Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

ПАВЛИШЕНКО БОГДАН МИХАЙЛОВИЧ

Підпис

УДК 004.89:519.765

МЕТОДИ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ КОНСОЛІДОВАНИХ ДАНИХ ДЛЯ ПІДТРИМКИ ПРИЙНЯТТЯ РІШЕНЬ

05.13.23 - системи та засоби штучного інтелекту

АВТОРЕФЕРАТ
дисертації на здобуття наукового ступеня
доктора технічних наук

Харків – 2021

Дисертацією є рукопис.

Роботу виконано у Львівському національному університеті імені Івана Франка Міністерства освіти і науки України.

Науковий консультант – доктор фізико-математичних наук, професор
Дияк Іван Іванович,
Львівський національний університет імені Івана Франка, декан факультету прикладної математики та інформатики.

Офіційні опоненти: доктор технічних наук, професор
Литвиненко Володимир Іванович,
Херсонський національний технічний університет, завідувач кафедри інформатики та комп'ютерних наук;

доктор технічних наук, професор
Пелешко Дмитро Дмитрович,
ДВНЗ "Університет банківської справи", професор кафедри кібербезпеки;

доктор технічних наук, професор
Бідюк Петро Іванович,
Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського", професор кафедри математичних методів системного аналізу Інституту прикладного системного аналізу.

Захист відбудеться 14 квітня 2021 року о 13:00 годині на засіданні спеціалізованої вченої ради Д64.052.01 у Харківському національному університеті радіоелектроніки за адресою: 61166, місто Харків, пр. Науки, 14.

З дисертацією можна ознайомитися в бібліотеці Харківського національного університету радіоелектроніки за адресою: 61166, місто Харків, пр. Науки, 14.

Автореферат розіслано 4 березня 2021 року.

Учений секретар
спеціалізованої вченої ради *Підпис* Є. І. Литвинова

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

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 лінвістичної 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 невизначенність 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 SoftServe 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.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У **вступі** обґрунтовано актуальність роботи, поставлено мету та завдання дисертаційної роботи, описано методи дослідження, наукову новизну, практичне значення отриманих результатів та апробацію роботи.

У **першому розділі** наведено літературний огляд основних моделей, методів та підходів, які використовуються в інтелектуальному аналізі даних. Складовими інтелектуального аналізу є: прогнозна аналітика, методи та алгоритми машинного навчання, зокрема, класифікація, регресія, кластеризація, глибоке навчання, теорія частих множин та асоціативних правил та інші розділи сучасних інформаційних технологій. В інтелектуальному аналізі даних часто використовують алгоритмічні прогнозні моделі. Алгоритмічну модель можна розглядати як деяку 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 у 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. Проаналізовано методи

та підходи в аналітиці даних табличного типу, зокрема, розглянуто лінійні моделі, ймовірнісні моделі на основі байєсівської регресії 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 дозволяють отримати щільність розподілу ймовірностей для цільової змінної, що є важливим в аналітиці ризиків та оцінці невизначеності прогнозування. Методами машинного та глибокого навчання можна виявити складні закономірності між ознаками, однак їх можна ефективно застосовувати лише у випадку стаціонарного розподілу ознак. Розглянуто методи глибокого Q-навчання. Різні методи у прогнозній аналітиці мають свої позитивні сторони та недоліки, тому доцільним є формування таких аналітичних методів та підходів, у яких можна було б об'єднати різні моделі і, як наслідок, отримати точні та надійні результати інтелектуального аналізу. Розглянуто основні методи аналізу текстових масивів. Наведено основні положення лексемної семантики. Розглянуто семантичне структурування лексемного словника на основі семантичної системи WordNet. Проаналізовано лінгвістичні концепції семантичних лексикографічних полів із точки зору їхнього використання в алгоритмах інтелектуального аналізу текстових масивів. Семантичні поля глибоко вивчені у лінгвістичних працях, однак існує необхідність розробки формалізованих математичних моделей для їхнього впровадження в алгоритми інтелектуального аналізу текстових масивів.

Проведений аналіз виявив необхідність подальшого дослідження комбінування різних моделей, зокрема, за допомогою ансамблів моделей для того, щоб ефективно використовувати 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. На основі наведених літературних даних зроблено висновки, сформульовано актуальні завдання. Обґрунтовано, що для реалізації мети дисертаційного дослідження необхідно розробити нові та удосконалити наявні методи інтелектуального аналізу структурованих даних табличного типу та даних текстового типу на основі розгляду типових задач із різних предметних областей. Для даних табличного типу необхідно розробити методи стекінгового об'єднання різнотипних моделей у прогнозні ансамблі і методи поєднання параметричних, ймовірнісних та машино-начальних моделей. Для інтелектуального аналізу даних текстового типу необхідно розробити методи формування семантичних ознак на основі концепції семантичних полів та використання цих ознак в аналізі даних на основі машинного навчання, методи застосування семантичних аналітичних ознак на основі теорії частих множин і асоціативних правил і методи використання теорії формальних концептів в аналітиці текстових потоків соціальних мереж.

У **другому розділі** розглянуто методи виявлення та формування ознак та інтелектуальний аналіз даних табличного типу. Консолідовані дані з різною структурою та з різних джерел можна представити 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. Для експериментального аналізу розглянуто історичні дані часових рядів, які описують динаміку продажів у роздрібній мережі [1]. Проаналізовано підходи на основі лінійних, байєсівських та машинно-навчальних моделей [46,49,50]. Використання регресійних підходів до 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. Лінійні моделі дають можливість аналізувати вплив зовнішніх факторів, однак не дають змогу виявляти складні патерни міжфакторної взаємодії. Для врахування такої взаємодії проведено дослідження використання методів машинного навчання із учителем. Одними із найбільш ефективних методів є такі, які базуються на деревах рішень, оскільки вони нечутливі до монотонних перетворень кількісних значень аналітичних ознак, що є актуальним за наявності зовнішніх факторів різної природи. Обмеженням методів машинного навчання є те, що їхнє застосування ефективне лише для аналізу стаціонарних процесів, припускаючи, що факторні аналітичні ознаки майбутніх даних мають такий самий розподіл, як і дані для тренування прогнозної моделі. Розглянуто підхід на основі алгоритму машинного навчання Random Forest із використанням історичних даних часових рядів продажів. Як незалежні регресійні змінні використано категоріальні ознаки – наявність промо акції, день тижня, день місяця, місяць. Для категоріальних ознак застосовано one-hot кодування, коли одна категоріальна змінна замінюється на n бінарних змінних, де n – кількість унікальних значень категоріальних змінних. На рис. 1 показано ковзне середнє залишків, а на рис. 2 – ковзне стандартне відхилення залишків прогнозу. Вертикальна лінія на рисунках розділяє часові проміжки для тренувальної та валідаційної вибірок. В отриманих прогнозних результатах спостерігається зміщення цільової змінної на валідаційній вибірці, яке часто виникає при застосуванні методів машинного навчання до нестаціонарних часових рядів. Корекцію такого зміщення можна здійснити, використовуючи лінійну регресію на валідаційній вибірці. Досліджено ефект генералізації у методах машинного навчання, який полягає у тому, що алгоритм регресії знаходить патерни, властиві для цілої вибірки даних. Якщо, наприклад, продажі мають виражені закономірності, то генералізація дозволяє отримати точніші результати, на які не впливають випадкові відхилення значень ознак. 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.

Розглянуто об’єднання моделей різних типів у прогнозний ансамбль на основі стекінгового підходу шляхом використання моделей другого рівня, які використовують прогнозні результати моделей першого рівня. При такому підході прогнозні значення різних моделей, які було отримано на валідаційній вибірці, розглянуто як прогнозні ознаки для моделей другого рівня.

