ВСЕУКРАЇНСЬКИЙ КОНКУРС СТУДЕНТСЬКИХ НАУКОВИХ РОБІТ
ЗІ СПЕЦІАЛЬНОСТІ “**КОМП’ЮТЕРНІ НАУКИ**”

Шифр «Emonito»

СТУДЕНТСЬКА НАУКОВА РОБОТА
ТЕМА: «МЕТОД ТА ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ СЕНТИМЕНТ-АНАЛІЗУ ТЕКСТОВОГО КОНТЕНТА ІЗ СОЦІАЛЬНИХ МЕРЕЖ НА ОСНОВІ КЛАСИФІКАЦІЇ ЧАСОВИХ РЯДІВ СЕНТИМЕНТ-ОЦІНОК»

2020

ЗМІСТ

ВСТУП

Наразі соціальні мережі – невід'ємна частина нашого життя і це є незаперечним фактом [1]. За даними соціальних досліджень людина проводить у мережі Інтернет до дев'яти годин на добу, з них третина – на спілкування в мережах [2]. У процесі спілкування користувачі соцмереж обмінюються різноплановим контентом, який включає в себе як семантичну, так і сентиментальну складову текстових повідомлень, емотикони, емодзі, фото і відео. Тому актуальним є розвиток інформаційних технологій аналізу процесів, що відбуваються в соціальних мережах, зокрема обробки та аналізу такого контенту.

Наприклад, наявність таких інструментів у сфері обслуговування дозволяє організувати ефективний зворотний зв'язок з клієнтами з метою аналізу та оптимізації якості відповідного сервісу. При цьому слід виділити два важливих аспекти. Перший з них пов'язаний з необхідністю аналізу клієнтського текстового контенту фахівцями обслуговуючих компаній – маркетологами, логістами, фахівцями з реклами, аналітиками і т. д. Другий аспект зачіпає проблему розвитку автоматизованих сервісів з використанням, наприклад, чат-ботів, здатних самостійно аналізувати клієнтські повідомлення і приймати відповідні рішення.

Практика показує, що при отриманні зворотного зв'язку від клієнтів важливо якомога раніше відстежити і виявити тенденції в їх емоційних реакціях на події, що нас цікавлять з подальшим моніторингом динаміки їх емоцій. При цьому таке завдання може вимагати одночасного аналізу інформації з багатьох сотень або тисяч джерел клієнтського контенту. Розробка методу та інформаційної технології на його основі, що дозволяють справлятися з такими викликами і при цьому бути простими для застосування на практиці, є актуальним і головним завданням даної роботи.

Мета − удосконалення існуючих підходів обробки неструктурованого текстового контенту з соціальних мереж шляхом розробки методу і

інформаційної технології класифікації часових рядів тональних оцінок, які несуть основну інформацію щодо настрою користувачів і дають уявлення про їх динаміку.

Для досягнення поставленої мети необхідно вирішити **наступні задачі:**

1. Розробка методу сентимент-аналізу текстового контенту для багатопоточного аналізу динаміки часових оцінок текстового контенту користувачів соціальних мереж.
2. Експериментальні дослідження запропонованого методу.
3. Аналіз і синтез інформаційної технології.

Об'єкт досліджень – процес сентимент-аналізу текстового контенту із соціальних мереж.

Предмет досліджень – метод та інформаційна технологія сентимент-аналізу текстового контенту із соціальних мереж на основі класифікації часових рядів сентимент-оцінок.

Методи досліджень. При виконанні поставлених задач використовуються вибірковий метод, методи соціально-мережевого аналізу (Social Network Analysis, SNA), обробки природної мови (Natural Language Processing, NLP), інтелектуального аналізу даних (Data Mining, DM).

Наукова новизна:

– вперше запропоновано метод сентимент-аналізу текстового контенту із соціальних мереж, який, на відміну від відомих, дозволяє визначати тональність текстового контенту при паралельному прослуховуванні великої кількості акаунтів, шляхом класифікації коротких ділянок часових рядів сентимент-оцінок з використанням аналізу головних компонент.

Практична цінність:

– розроблені метод і інформаційна технологія можуть бути покладені в основу DAAS та SAAS проектів, які дозволять автоматично проводити збір, аналіз та візуалізацію інформації із акаунтів соціальних мереж та месенджерів з використанням сентимент-складової текстового контенту;

– запропоновані наукові та інженерні рішення можуть бути використані при розв'язанні задач інформаційного моніторингу, впливу та протиборства у сфері маркетингу, реклами та інформаційних війнах.

Наукові публікації та особистий внесок авторів. У роботі [4] авторам належать такі результати: розроблено модель розширеної думки для задач семантичного та сентимент-аналізу текстового контенту із соціальних мереж. У роботі [26] – запропоновано концепцію збору та обробки неструктурованого текстового контенту з соціальних мереж у задачах сентимент-аналізу. У роботі [27] – розглянуті питання оцінки атрибутів автора безпосереднього думки в задачах сентимент-аналізу текстового контенту з соціальних мереж. У роботі [25] – розроблено метод сентимент-аналізу текстового контента із соціальних мереж на основі класифікації часових рядів сентимент-оцінок.

Апробація результатів. Основні положення дослідження обговорено на XXV міжнародній науково-технічній конференції студентів, аспірантів та молодих учених «Актуальні проблеми життєдіяльності суспільства», 25–26 квітня, 2018 р., м. Кременчук, XXVI міжнародній науково-технічній конференції студентів, аспірантів та молодих учених «Актуальні проблеми життєдіяльності суспільства», 24–25 квітня, 2019 р., м. Кременчук, 20th International Scientific Conference LOGI 2019, Institute of Technology and Business in České Budějovice, Faculty of Technology, Department of Transport and Logistics Czech Republic, November 14th-15th 2019.

Публікації. Основні наукові результати дослідження було опубліковано в 3-ох тезах доповідей на міжнародних конференціях та у одній науковій статті в закордонному виданні (індексується у WoS).

Структура і обсяг роботи. Робота складається зі вступу, 4-ох розділів, висновків та додатку, що складає 35 сторінок друкованого тексту, включає 7 рисунків, список використаної літератури на 27 назв та 1 додаток.

1 АНАЛІЗ ІСНУЮЧИХ МЕТОДІВ СЕНТИМЕНТ-АНАЛІЗУ ТЕКСТОВОГО КОНТЕНТУ

1.1 Процес сентимент-аналізу та сфери його застосування

Як було зазначено вище, завдання аналізу емоційного забарвлення тексту в мережі Інтернет набувають все більшої актуальності в зв'язку з величезною аудиторією мережі, зростаючим середнім часом перебування в ній. Аналітика та моніторинг соціальних мереж становить величезний інтерес для соціологів, лінгвістів, психологів, маркетологів і державних структур.

Для вирішення завдань аналізу емоційного забарвлення тексту в комп'ютерній лінгвістиці використовуються методи контент-аналізу, загальна назва для яких – Sentiment Analysis (аналіз тональності тексту) [6, 7].

Великі компанії використовують соціальні мережі для дослідження думок про свої продукти та послуги [17]. Такий підхід, на відміну від опитувань, наприклад, на сайті виробника, забезпечує більшу широту дослідження думок.

В органах, що забезпечують державну безпеку, контент-аналіз використовується для фільтрації та виявлення повідомлень, що містять інформацію про протиправні дії (терористичні загрози, аналіз забороненого контенту на сайтах тощо) [18].

