE=0,28. На наступному кроці ми додали бінарні ознаки, які позначають наявність розглянутих товарів на полиці. Для аналізу було взято 3048 товарів того ж магазину. На рис. А.9 наведено результати прогнозування для цього випадку, оцінка похибки – RMSE=0,23. Як випливає із отриманих результатів, використання ознак наявності товарів покращує точність прогнозування. На рис. А.10 наведено найбільш від'ємні коефіцієнти регесії для аналізованих товарів. Товари які відповідають цим коефіцієнтам, зменшують продажі певного товару у тому випадку, коли вони продаються в одному магазині із цим товаром. На рис. А.11 наведено найбільш додатні коефіцієнти регесії для продуктів. Підхід на основі машинного навчання дозволяє знайти складніші патерни в даних порівняно з лінійною регресією. За допомогою регресії на основі алгоритму Random Forest отримано результати прогнозування із похибкою RMSE=0.21, які наведено на рис. А.12. Отже, отримані результати показують, що фактор одночасної наявності деяких товарів на полиці може вплинути на продаж заданого товару. Цей фактор може розглядатися як додаткова прогнозна ознака в аналітиці продажів.

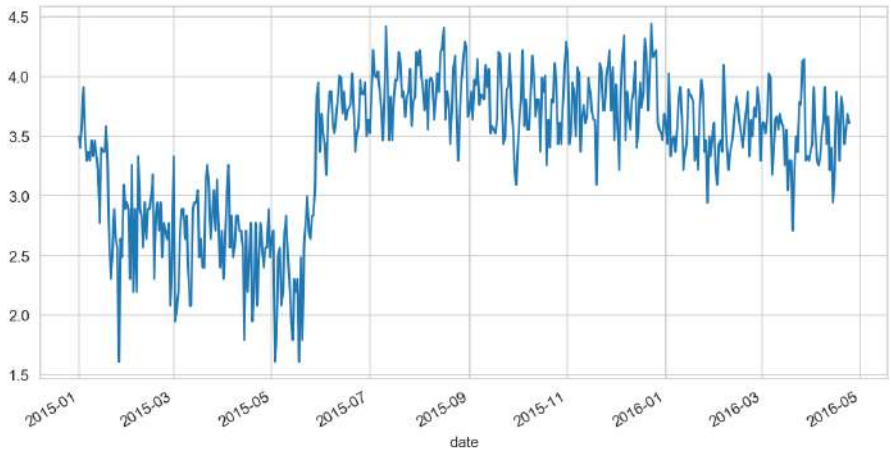

Рисунок А.7 – Часовий ряд продажів у логарифмічному масштабі для довільно вибраного товару

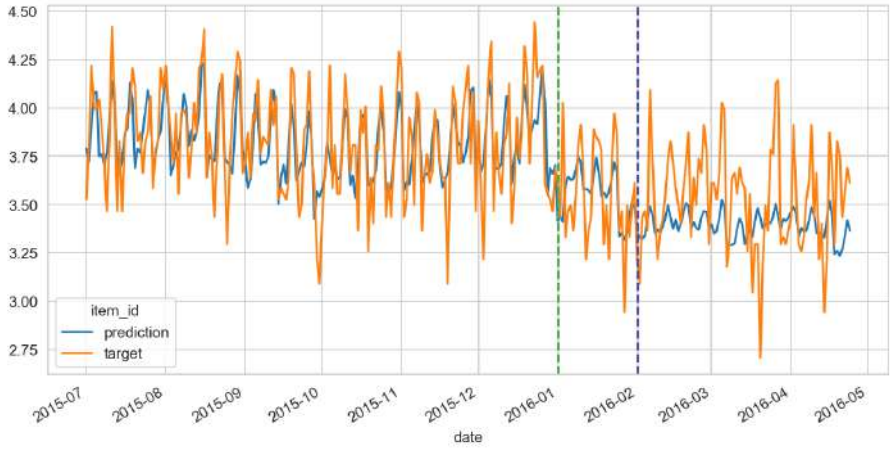

Рисунок А.8 – Результати прогнозування із використанням ознак сезонності на основі дати

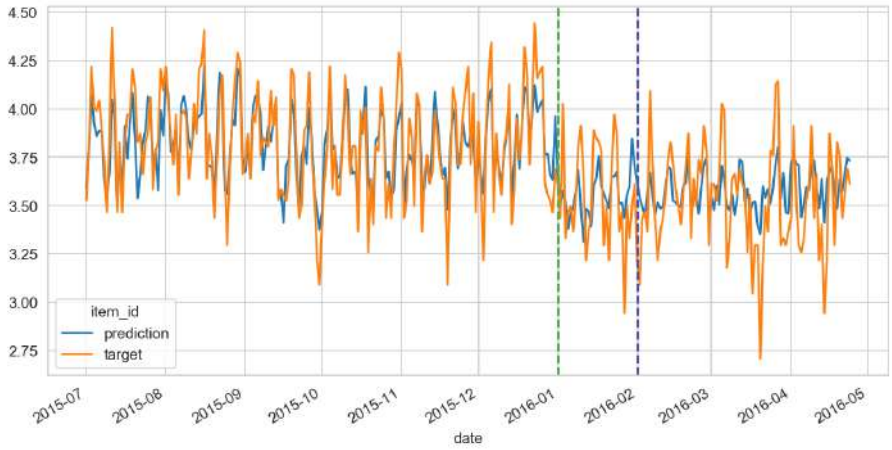

Рисунок А.9 – Результати прогнозування при використанні ознак наявності товарів

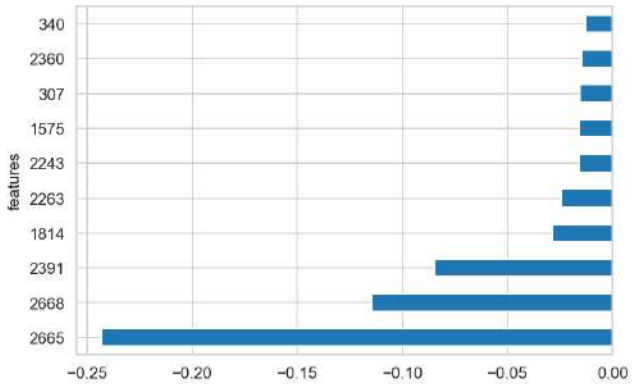

Рисунок А.10 – Від'ємні коефіцієнти регресії для товарів

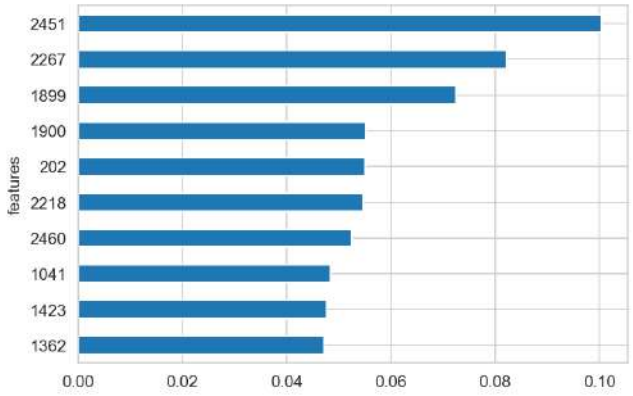

Рисунок А.11 – Додатні коефіцієнти регесії для товарів

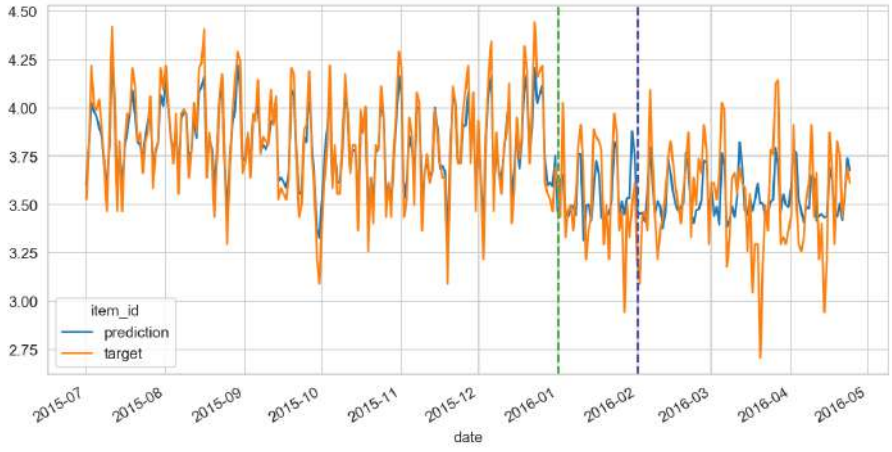

Рисунок А.12 – Результати прогнозування із використанням регресії на основі алгоритму Random Forest

А.3 Моделювання поширення COVID-19 на основі байєсівської регресії

В даний час існують різні методи, підходи та масиви даних для моделювання поширення COVID-19 [356, 357, 358, 359, 360, 361]. Для прогнозної аналітики поширення COVID-19 використано модель логістичної кривої. Для оцінки параметрів моделі використано байєсівську регресію [59, 58, 227]. Цей підхід дозволяє отримати постеріорний розподіл ймовірності для параметрів моделі. У байєсівському виведенні можна використовувати інформативні апріорні розподіли, які може встановити експерт. Отже, результат можна розглядати як компроміс між історичними даними та думкою експерта. Це важливо в тих випадках, коли ми маємо невелику кількість історичних даних. Імовірнісний підхід дозволяє отримати функцію щільності розподілу ймовірностей для цільової змінної. Модель логістичної кривої на основі байєсівської регресії можна записати так:

$$\begin{aligned} & n \sim \mathcal{N}(\mu, \sigma), \\ & \mu = \frac{\alpha}{1 + exp(-\beta(t - t_0))} 10^5, \\ & t = t_{weeks}(Date - Date_0), \end{aligned} \tag{А.1}$$

де $Date_0$ - день початку спостережень у масиві історичних даних, який вимірюється тижнями. Дані для аналізу було взято з [357]. Параметр α визначає максимальні значення поширення коронавірусу, параметр β - емпіричний коефіцієнт, який визначає швидкість поширення коронавірусу. Для аналізу байєсівських моделей використовуються числові методи Монте-Карло [59, 58, 227]. Баєсівське виведення дозволяє отримати функції щільності розподілу ймовірностей для параметрів моделі та оцінити невизначеність, яка є важливою в аналітиці оцінки ризику. У байесівській регресії можна враховувати думки експертів шляхом задання інформативного апріорного розподілу для параметрів моделі. Для чисельних розрахунків байєсівської регресії використано пакет *pystan* для платформи статистичного моделювання *Stan* [227]. На малюнку А.13 наведено коробкові графіки для параметрів β в моделі поширення коронавірусу для різних країн. На рис. А.14, А.15, А.16, А.17, А.18, А.19, А.20 показано результати прогнозування випадків поширення коронавірусу. Розрахунки зроблені із використанням доступних на момент прогнозування історичних даних. У практичній аналітиці важливо знайти максимум випадків виявлення випадків COVID-19 за добу, який відображає половину часового періоду активного поширення коронавірусу в досліджуваному регіоні. Розглянута феноменологічна модель не враховує механізмів поширення коронавірусу, які постійно змінюють свою динаміку. Тому прогнозування можуть змінюватись по мірі надходження нових даних та внаслідок експертного формування апріорних розподілів параметрів моделі.

Отже, байесівська регресійна модель із використанням логістичної кривої може

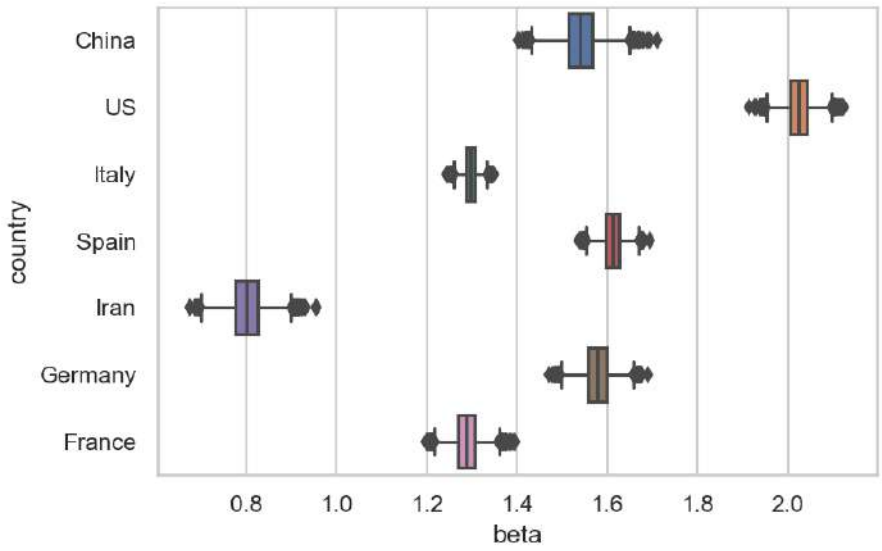

Рисунок А.13 – Коробкові графіки розподілів ймовірності для коефіцієнтів β для різних країн

