ДЕРЖАВНИЙ УНІВЕРСИТЕТ ТЕЛЕКОМУНІКАЦІЙ

Навчально–науковий інститут Інформаційних технологій

Кафедра Інженерії програмного забезпечення

Пояснювальна записка

до магістерської роботи
на ступень вищої освіти магістр

на тему «**ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ДЛЯ ОБРОБКИ ПОВІДОМЛЕНЬ TWITTER НА ОСНОВІ МЕТОДІВ СЕНТИМЕНТ АНАЛІЗУ»**

Виконав: студент 6 курсу, групи ПДМ - 62
спеціальності
121 Інженерія програмного забезпечення
(шифр і назва спеціальності)
Чуб Євгеній Михайлович
(прізвище та ініціали)
Керівник Трінтіна Н.А.
(прізвище та ініціали)
Рецензент
(прізвище та ініціали)

КИЇВ – 2023

ДЕРЖАВНИЙ УНІВЕРСИТЕТ ТЕЛЕКОМУНІКАЦІЙ

НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

Кафедра Інженерії програмного забезпечення
Ступінь вищої освіти -«Магістр»
Спеціальність підготовки – 121 «Інженерія програмного забезпечення»

ЗАТВЕРДЖУЮ
Завідувач кафедри
Інженерії програмного забезпечення
_____ Негоденко О.В.
"____" _____2022 року

З А В Д А Н Н Я
НА МАГІСТЕРСЬКУ РОБОТУ СТУДЕНТА

ЧУБА ЄВГЕНІЯ МИХАЙЛОВИЧА
(прізвище, ім'я, по батькові)

1. Тема роботи: «Інформаційна технологія для обробки повідомлень Twitter на основі методів сентимент аналізу»
Керівник роботи: Трінтіна Н.А. к.т.н., доц., доцент кафедри ІПЗ
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

Затверджені наказом вищого навчального закладу від «12» жовтня 2022 року №122.
2. Строк подання студентом роботи «30» грудня 2022 року
3. Вхідні дані до роботи
Науково-технічна література з питань, пов'язаних з програмним забезпеченням щодо розробки та проведення математичних досліджень

4. Зміст розрахунково-пояснювальної записки(перелік питань, які потрібно розробити).
4.1 Аналіз наявних математичних методів та технологій для розробки систем аналізу тональності тексту.
4.2 Дослідження інформаційних технологій за темою.
4.3 Реалізація інформаційної системи для аналізу тональності тексту.
4.4 Опис проектування системи.
4.5 Опис використаних технологій.

5. Перелік демонстраційного матеріалу (назва основних слайдів)

1. Титульний слайд
2. Мета, об’єкт та предмет дослідження
3. Актуальність роботи
4. Аналіз існуючих рішень
5. Результати дослідження
6. Апробація результатів дослідження
7. Висновки

6. Дата видачі завдання «14» жовтня 2022

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів бакалаврської роботи	Строк виконання етапів роботи	Примітка
1	Вивчення теми магістерської роботи	14.10.2022	Виконано
2	Вивчення літературних джерел	17.10.2022	Виконано
3	Складання плану роботи	22.10.2022	Виконано
4	Узгодження плану роботи та списку використаних джерел з науковим керівником	01.11.2022	Виконано
5	Аналіз існуючих математичних рішень	03.11.2022	Виконано
6	Дослідження інформаційних технологій	07.11.2022	Виконано
7	Розробка модифікованої математичної моделі та проведення дослідження	13.11.2022	Виконано
8	Вступ, висновки, реферат	03.12.2022	Виконано
9	Розробка обов‘язкових демонстраційних матеріалів	12.12.2022	Виконано
10	Попередній захист роботи	22.12.2022	
11	Здача роботи		

Студент _____ Чуб Є. М.
(підпис) (прізвище та ініціали)

Керівник роботи _____ Трінтіна Н. А.
(підпис) (прізвище та ініціали)

РЕФЕРАТ

Текстова частина бакалаврської роботи 71 с., 13 рис., 1 фор., 7 табл. , 36 джерел.

Об'єкт дослідження – процес аналізу тональності тексту.

Предмет дослідження – методи сентимент аналізу.

Мета роботи – вдосконалення наявних методів використання алгоритмів сентимент аналізу для аналізу неструктурованих текстових даних.

Результати проведеного наукового дослідження дають зрозуміти, що використані математичні методи, алгоритмічні моделі, та програмні засоби надають можливість вдосконалити процес аналізу полярності тексту.

Практичні значення отриманих результатів полягає в тому, що на основі проведених теоретичних досліджень було розроблено нову методику обробки повідомлень на основі синсетів WordNet у поєднанні з методами машинного навчання та обробки природної мови. Результати цього дослідження надалі дають змогу розробити нові інструменти та методи обробки повідомлень з урахуванням результатів цього дослідження, перш за все – механізмів запобігання появи спотворень (артефактів) інформації при витягу з Twitter.

ЗМІСТ

ВСТУП

Актуальність дослідження. У відповідь на зростаючу доступність неструктурованих текстів, таких як дописи в блогах та соціальних мережах і вебсайт з оглядами продуктів, за останнє десятиліття виникла галузь аналізу настроїв, щоб відповісти на питання, що люди думають про певну тему. Об’єднуючи дослідників у галузі інформатики, комп’ютерної лінгвістики, інтелектуального аналізу даних, психології та навіть соціології, сентимент-аналіз (аналіз настроїв) розширює традиційний аналіз тексту на основі фактів, щоб створити інформаційні системи, орієнтовані на думку та погляди.

Проводячи серйозні дослідження або приймаючи повсякденні рішення, ми часто шукаємо думки інших людей. Ми консультуємося на форумах політичних дискусій, коли голосуємо за політику, читаємо звіти споживачів, коли купуємо техніку, просимо друзів порекомендувати ресторан або фільм на вечір. У цьому контексті Інтернет (та Twitter зокрема) дав змогу дізнатися думки мільйонів людей про все, від останніх гаджетів до політичних філософій. При цьому широка доступність до цих даних створила нову сферу в аналізі тексту, розширивши предмет дослідження від традиційного фактоцентричного та інформаційно-орієнтованого погляду на текст на емоційноцентричний.

У цьому контексті за останнє десятиліття витяг настрою з тексту привернув багато уваги як у промисловості, так і в наукових колах. Дедалі більше компаній та некомерційних інституцій (особливо державних установ) усвідомлюють важливість думки користувачів Інтернету про продукти, послуги, ситуації та особистостей тощо.

Однак, проблеми розвитку нових методів витягу даних та їх аналізу на основі методів сентимент-аналізу, пов'язані з надмірною складністю процесу підготовки та обробки тексту, спотворенні та ненадійності інформації зумовлюють необхідність трансформації наявних підходів до обробки та аналізу тексту. В цьому

контексті, актуальним стає дослідження методів сентимент-аналізу, їх класифікації, проблем функціонування та потенціалу до розвитку як інструменту обробки повідомлень у соціальних мережах, зокрема твітів на соціальній платформі Twitter.

Мета дослідження – вдосконалення наявних методів використання алгоритмів сентимент аналізу для аналізу неструктурованих текстових даних.

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

- встановити сутність поняття «сентимент-аналізу» і визначити етапи становлення та класифікацію методів;
- висвітлити найбільш поширені інструменти та програми стенографії для проведення аналізу настроїв в Інтернеті;
- проаналізувати обмеження Twitter API як потенційної бази для аналізу настроїв;
- розробити схему нової методики обробки повідомлень на основі сентимент-аналізу у поєднання з методами машинного навчання;
- проаналізувати настроїв повідомлень у Twitter на тему війни в Україні;
- виокремити основні переваги розробленого підходу та шляхи подальшої модернізації.

Об'єктом дослідження є процес аналізу тональності тексту.

Предметом дослідження є методи сентимент аналізу.

Методи дослідження. Для досягнення поставленої мети, розв’язання сформульованих завдань та отримання відповідних результатів використано сукупність загальнонаукових та спеціальних методів, які сприяли забезпеченню концептуальної єдності дослідження: методи наукової абстракції (виділення найбільш суттєвих особливостей розвитку методів сентимент-аналізу), аналізу (висвітлення переваг та недоліків Twitter API); метод формалізації (розробка методики обробки повідомлень та пропозицій щодо модернізації у майбутньому).

Інформаційно-фактологічною базою дослідження виступили спеціальні монографії, посібники та спеціалізована література, а також окремі розділи з математичної статистики, математичних і статистичних методів шифрування, загальної статистики та її методів аналізу; публікації вітчизняних та закордонних журналів, що містять елементи спостережень у сфері психології, комп'ютерної лінгвістики, обробки природної мови; комп'ютерні опитування спеціалізованими компаніями.

Зокрема, питанням розробки методів шифрування повідомлень на основі методів стенографії та процесів удосконалення подібних підходів ретельно досліджено зарубіжними та вітчизняними спеціалістами: Х. Дугласом, Г. Браяном, А. В. Васильковим, Є. Б. Герасимовою, Дж.Ф.Девліном, Чжаном А. Ву Р, Цай та іншими.

Наукова новизна одержаних результатів полягає в тому, що на основні наукових здобутків проведено дослідження підходів до обробки повідомлень на основі методів сентимент-аналізу та пов'язаних з ними процесів і визначено особливості їх застосування не тільки з технічної точки зору (використання алгоритмів для обробки повідомлень), а й виокремлено цільові функції існування різноманітних методів аналізу та їх недоліки і переваги..

Практичне значення одержаних результатів дослідження полягає в тому, що на основі проведених теоретичних досліджень було розроблено нову методику обробки повідомлень на основі синсетів WordNet у поєднанні з методами машинного навчання та обробки природної мови. Результати цього дослідження надалі дають змогу розробити нові інструменти та методи обробки повідомлень з урахуванням результатів цього дослідження та міжнародного досвіду, перш за все – механізмів запобігання появи спотворень (артефактів) інформації при витягу з Twitter.

РОЗДІЛ 1. ТЕОРЕТИЧНІ ОСНОВИ ОБРОБКИ ІНФОРМАЦІЇ НА ОСНОВІ МЕТОДІВ СЕНТИМЕНТ-АНАЛІЗУ

1.1 Генезис становлення поняття сентимент-аналізу

Як галузь дослідження сентимент-аналіз тісно пов’язаний з комп’ютерною лінгвістикою, обробкою природної мови та аналізом тексту і навіть може розглядатися як частина кожного з цих сфер дослідження. Виходячи з вивчення афективного стану (психологічна теорія) і судження (теорія оцінки), ця сфера прагне відповісти на питання, які давно вивчаються в інших сферах дискурсу, використовуючи нові інструменти, надані аналізом даних і комп’ютерною лінгвістикою. Аналіз настроїв мав і досі має багато назв. Його часто називають аналізом суб’єктивності, дослідженням думок і вилученням оцінок (що дещо пов’язує його з емоційним обчисленням, комп'ютерним розпізнаванням та вираження емоцій).

Одним із завдань аналізу настроїв є визначення об’єктів дослідження – думок і суб’єктивності. Спочатку суб’єктивність була визначена лінгвістами, найвідомішим з яких є Рендольф Квірк.

Так, Доценко І. У своїй роботі цитуючи саме Квірка визначав суб’єктивність як приватний стан, як щось, що не піддається об’єктивному спостереженню чи перевірці. Ці приватні стани охоплювати серед іншого емоції, думки та спекуляції. Однак важливо зауважити, що суб’єктивне за Квірком не означає неправдиве. Наприклад, речення «Марія любить шоколад» виражає ставлення Марії до шоколаду, але воно не означає, що інформація неправдива. Так само не всі об’єктивні речення та інформація є правдивими [8].

Щоб підкреслити цю неоднозначність концепції суб’єктивності, надалі Доценко у своєму дослідженні перераховує визначення термінів, тісно пов’язаних із поняттям настрою [8]:

- думка передбачає продуманий висновок, але відкритий для суперечок («кожен експерт, здавалося, мав іншу думку»);
- огляд (погляд) передбачає суб'єктивну думку («дуже наполегливий у висловленні своїх поглядів»);
- віра часто передбачає навмисне прийняття та інтелектуальну згоду («тверда віра в програму своєї партії»);
- переконання стосується твердого й серйозного ставлення до ситуації чи ідеї («переконання, що життя тварин таке ж священне, як і людське»);
- упередженість передбачає віру, яка ґрунтується на впевненості («був переконаний, що ситуація змінюється на краще»);
- почуття передбачає усталену думку, що відображає певну емоцію («її феміністичні почуття добре відомі»).

Надалі, Блейк, відомий дослідник обробки природної мови (NLP), використовував визначення приватного стану Квірка, відстежуючи точки зору (погляди) в наративі. Для цього вона визначили приватний стан Квірку як трикомпонентний кортеж, який складається з [14]:

- досвіду;
- ставлення;
- об’єкту.

Безпосередньо приватний стан у цій моделі пов’язує людину з його або її ставленням до певному об'єкту. Однак, на практиці, при проведенні сентимент-аналізу зазвичай використовується спрощена версія цієї моделі, де ми розглядаємо лише полярність і ціль почуття. Так насправді багато дослідників визначають настрої вільно, як негативну або позитивну думку. Почуття також має кілька унікальних властивостей, які відрізняють його від інших якостей, які ми можемо відстежувати та аналізувати у тексті. Наприклад, ми можемо класифікувати почуття у тексті не тільки за настроями, а і за темами, що може передбачати роботу з цілими таксономіями тем.

Крім того, сентимент-аналіз зазвичай вивчає і менші суб'єктивні елементи, визначені Блейк як «лінгвістичні вираження приватних станів у контексті». Зазвичай це окремі слова, фрази чи речення. І хоча іноді цілі тексти та документи можуть вивчатися як одиниця настрою, загальновизнаним залишається теорія, що настрої містяться в менших мовних одиницях [14].

Оскільки настрої та думки часто стосуються однієї ідеї, у цій роботі надалі ці терміни використовуватимуться як синоніми. Настрої, які з’являються в тексті, бувають двох видів:

- експліцитні, коли суб’єктивне речення прямо виражає думку («Це чудовий день!»);
- імпліцитні, коли текст має на увазі думку («Навушник зламався за два дні»).

Більшість робіт та досліджень у галузі сентимент-аналізу зосереджена на першому типі настроїв, оскільки його легше аналізувати внаслідок чітко визначеної полярності настроїв (позитивне чи негативне ставлення до об'єкта). Але полярність також можна розглядати як діапазон. Наприклад, повідомлення, що містить кілька тверджень, що ґрунтуються на думках, зазвичай має комплексну та змішану полярність. Крім того, слід розрізняти полярність почуття та його силу. Хтось може бути переконаним, що продукт (наприклад, телефон) є не дуже хорошим чи поганим, оскільки, можливо, володів ним занадто короткий час, щоб сформувати тверду думку.

Іншою важливою частиною почуття є його ціль – об'єкт, концепція, ситуація, людина, будь-що. Найбільше досліджень в цьому контексті виконано над рецензіями на товари та фільми, де легко визначити тему тексту. Але часто варто звернути увагу на те, про яку особливість цього об’єкта говорить автор. Згадування цих функцій у тексті також може бути явним («Робота оператора у фільмі чудова») або неявним («План сцени прощання завеликий»). Крім того, на відміну від звичайного тематичного аналізу, авторство настрою може бути невід’ємною частиною проблеми. Політичні коментарі та новини переповнені цитатами і

можуть мати заплутану структуру, яку важко розрізнити. Наприклад, у новинній статті про політичні дебати буде поєднання цитат учасників дебатів, експертів, які коментують дебати, і, можливо, навіть позиція автора щодо проблем.

Як згадувалося раніше, деякі з текстів, які найбільше вивчаються в аналізі настроїв, – це огляди продуктів і фільмів. Перевагою є те, що вони вже мають чітко визначену тему, і часто припускається, що почуття, висловлені в оглядах, пов'язані з темою. У багатьох також є система рейтингу зірок, яка служить кількісним показником думки. Такі дані часто використовуються як золотий стандарт під час оцінки вилучення/ідентифікації настроїв. У подібній ситуації, загальним завданням, спрямованим на дослідження настроїв, було б знайти думки про певний продукт у будь-якому вебвміст. У цьому контексті кілька компаній пропонують послуги з відстеження брендів і сприйняття ринку з використанням методів аналізу настроїв. Наприклад, Еркісон у своїй роботі згадує OpSec Security, яка забезпечує «моніторинг, вимірювання та аналіз відгуків споживачів» своїм клієнтам, допомагаючи їм зрозуміти потреби ринку, цільові сегменти клієнтів і їхню позицію проти конкурентів. З іншого боку, однією з найскладніших сфер для методів аналізу настроїв є політика. Політичні дискусії рясніють цитатами, сарказмом і складними посиланнями на осіб, організації та ідеї [17].

Через складність проблеми (основні поняття, вирази в тексті тощо) сентимент-аналіз охоплює кілька окремих завдань. Зазвичай вони поєднуються, щоб отримати певні уявлення та факти про думки, які містяться в тексті.

Першим завданням є виявлення настроїв або думок, що можна розглядати як класифікацію тексту як об'єктивного чи суб'єктивного. Зазвичай визначення думки ґрунтується на перевірці прикметників у реченнях. Наприклад, полярність «це гарна картинка» можна легко визначити, подивившись на прикметник. У ранньому дослідженні Хацівасіглоу якраз розглядався вплив прикметників на суб'єктивність речення. Останні дослідження показали, що прислівники також можуть використовуватися з подібною метою [19].

Друге завдання – класифікація полярності. Враховуючи уривок тексту з думкою, мета полягає в тому, щоб класифікувати думку як таку, що підпадає під одну з двох протилежних полярностей настроїв, або визначити її позицію в континуумі між цими двома полярностями. Якщо розглядати це завдання кількісно, то класифікація полярності – це завдання бінарної класифікації, яка полягає в позначенні ярликового тексту як такого, що виражає загальну позитивну або загальну негативну думку. Більшість подібних досліджень було проведено на оглядах продукту, де визначення «позитивного» та «негативного» є чіткими. Інші завдання, такі як класифікація новин як «хороші» чи «погані», викликають деякі труднощі. Врешті решт, стаття новин може містити «погані» новини без фактичного використання будь-яких суб'єктивних термінів. Крім того, ці класи зазвичай змішуються, коли документ виражає як позитивні, так і негативні почуття. Тоді завдання може полягати в тому, щоб визначити головний настрій певного абзацу чи тексту загалом. Щоб розрізнити різні суміші двох протилежностей, у таких випадках при класифікації полярності використовується багатобальна шкала (наприклад, кількість зірок для огляду фільму). Тут завдання перетворюється на проблему категоризації тексту з кількома класами. Але на відміну від задач класифікації на основі кількох класів, де словники відрізняються для кожного класу (або трохи накладаються), словники для позитивних, нейтральних і негативних класів можуть бути дуже схожими і відрізнятися лише кількома важливими словами. Оскільки багато текстів мають «змішану» точку зору, цей клас насправді є поєднанням позитивного та негативного. Крім того, заперечення, які, як правило, ігноруються в аналізі тексту як несуттєві, відіграють важливу роль у настрої, перетворюючи початково позитивний термін на негативний і навпаки.

Наведені вище два завдання можна виконувати на кількох рівнях:

- термін;
- фраза;
- речення;
- текст (документ).