У соціологічних дослідженнях методи контент-аналізу дозволяють не тільки встановити поточне ставлення населення до об'єкта дослідження, а й прогнозувати подальше ставлення до об’єкта. Наприклад, у США існує проект Pulse of the Nation [6], метою якого є визначення настрою громадян країни протягом дня шляхом дослідження та аналізу їх записів у популярній соціальній мережі Twitter.

Визначення тональності текстових повідомлень здатне виявити емоційно забарвлену лексику і проаналізувати оцінку автора по відношенню до об'єктів, мова про які йде в тексті.

Аналіз тональності тексту (англ. *Sentiment analysis*) – вид методів контент-аналізу в комп'ютерній лінгвістиці, призначений для автоматизованого виявлення в текстах емоційно забарвленої лексики і емоційної оцінки авторів (думок) по відношенню до об'єктів, мова про які йде в тексті [6].

Тональність – це емоційне ставлення автора висловлювання до деякого об'єкту, виражене в тексті. Тональність всього тексту в цілому можна визначити як функцію (в найпростішому випадку суму) лексичних тональностей складових його одиниць (речень) і правил їх поєднання [5].

У сучасних системах автоматичного визначення емоційної оцінки тексту найчастіше використовується одновимірний емотивний простір: позитив чи негатив (добре або погано) [6]. Однак відомі успішні випадки використання і багатовимірних просторів [5].

У контексті задач, які перед собою ставлять автори даної роботи, використовується дихотомічна сентимент-оцінка, основана на оцінці ймовірності відношення висловлювання до позитивного чи негативного (див. п. 3) [11].

1.2 Методи сентимент-аналізу та їх застосування для аналізу тональності тексту із соціальних мереж

Існує декілька видів класифікації методів сентимент-аналізу текстового контенту (рис. 1.1) [6].

При оцінюванні тональності за бінарною шкалою для визначення полярності документа використовується два класи оцінок: позитивна чи негативна. Одним з недоліків цього підходу є те, що емоційну складову документа не завжди можна однозначно визначити, тобто документ може містити ознаки як позитивної, так і негативної оцінки.

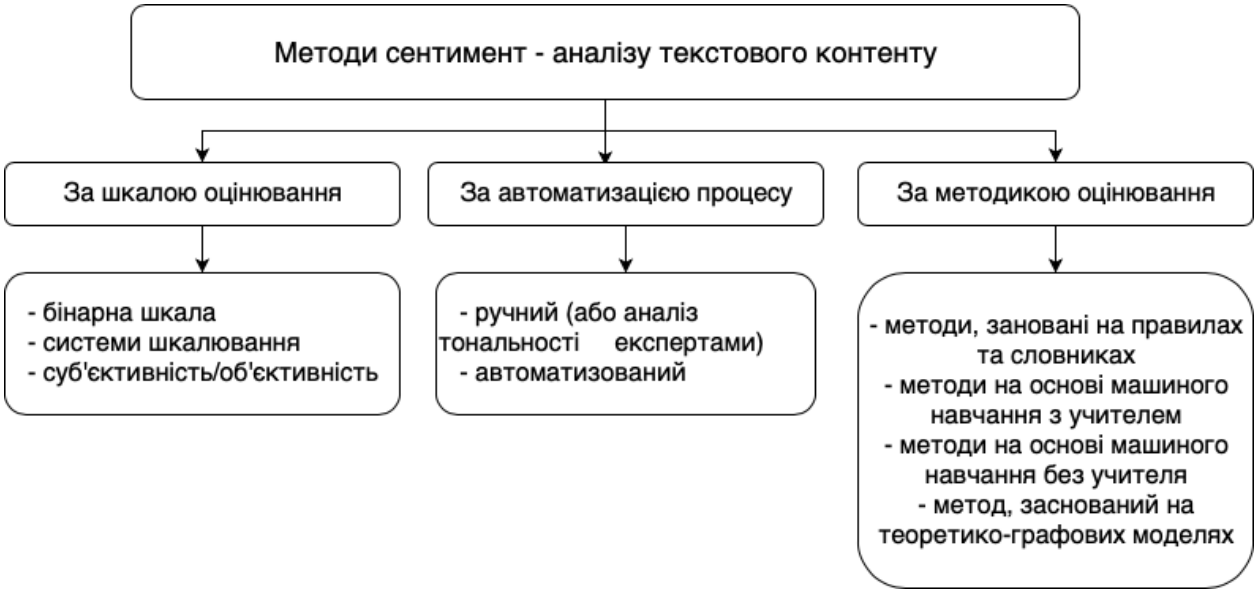

Рисунок 1.1 – Класифікація методів сентимент-аналізу текстового контенту

При оцінюванні з використанням методів шкалювання словам, зазвичай пов’язаних з негативними, нейтральними або позитивними тональностями, ставляться відповідно числа за шкалою від -10 до 10 (від негативного до самого позитивного). Спочатку фрагмент неструктурованого тексту досліджується з допомогою інструментів та алгоритмів обробки природної мови, а потім виділені з цього тексту об'єкти та терміни аналізуються з метою розуміння значення цих слів [19].

Інший дослідницький напрямок це ідентифікація суб'єктивності/об’єктивності. Це завдання зазвичай визначається як віднесення даного тексту до одного з двох класів: “суб’єктивний” або “об’єктивний”. Ця проблема іноді може бути більш складною, ніж класифікація полярності: суб'єктивність слів і фраз може залежати від контексту, а об'єктивний документ може містити в собі суб’єктивні пропозиції (наприклад, новина або стаття, що цитує думки людей).

За автоматизацією процесу методи сентимент-аналізу текстового контенту поділяють на автоматизовані та ручні. Найбільш помітні відмінності між ними лежать в ефективності системи і точності аналізу. У комп’ютерних програмах автоматизованого аналізу тональності застосовують алгоритми

машинного навчання, інструменти статистики і обробки природної мови, що дозволяє обробляти великі масиви тексту, включаючи веб-сторінки, онлайн-новини, тексти дискусійних груп у мережі Інтернет, онлайн-огляди, веб-блоги та соціальні медіа.

За методикою оцінювання методи сентимент-аналізу текстового контенту поділяють на наступні категорії.

Методи, засновані на правилах і словниках. Ця група методів заснована на пошуку *емотивної лексики* (лексичної тональності) в тексті по заздалегідь складеним тональним словникам і правилам із застосуванням лінгвістичного аналізу. За сукупністю знайденої емотивної лексики текст може бути оцінений за шкалою, що містить кількість негативної та позитивної лексики. Даний метод може використовувати як списки правил, так і спеціальні правила з'єднання тональної лексики всередині речення. Щоб проаналізувати текст, можна скористатися наступним алгоритмом: спочатку кожному слову в тексті привласнити його значення тональності зі словника (якщо воно присутнє в словнику), а потім обчислити загальну тональність всього тексту шляхом підсумовування значення тональностей кожного окремого речення.

Основною проблемою методів, заснованих на словниках і правилах, вважається трудомісткість процесу складання словника. Для того, щоб отримати метод, що класифікує документ з високою точністю, терміни словника повинні мати оцінки, адекватні предметній області документа. Наприклад, слово «величезний» по відношенню до обсягу пам’яті жорсткого диска є позитивною характеристикою, але негативною по відношенню до розміру мобільного телефону. Тому даний метод вимагає значних трудовитрат, так як для хорошої роботи системи необхідно скласти велику кількість правил. Існує ряд підходів, що дозволяють автоматизувати складання словників для конкретної предметної області (наприклад, тематика ресторанів або тематика мобільних телефонів).