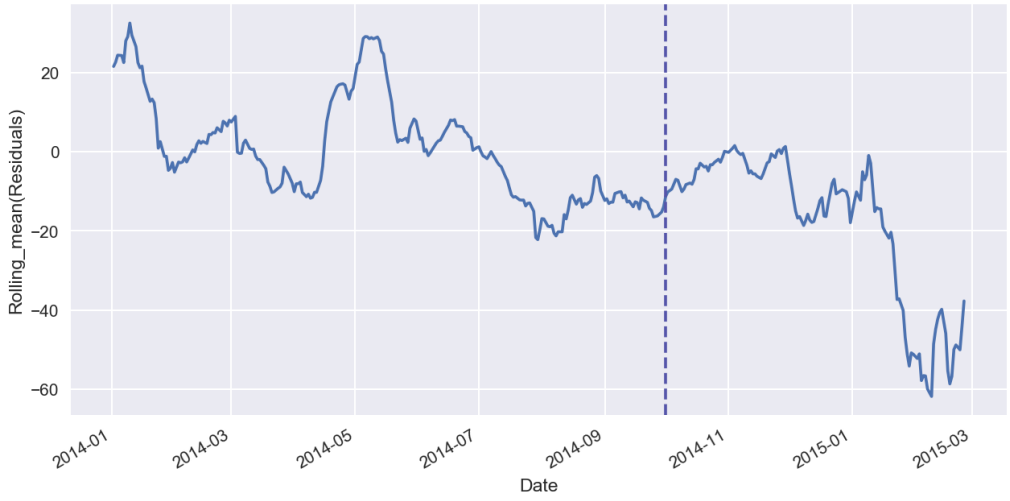

Рисунок 1 – Ковзне середнє для залишків прогнозування

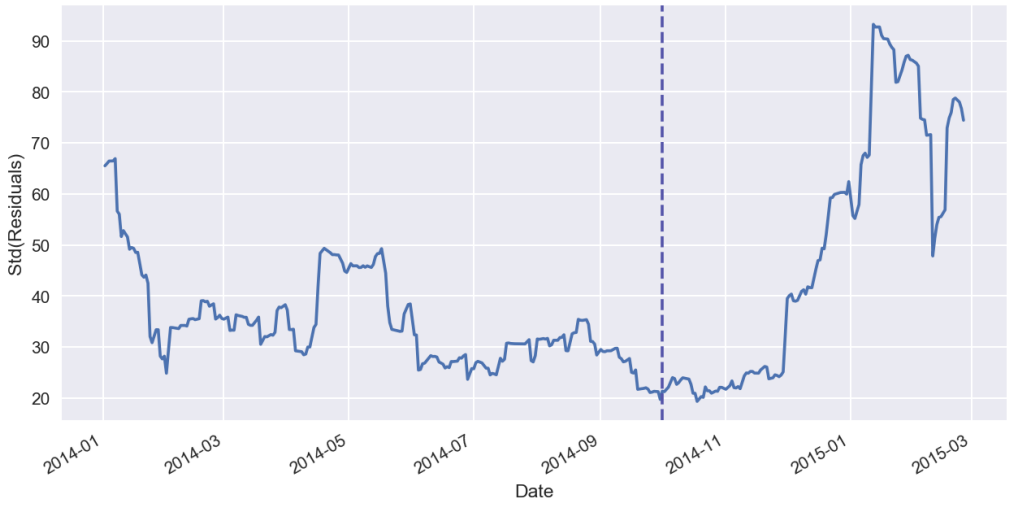

Рисунок 2 – Ковзне стандартне відхилення для залишків прогнозування

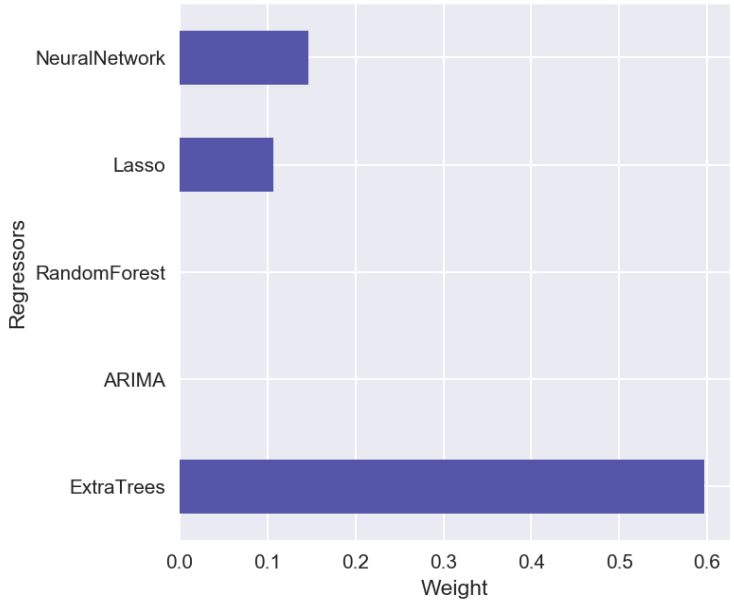

Рисунок 3 – Стекінгові коефіцієнти для прогнозних моделей

Таблиця 1 – Похибки прогнозування різних моделей стекінгового ансамблю

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%

У роботі проведено експериментальне дослідження стекінгу моделей з використанням LASSO регресії як стекінгової моделі другого рівня. На рис. 3 показано регресійні коефіцієнти для прогнозних моделей, отримані на другому стекінговому рівні ансамблю моделей [1]. Лише три моделі першого рівня (ExtraTree,LASSO, Neural Network) мають ненульові коефіцієнти у регресійній моделі. Для інших випадків історичних даних про продажі результати можуть бути іншими, коли інші моделі можуть відігравати також важливу роль у прогнозуванні. Отримані результати показують вищу точність прогнозування стекінгового ансамблю у порівнянні з точністю складових моделей [1]. Збільшення точності стекінгового ансамблю зумовлене тим, що різні моделі використовують різні алгоритми та набори прогнозних ознак, що зумовлює низьку кореляцію прогнозних похибок, які на стекінговому рівні взаємно компенсуються при комбінуванні прогнозних значень різних моделей. У таблиці 1 показано відносні абсолютні похибки у процентах на валідаційній вибірці та на вибірці даних, які не входять у тренувальний та валідаційний сети (out-of-sample sets). Ці результати показують, що використання стекінгу дозволяє врахувати відмінності у результатах для різних моделей з різними наборами параметрів та ознак і підвищити точність прогнозування на валідаційних даних, а також на позавибіркових даних. Крім того, такий ансамбль моделей є стабільнішим, ніж використання одиничних моделей, оскільки різні моделі можуть використовувати різні набори прогнозних ознак, а такі ознаки можуть з часом зменшувати свій прогнозний потенціал, який при стекінговому підході може компенсуватися за рахунок інших моделей. Використання регресії LASSO на стекінговому рівні дає можливість оптимізувати і зменшити набір складових моделей стекінгового ансамблю у деяких випадках на 30%. На другому рівні багаторівневого ансамблю можна застосувати одночасно різні алгоритми стекінгової регресії. Результати прогнозування різних стекінгових моделей можуть бути об’єднані, наприклад, за допомогою зваженої суми на наступному рівні стекінгового ансамблю моделей. На рис. 4 наведено приклад багаторівневої стекінгової моделі.

У роботі проведено дослідження застосування ймовірнісних методів регресії, яка дає можливість оцінити невизначеність складових факторів аналізу та оцінити прогнозні ризики [52]. Такий підхід дає можливість врахувати екстремальні значення при використанні негаусових розподілів із ”товстими хвостами” для цільової змінної. Використання експертних апріорних розподілів параметрів моделі є ефективним у випадках малої кількості історичних даних і дає

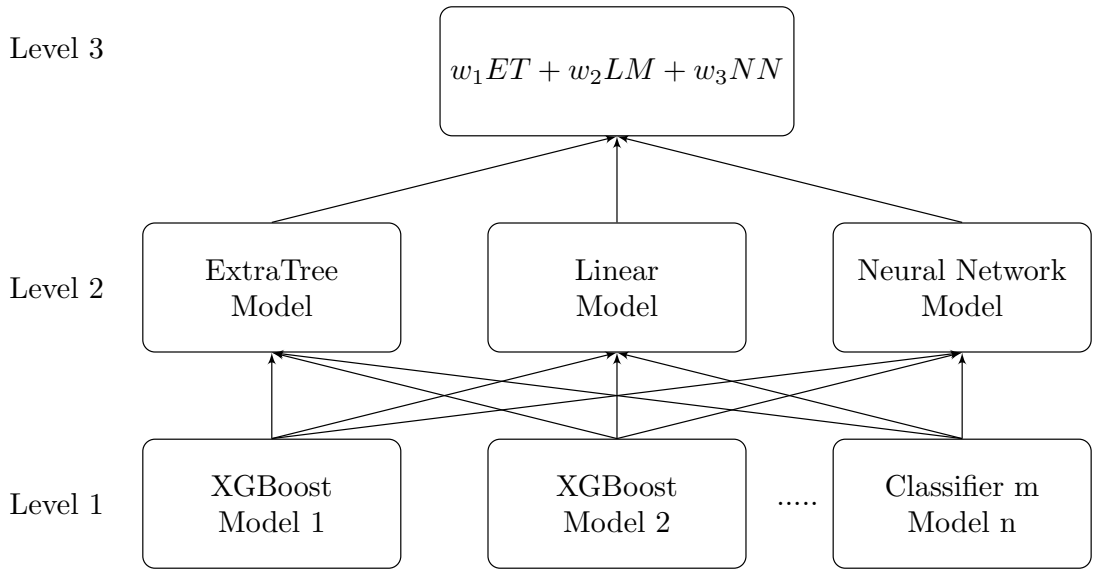

Рисунок 4 – Багаторівневий стекінговий ансамбль прогнозних моделей

можливість отримати прогнозування цільової змінної як результат компромісу між історичними даними та експертними висновками. Імовірнісний підхід на основі байєсівського висновування (Bayesian inference) у прогнозній аналітиці дає можливість отримувати щільність розподілу ймовірностей для цільової змінної. Маючи таку функцію, можна зробити оцінку різних ризиків, зокрема, обчислити кількісну величину ризику VaR (Value at Risk), яка дорівнює 5%-му перцентилю. Байєсівське висновування дозволяє реалізувати нелінійну регресію. Наприклад, у випадку трендів часових рядів із насиченням можна використати модель логістичної кривої і знайти її параметри. Для аналізу байєсівських моделей використовуються чисельні методи Монте-Карло. Для байєсівської регресії використано платформу Stan для статистичного моделювання. Розглянуто випадок нелінійної регресії для часового ряду із трендом, який виходить на насичення. Для моделювання розглянуто часові ряди продажів, які аналізувалися раніше. Таку модель можна описати так:

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

Досліджено використання байєсівської регресії для аналізу ієрархічних моделей, коли деякі параметри моделі є загальними для всіх даних, а інші – різними для різних груп вибірки даних. Наприклад, параметри сезонності можуть бути загальними для всіх товарів у аналітиці продажів, а вільний член у регресійній моделі може бути різним для різних магазинів. Для числового аналізу розглянуто випадок із п'ятьма різними магазинами. Тренди, промо-ефект та сезонність розглянуто однаковими для всіх магазинів, а вплив заданого магазину на продажі описується вільним членом регресії, тому цей параметр буде різним для різних магазинів. Показано,

що 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. Запропоновано метод використання байєсівської регресії для стекінгу прогнозних моделей. Імовірнісний підхід для стекінгу прогнозних моделей дозволяє зробити оцінку ризиків для прогнозів, що є важливим у процесі прийняття рішень. Стекінгову модель на основі байєсівскої регресії розглянуто у вигляді

$$y \sim Student_t(\nu, \mu, \sigma), \mu = \alpha + \sum_i \beta_i x_i, \tag{2}$$

де x_i – прогнозні результати складових моделей ансамблю на валідаційному сеті Розглянуто дворівневий ансамбль прогнозних моделей для часових рядів. Для прогнозування на першому рівні ансамблю моделей було використано моделі ARIMA, Neural Network, Random Forest, Extra Tree. На рис. 5 наведено результати прогнозування різних моделей на валідаційному сеті. На другому стекінговому рівні ансамблевої моделі було здійснено байєсівську регресію результатів прогнозування моделей на валідаційному сеті. Цей підхід дає можливість отримати щільність розподілу ймовірностей 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 із предметної області, у якій проводиться прогнозна аналітика. На рис. 6 зображено коробкові графіки розподілів значень коефіцієнтів складових моделей стекінгового ансамблю. Імовірнісний підхід для стекінгу прогнозних моделей дозволяє зробити оцінку ризиків для прогнозів, що є важливим у процесі прийняття рішень. Використання байєсівського висновування для стекінгової регресії може бути корисним у випадках невеликих сетів даних та допоможе експертам вибрати набір моделей для стекінгу, а також оцінити ризики різного типу 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.