При цьому зазвичай вихідні дані одного рівня використовують як вхідні дані для вищих рівнів. Наприклад, ми можемо застосувати аналіз настроїв до фраз, а потім використовувати цю інформацію для оцінки речень, абзаців тощо. Для різних рівнів підходять різні техніки. Техніки з використанням класифікаторів n-грамів або лексиконів зазвичай працюють на рівні термінів, тоді як позначення частини мови використовується для аналізу фраз і речень. Евристика часто використовується для узагальнення настроїв на рівні всього документу.

Третім завданням, яке доповнює ідентифікацію настроїв, є виявлення мети думки. Складність цього завдання значною мірою залежить від області аналізу. Як згадувалося раніше, зазвичай можна з упевненістю припустити, що відгуки про продукт зазвичай говорять про вказаний продукт. З іншого боку, загальні тексти, такі які ми можемо знайти у соціальних мережах та блогах, не завжди мають заздалегідь визначену тему та часто згадують багато об’єктів.

Іншою жвавою сферою та завданням дослідження є виділення ознак з огляду на об’єкт або тему тексту. Фінеган у своєму дослідженні 1984 року визначив ознаки як компоненти або атрибути об’єкта, які дозволяють точніше проаналізувати настрої та більш детально підсумувати результати. Наприклад, іноді в реченнях настрою є кілька цілей, як це має місце в порівняльних реченнях. Суб’єктивне порівняльне речення впорядковує об’єкти в порядку переваг, наприклад, «цей фотоапарат кращий за мій старий». Ці речення можна ідентифікувати за допомогою порівняльних прикметників і прислівників (більше, менше, краще, довше), прикметників найвищого ступеня (найбільше, найменше, найкраще) та інших слів, таких як однакові, відрізняються, виграють, віддають перевагу тощо. Відповідно об’єкти можна розташувати в порядку, який найбільше відповідає їхнім перевагам, як описано в тексті [17].

Таким чином, тематичний та міжтематичний аналіз настроїв вивчається з метою покращення продуктивності в певній області. Важливим питанням тут є поєднання загальних знань про вираження почуттів і тих, що стосуються

конкретної теми. У кростематичному аналізі ідея полягає в тому, щоб використовувати знання, зібрані про одну область, в іншій.

1.2. Класифікація методів сентимент-аналізу

Для досягнення описаних вище цілей використовується широкий спектр інструментів і технік. При цьому багато завдань в аналізі настроїв можна розглядати як основу для унікальної класифікації методів.

Так, наприклад машинне навчання пропонує багато алгоритмів, призначених саме для сентимент-аналізу:

1. Визначення частоти терміну. Традиційні системи пошуку інформації давно підкреслити важливість частоти терміну. Показник TF-IDF (частота термінів – зворотна частота тексту) добре відомий і активно використовується в моделюванні та витязі даних з документів.

Основа цього підходу полягає в тому, що терміни, які часто зустрічаються в тексту є більш інформативними щодо того, про що йдеться в документі, порівняно з термінами, згаданими лише один раз. Однак, у сфері аналізу настроїв замість того, щоб звертати увагу на найчастіші терміни, корисніше шукати найбільш унікальні. Внаслідок чого метод визначення частоти перетворюється на метод визначення присутності терміну. Наприклад, алгоритм аналізу автоматично сигналізує про наявність обраного терміну.

2. N-грами. Позиції термінів також важливі для представлення документа для аналізу настрою. Позиція термінів визначає, а іноді змінює полярність фрази загалом. Таким чином, інформація про позицію іноді кодується у вектор ознак (або n-грами) на основі точності, обчисленої за допомогою анотованих документів. N-грами – це пара слово-основа, частина мови.

3. Визначення частини мови Як згадувалося раніше, було встановлено, що прикметники добре вказують на настрої в тексті, і в останнє десятиліття їх часто використовували. в аналізі настроїв. Це актуально і для інших областей текстового аналізу, оскільки теги частини мови можна вважати грубою формою усунення неоднозначності слів. Наприклад, Гарсія у роботі 2020 року використовував шаблони частин мови, більшість з яких охоплювала прикметник або прислівник, для виявлення почуття на рівні цілого тексту [20].

4. Визначення синтаксису. Інформацію про синтаксис також використовували в наборах функцій, хоча все ще ведуться дискусії про переваги цього методу при проведенні сентимент-аналізу. Загалом, ця інформація може охоплювати такі важливі функції тексту, як заперечення, які своєю чергою є невід'ємною частиною аналізу настроїв.

Звичайне представлення тексту за допомогою пакета слів роз'єднує всі слова та вважає такі речення, як «Мені подобається ця книга» та «Мені не подобається ця книга», дуже подібними, оскільки лише одне слово відрізняє одне від іншого. Але коли йдеться про почуття, заперечення змінює полярність цілої фрази. Заперечення часто враховуються під час постобробки результатів, тоді як оригінальне представлення тексту їх ігнорує. Хоча насправді на практиці включити заперечення у підготовчий етап, додавши їх до термінів, близьких до заперечень, що може дозволити економити ресурси та час на проведення аналізу. Хоча використання подібного спільного розташування може бути занадто грубою технікою. Було б неправильно заперечувати настрої в такому реченні, як «Не дивно, що це всім подобається». Тому для розв'язання подібних проблем використовують специфічні шаблони тегів частини мови, щоб ідентифікувати заперечення, пов'язані з полярністю настроїв фрази.

Як вже згадувалось раніше, одним з основних завдань аналізу настроїв є визначення семантичної орієнтації (полярності та об'єктивності) слова. Для цього

використовують різноманітні методи, які можна приблизно розділити на наступні категорії:

- використання лексикону, створеного вручну або автоматично;
- використання деяких статистичних методів, таких як перегляд збігу слова зі словом відомої полярності;
- використання навчальних документів, позначених чи не позначених, як джерело знань про полярність термінів у колекції.

Кожен із цих прийомів має свої переваги та труднощі, які детально обговорюватимуться далі:

1. Лексикони є фундаментальною частиною сентимент-аналізу, але не всі вони однакові. Найпростішими є ті, що мають бінарну класифікацію слів на позитивні та негативні полярності або об'єктивні та суб'єктивні. Більш точне розрізнення між класами можна зробити за допомогою нечітких лексиконів, де кожна мітка має пов'язану з нею оцінку, що передає «силу» мітки. Ще більш складний підхід полягає у прийнятті будь-якої з більш детальних афективних класифікацій, розроблених у таких областях психології, як модель емоцій Плутчика або теорій «фундаментальних» або «базових» емоцій.

Нарешті, можна використовувати навіть більш складні та комплексні методи, як-от бази знань здорового глузду, розроблені дослідниками штучного інтелекту. Наприклад, Капріке і Грмімальді вручну створили лексикон, який асоціює слова з категоріями афекту, вказуючи інтенсивність (рівень сили афекту) і центральність (ступінь спорідненості з категорією) [22]. Цей лексикон можна назвати «нечітким», оскільки він здатний обробляти неоднозначність терміна, відносячи його до кількох семантичних категорій.

Крім анотації вручну, для створення лексиконів можна використовувати інші ресурси. Існуючі лексикони можна доповнити, включивши інформацію про настрої. Наприклад, лексикон WordNet Прінстонського університету (який розглянуто більш детально у розділі 1.3) є одним із найпопулярніших інструментів,

який використовувався для аналізу настроїв. Існують також способи визначення сентиментальної орієнтації слів за допомогою статистичного аналізу великих корпусів тексту. Наприклад, використовують спільне використання слів, щоб зробити висновок про семантичну орієнтацію слів, для чого вони досліджують два методи: поточкову взаємну інформацію (PMI) та латентний семантичний аналіз (LSA). Ідея полягає в тому, що «семантична орієнтація слова прагне відповідати семантичній орієнтації його сусідів» [22].

Техніки, які дослідники використовують, насправді досить різні за своєю природою: PMI обчислює спів поширеність слів, надсилаючи запит до пошукової системи, а LSA використовує метод матричної факторизації Singular Value Decomposition для аналізу статистичного зв'язку між словами. Оскільки ресурси рідко порівнюють, досі залишається відкритим питання про те, який із них є найбільш корисним для створення анотованих лексиконів.

Після визначення семантичної орієнтації окремих слів часто бажано поширити її на фразу чи речення, у якому це слово з'являється. Один із найпростіших способів досягти цього – взяти середнє значення полярностей слів у реченні. Левченко пише: «якщо переважає позитивна/негативна думка, речення відповідно вважається позитивним/негативним» [27].

У випадку, якщо кількість слів позитивної та негативної думки однакова, вони приймають орієнтацію найближчого речення думки. Можливо також використовувати наявність у реченні слів із відомими полярностями як вказівку на те, що речення є суб'єктивне. Попри те, що ця евристика доволі спрощена, вона працює в більшості випадків. Ще більш витончена комбінація міток настрою можлива, використовуючи переваги синтаксичних зв'язків між словами. Наприклад, Віджаялакшмі використовував техніку неконтрольованої класифікації Relaxation Labeling, яка розширює мітку, приписану слову, до речення, у якому воно з'являється. Цей підхід використовує, серед іншого, модифікатори заперечення, значення який обговорювалось вище [30].

Ще один метод сентимент-аналізу базується на нещодавньому відкритих техніках машинного перекладу. Використовуючи механізм перекладу з японської на англійську, дослідник зміг побудувати дерева речень, а потім застосувати зіставлення за зразком, щоб виявити сентиментальну орієнтацію речень. Завдяки складності методу вони змогли включити в процес багато лінгвістичних ознак, включаючи заперечення [34].

Хоча більша частина роботи виконується під час визначення семантичної орієнтації слів і фраз, деякі завдання, наприклад резюмування та пошук тексту, можуть вимагати семантичного маркування всього документа. Можливо, не має сенсу робити це для довгих текстів, таких як статті чи книги, які були ключовою формою традиційного інформаційного пошуку. Крім того, в епоху соціальних мереж та інтернет-торгівлі ми бачимо значне збільшення кількості та різноманітності коротких документів, які часто містять лише кілька речень. Це можуть бути огляди продуктів, електронні листи, публікації в блогах тощо. Подібно до підходів до визначення семантичної орієнтації слів, підходи для документів також варіюються від простих статистичних до підходів, які використовують складні структури знань для керування процесом. Одним з найпопулярніших і простих методів в цьому контексті є лінійна комбінація всіх полярностей, яка використовують усереднення для визначення полярності документів може бути виражена як формула 1.1.

$$class(d_i) = \begin{cases} C, & eval(d_i) > 0 \\ C', & eval(d_i) < 0 \end{cases}$$

(1.1)

де клас C – сума балів усіх термінів;

A клас D – усередненне для визначення полярності.

Якщо сума додається до позитивного числа, документ отримує позитивну мітку, інакше – негативну (eval більше або менше). Однак, цей підхід є суворо

бінарним – він не бере до уваги той факт, що якщо документ може мати сильні думки в обох напрямках, загалом його слід розглядати як такий, що має мітку змішаної думки, замість того, щоб покладатися на невеликі відмінності у вимірюванні для призначення будь-яка з полярностей.

При цьому маркування документа може містити інформацію, відмінну від семантичної орієнтації його складових частин. Наприклад, ми можемо використати повідомленнях груп новин, щоб розділити групу на підгрупи, які виступають за або проти проблеми. Для цього робиться чітке припущення, що цитати представляють антагоністичні точки зору, тобто більш імовірно, що відповідь буде суперечити попередньому допису (новині), ніж інакше. Хоча це, ймовірно, надмірне спрощення фактичної роботи груп новин, це перші кроки в оцінці документів у їхньому ширшому контексті.

Тепер ми переходимо до іншої важливої частини аналізу настрою – його цілі. У коротших, більш зосереджених текстах часто можна з упевненістю припустити, що автор говорить лише про тему документа. Але часто недостатньо знати загальну тему твору. Компанія, яка виробляє продукт, напевно хотіла б знати не лише те, що люди думають про цей продукт загалом, а й які особливості їм подобаються/не подобаються зокрема. Таким чином, завдання вилучення ознак (де ознака може бути будь-якою ціллю впевненого твердження) набуває популярності в області аналізу настроїв. Загальний підхід полягає у використанні тегів частини мови (POS) для створення шаблонів того, як настрої застосовуються до об’єктів. Наприклад, можна використовувати цей процес за наступним алгоритмом:

1. Виконати позначення частини мови (POS) і видалити цифри.

2. Замініть фактичні слова-особливості в реченні на тег/маркування ознаки.

3. Використати n-gram для створення коротших сегментів.

4. Розрізнити повторювані теги, даючи їм номери.

5. Використати систему аналізу асоціацій, наприклад CBA для вилучення решти функцій. Коли функції знайдено, вони групуються за допомогою синсетів

WordNet (розділ 1.3). Наприклад, слова «фото», «зображення» та «картинка» стосуються однієї функції цифрової камери. Отже, якщо вони виявляються синонімами, вони стають частиною однієї функції. Цей алгоритм загалом можливо модернізувати, наприклад видалити ті іменникові фрази, які можуть не бути характеристиками продукту, але це потребує більш комплексного підходу та довшого етапу підготовки.

Але наведені вище підходи виявляють лише явні функції, ті, що згадуються в тексті. У реченнях на зразок «ця камера завелика» є багато неявних ознак, які стосуються розміру камери. Їх можна отримати за допомогою контексту вже відомих функцій. Для цього техніка аналізу правил, описана вище, може бути розширена до неявних функцій шляхом позначення кожного специфічного для функції шаблону відповідною функцією. Зрештою, функції можна використовувати для ефективного узагальнення почуттів у тексті.

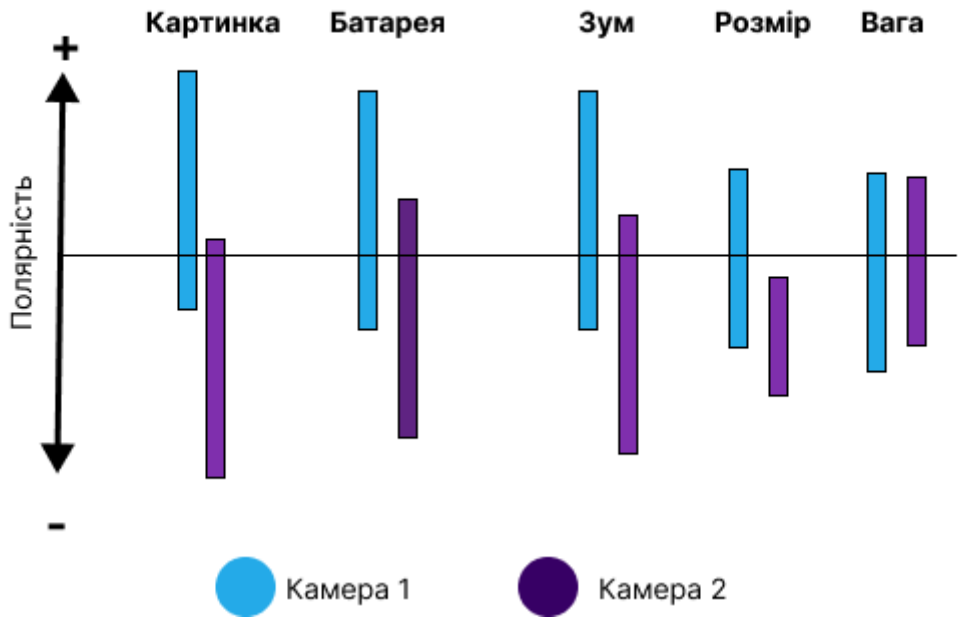


Рисунок 1.1 – Візуальне порівняння думок споживачів про два продукти

На рис. 1.1, наприклад, порівнюються різні характеристики двох камер. Тут кожна смужка вказує на діапазон думок (від негативних до позитивних) щодо

функцій камери. З першого погляду ми можемо сказати, що є більше позитивних думок щодо якості зображення камери 1, ніж камери 2.

Останнім важливим методом сентимент-аналізу є вивчення порівняльних речень. Порівняльне речення можна визначити як речення, яке виражає відношення, засноване на подібності чи відмінності більш ніж одного об'єкта. Їх можна класифікувати за типами, наприклад порівняння з градуюванням і без градуювання. Градаційне порівняння базується на співвідношенні більше, дорівнює або менше. Наприклад, «чіп Intel швидше, ніж AMD» ранжує об'єкти за якістю. Порівняння, яке не підлягає оцінці, порівнює характеристики, але не ранжує їх у порядку переваги: «Смак кока-коли відрізняється від смаку пепсі». Обидва типи речень говорять нам про відношення між різними об'єктами. Таким чином, одним із результатів системи порівняльного аналізу речень може бути ранг продуктів, визначений носіями думки. Однак поки що ідентифікація порівняльних речень була основною метою спільноти комп'ютерної лінгвістики, а не сентимент-аналізу. У цій сфері знань, використовується аналіз послідовних правил класу (CSR), щоб ідентифікувати порівняльні речення в відгуках клієнтів, обговореннях на форумах і статтях новин. Вони використовують відносно невеликий список слів (за допомогою WordNet) як перший крок в ідентифікації речень, оскільки він успішно ідентифікує майже всі порівняльні речення. У контексті сентимент-аналізу доречніше було б застосовувати іншу техніку інтелектуального аналізу даних, а саме техніку умовних випадкових полей (CRF) до анотованого вручну корпусу порівняльних речень. Так, ми зможемо ідентифікувати основні семантичні частини порівняльної думки:

- власник (автор);
- суб'єкт 1;
- порівняльні предикати;
- суб'єкт 2;
- атрибути та настрої.

1.3. Інструменти та програми для реалізації сентимент-аналізу

Як вже згадувалось у розділі 1.2. лексикон WordNet Принстонського університету є одним із найпопулярніших інструментів, який використовувався для аналізу настроїв. Як описано на їх офіційному сайті, WordNet R – це велика лексична база даних англійської мови, розроблена під керівництвом Джорджа А. Міллера. Іменники, дієслова, прикметники та прислівники згруповані в набори когнітивних синонімів (синсети), кожен з яких виражає окреме поняття. Синсети пов'язані між собою за допомогою понятійно-семантичних і лексичних відношень [15].

Есулі та Себастьяні розширюють WordNet, додаючи позначки полярності (позитивний-негативний) та об’єктивності (суб’єктивний-об’єктивний) для кожного терміна. Щоб позначити кожен термін, вони класифікують синсет (групу синонімів), до якого належить цей термін, використовуючи набір потрійних класифікаторів (пристрій, який прикріплює до кожного об’єкта рівно одну з трьох міток), кожен із яких здатний вирішити, чи синсет є позитивним, негативним або цільовим. Отримані бали варіюються від 0,0 до 1,0, що дає градовану оцінку властивостей термінів, пов’язаних із думкою. Їх можна підсумувати візуально, як показано на рис. 1.2 [15].

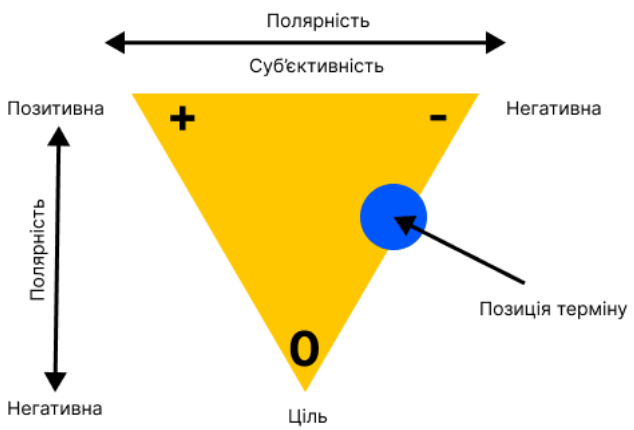


Рисунок 1.2 – Графічне представлення пов’язаних з думкою властивостей терміна

Ребра трикутника представляють одну з трьох класифікацій (позитивну, негативну та об’єктивну). Термін може розташовуватися в цьому просторі як точка, що представляє ступінь його належності до кожної з класифікацій.