Машинне навчання з вчителем. У наш час найбільш часто використовуваними в дослідженнях методами є методи на основі машинного

навчання з учителем [20]. Суттю таких методів є те, що на першому етапі навчається машинний класифікатор (наприклад, байєсівський) на заздалегідь розмічених текстах, а потім використовують отриману модель при аналізі нових документів.

Машинне навчання без вчителя. В основі цього підходу лежить ідея, що терміни, які найчастіше зустрічаються в цьому тексті і в той же час присутні в невеликій кількості текстів у всій колекції мають найбільшу вагу в тексті [20] Виділивши ці терміни, а потім визначивши їх тональність, можна зробити висновок щодо тональності всього тексту.

Метод, заснований на теоретико-графових моделях. В основі цього методу використовується припущення про те, що не всі слова в текстовому корпусі документа рівнозначні [15]. Певні слова мають більшу вагу і сильніше впливають на тональність тексту. Для класифікації слів використовується тональний словник, в якому кожне слово співвідноситься з оцінкою, наприклад «позитивна», «негативна» або «нейтральна».

При поточному розвитку машинного навчання та автоматизації процесів, вважається доцільним використання машинного навчання для аналізу тональності текстового контенту. При цьому машинне навчання з учителем і без нього активно використовується спеціалістами даної області.

При цьому активно використовуються метод опорних векторів [21], наївний класифікатор Байєса [16], дерева прийняття рішень [22] та метод максимальної ентропії [23]. Також для вирішення даних задач використовують нейронні мережі [24].

Важливо зазначити, що прикладні задачі, які потребують застосування методів сентимент-аналізу у соціальних мережах та каналах месенджерів, мають низку особливостей.

По-перше, нерідко виникає потреба «прослуховувати» *велику кількість інформаційних каналів* одночасно з реакцією від клієнтів, яким поставляється певна послуга, для реалізації зворотного зв’язку і покращення якості та оптимізації послуг, які їм надаються. Аналогічна ситуація може виникати і при

обробці результатів опитувань великої кількості респондентів (щодо якості послуги, товару, кандидата в президенти, політичної партії тощо). У даному випадку першочерговою задачею є оцінка емоційної компоненти (негатив/позитив) з послідуючою ідентифікацією власне об’єкта негативної (позитивної) реакції, чи його властивостей.

По-друге, досвід спілкування авторів даної роботи з представниками бізнесу показує, що ефективний зворотний зв’язок з клієнтом чи керуючий вплив на електорат має місце тоді, коли він здійснюється якомога раніше. Тому *своєчасний аналіз текстового контенту* з точки зору його як тональної, так і смислової компоненти, є вкрай важливим.

По-третє, в процесі моніторингу важливо відслідковувати *динаміку емоційної складової* клієнтів (читачів, електорату і т. п.), що дозволяє оцінювати і прогнозувати розвиток ситуації і, таким чином, своєчасно на неї реагувати.

І, нарешті, по-четверте, метод та технологія повинна давати *можливість доступної і зрозумілої інтерпретації результатів* пересічному користувачу.

У даній роботі зроблено спробу розробити метод та інформаційну технологію сентимент-аналізу текстового контенту, яка задовольняє вищевказаним вимогам.

2 РОЗРОБКА МЕТОДУ ТА ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ ОЦІНКИ ТОНАЛЬНОСТІ ТЕКСТОВИХ ПОВІДОМЛЕНЬ НА ОСНОВІ КЛАСИФІКАЦІЇ ЧАСОВИХ РЯДІВ СЕНТИМЕНТ-ОЦІНОК

2.1 Задача класифікації часових рядів сентимент-оцінок

2.1.1 Базові визначення та припущення

Основним завданням семантичного і сентимент-аналізу неструктурованого контенту з соціальних мереж є семантичний і сентимент-аналіз *повідомлень* (*меседжів*) і виявлення *безпосередніх думок* певних *авторів* (*суб'єктів*) щодо певних *об'єктів* або їх *властивостей* [5]. У контексті нашої задачі будемо говорити про безпосередню думку клієнтів компанії або пересічних користувачів соціальної мережі, яку вони висловлюють у текстових повідомленнях.

Публікацією (*publication*) будемо називати текстове повідомлення, що має певний сенс, рекламного або іншого характеру, опубліковане в акаунті соціальної мережі і т. п.

Текстовим *повідомленням* (*text message*) tm будемо називати текстовий контент, який згенерований тим чи іншим суб'єктом. Меседж розглядається як реакція на публікацію, що несе семантичну і тональну інформацію.

Суб'єктом (*subject*) sb будемо називати сутність (людину або віртуального співрозмовника (чат-бот)), яка згенерувала меседж. Меседж може (але не обов'язково) містити безпосередню думку щодо певного об'єкта або його властивостей. Тоді суб'єкт виступає в ролі автора думки.

Автором (*holder*) h думки називається суб'єкт, якому належить безпосередня думка.

Об'єктом тональності (*entity*), або *сутністю* e називається сутність, щодо якої автор висловив свою думку.

Властивістю об'єкта тональності (*feature*) f називається частина або

атрибут об’єкта тональності.

Тональної оцінкою (*orientation*, *polarity*) або *сентимент-оцінкою* op називається тональна оцінка, тобто емоційна позиція автора щодо згаданого об’єкту, або його властивості.

Таким чином, *безпосередньою думкою* (*direct opinion*) називається висловлювання автора про об’єкт, який представляється кортежем з п’яти елементів [5]:

$$do = (e,\ f,\ op,\ h, t), \tag{2.1}$$

де t – момент часу, коли було сформовано безпосередню думку.

Практика показує, що для сучасного стилю спілкування основної маси користувачів мережі характерні емотикони і емодзі як для семантичної та емоційної «забарвленості» основного текстового повідомлення, що передає зміст, так і при формуванні самостійних «образних» меседжів [3]. Однак «класичним» джерелом семантичної і сентимент-інформації є звичайний текст. У даній роботі розглядаються сентимент-оцінки традиційного тексту. Однак даний підхід може бути застосовано й до поняття розширеної думки, введеного авторами в роботі [4].

Оцінка сентимент-компоненти є нетривіальним завданням. Відомі різні підходи до оцінки сентимент-складової, що ґрунтуються на різних шкалах – як бінарних [6-8], так і багатополярних [9]. У контексті питань, розглянутих у даній статті, нами буде використана трирівнева шкала тональних оцінок (див. нижче) з наступними значеннями: «негативний», «нейтральний», «позитивний». Під «нейтральною» оцінкою мається на увазі оцінка, яка виникає у випадку, коли меседж не містить емоційного забарвлення.

Відштовхуючись від визначення (2.1), введемо поняття *часового ряду тональних оцінок* (*time series of tonal estimates*, *TETS*) [25], як впорядкованої за часом послідовності $op(t_1), op(t_2), \dots, op(t_i), \dots, op(t_n)$ тональних оцінок n авторів щодо одного і того ж об’єкта або його властивості в складі

безпосередньої думки. Важливо відзначити, що далеко не кожен текстовий меседж породжує безпосередню думку [4] і на практиці це може створювати певні труднощі під час аналізу. Трирівнева шкала сентимент-оцінок вирішує це питання просто: якщо меседж не породжує безпосередню думку, він отримує нейтральну сентимент-оцінку.

Перейдемо безпосередньо до постановки задачі в термінах, введених вище визначень та припущень.

