Міністерство освіти і науки України

Харківський національний університет радіоелектроніки

Факультет навчально-науковий центр заочної форми навчання
(повна назва)

Кафедра електронних обчислювальних машин
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА

Пояснювальна записка

Рівень вищої освіти другий (магістерський)

Метод аналізу тональності цитування публікації на основі контексту

(тема)

Виконав:

студент II курсу, групи СПзм-21-1

Рисенко О. С.
(прізвище, ініціали)

Спеціальність

123 «Комп'ютерна інженерія»
(код і повна назва спеціальності)

Тип програми освітньо-наукова
(освітньо-професійна або освітньо-наукова)

Освітня програма

Системне програмування
(повна назва освітньої програми)

Керівник: доц. Ільїна І. В.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри ЕОМ _____ Коваленко А.А.
(підпис) (прізвище, ініціали)

2023 р.

Харківський національний університет радіоелектроніки

Факультет навчально-науковий центр заочної форми навчання

Кафедра електронних обчислювальних машин

Рівень вищої освіти другий (магістерський)

Спеціальність 123 «Комп'ютерна інженерія»
(код і повна назва)

Тип програми освітньо-наукова
(освітньо-професійна або освітньо-наукова)

Освітня програма Системне програмування
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
"____" _____ 20__ р.

ЗАВДАННЯ

НА КВАЛІФІКАЦІЙНУ РОБОТУ

студенту Рисенку Олександру Сергійовичу
(прізвище, ім'я, по батькові)

1. Тема роботи Метод аналізу тональності цитування публікації на основі контексту

затверджена наказом по університету від " 24 " березня 2023 р. № 60Стз

2. Термін подання студентом роботи до екзаменаційної комісії 17 травня 2023 р.

3. Вхідні дані до роботи набір даних для експерементів

4. Перелік питань, що потрібно опрацювати у роботі

Проаналізувати існуючі методи та рішення для визначення тональності текстів.

Вивчити можливості збільшення ефективності існуючих методів.

Проаналізувати ефективність запропонованих покращень.

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (слайдів) 18 слайдів

6. Консультанти розділів роботи (заповнюється за наявності консультантів згідно з наказом, зазначеним у п.1)

Найменування розділу	Консультант (посада, прізвище, ім'я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Термін виконання етапів роботи	Примітка
1	Огляд літератури	30.03.23-05.04.23	
2	Аналіз алгоритмів класифікації	06.04.23-16.04.23	
3	Огляд методів та рішенньь для визначення	17.04.23-22.04.23	
	тональності текстів		
4	Розробка методу визначення тональності	23.04.23-29.04.23	
5	Проведення експеременту	30.04.23-07.05.23	
6	Аналіз результатів	08.05.23-11.05.23	
7	Подання кваліфікаційної роботи керівникові	12.05.23-13.05.23	
	її попередній захист		
8	Подання кваліфікаційної роботи на	14.05.23-15.05.23	
	рецензування		

Дата видачі завдання 27 березня 2023 р.

Студент _____
(підпис)

Керівник роботи _____ доц. Ільїна І.В.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ

Пояснювальна записка кваліфікаційної роботи: 70 с., 16 рис., 7 табл., 1 дод., 17 джерел.

ОБРОБКА ПРИРОДНОЇ МОВИ, АНАЛІЗ ТОНАЛЬНОСТІ ТЕКСТУ, НАІВНИЙ БАЙЄСОВИЙ КЛАСИФІКАТОР, МЕТОД ОПОРНИХ ВЕКТОРІВ.

Метою кваліфікаційної роботи є дослідження і розробка методу аналізу тональності цитат в наукових публікаціях.

Для досягнення поставленої мети необхідно вирішити такі завдання:

- провести огляд існуючих методів автоматичного аналізу емоційного забарвлення текстів;
- провести дослідження текстових особливостей цитат в наукових публікаціях в контексті розробки методів аналізу їх емоційного забарвлення;
- розробити метод автоматичного аналізу емоційно забарвлених цитат.

ABSTRACT

Master’s thesis: 70 pages, 16 figures, 7 tables, 1 appendices, 17 sources.

NATURAL LANGUAGE PROCESSING, TONE ANALYSIS OF TEXT, NAIVE BAYES CLASSIFIER, SUPPORT VECTOR METHOD.

The major goal of this thesis is to research and develop a method of analyzing the tonality of quotations in scientific publications.

To achieve the goal, the following tasks must be solved:

- conduct an overview of existing methods of automatic analysis of the emotional coloring of texts;
- conduct a study of the textual features of quotations in scientific publications in the context of developing methods for analyzing their emotional coloring;
- to develop a method of automatic analysis of emotionally colored quotes.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ 7
ВСТУП 8
1 АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ ТА ПОСТАНОВКА ЗАДАЧІ ДОСЛІДЖЕННЯ 10
1.1 Обґрунтування дослідження 10
1.2 Постановка проблеми наукового дослідження 13
2 ОГЛЯД СУЧАСНИХ ПІДХОДІВ ДО КЛАСИФІКАЦІЇ НАСТРОЮ ЦИТУВАНЬ 17
2.1 Наївний метод Байеса 18
2.2 Дерева рішень 19
2.3 Random Forest 21
2.4 Метод опорних векторів (SVM) 22
2.5 Модель мішок слів 24
2.6 Класифікація настрою цитування 25
3 МЕТОДОЛОГІЯ ДОСЛІДЖЕННЯ 30
3.1 Вилучення N-грамів 30
3.3 Скіп-грам 33
3.4 Класифікація настрою цитування на рівні документа 34
3.5 Виявлення аспекту 38
4 РЕЗУЛЬТАТИ ЕКСПЕРЕМЕНТАЛЬНИХ ДОСЛІДЖЕНЬ 43
4.1 Опис експериментального набору даних 43
4.2 Аналіз настрою цитування 46
ВИСНОВКИ 58
ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ 59
ДОДАТОК А Графічний матеріал кваліфікаційної роботи 61

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

NLP (Natural Language Processing) - обробка природної мови

SA (Sentiment Analysis) - аналіз тональності

CCA (Contextual Citation Analysis) - контекстуальний аналіз цитування

NER (Named Entity Recognition) -розпізнавання іменованих сутностей

ML (Machine Learning) - машинне навчання

SVM (Support Vector Machines) - метод опорних векторів

LSTM (Long Short-Term Memory) - довготривала короткочасна пам'ять

CNN (Convolutional Neural Network) - згорткова нейромережа

Word Embedding-векторне представлення слів

POS (Part of Speech) - частина мови

TF-IDF (Term Frequency-Inverse Document Frequency) - частота термів-зворотна частота документа

BERT (Bidirectional Encoder Representations from Transformers) - двонапрямлені вкодування на основі трансформаторів

ВСТУП

З кожним роком кількість наукових публікацій зростає експоненціально. Вивчення актуального стану публікацій в науковій літературі є основним напрямком. За рахунок наявності великої кількості наукових досліджень дослідникам і студентам важко бути в курсі думок, висловлених у науковій літературі. Тому важливим елементом цього наукового аналізу великих даних є тональність цитувань в науковій літературі. Велику роботу було проведено для аналізу настроїв оглядів продуктів, оглядів фільмів, текстів розповідей, публікацій у блогах, форумів, відгуків, рекомендацій, рейтингів, коментарів тощо, [1] тоді як менше уваги було приділено отримання інформації про аналіз думок із цитат [2]. Поточні репозиторії досліджень не надають жодних можливостей для узагальнення цитати на основі висловлених у них настроїв (аналізу думок) [3]. Наприклад, перш ніж цитувати статтю, дослідник може захотіти дізнатися, що інші дослідники та наукове співтовариство говорять про цей конкретний науковий внесок. Таким чином, він підтверджує потребу в системі аналізу настроїв, яка може аналізувати основну масу таких цитат і надавати значущу інформацію дослідникам і студентам. Визначення того, чому люди цитують певну статтю, викликало інтерес до дослідження в різноманітних областях інформаційних наук, аналізу дискурсу та соціальних наук.