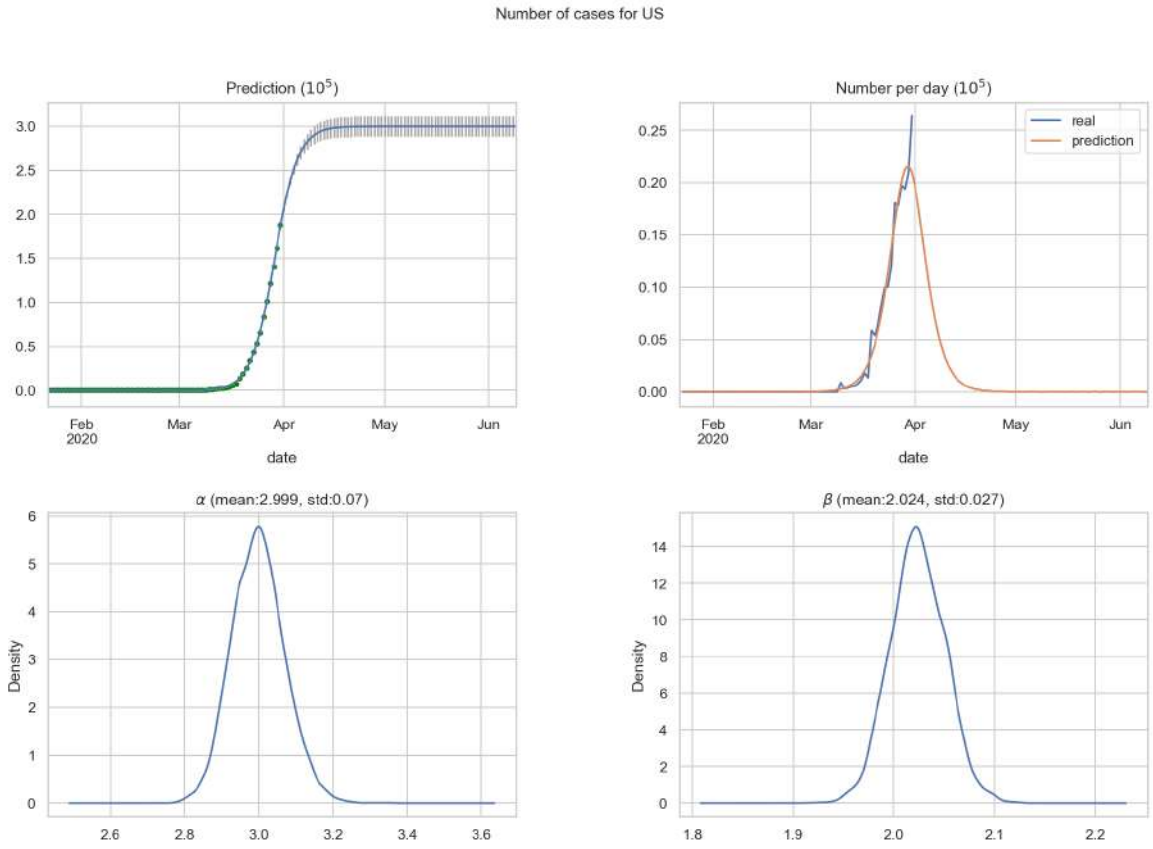

Рисунок А.14 – Моделювання поширення COVID-19 для США

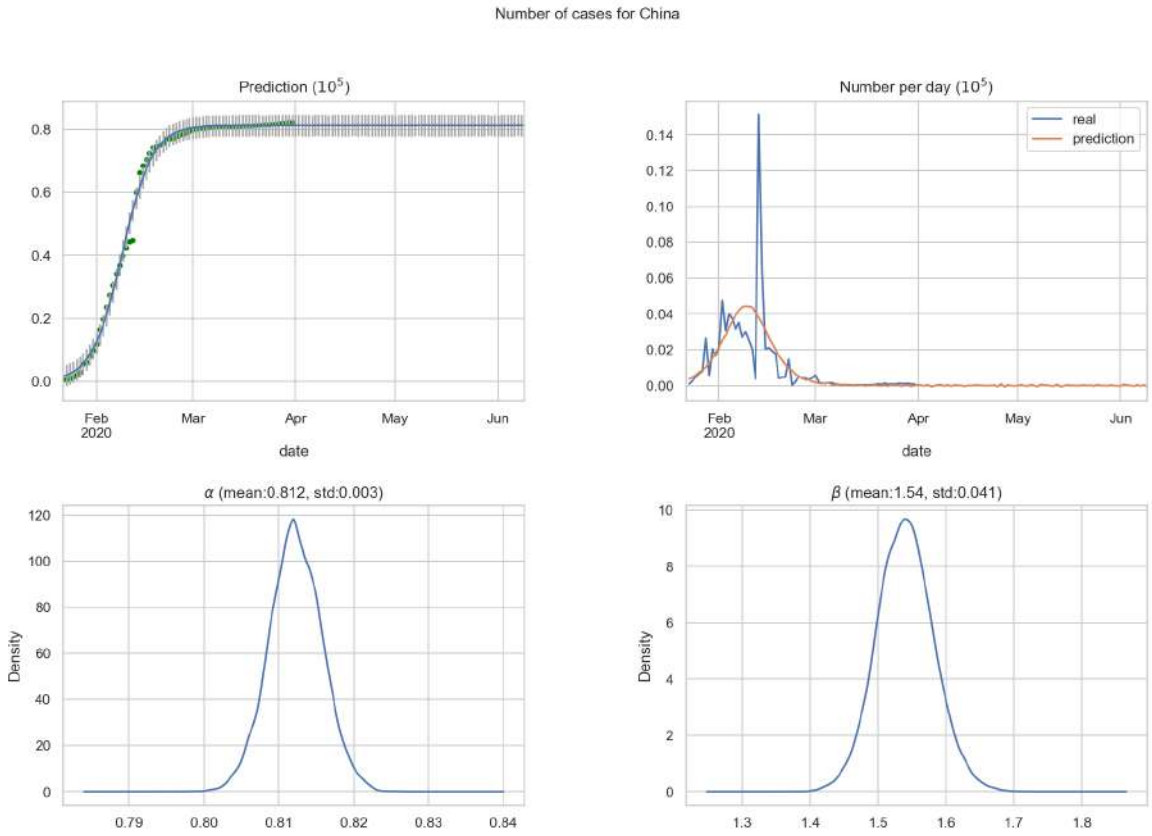

Рисунок А.15 – Моделювання поширення COVID-19 для Китаю

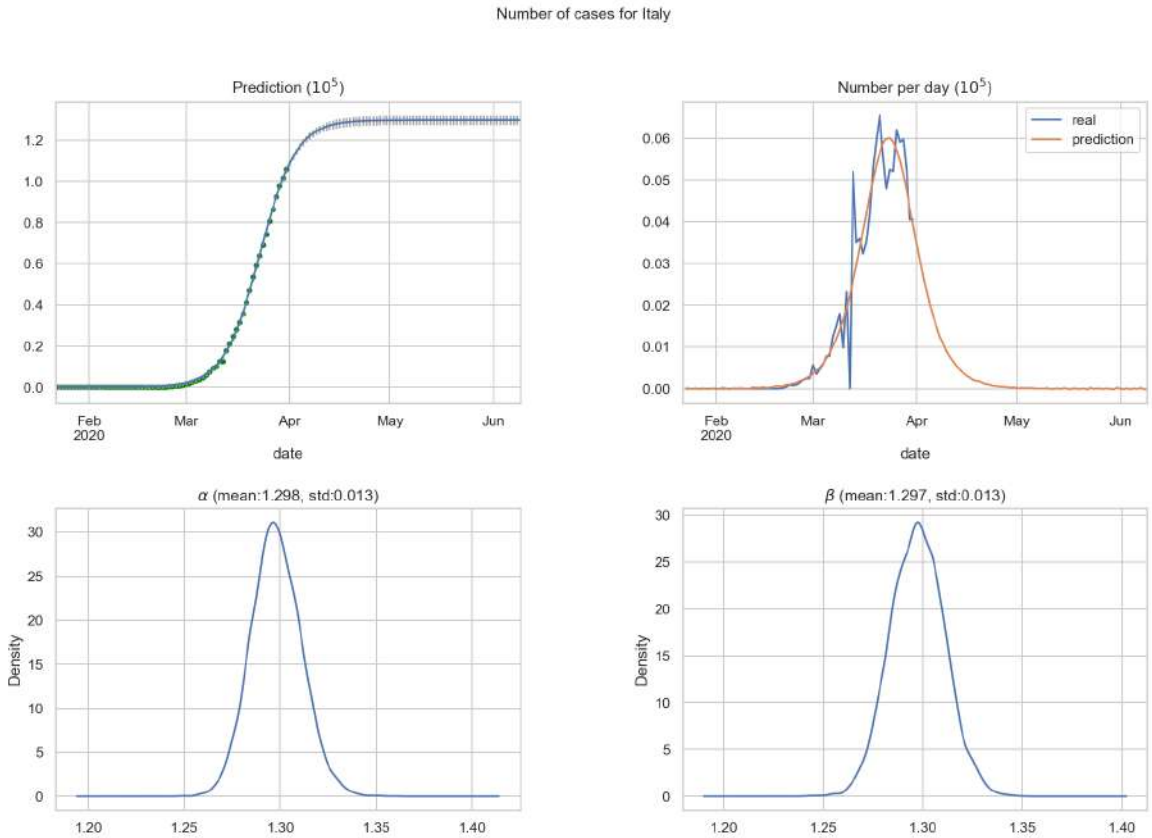

Рисунок А.16 – Моделювання поширення COVID-19 для Італії

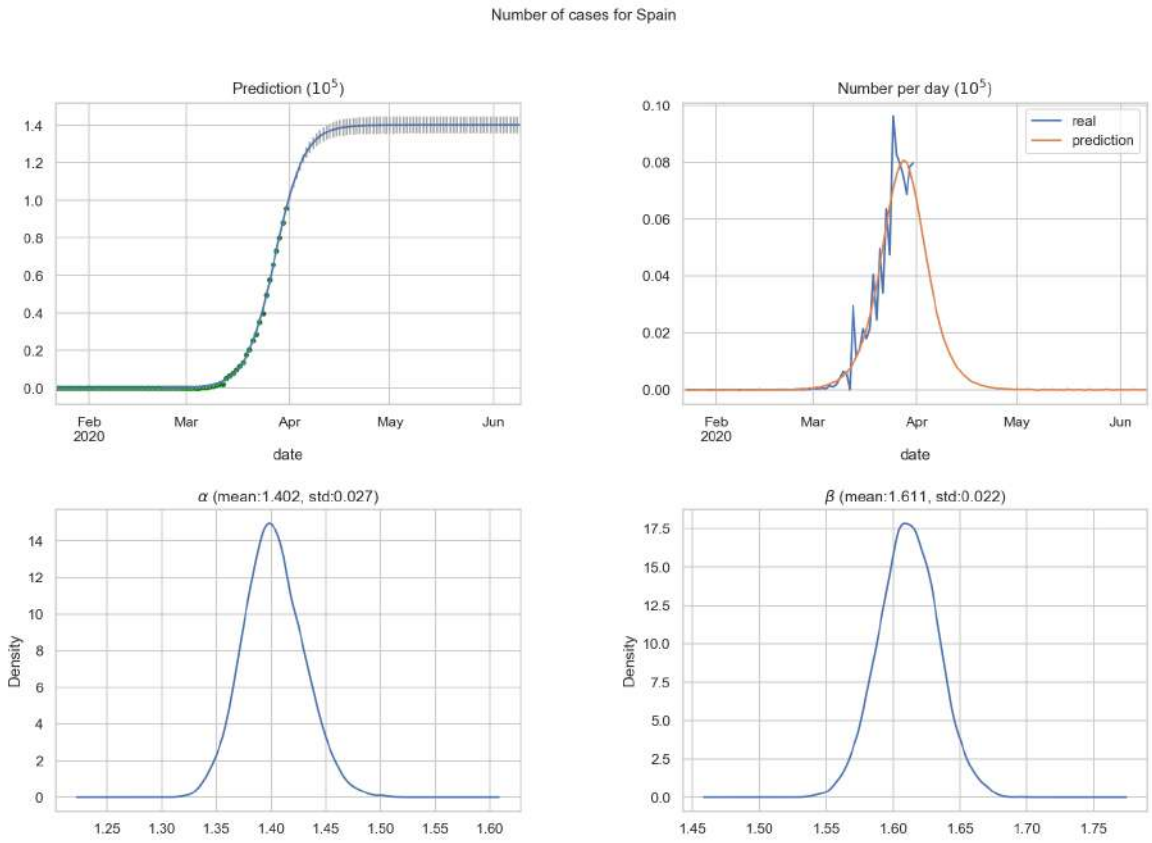

Рисунок А.17 – Моделювання поширення COVID-19 для Іспанії

бути використана для прогнозної аналітики поширення коронавірусу. Така модель може бути ефективною, коли є експоненційне зростання кількості підтверджених випадків COVID-19 [246].

А.4 Порівняльний аналіз впливу економічних криз на фінансовий ринок

Пандемія COVID-19 має великий вплив на фінансовий ринок. Аналіз такого впливу є важливим, зокрема при формуванні стабільних інвестиційних портфелів. Розглянемо вплив кризи, зумовленої пандемією COVID-19 на акції компаній на фінансовому ринку та порівняємо цей вплив із впливом кризи 2008 року та спадом ринку 2018 року. Для цього використаємо лінійну класичну та байєсівську регресії. Використання байєсівської регресії дозволяє отримати функції щільності розподілу ймовірностей для коефіцієнтів досліджуваних факторів та оцінити невизначеність. Ми взяли такі періоди часу для кожної з криз - crisis_2008: [2008-01-01,2009-01-31], down_turn_2018: [2018-10-01,2019-01-03], coronavirus: [2020-02-18,2020-03-25]. Для кожного з цих криз ми створили змінну регресії, яка дорівнює 1 у період кризи та 0 в інших випадках. На рис. А.21 показано часовий ряд для фінансового індексу S&P500. В якості цільової змінної розглянуто щоденну зміну ціни акції. Знаючи щоденну зміну цін, зміни в кризові періоди, можна

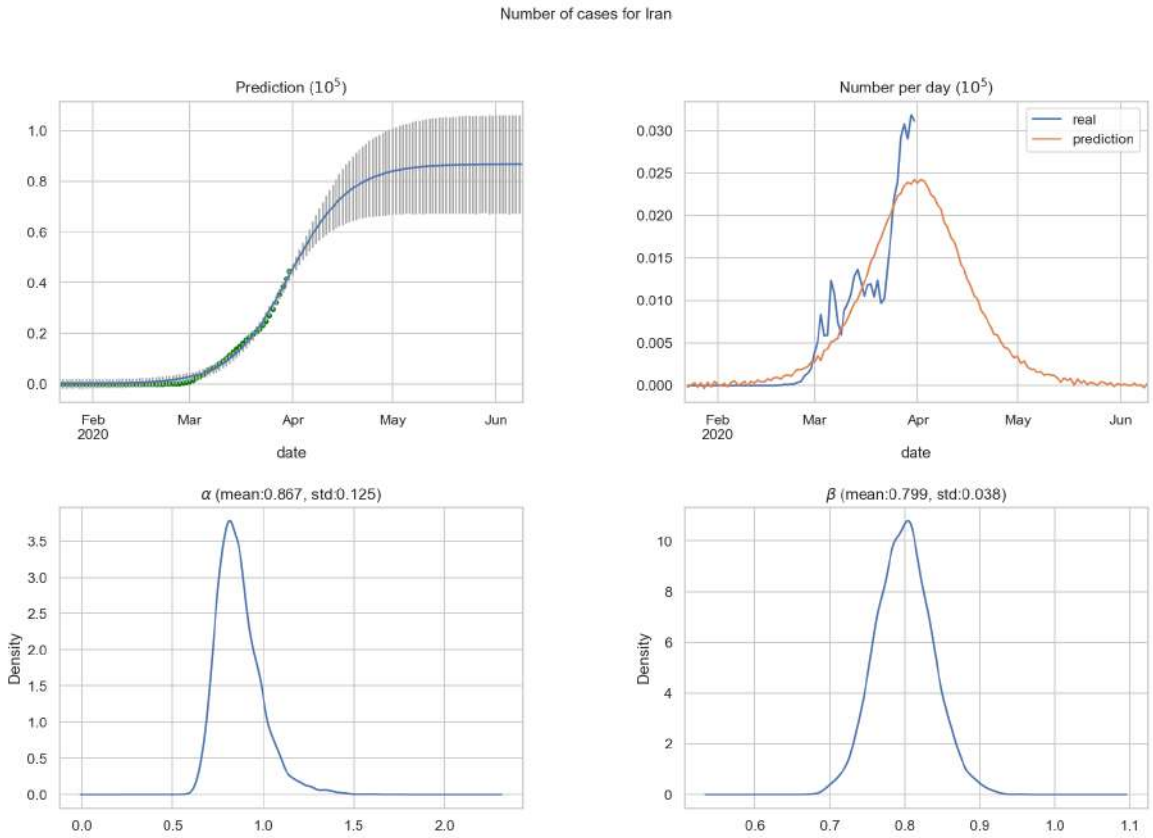

Рисунок А.18 – Моделювання поширення COVID-19 для Ірану

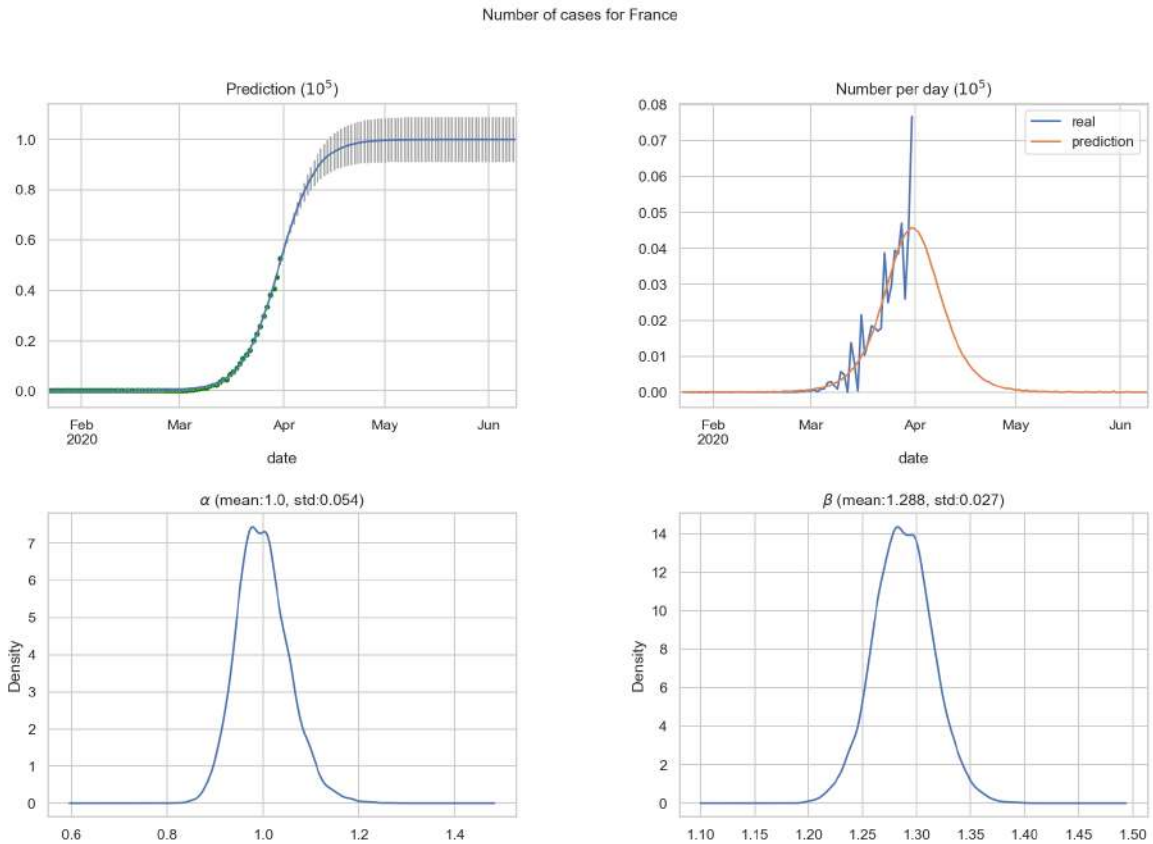

Рисунок А.19 – Моделювання поширення COVID-19 для Франції

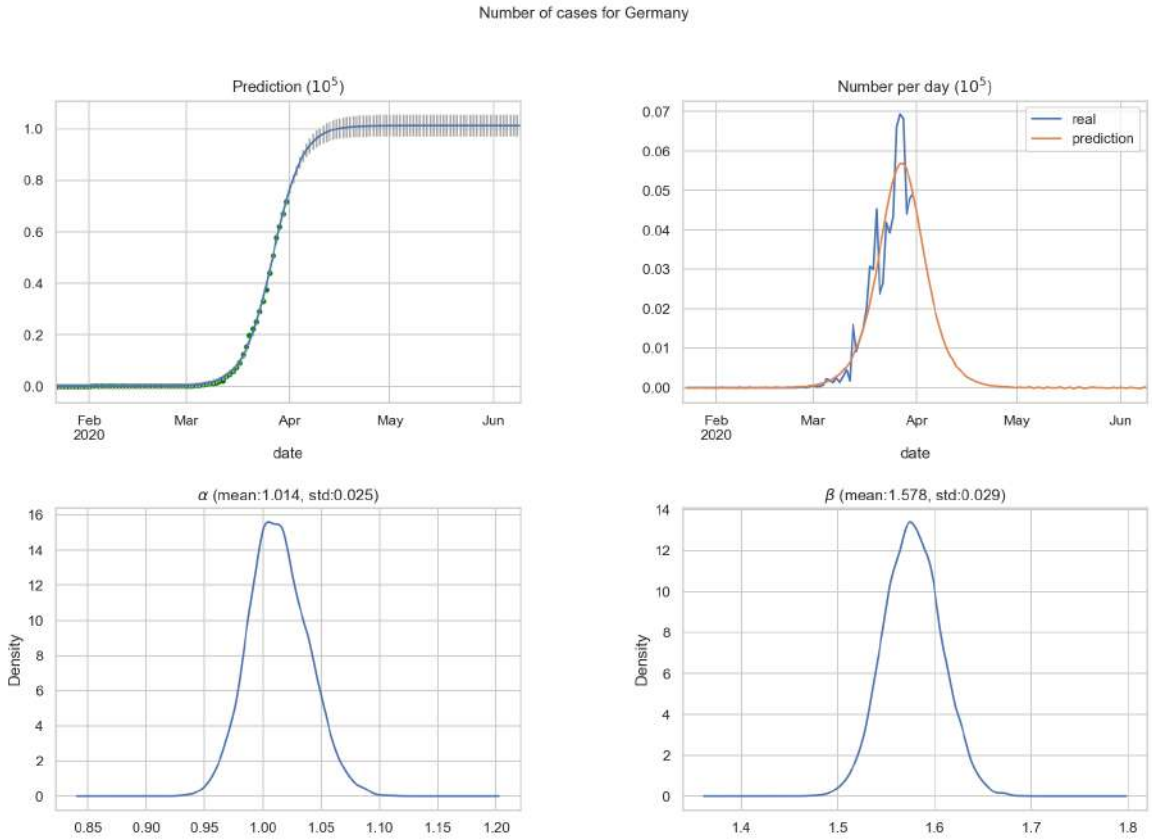

Рисунок А.20 – Моделювання поширення COVID-19 для Німеччини

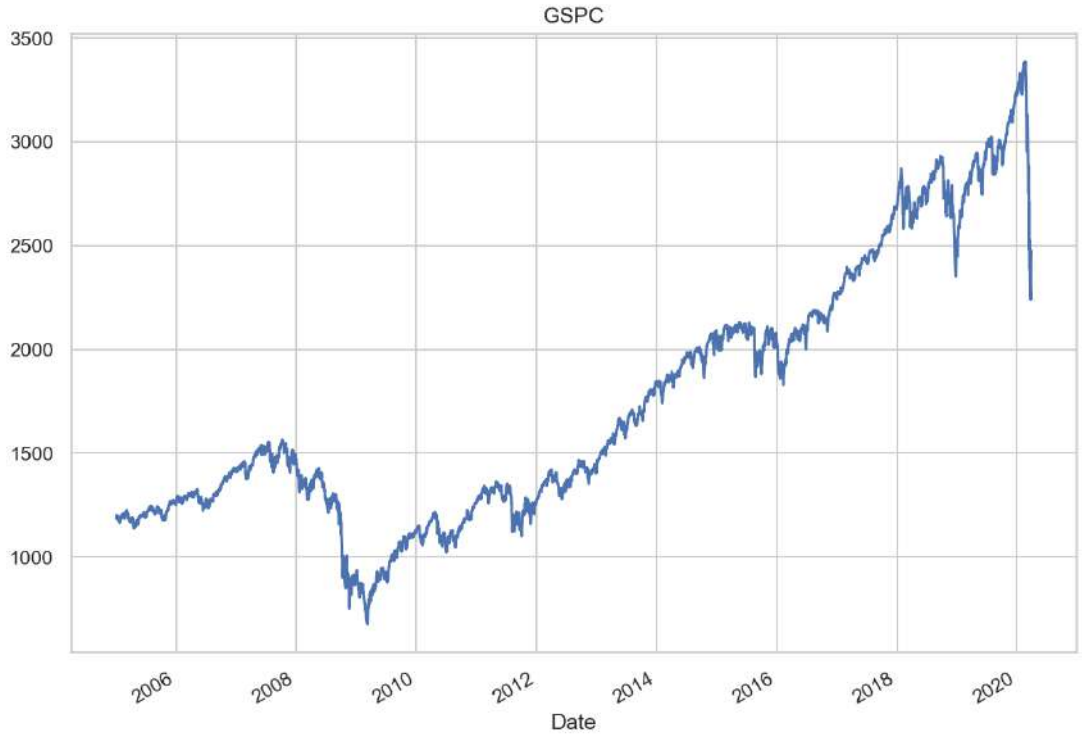

Рисунок А.21 – Часовий ряд для індексу S&P500

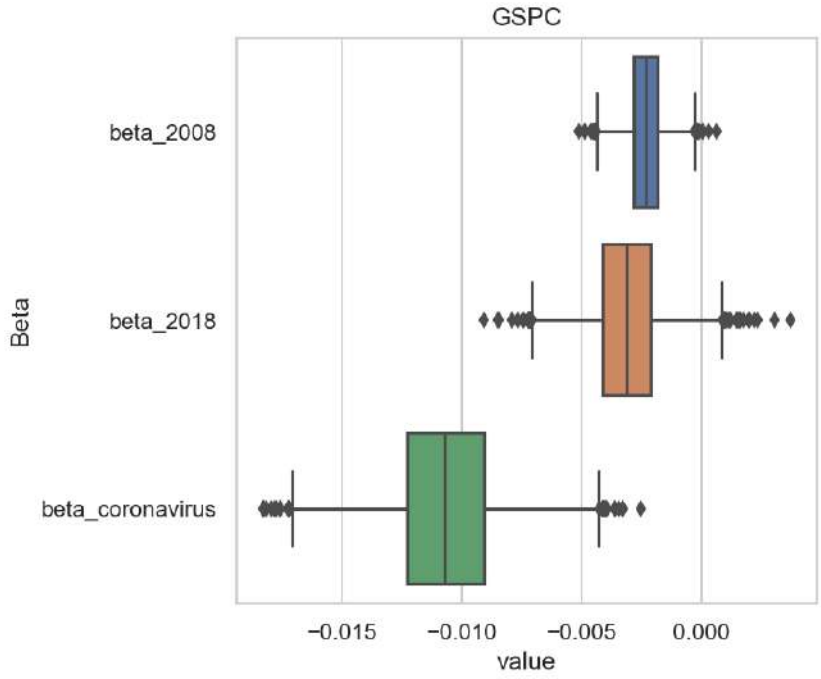

Рисунок А.22 – Коробкові графіки вагових коефіцієнтів вливу криз на індекс S&P500

оцінити здатність інвесторів розуміти тенденції та формувати інвестиційні портфелі. Для розрахунків байєсівської регресії було використано пакет *pystan* для платформи статистичного моделювання *Stan* [227]. На рис. А.22 показано коробкові графіки вагових коефіцієнтів кожної кризи для індексу S&P500. На рис. А.23 показано коробкові графіки розподілів вагових коефіцієнтів впливу для різних акцій. Для досліджень було взято випадковий набір акцій компаній зі списку S&P500. На рис. А.24 показано акції з найбільш негативними значеннями регресійних коефіцієнтів впливу COVID-19 кризи на денну зміну ціни. На рис. А.25 показано акції з позитивними значеннями регресійних коефіцієнтів впливу COVID-19 кризи на денну зміну ціни. На рис. А.26 показано вагові коефіцієнти впливу криз на довільно обрані акції.

Підхід із використанням байєсівської регресії дозволяє аналізувати невизначеність впливу різних фінансових криз. Отримані результати показують, що різні кризи порізному впливають на динаміку цін акцій внаслідок реалізації різних механізмів впливу. Невизначеність коронавірусної кризи більша порівняно з іншими кризами. Розрахунок невизначеності дозволяє робити оцінку ризиків для інвестиційних портфелів та бізнес-процесів.