Розглянуто лінійну модель для ціни біткоїна з регресійними ознаками, які базуються на статистиці біткоїна, характеристиках процесів видобутку біткоїна, трендах пошукових запитів Google, візитах на сторінки Вікіпедії, а також змінній, яка описує експертну корекцію [51]. 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. Розглянуто вплив пандемії COVID-19 на зміни на фондовому ринку із використанням баєсівської регресійної моделі, що дозволяє отримати щільність розподілу ймовірностей для параметрів моделі та зробити оцінку невизначеності факторів, які мають вплив на цільову змінну [30]. Як регресійні ознаки використано z-оцінки часових рядів кількості відвідувань сторінок Вікіпедії. Як цільову змінну використано z-оцінки індексу S&P-500. На рис. 7 показано розраховані середні значення та 0.01, 0.99 квантилі розподілу

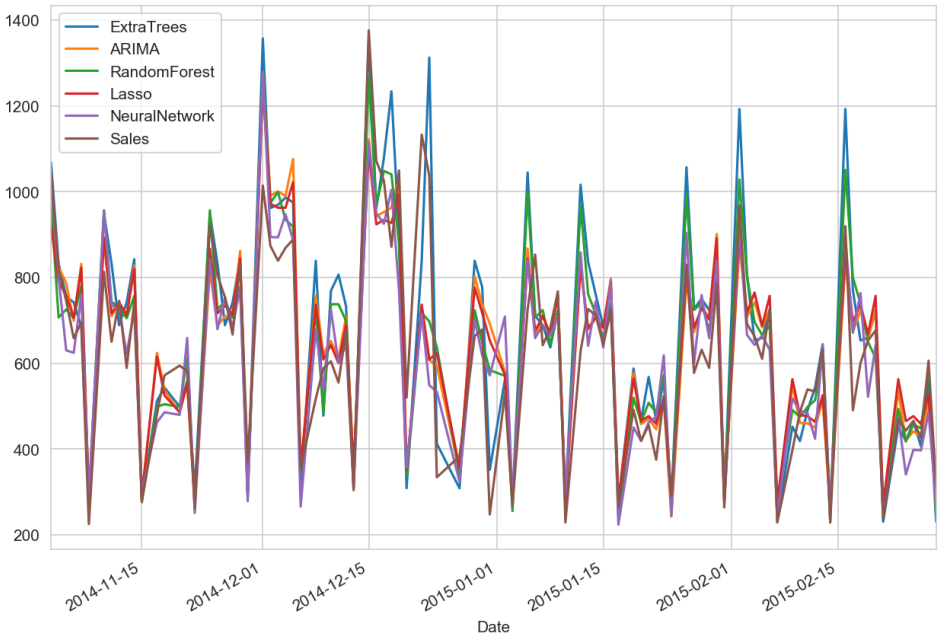

Рисунок 5 – Результати прогнозування різних моделей на валідаційному сеті

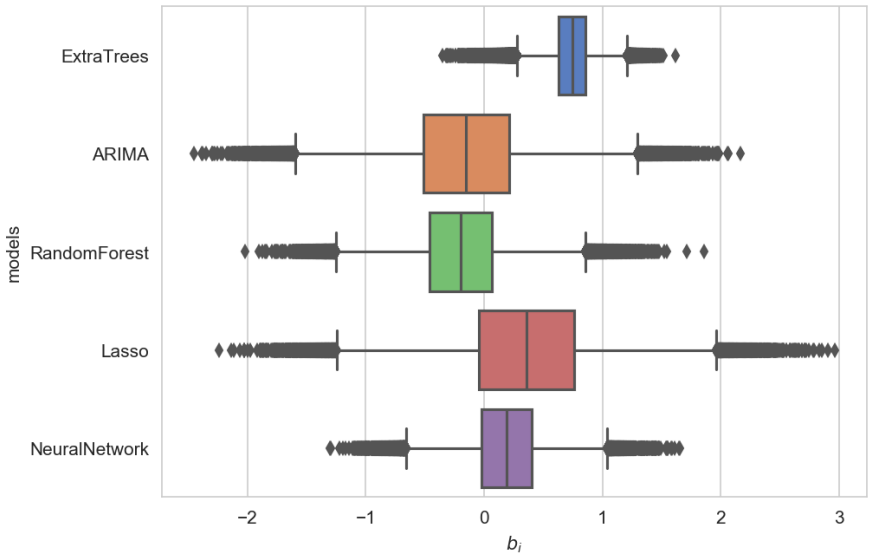

Рисунок 6 – Коробкові графіки розподілів значень коефіцієнтів складових моделей стекінгового ансамблю

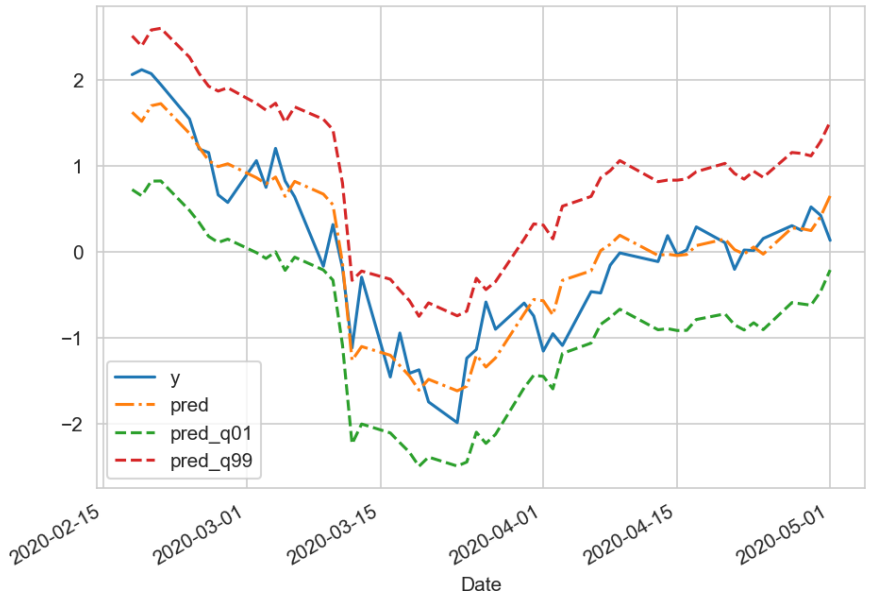

Рисунок 7 – Середні значення (*pred*) та 0.01 (*pred_q01*), 0.99 (*pred_q99*) квантилі розподілу прогнозів для індексу S&P-500 в період пандемії COVID-19

прогнозів для індексу S&P-500 під час пандемії COVID-19. На рис. 8 показано коробкові графіки розподілів значень коефіцієнтів лінійної регресії для аналітичних ознак.

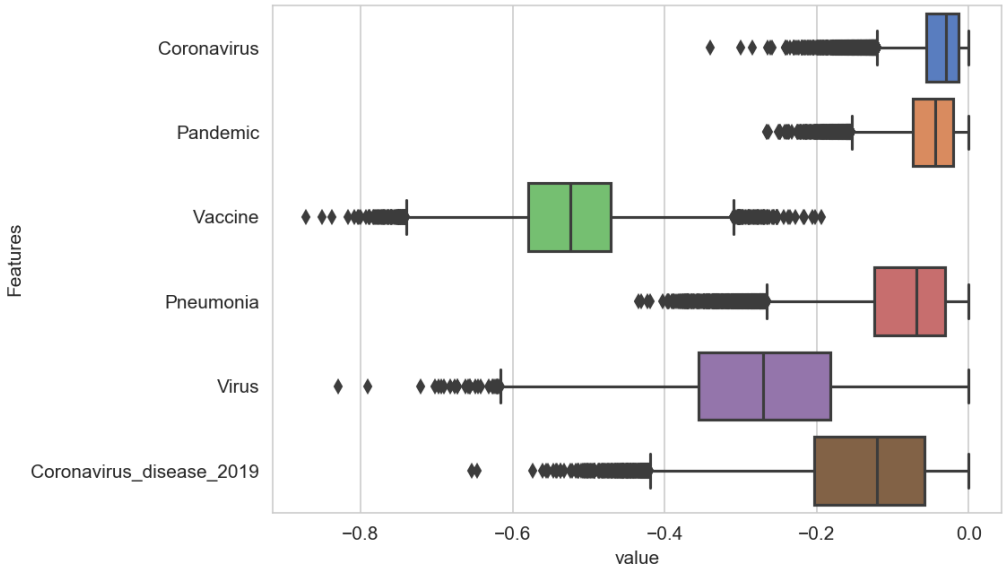

Рисунок 8 – Коробкові графіки розподілів значень коефіцієнтів лінійної регресії

Розглянуто використання лінійних, ймовірнісних та машинно-навчальних моделей для логістичної регресії на прикладі проблеми виявлення відмов на виробничих лініях [28, 47]. Проаналізовано стекінговий підхід на основі узагальненої лінійної моделі для логістичної регресії та на основі байєсівського висновування. Машинне навчання дозволяє виявити складні патерни виникнення відмов у виробничих процесах. Узагальнена лінійна модель для логістичної регресії дає можливість досліджувати фактори впливу на виникнення відмов. Використовуючи байєсівську модель, можна отримати статистичні розподіли для параметрів моделі, які можна використати при аналізі ризиків. Використання байєсівської моделі на другому рівні модельного ансамблю з незалежними змінними, які відображають ймовірності відмов, отриманих на основі прогнозних моделей першого рівня, дозволяє проаналізувати невизначеності моделей першого рівня та оцінити ризики, що виникають при похибках прогнозування модельного ансамблю.

Розглянуто використання моделей глибокого Q-навчання у задачах часових рядів продажів. Q-навчання належить до навчання з підкріпленням [7]. На відміну від машинного навчання з учителем, яке можна розглядати як форму пасивного навчання з використанням історичних даних, Q-навчання – це варіант активного навчання з використанням інтелектуального агента, який взаємодіє із середовищем. Мета такого навчання полягає у виявленні оптимальної послідовності взаємодій інтелектуального агента із середовищем з метою досягнення максимальної винагороди. Розглянуто безмодельний підхід Q-навчання для аналізу задачі оптимальних стратегій ціноутворення та задачі попиту та постачання.