Існуючі методи аналізу цитування, бібліометрія, h-індекс, g-індекс, a-індекс, імпакт-фактор тощо, схильні до певних внутрішніх обмежень, наприклад, вони ігнорують намір цитувати конкретну наукову статтю, і більшість із них є кількісними [4]. Через ці причини вони не можуть отримати доступ до справжнього впливу, позиції та думки автора щодо змісту попередньої роботи. Таким чином, існує кінцева потреба в аналізі настрою цитування/контенту на основі методів аналізу тексту та лексичного

аналізу. Текст цитування містить точний і стислий аналіз статті через обмеженість обсягу статей і високу якість статті з точки зору її коректності [5]. Існуючі підходи до виявлення настрою цитат покладаються лише на текст цитати. Але існують певні складності, пов’язані з виявленням настроїв у тексті цитати, наприклад:

- приховані настрої;
- неупередженість тексту цитати;
- контрастне вираження негативних настроїв;
- варіація лексичних термінів, які використовуються для вираження почуття в науковому тексті;
- використання технічних термінів;
- обсяг цитованого тексту;
- більш вибірковий словниковий запас.

Ще одна відмінність між загальним аналізом настрою та визначенням настрою цитування полягає в унікальних характеристиках цитат [6]. Усі ці причини роблять аналіз настрою цитування однією з найскладніших і складних сфер застосування аналітики настроїв [1, 2].

Існують різні аспекти дослідження аналізу цитування. Аналіз настрою цитування є новою та цікавою областю, яка спонукає досліджувати думку автора цитування щодо цитованого тексту. Спочатку він виділяє думку цитованої статті щодо цільової статті, тобто цитованої статті, і класифікує її полярність настроїв як позитивну, негативну або нейтральну [7, 8]. Крім простої класифікації настроїв цитування, існують численні дослідження щодо класифікації функцій цитування як «впливовий/невпливовий», «функціональний/поверхневий» і «контраст/конфлікт» [2, 5]. Одним з вимірів є узагальнення наукового внеску на основі цитат. Подібним чином, інший вимір полягає в розробці якісних показників для оцінки наукового внеску порівняно з кількісними показниками. Це призведе до розробки нових застосувань у бібліометрії та бібліографії дослідження.

1 АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ ТА ПОСТАНОВКА ЗАДАЧІ ДОСЛІДЖЕННЯ

1.1 Обґрунтування дослідження

Процес визначення актуальності та частоти цитувань з різних літературних матеріалів, включаючи: журнали, дослідницькі статті, книги, відомий як аналіз цитувань. Як правило, важливість або вплив дослідницької роботи зазвичай розраховується шляхом підрахунку кількості цитувань, які вона отримує за період. Але це в основному кількісна оцінка літературного твору може бути необ’єктивною через самоцитування або випадкове цитування. Для належної оцінки статті цитований текст має величезне значення, оскільки дослідники беруть уривки із загальних ідей і концепцій із різних робіт і згадують результати та висновки статті під час обговорення.

Але це не означає, що референтний документ завжди буде описувати документ позитивними словами. Іноді, якщо референтна стаття досягла певних результатів, порівнянних із реферованою статтями, автори не дадуть позитивної реакції на попередню. Тому для вирішення цієї проблеми необхідний аналіз настроїв тексту цитат для оцінки цитат на основі рівня їх полярності.

В існуючій літературі різні методи машинного навчання застосовувалися для визначення настрою цитування, наприклад, Machine Support Vector [9], наївний байєсівський класифікатор [5], класифікатор максимальної ентропії [10], логістична регресія [11]. Різні функції, такі як n-грами [1,5], відношення залежності, поділ речення, ознаки заперечення, фізичні особливості (місце розташування, популярність, щільність) [2,19–11], у літературі широко використовуються особливості контекстуальної полярності тощо. Існують гібридні методи кращі, ніж єдиний алгоритм. Всі перераховані вище технології мають свої переваги і недоліки.

Таким чином, одним з аспектів запропонованої дослідницької роботи є розробка гібридної метода за допомогою машинного навчання та методів інтелектуального аналізу даних для класифікації настрою цитування. Далі встановлюється в літературі, що n-грами вищого порядку, тобто довші фрази, як правило, менш двозначні з точки зору їхньої полярності, оскільки вони вловлюють заперечення на короткій відстані та позиційний контекст. N-грами вищого порядку як ознаки можуть досягти порівнянної або покращеної точності класифікації, ніж сучасні на великомасштабних наборах даних.

Далі цей метод можна пояснити на наступному прикладі. Розглянемо приклад цитати «The proposed approach yields better accuracy». Його уніграма: «The», «proposed», «approach», «yields», «better», «accuracy», у якій розглядається одне слово. Його біграма: «The proposed», «proposed approach», «approach yields», «yields better», «better accuracy» , у якій розглядаються пари слів. Його триграма: «The proposed approach», « proposed approach yields», «approach yields better», «yields better accurac», де розглядається послідовність слів, кількість яких дорівнює 3. n-грами вищого порядку відносяться до 4-грамів, 5-грамів і так далі. Таким чином, аналізуючи результати дослідження кількох авторів , буде розширено класифікацію настрою цитування за допомогою уніграми, біграми, триграми та пентаграми прикметника та прислівника та їх комбінацій. Враховуючи дослідницьку статтю з довгим списком цитувань, модель може ідентифікувати найвпливовіші аспекти та створити підсумок настроїв наукової статті на основі аспектів. Розглянемо наступний приклад цитатного речення: «The technique is efficient but the dataset is not very comprehensive». Тут цитуючий автор висловлює суперечливі почуття щодо двох аспектів дослідження – техніка аспекту означає позитивну думку, тоді як набір даних другого аспекту називається негативним. Цей приклад показує важливість тонких аспектів, у тому, що вони зображують уподобання авторів під час цитування дослідницької статті, що забезпечує зв’язок між двома статтями.

Метою є виділення всіх можливих і релевантних аспектів із цитованих

речень, а потім групування синонімів. Наприклад, «Technique is efficient» та «Approach yields the better result», виділіть слова «“technique» та «“approach», які представляють аспект цитованого дослідження, а потім згрупувати їх в одну категорію, оскільки обидва вони вказують на одне й те саме.

Наступним завданням є визначення відповідного настрою для кожного ідентифікованого та виділеного аспекту. Це полегшить вимірювання настрою на основі аспектів або інтенсивності цитувань (рівень позитивних і негативних цитат і аспектів). Результат цього має на меті не лише скоротити основну ідею цитованої статті, а також оцінити конкретні аспекти дослідження (наприклад, методику, набір даних, результати, алгоритм тощо). Мотивація полягає у виявленні фальшивих думок щодо цитованого дослідження. Приклад, наведений на рисунку 1.1, описує аспектний підсумок настрою дослідницької роботи. Це допоможе краще зрозуміти дослідницьку статтю та її цитовані аспекти.