2.1.2 Формальна постановка задачі

Нехай у певному акаунті соціальної мережі або в сеансі чату дискусії як реакція на публікацію сформувався набір q меседжів $TM = \{tm_1, tm_2, \dots, tm_i, \dots, tm_q\}$, на основі якого можна сформувати набір з $n \leq q$ безпосередніх думок $DO = \{do_1, do_2, \dots, do_i, \dots, do_n\}$. Набір безпосередніх думок DO породжує безпосередньо $op(t) = \big(op(t_1), op(t_2), \dots, op(t_i), \dots, op(t_n)\big)$.

Розглянемо g довільних публікацій довільних акаунтів і всі набори меседжів під кожною публікацією. В результаті матимемо колекцію наборів меседжів $TM_1, TM_2, \dots, TM_j, \dots, TM_g$, відповідно з довжинами $q_1, q_2, \dots, q_j, \dots, q_g$, які, в свою чергу, породжують колекцію наборів безпосередніх думок $DO_1, DO_2, \dots, DO_j, \dots, DO_g$ з довжинами $n_1, n_2, \dots, n_j, \dots, n_g$. Зауважимо, що $n_j \leq q_j$, так як не всі меседжі можуть породжувати безпосередні думки. Колекція наборів безпосередніх меседжів, в свою чергу, формує структуру даних, яку ми будемо називати колекцією наборів часових рядів сентимент-оцінок $OP = \Big(op_1(t), op_2(t), \dots, op_j(t), \dots, op_g(t)\Big)$ з довжинами $n_1, n_2, \dots, n_j, \dots, n_g$ відповідно і зберігати в таблиці розмірністю $\max(n_j) \times g$.

Відштовхуючись від властивостей колекції OP, необхідно запропонувати метод, який би дозволив класифікувати динаміку тональних оцінок авторів

думок у процесі їх реакції на публікацію за трирівневою шкалою: «негативна», «нейтральна», «позитивна», на основі невеликої ділянки часового ряду, довжина якої повинна бути обґрунтована.

Природно, першочерговим завданням є отримання оцінки колекції *OP* на реальному об'єкті і відповідей на ряд питань:

– які статистичні властивості часових рядів тональних оцінок і яка адекватна модель, що їх описує?

– чи існує природна сегментація в просторі часових рядів?

Для вирішення цього завдання був проведений експеримент в рамках ще одного важливого допущення: автори вважають, що є певні загальні властивості, характерні для всіх часових рядів тональних оцінок незалежно від мови і культури спілкування, типу соціальної мережі, чату і т. п.

3 ЕКСПЕРИМЕНТАЛЬНІ ДОСЛІДЖЕННЯ

3.1 Структура експериментального дослідження

На рис. 3.1 наведено структуру процесу проведення експериментального дослідження, що використовувалася для проведення експериментального дослідження. В якості джерела інформації було використано блог-платформу для ведення щоденників LiveJournal, яка надає API для роботи з її даними. На стадії «Інформаційний пошук» було взято список топових російськомовних блогерів [10] і в рамках перших 119-ти з них був виконаний імпорт/парсинг 2542-ох їх публікацій за період 2015-2019 рр. («Імпорт меседжей і парсинг»). Потім на їх основі створено колекцію меседжей («Побудова ТМ колекції»).

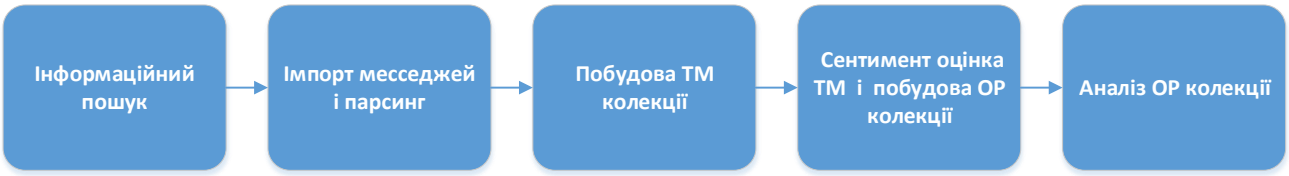

Рисунок 3.1 – Структура процесу проведення експериментального дослідження

Для оцінки сентимент-складової були використані ресурси проекта INDICO [11] і на основі результатів сформовано колекцію сентимент-оцінок часових рядів («Сентимент-оцінка ТМ і побудова ОР колекції»). Тональна оцінка визначалася як імовірність того, що месседж має позитивне або негативне забарвлення: значення ймовірностей більші за 0,5 вказують на позитивні настрої автора думки, тоді як значення менші за 0,5 – на негативні; значення, наближені до нуля, вказують на нейтральні настрої повідомлення.

На стадії «Аналіз ОР колекції» було виконано розвідувальний аналіз даних і запропоновано метод класифікації TETS.

3.2 Розвідувальний аналіз даних

На першому етапі були досліджені статистичні властивості часових рядів. Результати показали (рис. 3.2, табл. 3.1), що середня довжина набору меседжів дорівнює близько 24, а довжина 75% всіх наборів меседжів не перевищує 26. Виходячи з цього можна зробити висновок, що основна маса дискусій швидко згасає. У той же час, як вже зазначалося, що при спілкуванні з клієнтами, або зі споживачами послуги, важливо знати їх реакцію настрою на самому початку дискусії. Саме це дає можливість вчасно розуміти їх запити і приймати відповідні рішення.

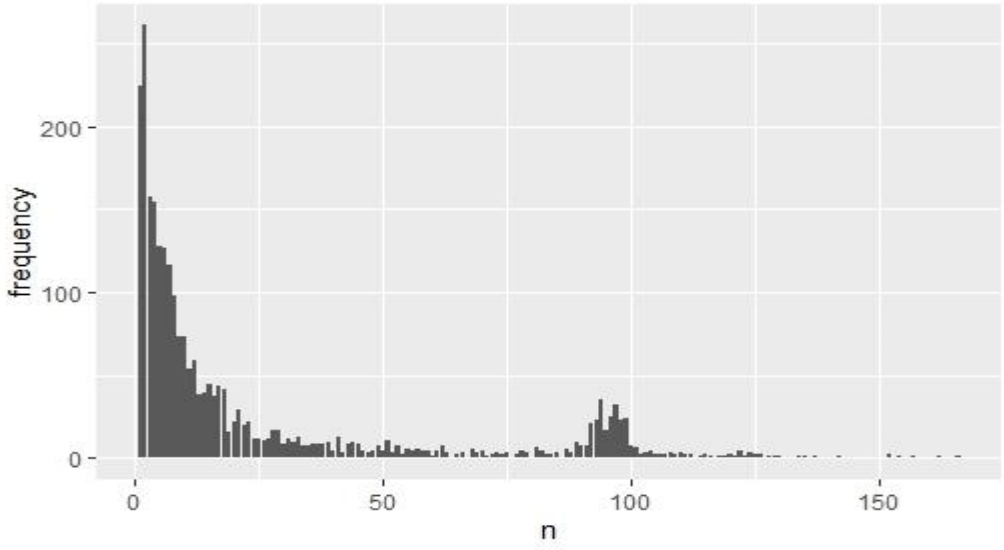

Рисунок 3.2 – Вибірковий розподіл частот довжини часових рядів тональних оцінок за даними 2542-ох спостережень

Таблиця 3.1 – Вибіркові характеристики часових рядів тональних оцінок

Стат. характе-ристики	мінімум	$1st\ Qu.$	медіана	середнє	$3st\ Qu.$	максимум
Значен-ня	1.00	3.00	9.00	23.63	26.00	166.00

На рис. 3.3 представлений найдовший ряд сентимент-оцінок з наявних у колекції й оцінки його автокореляційної (АКФ) і частинної автокореляційної (ЧАКФ) функцій, які відповідають характеристикам «білого шуму». Дослідження інших наборів у колекції показало, що більшість з них має аналогічні властивості. На підставі цього можна зробити висновок, що для природної дискусії (в чаті) по всій її довжині характерне хаотичне коливання тональності думок авторів, що не підтверджує наявність явно виражених класів часових рядів, наприклад, таких, як наводять автори в [12].