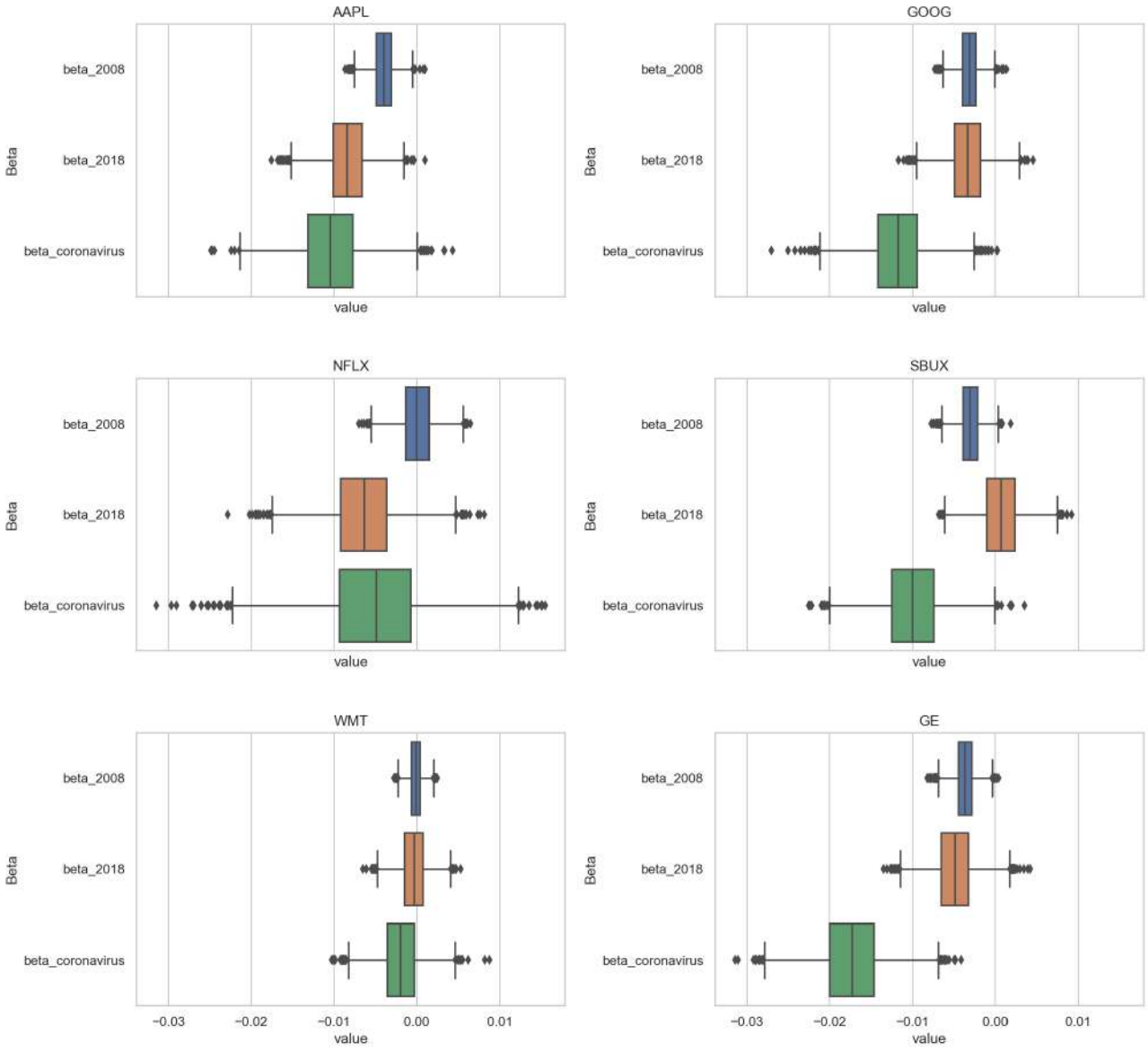

Рисунок А.23 – Коробкові графіки розподілів вагових коефіцієнтів для різних акцій

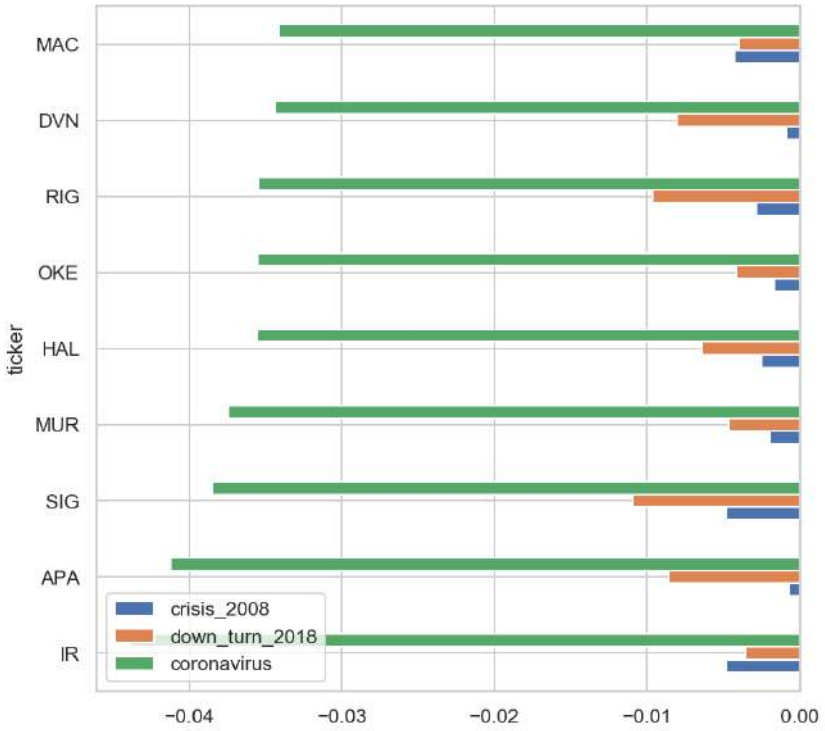

Рисунок А.24 – Акції з найбільш негативними значеннями регресійних коефіцієнтів впливу COVID-19 кризи на денну зміну ціни

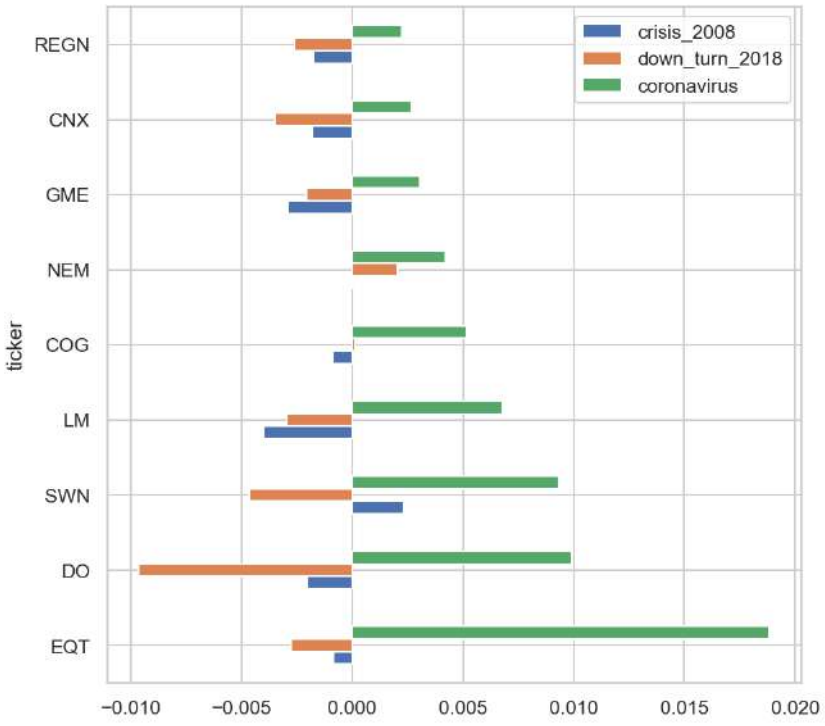

Рисунок А.25 – Акції з позитивними значеннями регресійних коефіцієнтів впливу COVID-19 кризи на денну зміну ціни

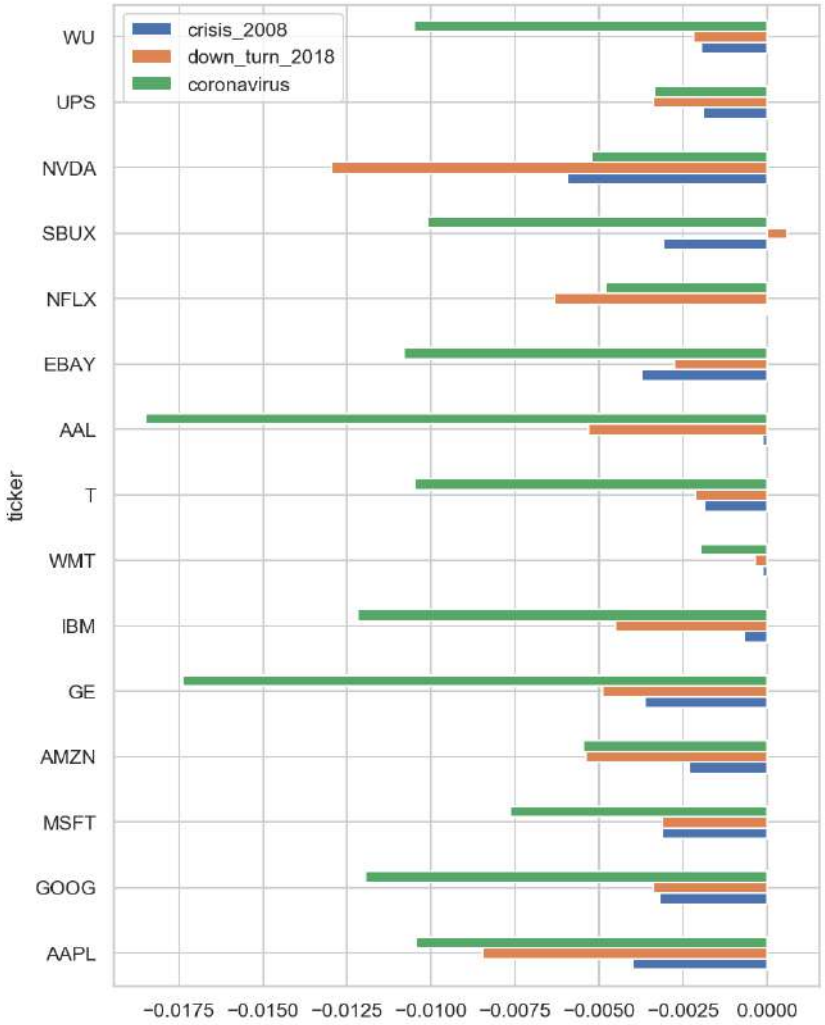

Рисунок А.26 – Вагові коефіцієнти впливу криз на довільно обрані акції

А.5 Розподіли семантичних ознак у текстових вибірках

А.5.1 Розподіли семантичних та тематичних полів у текстах повідомлень груп новин

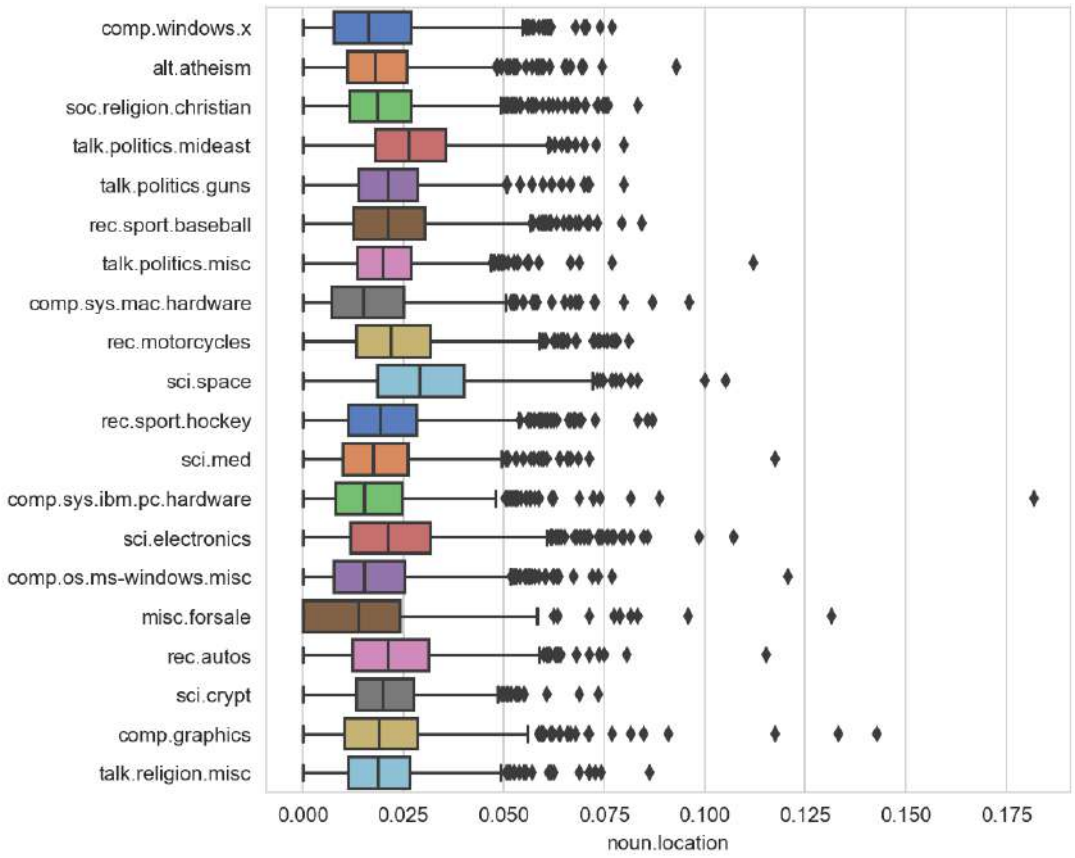

Рисунок А.27 – Розподіл частот семантичного поля *noun.location* по класах документів

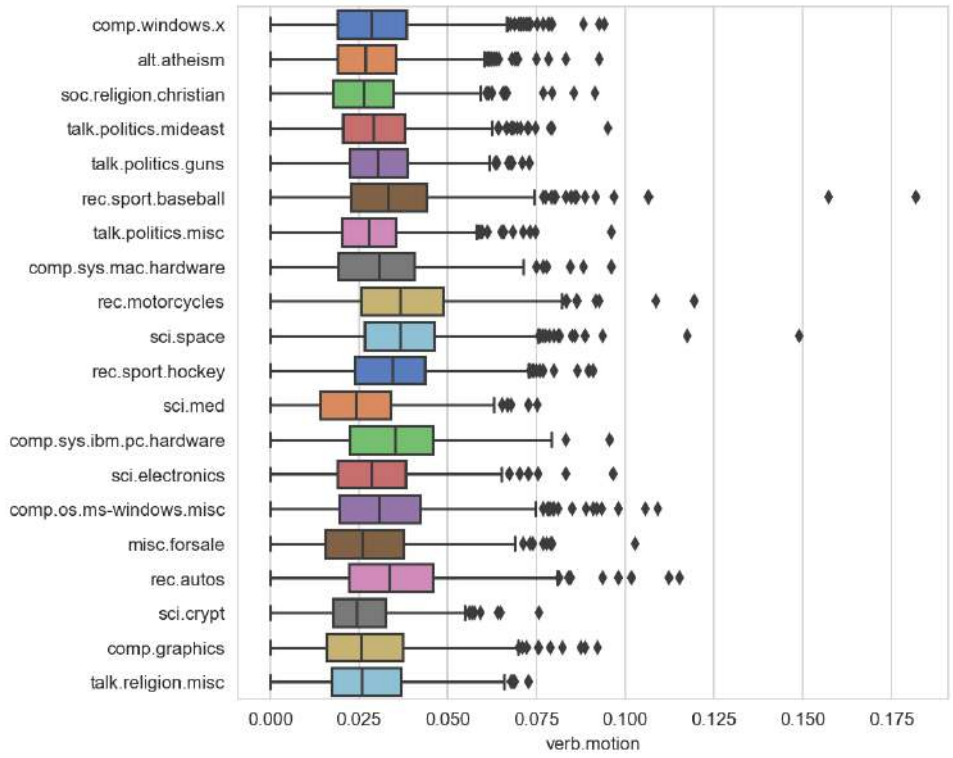

Рисунок А.28 – Розподіл частот семантичного поля *verb.motion* по класах документів

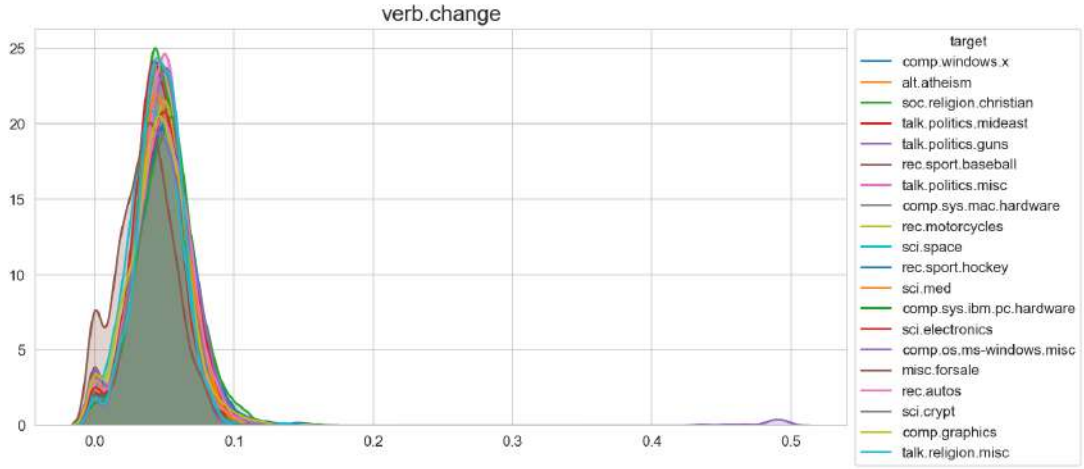

Рисунок А.29 – Щільність розподілу ймовірностей частоти семантичного поля *verb.change* для різних класів документів

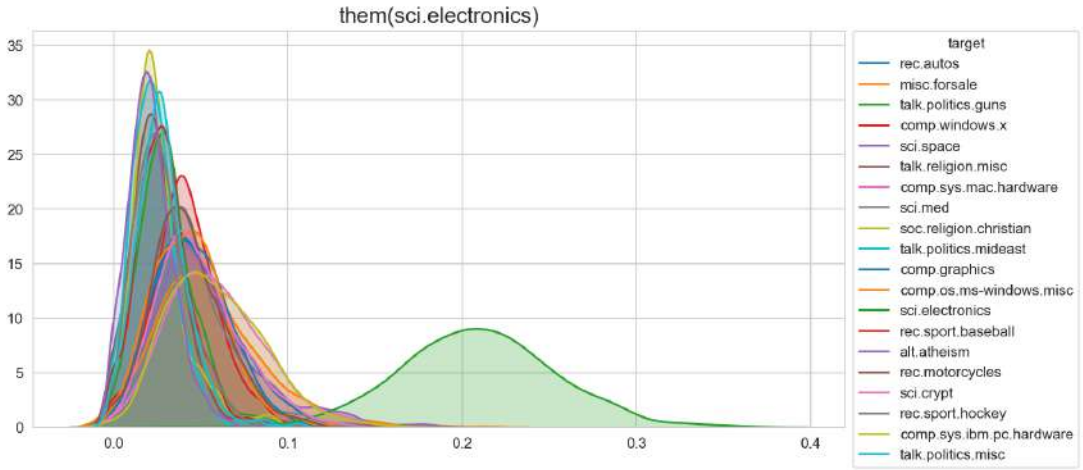

Рисунок А.30 – Щільність розподілу ймовірностей тематичного поля *them(sci.electronics)* для різних класів документів

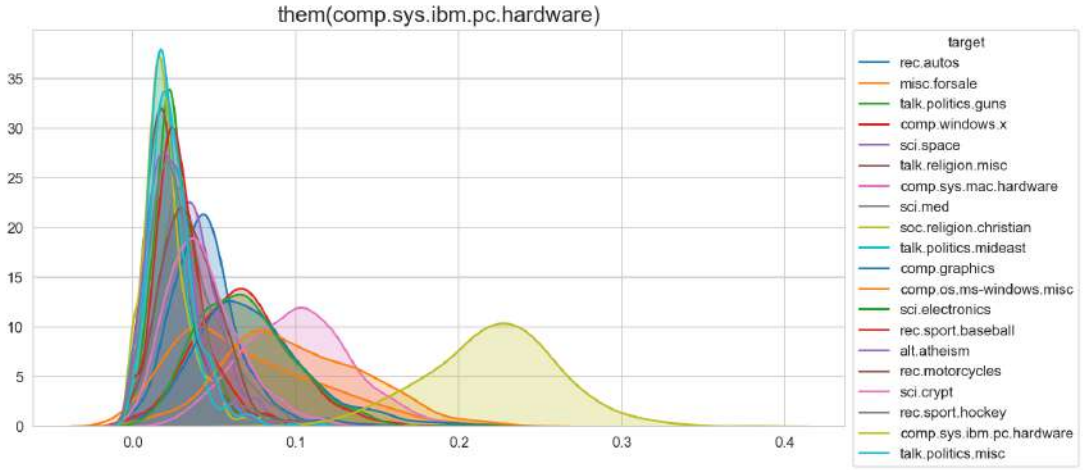

Рисунок А.31 – Щільність розподілу ймовірностей частоти тематичного поля *them(comp_sys_ibm_pc_hardware)* для різних класів документів