Q-функцію можна визначити за допомогою ітераційного процесу, який базується на рівнянні Беллмана

$$Q_{i+1}(s,a) = \mathbb{E}\left[r + \gamma \max_{a'} Q_i(s',a') \mid s,a\right] \quad (3)$$

Значення Q-функції можуть бути представлені у вигляді Q-таблиці, у якій рядки описують стани інтелектуального навчального агента, а стовпці – дії агента у середовищі. Стани навчального агента визначаються кількісними характеристиками продажів та очікуваними промо акціями. У випадку великої кількості станів Q-таблиця апроксимується за допомогою нейронної мережі. Використовуються запам'ятовування останніх станів інтелектуального агента та перетреновування нейронної мережі на мініпакетній вибірці після кожної взаємодії агента та середовища. Параметри нейронної Q-мережі можна знайти за допомогою градієнтних методів мінімізації функції втрат:

$$L_i(\theta_i) = \mathbb{E}_{s,a\sim\rho}\left[(y_i - Q(s,a;\theta_i))^2\right] \quad (4)$$

$$y_i = \mathbb{E}_{s'\sim\varepsilon}\left[r + \gamma \max_{a'} Q(s',a',\theta_{i-1}) \mid s,a\right] \quad (5)$$

де ρ – характеристика поведінки інтелектуального агента, θ – вагові параметри Q-мережі. На рис. 9 наведено зміну значень функції винагороди на часових епізодах.

Рисунок 9 – Динаміка винагород на епізодах навчання інтелектуального агента

Як випливає із отриманих результатів, побудований інтелектуальний агент може навчатися та виявляти закономірності у даних продажів та виробляти оптимальну стратегію взаємодії із середовищем. 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, 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 можна використати для холодного старту навчання інтелектуального агента. 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 адаптовуватися у процесі його використання до змін у бізнес-процесах.

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. Під парадигмою можна розуміти, наприклад, спектр семантичних або тематичних понять, які відображені у структурі лексикографічних значень лексем. На основі концепцій семантичних полів створено теоретико-множинну модель, яка об'єднує поняття семантичного та тематичного лексемних полів [3, 4, 5, 24, 25, 29]. Лексикографічні поля утворено на основі експертного семантичного групування лексемного складу словника. Тематичні поля утворені на основі лексем, характерних для тематично категоризованих текстових документів і визначено 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. Базис лексикографічних семантичних полів є незалежним від вибірки, а базис тематичних полів є індивідуальним для кожної текстової вибірки. Розглянуто векторну модель текстових документів у семантичному просторі, базис якого утворено частотно-дистрибутивними характеристиками семантичних та тематичних полів [24]. Лексемний склад семантичного поля s_k визначено на основі експертного лексикографічного аналізу як

$$W_k^s = \left\{ w_i \mid w \xrightarrow{U_{ws}} s_k, i = 1, 2, ..., N_w \right\}. \tag{6}$$

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

$$U_{sd} : s_k \to p_{kj}^{sd}, k = 1, 2, ...N_s, j = 1, 2...N_d. \tag{7}$$

Величина p_{kj}^{sd} визначає частоту семантичного поля s_k у текстовому документі d_j

$$p_{kj}^{sd} = \sum_{i=1}^{N_w} p_{ij}^{wd} f_s(w_i, s_k), \quad f_s(w_i, s_k) = \begin{cases} 1, & w_i \in W_k^s \\ 0, & w_i \notin W_k^s \end{cases}, \tag{8}$$

де p_{ij}^{wd} - текстова частота лексеми w_i у документі d_j. Сукупність значень p_{kj}^{sd} утворюють матрицю типу ознака-документ, у якій ознаками виступають частоти семантичних полів у документах:

$$M_{sd} = \left(p_{kj}^{sd}\right)_{k=1,j=1}^{N_s,N_d}. \tag{9}$$

Вектор

$$V_j^s = \left(p_{1j}^{sd}, p_{2j}^{sd}, ..., p_{N_s j}^{sd}\right) \tag{10}$$

відображає документ d_j в N_s-мірному просторі текстових документів. Використання концепції семантичних полів є ефективним у векторній моделі

текстових документів внаслідок зменшення розмірності фазового простору представлення документів. Це дає можливість зменшити кількість необхідних обчислень в алгоритмах аналізу текстових даних. Для чисельного аналізу розглянуто текстові вибірки різних типів, зокрема, масив авторських текстів англомовної художньої літератури та повідомлення груп новин. В аналізі використано лексемний склад семантичних полів іменників та дієслів семантичної системи WordNet. На рис. 10 показано щільність розподілу ймовірностей частот семантичного поля *verb.communication* для різних класів документів у текстовій вибірці художніх творів англомовної прози. На рис. 11 показано щільність розподілу ймовірностей частот заданого тематичного поля для вибірки груп новин. На рис. 12 показано розподіл тематичних полів у двовимірному просторі головних компонент (PCA), а на рис. 13 у двовимірному t-SNE просторі для вибірки повідомлень груп новин. Як випливає з отриманих даних, семантичні та тематичні поля мають значний класифікаційний потенціал. Це дає можливість використовувати кількісні ознаки на основі тематичних полів у задачах інтелектуального аналізу текстових даних, зокрема у класифікації текстових документів [4, 29]. Деякі семантичні та тематичні поля мають високий розділювальний потенціал для диференціювання авторського стилю. На основі теорії нечітких множин створено модель нечіткого семантичного поля лексемного складу текстових масивів [12]. Визначено характеристики для нечіткого семантичного поля – функцію приналежності, міру нечіткості семантичного поля, семантичне поле α-рівня. 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, а також на основі сингулярного розкладу матриць типу *частоти_семантичних_полів-документи* [25]. Розмірність простору вторинних семантичних полів є суттєво меншою за розмірність простору первинних семантичних полів унаслідок заміни взаємопов'язаних складових некорельованими семантичними характеристиками. 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 [29]. Ці компоненти можуть бути використані як додаткові семантичні ознаки текстових документів у задачах інтелектуального аналізу текстових даних.

У **четвертому розділі** розглянуто методи машинного навчання в аналітиці текстових даних із використанням концепції семантичних полів. Для кластерного та класифікаційного аналізу методами машинного навчання розглянуто текстові вибірки різних типів, зокрема, масив авторських текстів англомовної художньої літератури, повідомлення груп новин та короткі повідомлення Твіттера.

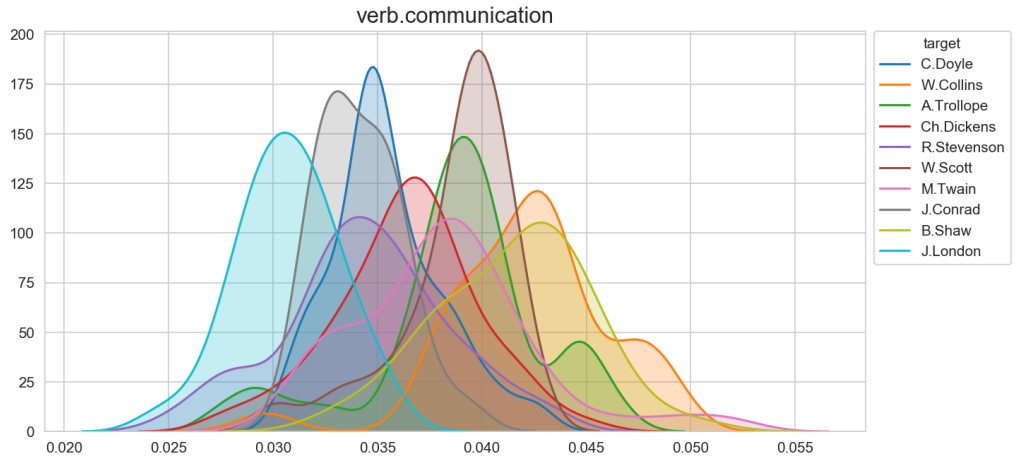

Рисунок 10 – Щільність розподілу ймовірностей частот семантичного поля *verb.communication* для різних класів документів у текстовій вибірці художніх творів англомовної прози

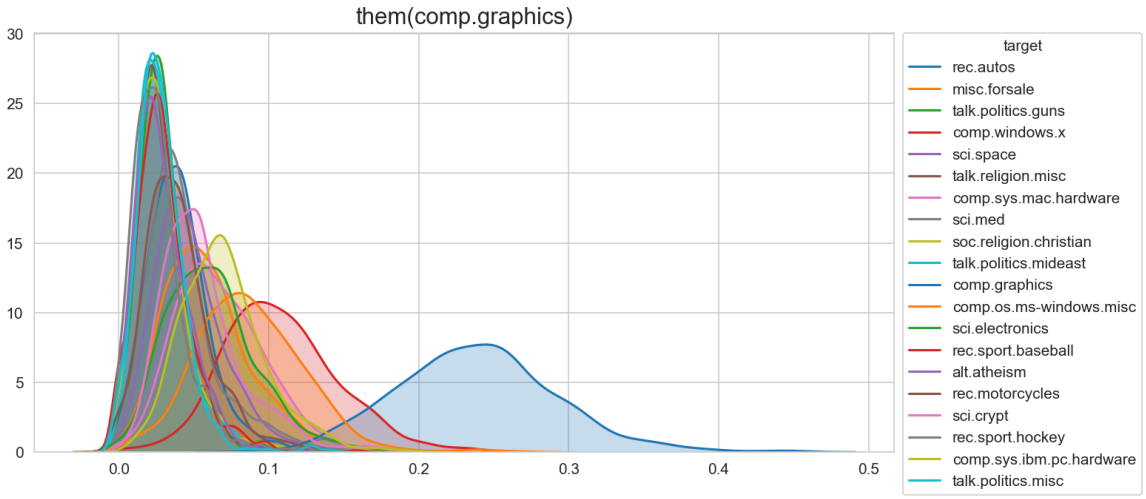

Рисунок 11 – Щільність розподілу ймовірностей частот тематичного поля *them(comp_graphics)* для різних класів документів у вибірці повідомлень груп новин

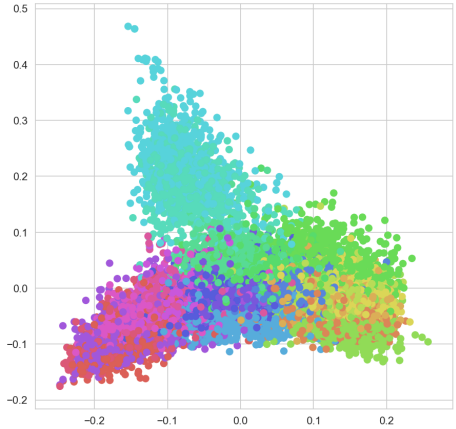

Рисунок 12 – Розподіл тематичних полів у двовимірному PCA просторі у вибірці повідомлень груп новин