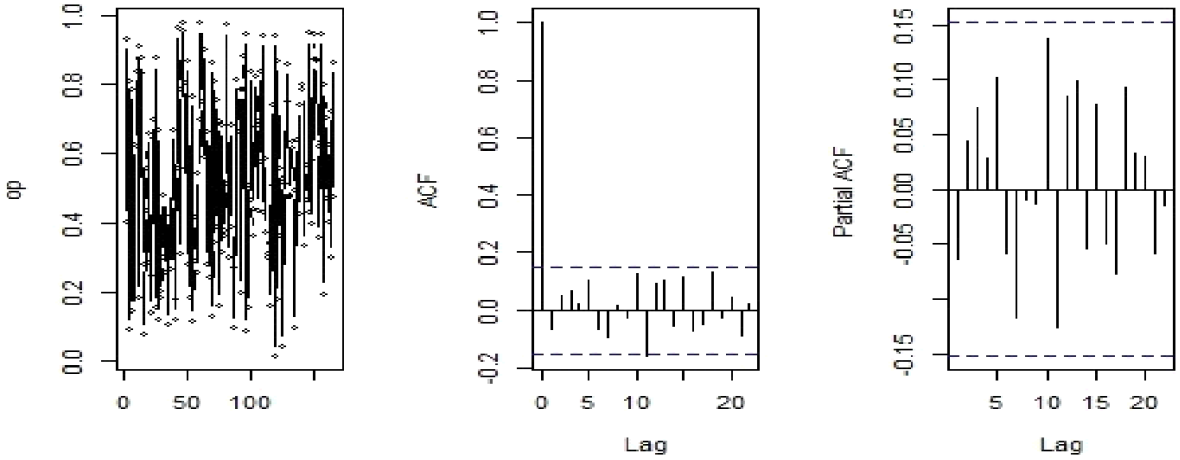

Рисунок 3.3 – Найбільший (*n = 166*) з тональних часових рядів має характеристики «білого шуму»

Так як довжини колекцій меседжів і, як наслідок, довжини часових рядів сентимент-оцінок, розподілені вкрай нерівномірно, мають характеристики «білого шуму» і основна маса з них має вкрай малу довжину, застосування відомих методів кластеризації часових рядів, наприклад, на основі алгоритму DTW [13], є малоефективним.

3.3 Класифікація часових рядів сентимент-оцінок на основі PCA-розкладу коротких ділянок

Суть пропонованого методу полягає в тому, щоб розглядати досліджувані

часові ряди через рухоме вікно фіксованої ширини w (у нашому експерименті $w = 10$) і оцінювати динаміку ряду в рамках цього вікна, попередньо ввівши ряд простих агрегатів, що характеризують міри центральної тенденції та міри розсіювання на даній ділянці ряду: X_1 – вибіркове математичне сподівання (*mean*), X_2 – вибіркова медіана (*median*), X_3 – стандартне відхилення (*sd*), X_4 – мінімальне значення вибірки (*min*), X_5 – максимальне значення вибірки (*max*) і X_6 – розмах (*range*). Образ $(x_1, x_2, x_3, x_4, x_5, x_6)$ навмисно містить робасні і не робастні характеристики, що корелюють між собою з метою оцінки впливу окремих викидів.

У процесі експерименту були виконані оцінки образу $(x_1, x_2, x_3, x_4, x_5, x_6)$ з подальшою його редукцією методом головних компонент (PCA) [14] і візуалізацією на дві перші головні компоненти Y_1, Y_2 для всієї колекції TETS. PCA-модель є лінійною комбінацією агрегатів і у нашому випадку має вигляд (3.1):

$$\begin{aligned} Y_1 &= a_{11}X_1 + a_{12}X_2 + \cdots + a_{16}X_6 \\ Y_2 &= a_{21}X_1 + a_{22}X_2 + \cdots + a_{26}X_6 , \end{aligned} \tag{3.1}$$

де $(a_{i1}, a_{i2}, \ldots, a_{i6})$ – навантаження Y_i, $i = \overline{1,2}$ (ваги, на які слід помножити кожну стандартизовану оригінальну змінну, щоб отримати значення головних компонент). Головна компонента Y_i – це лінійна комбінація, яка має максимальну дисперсію, за умови обмеження, що вектор коефіцієнтів має одиничну довжину, тобто $\sum_{j=1}^{6} a_{ij}^2 = 1$. Якщо матриця коваріації X дорівнює Σ, то дисперсія Y_i, $i = \overline{1,2}$ дорівнює

$$Var(Y_i) = a'_i \Sigma a_i = \lambda_i, i = \overline{1,2} . \tag{3.2}$$

За цією моделлю коефіцієнти a відповідають власним векторам Σ, тоді як дисперсія Y дорівнює власним значенням λ_i.

У процесі експерименту були послідовно отримані PCA-декомпозиції для кількох перших вікон шириною $w = 10$ і кроком 10. При цьому віконна обробка здійснювалася паралельно для всієї колекції наборів часових рядів сентимент-оцінок OP $(g = 2542)$. Потім досліджувалися характеристики головних компонент. Результати показали, що дві перші головні компоненти дають можливість адекватно описати сентимент-ситуацію в обраному вікні і класифікувати ситуації з наявністю великої кількості негативних або позитивних висловлювань.

У табл. 3.2 наведено приклади оцінки перших трьох головних компонент, отриманих для всіх 735-ох із TETS, довжина яких більше за 20 для другого вікна (номера елементів ряду $i = 11..20$). Видно, що дві головні компоненти описують близько 85% загальної дисперсії, що цілком достатньо для практичних задач.

Таблиця 3.2 – Власні значення та загальний відсоток поясненої дисперсії

Компоненти	Власні значення (λ_i)	Відсоток дисперсії	Сукупний відсоток дисперсії
Y_1	**2.7542065**	**45.9034419**	**45.90344**
Y_2	**2.3807033**	**39.6783890**	**85.58183**
Y_3	0.5973904	9.9565064	95.53834

Аналіз структури навантажень головних компонент (табл. 3.3) показав, що перша головна компонента Y_1 характеризує в основному варіацію сентимент-оцінок, в той час як друга Y_2 – в основному їх середнє значення. Ця особливість дає можливість легко візуалізувати і інтерпретувати результати (рис. 3.4).

У даному випадку ми бачимо, що тональні оцінки для всієї колекції в

досліджуваному діапазоні не сильно відрізняються одна від одної, тому точки утворюють один щільний сегмент. При цьому інтерес представляють окремі викиди. Структура основних компонент така, що до верхньої напівплощини потрапляють позитивні тональні оцінки, а до нижньої – негативні.