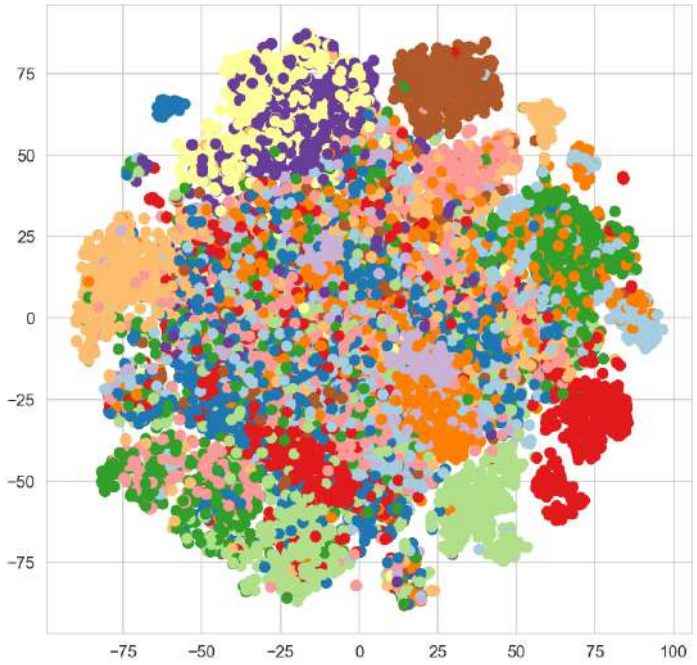

Рисунок А.32 – Розподіл тематичних полів у двохвимірному t-SNE просторі

A.5.2 Розподіли компонент сингулярного розкладу матриць TF-IDF в текстах груп новин

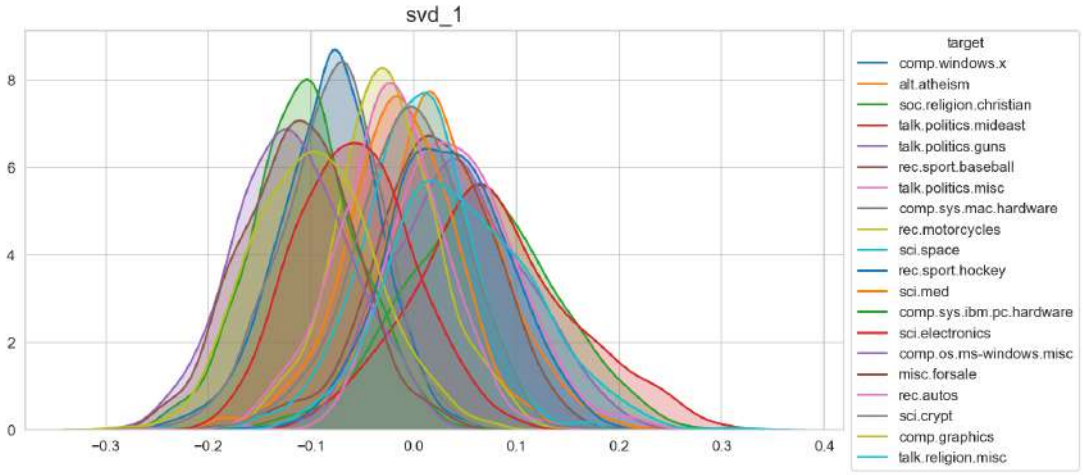

Рисунок А.33 – Щільність розподілу ймовірностей значень заданої компоненти сингулярного розкладу TF-IDF матриці для різних класів документів

A.5.3 Розподіли компонент латентного розміщення Діріхле в текстах груп новин

Для формування компонент латентного розміщення Діріхле (LDA) вибрано 1000 лексем на основі впорядкованих за спаданням частот на основі TF-IDF матриці. Було вибрано 30 LDA тематик, які були сформовані із використанням пакету *gensim* [288]. Структура утворених LDA тематик зображена на рис. А.34

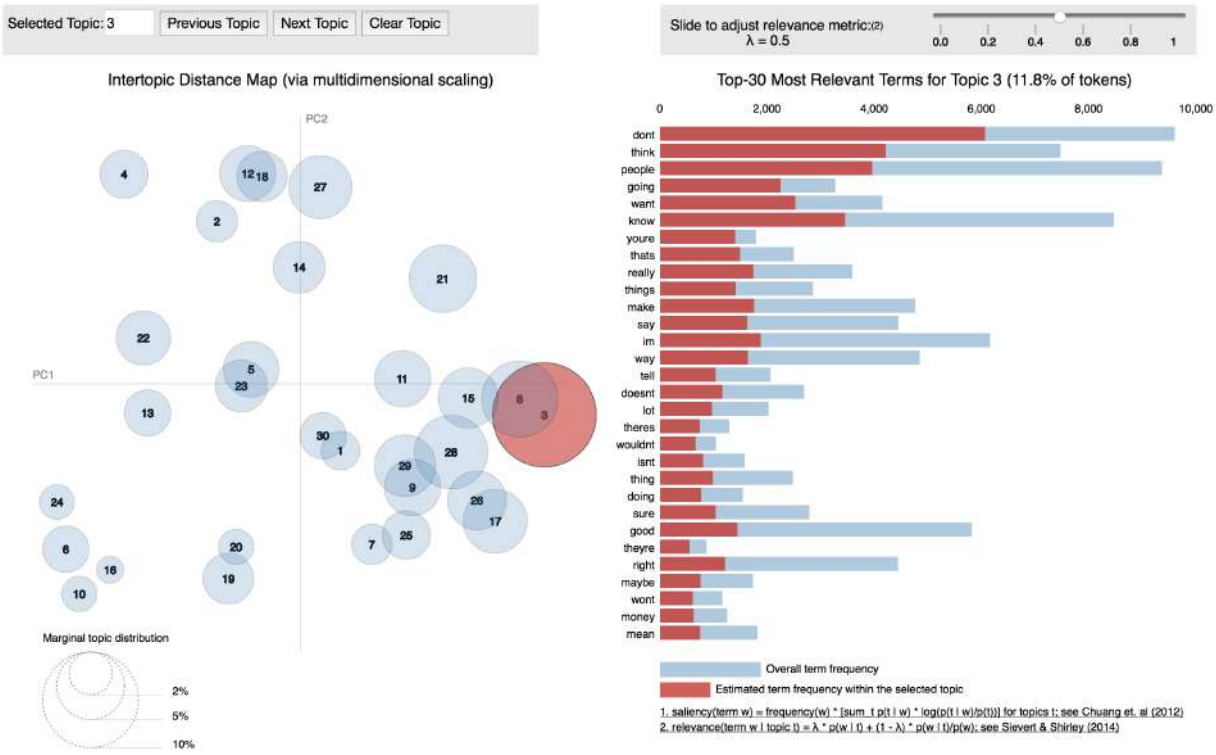

Рисунок А.34 – Структура утворених LDA тематик

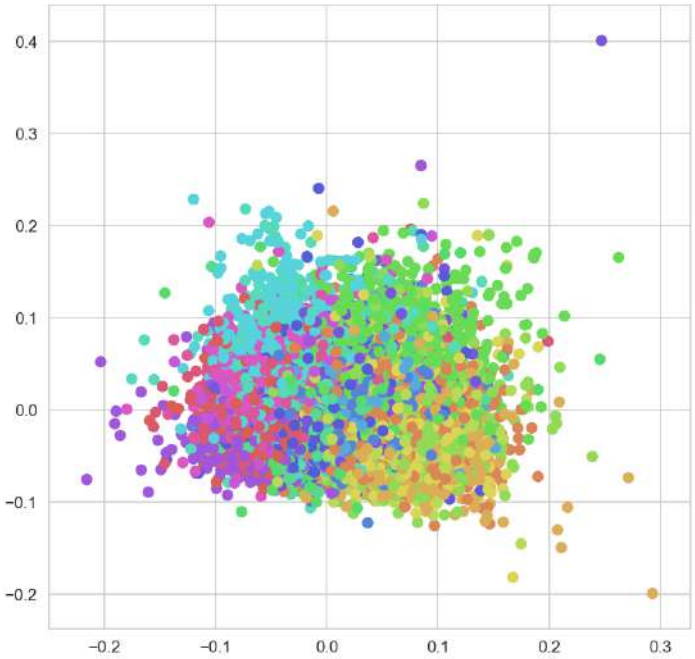

Рисунок А.35 – Розподіл LDA компонент у двохвимірному PCA просторі

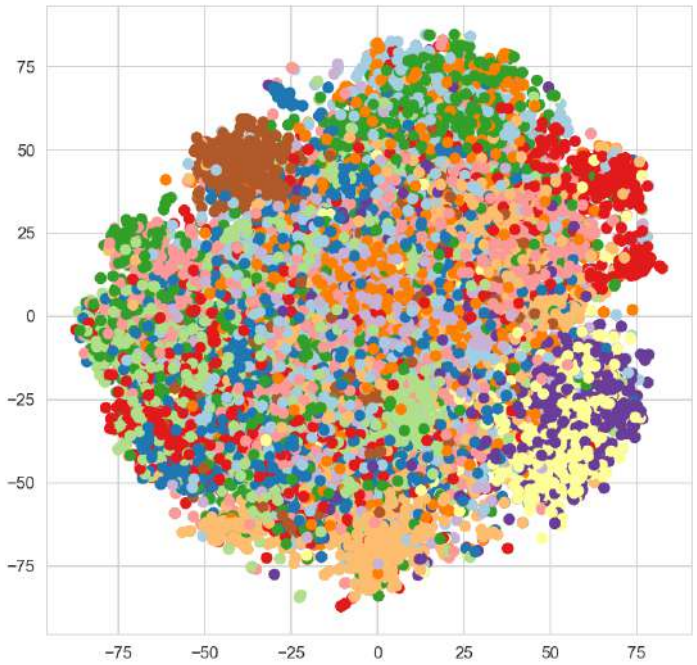

Рисунок А.36 – Розподіл LDA компонент у двохвимірному t-SNE просторі

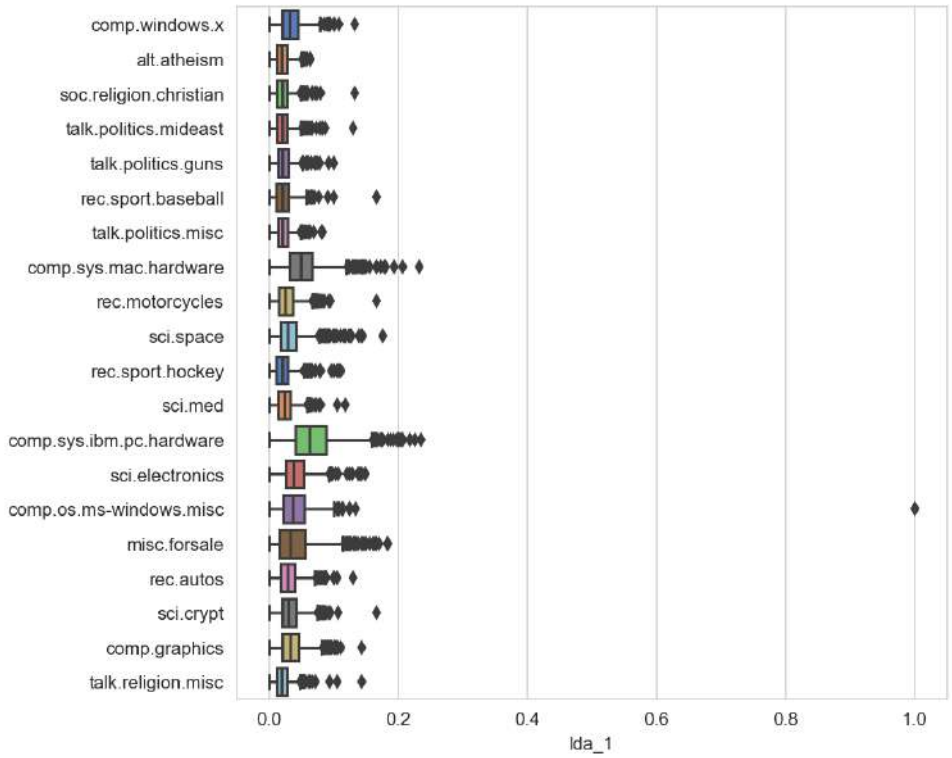

Рисунок А.37 – Приклад розподілу LDA компоненти для різних класів документів

A.6 Аналіз текстових даних за допомогою алгоритмів машинного навчання

A.6.1 Кластеризація авторських текстів за різними семантичними ознаками

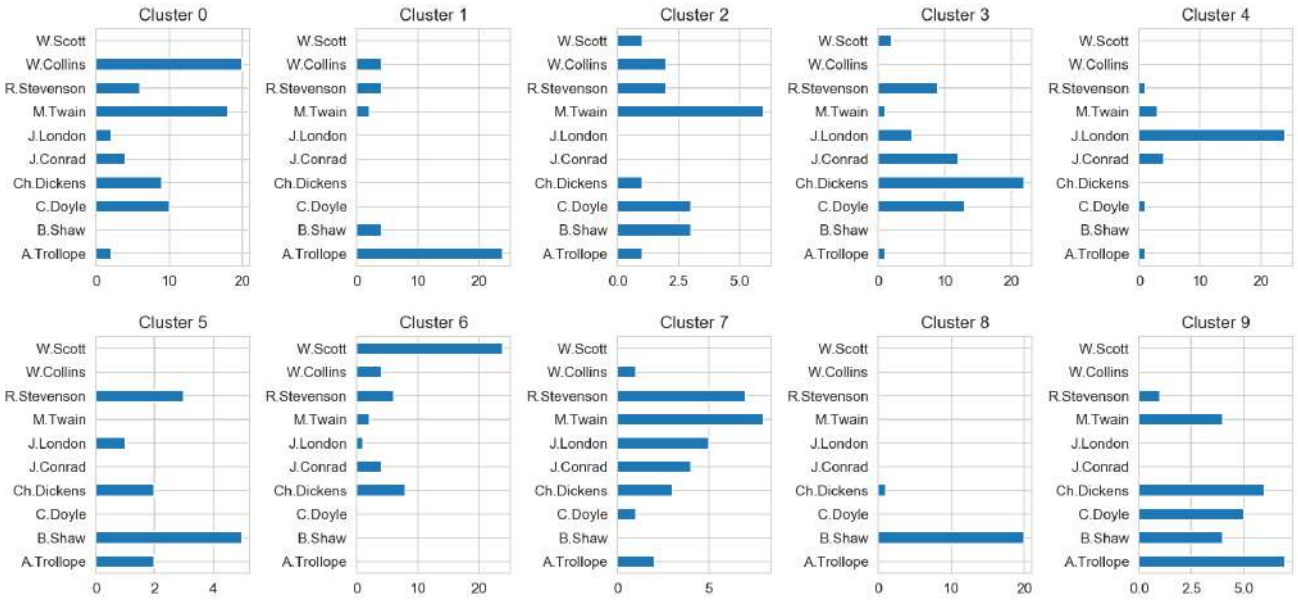

Рисунок А.38 – Розподіл документів у кластерах за авторами в алгоритмі k-means у просторі семантичних полів

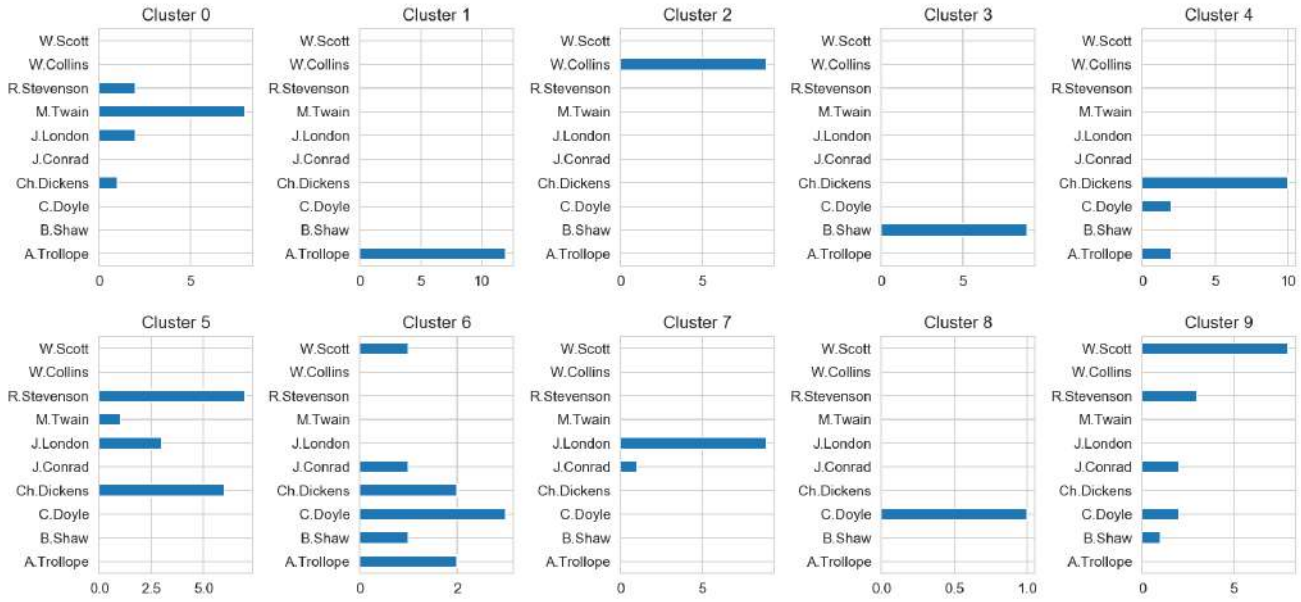

Рисунок А.39 – Розподіл документів у кластерах за авторами в алгоритмі k-means у просторі SVD компонент

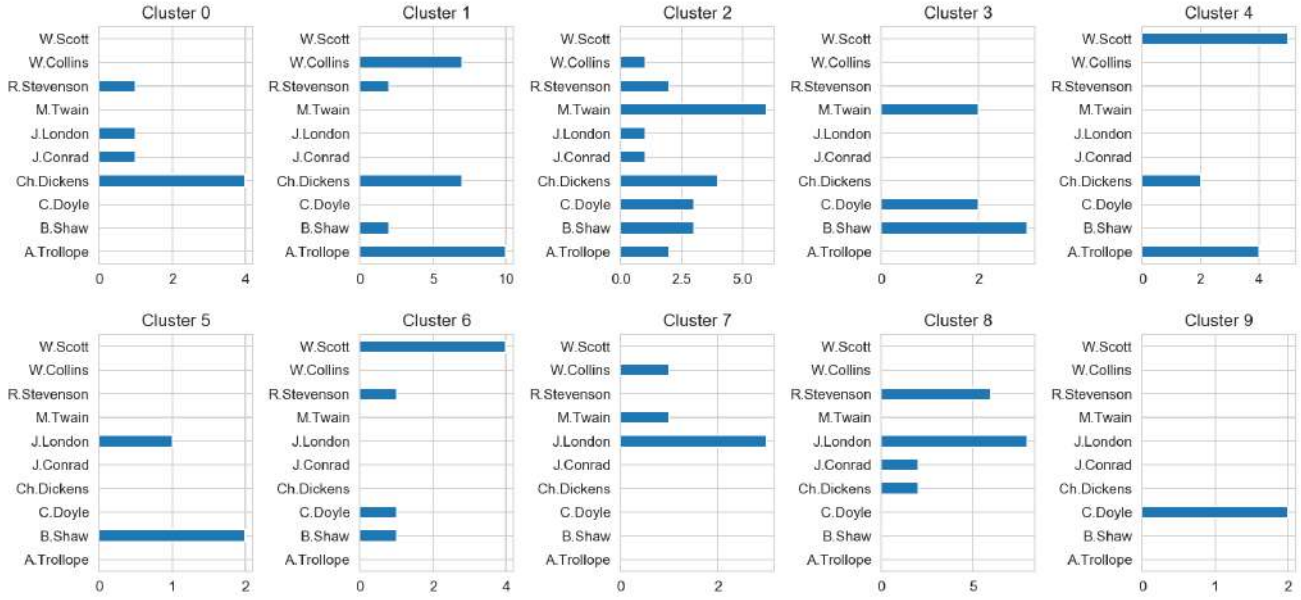

Рисунок А.40 – Розподіл документів у кластерах за авторами в алгоритмі k-means у просторі LDA компонент