Іншим розширенням для WordNet є WordNet-Affect, розроблений Страппарава та Валітутті у 2004 році. Вони позначають синсети WordNet за допомогою афективних міток (а-міток), що представляють різні афективні категорії, такі як емоції, свідомий стан, ставлення, почуття тощо. WordNet також безпосередньо використовувався в аналізі настроїв. Наприклад, у 2005 було створено лексикони позитивних і негативних термінів, починаючи з невеликого списку «початкових» термінів відомих протилежностей (наприклад, любов, як щось приємне, а ненависть як негативне), а потім використовуючи властивості антонімії та синонімії термінів, щоб згрупувати їх в одну з категорій полярності. Інші ресурси були використані для створення лексиконів. Велика робота була проведена для створення баз знань здорового глузду в галузі штучного інтелекту [15].

Але не тільки WordNet використовується для створення лексиконів. Так, наприклад велика робота була проведена для створення баз знань здорового глузду в галузі штучного інтелекту. Одними з найвідоміших проектів є Cyc4, Open Mind Common Sense5 і ThoughtTreasure6.

Гарсія у дослідженні 2020 року, яке було згадано у розділі 1.2, використав саме базу знань Open Mind Common Sense, яка містить близько півмільйона речень, щоб створити кілька моделей, що відображають різні концепції шести «базових» емоцій – щастя, суму, гніву, страху, огиди та здивування – на основі досліджень Данкомбу [16].

Крім того, Альтенберг розробив підхід до семантичного маркування з підтримкою онтології полярності (OSPM). Він вручну створив онтологію для оглядів фільмів, що значно підвищило продуктивність порівняно зі стандартною базовою лінією [13].

Також можна виконати класифікацію настроїв за допомогою статистичного аналізу та інструментів машинного навчання, які використовують переваги величезних ресурсів доступних документів з мітками (вручну за допомогою анотаторів або за допомогою системи зірка/точка). Веб-сайти з оглядами продуктів, такі як C-NET7, Ebay8, RottenTomatoes та Internet Movie Database (IMDB), широко використовувалися як джерела анотованих даних. Система «зірок» забезпечує чітке визначення загальної полярності огляду, і її часто сприймають як золотий стандарт в оцінці алгоритму.

Різноманітні дані, позначені вручну, також доступні за допомогою інструментів та програм з оцінки, таких як Text REtrieval Conference (TREC), NII Test Collection for IR Systems (NTCIR) і Cross Language Evaluation Forum (CLEF). Набори даних зібрані цими інструментами, часто служать стандартом у спільноті пошуку інформації, зокрема для дослідників аналізу настроїв. Окремі дослідники та дослідницькі групи також створили власні унікальні та цікаві бази даних, які знаходяться у відкритому доступу, наприклад:

- стенограми дебатів у Конгресі – опубліковані Томасом і Пангом у 2006 році; ці документи містить політичні промови, які позначені, щоб вказати, чи підтримував оратор обговорюване законодавство, чи виступав проти нього;
- Economining – опубліковане Школою Стерна при Нью-Йоркському університеті, що містить повідомлення для продавців на Amazon.com;
- набори даних огляду фільмів Корнелла – представлені Пангом і Лі та містять 1000 позитивних і 1000 негативних автоматично отриманих міток на рівні документа, а також 5331 позитивне і 5331 негативне речення/фрагмент;
- MPQA Corpus – системний корпус, що містить 535 анотованих вручну статей новин із різноманітних джерел, що містять позначки для думок і приватних станів (переконання, емоції, припущення тощо);
- багатоаспектні огляди ресторанів – представлені Снайдером і Барзілеєм у 2007 році; містять 4488 відгуків з чіткою оцінкою від 1 до 5 за п’ятьма

різними аспектами – їжею, атмосферою, обслуговуванням, цінністю та загальним враженнями клієнтів.

Після того, як бажаний набір даних отримано, можна використовувати різні алгоритми машинного навчання для тренування настрою класифікатори. Одними з найпопулярніших алгоритмів є опорні векторні машини та класифікатори на основі максимальної ентропії.

Крім того, попри те, що сфера аналізу настроїв є відносно молодою, уже є багато компаній, які використовують методи, розроблені в цій галузі, для клієнтів, зацікавлених у відстеженні бренду та сприйнятті ринку. Наприклад, у рамках своїх послуг із боротьби з підробкою та зловживанням брендом в Інтернеті OpSec Security надає такі послуги аналізу настроїв, як моніторинг, вимірювання та аналіз відгуків споживачів, зокрема:

- відстеження колективних думок користувачів і оцінок продуктів і послуг;
- аналіз споживчих тенденцій і галасу на ринку;
- вимірювання реакції на події та інциденти, пов’язані з компанією;
- моніторинг критичних проблем для запобігання негативні вірусні наслідки;
- оцінка відгуків кількома мовами.

Як джерело думок подібні компанії розглядають:

- інтернет-спільноти;
- дошки обговорень;
- вебжурнал;
- сайти з рейтингами продуктів ;
- чати;
- портали порівняння цін;
- групи новин.

Збираючи, оцінюючи та інтерпретуючи дані, знайдені на цих вебсайт, подібні сервіси надають інформацію та рекомендації та прогнозують тенденції щодо продуктів і брендів.

Постачальник аналізу тексту Lexalytics, з іншого боку, працював із Cisco, де вони «використовували механізм оцінки, щоб визначити, які керівники мають найвищу кореляцію щодо позитивного руху курсу акцій», як зазначає Друз у своєму огляді [15].

Відкриття «лідерів думок», як вони стверджують, допомагає компаніям виявити свої сильні сторони. Ці послуги також можуть бути корисними для урядової розвідувальної служби. Моніторинг комунікацій на наявність сплесків негативних настроїв може бути корисним для таких агенцій, як СБУ. Але крім компаній і державних установ, звичайні користувачі Інтернету можуть отримати вигоду від інструментів, що оцінюють почуття. В Інтернеті є кілька пошукових систем, орієнтованих на думку, таких як Opinmind. Завдяки попередньому позначенню вебсторінка і блогів ці служби забезпечують кластерний перегляд результатів, що покращує розуміння результатів користувачами. Теми не обов’язково обмежуються оглядами продуктів – це можуть бути політичні питання або думки про кандидатів, які балотуються на певну посаду.

Нарешті, виявлення думок може бути корисним компонентом іншого сервісу. Системи рекомендацій можуть отримати значну користь від отримання оцінок користувачів із тексту. Інформаційно-пошукові системи також можуть використовувати показники суб’єктивності при роботі з певним типом інформації.

І звісно варто зазначити відносно нову галузь та методику сентимент-аналізу, яка використовує обробку природної мови, інтелектуальний аналіз даних і інструменти пошуку тексту для розв’язання проблеми вилучення думок із тексту. Перші спроби розв'язання цієї проблеми запозичували прийоми з суміжних напрямів досліджень використовували:

- статистичні методи, щоб відстежувати впевнені слова, алгоритми;
- машинного навчання були застосовані до тексту з мітками для створення класифікаторів полярності тощо.

Але складний характер завдання вимагає ще більш складних підходів (можливо, комбінації відомих). Залишаються дві головні проблеми:

- алгоритми NLP не модернізовані для проведення досліджень у складних умовах сентимент-аналізу (наприклад, витягнути настрої з більш дискурсивних текстів);
- настрої залежать від теми: значення слів змінюється, і іноді вони стають протилежними через різницю в контексті (наприклад, фраза «йди почитай книгу» була б позитивним твердженням у рецензії на книгу, але якщо вона була б сказана в рецензії на фільм, це могло б свідчити про те, що книга краща за фільм, і, отже, мати протилежний ефект).

Загальні лексикони та алгоритми повинні бути скориговані та розширені, щоб відповідати кожній темі та її особливостям. Як згадувалося раніше, багато підходів до аналізу настроїв базуються на словниках, які не враховує лексичні зв’язки між словами. Таким чином, через складну природу настроїв потрібні більш складні інструменти, щоб повною мірою використовувати переваги семантичної інформації в тексті.

Висновки до розділу 1

У цьому розділу було висвітлено сферу аналізу настроїв, основні методи сентимент-аналізу та останні тенденції та інструменти. Об’єднуючи дослідників з інформатики, інтелектуального аналізу даних, пошуку тексту та комп’ютерної лінгвістики, ця сфера надає широкі можливості як для кількісної, так і для якісної роботи з текстом. Розглядаючи розмиті визначення почуттів і складність їх прояву в тексті, сентимент-аналіз також надає можливості для використання вже розроблених методів інтелектуального аналізу даних і аналізу тексту, а також піднімає нові питання, спонукаючи до розробки ще кращих інструментів.

РОЗДІЛ 2. АНАЛІЗ МЕТОДИК СЕНТИМЕНТ-АНАЛІЗУ ПОВІДОМЛЕНЬ В TWITTER

2.1. Аналіз обмежень Twitter як бази для сентимент-аналізу

За останні кілька років відбулося величезне зростання використання платформ мікроблогів, таких як Twitter. Спонукані цим зростанням, компанії та медіаорганізації все частіше шукають способи видобутку Twitter для отримання інформації про те, що люди думають і відчувають. Такі компанії, як Twitratr, Tweetfeel і Social Mention – це лише деякі з них, які рекламують аналіз настроїв у Twitter як одну зі своїх послуг.

Перебуваючи в десятці найбільш завантажуваних програм і маючи понад 400 мільйонів користувачів, Twitter є однією із найбільших наявних соціальних мереж у світі. Вона має багато зручних функцій, таких як автономний клієнт для робочого столу, а також широку підтримку для передачі текстових повідомлень та зображень.

При цьому Twitter містить дуже велику кількість коротких повідомлень (мікроблогів або твітів), створених користувачами цієї платформи. Зміст повідомлень варіюється від особистих думок до публічних заяв. Оскільки аудиторія платформ і сервісів мікроблогів зростає щодня, дані з цих джерел можна використовувати для аналізу думок і настроїв. Наприклад, компанії-виробники можуть цікавити такі питання:

- Що люди думають про продукт (послугу, компанію тощо)?
- Наскільки позитивно (чи негативно) люди ставляться до продукту?
- Які люди віддали б перевагу продукту?

Політичним партіям може бути цікаво знати, чи підтримують люди їхню програму, чи ні. Громадські організації можуть запитувати думку людей щодо поточних дебатів.

Аудиторія Twitter варіюється від звичайних користувачів до знаменитостей, представників компаній, політиків і навіть президентів країн. Таким чином, можна збирати текстові пости користувачів з різних соціальних груп та груп інтересів. І хоча в Twitter здебільшого переважають користувачі з США, дані можливо збирати різними мовами. Це робить Twitter реалістичною базою даних для проведення сентимент–аналізу. У цьому контексті вже було проведено досить багато досліджень про те, як настрої виражаються в таких жанрах, як онлайн-огляди та новинні статті, те, як настрої виражаються з огляду на неформальну мову та обмеження щодо довжини повідомлення (твітів).

Ще однією особливістю-проблемою Twitter є неймовірна широта теми, яка охоплюється. Не буде перебільшенням сказати, що люди пишуть у Твіттері про що завгодно. Таким чином, щоб мати можливість створювати системи для визначення настроїв Twitter щодо будь-якої теми, нам потрібен метод швидкої ідентифікації даних, які можна використовувати для навчання. Надалі ми розглянемо один із методів створення таких даних внаслідок використання хеш-тегів Twitter (наприклад, #bestfeeling, #epicfail).

Враховуючи обмеження символів у твітах, класифікація настрою повідомлень Twitter найбільш подібна до аналізу настрою на рівні речення; однак неформальна та спеціалізована мова, яка використовується у твітах, а також сама природа домену мікроблогів роблять аналіз настроїв у Твіттері унікальним завданням. Залишається відкритим питання, наскільки добре функції та методи, що використовуються для якісно сформованих даних, перенесуться в домен мікроблогів.

Варто зазначити, що завдяки доступному та легкому у використанні API Twitter стало звичним використовувати платформу соціальних медіа як джерело

даних для наукових проєктів. Але хоча Twitter загалом є чудовим джерелом даних, ми повинні бути обережними щодо висновків, які ми робимо при аналізу тексту взятого з цієї платформи. Наприклад, демографічні показники бази користувачів Twitter не обов'язково є репрезентативними для загального населення. Можливо, це не така вже й велика проблема. Зрештою, коли ми використовуємо такі бібліотеки, як Tweepy, ми зазвичай отримуємо набори твітів, а не користувачів для аналізу настроїв. Але твіти та користувачі не підтримують стале співвідношення на платформі. Деякі люди багато твітують, інші – ні. Фактично, за даними PEW Research Center, 80% твітів надходять від 10% найактивніших користувачів. Це означає, що коли ми формуємо нашу колекцію твітів для аналізу, ми не обов'язково отримуємо добірку думок різних користувачів. Натомість більш імовірно, що ми бачимо думки невеликої групи користувачів, які повторюються знову і знову.

Крім того, основною частиною Twitter є можливість «ретвітнути» певний твіт, часто з додаванням до нього власних коментарів. Ретвіти часто розуміють як схвалення позиції, але нерідко можна побачити ретвіт разом з його критикою. Отже, якщо ми використовуємо дані Twitter, ми повинні вирішити, як розглядати ретвіти – чи вважати ми ретвіти за згоду, щоб багаторазові ретвіти ідеї збільшували вагу цієї ідеї? Або ми повинні рахувати твіт лише один раз, незалежно від того, як часто його ретвітять? Перший варіант не зможе повністю врахувати ті ретвіти, які фактично функціонують як критика (негативна полярність), а другий згладить розмову, щоб впливові ретвітні заяви залишалися на тому ж рівні, що й випадкові коментарі з периферії. Будь-який вибір залишає бажати кращого, але вибір потрібно зробити.

Також Twitter наповнений нелюдськими співрозмовниками (ботами), а також створеним людьми шумом. Хоча автентичні твіти людей значно переважають над твітами ботів, проблема все одно може забруднити дані майбутнього сентимент-аналізу. В цьому контексті твіти ботів може бути важко виявити, особливо з невеликими розмірами вибірки. Проте вони можуть спотворити наші дані, так само як іноді вони призначені для навмисного спотворення розмови. Дійсно, боти, як

правило, пишуть твіти та використовують ключові терміни частіше, ніж звичайні люди, оскільки вони зосереджені на конкретних елементах розмови. Це означає, що вони створюють більше шуму в даних, ніж звичайна людина.

Окрім ботів, іноді люди намагаються маніпулювати алгоритмами Twitter для незвичних цілей, будь то комедійні, політичні чи самореклама. Наприклад, у січні 2021 року шанувальники K-pop помітили, що хештег #ImpeachBidenNow є популярним, і використали його, щоб поділитися своїми улюбленими корейськими попзірками. Таким чином вони заглушили початкову розмову про політику Байдена (фанати K-pop самі, як правило, ліві) і замінили її своєю улюбленою темою. Будь-хто, хто намагається вивчити цей конкретний хештег, буде змушений боротися з цією унікальною формою творчого активізму в Twitter.

Варто також зазначити, що більшість твітів безглузді. У контексті гігантського фрейму даних твітів багато з них позбавлені контексту або занадто короткі, щоб служити справжніми носіями сенсу чи настрою. В інших випадках будь-яке значення твіту приховано за посиланням на зображення чи вебсайт, що створює цілий ряд проблем. Цю проблему можна розділити на дві підкатегорії.

Твіти, рядки яких містять інформацію, яку важко розібрати (з деякими з них легше впоратися, ніж з іншими); емодзі, наприклад, можна врахувати в багатьох бібліотеках аналізу настроїв, однак складнішими є посилання та зображення. Ми можемо виключити їх зі свого набору даних (найпоширеніший метод у сучасних дослідженнях), але при цьому ми втратимо вагому частину даних і ризикуємо спотворити результати.

По-друге, багато твітів у розмові є відповідями на попередні твіти. Уявіть собі твіт «@aslidsiksoraksi повністю згоден». Щоб дізнатися з чим згоден користувач, нам потрібно знайти твіт, на який він або вона відповідає, що може бути дещо складно. І після цього нам потрібно розробити алгоритм, який може використати попередній твіт, щоб контекстуалізувати цей. Для цього прийдеться скористатись громіздкими моделюваннями теми, щоб з’ясувати, про що йдеться у

першій повідомленні, а потім за допомогою аналізу настроїв вже проаналізувати друге. Вся ця робота лише для того, щоб зрозуміти значення однієї відповіді.

Ще однією проблемою Twitter як бази даних для сентимент-аналізу є текстові девіації. Соціальна платформа вимагає від користувачів дотримуватися обмеження кількості символів. Крім того, Twitter використовується в переважній більшості на мобільних пристроях, де користувачі мають доступ до систем автовиправлення, які впливають на процес написання твіту. Обидва вони можуть створювати дивацтва в тому, як написані твіти, яких не було б у звичайній мові. Так, наприклад, якщо ми візьмемо n-грами GloVe, які були навчені на газетах, в нас не буде гарантії, що вони так само будуть працювати у твітах. Тому що, обмеження символів у твітах часто означає, що користувачі скорочують слова. Ці слова, крім того, люди часто вигадують абревіатури на ходу або пропускають слова, про наявність яких люди можуть здогадатися. Ці абревіатури та пропуски можуть бути виявлені технікою машинного навчання, але це зовсім інший рівень, який ускладнює аналіз.

Крім того, дані Twitter матимуть деякі помилки через автовиправлення (і звичайні орфографічні помилки). На відміну від орфографічних помилок, помилки автовиправлення важче виявити, оскільки проблемне слово є справжнім словом, яке просто вжито неправильно. Знову ж таки, безумовно, існують інструменти, які можуть це виправити, але їх використання додає додатковий рівень складності будь-якому дослідницькому проєкту на основі Twitter.

Більш загальною проблемою також залишається точне виявлення сарказму та гумору. Люди самі по собі не дуже добре розпізнають сарказм, особливо коли єдиною підказкою є кілька рядків тексту. Але сарказм є звичайною частиною розмов, особливо онлайн. Тому аналізуючи дані у Twitter система повинна враховувати сарказм та жарти.

Не можна заперечувати, що Twitter є вагомим джерелом даних для проведення сентимент-аналізу, але процес отримання результатів, які є інформативними, репрезентативними та корисними дуже складний. Природа

соціальних мереж і Twitter зокрема виступає проти простої та доступної і будь-який аналіз настроїв у Twitter має враховувати їх, щоб отримати кращі та об’єктивні результати.

Активне вивчення твітів з погляду сентимент-аналізу розпочалося близько 2010 року. Вивчення полярності текстів користувачів, особливо коротких текстів Twitter, як свідчать роботи багатьох вчених, значно ускладнено через кілька причин.

1. Вибір одиниці аналізу. Як зазначено у першому розділі, сентимент-аналіз може проводитись на рівнях від леми до всього документа. Але до Твіттеру можливо застосування класифікації як на рівні пропозиції, так і на рівні документа, що в такому випадку не важливо по суті (адже твіт вкрай рідко містить більше однієї пропозиції), але змінює доступні інструменти, результати аналізу та їхню можливу інтерпретацію. Так, наприклад, у воркшопі SemEval було запропоновано поділ семантичного аналізу для Твіттеру на аналіз всього повідомлення (message-level analysis) та термінів (слів, словосполучень, послідовностей слів; term-level analysis).