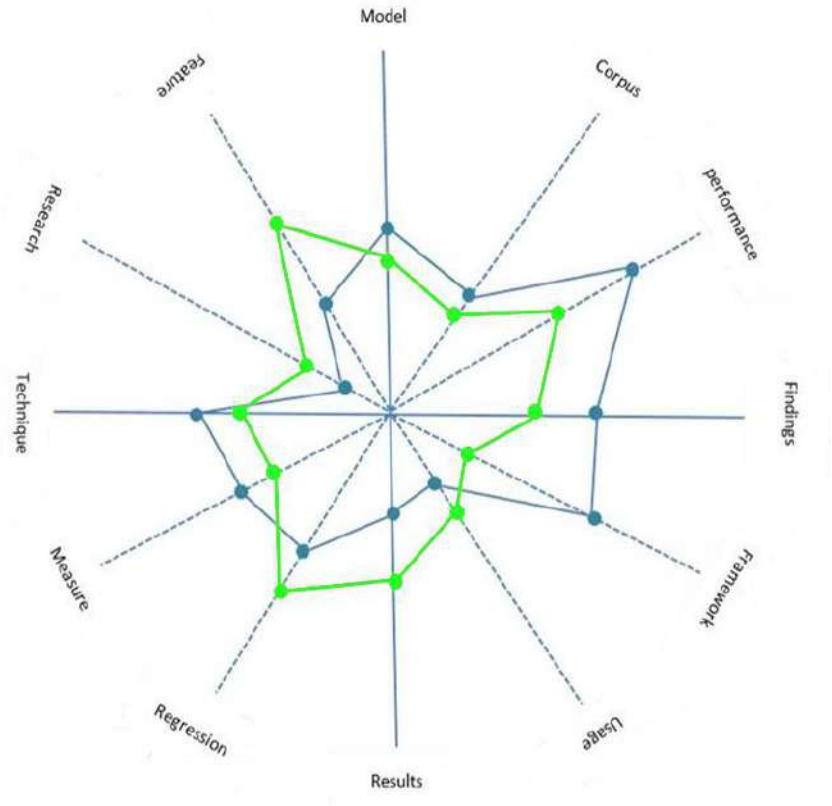

Рисунок 1.1 - Профіль настрою цитування на основі аспектів

Є лише кілька досліджень, які зосереджуються на силі настроїв

шляхом ідентифікації прикметників і ранжування цитат на основі вилученої полярності цитат. Іншою причиною дослідження цієї області є розробка деякого автоматизованого підходу до ідентифікації настрою автора цитування щодо цитованої літератури, виражена в цитаті з використанням різних методів машинного навчання. Крім того, можна визначити полярність цитати, коли вона використовується в певному контексті.

1.2 Постановка проблеми наукового дослідження

У цій дослідницькій роботі я розглянув три важливих проблем у сфері аналізу настрою цитування. Це допомагає краще зрозуміти проблему. Цей критичний аналіз стає шляхом до глибокого занурення в дослідницьку область, яка є критичною для аналізу та розуміння.

У сфері обробки природної мови та ймовірності n-грами широко використовуються в функціях для аналізу думок. N-грами - це безперервна послідовність n-елементів, термінів або об’єктів із даного документа, послідовності тексту чи мовлення, яка використовується як функція для отримання підказок настрою з тексту [12, 13]. N-грами можуть мати фіксовану або змінну послідовність на рівні символів або маркерів. Існує кілька переваг використання n-грам:

- використання n-грам як базового терміна забезпечує відмовостійкість у орфографічних помилках і не вимагає попередніх знань;
- за допомогою n-грам система може досягти мовної незалежності;
- словники, граматики та правила стають непотрібними;
- системи, що розглядають n-грами, стають вільними від стоп-слів;
- немає необхідності відокремлювати символи від слів.

Основне обмеження, пов’язане з використанням n-грам для класифікації настроїв, полягає в тому, що семантична інформація частково втрачається, і вона не може ефективно обробляти заперечення [14]. N-грами

розмірів два і три найчастіше використовуються в сучасній літературі [2, 12]. N-грами розміру 1 або 2 забезпечують кращі результати в поєднанні з деякими іншими функціями, але в деяких випадках вони не дають результатів. Наприклад, якщо ми проаналізуємо наступний текст цитати за допомогою уніграмного підходу. «The model is not applicable» Це призведе до нейтральної полярності тексту цитати через присутність одного слова позитивної полярності, тобто «applicable», і одного слова негативної полярності, тобто «not». Подібним чином, коли аналізуємо наведений вище текст за допомогою біграмного підходу, це класифікує його як негативний через наявність слова «not applicable», що є правильною класифікацією. Розглянувши приклад «This approach is not very impactful but produce excellent word alignments», висвітлюється заперечення, посилення та контраст щодо внеску дослідження. Так, цей сценарій показує, що коли розглядаємо використання n-грамів вищого порядку, очікується, що результати будуть кращими. Нюанс використання n-грам вищого порядку полягає в тому, що вони не лише враховують позначене слово, але й фіксують короткочасну позиційну інформацію, яка допомагає досягти кращіх результатів. По-перше використання високого порядку n-грами можуть допомогти виявити настрій цитування. По-друге, в літературі зазначається, що технічні терміни відіграють домінуючу роль у тексті наукового цитування [1]. Іноді з цими термінами пов’язують почуття. Наприклад, якщо розглянемо текст цитати, «the state of the art sentiment classification systems use n-gram features ...». У такій ситуації розгляд лише n-грам нижчого порядку не підійде для класифікації цитованого речення, оскільки коротші фрази мають тенденцію бути неоднозначними з точки зору їх полярності. Тому в роботі запропоновано класифікувати цитовані речення за допомогою n-грам вищого порядку та вивчити їхній вплив на виявлення настрою цитованих. N-грами вищого порядку є більш точними та детермінованими порівняно з n-грамами нижчого порядку. Використання довгих фрагментів може допомогти отримати інформацію про полярність, оскільки існує прямий зв’язок між

довжиною фрагмента та точністю класифікації настрою. Для експерименту було встановлено максимальну довжину фрагмента на 5. В експериментах із ознаками біграми, триграми та пентаграми та їх комбінаціями я також включалися n-грами нижчого порядку (уніграми та біграми). Причина для цього полягає в тому, щоб продемонструвати точність передбачення використання n-грам вищих порядків у широкомасштабних задачах аналізу настрою. Ця задача передбачає ідентифікацію настроїв автора, що цитує, щодо цитованої статті шляхом вилучення моделей прикметників і прислівників уніграми, біграми, триграми та пентаграми з тексту цитати. Після того, як POS тегував текст цитати, було використано синтаксичний аналізатор речень для виділення лінгвістичних особливостей, що складаються з прикметників, прислівників і n-грам з тексту цитати. У цьому процесі нерелевантна інформація, наприклад спеціальні символи, теги HTML, орфографічні помилки тощо, відфільтровується. На цьому кроці ця стороння інформація видаляється. Наступним кроком є видалення стоп-слів із цитованих речень. Наступною функцією, яка виконується, є виділення корня, у якому слова скорочуються до їх термінів/кореневих форм. Було розділено текст цитати на лексеми високого порядку n-грамів, тобто n=5 зліва направо. Потім знаходиться оцінка настрою цих термінів на основі словника SentiWordNet 3.0, який пізніше класифікується як позитивний, негативний і нейтральний на основі оцінки настрою. Це корисно для визначення правильної орієнтації настрою. Сила настрою також необхідна для аналізу цитування.