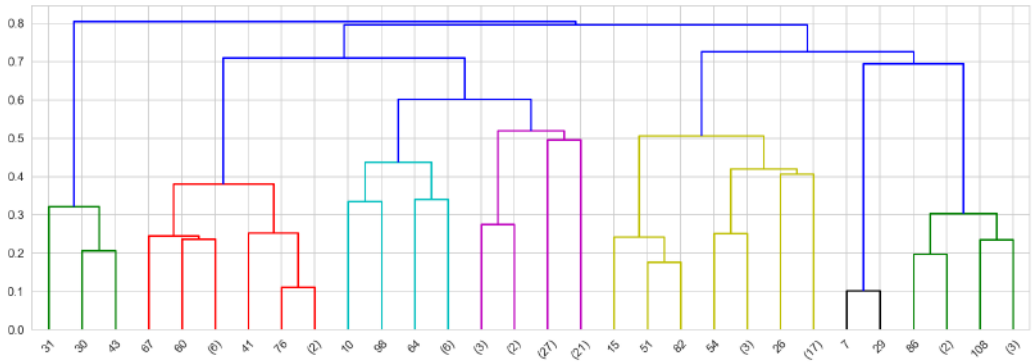

Рисунок А.41 – Дендрограма агломеративної кластеризації у просторі LDA компонент

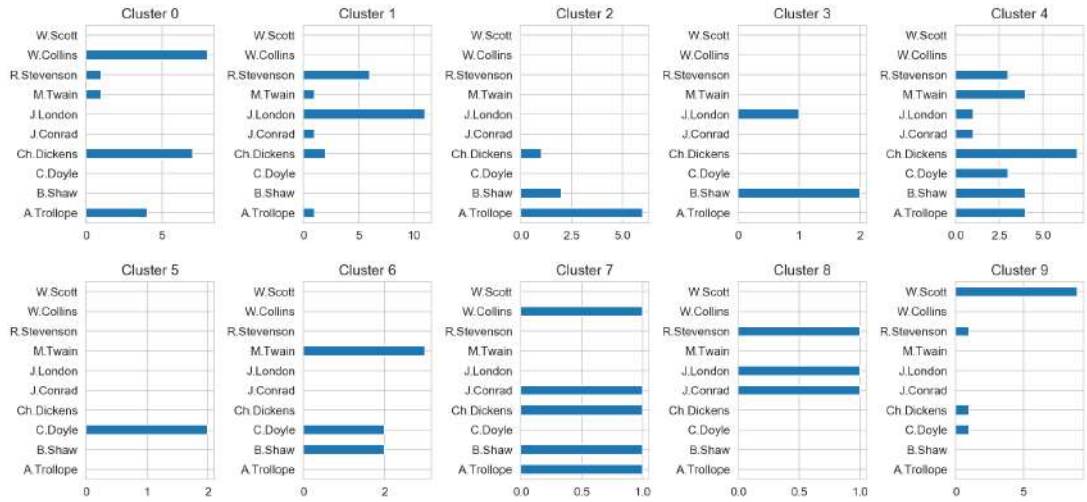

Рисунок А.42 – Розподіл документів у кластерах за авторами в алгоритмі агломеративної кластеризації у просторі LDA компонент

A.6.2 Кластеризація текстів груп новин за різними семантичними ознаками

Рисунок А.43 – Розподіл документів у кластерах за групами новин в алгоритмі агломеративної кластеризації у просторі семантичних полів

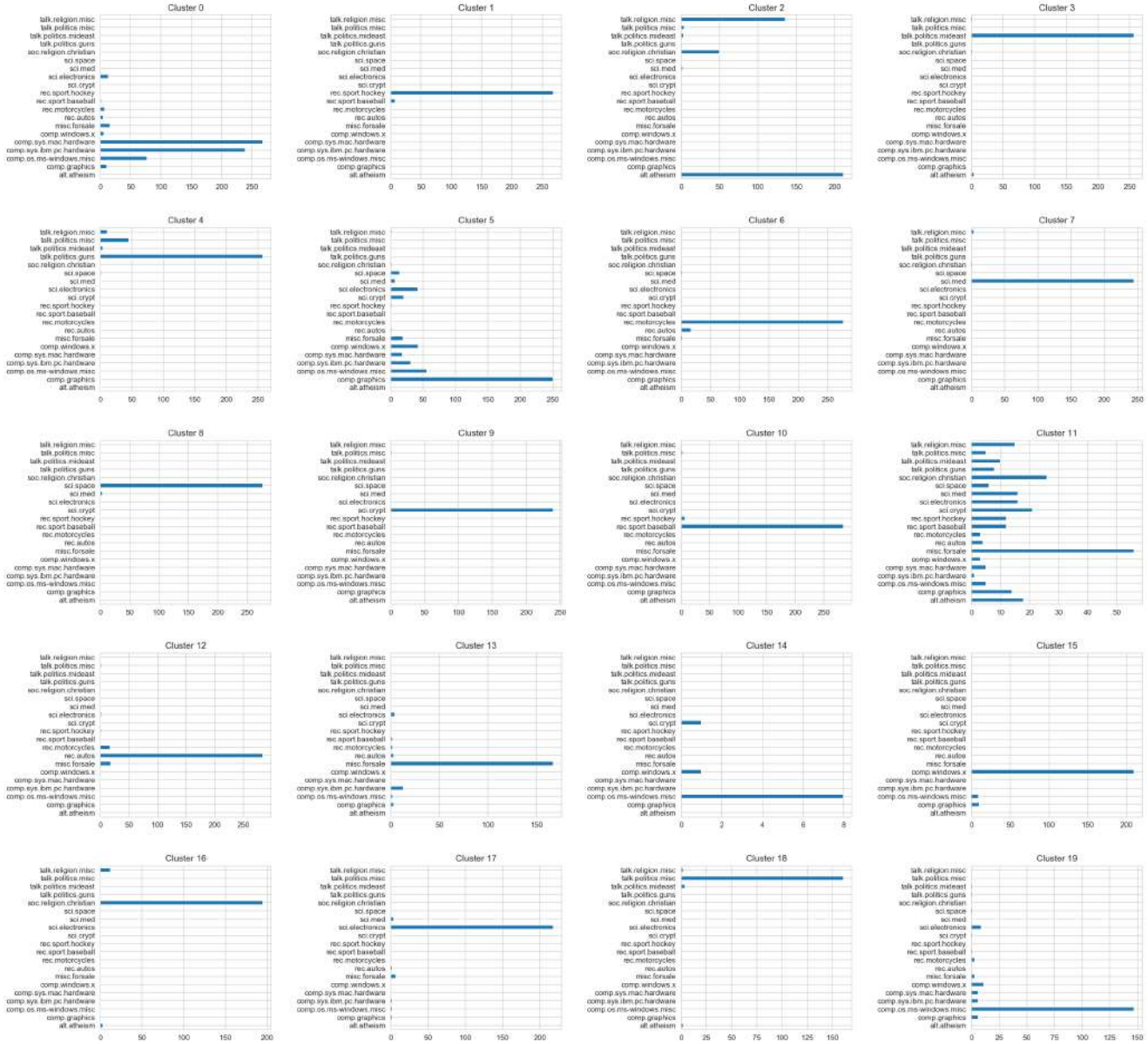

Рисунок А.44 – Розподіл документів у кластерах за групами новин в алгоритмі агломеративної кластеризації у просторі тематичних полів

Рисунок А.45 – Розподіл документів у кластерах за групами новин в алгоритмі агломеративної кластеризації у просторі SVD компонент

А.6.3 Сингулярна декомпозиція матриці семантичних ознак в алгоритмі ієрархічної кластеризації

Текстовий масив можна представити у вигляді матриці ознак слів (термів) та документів. Такими ознаками можуть бути текстові частоти лексем. Такий підхід має також ряд проблем, зокрема, розмірність аналізованого простору є великою, оскільки зумовлена розміром словника. Одним із шляхів вирішення цієї проблеми є використання латентного семантичного аналізу [151, 277, 278]. Суть такого аналізу полягає в сингулярному розкладі матриці ознак типу "ключові_слова–документи" і аналізі текстових масивів у новому векторному просторі меншої розмірності. Базис цього простору побудований на лінійних комбінаціях квантитативних характеристик лексем словника. Такий новий векторний простір часто називають простором концептів. Розмірність нового простору визначається кількістю найбільших сингулярних чисел – елементів діагональної матриці сингулярного розкладу. Документи також можуть бути квантитативно близькими не тільки за частотами окремих лексем, а також за характеристиками заданих лексемних об'єднань, зокрема семантичних полів. Розмірність матриці ознак «семантичні_поля–документи» є суттєво меншою у порівнянні із матрицею ознак для лексем словника текстових масивів. Семантичні поля формуються на основі експертного аналізу, одні і ті ж лексеми можуть одночасно належати до різних семантичних полів. Сингулярна декомпозиція матриці семантичних ознак дасть можливість аналізувати текстові масиви в ще меншому векторному просторі. Визначити ефективність такої декомпозиції можна аналізуючи утворення кластерної структури в новому семантичному просторі концептів для класифікованих за певною ознакою текстових документів. Такою ознакою може бути, наприклад, спільний стиль або автор. Сингулярна декомпозиція матриці семантичних ознак буде ефективною у випадку відображення класифікаційної структури в кластерній структурі, утвореній у новому векторному просторі семантичних концептів. Для аналізу ефективності сингулярної декомпозиції матриці семантичних ознак розглянемо сингулярний розклад матриці "частоти_семантичних_полів–документи" [280]. На прикладі тестової вибірки текстових документів проаналізуємо утворення ієрархічної кластерної структури у векторних просторах семантичних концептів різної розмірності. Далі співставимо класифікаційний розподіл текстових документів за авторами та утворену кластерну структуру в новому просторі семантичних концептів. Розглянемо сингулярний розклад матриці частот семантичних полів M_{sd} (3.17). Для експериментального аналізу було сформовано матрицю типу частоти_семантичних_полів-документи використовуючи текстову вибірку та параметри семантичних полів, які були вибрані для ієрархічної кластеризації у попередніх дослідженнях. Кожний документ розглядався як вектор в семантичному просторі. Далі було проведено сингулярний розклад матриці семантичних ознак. На наступному етапі була проведена агломеративна ієрархічна кластеризація документів у просторах семантичних концептів різної розмірності. Для оцінки міжкластерних відстаней

використовувалась евклівдова відстань, а кластеризація проведена методом Варда. Розглянемо результати чисельного кластерного аналізу авторських текстів у просторі SVD компонент. На рис. А.46 показана дендрограма агломеративної кластеризації у просторі SVD компонент. На рис. А.47 показано розподіл документів у кластерах по авторах в алгоритмі агломеративної кластеризації у просторі SVD компонент. Наведені дендрограми обмежені рівнем із 10-ма кластерами. Розглянемо кластеризацію авторських текстів у ортогональному низько розмірному просторі вторинних семантичних полів утворених на основі SVD факторизації матриці семантичних полів. Для утворення ортогонального семантичного підпростору взято координати вторинних семантичних полів, які відповідають першим 30 сингулярним числам матриці. На основі утвореного низькорозмірного ортогонального простору проведені аналогічні розрахунки кластерних структур. Як випливає із отриманих даних спостерігаються кластери із домінуванням окремих авторів. Такі кластери характеризують семантичну область авторського ідіолекта у низькоромірному семантичному просторі із ортогональним базисом. Формування кластерів із домінуванням ідіолекту одного автора визначається як вибором базису семантичного простору, так і методом кластеризації. Аналізуючи кластери із домінуванням текстів одного автора, які отримані методом агломеративної кластеризації у просторі семантичних полів в ортогональному семантичному просторі та методом k-means в ортогональному просторі, можна побачити, що є автори, тексти яких домінують у всіх цих випадках. Семантичні кластери із домінуванням авторського ідіолекту цих авторів можна розглядати як семантично інваріантні і незалежні від розглянутих семантичних просторів та методів кластеризації.

Отже, формування простору семантичних полів дає можливість отримувати новий структурний поділ документів за семантичними ознаками. Сингулярний розклад матриці семантичних ознак типу “частоти_семантичних_полів–документи” дає можливість аналізувати текстові документи у новому просторі семантичних концептів. Ієрархічна кластеризація документів у такому просторі відображає класифікаційну структуру документів за різними ознаками, зокрема за авторством текстів. Розмірність простору семантичних концептів визначається рангом апроксимації матриці семантичних ознак при сингулярному розкладі і може бути суттєво меншою за розмірність простору семантичних полів. У випадку дослідження авторства текстів вибір розмірності простору семантичних концептів зумовлений рівнем відображення класифікаційного поділу документів за авторами в кластерній структурі, що визначається наявністю домінуючих кластерів для документів окремих авторів. Сингулярний розклад матриці семантичних полів дає можливість суттєво зменшити розмірність семантичного простору в кластерному аналізі текстових даних [280].

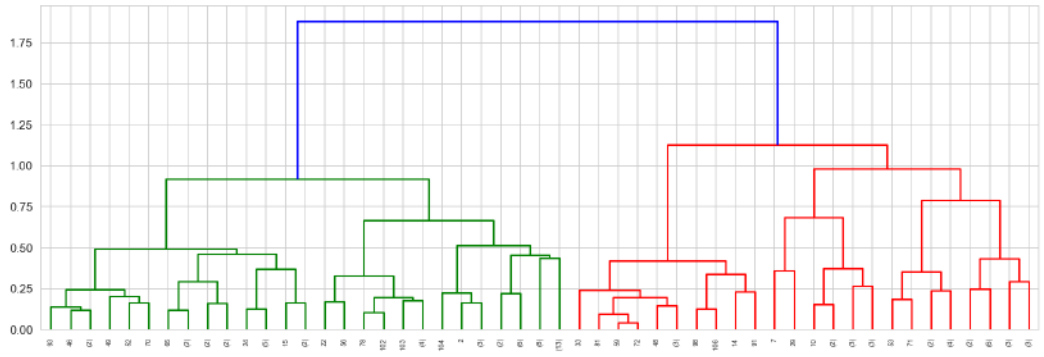

Рисунок А.46 – Дендрограма агломеративної кластеризації у просторі SVD компонент

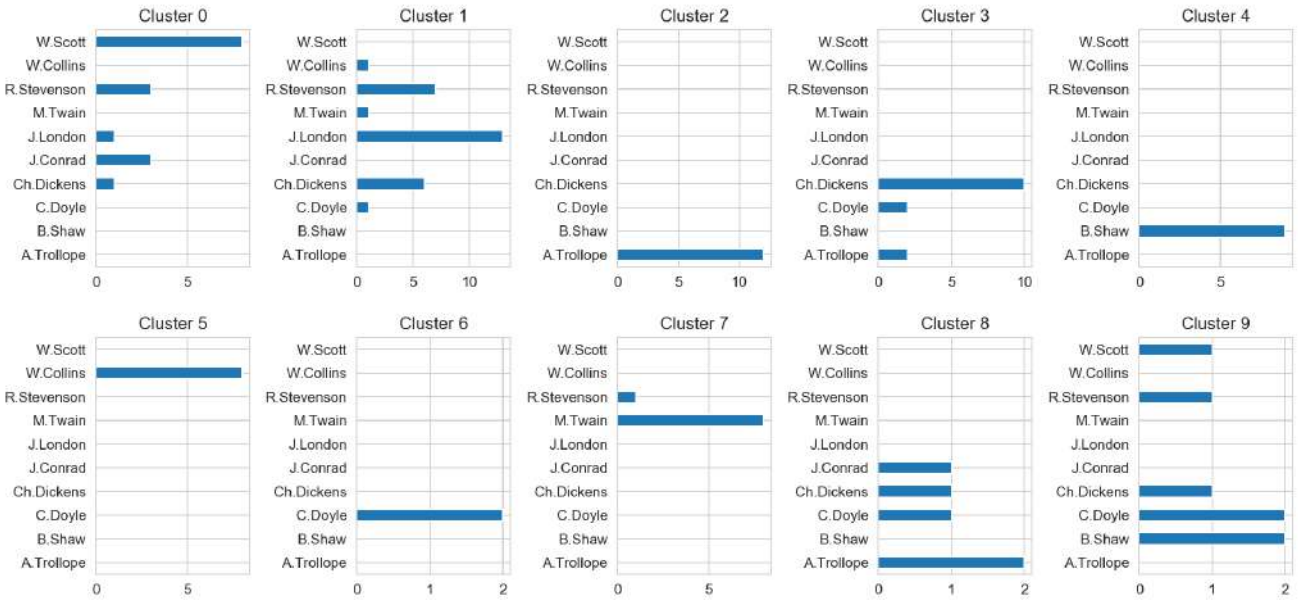

Рисунок А.47 – Розподіл документів у кластерах по авторах в алгоритмі агломеративної кластеризації у просторі SVD компонент

А.6.4 Класифікаційний аналіз текстових даних при використанні різних семантичних ознак

Класифікаційний аналіз здійснювався за допомогою алгоритму Random Forest пакету *scikit-learn* [217]. На рис. А.48,А.49 наведено оцінки класифікаційного аналізу авторських текстів, а на рис. А.50–А.53 – груп новин.

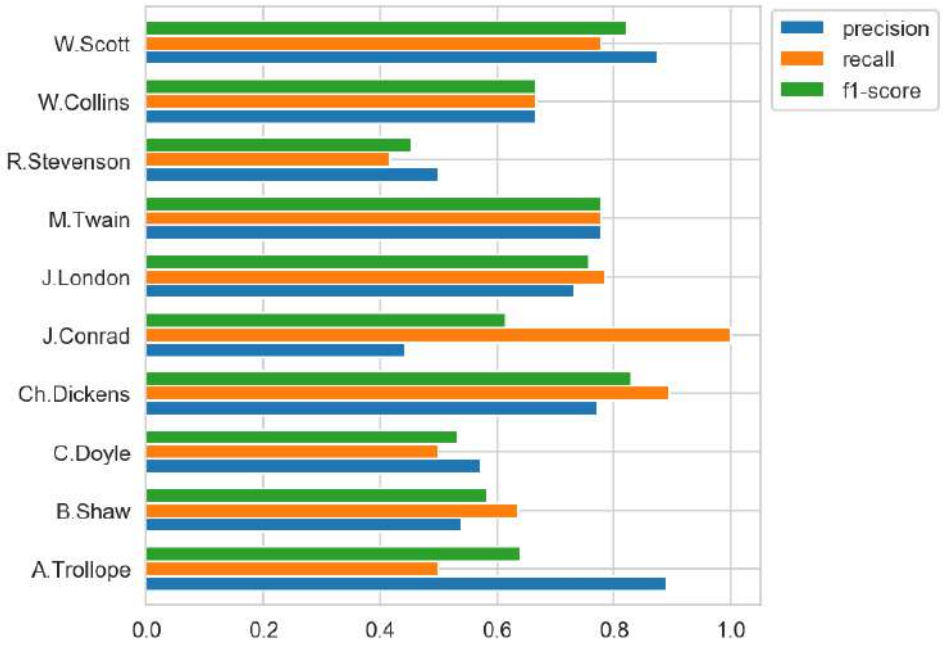

Рисунок А.48 – Оцінки класифікації авторських текстів при використанні ознак на основі компонент LDA

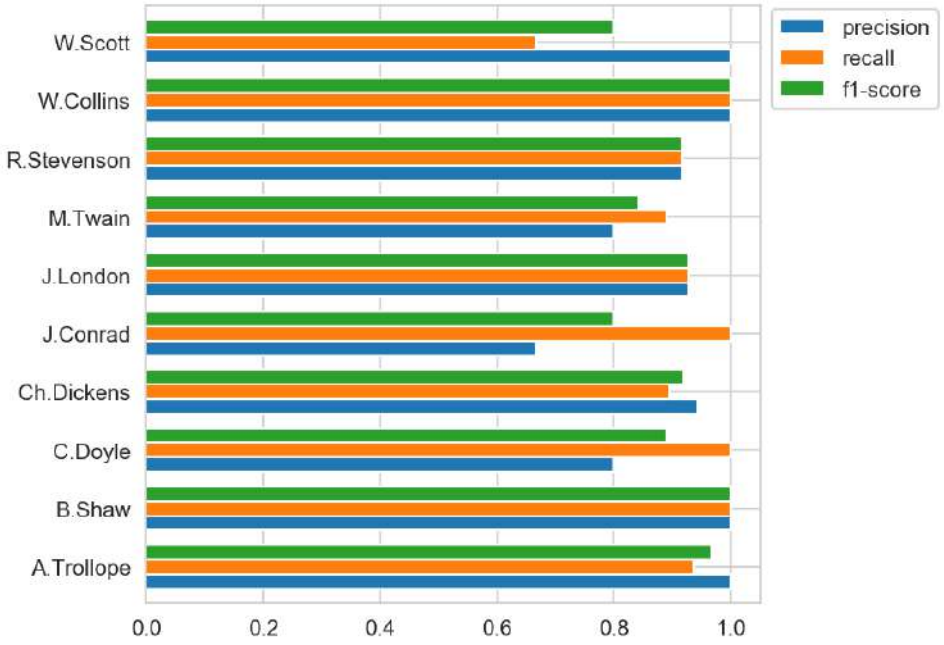

Рисунок А.49 – Оцінки класифікації авторських текстів при використанні ознак на основі компонент SVD

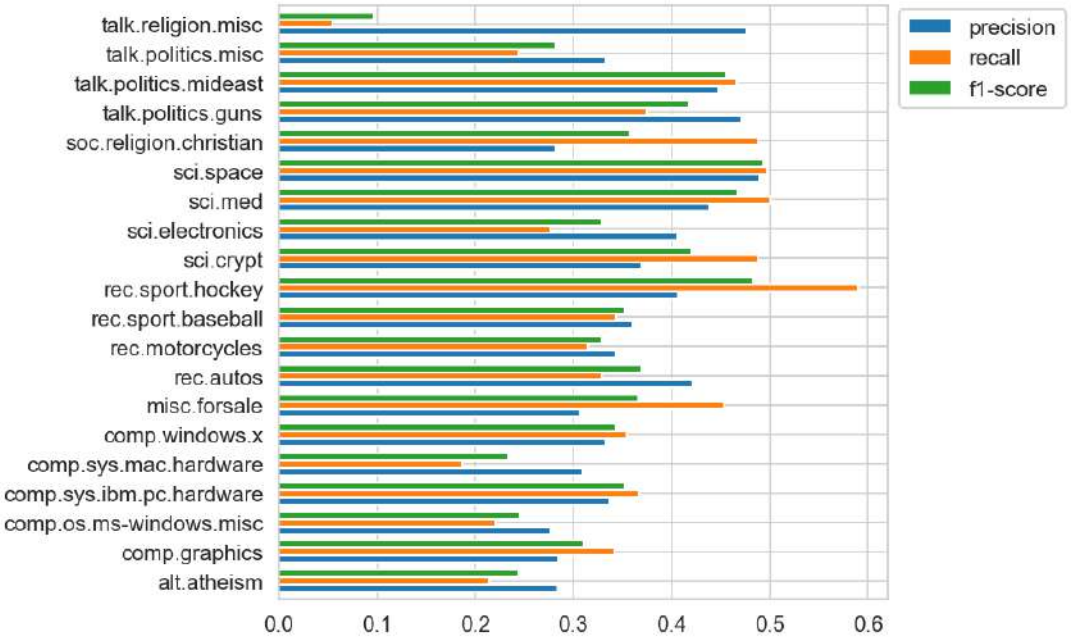

Рисунок А.50 – Оцінки класифікації текстів груп новин при використанні ознак на основі семантичних полів