Таблиця 3.3 – Навантаження

	Y_1	Y_2
x_1, *mean*	0.5859609	**0.7795309**
x_2, *median*	0.4954367	**0.7341407**
x_3, *sd*	**0.7938915**	-0.4707043
x_4, *min*	-0.1605115	**0.8639686**
x_5, *max*	**0.8933787**	0.1270330
x_6, *range*	**0.8433539**	-0.4999305

На рис. 3.4 викид відповідає ділянці одного зі спостережень колекції, в якому присутні меседжі з яскраво вираженою позитивною тональністю. Нейтральні коментарі групуються в районі осі абсцис.

Таким чином, такий підхід дає можливість здійснювати моніторинг будь-якої кількості постів або каналів і, отримуючи послідовні «snapshots» для даних вікон TETS та виконувати їх класифікацію за трирівневою шкалою, звертаючи особливу увагу на викиди, чи, навіть, появу невеликих кластерів з позитивною чи негативною тональністю, і відстежувати таким чином динаміку тональності думок клієнтів у часі.

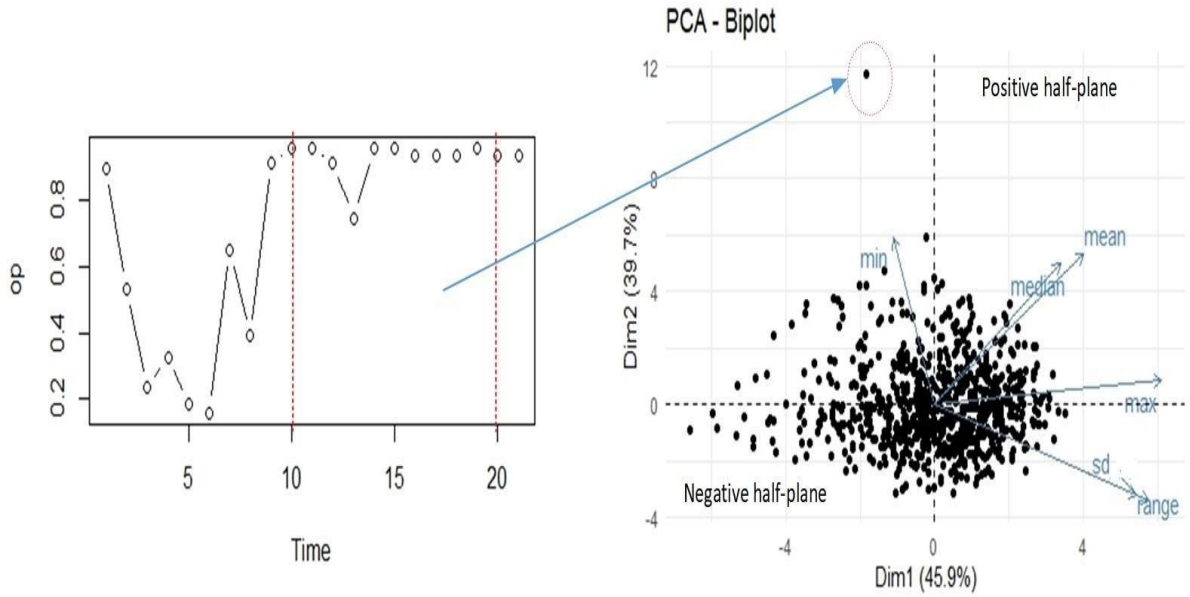

Рисунок 3.4 – Візуалізація PCA-декомпозиції на дві перші головні компоненти для другого вікна для 735 тональних оцінок часових рядів, довжина яких перевищує 20. Викид на графіку відповідає частині TETS з високими позитивними оцінками тональності. Тут $\mathrm{Dim1} - Y_1$, $\mathrm{Dim2} - Y_2$. Стрілки показують внесок змінних у головні компоненти (див. табл. 3.3)

У той же час, з’ясовувати, які властивості об’єкта викликали емоційну реакцію автора думки, логічно відштовхуючись саме від оцінки сентимент-ситуації безпосередньої думки.

4 АНАЛІЗ І СИНТЕЗ ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ

4.1 Синтез методу класифікації часових-рядів сентимент-оцінок на основі аналізу динаміки коротких ділянок

Проведені теоретичні та експериментальні дослідження дають можливість виконати синтез методу класифікації часових рядів сентимент-оцінок, який може бути представлений у вигляді структурно-функціональної схеми (рис. 4.1).

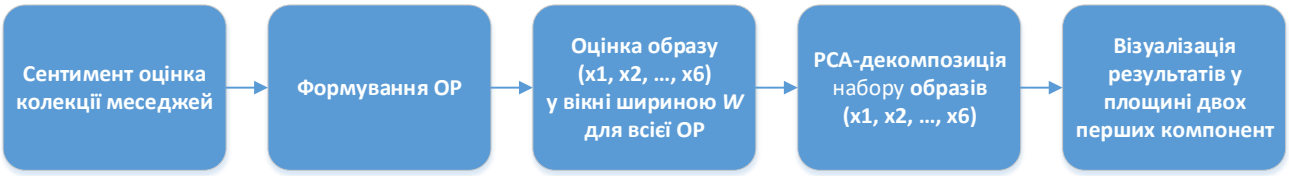

Рисунок 4.1 – Структурно-функціональна схема методу класифікації часових-рядів сентимент-оцінок на основі аналізу динаміки коротких ділянок

Після етапу парсингу і сентимент оцінки колекції текстових меседжей із заданої кількості акаунтів, що піддаються моніторингу, формується колекція часових рядів сентимент оцінок довжиною, що дорівнює наперед заданою шириною вікна і виконується оцінка образу, складовими якого є шість агрегатів (див. вище), що характеризують поведінку ряду всередині вікна. Після того, виконується PCA-декомпозиція набору даних образів с послідуючою візуалізацією результатів у площині двох перших компонент, що дає можливість легко класифікувати меседжи за тональністю тексту і виявити наявність різко-позитивних та різко-негативних за тональністю думок авторів.

Даний метод є складовою частиною модулю інтелектуального аналізу та візуалізації даних (див. рис. 4.2).

4.2 Структурно-функціональна схема інформаційної технології моніторингу і сентимент-аналізу

Нижче зроблено спробу запропонувати один з можливих варіантів реалізації інформаційної технології, в основі якої лежить вищеописаний метод з використанням сучасних і доступних програмно-апаратних засобів. Такий підхід дає можливість використання даного підходу на практиці широкому колу користувачів з різних сфер.

В основі інформаційної технології сентимент-аналізу текстового контента із соціальних мереж на основі класифікації часових рядів сентимент-оцінок лежить клієнт-серверна архітектура. Призначенням технології є моніторинг сентимент-компоненти вибраних інтернет-ресурсів у режимі реального часу.

Для реалізації клієнтської частини автори використали можливості мови Kotlin для створення мобільного додатку. Серверна частина використовує можливості AWS платформи для створення сервісів для обробки і візуалізації даних у режимі реального часу.

Серверна частина додатку використовує AWS API Gateway для забезпечення додатку двостороннього зв'язку в реальному часі, AWS Lambda – для збору та обробки даних, DynamoDB – для збереження записів у базі даних NoSQL.

У даному проекті пропонується використання двох типів лямбда-функцій. Перший, для збору даних, реалізується на Node.js з використанням API сервісів та технологій безголового браузера (Puppeteer) для сервісів, які не мають можливості представити необхідні дані через API. Другий тип, для обробки даних, реалізується засобами мови Python. Цей модуль представляє основу проекту, де виконуються процедури очищення та маніпуляції даними, розрахунку сентимент-оцінок та проводиться інтелектуальний аналіз даних.