2. Мультилінгвальність твітів: часто хештегами, написаними однією мовою (наприклад, #JeSuisCharlie), маркуються твіти іншою мовою. Аналіз таких твітів вимагає багатомовних лексиконів або корпусів текстів для машинного навчання, або знаходження алгоритму фільтрації багатомовних твітів із корпусів текстів.

3. Мовні особливості твітів. По-перше, це порушення граматичних та інших мовних правил користувачами платформи. По-друге, відсутність складних пропозицій через ліміт на число символів (140 знаків). По-третє, велика кількість жаргонної та просторічної лексики, абревіатур та емотиконів (емоджі). На винятковому або частковому використанні емотиконів для оцінки тональності побудовано вже більше десятка серйозних досліджень тональності у Твіттері, але інтерпретація емотиконів у крослінгвистичному середовищі поки що мало вивчена, попри їхню одноманітність.

4. Високий відсоток сарказму в корпусах твітів. Такі твіти заплутують класифікатор, оскільки позитивні лексеми разом створюють негативне висловлювання з допомогою інтонації та/або зміни порядку слів. У разі конфліктних дискусій у Твіттері кількість таких твітів ще більше зростає. Для усунення збоїв класифікатора через сарказм рекомендується застосування методів машинного навчання на розмічених колекціях твітів.

5. Особливості попередньої обробки наборів даних. Як правило, процедури обробки включають токенізацію (tokenization), нормалізацію (normalization) і розмітку частинами мови (POS tagging). Однак деякі роботи показали, що розмітка частинами мови не працює для твітів [15]. У силу зазначених причин інструменти, розроблені для корпусів текстів більшого обсягу та більшої орієнтації на письмовий стиль викладу, дають знижений результат аналізу (low recall) для корпусів твітів. Тому для Твіттеру, по-перше, запропоновано використовувати найпростіші методи аналізу, які зазнають постійної ручної перевірки, по-друге, представлено кілька (поки недостатньо апробованих) багатоступеневих моделей тонального аналізу на основі відносно простих кроків.

Комплексні алгоритми комбінують виділення ознак об'єкта, що вивчається (feature-based models, tree kernel-based models та ін.) з n-грамами і словниками, а також з машинним навчанням або використовують побудову онтологій на базі формального аналізу концептів (formal concept analysis).

Однак, загалом сентимент-аналіз для Твіттеру поки що знаходиться на початковій стадії розвитку. Часто оцінка сентименту твітів сприймається як самоціль. Лише у кількох роботах сентимент-аналіз твітів застосовується для аналізу, передбачення (прогнозування) чи описи інших явищ. Так, сентимент твітів пов'язується з рухом ринку акцій, політичними уподобаннями громадян, обізнаністю під час техногенної катастрофи.

При цьому сентимент-аналіз повинен насамперед служити цілям опису та передбачення соціальних явищ та запобігання руйнуванням, жертвам і паніці під

час природних та техногенних катастроф. Впадає в око відсутність досліджень, що пов'язують настрій, мову користувачів та їх статус у дискусії або метадані користувача (регіон проживання, соціальний статус тощо) з його емоційною стратегією. Так, тільки в одній роботі перевірено зв'язок настрою користувачів та їх популярності за метрикою ретвітів [11].

Крос-лінгвальний сентимент-аналіз полягає у визначенні сентименту у кількох паралельних корпусах текстів різними мовами за умови, що еталонна вибірка розмічена тільки для однієї мови. Сентимент-аналіз залежить від якості машинного перекладу і вимагає як мінімум перевірки незалежності від перекладу. Іноді застосовується зворотна стратегія перекладу: перекладаються не тезауруси (з англійської на цільові мови), а самі тексти користувача (з цільових мов на англійську), після чого до них застосовуються інструменти для роботи з англійською мовою. При цьому порівняння результатів сентимент-аналізу текстів після машинного перекладу майже не відрізняється від результатів аналізу оригінальних неангломовних текстів, машинний переклад із використанням різних перекладачів також дає подібний результат.

Наприкінці 2000-х було зроблено багато спроб покращити якість крос-лінгвального аналізу настрою шляхом застосування комбінації тематичного моделювання, методів синхронного та спільного машинного навчання та ін. з сентимент-аналізом на основі машинного перекладу. Але на початку 2010-х вийшла низка робіт, яка показала, що в таких системах поки що погано враховується різниця в оформленні думки та суб'єктивності в різних мовах. Для Твіттеру, з його платформними особливостями, лінгвістичний аналіз ускладнюється ще більше. Окремі роботи присвячені сентимент-аналізу у перекладах одного корпусу текстів різними мовами? у таких роботах часто використовується метод тріангуляції, що включає етапи перекладу та ручного коригування. Але для завдань зіставлення реальних дискурсів у соцмережах цей метод не застосовується. При цьому очевидна необхідність враховувати не лише

мовні, а й культурні та контекстуальні особливості корпусів текстів, що вивчаються, при складанні словників та розмітці еталонних вибірок.

Сьогодні вкрай рідкісні роботи, які б застосовували сентимент-аналіз різними мовами (включаючи українську) до текстів ЗМІ чи соціальних медіа та аналізу соціальних конфліктів, гострих питань порядку денного, воєн, антропогенних і природних катастроф. Розроблені на цей момент програми та алгоритми для крос-лінгвального сентимент-аналізу у Твіттері, наприклад SentiSAIL або B4MSA, поки не застосовувалися до аналізу реальних тематично подібних вибірок різними мовами. Загалом при аналізі твітів різними мовами слід йти простим, але трудомістким шляхом і використовувати працю кодувальників для створення розмічених кейс-специфічних еталонних вибірок. І тут підвищення якості аналізу настрою критично вагомим залишається число закодованих для еталонної вибірки твітів і ступінь згоди між експертами, а чи не сам алгоритм сентимент-анализа.

Висновок, який можна зробити з даного огляду, полягає в тому, що поки не розроблена методика сентимент-аналізу на основі паралельних кейс-орієнтованих лексиконів, сформованих вручну за одним і тим же принципом для подібних кейсів різними мовами (наприклад, для дискусій про міжнаціональні конфлікти, міграційну кризу або природні катастрофи в різних мовних сегментах Твіттеру).

Основою такої методики має стати єдина конструкція лексикону. Елементами такого лексикону для кожної мови могли б бути:

- серцевина – вузька група базових тонально розмічених лексем даної мови;
- розширення для цієї мови, що прийшло з єдиного джерела шляхом машинного перекладу (наприклад, з тезауруса WordNet та його похідних або тезаурусу SenticNet);
- корпус культурно-орієнтованої емоційної маркованої лексики, створений шляхом експертного очищення та полярної розмітки частотного словника

даного кейсу та зіставлений з такими самими корпусами для паралельних кейсів, щоб встановити ступінь подібності;

- подібні процедури перевірки якості сентимент-аналізу.

Такий дизайн дослідження дозволив би оцінити, наскільки в принципі можлива автоматизація та отримання подібних результатів сентимент-аналізу без застосування перекладу текстів англійською мовою, що для Твіттеру є неприйнятним. Він також дозволив би розробити нові, універсальні критерії оцінки якості аналізу настрої коротких текстів користувача в мережі Інтернет.

2.2. Особливості аналізу даних у Twitter за допомогою методів машинного навчання

Алгоритми машинного навчання більш адаптивні до змін вхідних даних, яку ми можемо часто спостерігати у Twitter. Методи машинного навчання зазвичай використовуються для бінарної класифікації та прогнозування настроїв як позитивних чи негативних. При цьому безпосередньо алгоритми машинного навчання далі класифікуються за такими категоріями.

1. Під наглядом: у цих алгоритмах надається навчальний набір даних із попередньо визначеними класами, і на основі цього навченого набору даних вхідні дані позначаються вихідним класом-результатом. Ці алгоритми класифікують вхідний набір даних за допомогою навченого класифікатора. Навчальні дані складаються з набору прикладів, кожен з яких містить вхідний об’єкт і бажані вихідні результати. Визначена функція створюється внаслідок аналізу навчальних даних під наглядом, які пізніше можуть бути використані для відображення нових вхідних даних, які також називаються тестовими даними. Переважно методи машинного навчання використовують саме цей контрольований підхід. Його можна додатково класифікувати за двома методологіями, а саме класифікацією та

регресією. Найпоширенішими прикладами керованих алгоритмів машинного навчання є лінійна регресія, випадковий ліс і опорні векторні машини.

2. Неконтрольоване: ці типи алгоритмів машинного навчання беруть неструктуровані вхідні дані, а потім за допомогою різних алгоритмів виявляють приховану структуру-шаблон. На відміну від керованого навчання, цей метод не використовує попередньо зазначені дані для навчання класифікатора. Неконтрольоване машинне навчання можна своєю чергою розділити на кластеризацію та асоціацію, найпоширенішим прикладом алгоритмів неконтрольованого машинного навчання є K-Means і Apriori Algorithm.

3. Напівконтрольований: ці типи алгоритмів працюють як з позначеними, так і з не позначеними наборами даних, наприклад інструменти та методики, засновані на словнику.

Найбільшою популярністю користуються саме методи контрольованого навчання, коли по корпусу розмічених даних (навчальній вибірці) будується модель (машинний класифікатор), яка застосовується до нових, нерозмічених текстів. Поряд із традиційними методами машинного навчання цього типу, такими як наївний баєсівський класифікатор, дерева рішень, метод опорних векторів, логістична регресія, все частіше застосовуються приховані моделі Маркова (HIDden Markov Models, HMM), метод умовних випадкових полів (condition random fields) та нейронні мережі. Дані, витягу яких необхідно навчитися, розмічені в текстах навчальної вибірки певним чином, а саме записані їх лінгвістичні і структурні ознаки, часом також найближчий контекст. Для спрощення роботи експерта, який розмічає тексти, нерідко попередньо проводиться їх графематичний та морфологічний аналіз, рідше синтаксичний. Використовувана розмітка і ознаки, що враховуються, а також методи, що застосовуються для навчання, багато в чому залежать від виду інформації, що видобувається.

Для розмітки категорій іменованих сутностей запропоновано різні схеми. Найпростішим є схема IO (I – inside, O – outside), при застосуванні якої токени, що

належать до імені (inside), розмічаються його категорією, а токени поза ім'ям (outside) розмічаються тегом O. Виходить, що якщо система навчається для розпізнавання імен персоналій (PERS), всі токени тексту діляться на два класи: які стосуються PERS і які стосуються O. У разі, якщо додатково враховуються географічні назви (LOC), то токени будуть ділитися вже на три класи: PERS, LOC або O.

Якщо система повинна навчитися розрізняти категорії PERS та LOC, то при навчанні всі токени тексту діляться на класи B-PERS, I-PERS, B-LOC, I-LOC та O, тобто класифікатор для кожного токена вчиться розпізнавати, чи є цей токен початком, чи продовженням іменованої сутності конкретної категорії (PERS чи LOC), і навіть випадок, коли токен взагалі є частиною найменування. Існують і більш складні схеми розмітки, що дозволяють розрізняти однослівні та багатослівні назви; для багатослівних назв передбачається розмітка як їх початку і продовження, а й закінчення. Для застосування машинного навчання дані після розмітки мають бути перетворені на набори ознак для кожного токена. Набір ознак зазвичай включає ознаки самого токена, а також ознаки, що ґрунтуються на знаннях, одержуваних із зовнішніх джерел, зокрема, зі словникових ресурсів. Причому ознаки вказуються як для значних (витягуваних) токенів, але й сусідніх, зазвичай беруться два токени зліва і справа.

До ознак токена зазвичай відносять:

- власне токен;
- вид токена: слово, розділовий знак, цифро-літерний комплекс тощо;
- довжину токена; · чи є початком/кінцем пропозиції;
- тег токена, отриманий ним під час розмітки.

Для токенів-слів додатково враховується:

- спосіб написання токена: тільки великими літерами, тільки малими, перша літера заголовна і т. п;
- лема, частина мови, значення морфологічних ознак;

- склад слова: коріння, суфікси та закінчення, типові для прізвищ, назв організацій та інших категорій сутностей.

Що стосується ознак, що отримуються зі словникових ресурсів, то вони, як правило, відповідають на питання, чи входить токен до певного словника. Використовувані словники можуть включати:

- частотні імена, по батькові та прізвища, назви компаній, фірм та організацій, географічні назви; ;
- слова, що є частинами найменувань, наприклад, типи організацій;
- слова-маркери, за якими зазвичай розташовуються іменовані сутності певних категорій (місто, вулиця, річка тощо).

Використання під час отримання іменованих сутностей великої кількості словникових ресурсів є особливістю сучасних систем, заснованих на машинному навчанні (а також систем, заснованих на інженерному підході).

Таким чином, завдання вилучення іменованих сутностей сприймається як класичне завдання класифікації токенів кількох класів. При цьому для навчання та використання класифікаторів можуть застосовуватись різні стратегії. Наприклад, можна навчити класифікатор одночасно розпізнавати сутності різних категорій, а можна побудувати окремі класифікатори кожної категорії і потім об'єднувати результати їх роботи. У той самий час, оскільки під час вирішення завдання вилучення сутностей активно використовується локальний контекст класифікованого токена, дуже часто логічніше використовувати некласичні методи навчання (байесовський класифікатор, дерева рішень тощо), а приховані марківські моделі (HMM) та метод умовних випадкових полів (CRF), розглядаючи категорії іменованих сутностей як приховані стани, а токени – як спостережувані.

Сучасною тенденцією у розв’язанні завдання отримання іменованих сутностей є також застосування методів навчання без нагляду (методів кластеризації), що дозволяють автоматично кластеризувати слова за схожими контекстами їх вживання. Важливо, що робота цих методів відбувається з

нерозміченим текстовим корпусом, як вже згадувалось раніше, що дозволяє долати обмеженість наявної текстової колекції. Зауважимо, що результати кластеризації іменованих сутностей можуть використовуватися також як додаткова ознака, що базується на знаннях, при застосуванні методів (часткового) нагляду.

В останні роки з'явилися роботи, в яких застосовуються нейронні мережі та використовуються підходи на основі глибокого навчання (deep learning), наприклад, технологія Word2vec, але загалом вони не дали суттєвого приросту якості вилучення. Особливістю саме методів, що використовують нейронні мережі, є те, що вони дозволяють досягти якості, порівнянної з найкращими сучасними методами, але з мінімальним набором додаткової інформації: ознак токенів, словникових ресурсів та ін [11].

У завданнях розпізнавання відносин і фактів через складності розмітки даних використовуються дуже рідко. Найбільш типове використання методів з урахуванням часткового навчання – це так званий підхід distant supervision, у якому для навчання береться велика кількість прикладів сутностей (сотні та тисячі), пов'язаних певним ставленням або фактом. Джерелом цієї інформації може бути зовнішня база знань типу Twitter. Для формування навчальної вибірки у цьому контексті робиться досить грубе припущення, що обрані, що містять пов'язані певним ставленням сутності, є позитивними прикладами, а пропозиції, що містять сутності, але не пов'язані цільовими відносинами, є негативними. Таким чином, автоматично готується навчальна вибірка пропозицій, до якої можна застосувати машинне навчання.

Ознаки, які застосовуються при вилученні відносин і фактів, що пов'язують іменовані сутності, в основному враховують контекст навколо сутностей:

- список лем слів, що стоять між сутностями, та їх частини мови;
- слова та їх частина мови зліва від лівої та праворуч від правої сутності;
- синтаксичний шлях між сутностями та його довжину;
- категорії іменованих сутностей.

Зазначимо також, що для коректної роботи систем, що базуються на машинному навчанні, розмічений корпус (навчальна вибірка) повинна мати досить великий обсяг, а також високу якість розмітки. Розмітка тексту є непростим і трудомістким процесом, який породжує досить високий відсоток помилок. Додатковою складністю може стати вибір відповідного методу навчання. Ще одне слабке місце підходу на основі машинного навчання пов'язане з тим, що результати роботи методів машинного навчання зазвичай погано зрозумілі, тому локалізувати і виправити помилки, що виникають практично неможливо.

При переході на інше завдання та предметну область системи, що використовують машинне навчання, стикаються з тими ж проблемами, що й системи, що ґрунтуються на правилах: систему необхідно налаштовувати заново. Залежно від використовуваного методу може знадобитися навчання системи на новому Twitter-корпусі та/або коригування безлічі ознак, які він враховує. Проте вже розмічений корпус і створений набір ознак можна використовувати багаторазово, пробуючи на ньому різні методи і стратегії навчання, не залучаючи лінгвіста для кропіткої роботи з аналізу текстів предметної області та написання правил і шаблонів. Ця обставина пояснює широке використання машинного навчання в дослідницьких роботах із вилучення інформації у Twitter в останні роки.

2.3. Порівняльний аналіз інструментів аналізу настроїв у Twitter

За останні роки було розроблено багато інструментів для аналізу настроїв у коротких неформальних текстах у соціальних мережах. Ми перевірили досить різноманітний набір із 19 інструментів аналізу настроїв Twitter. Ці інструменти включали вільно доступні системи, розроблені в академічних умовах, комерційні інструменти на основі API, які вимагали щомісячної підписки, і кілька алгоритмів, опублікованих у літературі з NLP. Хоча 19 включених аж ніяк не є вичерпним списком, вони включають інструменти з сотнями регулярних користувачів, ті, які

були завантажені кілька тисяч разів, і ті, що з’являються в статтях, які цитуються сотні разів, і навіть використовують великі компанії.

Ці 19 оцінюваних інструментів можна загалом згрупувати у дві категорії: автономні комерційні інструменти та навчені робочі інструменти. Окремі інструменти використовують моделі текстової аналітики, які можна застосувати безпосередньо до документів без міток відразу «з коробки». Ці інструменти включають пропозиції на основі API та ті, які можна завантажити як настільні програми. Ми включили в оцінку 15 таких інструментів. Готовий до використання характер цих інструментів, без потреби в розробці предметно-спеціальної моделі або навчанні, полегшує процес їх застосування. Однак відсутність специфіки домену також може бути шкідливим з точки зору продуктивності, оскільки моделі, що лежать в основі інструментів, можуть включати правила/припущення, які є помилковими або незастосовними в контексті конкретного тестовому стенду (зокрема аналізу Twitter).

Окремо оцінено наступні комерційні інструменти: uClassify, ChatterBox, Sentiment140, Textalytics, Intridea, AiApplied, ViralHeat, Lymbix, SentimentAnalyzer, TextProcessing, Semantria, SentiStrength, MLAnalyzer і Repustate. Більшість автономних інструментів, включених у дослідження, є комерційними пропозиціями, доступ до яких здійснюється безпосередньо через API постачальника або через сторонній ринок API, такий як MashApe. Двома винятками є Sentiment140 і SentiStrength, обидва з яких були розроблені в результаті опублікованих академічних досліджень.

SentiStrength є популярним автономним інструментом аналізу настроїв. Він використовує лексикон настроїв для присвоєння балів негативним і позитивним фразам у тексті. Щоб визначити полярності на рівні речень або документів, бали на рівні фрази можна агрегувати. Природа такого підходу з неконтрольованим навчанням робить його легко застосовним до будь-якого набору даних у Twitter.

Sentiment140 використовує навчений класифікатор машинного навчання, створений на основі великого корпусу Twitter позитивних і негативних твітів, автоматично створених на основі наявності смайлів. Інструмент використовує n-грами тегів слів і частин мови в поєднанні з класифікатором машинного навчання на основі максимальної ентропії.