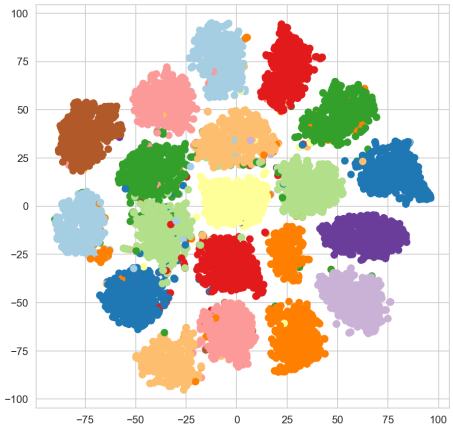

Рисунок 13 – Розподіл тематичних полів у двовимірному t-SNE просторі у вибірці повідомлень груп новин

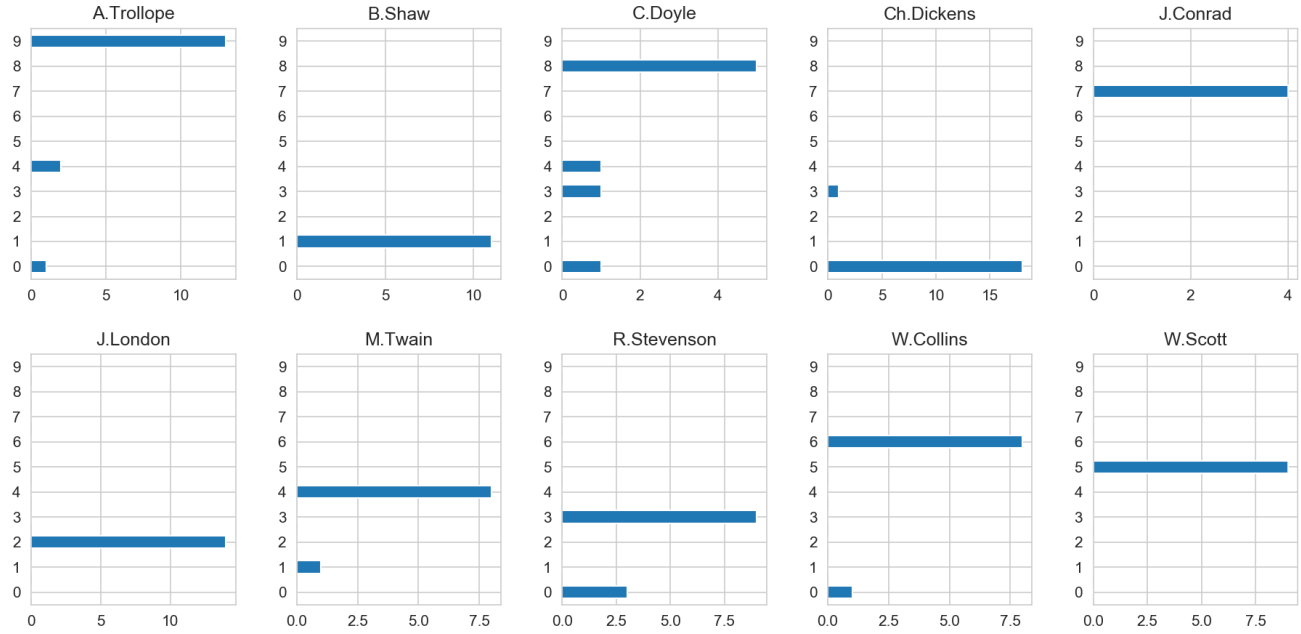

Рисунок 14 – Розподіл авторських документів за кластерами в алгоритмі агломеративної кластеризації у просторі тематичних полів

Запропонований метод кластеризації текстових документів у просторі семантичних та тематичних полів дає можливість отримувати новий структурний поділ документів за семантичними ознаками у просторі суттєво меншої розмірності, ніж простір, утворений лексемним складом текстової вибірки [3, 9, 15, 20]. Такий структурний поділ відображає групування документів за новими ознаками, 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 та тематичних полів. Для кластеризації використано алгоритми агломеративної кластеризації та алгоритм k-means. Як семантичні ознаки розглянуто частотні характеристики семантичних та тематичних полів, компоненти сингулярного розкладу TF-IDF матриці, а також компоненти тематик латентного розміщення Діріхле. 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. Приналежність 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. На рис. 14 показано розподіл авторських документів за кластерами в алгоритмі агломеративної кластеризації 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. Також проведено кластерний аналіз текстової вибірки повідомлень груп новин у семантичному просторі. 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 відображено 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 з кластерною структурою у просторі семантичних полів. Запропоновано використання комбінації різних семантичних ознак, зокрема, семантичних та тематичних полів, компонент сингулярного розкладу TF-IDF матриці та компонент латентного розміщення Діріхле. Такий підхід дає можливість отримати вищу точність у задачах класифікації текстових документів. Застосування 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. Експериментальний класифікаційний аналіз тестової вибірки текстових документів у векторному просторі семантичних та тематичних полів показав високу ефективність використання лексемних полів у класифікаційному аналізі [4, 19, 21, 29]. Використання ознак на основі експертно сформованих семантичних полів дає можливість створювати стабільні прогнозні моделі з урахуванням можливої зміни розподілу лексем тестових вибірок. Висока точність класифікації авторських текстів у векторному просторі семантичних полів свідчить про наявність у цьому просторі відокремлених областей авторського ідіолекта, які характеризують індивідуальний стиль авторів. Сингулярний розклад матриці семантичних ознак типу *"частоти_семантичних_полів–документи"* дає можливість аналізувати текстові документи у новому просторі семантичних концептів, суттєво зменшити розмірність задач класифікації текстових документів та оптимізувати задачі інтелектуального аналізу даних текстового типу [15]. Використання ознак на основі текстових частот семантичних полів дає можливість суттєво зменшити розмірність семантичного простору в порівнянні з використанням TF-IDF матриць [4, 29] для заданих умов класифікаційного аналізу. Класифікація текстових даних за ознаками семантичних полів дозволяє провести інтелектуальний аналіз текстів у експертно створеному векторному просторі з відповідними семантичними акцентами, які відображають семантичну сторону предметної області аналізу. Прогнозні моделі класифікаційного аналізу текстових даних за семантичними та тематичними полями можуть бути складовими прогнозних

багаторівневих ансамблів на основі стекінгового підходу.

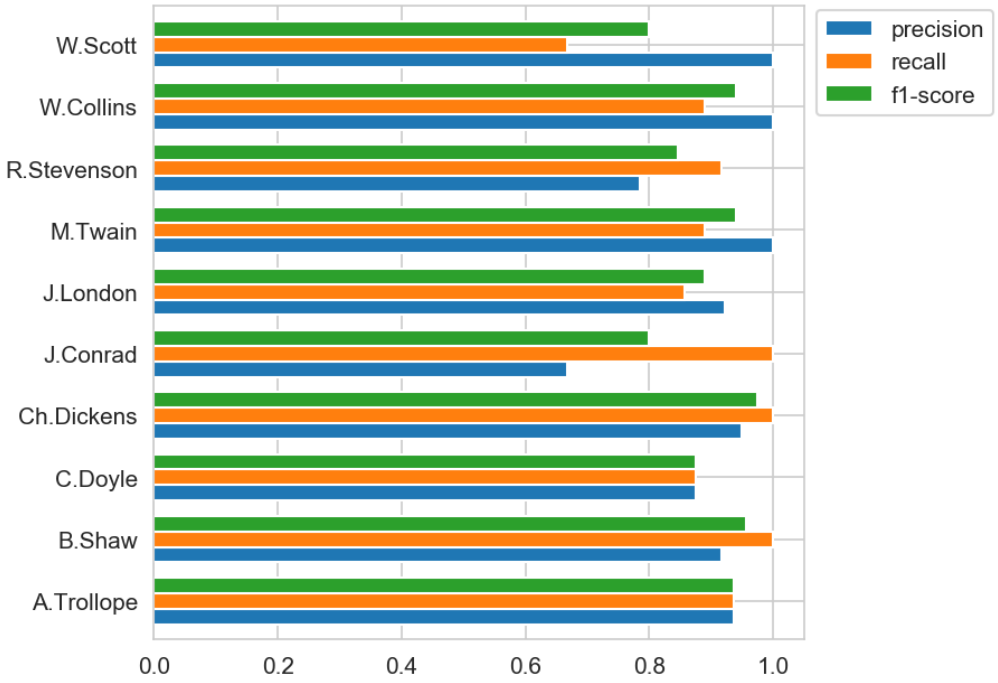

Рисунок 15 – Оцінки класифікації при використанні сукупних семантичних ознак в алгоритмі XGBoost

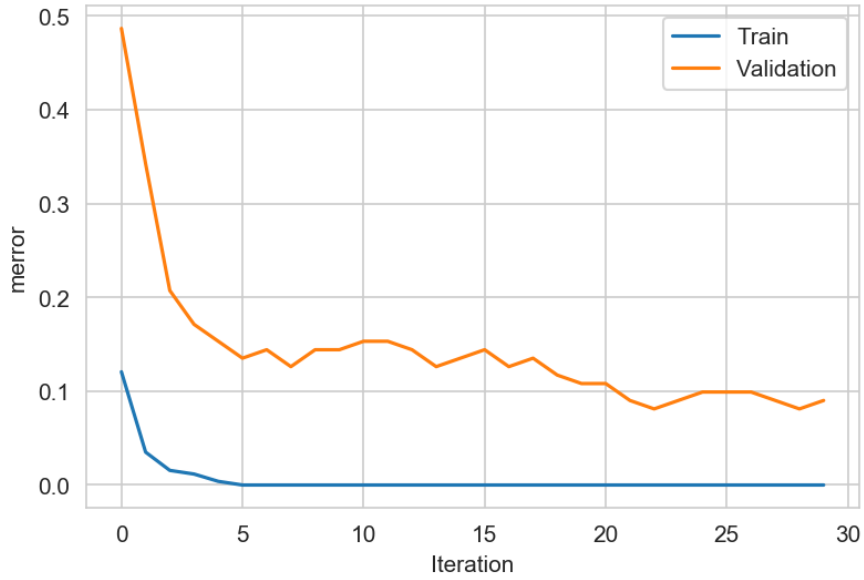

Рисунок 16 – Динаміка багатокласової класифікаційної похибки на тренувальному та валідаційному сетах в алгоритмі XGBoost

Проаналізовано застосування різних алгоритмів машинного навчання в інтелектуальному аналізі текстових даних, зокрема, Random Forest, XGBoost та нейронних мереж прямого поширення [29]. Для вивчення ефекту генералізації класифікаційних алгоритмів текстову вибірку було розділено на тренувальний та валідаційний сети. На рис. 15 наведено оцінки класифікації при використанні сукупних семантичних ознак на основі семантичних і тематичних полів, компонент сингулярного розкладу TF-IDF матриці та компонент латентного розміщення

Діріхле в алгоритмі XGBoost. На рис. 16 відображено динаміку багатокласової класифікаційної похибки на тренувальному та валідаційному сетах в алгоритмі XGBoost при використанні сукупних семантичних ознак. Розглянуто випадок використання комбінованої нейронної мережі, яка складається із рекурентної нейронної підмережі для аналізу текстових даних та підмережі для числових семантичних ознак текстових документів. Структуру такої нейронної мережі зображено на рис. 17. Проаналізовано числову регресію, у якій як вхідні розглядалися текстові дані для випадку аналізу цін за текстовим описом товарів. Для аналізу було вибрано аналогічну до зображеної на рис. 17 комбіновану нейронну мережу із LSTM підмережею для текстових даних і підмережею із повністю з'єднаними шарами для числових компонент сингулярного розкладу матриці TF-IDF.