Огляд літератури показує, що стиль написання наукової статті є об’єктивним, що призводить до класифікації більшості цитат як об’єктивних або нейтральних [2, 15]. Більшість авторів стверджують, що пишуть об’єктивні чи загальні рецензії на цитовану статтю. Вони зазвичай приховують негативні настрої як двозначну похвалу, що робить виявлення настроїв ще більш заплутаним [1, 2]. У такому сценарії розгляду лише цитатного речення буде недостатньо. Щоб вирішити цю проблему, береться

до уваги контекст цитування як розширений простір для виявлення настрою цитування. Контекст цитування відноситься до теми, тексту або речень, які оточують текст цитати або посилання в наукових статтях. Текст, який оточує цитоване речення, вважається його локальним контекстом. Раніше використовувався контекст цитування з різним розміром вікна від кількох слів до кількох речень, зосереджених навколо цитованих речень.

Існуючі підходи до виділення контексту цитування можна розділити на дві групи. Підхід із симетричним вікном, який розглядає вікно слів, тобто n токенів перед і після цитованого речення. Другий підхід - це підхід, заснований на реченнях, коли контекст розглядається на основі n речень перед і після цитатного речення. Відповідно до наявної літератури, контекст цитування, що складається з вікна з 3 речень, вважається достатнім для класифікаційного рішення. Більшість існуючих робіт з виявлення настрою цитування розглядають лише цитатне речення та просто ігнорують його локальний контекст, що може підірвати його точність класифікації. Врахування контексту цитування допоможе покращити модель BOW (bag-of-words) і розкрити приховані настрої щодо цитованої роботи. Наприклад, розглянемо цитоване речення «[] has recently proposed a simpler SVM-based algorithm for analogical classification called Pair-Class» та ігнорування його подальшої анафори не вказуватиме на чітке ставлення до цитованої роботи. Текст перед цитатою «“it does not adopt a set-based or distributional model of relational similarity» має негативну орієнтацію. Це свідчить, що настрої пов’язані з цитатою не є статичним, а це залежить від конкретного контексту, в якому воно цитувалося. Підводячи підсумок, можу сказати, що настрій цитатного тексту можна передати як через його контекст, так і через концептуальну семантику.

Отже досліджуючи контекст настроїв і його розташування, ми можемо не тільки зрозуміти намір, чому автори цитування цитують таким чином, але й з’ясувати закономірність розподілу цих настроїв у IMRaD.

2 ОГЛЯД СУЧАСНИХ ПІДХОДІВ ДО КЛАСИФІКАЦІЇ НАСТРОЮ ЦИТУВАНЬ

На даний момент існує два основних підходи до проблеми аналізу тональності текстів: підхід, заснований на методах машинного навчання і підхід, заснований на використанні словників тональної лексики. [9]. В основі підходу, заснованому на використанні словників тональності, лежить аналіз тональності окремих слів (термів) в тексті та подальше визначення тональності всього тексту згідно з оцінками окремих слів, що входять в цей текст. Для цього в основному використовуються словники тональності, в яких кожному слову ставиться у відповідність величина, яка відображає «вага» слова в тональності всього тексту. Надалі, згідно із запропонованим методом будується функція, яка на вхід приймає кількість входжень в текст кожного слова і обчислює агрегированную величину тональності всього тексту [2, 3, 4, 5]:

$$W_i = f\{n_i(w_i),...,n_i(w_\alpha)\}, \quad (2.1)$$

le n_i (w_j) - кількість входжень слова j в текст i. α - кількість слів у словнику тональності. W_i - тональність тексту i

У підході, заснованому на машинному навчанні завдання аналізу тональності зводиться до задачі класифікації текстів, яка може бути вирішена шляхом навчання класифікатора на заздалегідь розміченій колекції текстів [6, 7, 8].

Кожен підхід має свої переваги й недоліки. Так, наприклад, методи, засновані на використанні словників тональності не потребує навчальної колекції, тобто немає необхідності в ручному розмітці текстів. Також дані методи не потребують складанні навчальної функції. Таким чином, «рішення», прийняті класифікатором можуть бути легко пояснені. У той же

час для даних методів потрібні заздалегідь розмічені словники тональності, які крім усього іншого повинні враховувати предметну область досліджуваного тексту.

У методах машинного навчання не потрібні словники тональності, і на практиці класифікатори демонструють високу якість класифікації. Більш того, якість класифікації можна поліпшити за рахунок вибору параметрів (ознак) класифікації і правильно підібрані комбінації текстових документів у навчальній вибірці. У той же час класифікатор, навчений на текстах однієї предметної області, може не справлятися зі своїм завданням для текстів з іншої предметної області. [9].

Нижче наведені методи, використовувані для вирішення завдання аналізу тональності.

2.1 Наївний метод Байеса

Нехай P (c | d) – Імовірність того, що документ d належить класу c. У наївною байєсівської класифікації завдання полягає в знаходженні найбільш ймовірного класу c_m для даного документа d

$$c_m = \arg\max_{c \in C} P(c \mid d) \tag{2.2}$$

використовуючи формулу Байеса, можна переписати вираз для P (c | d)

$$c_m = \arg\max_{c \in C} \frac{P(d \mid c)P(c)}{P(d)} \tag{2.3}$$

де знаменник P (d) не залежить від c і, отже, не впливає на знаходження максимуму, тому його можна опустити; P (c) – ймовірність того, що зустрінеться клас c, незалежно від розглянутого документа. P (d | c) –

ймовірність зустріти документ d серед документів класу c.

Документ можна представити у вигляді вектора вхідних в нього умов (слів). Тоді величина P (d | c) обчислюється таким чином:

$$P(d \mid c) = P(t_1, t_2, ...t_{n_p} \mid c) = \prod_{k=1}^{n_d} P(t_k \mid c) \quad (2.4)$$

Отже, формула для обчислення найбільш ймовірного класу приймає наступний вигляд:

$$c_m = \arg\max_{c \in C} \hat{P}(d \mid c)\hat{P}(c) = \arg\max_{c \in C} \hat{P}(c) \prod_{k=1}^{n_d} P(t_k \mid c) \quad (2.5)$$

Метод наївної байєсівської класифікації будується на двох припущеннях:

- про умовної незалежності термів і;
- про незалежність позицій термів.

Іншими словами, не береться до уваги той факт, що в тексті на природній мові поява двох різних слів часто взаємопов'язане (наприклад, найімовірніше, що слово «удар» зустрінеться в одному тексті зі словом «бокс», ніж зі словом «балет»), і, по-друге нехтування тим фактом, що ймовірність зустріти одне і те ж слово на різних позиціях (місцях) у тексті різна. Саме тому дана модель природної мови називається наївною. Незважаючи на це, модель показує досить точні результати класифікації.