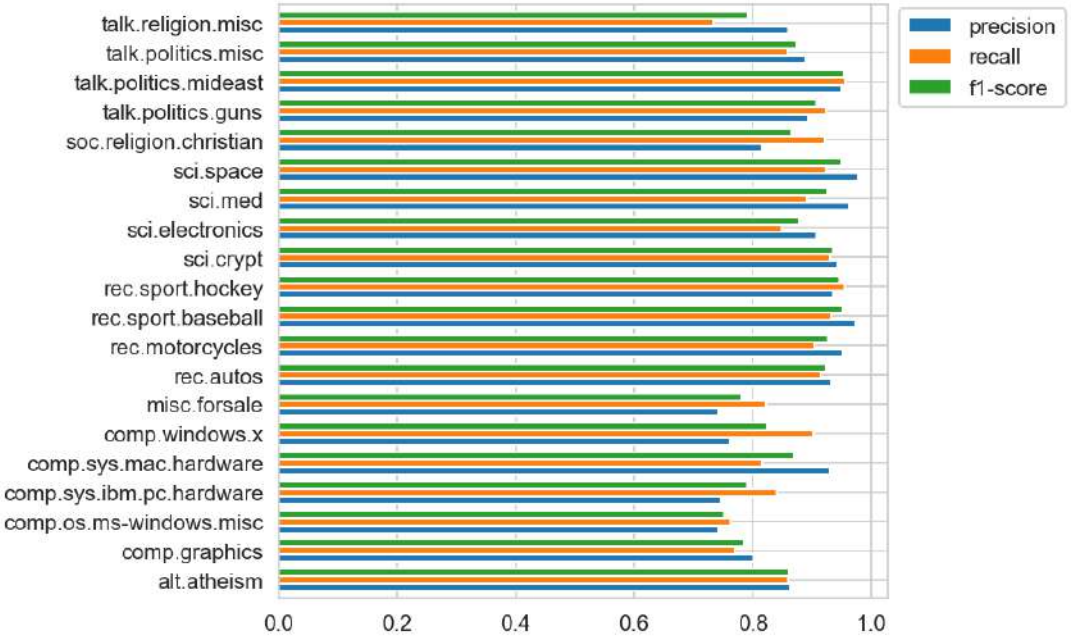

Рисунок А.51 – Оцінки класифікації текстів груп новин при використанні ознак на основі тематичних полів

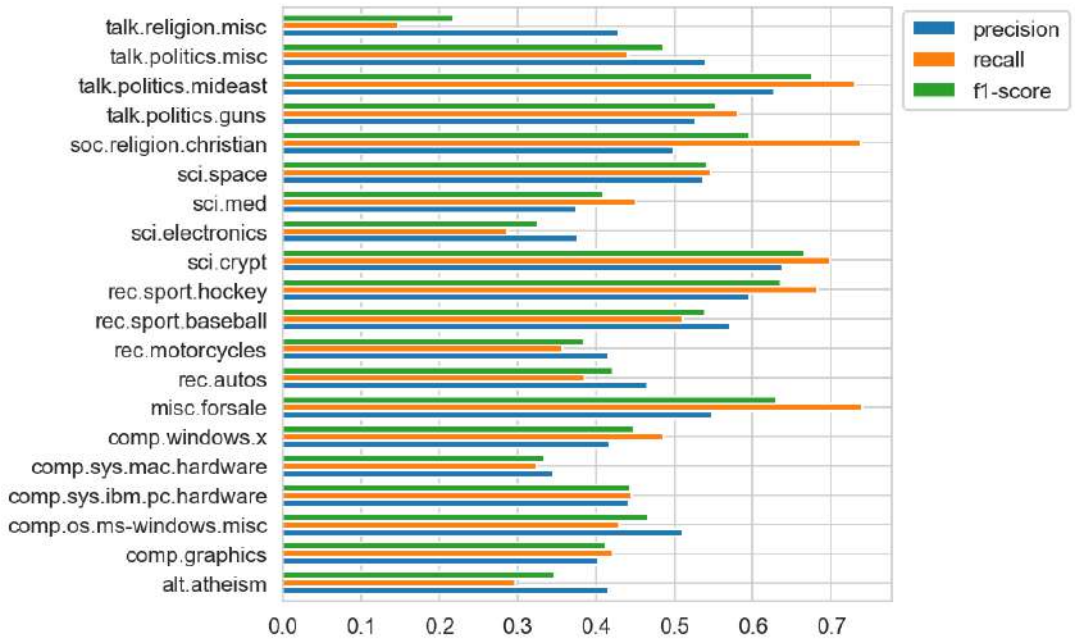

Рисунок А.52 – Оцінки класифікації текстів груп новин при використанні ознак на основі компонент LDA

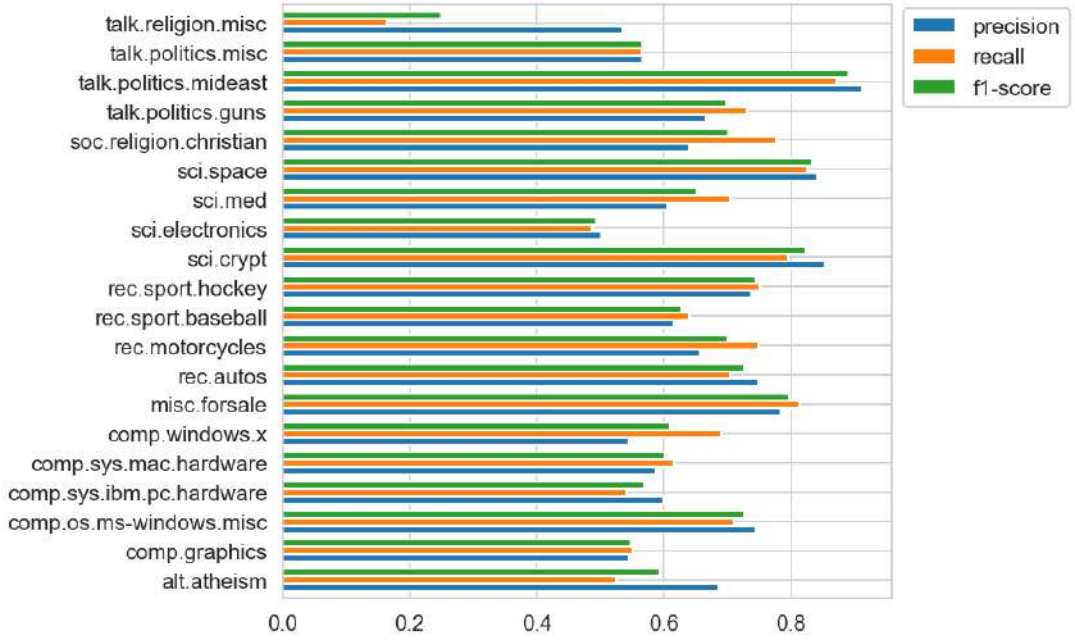

Рисунок А.53 – Оцінки класифікації текстів груп новин при використанні ознак на основі компонент SVD

Розглянуто випадок використання сукупного набору семантичних ознак за допомогою класифікатора на основі нейронної мережі. Структура нейронної мережі наведено на рис. А.54. Результати класифікаційного аналізу наведено на рис. А.55.

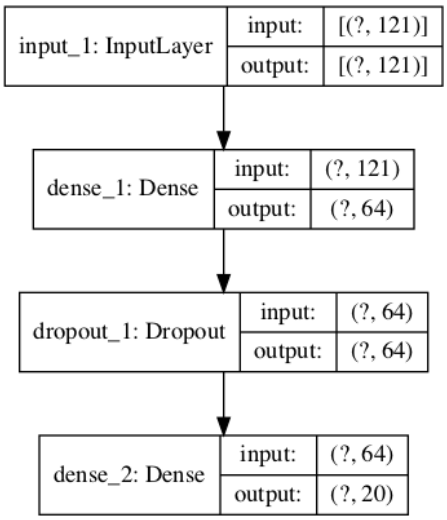

Рисунок А.54 – Структура нейронної мережі

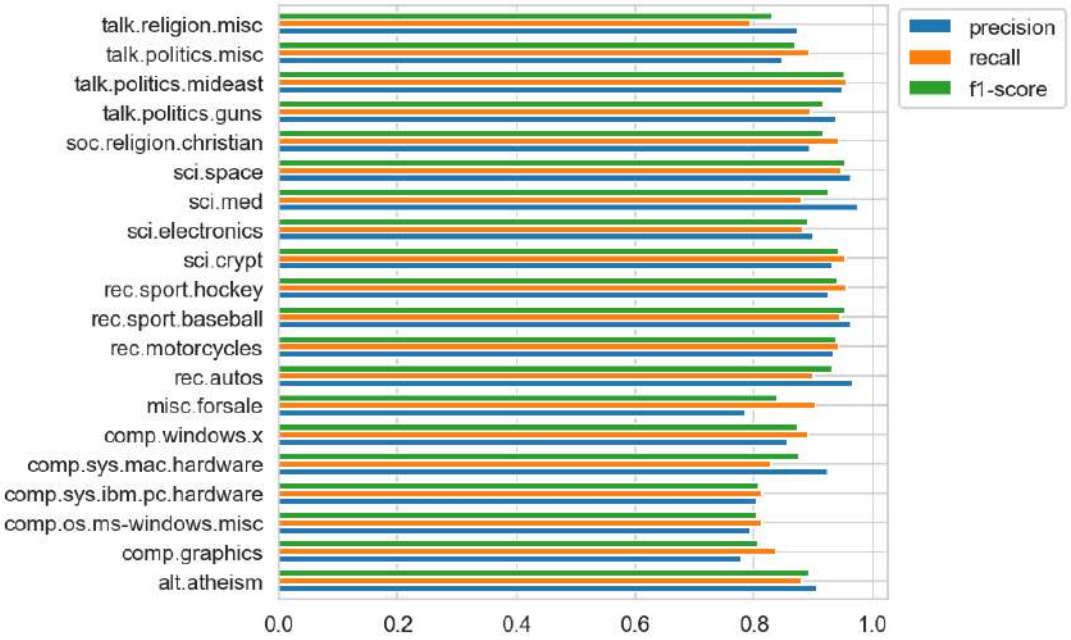

Рисунок А.55 – Оцінки класифікації текстів груп новин нейронною мережею при використанні сукупних семантичних ознак

А.7 Квантові обчислення

Квантові комп'ютери та алгоритми дозволяють суттєво пришвидшити розв'язок деяких класів задач внаслідок реалізації квантового паралелізму та заплутаності квантових станів [308, 309, 310, 311, 312, 313, 314, 315].

В основі квантових логічних елементів лежить поняття квантового біту - кубіту, який є вектором одиничної довжини в 2-вимірному комплексному векторному просторі з базисом $\{|0\rangle, |1\rangle\}$ [313, 314]. В класичному випадку n двохстанових елементів утворюють $2n$-мірний простір. В квантових системах n кубітів утворюють простір вимірності 2^n.

Розглянемо базові операції над кубітами. Оператор тотожного перетворення не змінює значення кубітів і в матричному записі має вигляд

$$i = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{A.2}$$

Оператор заперечення X використовують для реалізації інверсії значень кубітів і його визначають так:

$$X = |0\rangle\langle 1| + |1\rangle\langle 0|. \tag{A.3}$$

У матричному зображенні оператор заперечення має вигляд

$$X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \tag{A.4}$$

Одним із важливих елементів є 'контрольоване НЕ', яке здійснюється над двома кубітами і змінює значення другого кубіта на протилежне, якщо значення першого кубіта рівне 1. Цей логічний елемент може бути визначений як

$$U_{CNOT} = |0\rangle\langle 0| \otimes i + |1\rangle\langle 1| \otimes X, \tag{A.5}$$

а матриця оператора унітарного перетворення 'контрольоване НЕ' має вигляд

$$U_{CNOT} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}. \tag{A.6}$$

Дію вентиля 'контрольоване НЕ' можна зобразити так

$$U_{CNOT} : |a, b\rangle \to |a, a \oplus b\rangle, \tag{A.7}$$

де $\oplus$ означає додавання за модулем 2. Ще одним важливим логічним елементом є вентиль Тоффолі, який діє на три кубіти і змінює значення третього кубіта на протилежне, якщо значення першого та другого кубіта рівно 1. Від логічного елемента 'контрольоване НЕ' вентиль Тофолі відрізняється наявністю ще одного додаткового керуючого кубіта. Цей вентиль можна визначити як

$$T = |0\rangle\langle 0| \otimes i \otimes i + |1\rangle\langle 1| \otimes U_{CNOT} \tag{A.8}$$

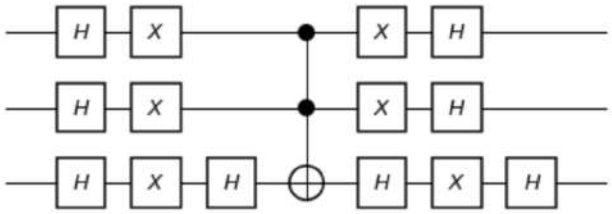

Рисунок А.56 – Приклад зображення квантової схеми

Перетворення Тофолі можна зобразити так

$$T : |a, b, c\rangle \rightarrow |a, b, c \oplus ab\rangle, \tag{A.9}$$

Вентиль Тофолі є універсальним квантовим логічним елементом на основі якого можна побудувати оборотну квантову машину Тюрінга [313, 314].

Одним із ефективнихз квантових алгоритмів є алгоритм Гровера [310, 311, 312], який дає можливість реалізувати пошук у невпорядкованій вибірці даних поліноміально швидше у порінянні з класичними алгоритмами внаслідок реалізації квантового паралелізму. Актуальним є розгляд можливості використання квантових алгоритмів в інтелектуальному аналізі даних слабоструктурованого типу.

На даний час вже існують реальні квантові комп'ютери, зокрема фірми IBM, які дають можливість проводити прості експериментальні квантові обчислення. Фірма IBM надає вільний доступ до декількох квантових комп'ютерів через хмарковий сервіс. Реалізувати такий доступ та провести прості квантові обчислення можна за допомогою мови програмування Python із використанням пакету Qiskit [318, 319, 320, 321, 322]. Квантові алгоритми часто представляють за допомогою квантової схеми. Приклад такої квантової схеми показано на рис. А.56. На рис. А.57, А.58 показана реалізація заплутаних квантових станів на на двох різних реальних квантових комп'ютерах IBM Q. Можна побачити, що результати дещо відрізняються. Це зумовлено наявністю квантових помилок при реалізації обчислень на реальних квантових комп'ютерах. Такі помилки виникають тому, що внаслідок наявних шумів зумовлених взаємодією квантової системи із середовищем зникають чисті квантові стани і стає неможливим реалвзація унітарних перетворень у квантових алгоритмах. Результати наведені на цих рисунках отримані при реалізації великої кількості спроб і відображають статистичний результат реалізації квантових алгоритмів. На рис.А.59 показано результат реалізації алгоритму Гровера на симуляторі квантових обчислень IBM Q.

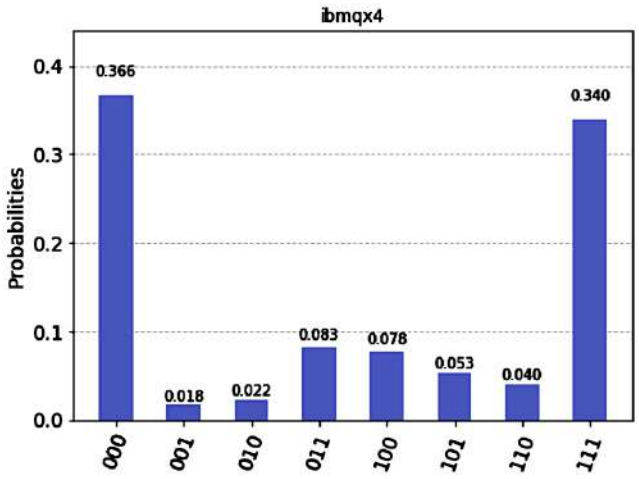

Рисунок А.57 – Реалізація заплутаних квантових станів на реальному квантовому комп'ютері IBM Q

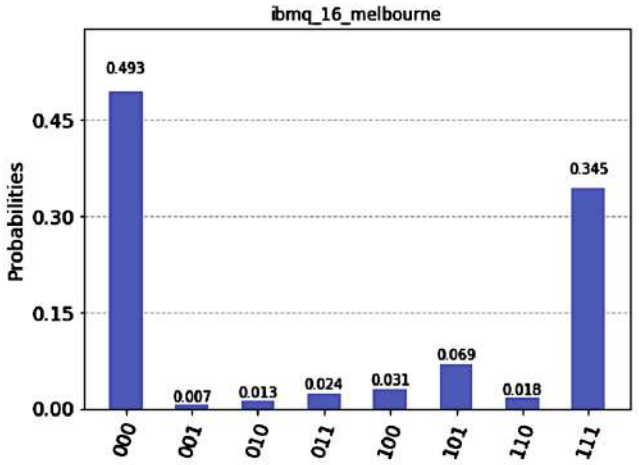

Рисунок А.58 – Реалізація заплутаних квантових станів на реальному квантовому комп'ютері IBM Q

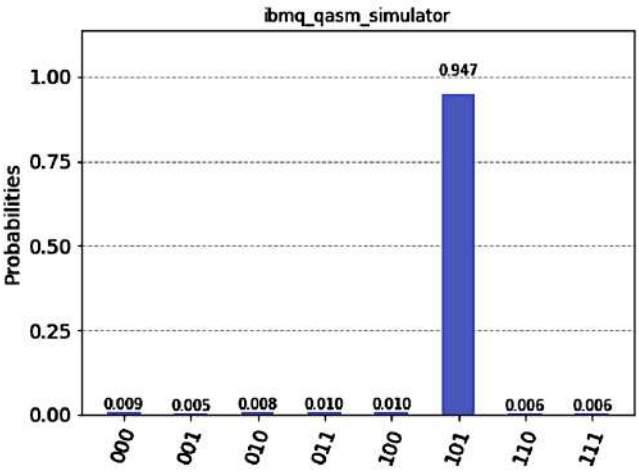

Рисунок А.59 – Реалізація алгоритму Гровера на симуляторі квантових обчислень IBM Q

A.8 Формування прогнозних ознак на основі трендів у спільнотах соціальних мереж

Тренди соціальних мереж, пов'язані з деякими бізнес сутностями, мають великий вплив на споживання послуг або продуктів, пов'язаних з цими сутностями. Можна спостерігати штучно створені тренди з метою сформувати певне ставлення користувачів до деяких сервісів компаній та соціальних і бізнес-процесів. Важливим є виявлення та аналіз таких трендів та спільнот користувачів, які формують ці тренди. Кількісні ознаки, які описують діяльність спільноти, можуть мати прогнозний потенціал і можуть бути використані у прогнозній аналітиці. Проаналізуємо спільноти користувачів у соціальній мережі Твіттер, які виникають навколо обговорення питань, пов'язаних із компанією Zoom. Твіти з ключовим словом "zoom" завантажувались в період часу з 25.04.2020 р. по 04.06.2020 р. Спільноти користувачів формуються внаслідок взаємодії користувачів, зокрема, за допомогою ретвітів, відповідей тощо. У твітах можна знайти семантичні сутності, пов'язані з аналізованим процесом. На рис. А.60 показані частоти лексем сформованого тематичного поля, пов'язаного з компанією Zoom. Використовуючи теорію частих множин та асоціативних правил можна знайти семантичну структуру твітів, яка розкриває зв'язок між сутностями в аналізованих трендах. На рис. А.61-А.64 показано часті множини, а на рис. А.65-А.66 – асоціативні правила, виявлені в трендах Твіттера.
Розглянемо спільноти користувачів. Як ознаки, використано відношення користувачів до твітів, які представлені двійковою матрицею. Ми застосували перетворення SVD та TSNE до цієї матриці і отримали структуру твітів у представленні TSNE, яке зображено на рис. А.67. Отримані результати показують, що існує структура твітів, яка зумовлена активністю спільнот користувачів, що формують тренди в обговоренні питань, пов'язаних із компанією Zoom. Проведено спектральна бікластеризація та виявлено бікластери, які складаються з користувачів та твітів, пов'язаних із цими користувачами. Кожен з цих кластерів описує групи користувачів, які формують різні тенденції думок. Кількісні характеристики активності цих груп можуть бути використані як ознаки в прогнозних моделях для аналізу різних бізнес-процесів. На рис. А.68 показано часовий ряд кількісної характеристики активності різних груп користувачів у логарифмічній шкалі. Аналізуючи твіти різних бікластерів, можна побачити, що різні кластери мають різні тренди настроїв по відношенню до заданих сутностей. Для визначення характеристик настрою та характеристик користувачів у спільнотах використано сервіс API IBM Watson Personality Insights [329]. На рис. А.69-А.71 показано характеристики настроїв для твітів, пов'язаних із різними кластерами користувачів. На рис. А.72-А.75 показано характеристики спектру емоцій для твітів, пов'язаних із різними кластерами користувачів.
Отже, як показують отримані результати, твіти різних груп користувачів мають різну структуру характеристик настроїв та емоцій. Кількісні ознаки частих множин твітів та характеристики настроїв і емоцій різних спільнот користувачів можна розглядати як додаткові ознаки в прогнозних моделях для бізнес аналітики.