Деякі з автономних інструментів виводять лише безперервні оцінки полярності на відміну від класифікації дискретної полярності. Для таких інструментів (наприклад, ChatterBox) ми дискретизували безперервні оцінки на три категорії (позитивні, нейтральні та негативні), використовуючи контейнери, які максимізували ефективність інструментів щодо загальної точності та запам'ятовування на рівні класу. Хоча такий підхід міг збільшити продуктивність певних автономних інструментів, ми вважаємо, що максимізація продуктивності інструментів, наскільки це можливо, втілює перспективу, орієнтовану на користувача.

Інструменти верстака (Workbench) – це ті, які потребують контрольованої розробки моделі на основі навчання на визначеному навчальному наборі. Вони надають параметри для створення основи, токенізації, включення різних представлень функцій і параметрів для кількості функцій, які потрібно включити в моделі. Серед 5 інструментів робочого столу були LightSide, BPEF, EWGA, FRN і базовий запуск слова n-грам за допомогою розширення обробки тексту в RapidMiner.

Інструменти Workbench вимагають детального налаштування параметрів і перевірки в навчальному середовищі, але мають потенціал для включення точних завдань і предметних знань. Так, EWGA використовує ентропійно-зважений генетичний алгоритм для ефективного вибору ознак для класифікації настроїв за допомогою моделі-обгортки, де продуктивність підмножини ознак використовується як значення функції відповідності в рамках генетичного алгоритму. FRN використовує мережу відношень ознак, що складається з двох

ключових синтаксичних відношень n-грам: субсумпції та паралельних відношень. Ці відносини використовуються для ефективного вибору ознак із багатих просторів ознак, що охоплюють різноманітні типи n-грам. Скорочений набір функцій вводиться в класифікатор SVM. BPEF використовує структуру початкового параметричного ансамблю. Ансамбль, що охоплює десятки тисяч бінарних класифікаторів «один проти одного», створено з використанням різних комбінацій наборів даних, наборів функцій і машин. Евристика пошуку на метарівні використовується для ідентифікації невеликої підмножини моделей, які зрештою зберігаються для класифікації. Базовий рівень n-gram складається з уніграм, біграм і триграм, вибраних за допомогою евристики отримання інформації та в поєднанні з класифікатором SVM, як це було зроблено в попередніх дослідженнях.

Для порівняльного аналізу ми використовували твіти, що стосуються 5 широких тем: політики, фармацевтики, безпеки, технологій та комерції. Вибрані теми стосуються низки сфер застосування, включаючи настрої споживачів, соціальні медіа для розумного здоров’я, інформаційну безпеку та досвід користувачів. Усі набори даних були позначені полярністю золотих стандартів за допомогою Amazon Mechanical Turk (AMT). До використання AMT використовувалися ручні та автоматизовані методи попередньої обробки для видалення нерелевантних твітів (наприклад, неанглійською мовою або не пов’язаних з темами, що цікавлять). В рамках AMT використовувався модуль Sentiment Rating. Крім того, лише твіти з настроями, пов’язаними з цілями, були позначені як позитивні чи негативні.

Таблиця 2.1 – Огляд випробувальному стенду

Тема	Опис твітів	Кількість твітів			
		Усього	Позитивні	Негативні	Нейтральні
Політика	Загальне обговорення досвіду, новин і конкретних подій.	5281	20.9%	8.9%	70.2%
Фарма	Пов’язані з досвідом користувачів із фармацевтичними препаратами.	5009	15.6%	11.1%	73.3%
Безпека	Пов’язані з продуктами та послугами основних компаній, що займаються охороною, включно з інцидентами безпеки та новими випусками програмного забезпечення та/або виправленнями безпеки.	5086	24%	11.1%	64.8%
Технології	Включіть обговорення продуктів технологічних компаній, послуг, політики та загального досвіду користувачів.	3502	15.1%	16.9%	68%
Комерція	Охоплюють обговорення конкретної категорії товарів для роздрібної торгівлі (побутова техніка) та досвід користувачів, пов’язаний із цими продуктами.	3750	42.7%	9%	48.3%

У таблиці 2.1 наведено огляд випробувальному стенду. Більшість твітів у кожному тестовому стенді були нейтральними, оскільки більшість повідомлень у

соціальних мережах, пов’язаних із певною темою, не містять явно виражених позитивних чи негативних настроїв. Єдиним винятком був тестовий стенд комерції, де люди мають переважно або позитивні, або негативні настрої. З усім тим, усі п’ять наборів даних були досить незбалансованими щодо розподілу твітів між класами позитивної, негативної та нейтральної полярності. Було використано кілька стандартних показників оцінки. Вони включали загальну точність і запам’ятовування на рівні класу.

Загальна точність – це відсоток загальної кількості твітів, класифікованих правильно (як позитивні, нейтральні чи негативні). Запам'ятовування на рівні класу – це відсоток твітів, пов'язаних з певним класом, які були класифіковані як такі. Наприклад, негативне запам’ятовування – це відсоток усіх негативних твітів у тестовому стенді, які класифікуються як негативні.

Таблиця 2.2 – Загальна точність для 14 автономних інструментів на 5 тестових стендах

Інструмент	**Усього**	**Фарма**	**Комерція**	**Безпека**	**Техніка**	**Політика**
SentiStrength	67.49	74.68	56.35	65.51	69.61	71.31
Chatterbox	67.43	75.04	53.19	67.20	69.73	71.99
Sentiment140	66.46	62.09	61.77	68.84	67.82	71.79
Textalytics	66.22	70.33	55.14	66.33	68.29	71.02

Продовження таблиці 2.2 – Загальна точність для 14 автономних інструментів на 5 тестових стендах

Інструмент	**Усього**	**Фарма**	**Комерція**	**Безпека**	**Техніка**	**Політика**
Intridea	63.31	64.18	47.37	62.63	75.19	67.20
AiApplied	61.84	69.59	47.99	64.05	60.39	67.20
ViralHeat	61.16	63.77	48.42	61.94	64.12	67.56
Lymbix	56.63	52.03	54.81	47.60	63.45	65.25
SentimentAnalyzer	55.15	55.33	51.36	54.83	56.50	57.75
TextProcessing	54.06	49.68	50.01	58.40	52.40	59.79
Semantria	53.50	44.68	56.33	45.46	60.99	60.06
uClassify	47.22	51.70	42.12	47.51	50.31	44.47
MLAnalyzer	45.20	37.95	52.15	41.35	48.06	46.47
Repustate	43.98	35.80	41.06	31.93	40.90	59.79

У таблиці 2.2 наведено результати експерименту для 15 автономних комерційних інструментів, а в таблиці 2.3 зображено загальну точність для 5 методів верстака. У таблиці також включено середню точність на 5 тестових стендах як показник загальної ефективності. Що стосується автономних інструментів, SentiStrength, ChatterBox, Sentiment140 і Texalytics забезпечили найкращу загальну продуктивність на тестових стендах із середньою точністю понад 66%. Серед найефективніших автономних інструментів продуктивність

Sentiment140 була найбільш збалансованою серед тестових стендів із точністю від 61% до 71%.

З іншого боку, середня точність чотирьох інструментів була нижчою за 50%, а середня точність для 15 інструментів становила лише 56%. Ці результати свідчать про те, що якість інструментів аналізу настроїв Twitter значно відрізняється, і ця варіація продуктивності може мати важливі наслідки для різних програм аналітики соціальних мереж на основі Twitter.

Таблиця 2.3 – Загальна точність для 5 робочих інструментів на 5 випробувальних стендах

Інструмент	**Усього**	**Фарма**	**Комерція**	**Безпека**	**Техніка**	**Політика**
BPEF	71.38	67.81	65.24	75.32	76.30	72.21
Lightside	69.35	70.71	58.22	69.86	76.99	70.99
FRN	69.17	72.60	59.96	69.98	71.00	72.30
EWGA	68.12	70.21	60.00	68.50	70.50	71.41
RapidMiner	66.86	67.50	59.52	66.02	70.02	71.22

Виходячи з результатів, наведених у таблиці 2.3, не дивно, що методи верстака забезпечили вищу середню точність на тестових стендах, оскільки вони навчалися на подібних даних, що й набори для оцінювання. Середня точність цих методів становила від 67% до 71% із загалом меншою варіацією між тестовими стендами (наприклад, точність BPEF змінювалася лише на 6% між тестовими стендами). Результати в таблицях 2.2 і 2.3 також ілюструють можливі компроміси між використанням автономних і робочих інструментів. Щоб глибше дослідити ці

потенційні компроміси, ми перевірили показники запам’ятовування на рівні класу для найефективніших автономних і робочих інструментів. Були включені п’ять кращих автономних і три кращих робочих інструментів. Ці результати представлені в таблицях 2.4 і 2.5.

Таблиця 2.4 – Відкликання на рівні класу для окремих автономних та Workbench-інструментів у тестових стендах фармацевтики, комерції та безпеки

Інструмент	**Фармацевтика**			**Комерція**			**Безпека**		
	Поз.	**Нег.**	**Нейт.**	**Поз.**	**Нег.**	**Нейт.**	**Поз.**	**Нег.**	**Нейт.**
SentiStrength	46.98	90.47	29.29	53.28	38.10	62.44	87.88	90.93	0.15
Chatterbox	37.23	56.65	62.47	57.55	11.01	52.79	63.39	30.07	66.52
Sentiment140	44.03	62.59	65.84	43.76	11.01	87.09	61.02	25.98	79.02
Textalytics	28.75	23.74	86.20	21.04	10.39	93.60	24.80	16.01	90.25
Intridea	26.19	68.71	37.45	32.52	63.69	37.34	32.92	62.23	39.89
BPEF	63.18	61.33	69.76	60.74	64.58	69.33	75.92	72.42	75.60
Lightside	44.93	28.06	82.63	57.24	34.23	63.54	56.59	47.33	78.60
FRN	39.15	21.04	84.86	58.99	24.70	67.35	55.20	35.94	81.23

З таблиці 2.4 видно, що показники запам’ятовування на рівні класу значно відрізнялися для певних інструментів у різних наборах даних. Що стосується автономних інструментів, SentiStrength досяг хорошої загальної точності завдяки вищим позитивним і негативним рівням відкликання наборів даних, таких як

фармацевтика, безпека та політика (від 80% до 90% при негативному відкликанні). Однак ці вищі позитивні/негативні показники супроводжувалися помітно нижчими показниками запам’ятовування нейтрального класу (тобто багато твітів з хибно позитивними/негативними настроями). Подібним чином ChatterBox мав тенденцію до кращого запам’ятовування позитивних і нейтральних твітів, тоді як Textalytics був найефективнішим з точки зору нейтрального запам’ятовування (від 86% до 90% на тестових платформах).

Таблиця 2.5 – Відкликання на рівні класу для окремих автономних та Workbench-інструментів у тестових стендах технологій та політики

Інструмент	**Технології**			**Політика**		
	Поз.	**Нег.**	**Нейт.**	**Поз.**	**Нег.**	**Нейт.**
SentiStrength	62.00	45.27	61.89	59.67	80.56	42.47
Chatterbox	52.64	45.10	56.51	53.16	33.97	64.39
Sentiment140	55.09	36.82	78.36	46.38	28.42	84.84
Textalytics	26.98	18.04	90.00	25.07	29.06	90.02
Intridea	69.81	81.76	74.75	32.73	68.38	46.32
BPEF	69.25	78.55	77.31	57.01	74.52	76.47

Продовження таблиці 2.5 – Відкликання на рівні класу для окремих автономних та Workbench-інструментів у тестових стендах технологій та політики

Інструмент	**Технології**			**Політика**		
	Поз.	**Нег.**	**Нейт.**	**Поз.**	**Нег.**	**Нейт.**
Lightside	47.92	55.91	84.29	38.70	49.79	83.30
FRN	37.17	42.06	80.92	33.30	37.26	81.10

Щодо верстакових інструментів, LightSide та FRN продемонстрували подібну упередженість щодо певного класу полярності (табл. 2.5). Обидва методи досягли вищих нейтральних показників запам'ятовування, але з помітно нижчими позитивними та негативними значеннями запам'ятовування. І навпаки, BPEF досяг відносно збалансованих значень запам’ятовування на рівні класу в позитивних, негативних і нейтральних класах. Швидкість запам'ятовування на рівні класу загалом була в межах 10% один від одного, за винятком набору даних політика, де показник позитивного відкликання був приблизно на 17%–19% нижчим, ніж у двох інших класах. Показники запам’ятовування на рівні класу загалом і збалансованість зокрема мають важливе значення для програм аналітики соціальних медіа, що включають настрої.

Наприклад, дослідження Хассана у 2013 показали, що інструменти з нижчими позитивними та негативними показниками запам’ятовування сприйнятливі до створення часових рядів індексу настроїв, які є «плоскими» та менш ефективними для представлення подій із надзвичайно позитивними чи негативними настроями. Вони проілюстрували це за допомогою часових рядів настроїв для великого північноамериканського телекомунікаційного провайдера.

Різниця в продуктивності між точністю окремих інструментів (у середньому понад 26%) і загалом низька продуктивність інструментів (переважно нижче 70%)

підкреслюють проблеми, пов’язані з ефективною класифікацією настроїв Twitter. Щоб краще зрозуміти ці проблеми, було проведено детальний аналіз помилок на 5 тестових стендах. Ми перевірили найбільш неправильно класифіковані твіти у кожному тестовому стенді. Помилки були згруповані у дворівневу ієрархічну таксономію (представлену на рис. 2.1).

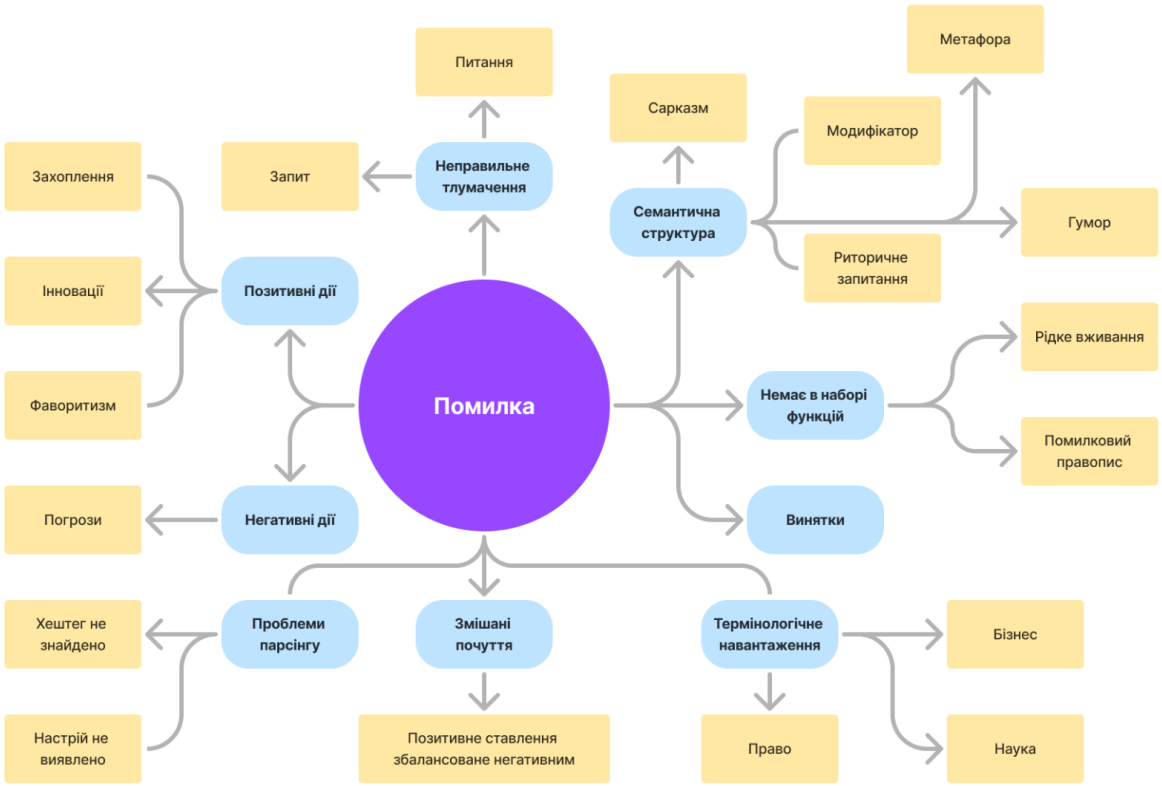


Рисунок 2.1 – Таксономія помилок аналізу настрою Twitter

Таксономія містить 9 категорій верхнього рівня. Категорія неправильно витлумачених цілей користувача включає нейтральні запитання чи прохання, які помилково сприймаються як доповнення чи критику. Категорія семантики-структури речення включає жарти, сарказм, риторику та пов’язані літературні прийоми, які були добре задокументовані як проблематичні для інструментів аналізу настроїв. Неправильно витлумачене призначення користувача включає запити або запитання (наприклад, «Було б чудово, якби ми могли…»). Проблеми аналізу включають почуття, виражені в хеш-тегах, які часто не аналізуються

належним чином. Виняток із категорії звичайних сигналів настрою включає помилки, пов’язані з наявністю термінів, які використовуються в нетиповому відношенні, наприклад, використання лайки для вказівки на позитивний результат.

Інші поширені категорії помилок включали твіти, що містять змішані почуття, де автори виражають як позитивні, так і негативні почуття щодо різних тем у межах 140 символів. Подібним чином відсутність відповідності висловлених настроїв цілям була ще однією причиною неправильних класифікацій. Цікавою також була категорія помилок позитивних рішень. Це були твіти, що містили тонкі позитивні настрої, такі як згадки про пожертви, благодійні організації та інші події чи дії з певною мірою прихованою позитивною конотацією. Інструменти, які не могли ідентифікувати такі твіти, як правило, були тими, у яких, імовірно, бракувало лексикону термінів позитивних дій та/або ключових слів. Використовуючи цю таксономію, ми перевірили частоту помилок для 19 інструментів за категоріями на 5 тестових сценах.

Очевидно, що певні типи помилок були найбільш поширеними. Не дивно, що проблеми із семантикою/структурою речень (наприклад, сарказм, модифікатори, жарти, риторика тощо) спричинили найбільший відсоток дуже помилкових твітів для багатьох тестових стендів. Ця категорія постійно охоплювала 10-15% від загальної кількості помилок. Проте кілька інших категорій також були досить поширеними. Позитивні рішення за цільовою категорією спричинили понад 10% помилок у чотирьох із п’яти наборів даних. Нерелевантні категорії позитивних і негативних настроїв (стосовно цілей) також спричинили від 5% до 15% помилок кожна. Цікаво, що помилки, що стосуються змішаних настроїв, були дуже рідкісними, попри те, що вони становили серйозну проблему в інших каналах соціальних мереж, таких як вебфоруми та блоги. Цей висновок свідчить про те, що обмеження в 140 символів накладає певні обмеження на здатність користувачів формулювати складні думки, що охоплюють численні протилежні почуття. Крім того, існувала взаємодія між категорією помилок і набором даних, причому певні помилки були поширеними в окремих доменах.