2.2 Дерева рішень

Ідея даного методу полягає в побудові дерева рішень на «навчальному» наборі документів. Дерево будується за наступним правилом: вибираємо

терм, документи, що містять цей терм кладемо направо, інші наліво. Таким чином, документи розділилися на дві непересічні колекції. Для кожної колекції вибирається новий терм і повторюється описана вище процедура. Так триває до тих пір, поки не вийде однорідна колекція, тобто колекція, в якій або всі документи відповідають категорії, або всі документи відповідають її доповнення.

Нехай є Q – навчальна вибірка документів, а $C = \{c_1 \dots c_{|C|}\}$ – множина класів. Нехай множина класів складається з $|\ C\ | = m$ елементів. Для кожного прикладу з Q відома його приналежність до якої-небудь з класів $\{\ c_1 \dots c|\ C\ |\}$, тобто відомо значення цільової функції для всіх документів в навчальній вибірці. $\forall d \in Q$ Ф: D × C → {0,1}.

У листі дозволяє дерева розміщуються значення цільової функції, в інших вузлах - умови переходу, що визначають напрямок руху вздовж ребер дерева. Для класифікації кожного прикладу алгоритму необхідно пройти всі дерево від кореня до одного з листів. Тим самим отримати значення цільової функції. Алгоритм C4.5 є одним з найпопулярніших алгоритмів побудови дерев рішень.

Є корінь і пов'язане з ним безліч Q, яке необхідно розділити на підмножини. Для цього використовується один з атрибутівwjв якості перевірки. Обраний атрибутwjприймає k значень, що дозволяє розділити безліч на k підмножин. Після чого створюються k нащадків кореня. Кожному нащадку відповідає деяка підмножина, отримане після розбиття множини Q.

Функція вибору атрибута w_j і подальшого розбиття по ньому рекурсивно застосовується до всіх k нащадкам і завершує свою роботу в двох випадках:

- після чергового поділу на підмножини в вершині і виявляються документи з одного класу (тоді вершина відзначається як лист. Клас, якому належать документи в цій вершині, є позначенням листа);

- вершина після виконання чергового кроку виявилася асоційованої з порожнім безліччю (в такому випадку вона призначається листом, а в якості

рішення вибирається найбільш частий клас у безпосереднього предка).

Приклад дерева рішень представлено на рис.2.1.

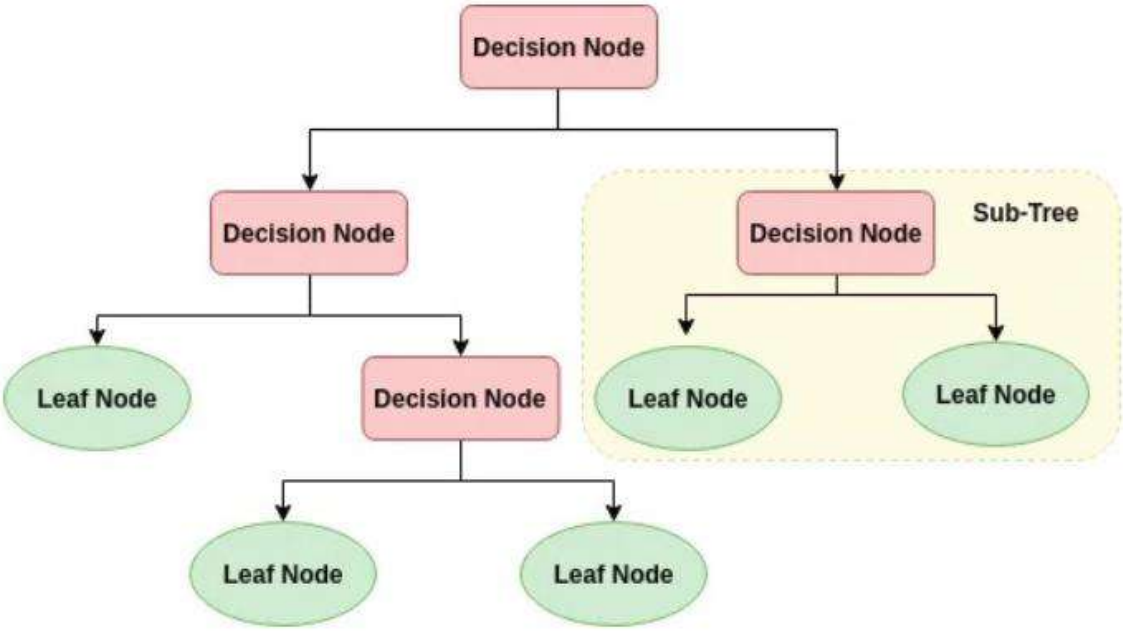

Рисунок 2.1 – Приклад Дерева рішень

2.3 Random Forest

Алгоритм Random Forest – ансамблевий метод машинного навчання, який використовує ансамбль дерев рішень. Він ґрунтується на основних підходах беггінга і вибору випадкових підмножин ознак. Цей алгоритм дозволяє досягти високої точності класифікації. Дерева в ансамблі будуються один від одного незалежно. [10]

Фінальна класифікація документів проводиться за допомогою «Голосування», тобто підсумковим класом об'єкта оголошується той клас, за який проголосувала найбільша кількість дерев.

Метод демонструє високу якість класифікації, порівнянне з SVM або бустінгом, Але в той же час володіє високою складністю по пам'яті для

зберігання дерев - $O\ (N * K)$. [22]

Приклад Random Forest представлено на рис. 2.2.