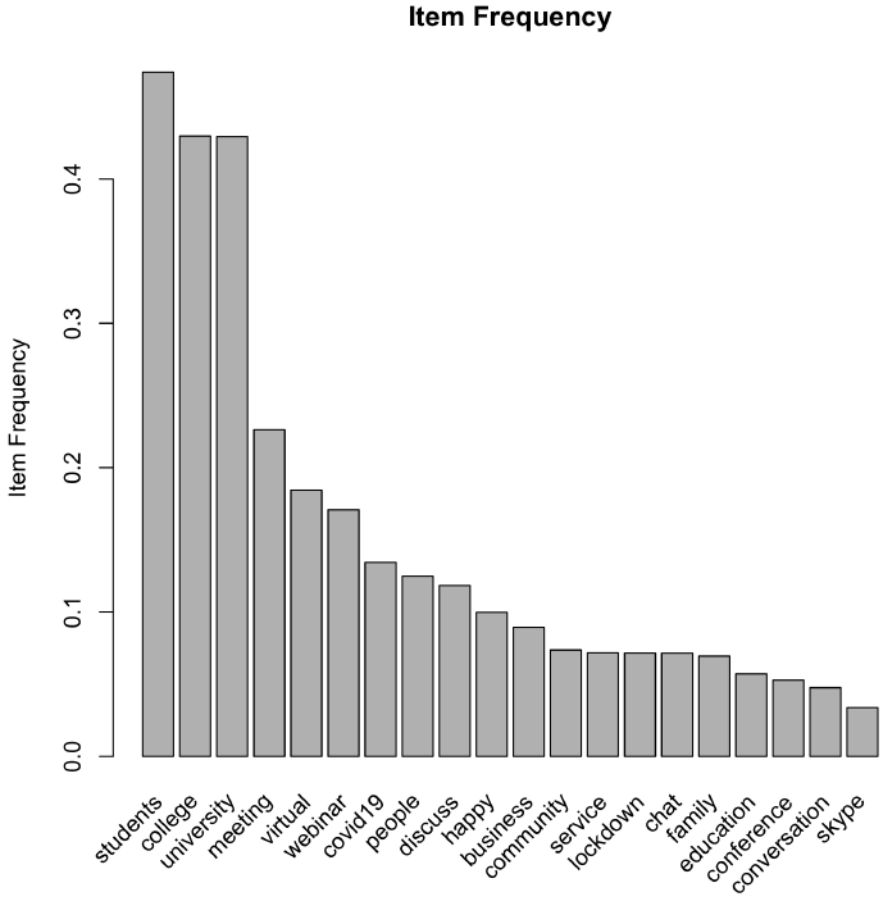

Рисунок А.60 – Частоти лексем сформованого семантичного поля, пов'язаного з компанією Zoom

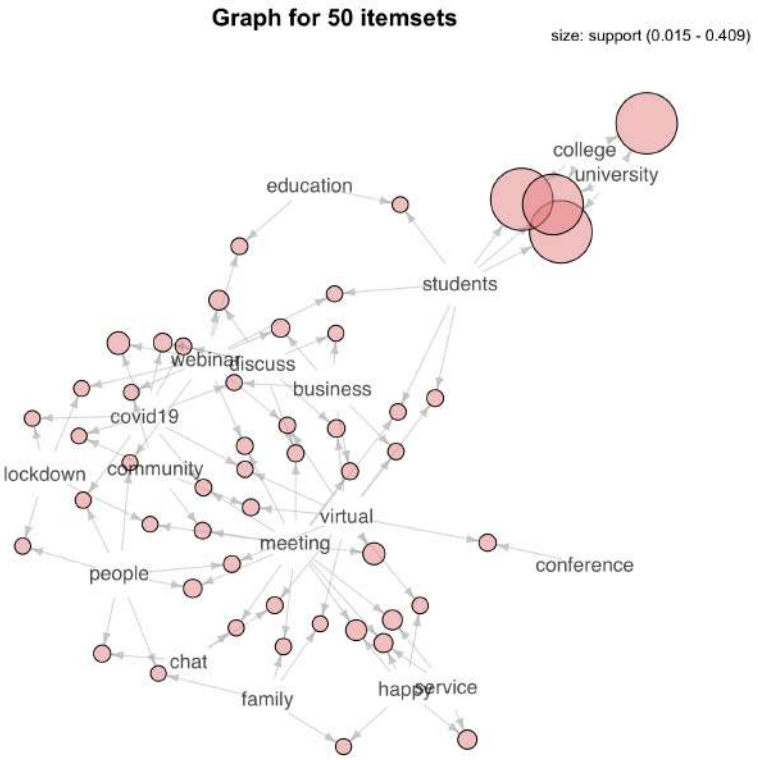

Рисунок А.61 – Виявлені часті множини в трендах Твіттера

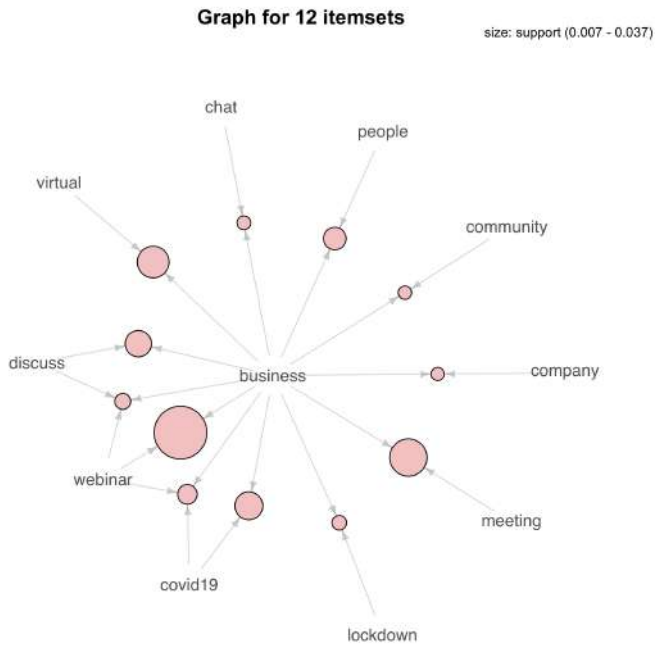

Рисунок А.62 – Виявлені часті множини в трендах Твіттера

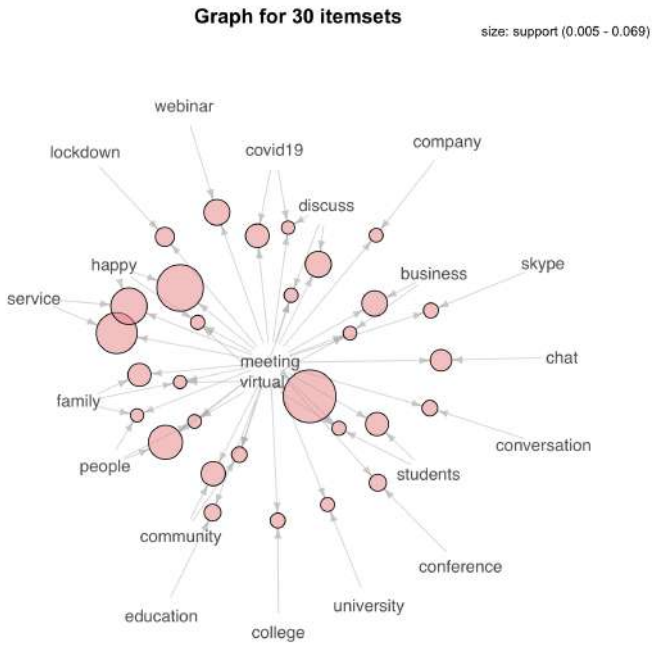

Рисунок А.63 – Виявлені часті множини в трендах Твіттера

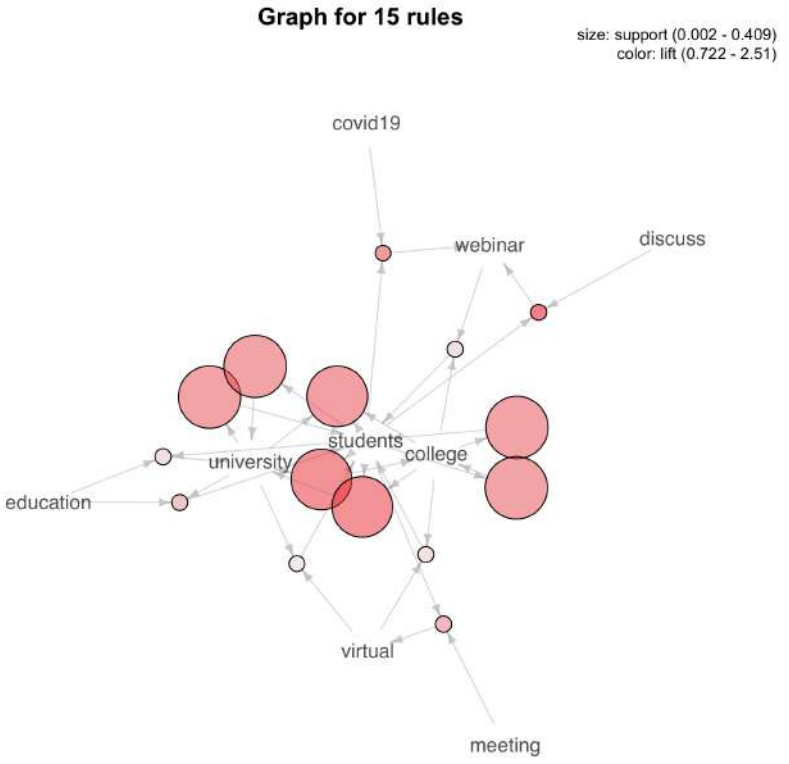

Рисунок А.64 – Виявлені часті множини в трендах Твіттера

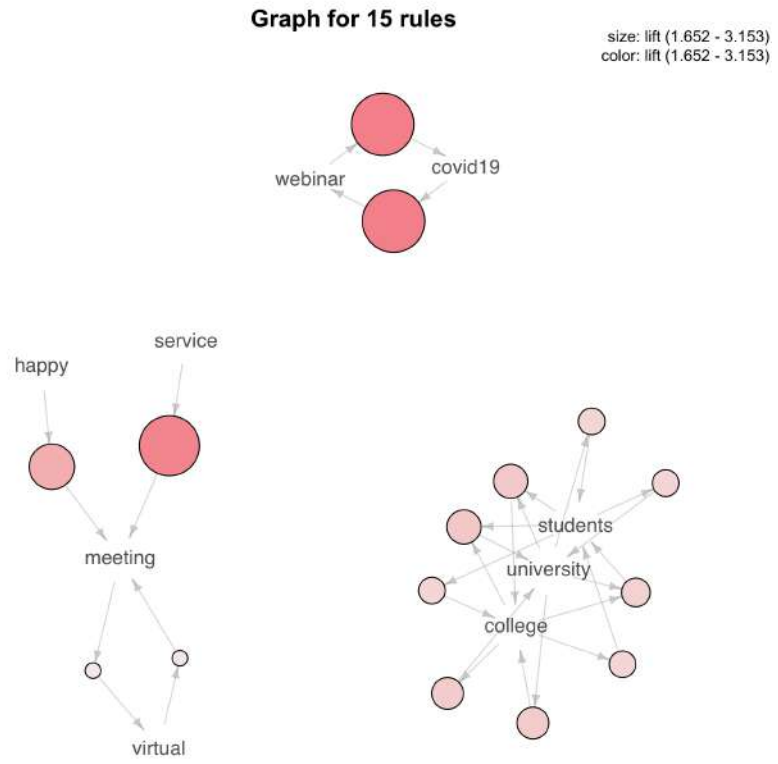

Рисунок А.65 – Виявлені асоціативні правила в трендах Твіттера

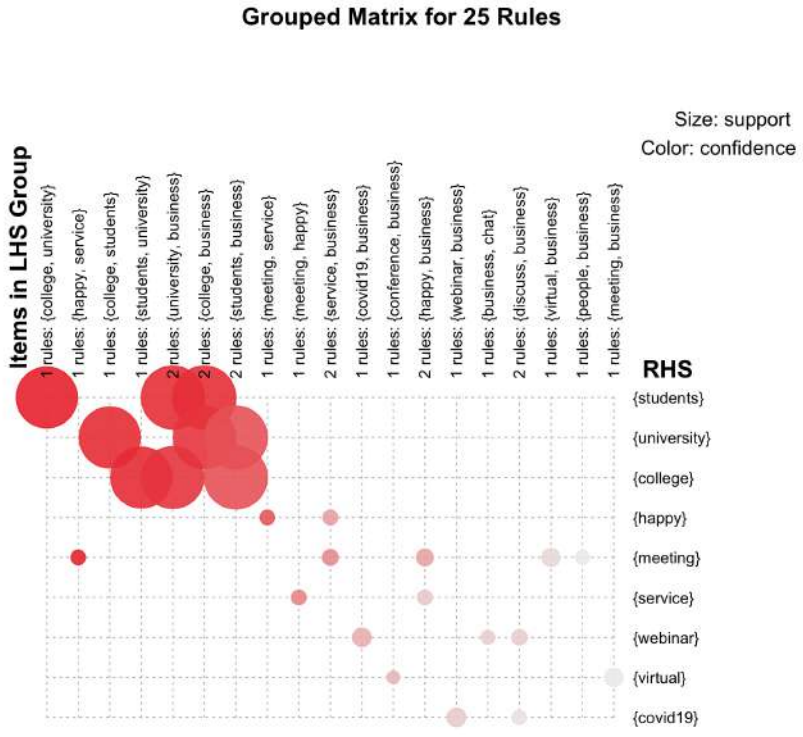

Рисунок А.66 – Виявлені асоціативні правила в трендах Твіттера, згруповані в таблицю

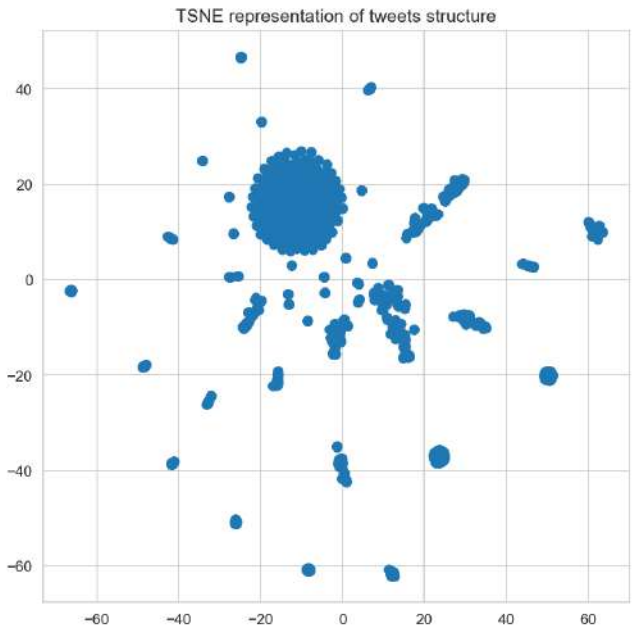

Рисунок А.67 – Структуру твітів у представленні TSNE

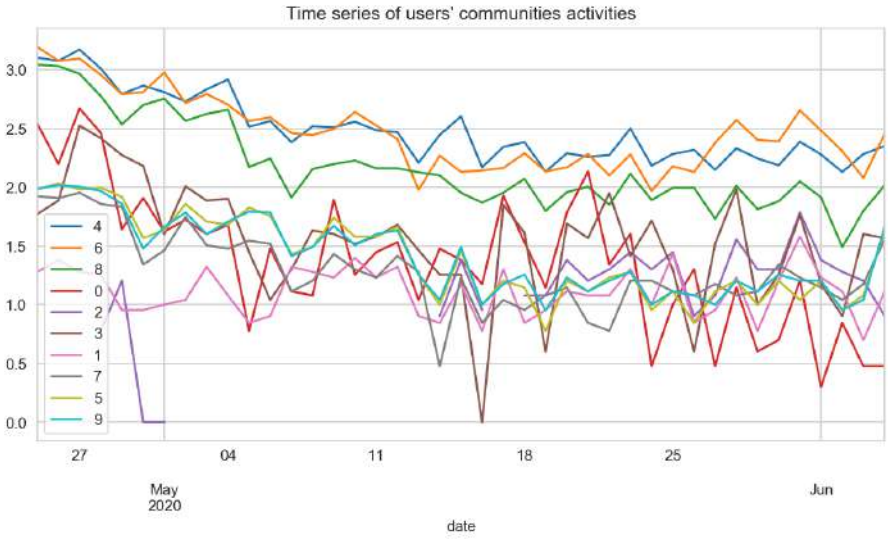

Рисунок А.68 – Часовий ряд активності груп користувачів

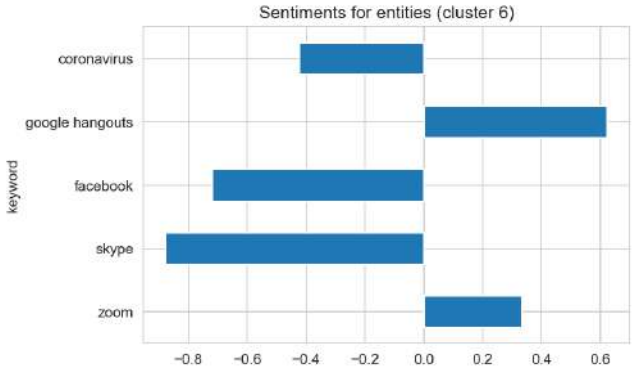

Рисунок А.69 – Характеристики настроїв у масивах твітів заданого кластера користувачів

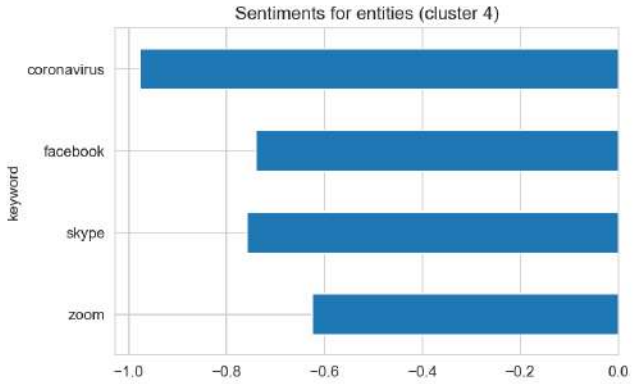

Рисунок А.70 – Характеристики настроїв у масивах твітів заданого кластера користувачів

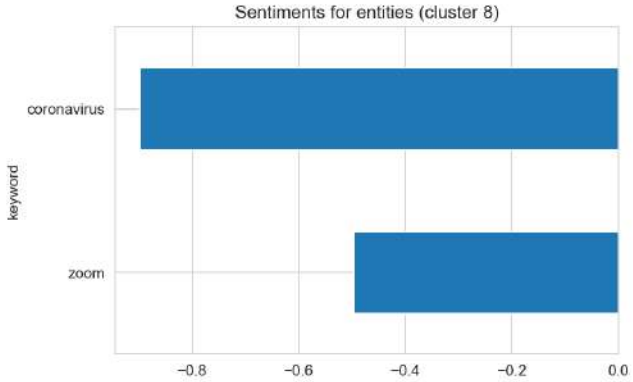

Рисунок А.71 – Характеристики настроїв у масивах твітів заданого кластера користувачів

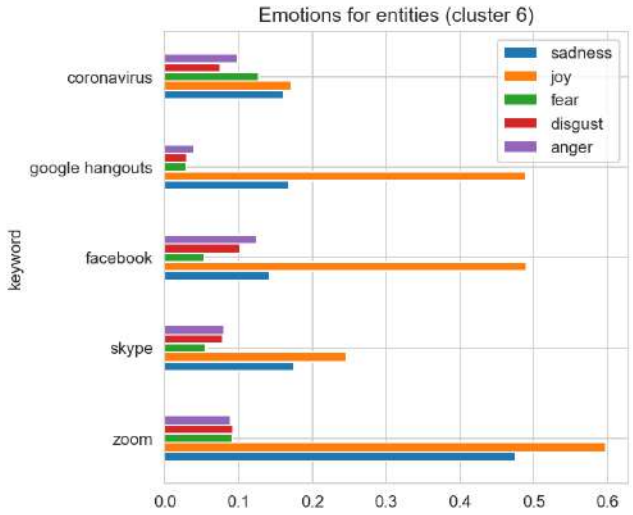

Рисунок А.72 – Характеристики спектру емоцій у масивах твітів заданого кластера користувачів

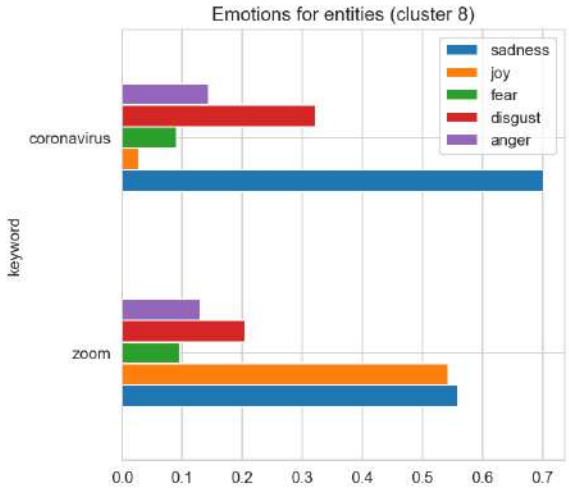

Рисунок А.73 – Характеристики спектру емоцій у масивах твітів заданого кластера користувачів

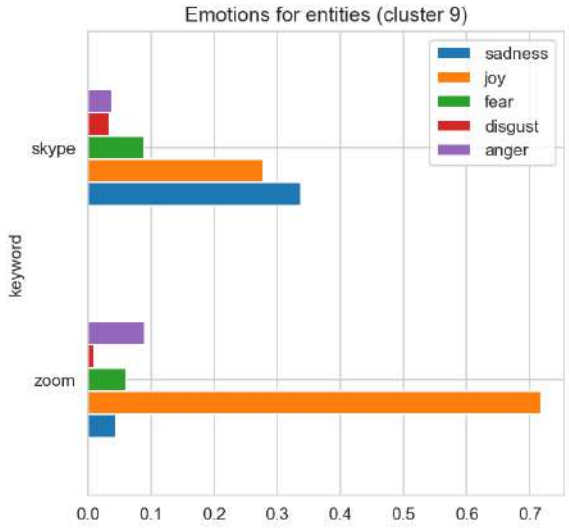

Рисунок А.74 – Характеристики спектру емоцій у масивах твітів заданого кластера користувачів

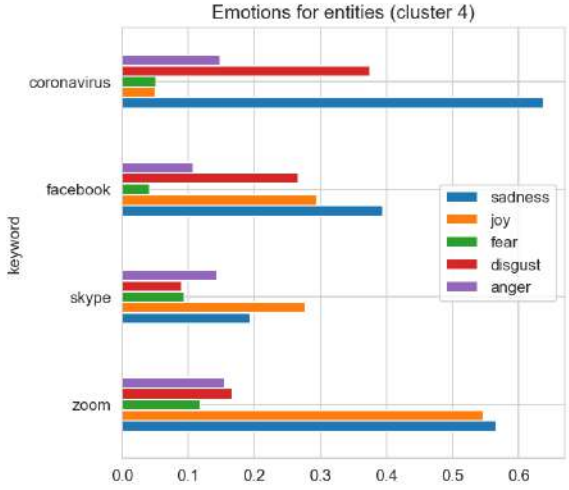

Рисунок А.75 – Характеристики спектру емоцій у масивах твітів заданого кластера користувачів

А.9 Акти впровадження дисертаційних досліджень

softserve

16 листопада 2020

АКТ
про використання результатів дисертаційної роботи “Методи інтелектуального аналізу консолідованих даних для підтримки прийняття рішень” Павлишенка Богдана Михайловича при розробці програмного забезпечення в компанії SoftServe Inc.