У більшості випадків 2-3 найпопулярніші категорії помилок становлять більшість помилок у цьому конкретному наборі даних. Наприклад, неправильно витлумачена мета користувача була значним джерелом помилок у наборах даних фармацевтики та комерції (понад 20% і 40% відповідно), де запитання щодо досвіду використання ліків, що відпускаються за рецептом, або якості товарів кількома інструментами неправильно класифікувалися як негативні настрої. Подібним чином багато твітів, що стосуються благодійної діяльності, заходів зі збору коштів і пожертвувань у наборі даних політики, були неправильно класифіковані декількома інструментами, що спричинило понад 35% помилок. Навіть категорія семантики-структури речення, хоч і поширена в усіх п’яти наборах даних, була більш поширеною в наборах даних технології і безпеки. У цих двох сферах літературні прийоми, такі як жарти, сарказм і риторика, були набагато поширенішими (наприклад, коли йдеться про програмне забезпечення безпеки, виробників смартфонів або великі технологічні фірми). Загалом результати проливають світло на те, як помилки інструменту аналізу настроїв проявляються у твітах, що стосуються різних тем.

Таким чином, результати нашого порівняльного аналізу мають важливе значення для кількох груп зацікавлених сторін. Дослідники аналітики соціальних медіа можуть використовувати результати, щоб приймати більш обґрунтовані рішення щодо вибору конкретних інструментів для проєкту. Вони також можуть зважити компроміси між використанням автономних і робочих інструментів. Дослідники та розробники NLP і текстового аналізу можуть використовувати результати аналізу помилок для вдосконалення майбутніх комерційних автономних інструментів. Керівники галузі можуть краще знати про можливі сильні та слабкі сторони основної текстової аналітики та про те, як це може вплинути на якість і надійність вхідних даних соціальних медіа, які використовуються для прийняття рішень. Крім того, дані аналізу помилок наведені вище допомогли нам розробити алгоритм аналізу настроїв Twitter, зазначений у розділі 3 цієї роботи.

РОЗДІЛ 3. РОЗРОБКА ТЕХНОЛОГІЇ НА ОСНОВІ NLP ДЛЯ ТЕМАТИЧНОГО СЕНТИМЕНТ-АНАЛІЗУ ДИСКУРСУ ВІЙНИ В УКРАЇНІ У TWITTER

3.1. Методологія збору даних для сентимент-аналізу

Щоб охарактеризувати російсько-український конфлікт 2022 року у Твіттері, ми розробили методологію для:

- створення відповідного набору даних;
- кількісного та якісного аналізу всіх зібраних твітів на цю тему.

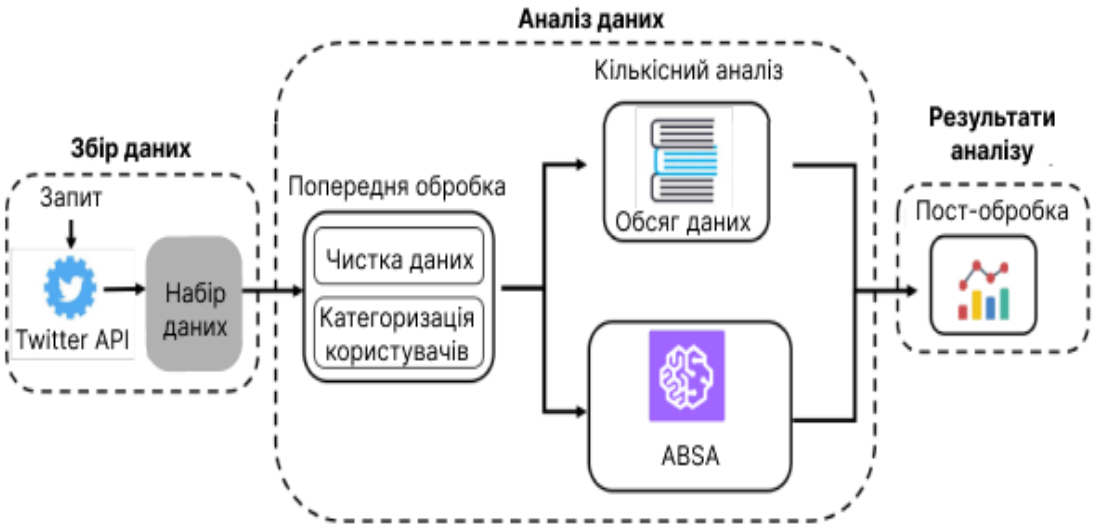


Рисунок 3.1 – Методологія розробленого методу

Рисунок 3.1 детально ілюструє наш підхід, що складається з трьох різних етапів. Перший крок – збір даних. Щоб створити наш набір даних, ми зібрали твіти, пов’язані з цільовою темою, через Twitter API. Далі, після попередньої обробки наших даних, ми аналізуємо зібрані твіти як з кількісної, так і з якісної точки зору.

Зокрема, ми оцінили обсяг наших даних і дослідили настрій користувача з часом за допомогою аналізу настрою на основі аспектів.

Щоб створити наш набір даних, ми зібрали твіти, пов’язані з російсько-українським конфліктом, через Twitter API v2. Зокрема, ми запитали кінцеву точку Twitter «tweets/search/all», яка повертає повну історію загальнодоступних твітів, що відповідають пошуковому запиту, від самого першого твіту на платформі (26 березня 2006 р.) до останнього (час виконання запиту). Щоб відібрати лише ті твіти, які найкраще висвітлити зміни в суспільних настроях, по-перше, ми звузили простір пошуку до двох місяців, за один місяць до та один місяць після 24 лютого 2022 року. Того дня російський президент Володимир Путін оголосив про початок військової спеціальної операції в Україні. Потім ми вибрали кілька ключових слів, наведених на рис. 3.2, щоб визначити сферу наших інтересів.

```
['zelensky', 'ukraine', 'ukrainian', 'russia', 'russian', 'putin']
```

Рисунок 3.2 – Вибрані ключові слова для збору даних

Оскільки ми досліджуємо тему російсько-українського конфлікту, нас також цікавить сприйняття користувачами Організації Північноатлантичного договору (НАТО). Однак ми не використовували це ключове слово для збору даних, оскільки воно безпосередньо не стосується досліджуваної теми. Насправді включивши ключове слово НАТО, ми також зібрали б твіти, які ні прямо, ні опосередковано не стосуються російсько-українського конфлікту. Тим не менш, ми розглянули ключове слово НАТО в аналізі даних, перевіривши як обсяг, так і настрої зібраних твітів, у яких згадувалося Атлантичний альянс разом з одним із ключових слів (рис 3.6).

Крім того, ми вирішили орієнтуватися лише на оригінальний вміст, уникаючи ретвітів, відповідей і цитат. Крім того, оскільки нас цікавить лише

текстовий вміст, ми виключили твіти, що містять посилання та медіа. І враховуючи інструменти та обсяг цього дослідження, ми аналізували лише твіти англійською мовою (розробити корпус для об'єктивного аналізу українською зайняло б багато часу і ресурсів).

```
1 start_time = "2022-01-27T00:00:00.000Z"
2 end_time = "2022-03-23T00:00:00.000Z"
3
4 query = "(zelensky OR zelenskyy OR ukraine OR ukrainian
5 OR russia OR russian OR putin)"
6 query += " -has:links -has:media -is:retweet -is:reply -is:quote lang:en"
```

Рисунок 3.3 – Запит, який ми надіслали через запит GET до API Twitter версії 2, кінцева точка «tweets/search/all»

Запит, який ми використовували для створення нашого набору даних, представлений на рис 3.3. Ми вставили два параметри в запит GET, щоб обмежити часовий простір пошуку з 27 січня по 23 березня, тобто за 1 місяць до і 1 місяць після ініціювання конфлікту (рядки 1 і 2). Наші ключові слова використовуються за логікою АБО (OR), щоб перехопити всі твіти, які містять хоча б один із них (рядки 4 і 5). Ми виключили ретвіти, відповіді та цитати, використовуючи оператор «є» (is), перед яким ставиться тире, щоб заперечити його (рядок 6). Подібним чином ми використали (заперечений) оператор «має» (has), щоб виключити твіти, які містять будь-які посилання або поділяють будь-які медіа (рядок 6). Виконуючи запит, описаний вище, ми зібрали 5 583 168 твітів від 1 858 605 різних користувачів.

Після створення нашого набору даних ми спочатку очистили дані, які повертає Twitter API, щоб видалити твіти, які виходять за рамки цього дослідження або містять будь-що, крім англійського тексту, що може завадити подальшому аналізу. Ми видалили загалом 1144 твіти, 856 з яких, попри те, що були повернуті

Twitter API, насправді не містили жодного з ключових слів, використаних у запиті. Потім ми класифікували облікові записи Twitter на основі різних показників, щоб зрозуміти, які профілі користувачів більше зацікавлені цією темою.

Зокрема, ми розділили всіх користувачів, включених до нашого набору даних, на такі п’ять категорій:

1. Довірені облікові записи: ця категорія включає всі облікові записи, справжня особа власника яких якимось чином відома. Ми помічаємо в цю категорію облікові записи, позначені Twitter як «перевірені» та/або облікові записи з великою кількістю підписників. Зокрема, ми розглядали користувачів з дуже популярним обліковим записом (від 900 тисяч підписників).

2. Дитячі облікові записи: ця категорія включає всі облікові записи, які можна вважати «молодими» під час цього дослідження, тобто облікові записи, створені з вересня 2021 року. Ми вважаємо цю категорію цікавою, оскільки вона може включати підмножину облікових записів, явно створених для твітів про ескалацію російсько-українського конфлікту.

3. Підозрілі облікові записи: ця категорія містить облікові записи, які не поводяться так, як звичайні користувачі. Є підозри, що такими обліковими записами керують боти, тролі та, можливо, інші нетрадиційні користувачі. Щоб ідентифікувати цей тип облікових записів, ми використали стандартну метрику Friend Ratio (FR), яка використовує два з небагатьох показників користувачів, наданих Twitter, – кількість підписників і кількість підписок.

Зокрема, для кожного облікового запису ми обчислюємо коефіцієнт друзів (FR) як відношення кількості підписників/підписок. Після цього обліковий запис міг бути класифікований «підозрілий» за трьох різних сценаріїв, які, на нашу думку, малоймовірні для облікових записів, якими керують приватні особи:

- облікові записи з нульовою підпискою;

- облікові записи з надзвичайно низьким показником FR, тобто менше або дорівнює 0,02, оскільки наявні роботи вважають акаунти з таким показником ботами з високою ймовірністю [35];
- облікові записи з урівноваженим FR, тобто понад 0,99 і менше ніж 1,1, оскільки звичайні користувачі, як правило, мають більшу кількість підписників щодо своїх підписників, оскільки вони стежать за знаменитостями чи іншими популярними обліковими записами, які не слідкують за ними.

4. Невідомі облікові записи: ця категорія включає облікові записи, для яких Twitter не повернув інформацію, оскільки акаунт користувача було призупинено або видалено з платформи на момент збору даних.

5. Звичайні облікові записи: ця категорія включає всі інші облікові записи, які не підпадають під категорії, розглянуті вище.

3.2. Вибір засобів для проведення сентимент-аналізу

Після того як ми зібрали відповідно дані (розділ 3.1), ми використовували аспектний аналіз настроїв (ABSA), який відповідно допомагає визначити точну полярність думок щодо певного аспекту, пов’язаного з певною ціллю. На відміну від загального аналізу настроїв, ABSA може надати більш детальну інформацію, що допомагає виконати більш складний аналіз.

Рисунок 3.4 – Приклад аналізу ABSA

На рис. 3.4 показано, як працює ABSA на прикладі одного речення. У цьому прикладі результатом загального аналізу настрою є «Змішаний», оскільки є один негативний і один позитивний настрій. Однак, як показано на рис. настрої щодо терміна «Зеленський» в аспекті ABSA є позитивними, а щодо «Путіна» – негативними. Таким чином, коли загальний аналіз настроїв дозволяє нам визначити лише переважаючі настрої одного твіту, ABSA дозволяє нам витягти різні настрої, пов'язані з об'єктами, визначеними в одному твіті, таким чином уможливлюючи більш повний, детальний аналіз.

Оскільки наша мета полягає в тому, щоб зрозуміти настрої, пов'язані з різними суб'єктами, залученими протягом зазначеного періоду (починаючи з одного місяця до конфлікту, до одного місяця після), ми вирішили використовувати ABSA, щоб отримати більше інформації з кожного окремого твіту.

З огляду на те, що ABSA все ще є актуальною темою для досліджень у сфері NLP, існує не так багато рішень із відкритим кодом або сторонніх розробників, які надають цю послугу. Тому для цього дослідження ми розглянули два основних рішення:

- бібліотеку Python під назвою aspect-based-sentiment-analysis 2.0.35;
- Amazon AWS Comprehend Targeted Sentiment Analysis6.

AWS Comprehend надає більш детальні результати, наприклад оцінки настрою, тип аспектного терміна, місце розташування аспектного терміна тощо, тоді як бібліотека Python надає лише аспектні терміни та відповідні настрої. Тому ми вирішили провести наші експерименти з AWS Comprehend, який широко використовувався дослідниками [20].

Вхідним набором даних для AWS Comprehend є файл CSV, який містить список текстів твітів, які потрібно проаналізувати. З огляду на обмеження вхідних даних AWS Comprehend у 30 000 рядків у кожному вхідному файлі, ми розділяємо наш набір даних на групу файлів CSV і надсилаємо їх до AWS Comprehend один за одним. Вихідні дані AWS Comprehend – це файли JSON. Для кожного твіту AWS

Comprehend надає необроблений об’єкт JSON, що включає кілька аспектних термінів (суб’єкти, об’єкти, іменники тощо) цього твіту, їхні відповідні настрої та бали настроїв.

{'Entities': [{'DescriptiveMentionIndex': [0, 1, 2], 'Mentions': [{'BeginOffset': 260, 'EndOffset': 266, 'Score': 0.978997, 'GroupScore': 0.999964, 'Text': 'Russia', 'Type': 'ORGANIZATION', 'MentionSentiment': {'Sentiment': 'NEUTRAL', 'SentimentScore': {'Mixed': 0.0, 'Negative': 4e-06, 'Neutral': 0.999996, 'Positive': 0.0}}}, {'BeginOffset': 267, 'EndOffset': 273, 'Score': 0.833712, 'GroupScore': 0.601097, 'Text': '#WWIII', 'Type': 'DATE', 'MentionSentiment': {'Sentiment': 'NEUTRAL', 'SentimentScore': {'Mixed': 0.0, 'Negative': 1e-06, 'Neutral': 0.999998, 'Positive': 0.0}}}, {'BeginOffset': 1, 'EndOffset': 7, 'Score': 0.994837, 'GroupScore': 1.0, 'Text': 'Russia', 'Type': 'ORGANIZATION', 'MentionSentiment': {'Sentiment': 'NEGATIVE', 'SentimentScore': {'Mixed': 0.0, 'Negative': 0.999999, 'Neutral': 0.0, 'Positive': 0.0}}}]}, {'DescriptiveMentionIndex': [0], 'Mentions': [{'BeginOffset': 97, 'EndOffset': 103, 'Score': 0.970907, 'GroupScore': 1.0, 'Text': 'Moscow', 'Type': 'ORGANIZATION', 'MentionSentiment': {'Sentiment': 'NEGATIVE', 'SentimentScore': {'Mixed': 0.0, 'Negative': 1.0, 'Neutral': 0.0, 'Positive': 0.0}}}]}, {'DescriptiveMentionIndex': [0], 'Mentions': [{'BeginOffset': 230, 'EndOffset': 237, 'Score': 0.99953, 'GroupScore': 1.0, 'Text': 'Ukraine', 'Type': 'LOCATION', 'MentionSentiment': {'Sentiment': 'NEUTRAL', 'SentimentScore': {'Mixed': 4e-06, 'Negative': 0.064902, 'Neutral': 0.935079, 'Positive': 1.5e-05}}}]}, {'DescriptiveMentionIndex': [0], 'Mentions': [{'BeginOffset': 245, 'EndOffset': 248, 'Score': 0.92072, 'GroupScore': 1.0, 'Text': 'gas', 'Type': 'OTHER', 'MentionSentiment': {'Sentiment': 'NEUTRAL', 'SentimentScore': {'Mixed': 0.0, 'Negative': 0.0, 'Neutral': 0.999999, 'Positive': 0.0}}}]}], 'File': '2wbb.csv', 'Line': 158498}

Рисунок 3.5 – Необроблений результат ABSA, що показує аспекти та відповідні настрої

На рис 3.5 показано знімок екрана результату виведення AWS Comprehend. Терміни, виділені жовтим, – це визначені аспектні терміни, а терміни, виділені зеленим – відповідні почуття.

```
zelensky/ukraine/ukrainian/russia/russian/putin/NATO
```

Рисунок 3.6 – Суб'єкти, які ми використовували для аспектного аналізу настрою

У цьому дослідженні ми вибрали п’ять ключових слів, наведених рис. 3.6, як основні аспектні терміни, пов’язані з темою, що нас цікавить. Потім ми використали AWS Comprehend ABSA, щоб отримати пов’язані з ними настрої з усіх твітів у нашому наборі даних. Для цього ми спочатку надіслали наш набір даних (5 583 168 твітів у форматі файлу CSV) на платформу AWS Comprehend як вхідні дані. Далі ми отримали відповідні результати ABSA (терміни аспектів, настрої, оцінки настроїв тощо) як вихідні дані.

Для обробки необроблених результатів аналізу ABSA ми використовуємо спеціальні сценарії на Python. Зокрема, для кожного твіту ми фільтруємо аспектні терміни, визначені AWS Comprehend, щоб отримати лише результати, наприклад, настрої та бали настроїв, пов’язані з нашими суб’єктами (рис 3.6). Таким чином, для кожного предмета (аспектного терміна) ми маємо всі твіти, у яких він згадується з відповідним настроєм.

Щоб уникнути кількох настроїв щодо одного конкретного аспекту в одному твіті, ми визначаємо такі правила відбору:

1. Якщо є 2 різні настрої для терміна аспекту:

- Нейтральний і Позитивний: результат є Позитивним
- Нейтральний і Негативний: результат є негативним
- Негативний і Позитивний: відкинути настрої взагалі.

2. Якщо є 3 різні настрої для певного аспекту, ми відкидаємо настрої.

Ці правила відбору гарантують, що кожен твіт створює лише один настрій для кожного аспекту нашого дослідження.

Надалі файли CSV із результатами сортуються в хронологічному порядку та наносяться на графік, щоб відобразити настрої аспектів у часі (наприклад, щогодини, щодня тощо). На останньому етапі постобробки ми нормалізуємо вміст (#нейтральний, #негативний, #позитивний) отриманих файлів CSV, щоб перетворити кожне значення в число від 0 до 1. Це завдання полегшує порівняння

різних графіків і вилучення інформації, що можна буде побачити у результатах в розділі 3.3).

Варто зазначити, що методи ABSA все ще мають недоліки, на які інші дослідники повинні звернути увагу у майбутньому. Так, наприклад у нашому дослідженні 10% настроїв відсутні, оскільки AWS Comprehend ABSA не зміг виявити відповідні терміни аспектів із вхідного набору даних взагалі.

3.3. Результати сентимент-аналізу твітів на тему війни в Україні

Виконуючи запит, описаний на рис. 3.2, ми зібрали 5 583 168 повідомлень, опублікованих 1 858 605 різними користувачами в Twitter протягом розглянутого періоду, тобто за місяць до і через місяць після початку збройного конфлікту, який стався 24 лютого 2022 року., щоб охарактеризувати російсько-український конфлікт 2022 року в Twitter. Далі ми проаналізували наш набір даних за допомогою кількох показників.