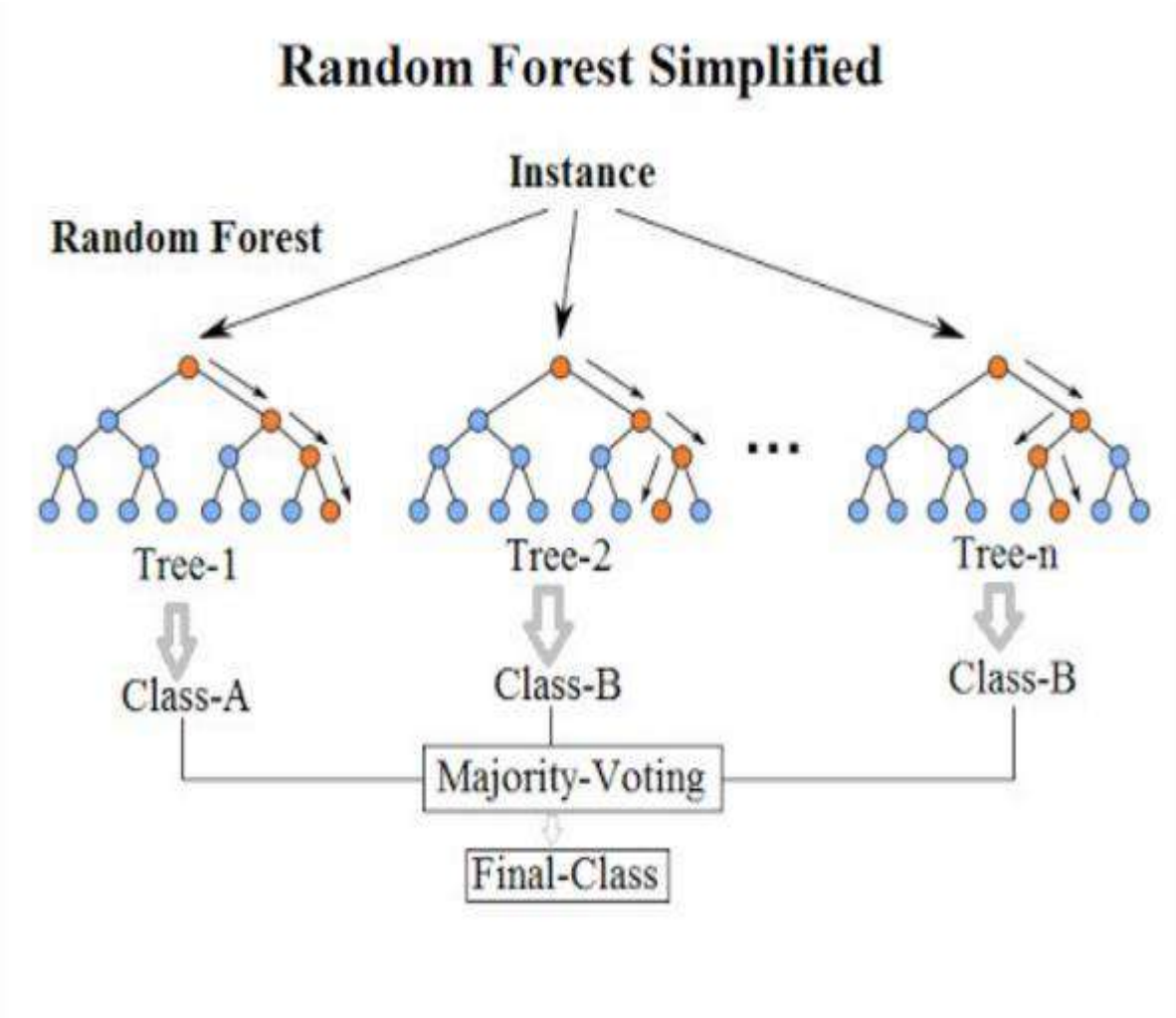

Рисунок 2.2 – Приклад Random Forest

2.4 Метод опорних векторів (SVM)

Метод опорних векторів (Support Vector Machine, SVM) – метод, в якому основою є побудов (Оптимальна) розділяючої гиперплощини. Деяка вибірка лінійно роздільна, якщо в ній можливо отримати (побудувати) лінійний пороговий класифікатор:

$$sign(\sum_{i=1}^{m} w_i * x^i - w_0) = sign\,(< w, x > - w_0) \quad (2.6)$$

де $x = (x_1, ..., x_n)$ – признаковая опис об'єкта x; вектор $w = (w_1, ..., w_n) \in \mathbb{R}n$ і скалярний поріг $w_0 \in \mathbb{R}$ є параметрами алгоритму. Таким чином завдання полягає в тому, щоб підібрати значення вектора w такі, при яких функціонал, що визначає число помилок дорівнює нулю:

$$\sum_{i=1}^{n} [y_i(< w, x_i > - w_0) \le 0] = 0 \quad (2.7)$$

де $<w, x> = w_0$ – розділяє гіперплощина. Більш докладний опис алгоритму можна знайти в [7].

Приклад класифікації наведено на рисунках 2.3 та 2.4.