Цей акт підтверджує, що результати дисертаційної роботи Павлишенка Богдана Михайловича на тему “Методи інтелектуального аналізу консолідованих даних для підтримки прийняття рішень”, поданої на здобуття наукового ступеня доктора технічних наук, використовуються в компанії SoftServe Inc. для розробки програмного забезпечення пов'язаного із аналізом даних.
Компанія SoftServe Inc. застосовує запропоновану автором методику інтелектуального аналізу даних із використанням методів машинного навчання для отримання інформативних результатів прогнозування та аналітики даних із різних предметних областей. Ці методики використовуються відділом Data Science Group у процесі розробки програмного забезпечення та надання консультаційних послуг кліентам компанії SoftServe Inc.

Сергій Газієв,
Старший віце-президент по провідним
технологіям компанії SoftServe Inc.

shaziyev@softserveinc.com, +1 239 703 4764,
Austin HQ, 201 W 5th St #1550, Austin, TX 78701

ЗАТВЕРДЖУЮ
Проректор з наукової роботи
Львівського національного
університету імені Івана Франка
проф. Гладишевський Р. Є.
"28" 12 2020 р.

АКТ
про використання результатів дисертаційної роботи
"Методи інтелектуального аналізу консолідованих даних для підтримки прийняття рішень" докторанта кафедри системного проектування Павлишенка Богдана Михайловича при виконанні держбюджетних науково-дослідних тем

Комісія у складі начальника НДЧ доктора фізико-математичних наук, старшого наукового співробітника Плевачука Ю.О., завідувача кафедри системного проектування доцента канд. фіз.-мат. наук Шувара Р.Я., заступника декана факультету елекетроніки і комп'ютерних технологій доцента канд. фіз.-мат. наук Вельгоша С.Р. підтверджує, що результати дисертаційної роботи "Методи інтелектуального аналізу консолідованих даних для підтримки прийняття рішень" докторанта кафедри системного проектування Павлишенка Б.М. поданої на здобуття наукового ступеня доктора технічних наук, були використані при виконанні науково-дослідної теми "Аналіз даних засобами машинного навчання" (номер держреєстрації 0119U002409, термін виконання 01.01.2019 - 31.12.2021).

Серед результатів отриманих Павлишенком Б.М. використано, зокрема підходи у використанні методів машинного навчання в аналізі даних, методи побудови ансамблів прогнозних моделей, методи прогнозної аналітики часових рядів, методи семантичного аналізу текстових масивів даних, методи аналізу соціальних мереж.

Голова комісії: Плевачук Ю.О.

Члени комісії: Шувар Р.Я.

Вельгош С.Р.

Затверджую
Проректор з наукової роботи
Львівського національного
університету імені Івана Франка
проф. Гладишевський Р. Є.

"12" 12 2020 р.

АКТ

Про впровадження результатів дисертаційної роботи "Методи інтелектуального аналізу консолідованих даних для підтримки прийняття рішень" докторанта кафедри системного проектування Павлишенка Богдана Михайловича в навчальний процес

Комісія у складі завідувача кафедри системного проектування доцента, канд. фіз.-мат. наук Шувара Р.Я., доцента кафедри системного проектування, канд. фіз.-мат. наук Ненчука Т.М., доцента кафедри системного проектування, канд. фіз.-мат. наук Демків Л.С. підтверджує, що результати дисертаційної роботи "Методи інтелектуального аналізу консолідованих даних для підтримки прийняття рішень" докторанта кафедри системного проектування Павлишенка Б.М., поданої на здобуття наукового ступеня доктора технічних наук, були використані в матеріалах лекційних курсів: "Системи опрацювання даних", "Аналіз даних", "Основи машинного навчання", "Аналітика даних"

Серед результатів отриманих Павлишенком Б.М. використано, зокрема, методи машинного навчання в аналізі даних, методи прогнозної аналітики, методи семантичного аналізу текстових масивів даних.

Голова комісії: Шувар Р.Я.

Члени комісії: Ненчук Т.М.

Демків Л.С.

Шувара Р. Ненчук Т. Демків Л.

Підпис ПІДТВЕРДЖУЮ
ВЧЕНИЙ СЕКРЕТАР
ЛЬВІВСЬКОГО НАЦІОНАЛЬНОГО
УНІВЕРСИТЕТУ ІМЕНІ ІВАНА ФРАНКА

АКТ
про впровадження результатів дисертаційної роботи
"Методи інтелектуального аналізу консолідованих даних
для підтримки прийняття рішень"
докторанта кафедри системного проектування
Львівського національного університету імені Івана Франка
Павлишенка Богдана Михайловича
в навчальний процес

Комісія у складі в.о. завідувача кафедри математичної статистики і диференціальних рівнянь доктора фіз.-мат. наук доц. Бугрія О.М., професора кафедри математичної статистики і диференціальних рівнянь доктора фіз.-мат. наук проф. Бокала М.М., професора кафедри математичної статистики і диференціальних рівнянь доктора фіз.-мат. наук проф. Єлейка Я.І., професора кафедри математичної статистики і диференціальних рівнянь доктора фіз.-мат. наук проф. Лопушанської Г.П. підтверджує, що результати дисертаційної роботи "Методи інтелектуального аналізу консолідованих даних для підтримки прийняття рішень" докторанта кафедри системного проектування Павлишенка Б.М., поданої на здобуття наукового ступеня доктора технічних наук були використані в матеріалах лекційних курсів: "Байєсівський аналіз даних" (для студентів-магістрів (1-й курс) спеціальності "112-Статистика"), "Комп'ютер в математичному дослідженні" (для студентів-бакалаврів (2-й курс) спеціальності "111-Математика"). Серед результатів отриманих Павлишенком Б.М. використано, зокрема, методи машинного навчання в аналізі даних, методи аналізу текстових масивів даних, методи побудови моделей прогнозної аналітики.

Голова комісії: Бугрій О.М.

Члени комісії: Бокало М.М.

Єлейко Я.І.

Лопушанська Г.П.

Підписи Бугрія О.М., Бокала М.М., Єлейка Я.І., Лопушанської Г.П. підтверджую.

Вчений секретар
Львівського національного університету
імені Івана Франка Грабовецька О.С.

А.10 Список публікацій здобувача за темою дисертації

Список публікацій здобувача, в яких опубліковано основні наукові результати дисертації:

1. Pavlyshenko B. M. Machine-learning models for sales time series forecasting // Data. 2019. Vol. 4, № 1. P. 15. (Входить до міжнародних наукометричних баз Web of Science та Scopus)

2. Pavlyshenko B. Genetic Optimization of Keyword Subsets in the Classification Analysis of Authorship of Texts // Journal of Quantitative Linguistics. 2014. Vol. 21, № 4. P. 341–349. (Входить до міжнародних наукометричних баз Web of Science та Scopus)

3. Pavlyshenko B. Clustering of Authors' Texts of English Fiction in the Vector Space of Semantic Fields // Cybernetics and Information Technologies. 2014. Vol. 14, № 3. P. 25–36. (Входить до міжнародних наукометричних баз Web of Science та Scopus)

4. Pavlyshenko B. Classification analysis of authorship fiction texts in the space of semantic fields // Journal of Quantitative Linguistics. 2013. Vol. 20, № 3. P. 218–226. (Входить до міжнародних наукометричних баз Web of Science та Scopus)

5. Pavlyshenko B. The Distribution of Semantic Fields in Author's Texts // Cybernetics and Information Technologies. 2016. Vol. 16, № 3. P. 195–204. (Входить до міжнародних наукометричних баз Web of Science та Scopus)

6. Павлишенко Б. Квантовий алґоритм еволюційного аналізу одновимірних кліткових автоматів // Журнал фізичних досліджень. 2011. Т. 15, № 3. С. 1–6. (Входить до міжнародної наукометричної бази Scopus)

7. Pavlyshenko B. M. Sales Time Series Analytics Using Deep Q-learning // International Journal of Computing. 2020. Sep. Vol. 19, № 3. P. 434–441. (Входить до міжнародної наукометричної бази Scopus)

8. Павлишенко Б. М. Модель семантичного контексту в алгоритмах інтелектуального аналізу текстів // Комп'ютинг. 2011. Т. 10, № 3. С. 216–222.

9. Павлишенко Б. Семантична кластеризація текстових документів методом k–середніх // Комп'ютерні науки та інформаційні технології. 2011. № 710. С. 215–218.

10. Павлишенко Б. М. Групування тегів користувачів мікроблогів на основі ґратки семантичних концептів // Комп'ютерні системи та мережі. 2011. № 717. С. 120–124.

11. Павлишенко Б. М. Пошук частих множин семантичних ознак та асоціативних правил в повідомленнях мікроблогів // Нові технології. 2011. № 3(33). С. 82–86.

12. Павлишенко Б. М. Моделювання нечітких семантичних полів у масивах текстових документів // Системи обробки інформації. 2011. № 8. С. 175–178.

13. Павлишенко Б. М. Квантовий алгоритм пошуку ключових слів у масивах текстових даних // Біоніка інтелекту. 2011. № 3(77). С. 157–161.

14. Павлишенко Б. Числове моделювання алгоритму Гровера для квантового пошуку даних // Теоретична електротехніка. 2010. № 61. С. 49–59.

15. Павлишенко Б. М. Сингулярна декомпозиція матриці семантичних ознак в алгоритмі ієрархічної кластеризації текстових масивів // Математичні машини і системи. 2012. Т. 1, № 1. С. 69–76.

16. Павлишенко Б. М. Групування текстових даних на основі моделі семантичного контексту // Східно-Європейский журнал передових технологій. 2011. № 5 (2). С. 39–42.

17. Павлишенко Б. М. Модель решітки семантичних концептів для інтелектуального аналізу мікроблогів // Штучний інтелект. 2012. № 1. С. 103–111.

18. Павлишенко Б. М. Часова залежність квантитативних характеристик ключових тегів у RSS каналах // Системи обробки інформації. 2012. № 3 (2). С. 199–202.

19. Павлишенко Б. Ймовірнісна класифікація текстових документів у просторі семантичних полів // Електроніка та інформаційні технології. 2012. № 2. С. 164–172.

20. Павлишенко Б. М. Кластерний аналіз повідомлень груп новин у просторі семантичних ознак // Комп'ютерні системи та мережі. 2012. № 745. С. 148–155.

21. Павлишенко Б. Класифікація повідомлень груп новин у векторному просторі семантичних полів // Комп'ютерні науки та інформаційні технології. 2012. № 744. С. 294–302.

22. Павлишенко Б. М. Аналіз семантичних образів у масивах текстових об'єктів за допомогою квантових обчислень // Математичні машини і системи. 2013. № 1. С. 34–43.

23. Павлишенко Б. М. Формування базису семантичного простору текстових документів за допомогою генетичних алгоритмів // Математичні машини і системи. 2013. № 2. С. 96–104.

24. Павлишенко Б. М. Використання лексемних полів у інтелектуальному аналізі текстових масивів // Штучний інтелект. 2013. № 1. С. 98–109.

25. Павлишенко Б. М. Модель вторинних некорельованих семантичних полів для аналізу текстових даних // Системні дослідження та інформаційні технології. 2014. № 3. С. 130–138.

26. Pavlyshenko B. M. Forecasting of Events by Tweets Data Mining // Electronics and information technologies. 2018. № 10. P. 71–85.

27. Pavlyshenko B. M. Can Twitter Predict Royal Baby's Name? // Electronics and information technologies. 2019. № 11. P. 52–60.

28. Pavlyshenko B. M. Detection of Technical Failures on Production Lines Using Machine Learning, Linear and Bayesian Models of Logistic Regression // Electronics and information technologies. 2019. № 12. P. 3–19.

29. Павлишенко Б. М. Використання методів машинного навчання та семантичних ознак в інтелектуальному аналізі текстових даних // Електроніка та інформаційні технології. 2020. № 13. С. 3–18.

30. Pavlyshenko B. M. Modeling COVID-19 Spread and Its Impact on Stock Market Using Different Types of Data // Electronics and information technologies. 2020. № 14. P. 3–21.

Публікації, які засвідчують апробацію матеріалів дисертації:

31. Павлишенко Б. М. Використання квантових алгоритмів в системах розпізнавання образів // Друга Всеукраїнська науково–практична конференція "Проблеми електроніки та інформаційні технології", 02–05 вересня 2010 р. – Львів–Чинадієво. 2010. С. А11.

32. Павлишенко Б. М. Алгоритми семантичної векторизації та кластеризації текстових масивів // Друга Всеукраїнська науково–практична конференція "Проблеми електроніки та інформаційні технології", 02–05 вересня 2010 р. – Львів–Чинадієво. 2010. С. А12.

33. Павлишенко Б. М. Кластерний аналіз текстових документів в просторі семантичних концептів // Збірник доповідей науково–практичної конференції з міжнародною участю "Системи підтримки прийняття рішень. Теорія і практика", 6 червня 2011 р. – Київ. 2011. С. 146–149.

34. Павлишенко Б. М. Алгоритми семантичного групування текстових документів // III науково–практична конференція "Електроніка та інформаційні технології

(ЕлІТ–2011)": тези доповідей, 01–04 вересня 2011 р. – Львів–Чинадієво. 2011. С. 22–23.

35. Павлишенко Б. М. Модель формального семантичного контексту в алгоритмах обробки текстових документів // III науково–практична конференція "Електроніка та інформаційні технології (ЕлІТ–2011)": тези доповідей, 01–04 вересня 2011 р. – Львів–Чинадієво. 2011. С. 24–27.

36. Павлишенко Б. М. Інтелектуальний аналіз мікроблогів за допомогою решітки семантичних концептів // 5-а міжнародна науково–технічна конференція ACSN–2011 "Сучасні комп'ютерні системи та мережі: розробка та використання": тези доповідей, 29 вересня – 1 жовтня 2011 р. – Львів. 2011. С. 85–87.

37. Павлишенко Б. М. Аналіз формальних семантичних понять в алгоритмах обробки даних // XVII Всеукраїнська наукова конференція "Сучасні проблеми прикладної математики та інформатики": тези доповідей, 6–7 жовтня 2011 р. – Львів. 2011. С. 80.

38. Павлишенко Б. М. Векторна модель текстових документів у семантичному ортонормованому базисі // XVIII Всеукраїнська наукова конференція "Сучасні проблеми прикладної математики та інформатики": тези доповідей, 4–5 жовтня 2012 р. – Львів. 2012. С. 127.

39. Павлишенко Б. М. Модель нечітких семантичних полів для інтелектуального аналізу текстових масивів // IV науково–практична конференція "Електроніка та інформаційні технології (ЕлІТ–2012)": тези доповідей, 30 серпня – 2 вересня 2012 р. – Львів–Чинадієво. 2012. С. 98.

40. Павлишенко Б. М. Аналіз семантичних асоціацій у веб–блогах за допомогою гратки формальних понять // Міжнародна науково–технічна конференція "Штучний інтелект. Інтелектуальні системи" (ШІ–2012): матеріали конференції, 1–5 жовтня, 2012 р. – Кацивелі, АР Крим. 2012. С. 118–122.

41. Павлишенко Б. М. Аналіз мікроблогів користувачів на основі гратки семантичних концептів // Збірник доповідей науково–практичної конференції з міжнародню участю "Системи підтримки прийняття рішень. Теорія і практика", 6 червня 2012 р. – Київ. 2012. С. 115–118.

42. Павлишенко Б. М. Прогнозування подій на основі інтелектуального аналізу повідомлень мікроблогів Twitter // XIII міжнародна наукова конференція імені Т. А. Таран "Інтелектуальний аналіз інформації" (ІАІ–2013): збірка праць, 15–17 травня 2013 р. – КПИ, Київ. 2013. С. 199–205.

43. Павлишенко Б. М. Чи може Твіттер передбачити ім'я британського принца? // XIX Всеукраїнська наукова конференція "Сучасні проблеми прикладної математики та інформатики": тези доповідей, 3–4 жовтня 2013 р. – Львів. 2013. С. 108.

44. Павлишенко Б. М. Використання інтелектуального аналізу повідомлень Twitter у прогнозуванні фінансових ринків // Матеріали 2–ї Міжнародної конференції "Інформація, комунікація, суспільство 2013" (ІКС–2013), 16–19 травня, 2013 р. – Львів–Славське. 2013. С. 86–87.

45. Павлишенко Б. М. Аналіз курсу акцій на основі твітів інформагенств // V науково–практична конференція "Електроніка та інформаційні технології" (ЕлІТ–2013): тези доповідей, 29 серпня–1 вересня 2013 р. – Львів–Чинадієво. 2013. С. 60.

46. Pavlyshenko B. M. Linear, machine learning and probabilistic approaches for time series analysis // Data Stream Mining & Processing (DSMP), IEEE First International Conference. 2016. P. 377–381. (Входить до міжнародної наукометричної бази Scopus)

47. Pavlyshenko B. Machine learning, linear and Bayesian models for logistic regression in failure detection problems // Big Data (Big Data), 2016 IEEE International Conference on, IEEE, Washington D.C. 2016. P. 2046–2050. (Входить до міжнародної наукометричної бази Scopus)

48. Pavlyshenko B. Using Stacking Approaches for Machine Learning Models // 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP). 2018. P. 255–258. (Входить до міжнародної наукометричної бази Scopus)

49. Pavlyshenko B. Predictive Analytics for Sales Time Series // Xth International Scientific and Practical Conference "Electronics and Information Technologies" (ELIT-2018) August 30 - September 2, 2018, Lviv, Karpaty village, Issue 10. 2018. P. 85–87.

50. Pavlyshenko B. M. Regression Approaches For Sales Time Series Forecasting // Матеріали XXIV Всеукраїнської наукової конференції "Сучасні проблеми прикладної математики та інформатики", APAMCS-2018 26-28 вересня 2018 року, Львів. 2018. С. 121–123.

51. Pavlyshenko B. Bitcoin Price Predictive Modeling Using Expert Correction // 2019 XIth International Scientific and Practical Conference on Electronics and Information Technologies (ELIT), September 16 – 18, 2019 Lviv, Ukraine. 2019. P. 163–167. (Входить до міжнародної наукометричної бази Scopus)

52. Pavlyshenko B. Using Bayesian Regression for Stacking Time Series Predictive Models // 2020 IEEE Third International Conference on Data Stream Mining & Processing (DSMP). 2020. P. 305–309. (Входить до міжнародної наукометричної бази Scopus)