Дані зберігаються у таблиці DynamoDB після збору в “сирому” вигляді. Після очищення та обчислення сентимент-оцінок ми отримуємо кортеж

безпосередніх думок, який також зберігається у DynamoDB для подальшої комфортної роботи.

Для зв’язку між лямбда-функціями використовується технологія AWS Step Functions.

На рис. 4.2 представлено структурно-функціональну схему інформаційної технології. Нижче описані основні її етапи.

1. Здійснення входу до мобільного додатку.
2. Здійснення реєстрації та авторизації користувача.
3. AWS API Gateway являє собою посередника між користувацькою частиною та бекендом сервісу, а також між модулями технології.
4. На даному етапі користувачу буде запропоновано зберегти дані акаунтів користувача відповідно до сервісів, які підтримуються технологією, та вибрати сервіси й публікації, що підлягають моніторингу. Дана процедура має сенс при першому вході користувача у систему, далі за необхідності користувач може змінити, додати чи припинити збір даних для своїх акаунтів.
5. База даних DynamoDB для збереження інформації у NoSQL базу.
6. AWS Step Functions являє собою стек або послідовність виконання лямбда-функцій для забезпечення логіки технології.
7. На даному етапі має бути запущений збір даних для вибраних сервісів та публікацій із збереженням їх у базу даних.
8. Модуль обробки та формування кортежу думок, який також буде збережений у БД. Для оцінки сентимент-складової використовуються ресурси проекту INDICO [11] і на основі результатів формується колекція сентимент-оцінок часових рядів.
9. Модуль для вибірки даних з БД.
10. Модуль для інтелектуального аналізу та візуалізації даних.

Користувацьку частину пропонується представити у базовому вигляді як динамічний емоджі. Також є розгорнутий вигляд ситуації, у якому можна побачити необхідні для детального аналізу дані у таблиці та візуалізації у

вигляді графіків. Технологія також дозволяє побачити звіт роботи по декільком акаунтам у паралельних графіках із зручним способом переключення та моніторингу.

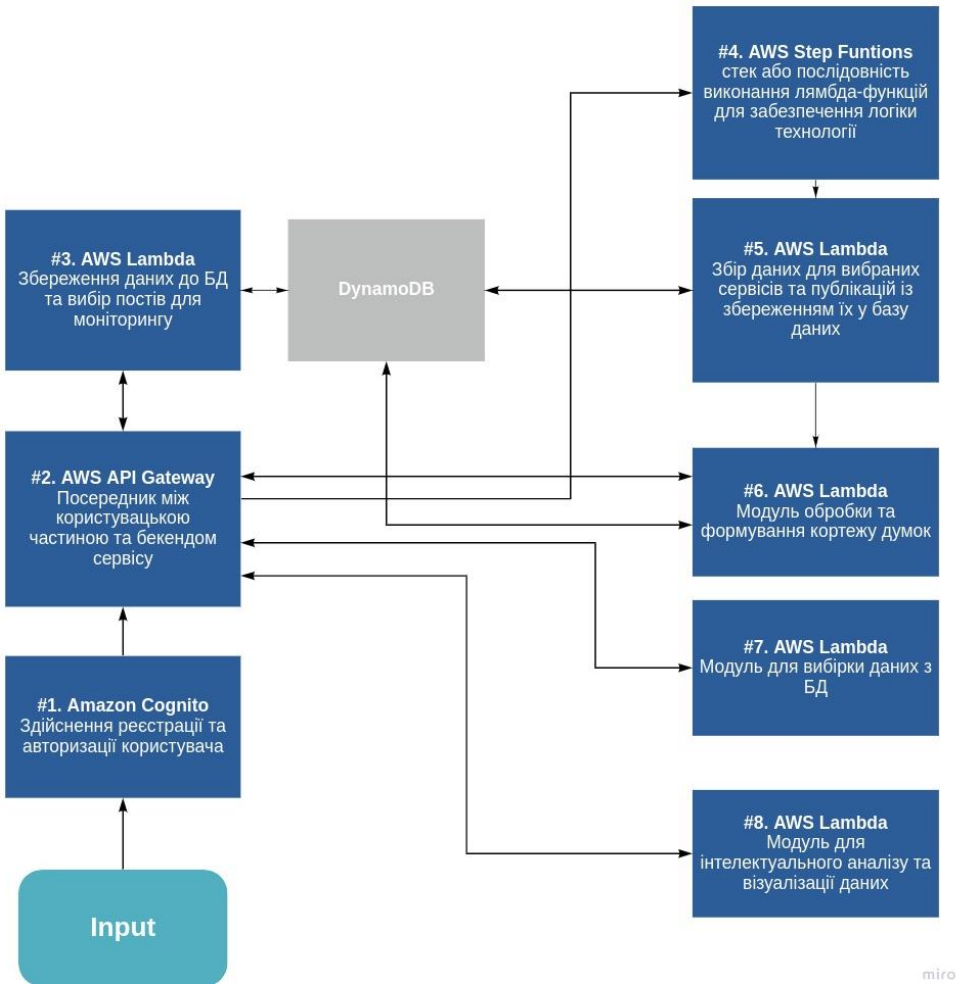

Рисунок 4.2 – Структурно-функціональна схема інформаційної технології

Однією з головних переваг запропонованого підходу на основі AWS Lambda є те, що, об’єм обчислювальних ресурсів, необхідних для вирішення задачі, визначається виключно системою, а не самим користувачем, і програмний застосунок активується тільки за умови настання певної події, або згідно з заздалегідь встановленим розкладом. Це суттєво здешевлює вартість використання сервісу.

ВИСНОВКИ

1. Введено поняття часового ряду сентимент-оцінок набору месседжей, які виникають у результаті дискусії в соціальній мережі з приводу певної публікації. На колекції рядів, розміром 2542 показано, що дані часові ряди в загальному випадку мають характеристики “білого шуму” з середньою довжиною близько 23, мають вкрай нерівномірний розподіл по довжині і до них не може бути застосовані відомі методи класифікації часових рядів для визначення характеру їх динаміки.

2. Показано, що в умовах суттєвої нерівномірності довжин часових рядів сентимент-оцінок безпосередніх думок, що генерується в акаунтах соціальних мереж і відсутності явних закономірностей в їх структурі, доцільно аналізувати короткі їх ділянки («вікна») з використанням простого способу, що описують центральну тенденцію і розсіювання всередині «вікна».

3. Запропоновано метод сентимент-аналізу текстового контенту із соціальних мереж, який, на відміну від існуючих, дозволяє здійснювати ефективний моніторинг динаміки тональності авторів думок одночасно великої кількості каналів текстової інформації та візуалізувати результати простим і зрозумілим способом. В основі методу лежить ідея класифікації часових рядів сентимент-оцінок за короткими часовими ділянками (вікнами), шляхом послідовної PCA-декомпозиції вектору простих статистичних оцінок, які характеризують динаміку ряду всередині короткого «вікна». Даний метод взято за основу інформаційної технології моніторингу авторів безпосередніх думок, який може бути застосований як для використання людиною, так і автоматизованими пристроями на основі штучного інтелекту.

4. Запропонована інформаційна технологія і метод, який покладено в його основу, на відміну від існуючих, дозволяє розв’язувати задачу моніторингу безпосередніх думок в умовах малої кількості даних, коли побудова більш складних моделей прогнозування динаміки є неможливою або недоцільною.