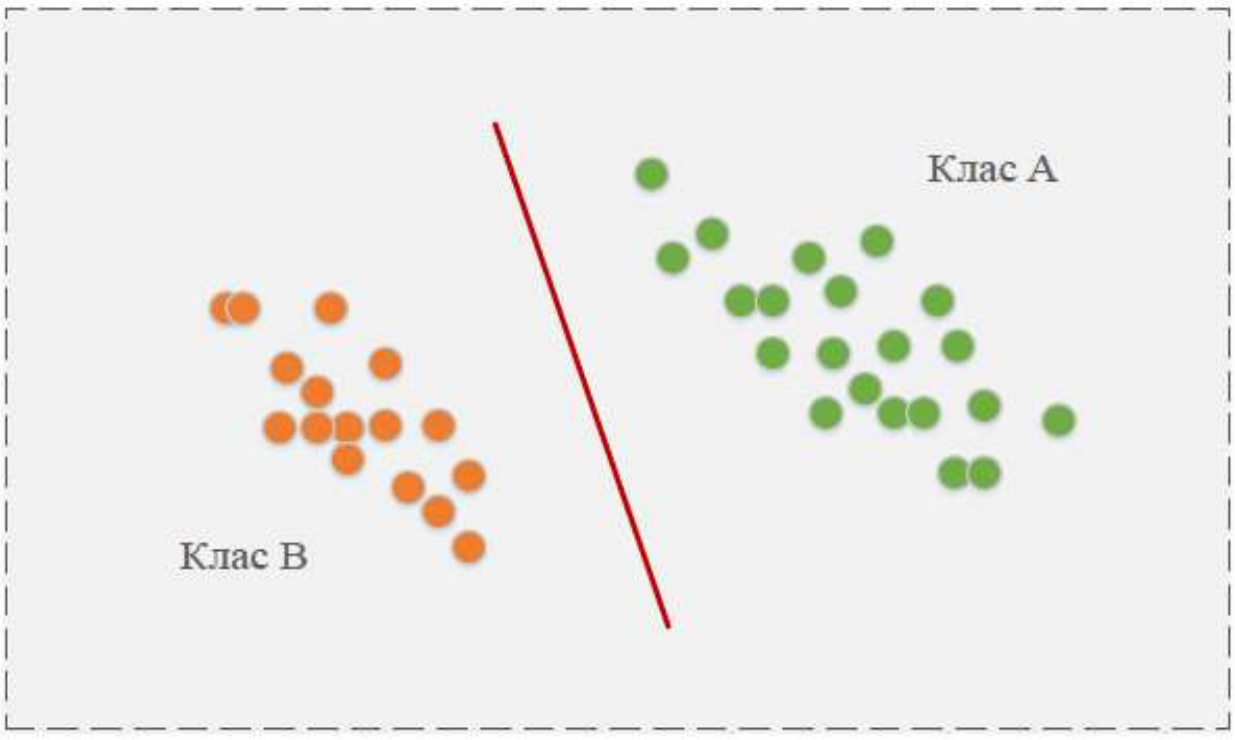

Рисунок 2.3 – Класифікація на два класи

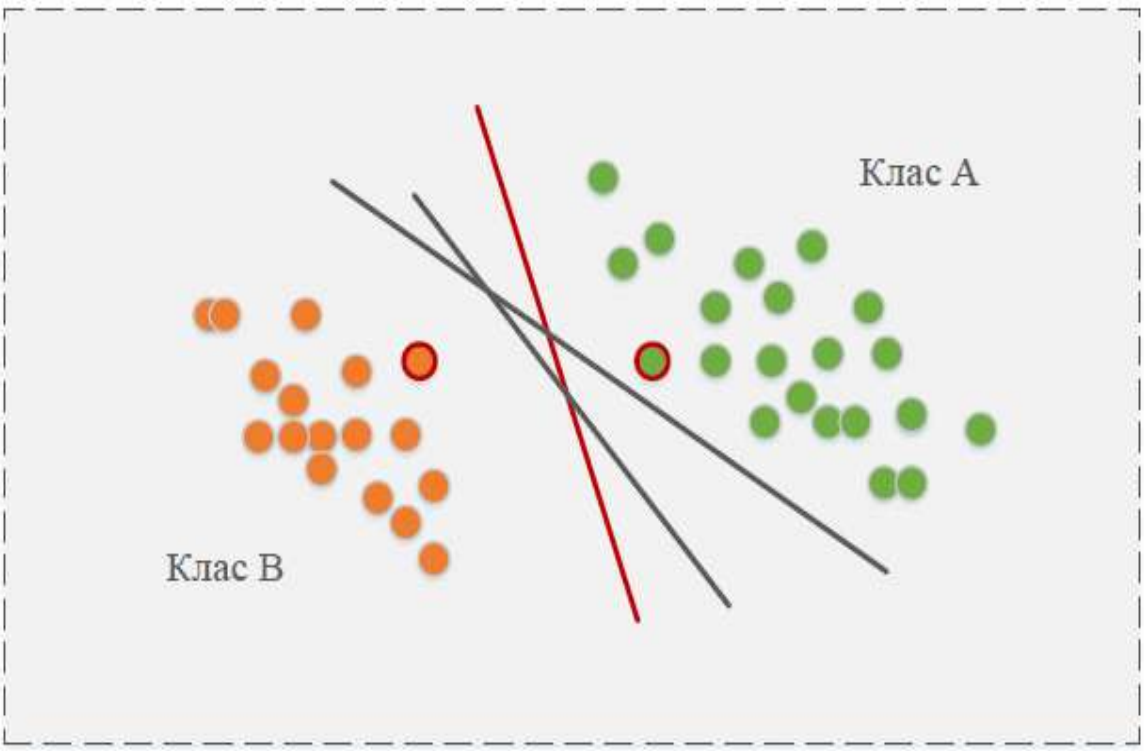

Рисунок 2.4 – Різні роздільні прямі

2.5 Модель мішок слів

Для реалізації методів машинного навчання існує класична модель "Мішок слів" (Bag of Words). Формальна постановка задачі виглядає наступним чином:

Нехай $f1$, ..., fm – множина, що складається з m ознак (атрибутів), які можуть бути присутніми в документі; нехай $ni\ (d)$ – це кількість входжень ознаки fi в документ d. Далі кожен документ d представляється у вигляді вектора наступним чином:

$$\vec{d} = (n_1(d), n_2(d), ..., n_m(d)) \tag{2.8}$$

Виділяють два основних типи атрибутів:

- частотні – кожне значення в векторі d відповідають кількості входжень ознак в документ d; тоді $ni\ (d) \in (0; +\infty)$

- бінарні (наявності / відсутності), кожне значення в векторі d бінарне (true / false або 0/1) і відображає факт присутності ознаки f_i в документі d. Тоді n_i (d) = {0,1}

Далі документи, представлені у вигляді векторів своїх ознак (атрибутів), використовуються для навчання класифікатора, реалізованого з допомогою одного з методів машинного навчання.

2.6 Класифікація настрою цитування

Існуючі дослідження розглядають різні методи підходу до аналізу настрою цитування. Дослідники наголошують на дослідженні змісту цитування та контексту цитування для кількісної оцінки наукового внеску. Для характеристики внеску цитованої статті в цитування вони мають використовували різні методи аналізу тексту та обробки природної мови. В даному розділі підсумовано відповідну літературу для класифікації настроїв цитат наукових праць. В роботі [1] представлено підхід автоматичної ідентифікації полярності настроїв у тексті цитати за допомогою різних мовних особливостей рівня слів. Використовуючи такі функції, як n-грами (n=3), відносини залежності, ознаки заперечення, науковий лексикон і поділ речень у структурі SVM.

SVM розділив цитати на три різні класи, тобто позитивні, негативні та нейтральні. Результати також підтвердили, що триграми та відносини залежності пропонують надійні результати в цьому відношенні та перевершують науковий лексикон і функції поділу речень. [5] спрямований на розробку автоматизованого методу виявлення настрою цитування за допомогою методів машинного навчання та лінгвістичних підказок. У своїй попередній роботі автори представили техніку класифікації тексту цитат на основі опорних векторних машин із використанням n-грамової (уніграми та біграми) статистики слів як вектор ознак. Запропонована техніка класифікації настрою цитування

класифікує текст на дві категорії, тобто позитивні та інші. У майбутній роботі вони планують удосконалити свій метод, використовуючи більш щільний набір наземних даних і збагачуючи набір функцій, враховуючи більше вхідних функцій. [7] розглядав техніку генерації структурованого резюме наукової статті на основі тексту цитат у цитованих статтях. Вони класифікували цитатний текст за допомогою класифікації з кількома мітками в один або більше з п’яти різних класів, тобто резюме, сильні сторони, обмеження, пов’язана робота та застосування. В якості базової лінії вони використали наївний алгоритм Байєса, розглядаючи комбінації прикметників, дієслів і n-грамів у кожному класі. За результатами експерименту, враховуючи прикметники, дієслова та біграми в комбінації досягли середньої точності 68,54%. В роботі [14] автор представив техніку виявлення упередженості цитування за допомогою ручного аналізу настрою цитування публікацій біомедичних досліджень. Автори окреслили деякі відмінності між підходами, що використовуються біомедичними дослідниками, та методами автоматизованої класифікації настрою цитування. Згідно з їхніми висновками, дослідники приділяли велику увагу таким аспектам настрою цитування, як сила та обґрунтованість, і менше приділяли увагу простому настрою, заснованому на полярності.

В роботі [7] автори описали, що багато з існуючих методів класифікації настрою цитування відповідають до предметних галузей, таких як медицина, біомедицина, інформатика, французькі гуманітарні статті тощо. У своїй роботі вони класифікували цитати на основі наївного байєсівського класифікатора, вибравши вікно з п’яти речень навколо цитованого місця та досягли точності на 80%. Для експериментальних цілей вони використовували узагальнену лексику, враховуючи цитати з багатьох областей, і це показує демонстрацію їх підходу в багатьох дисциплінах. В роботі [3] автори зосередились на поєднанні методів аналізу тексту та лексичного аналізу для з’ясування ставлення автора до

цитованого твору. Вони інкапсулювали отримані думки в об’єктивну міру для визначення впливу наукових публікацій. Одним із ключових висновків їхньої дослідницької роботи є те, що більшість цитат були нейтральними за своєю природою та виявили значну згоду у використанні термінології для вербалізації настроїв щодо цитованої роботи як з точки зору позитивних, так і негативних думок. У майбутньому вони хочуть розширити модель, враховуючи різні лексичні елементи та їх різні комбінації, такі як сполучники, прикметники, дієслова, прислівники та комбінації дієслів і прислівників.

У роботі [16] автори зосередилися на текстових (наказові фрази, кількість наказових фраз тощо), фізичних (розташування, популярність, щільність тощо) та синтаксичних особливостях у встановленні зв’язку між функцією цитування та її полярністю. Вони класифікували речення цитування за чотирма різними типами: фон, фундаментальна ідея, технічна основа та порівняння за допомогою різних класифікаторів навчання під наглядом, включаючи BayesNet, NaiveBayes, SMO, J48 і RandomForest. Вони запропонували модель класифікації, засновану на самонавчанні в стилі ансамблю, і провели свої експерименти з цитатами, витягнутими з робіт з комп’ютерної лінгвістики. Аналіз результатів показує ефективність використання запропонованого набору ознак у поєднанні з синтаксичними ознаками, отриманими з тегів частини мови (POS).