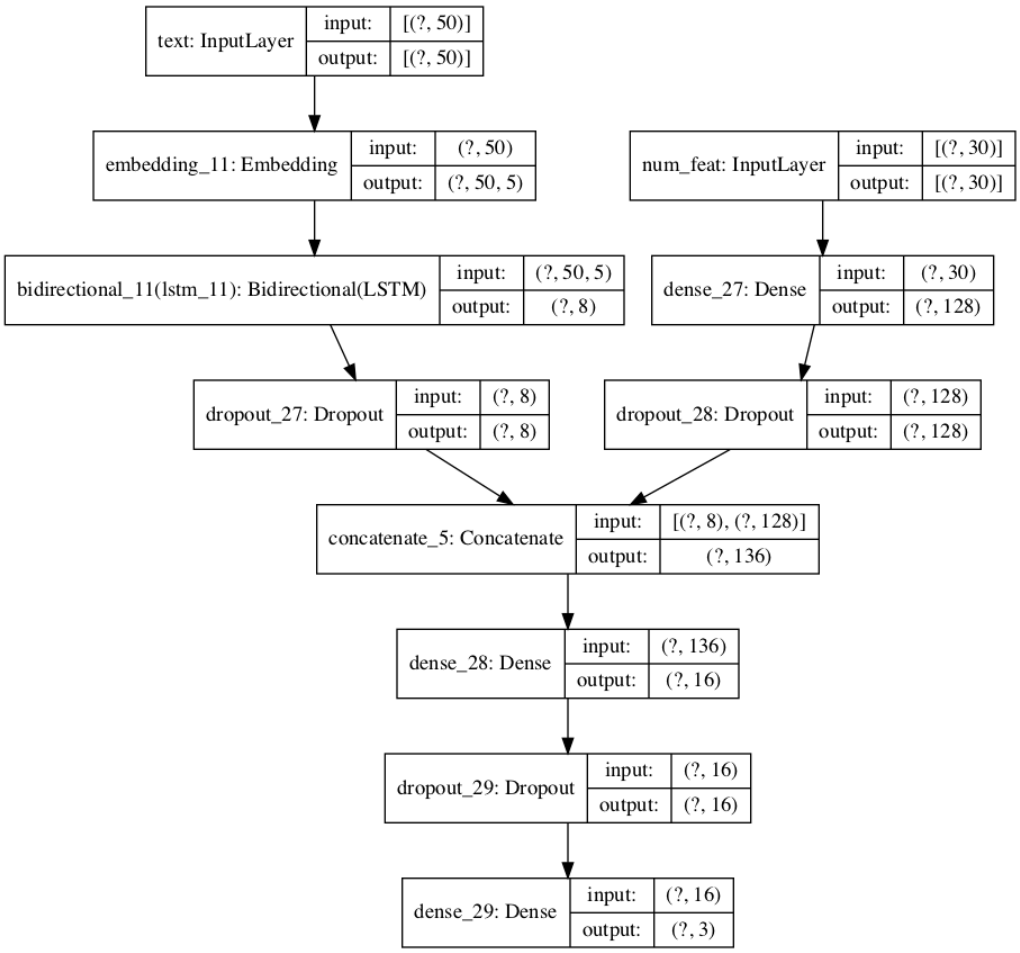

Рисунок 17 – Структура нейронної мережі з підмережами для текстових та числових даних

Результати показують, що у зразках даних із числовими та текстовими типами ознак та числовою цільовою змінною можуть бути виявлені відповідні патерни, які проявляються у зростанні точності прогнозування на ітераціях навчання комбінованої нейронної мережі [29].

Розглянуто використання генетичних алгоритмів для формування оптимального набору семантичних ознак у задачах машинного навчання [2, 23]. Показано, що 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.

Розглянуто використання квантових алгоритмів в аналізі даних, зокрема, модель квантового представлення семантичних векторів масиву текстових об'єктів, яка дає можливість експоненційно зменшити об'єм необхідної квантової пам'яті у порівнянні з класичним випадком [6, 13, 14, 22]. Запропоновано метод використання квантового алгоритму Гровера для пошуку 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.

У **п'ятому розділі** розглянуто використання теорії частих множин та асоціативних правил в аналітиці текстових повідомлень соціальних мереж, зокрема Твіттера [11, 26, 27, 30, 34, 44]. Пошук частих множин у повідомленнях мікроблогів 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. 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 можуть бути сформовані на основі тематичного поля, під яким 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. Розглянуто використання теорії графів для аналізу повідомлень мережі Твіттер, зокрема, для аналізу зв'язків між користувачами та виявлення різних спільнот. Зв'язки між користувачами визначаються на основі посилань у твітах. Різні спільноти можуть формувати різні інформаційні тренди. Приналежність користувача до спільноти може бути додатковою ознакою у прогнозній моделі. Показано, що в дискусіях різних тематик користувачі утворюють стійкі об'єднані групи, де існують свої впливові учасники, які формують напрям дискусії. Показано, що в потоках твітів, у яких обговорюються очікувані події, можна виявити ознаки на основі частих множин, які мають прогнозний потенціал стосовно цих подій. Проаналізовано твіти, пов'язані з компанією Tesla. На рис. 18 зображено граф структурної організації груп користувачів у масиві твітів, на рис. 19 наведено приклади виявлених асоціативних правил у структурі твітів.

Показано, що квантитативні характеристики частих множин та асоціативних правил, виявлених у масиві твітів, зокрема, величина підтримки, можуть бути використані в аналітиці як прогнозні ознаки у регресійних моделях. Використовуючи алгоритми аналізу графів, а також теорію частих множин та асоціативних правил, проведено інтелектуальний аналіз повідомлень мережі Твіттер, пов'язаних із пандемією COVID-19 [30]. На рис. 20 показано частоти ключових слів, які формують тематичне поле, що використовується для аналізу частих множин. Деякі ключові слова позначають сумарну частоту семантично близьких слів. На рис. 21 показано приклад виявлених частих множин. Знайдені часті множини та асоціативні правила відображають семантичну структуру в масиві твітів, пов'язаних із пандемією COVID-19. Прогнозну аналітику твітів із використанням теорії частих множин та асоціативних правил апробовано в аналізі та прогнозуванні подій, які відображають очікування користувачів соціальної мережі Твіттер. 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.

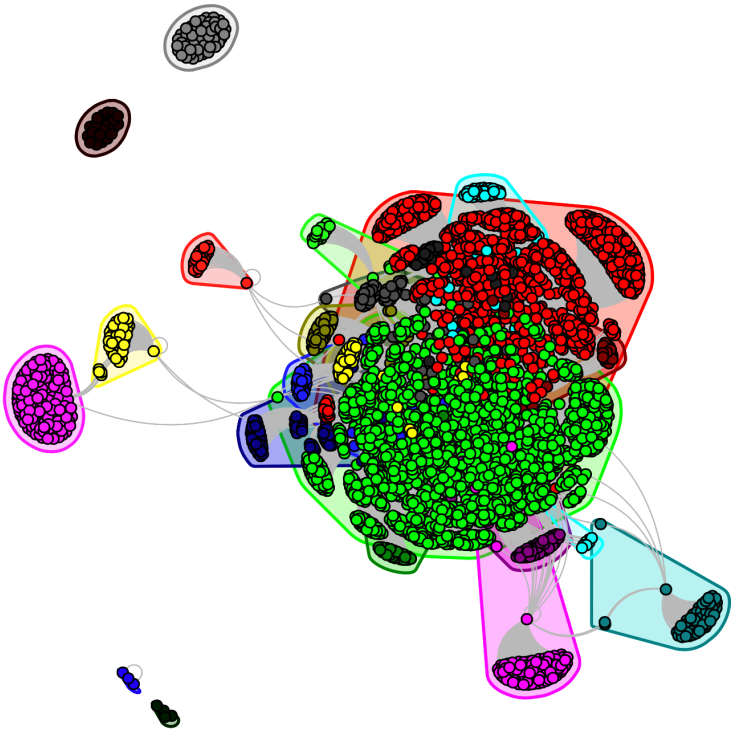

Рисунок 18 – Зображення графу структурної організації груп користувачів у масиві твітів