Спочатку ми підрахували, скільки разів наші ключові слова використовувалися в зібраних твітах, щоб визначити найпопулярніші теми (табл. 3.1).

Таблиця 3.1 – Кількість твітів, отриманих за нашими ключовими словами

Ключове слово	**Кількість твітів**	**Кількість ексклюзивних твітів**
Ukraine	1,907,016	1,202,640

Продовження таблиці 3.1 – Кількість твітів, отриманих за нашими ключовими словами

Ключове слово	**Кількість твітів**	**Кількість ексклюзивних твітів**
Putin	1,594,509	712,669
Russia	1,364,853	512,946
Russian	1,250,835	494,458
Ukrainian	544,308	166,628
Zelensky	193,683	60,028
NATO	330,638	0

Результати цього аналізу зображено в таблиці 3.1, де другий стовпець містить загальну кількість твітів, які містять виділене ключове слово, а третій стовпець містить кількість твітів, де це ключове слово використовується окремо, тобто без інших шести. Україна є, безперечно, найбільш використовуваним ключовим словом у нашому наборі даних, як у поєднанні з іншими (34% усього набору даних), так і окремо (22%).

У дискусії у Твіттері також звучали ключові слова, пов’язані з Росією, тобто Путін, Росія та російський (russian) відповідно. Натомість інші ключові слова, пов’язані з Україною, використовуються набагато рідше. Зеленський, наприклад, був названий лише в 3% зібраних твітів (лише 1%). Загалом ключові слова, пов’язані з Україною, були використані (тільки) в 1 429 296 твітах, що становить 26% від усього набору даних, тоді як ключові слова, пов’язані з Росією, використовувалися в 1 720 073 твітах, тобто 31%.

НАТО натомість було названо в 330 638 твітах, 6% від загальної кількості твітів, завжди в поєднанні з іншими ключовими словами. Важливо зазначити, що за задумом наш набір даних не включає жодного твіту лише з ключовим словом НАТО, як описано в розділі 3.1.

Обсяг твітів, пов’язаних з російсько-українським конфліктом 2022 року, опублікованих протягом досліджуваного періоду, наведено на рисунку 3.7.

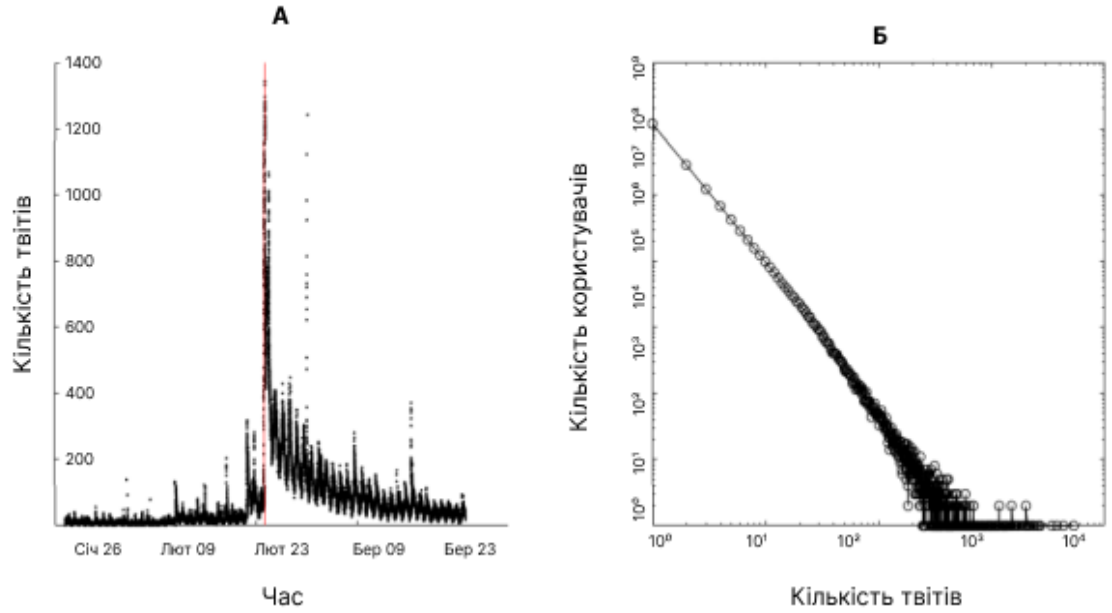


Рисунок 3.7 – Обсяг твітів, пов’язаних з російсько-українським конфліктом 2022 року, з 27 січня по 23 березня. Загальна кількість твітів за хвилину (А) і кількість твітів на обліковий запис (Б) у логарифмічно масштабованих осях; червона вертикальна лінія – 24 лютого

Протягом першого місяця, який розглядається, кількість твітів на хвилину стабільно становила близько 20, за винятком кількох окремих спалахів, як показано на рис 3.7-А. Потім він почав зростати приблизно за тиждень до центральної події в нашому наборі даних: прес-конференції, на якій президент Росії оголосив про початок військової спеціальної операції в Україні. Того дня кількість повідомлень, опублікованих у Twitter, раптово перевищила 1300 за хвилину. Згодом, протягом

наступних тижнів, він поступово зменшувався, поки не встановився між 20 і 100 твітами на хвилину. Переважна більшість користувачів опублікувала дуже мало твітів у розглянутому часовому вікні, як показано на малюнку 3.7-Б.

Таблиця 3.2 – Кількість користувачів, твітів і середня кількість твітів на користувача для кожної категорії

Тип акаунту		**Кількість акаунтів**	**Кількість твітів**	**Середня кількість твітів на одного користувача**
Дитячі		237,814	718,163	3.02
Довірені		7,601	80,400	41.50
Підозрілі	0 підсник.	0 підсник.	0 підсник.	0 підсник.
	Врівн. FR	83,589	285,023	3.41
	Низьк. FR	28,323	55,821	1.97
Невідомі	Заблоковані	7,166	62,783	8.76
	Видалені	1,358	3,366	2.48
Звичайні		1,489,940	4,360,028	2.93

Після категоризації користувачів, розглянутої в розділі 3.2, ми подивилися, наскільки кожна категорія брала участь в обговоренні ескалації конфлікту. У таблиці 3.2 наведено кількість користувачів, загальну кількість твітів і середню кількість твітів на користувача для кожної категорії – і підкатегорії, якщо такі є.

Користувачі, які зробили найбільший внесок в обговорення, є довіреними обліковими записами. Зокрема, знаменитості писали твіти в середньому 41 раз, тоді як дуже популярні та перевірені облікові записи публікували твіти 16 і 10 разів відповідно. Це спостереження не є дивним, враховуючи, що багато інформаційних агентств потрапляють до цієї категорії. Натомість звичайні облікові записи, якими, як передбачається, володіють і керують звичайні користувачі, твітнули в середньому лише двічі, відповідно до того, що можна побачити на рис. 3.7-Б.

Важливо зазначити, що аномальні (підозрілі) облікові записи, зокрема ті, що мають низький FR, суттєво відрізняються від звичайних облікових записів, оскільки в середньому вони твітнули близько восьми разів. Ці облікові записи можуть бути залучені до діяльності з дезінформації, і їх аналіз слід поглибити в майбутніх роботах. Ще одна особливо цікава категорія – дитячі облікові записи. Як описано в Розділі 3.1, ця категорія включає всі облікові записи, створені в період близького до розглядуваного часового вікна, тобто з 27 січня по 23 березня. Тому дуже ймовірно, що підгрупа цих акаунтів була явно створена для твітів про російсько-український конфлікт спеціально (рис. 3.8).

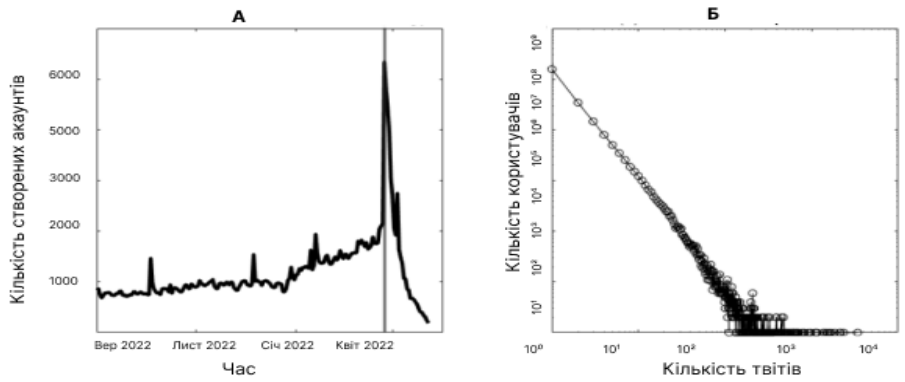


Рисунок 3.8 – Кількість створених облікових записів дітей (щодня) з вересня 2021 року по березень 2022 року (А) і кількість твітів на обліковий запис дитини (Б) у логарифмічно масштабованих осях

На малюнку 3.8-А показано кількість облікових записів, створених за певний час. У перший аналізований період створювалося близько 1000 акаунтів на день. Потім рівень створення збільшується приблизно за 1 місяць до 24 лютого (сіра вертикальна лінія), щоб досягти понад 5000 облікових записів, створених у цей день. Потім число раптово зменшується до значень, набагато менших за медіану до 24 лютого. Рисунок 3.8-Б замість цього показує кількість твітів на дитячий обліковий запис. Подібно до звичайних, переважна більшість дитячих облікових записів публікували дуже мало твітів у розглянутому часовому вікні відповідно до степеневого розподілу.

Застосування ABSA до набору даних російсько-українського конфлікту 2022 (твіти за 2 місяці) дає можливість отримати та представити приховану інформацію, яку нелегко досягти за допомогою інших методів сентимент-аналізу. Основна мета — дослідити настрої користувачів Twitter щодо основних гравців, залучених до ескалації конфлікту, і як ці настрої змінювалися з часом. Враховуючи наші технічні обмеження, ми повідомляємо лише про найбільш релевантні результати, тобто пов’язані з підмножиною ключових слів і категорій користувачів.

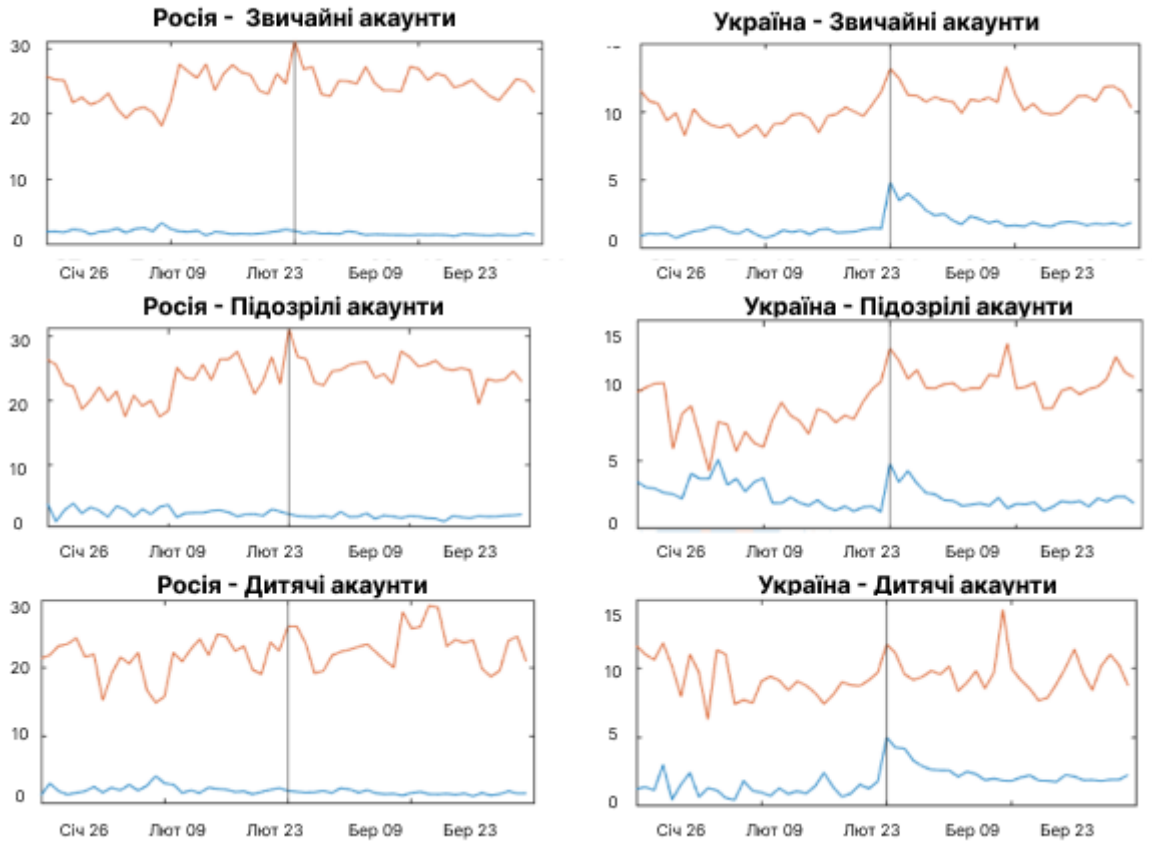


Рисунок 3.9 – Настрої щодо термінів «Росія» та «Україна» з часом; сіра вертикальна лінія – 24 лютого.

Рисунок 3.9 ілюструє наші результати для аспектів «Росія» та «Україна» в часі за 1 місяць до та 1 місяць після ескалації конфлікту, тобто 24 лютого. Еволюція настроїв для 2 розглянутих аспектів є досить різною. Щодо «України» кількість позитивних настроїв значно зростає після 24 лютого. Натомість ті самі настрої щодо «Росії» здаються відносно постійними протягом двох розглянутих місяців, не зазнаючи жодних змін навколо 24 лютого.

З іншого боку, кількість негативних настроїв зростає до «Росії» після ескалації конфлікту, тоді як для України темп зростання негативних настроїв набагато нижчий. Для аномальних (підозрілих) облікових записів графіки суттєво

відрізняються. Для аспекту «Україна» рівень позитивних настроїв є дуже високим, тоді як рівень негативних настроїв дуже низький.

Рисунок 3.10 – Настрої для суб’єктів «Путін» і «Зеленський» у часі; сіра вертикальна лінія – 24 лютого.

Рисунок 3.10 ілюструє результати для аспектів «Путін» та «Зеленський» у часі. Графіки та моделі еволюції абсолютно різні для двох аспектних термінів. Для «Путіна» нейтральний і негативний рейтинги зближуються після 24 лютого. Ми вважаємо, що нейтральні настрої в основному відображають новини та інформацію, пов’язану з цим конфліктом. Після ескалації конфлікту почуття людей до «Путіна» масово стають негативними. Щоправда, позитивний показник дуже низький в обох аспектах, а не лише у «путінському».

У кожній категорії можна побачити, що позитивні настрої здаються відносно постійними протягом усього періоду, що розглядається, з дуже невеликим піком 24 лютого. Натомість негативний і нейтральний тренди врівноважуються. Протягом майже всього періоду, що розглядається, коли один з двох зростає, інший падає приблизно на однакову величину, і навпаки. Це свідчить про те, що сильні

прихильники президента Путіна залишалися на одній думці протягом досліджуваного періоду, тоді як нейтральні настрої стали переважно негативними після ескалації конфлікту.

Іншим висновком є той факт, що президент Зеленський, попри те, що він має постійну популярність у стандартних медіа (телебачення, газети), не зміг зрівнятися з популярністю у Twitter. Дійсно, хоча загальні настрої щодо нього були позитивні, вони не досягають значних висот серед англомовних користувачів цієї соціальної мережі.

Нарешті, варто зауважити, що хоча настрої щодо президента Путіна загалом стали негативними, настрої щодо Росії не зазнали такого ж падіння популярності.

Варто також зазначити, що це дослідження містить деякі обмеження, які можна розглянути в майбутніх роботах.

По-перше, оскільки наш аналіз спрямований лише твіти англійською мовою. Аналіз даних іншими мовами може дозволити виявити нові тренди та особливо обробки повідомлень.

По-друге, наявні методи ABSA не здатні розрізнити різні емоції одного і того ж почуття, наприклад, радість і смуток, любов і ненависть. У результаті, якщо конкретний твіт є, наприклад, негативним як для Росії, так і для України, наш аналіз не показує, чи автор злий на Росію, але сумує за Україну, чи навпаки. Використання більш досконалих методів ABSA підвищить ефективність цього дослідження.

По-третє, категоризація користувачів базується лише на кількох атрибутах користувача, таких як кількість підписників та підписок. Використання більш просунутих методів може допомогти покращити поширені сьогодні категорії та/або створити нові.

Висновки до розділу 3

У цьому розділу ми розробили новий підхід для обробки повідомлень Twitter на основі методів сентимент-аналізу та проаналізували російсько-український конфлікт 2022 року з його допомогою. В основному ми досліджували обсяг даних і сприйняття конфлікту громадськістю, використовуючи методи статистичного аналізу та аналіз настроїв на основі аспектів (ABSA).

Ми виявили кілька цікавих аномалій у поведінці та настроях користувачів щодо деяких тем, які вимагають подальших досліджень у цій галузі. Зокрема, облікові записи з низьким коефіцієнтом друзів (FR) твітнули набагато більше, ніж звичайні користувачі, з тенденцією настрою для деяких ключових слів, яка відрізнялася від інших користувачів. Крім того, попри те, що нещодавно створені облікові записи поводяться як звичайні користувачі, їх щоденна швидкість створення свідчить про те, що їх підмножина була створена спеціально для твітів про конфлікт.

ВИСНОВКИ

У цій роботі розглядається потенціал використання інформаційної технології обробки повідомлень, метою якої є розуміння й аналіз емоцій, виражених у тексті. Розроблений нами метод об’єднує дослідження з різних галузей, зокрема інформатики, інтелектуального аналізу даних, пошуку та вилучення тексту та комп’ютерної лінгвістики.

1. Встановлено, що сентимент-аналіз – це область дослідження, яка спрямована на розуміння й аналіз емоцій і ставлень, виражених у тексті. Це швидко зростаюча сфера, яка залучає дослідників з різних дисциплін, включаючи інформатику, аналіз даних, пошук та вилучення тексту та комп’ютерну лінгвістику.

2. Оцінено 19 інструментів для сентимент-аналізу, яка загалом можна згрупувати у дві великі категорії: автономні комерційні інструменти та навчені робочі інструменти. Ці інструменти здебільшого включають пропозиції на основі API. Однак відсутність специфіки домену та те, що сучасні методи сентимент-аналізу погано розрізнити різні емоції, які підпадають під одну категорію почуттів заважають їх ефективному використанню

3. Розроблено технологію, яка базується як на методах кількісного, так і якісного дослідження тексту та дозволяє використовувати наявні методи інтелектуального аналізу даних та аналізу тексту для обробки повідомлень у соціальних мережах, зокрема Twitter.

4. Визначено, що природа соціальних мереж і Twitter зокрема виступає проти простого та доступного аналізу настроїв: під час проведення аналізу необхідно враховувати чисельні обмеження Twitter API: мультилінгвальність, розміри, мовні особливості обробки твітів тощо.

5. Розроблено схему нової методики обробки повідомлень на основі сентимент-аналізу у поєднання з методами машинного навчання, а саме на основі методів статистичного аналізу та аналізу настроїв на основі аспектів (ABSA).