5. Як перспективне дослідження передбачається розглянути застосування даного методу до розширеної думки [4] з перспективою розробки підходів до управління сентимент-компонентою в процесі формування безпосередніх думок у задачах інформаційного впливу і протиборства.

ПЕРЕЛІК ПОСИЛАНЬ

1. Sergeeva, Y. (2018, March). Social Networking in 2018: A Global Study. Retrieved October 20, 2019, from https://www.web-canape.ru/business/socialnye-seti-v-2018-godu-globalnoe-issledovanie/

2. Glyoza, O. (2017, April). Infographics: how much time users spend on social networks. Retrieved October 20, 2019, from http://mmr.ua/show/infografika_skolyko_vremeni_my_provodim_v_sotssetyah#483504973.1523294675

3. Emoticonr. (Copyright © 2014-2017). All emoticons and smileys. Retrieved October 20, 2019, from http://www.emoticonr.com/emoticons

4. Ryichok, Y., Kravchenko, S., Sydorenko, V. (2018). A direct opinion model based on in-depth semantic and sentiment analysis of unstructured content in monitoring tasks of social networks. Proceedings of the XXV International scientific conference of young scientists and researches «Topical problems of vital functions of society», 24-25 April (pp. 45-46). Kremenchuk, Ukraine: Kremenchuk Mykhailo Ostrohradskyi National University.

5. Liu, Bing (2010). "Sentiment Analysis and Subjectivity". Handbook of Natural Language Processing (Second ed.). Editors: In Indurkhya, N.; Damerau, F. J.

6. Pang, B., Lee, L. (2008). Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval (2), 1-135. DOI: 10.1561/1500000001

7. Turney, P. (2002). Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. Proceedings of the 40th Association for Computational Linguistics, (ACL), July 2002 (pp. 417–424). Philadelphia, USA.

8. Pang, B., Lee, L. & Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP), July 2002 (pp 79–86). Philadelphia, USA: Association for Computational Linguistics.

9. Snyder, B., Barzilay, R. (2007). Multiple Aspect Ranking using the Good Grief Algorithm. Proceedings of the Joint Human Language Technology/North American Chapter of the ACL Conference (HLT-NAACL), April 2007 (pp. 300–307). Rochester, New York: Association for Computational Linguistics.

10. LiveJournal (Copyright © 2019). Users ratings: Cyrillic. Retrieved October 20, 2019, from https://www.livejournal.com/ratings/users/authority/?country=cyr&page=

11. Wilde, T. (Copyright © 2019 Indico Data Solutions, INC.). Sentiment Analysis. Retrieved October 20, 2019, from https://indico.io/blog/docs/indico-api/text-analysis/sentiment-analysis/

12. Keogh, E., Pazzani, M. (1998, June). Synthetic Control Chart Time Series. The UCI KDD Archive Information and Computer Science University of California, Irvine Retrieved October 27, 2019, from http://kdd.ics.uci.edu/databases/synthetic_control/synthetic_control.html

13. Al-Naymat, G., Chawla, S., Taheri, J. (2008) Sparse DTW: A novel approach to speed up Dynamic Time Warping. Proceeding AusDM '09 Proceedings of the Eighth Australasian Data Mining Conference December 01 - 04, 2009 (pp. 117-127). Melbourne, Australia: Australian Computer Society, Inc. Darlinghurst.

14. Dubrov, A. (1978). Processing of statistical data by the method of principal components. Moscow, USSR: Statistics.

15. Усталов Д.А. Извлечение терминов из русскоязычных текстов при помощи графовых моделей//Теория графов и приложения: Graph theory and Applications: материалы конференции. Екатеринбург, 2012.

16. Hayashi, Y. A Dictionary Model for Unifying Machine Readable Dictionaries and Computational Concept Lexicons [Text] / Y. Hayashi, T. Ishida // LREC / Y. Hayashi, T. Ishida., 2006. – . 1–6.

17. SocialMedia Examiner (Copyright © 2019). Social Media Marketing Industry Survey Report – Digital Marketing 2019 from https://www.socialmediaexaminer.com

18. Прокопенко А.Н., Савотченко С. Е., Старостенко И. Н. Особенности осуществления правоохранительными органами мероприятий по противодействию информационным угрозам в социальных сетях. Вестник краснодарского университета МВД России №4. Краснодар, 2018.

19. Thelwall, Mike; Buckley, Kevan; Paltoglou, Georgios; Cai, Di; Kappas, Arvid (2010). "Sentiment strength detection in short informal text". Journal of the American Society for Information Science and Technology

20. М. В. Клековкина, Е.В. Котельников. Метод автоматической классификации текстов по тональности, основанный на словаре эмоциональной лексики (рус.) // RCDL-2012, Переславль-Залесский, Россия : конференция. — 2012.

21. Cortes, Corinna; Vapnik, Vladimir N. (1995). "Support-vector networks". Machine Learning

22. Акобир Ш. Деревья решений − общие принципы работы: [Електроний ресурс]. − (https://basegroup.ru/community/articles/description).

23. Классификация методом максимальной энтропии [Електроний ресурс]. - (http://bazhenov.me/blog/2013/04/23/maximum-entropy-classifier.html)

24. Zhang Y., Wallace B. A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification //arXiv preprint arXiv:1510.03820. — 2015.

25. Sydorenko, V., Ryichok, Y., Kravchenko, S. (2019). Method of classification of tonal estimations time series in problems of intellectual analysis of text content. : Transportation Research Procedia. Proceedings of the 2019 Conference on LOGI 2019 – Horizons of Autonomous Mobility in Europe (LOGI), November 2019 (pp ??–??)., Czech, České Budějovice: Institute of Technology and Business in České Budějovice.

26. Рычок Ю. С., Сидоренко В. Н. Сбор и обработка неструктурированного текстового контента из социальных сетей в задачах сентимент-анализа. Матеріали XXVI міжнародної науково-технічної

конференції студентів, аспірантів та молодих учених «Актуальні проблеми життєдіяльності суспільства», 24–25 квітня, 2019 р., м. Кременчук. – с. 24

27. Кравченко С. А., Сидоренко В. Н. Оценка атрибутов автора непосредственного мнения в задаче сентимент-анализа текстового контента из социальных сетей. Матеріали XXVI міжнародної науково-технічної конференції студентів, аспірантів та молодих учених «Актуальні проблеми життєдіяльності суспільства», 24–25 квітня, 2019 р., м. Кременчук. – с. 24

ДОДАТОК А

Матеріали впровадження наукової роботи

ЗАТВЕРДЖУЮ

Директор ТОВ «ЗЕ Грейдієнт»

Д. О. Скрипник

«01» січня 2020 р.

АКТ

впровадження результатів студентської наукової роботи Кравченко С. А. та Ричка Ю. С. **«Метод та інформаційна технологія сентимент-аналізу текстового контента із соціальних мереж на основі класифікації часових рядів сентимент-оцінок»**

Розглянуті результати студентської наукової роботи Кравченко С. А. та Ричка Ю. С. «Метод та інформаційна технологія сентимент-аналізу текстового контента із соціальних мереж на основі класифікації часових рядів сентимент-оцінок» прийняті для впровадження у процес розробки прикладного програмного забезпечення у вигляді прототипу програмного забезпечення, який реалізує розроблений метод та інформаційну технологію.

Застосування розробленого методу та інформаційної технології дозволить підвищити ефективність роботи аналітичного модулю розроблюваного програмного забезпечення і розширити функціонал програмного застосунку, призначеного для моніторингу соціальних мереж.

Менеджер проекту О. Е. Гасьошин