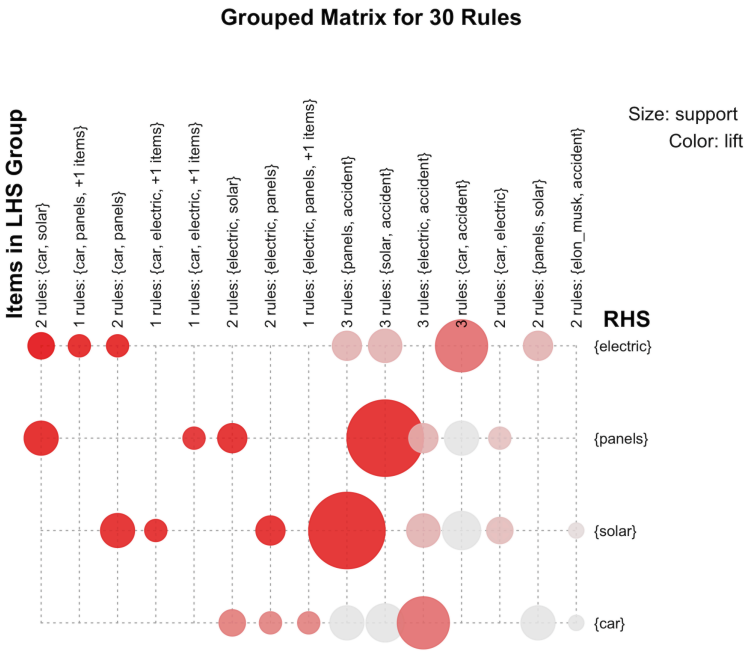

Рисунок 19 – Приклади асоціативних правил у структурі твітів

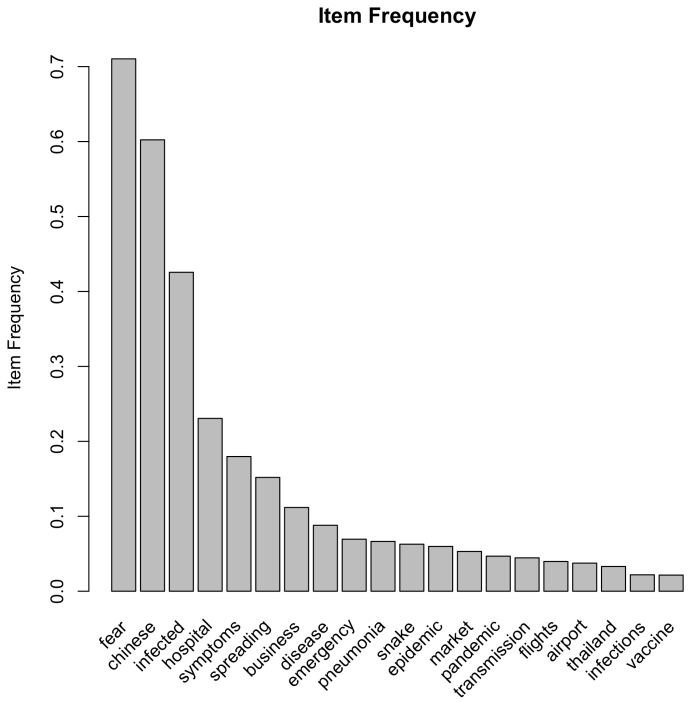

Рисунок 20 – Частоти ключових слів

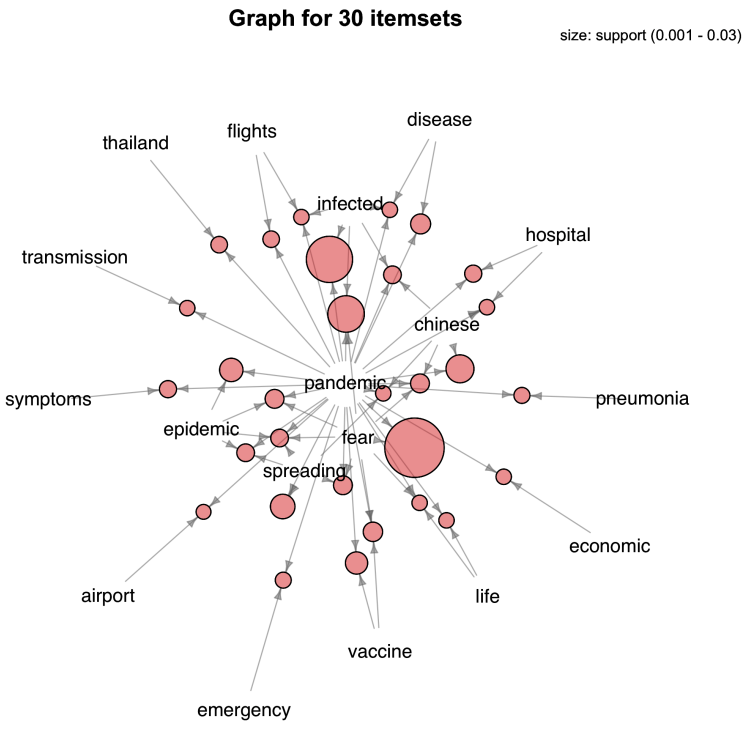

Рисунок 21 – Граф частих множин, які містять ключове слово *'pandemic'*

У **шостому розділі** на основі теорії аналізу формальних концептів запропоновано модель семантичного контексту, яка відображає структурну семантичну організацію текстових масивів [8,10,16,17,37,40,41]. 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, яку часто відображають за допомогою діаграми Гассе, дає можливість описувати ієрархічну семантичну структуру в масиві документів та виявляти групи текстових документів, об'єднаних спільними семантичними ознаками. На основі 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. Запропоновано застосування теорії аналізу формальних концептів в інтелектуальному аналізі текстових повідомлень Твіттера. Розглянуто модель ґратки семантичних концептів для аналізу твітів, згрупованих за користувачами Твіттера. Формальний об'єм такого семантичного концепту об'єднує користувачів, які часто вживають у своїх повідомленнях задані ключові слова, що утворюють множину формального змісту цього концепту. Уведення поняття семантичного поля як множини тематично об'єднаних лексем зменшує обсяг необхідних обчислень унаслідок фільтрації масиву повідомлень.

Формальний контекст для згрупованих за користувачами повідомлень розглянуто як

$$K_{usr}^{tw(kw)} = \left(TW_s^{usr(kw)}, Keywords, I_s^{usr} \right), \tag{11}$$

де I_s^{usr} – відношення $I_s^{usr} \subseteq TW_s^{usr(kw)} \times Keywords$, яке описує зв'язки згрупованих за користувачами повідомлень із ключовими словами в цих повідомленнях. Вважається, що $(tw_i^{usr(kw)}, keyword_j) \in I_s^{usr}$, якщо ключове слово $keyword_j$ зустрічається в масиві повідомлень користувача $tw_i^{usr(kw)}$ певну кількість разів n_{ij}^{usr}. Відношення I_s^{usr} можна розглядати як множину

$$I_s^{usr} = \left\{ (tw_i^{usr}, keyword_j) | keyword_j \in tw_i^{usr(kw)}, n_{ij}^{usr} > n_{th}^{usr} \right\}. \tag{12}$$

Уведення порогового значення n_{th}^{usr} є необхідними для того, щоб включити до розгляду лише ключові слова тематик, які активно обговорюються. Для деяких $Ext^{usr} \subseteq TW_s^{usr(kw)}$, $Int^{usr} \subseteq Keywords$ визначено такі відображення

$$\begin{aligned} &(Ext^{usr})' = \\ &\left\{ keyword_j \in Keywords | tw_i^{usr(kw)} \in Ext^{usr} : (tw_i^{usr(kw)}, keyword_j) \in I_s^{usr} \right\}, \\ &(Int^{usr})' = \\ &\left\{ tw_i^{usr(kw)} \in TW_s^{usr(kw)} | keyword_j \in Int^{usr} : (tw_i^{usr(kw)}, keyword_j) \in I_s^{usr} \right\}. \end{aligned} \tag{13}$$

Множина $(Ext^{usr})'$ описує ключові слова, властиві згрупованим за користувачами повідомленням множини Ext^{usr}, а множина $(Int^{usr})'$ описує повідомлення користувачів, які містять ключові слова множини Int^{usr}. Уведено 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 множини $Int^{usr} \subseteq Keyword$ з умовами $(Ext^{usr})' = Int^{usr}, (Int^{usr})' = Ext^{usr}$, де Ext^{usr} є формальний об'єм, а Int^{usr} – формальний зміст семантичного концепту $Concept^{usr}$. Формальний об'єм семантичного концепту об'єднує користувачів, які часто вживають у своїх повідомленнях ключові слова, які утворюють множину формального змісту цього концепту. В семантичному контексті $K_{usr}^{tw(kw)}$ утворюється частково впорядкована множина семантичних концептів повідомлень, згрупованих за користувачами

$$\Psi^{usr}(TW_s^{usr(kw)}, Keywords, I_s^{usr}) = \{ Concept_m^{usr} = (Ext_m^{usr}, Int_m^{usr}) \}. \tag{14}$$

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, які характеризують семантичні поняття у повідомленнях користувачів.

Проведено аналіз семантичних концептів у повідомленнях Твіттера заданих тематик, які визначаються наявністю ключових слів. На рис. 22 наведено діаграму Гассе, яка відображає утворену ґратку семантичних концептів для семантичного

поля, пов'язаного з програмним забезпеченням. На цій діаграмі наведено формальний зміст концептів верхнього рівня. Формальні змісти концептів нижніх рівнів є комбінаціями формальних змістів наведених концептів верхнього рівня відповідно до зв'язків на діаграмі.

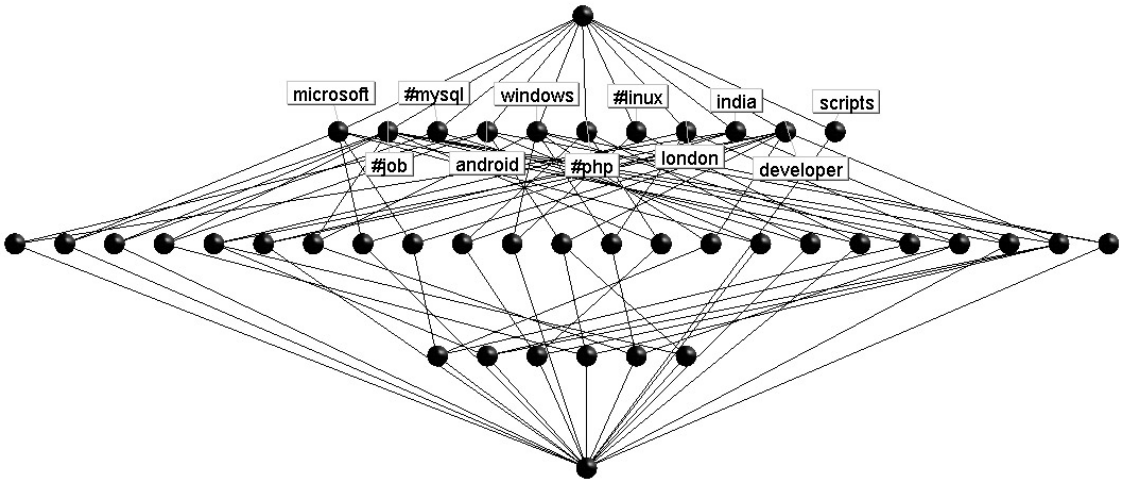

Рисунок 22 – Діаграма Гассе для ґратки семантичних концептів

Розглянуто поняття порядкового ідеала та фільтра для деякої частково впорядкованої множини $(P, \leq)$. Порядковим ідеалом називають підмножину $J \subseteq P$, для якої $\forall x \in J, y \leq x \Rightarrow y \in J$. Порядковим фільтром називають підмножину $F \subseteq P$, для якої $\forall x \in F, y \geq x \Rightarrow y \in F$. 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.