6. Проаналізовано настрої повідомлень у Twitter на тему війни в Україні. Для цього було досліджено твітів за один місяць до 24 лютого 2022 року і один місяць після. Можна побачити, що позитивні настрої здаються відносно постійними протягом усього періоду, що розглядається, з дуже невеликим піком 24 лютого. Натомість негативний і нейтральний тренди врівноважуються. Протягом майже всього періоду, що розглядається, коли один з двох зростає, інший падає приблизно на однакову величину, і навпаки. Це свідчить про те, що сильні прихильники президента Путіна та дій Росії у цій війні залишалися вірні своїм уподобанням протягом досліджуваного періоду, тоді як нейтральні настрої стали переважно негативними після ескалації конфлікту.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Іващенко Р., Рєзіна О. Аналіз тональності та об'єктивності тексту засобами мови програмування python. URL: https://phm.cuspu.edu.ua/ojs/index.php/SNYS/article/download/1938/1948.
2. Кохан В. В. Б., Kokhan V. V. Алгоритмічне та програмне забезпечення систем автоматизованого оцінювання емоційного нахилу статей про Україну : Master Thesis. 2021. URL: http://elartu.tntu.edu.ua/handle/lib/36642 (дата звернення: 27.12.2022).
3. Марченко А. В. Комп’ютерна автоматизована система визначення тональності тексту : Магістерська робота. 2020. URL: https://dspace.znu.edu.ua/jspui/handle/12345/4999 (дата звернення: 27.12.2022).
4. Мироненко С., Онищенко Є. Порівняльний аналіз методів для вирішення задачі сентимент аналізу тексту. Computer-integrated technologies: education, science, production. 2020. № 40. С. 140–145. URL: https://doi.org/10.36910/6775-2524-0560-2020-40-21 (дата звернення: 25.12.2022).
5. Огляд методів обробки та аналізу текстів на природних мовах / С. І. Доценко та ін. Інформаційно-керуючі системи на залізничному транспорті. 2018. № 6. С. 26–32. URL: https://doi.org/10.18664/ikszt.v0i6.151638 (дата звернення: 25.12.2022).
6. Пентилюк М. Наукові параметри аналізу тексту. Дивослово. 2017. № 9 (726), верес. С. 36–41.
7. Сентимент аналіз засобами нейронної мережі / К. Ялова та ін. Математичне моделювання. 2021. № 1(44). С. 30–37. URL: https://doi.org/10.31319/2519-8106.1(44)2021.235906 (дата звернення: 27.12.2022).
8. Серажим К. Інформаційний аспект аналізу тексту. Журналістика. 2008. Вип. 7 (32). С. 40–48.

9. Шингалов Д. В., Тріщ О. В., Минайленко Р. М. Методи автоматичного аналізу настроїв в соціальних мережах : Thesis. 2017. URL: http://dspace.kntu.kr.ua/jspui/handle/123456789/7489 (дата звернення: 27.12.2022).
10. A Study on text mining and text mining products. International journal of science, technology and humanities. 2014. Vol. 1, no. 1. P. 61–63. URL: https://doi.org/10.26524/ijsth11 (date of access: 25.12.2022).
11. A survey on twitter sentiment analysis / E.-U. Rahman et al. International journal of computer sciences and engineering. 2018. Vol. 6, no. 11. P. 644–648. URL: https://doi.org/10.26438/ijcse/v6i11.644648 (date of access: 27.12.2022).
12. Algorithmic amplification of politics on Twitter / F. Huszár et al. Proceedings of the national academy of sciences. 2021. Vol. 119, no. 1. URL: https://doi.org/10.1073/pnas.2025334119 (date of access: 27.12.2022).
13. Altenberg B. A bibliography of publications relating to English computer corpora. English computer corpora. Berlin, Boston. URL: https://doi.org/10.1515/9783110865967.355 (date of access: 25.12.2022).
14. Blake C. Text mining. Annual review of information science and technology. 2011. Vol. 45, no. 1. P. 121–155. URL: https://doi.org/10.1002/aris.2011.1440450110 (date of access: 25.12.2022).
15. Drus Z., Khalid H. Sentiment analysis in social media and its application: systematic literature review. Procedia computer science. 2019. Vol. 161. P. 707–714. URL: https://doi.org/10.1016/j.procs.2019.11.174 (date of access: 27.12.2022).
16. Duncombe C. The politics of twitter: emotions and the power of social media. International political sociology. 2019. Vol. 13, no. 4. P. 409–429. URL: https://doi.org/10.1093/ips/olz013 (date of access: 27.12.2022).
17. Eriksson M. Lessons for crisis communication on social media: a systematic review of what research tells the practice. International journal of strategic

communication. 2018. Vol. 12, no. 5. P. 526–551. URL: https://doi.org/10.1080/1553118x.2018.1510405 (date of access: 27.12.2022).

18. Exploring neural question generation for formal pragmatics: data set and model evaluation. Frontiers. URL: https://www.frontiersin.org/articles/10.3389/frai.2022.966013/full (date of access: 24.12.2022).
19. Finegan E., Johansson S. Computer corpora in english language research. Language. 1984. Vol. 60, no. 1. P. 190. URL: https://doi.org/10.2307/414219 (date of access: 25.12.2022).
20. Garcia M. B. Sentiment analysis of tweets on coronavirus disease 2019 (COVID-19) pandemic from metro manila, philippines. Cybernetics and information technologies. 2020. Vol. 20, no. 4. P. 141–155. URL: https://doi.org/10.2478/cait-2020-0052 (date of access: 27.12.2022).
21. Google trends. URL: https://trends.google.com/trends/ (date of access: 25.12.2022).
22. Karpicke J. D., Grimaldi P. J. Retrieval-Based learning: a perspective for enhancing meaningful learning. Educational psychology review. 2012. Vol. 24, no. 3. P. 401–418. URL: https://doi.org/10.1007/s10648-012-9202-2 (date of access: 25.12.2022).
23. Korney A. O., Kryuchkova E. N. Semantic-statistical algorithm for determining the categories of aspects in the problems of sentiment analysis. IZVESTIYA SFedU. ENGINEERING SCIENCES. 2021. No. 6. P. 66–74. URL: https://doi.org/10.18522/2311-3103-2020-6-66-74 (date of access: 27.12.2022).
24. Kumar A., Sharma S., Singh D. Sentiment analysis on twitter data using a hybrid approach. International journal of computer sciences and engineering. 2019. Vol. 7, no. 5. P. 906–911. URL: https://doi.org/10.26438/ijcse/v7i5.906911 (date of access: 27.12.2022).
25. Kyrychenko R. Типологія задач машинного аналізу текстів у сучасній соціології. Sociological studios. 2021. № 2(19). С. 53–62. URL:

https://doi.org/10.29038/2306-3971-2021-02-41-48 (дата звернення: 27.12.2022).

26. Lambert C. E. Earthquake country: a qualitative analysis of risk communication via facebook. Environmental communication. 2020. Vol. 14, no. 6. P. 744–757. URL: https://doi.org/10.1080/17524032.2020.1719176 (date of access: 27.12.2022).
27. Levchenko O., Povoroznik N. Features of sentiment analysis implementation (interpretation of sarcasm, word ambiguity, negation, and multipolarity). Young scientist. 2020. Vol. 11, no. 87. URL: https://doi.org/10.32839/2304-5809/2020-11-87-98 (date of access: 27.12.2022).
28. NLP-progress repository. NLP-progress. URL: https://nlpprogress.com/english/question_answering.html (date of access: 24.12.2022).
29. Nair R. B., E K S. Sentiment analysis using twitter data. International journal of computer trends and technology. 2019. Vol. 67, no. 5. P. 82–84. URL: https://doi.org/10.14445/22312803/ijctt-v67i5p114 (date of access: 27.12.2022).
30. Perspectives of semantic web in E commerce / B. VijayaLakshmi et al. International journal of computer applications. 2011. Vol. 25, no. 10. P. 52–56. URL: https://doi.org/10.5120/3172-4166 (date of access: 27.12.2022).
31. Rai S., S. B G., Kumar J. Sentiment analysis of twitter data. International research journal on advanced science hub. 2021. Vol. 2, Special Issue ICSTM 12S. P. 56–61. URL: https://doi.org/10.47392/irjash.2020.261 (date of access: 27.12.2022).
32. Srivastava A., Singh V., Drall G. S. Sentiment analysis of twitter data. International journal of healthcare information systems and informatics. 2019. Vol. 14, no. 2. P. 1–16. URL: https://doi.org/10.4018/ijhisi.2019040101 (date of access: 27.12.2022).
33. Twitter sentiment analysis approaches: a survey / O. Y. Adwan et al. International journal of emerging technologies in learning (ijet). 2020. Vol. 15, no. 15. P. 79. URL: https://doi.org/10.3991/ijet.v15i15.14467 (date of access: 27.12.2022).

34. Twitter sentiment analysis using machine learning techniques. International journal of engineering and advanced technology. 2020. Vol. 9, no. 3. P. 4205–4209. URL: https://doi.org/10.35940/ijeat.c6281.029320 (date of access: 27.12.2022).
35. Udodenko O., Dovgopoly D., Ivanenko V. Dynamics of emotional coloring of texts of venture investor publications in the “Twitter” social network. Bulletin of taras shevchenko national university of kyiv. economics. 2020. No. 211. P. 62–70. URL: https://doi.org/10.17721/1728-2667.2020/211-4/6 (date of access: 27.12.2022).

ДОДАТОК А
ДЕМОНСТРАЦІЙНІ МАТЕРІАЛИ

ДЕРЖАВНИЙ УНІВЕРСИТЕТ ТЕЛЕКОМУНІКАЦІЙ

НАВЧАЛЬНО–НАУКОВИЙ ІНСТИТУТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

Кафедра інженерії програмного забезпечення

МАГІСТЕРСЬКА РОБОТА

«РОЗРОБКА СИСТЕМИ ГОЛОСОВОГО УПРАВЛІННЯ РОЗУМНИМ БУДИНКОМ З ВИКОРИСТАННЯМ ШТУЧНОГО ІНТЕЛЕКТУ»

Виконав: студент групи ПДМ-62 Чуб Євгеній Михайлович

Керівник: к.т.н., доц., доцент кафедри ІПЗ, Трінтіна Н.А.

Київ - 2022

МЕТА, ОБ'ЄКТ, ПРЕДМЕТ ДОСЛІДЖЕННЯ

Мета дослідження: удосконалення наявних методів використання алгоритмів сентимент аналізу для аналізу неструктурованих текстових даних.

Об'єкт дослідження: процес аналізу тональності тексту.

Предмет дослідження: методи сентимент аналізу.

АНАЛІЗ НАЯВНИХ МЕТОДІВ ТА МОДЕЛЕЙ

	Класифікація	Категоризація	Кластеризація
Ціль	Автоматична класифікація для електронних листів і веб-сторінок	Тегування документів	Групування та формування таксономії
Процес / модель	• Класифікація нових/невідомих текстових документів • Визначення навчального набору та тестових наборів	• Методи машинного навчання • На основі векторної моделі простору • Нейронні мережі	• Попередня обробка даних • Видалення стоп-слів • Визначення основи • Виділення ознак • Лексичний аналіз
Алгоритм	• Наївний Байєс • Нейронні мережі • Дерева рішень • На основі правил асоціації	• Наївний Байєс • Нейронні мережі • Дерева рішень • На основі правил асоціації	• Послідовні алгоритми • Краяркічні алгоритми • Агломераційні алгоритми • Розділові алгоритми • Алгоритми нечіткого люстрування
Сфера використ-тання	• Інтерфейс електронної комерції (Amazon, Ebay) • Сітка медичної області • Геодемографічна класифікація ACORN • Видобуток даних	• Веб-сторінки • Статті у журналах (індексація) • Патентні архіви • Фільтри в службах електронної пошти	• Кдасифікація паттернів • Сегментація зображень та тексту на них • Вилучення даних (економічні науки)

3

АСПЕКТНИЙ АНАЛІЗ НАСТРОЇВ (ABSA)

4

{'Entities': [{'DescriptiveMentionIndex': [0, 1, 2], 'Mentions': [{'BeginOffset': 260, 'EndOffset': 266, 'Score': 0.978997, 'GroupScore': 0.999964, 'Text': 'Russia', 'Type': 'ORGANIZATION', 'MentionSentiment': {'Sentiment': 'NEUTRAL', 'SentimentScore': {'Mixed': 0.0, 'Negative': 4e-06, 'Neutral': 0.999996, 'Positive': 0.0}}}, {'BeginOffset': 267, 'EndOffset': 273, 'Score': 0.833712, 'GroupScore': 0.601097, 'Text': '#WWIII', 'Type': 'DATE', 'MentionSentiment': {'Sentiment': 'NEUTRAL', 'SentimentScore': {'Mixed': 0.0, 'Negative': 1e-06, 'Neutral': 0.999998, 'Positive': 0.0}}}, {'BeginOffset': 1, 'EndOffset': 7, 'Score': 0.994837, 'GroupScore': 1.0, 'Text': 'Russia', 'Type': 'ORGANIZATION', 'MentionSentiment': {'Sentiment': 'NEGATIVE', 'SentimentScore': {'Mixed': 0.0, 'Negative': 0.999999, 'Neutral': 0.0, 'Positive': 0.0}}}]}, {'DescriptiveMentionIndex': [0], 'Mentions': [{'BeginOffset': 97, 'EndOffset': 103, 'Score': 0.970907, 'GroupScore': 1.0, 'Text': 'Moscow', 'Type': 'ORGANIZATION', 'MentionSentiment': {'Sentiment': 'NEGATIVE', 'SentimentScore': {'Mixed': 0.0, 'Negative': 1.0, 'Neutral': 0.0, 'Positive': 0.0}}}]}, {'DescriptiveMentionIndex': [0], 'Mentions': [{'BeginOffset': 230, 'EndOffset': 237, 'Score': 0.99953, 'GroupScore': 1.0, 'Text': 'Ukraine', 'Type': 'LOCATION', 'MentionSentiment': {'Sentiment': 'NEUTRAL', 'SentimentScore': {'Mixed': 4e-06, 'Negative': 0.064902, 'Neutral': 0.935079, 'Positive': 1.5e-05}}}]}, {'DescriptiveMentionIndex': [0], 'Mentions': [{'BeginOffset': 245, 'EndOffset': 248, 'Score': 0.92072, 'GroupScore': 1.0, 'Text': 'gas', 'Type': 'OTHER', 'MentionSentiment': {'Sentiment': 'NEUTRAL', 'SentimentScore': {'Mixed': 0.0, 'Negative': 0.0, 'Neutral': 0.999999, 'Positive': 0.0}}}]}], 'File': '2wbb.csv', 'Line': 158498}

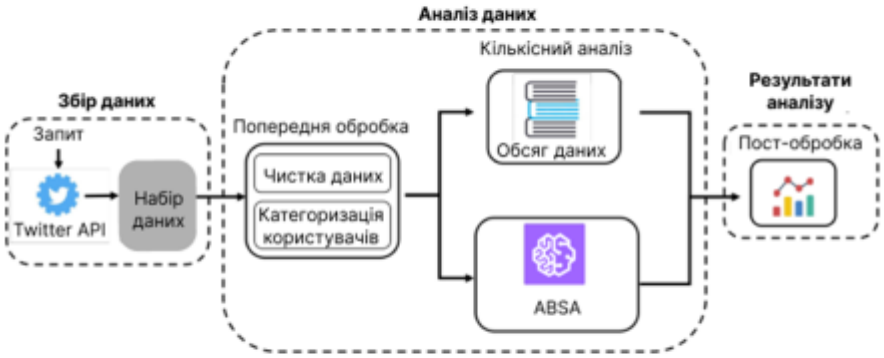
Необроблений результат ABSA, що показує аспекти та відповідні настрої

5

ПРАВИЛА ВИЗНАЧЕНІ НА ОСНОВІ ПРИНЦИПІВ РОБОТИ AWS Comprehend

6

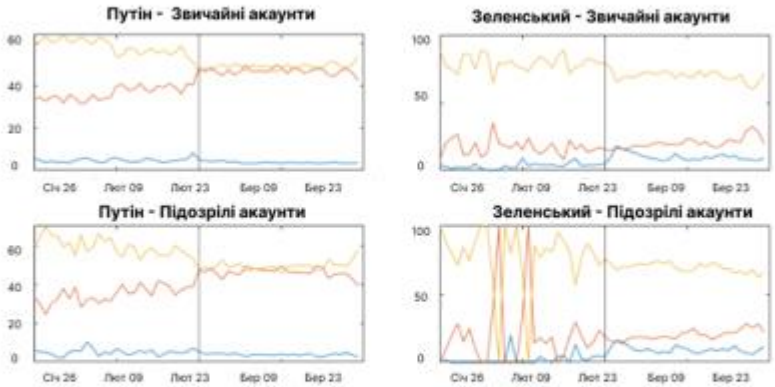
УЗАГАЛЬНЕНИЙ ПРАКТИЧНИЙ РЕЗУЛЬТАТ



Алгоритм роботи розробленої моделі сентимент-аналізу повідомлень у Twitter

7

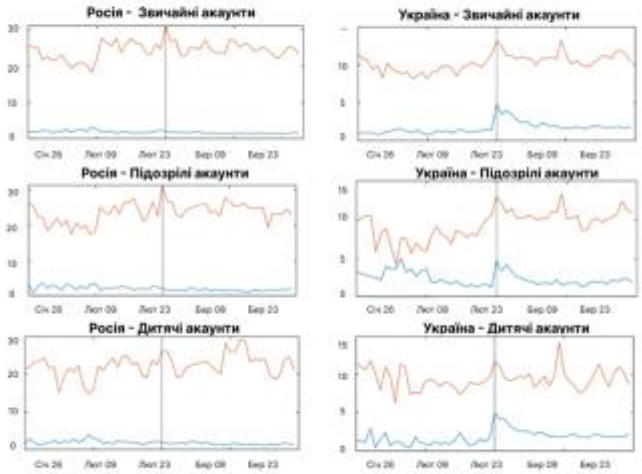
РЕЗУЛЬТАТИ МОДЕЛЮВАННЯ



Настрої для суб'єктів «Путін» і «Зеленський» у часі; сіра вертикальна лінія – 24 лютого

8

РЕЗУЛЬТАТИ МОДЕЛЮВАННЯ



9

ВИСНОВКИ

1. Проаналізовано область застосування аналізу тональності тексту, а також його основні завдання та цілі.

2. Проведено аналіз існуючих інструментів для аналізу тональності тексту.

3. Обрано соціальну мережу Twitter, як платформу для збору неструктурованих даних для проведення сентимент аналізу, та визначено проблеми та обмеження Twitter API.

4. Розроблено схему методики обробки повідомлень на основі сентимент-аналізу у поєднання з методами машинного навчання, а саме на методах статистичного аналізу та аналізу настроїв на основі аспектів (ABSA).

5. Проведено аналіз настроїв повідомлень у соціальній мережі Twitter на тему російсько-української війни за один місяць до 24 лютого та один місяць після.

10

ПУБЛІКАЦІЇ ТА АПРОБАЦІЯ РОБОТИ

Тези доповідей:

1. Чуб Є.М. Сентимент-аналіз повідомлень за допомогою обробки природної мови на прикладі Twitter//XV Конференції «Сучасні інфокомунікаційні технології", 9 грудня 2022 року, Державний університет телекомунікації, Київ, Україна.
2. Чуб Є.М. Сентимент аналіз: переваги, недоліки та методи оцінювання//Міжнародна науково-практична конференція «Актуальні питання забезпечення кібербезпеки та захисту інформації», 22-23 січня 2023, Національний Авіаційний Університет Київ, Україна.

11

ДЯКУЮ ЗА УВАГУ!

12