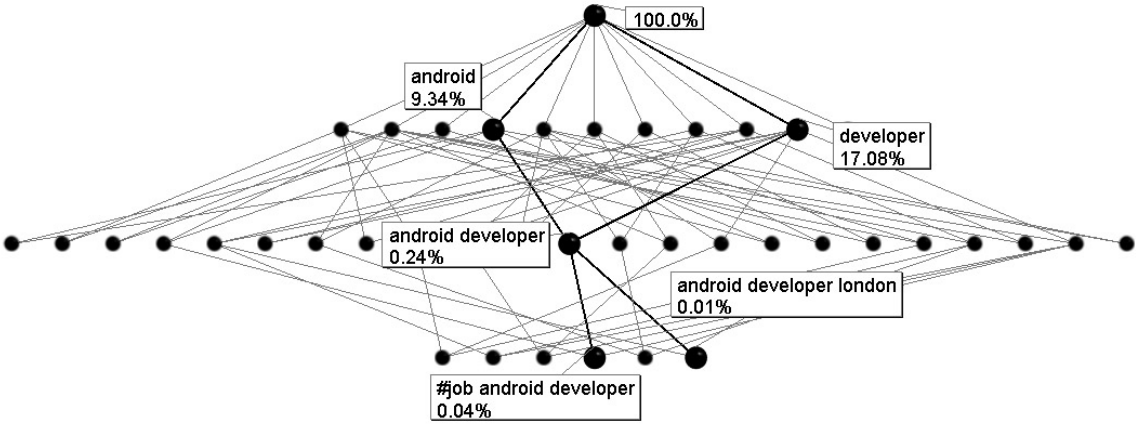

Рисунок 23 – Порядковий фільтр та ідеал для концепта *{android, developer}*

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. Множина формальних змістів такого об'єднання утворює деяке семантичне поле, яке відображає множину взаємопов'язаних понять, описаних ключовими словами. Одним із методів формування семантичних полів є пошук множини формальних змістів концептів деякого об'єднання порядкових ідеала та фільтра заданого формального контексту. На рис. 23 наведено приклад порядкового фільтра та ідеала для заданого семантичного концепта в масиві повідомлень Твіттера. На основі розрахованої ґратки семантичних концептів можна виявити асоціативні правила, які відображають семантичні структурні зв'язки між ключовими словами. Асоціативним правилом деякого контексту

$$K^{tw(kw)} = \left(TW_s^{(kw)}, Keywords, I_s\right)$$

є вираз

$$A \to B,\ A, B \subseteq Keywords. \tag{15}$$

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 обчислити за такими виразами:

$$Supp_{A\to B} = \frac{|(A \cup B)'|}{\left|TW_s^{(kw)}\right|}, \tag{16}$$

$$Conf_{A\to B} = \frac{|(A \cup B)'|}{|A'|}. \tag{17}$$

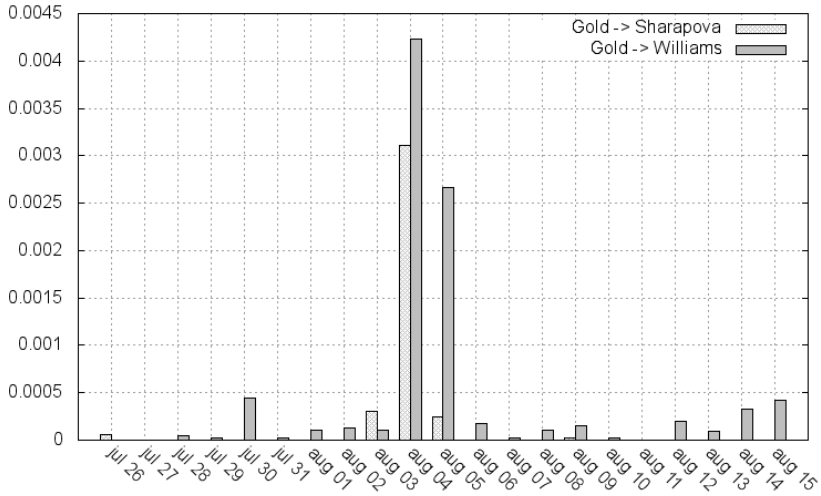

Рисунок 24 – Динаміка підтримки для асоціативних правил $Gold \to Sharapova$, $Gold \to Williams$

На основі змістів концептів, які відповідають заданій тематиці, можна сформувати базис семантичного простору текстових документів. 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. Розглянуту у роботі модель ґратки семантичних концептів масиву твітів апробовано в аналізі та прогнозуванні спортивних подій на прикладі фінального тенісного турніру олімпіади у Лондоні 2012 року. У спортивних подіях очікування користувачів можуть бути відкореговані випадковими факторами та реальним рівнем підготовки спортсменів, який може відрізнятися від очікувань болільників. На рис. 24 наведено динаміку підтримки для асоціативних правил $Gold \rightarrow Sharapova$, $Gold \rightarrow Williams$. Характерними для отриманих кривих є максимуми в день змагань та наступного дня. Підтримка та достовірність асоціативних правил у день змагань відображають очікування, а в день після змагань – реальний результат. Величини підтримки відповідних концептів для фіналістів були співмірними перед початком турніру, та суттєво відмінними після завершення турніру, що відповідає результатам фіналу. Використання моделі формальних концептів в аналізі текстових 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 та результати змагань.

У **додатках** наведено інші результати, які доповнюють основний матеріал дисертації.

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 ОПУБЛІКОВАНИХ ПРАЦЬ 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)

АНОТАЦІЯ

Павлишенко Б. М. Методи інтелектуального аналізу консолідованих даних для підтримки прийняття рішень. – На правах рукопису.

Дисертація на здобуття наукового ступеня доктора технічних наук за спеціальністю 05.13.23 – системи та засоби штучного інтелекту. – Харківський національний університет радіоелектроніки, Міністерство освіти і науки України, Харків, 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. Розроблено метод оптимізації прогнозної аналітики часових рядів з використанням стекінгового об’єднання та відбору різнотипних моделей на основі лінійної регресії LASSO та байєсівської регресії. Проаналізовано поєднання байєсівської, лінійної та машино-навчальної логістичних регресій у задачах виявлення технічних відмов. Розглянуто оптимізацію послідовності дій інтелектуального агента в задачах аналітики попиту з використанням глибокого Q-навчання та імітаційного моделювання середовища. Запропоновано модель векторного представлення текстових даних, у просторі семантичних семантичних та тематичних полів. Проведено аналіз текстових даних на основі алгоритмів машинного навчання з використанням кількісних семантичних ознак. Розроблено метод виявлення додаткових аналітичних ознак на основі лексемних поєднань у семантичних структурах текстових масивів. Запропоновано модель семантичних концептів текстових масивів на основі теорії аналізу формальних концептів.

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

АННОТАЦИЯ

Павлишенко Б. М. Методы интеллектуального анализа консолидированных данных для поддержки принятия решений. – На правах рукописи.

Диссертация на соискание ученой степени доктора технических наук по специальности 05.13.23 – системы и средства искусственного интеллекта. – Харьковский национальный университет радиоэлектроники, Министерство образования и науки Украины, Харьков, 2021.

Диссертационная работа посвящена разработке методов моделирования, формирования аналитических признаков, интеллектуального анализа табличных и текстовых консолидированных данных для повышения точности, достоверности и информативности результатов анализа, которые используются для поддержки принятия решений в информационно-аналитических системах. Разработан метод оптимизации прогнозной аналитики временных рядов с использованием стэкингового объединения и отбора разнотипных моделей на основе линейной регрессии LASSO и байесовской регрессии. Проанализировано сочетание байесовской, линейной и машинно-обучаемой логистических регрессий в задачах выявления технических отказов. Рассмотрена оптимизация последовательности действий интеллектуального агента в задачах аналитики спроса с использованием глубокого Q-обучения и имитационного моделирования среды взаимодействия. Предложена модель векторного представления текстовых данных в пространстве семантических и тематических полей. Проведен анализ текстовых данных на основе алгоритмов машинного обучения с использованием количественных семантических признаков. Разработан метод выявления дополнительных аналитических признаков на основе лексемных сочетаний в семантических структурах текстовых массивов. Предложена модель семантических концептов текстовых массивов на основе теории формальных концептов.

Ключевые слова: интеллектуальный анализ данных, методы машинного обучения, признаки данных, временные ряды, семантические поля, частые множества, ассоциативные правила, анализ формальных концептов.

ABSTRACT

Pavlyshenko B. M. Methods of intellectual analysis of consolidated data for decision-making support. – On the rights of the manuscript.

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

The thesis focuses on the development of methods of modeling, formation of analytical features, intellectual analysis of tabular and textual consolidated data for increasing 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 reinforcement machine learning 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 represent 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 intellectual analysis of text arrays; the method of classification and regression analysis of different types of consolidated data based on the combination of LSTM neural network for input text data and neural network with fully connected layers for 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 analysis has been developed, making it possible to identify effective analytical features taking into account the semantic structure of text datasets.

The results obtained in the thesis research and the developed methods are a component technology for decision-making support in complex information systems and they 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 analysis.

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: data mining, machine learning methods, data features, time series, semantic fields, frequent sets, associative rules, formal concept analysis.

Підписано до друку 03.03.21
Формат 60х84/16. Папір офсетний.
Друк на різографі. Зам. №03/03-1
Ум. друк. арк. 1,8
Наклад 100 прим.

Видавництво “Галич-Прес”
Видавець ФОП Король І.В.
м. Львів, вул. Гнатюка, 17
Ел. пошта: lvivprint@ukr.net. Тел. 096-59-88-924
Свідоцтво ДК №5353 від 24.05.2017 р.