В роботі [10] автор використовував класичну фасетну схему анотації Moravcsik і Murugesan (1975) і застосував стандартний класифікатор ME (максимальна ентропія). Вони класифікували речення з цитатами за чотирма різними аспектами, щоб вирішити, чи є цитована робота:

- ідеєю чи інструментом;
- точною чи помилковою (підтверджує чи заперечує);
- фундаментальний або поверхневий;

- заснований на або альтернативній роботі.

Контекстна довжина цитатних речень складала від 1 до 3 речень. Ряд особливостей подібно до уніграм, місце речення, мовні особливості, компаративи та лексика, що складається з позитивних і негативних речень, використовуються комбінаторно. З чотирьох міток анотації одна використовується для полярності настроїв, позначаючи позитивні випадки як підтверджувальні, а негативні речення як заперечні. Проте нейтрального класу немає. Одним із ключових висновків дослідження є те, що більшість цитат мали нейтральний характер із значною згодою у використанні термінології для вербалізації настроїв щодо цитованої роботи як з точки зору позитивних, так і негативних думок. Наступним напрямком роботи може бути розширення моделі шляхом розгляду різних лексичних елементів та їх різних комбінацій, таких як сполучники, прикметники, дієслова, прислівники та комбінації дієслів і прислівників.

Відповідно до висновків, у сфері аналізу настрою цитування ще є багато місця для подальших досліджень і розробок. Крім того, моделювання заперечень у цитованих реченнях для покращення продуктивності може бути можливим напрямком, який допоможе досліджувати довші контексти цитування. Це також може допомогти усунути залежність від ручного анотування та належного позначення та розпізнавання категорій цитат. Що стосується аналізу настрою цитування, то існують певні труднощі, зокрема там, де набори даних є нестандартизованими. Це може бути результатом наявності кількох форматів цитування та стилів посилань. Розмір контексту цитування, в якому існує прив’язка цитування, також є важливим і складним завданням для вивчення. Наявність шуму в навчальних наборах даних, невеликий набір даних і аномалії анотації даних також додають до цих проблем. Щоб стимулювати дослідницьку діяльність щодо настрою цитування на основі обробки природної мови та підходів, що базуються на машинному навчанні, автори-учасники в цьому домені повинні

зробити своє кодування та анотовані набори даних цитування легкодоступними. Область аналізу настроїв цитування може бути додатково розширена за допомогою приналежності авторів до установ, рейтингу та індексації журналів, щоб отримати більш сукупний рівень аналізу цитувань для надання нюансів дослідженням у бібліометрії. Що стосується кращого стимулювання впливу дослідження та його впливу на настрої, інший напрямок у майбутньому полягає в розгляді загального часу, що минув від публікації, та настроїв, отриманих протягом періоду часу.

3 МЕТОДОЛОГІЯ ДОСЛІДЖЕННЯ

3.1 Вилучення N-грамів

Текст цитат сегментується на речення, а теги частин мови виконуються на наборі даних золотого стандарту [1] для анотування кожного слова та символу. Центральним і важливим аспектом аналізу настроїв є вибір хорошого представлення ознак. Кожне цитатне речення розбивається на n-грами різного розміру (уніграма, біграма, триграма тапентаграма).Уніграми це витягнуті пакети слів, розділені пробілами. Наприклад, у реченні «The proposed method outperformed the class-based model», після видалення стоп-слова «the», слова «proposed», «method», «outperformed», «class», «based», «model» — усі вони є різними уніграмами. Крім того, деякі незалежні від теми слова, наприклад стоп-слова (I, you, an, the, we, you, my, to, his, for, nor, but, yet, or, тощо) видаляються з набору даних аби зробити його придатним для аналізу. Однак усі традиційні стоп-слова не видаляються, оскільки деякі з них, точніше заперечні слова, висловлюють почуття. Наприклад, слово «not», «like» зазвичай не розглядається як стоп-слово, але з огляду на речення «author do not like the performance of the classifier», тут ці слова мають важливу зеперечну силу. Біграми, триграми та пентаграми - це елементи, що складаються з двох, трьох і п'яти послідовних слів. Наприклад, у реченні «the results are not competitive to the state-of-the-art systems», після видалення стоп-слів «the» і «to», «results are», «are no», «‘not competitive», «competitive state-of-the-art», «state-of-the-art system » є відмінними характеристиками біграми. Перевага використання цих функцій полягає в тому, що вони здатні містити деяку контекстну інформацію. Розгляд контекстної інформації в термінах n-грам вищого порядку допоможе не тільки вловити заперечення, але й сприяти вловлюванню тонких значень у формі неявних заперечень. Тому було прийнято рішення на використанні функцій n-грамів вищого порядку

для класифікації настрою цитування.

Приклад цих функцій наведено в таблиці 3.1.

Таблиця 3.1 - Список тегів POS та їх значення

Теги части мови	Значеня
JJ, JJR,JJS	Прикметник, порівняльний прикметник, назывний відмінок прикметника
NN, NNS, NNP,NNPS	Іменник однина, іменник множина, власна назва
CC, PRP, CD	Узгоджувальний сполучник, особистий займенник, цифрове значення
RB, RBR, RBS	Прислівник, порівняльний прислівник, чудова ступінь прикметників
VB, VBD, VBG, VBN,VBP, VBZ	Дієслово, дієслово минулого часу, дієслово в теперішньому часі, дієприкметник минулого часу

Для аналізу настроїв цитування запропонований підхід, який полягає не тільки в розгляді прикметників і прислівників POS, позначених n-грамами самостійно, але й у оцінці їх безперервних і послідовних слів у формі n-грам. Він розглядає довгі прикметники, прислівники та їхні комбінаторні фрази, щоб вивести суб'єктивність, щоб продемонструвати їх доречність для аналізу настрою цитування. Причина використання біграми, триграми та пентаграми прикметників і прислівників полягає в тому, що вони більше несуть почуття та відображають якісне судження щодо фрагмента тексту. Для пентаграм він використовує (JJ/RB -*, *- JJ/RB або *-JJ/RB -*), триграми (JJ/RB -*, *- JJ/RB, *- JJ/RB- *), біграми (JJ/RB -*). Він розглядає довгі прикметники, прислівники та їхні комбінаторні фрази, щоб вивести суб'єктивність, щоб продемонструвати їх доречність для аналізу настрою цитування. Причина використання біграми, триграми та пентаграми прикметників і прислівників полягає в тому, що вони більше несуть почуття та відображають якісне судження щодо фрагмента тексту. Для пентаграм використовується (JJ/RB–*,

*–JJ/RB або *–JJ/RB–*), триграми (JJ/RB–*, *–JJ/RB, *–JJ/RB–*), біграми (JJ/RB–*). Розглядаються довгі прикметники, прислівники та їхні комбінаторні фрази, щоб вивести суб'єктивність та продемонструвати їх доречність для аналізу настрою цитування. Причина використання біграм, триграм та пентаграм прикметників і прислівників полягає в тому, що вони більше несуть почуття та відображають якісне судження щодо фрагмента тексту. Для пентаграм використовується (JJ/RB–*, *–JJ/RB або *–JJ/RB–*), триграми (JJ/RB–*, *–JJ/RB, *–JJ/RB–*), біграми (JJ/RB–*, *–JJ/RB), а уніграми складаються лише з JJ/RB. Тут * означає наявність будь-якої іншої частини мови в реченні.

Беручи до уваги речення «model produces excellent alignment results», модель з тегами POS для розгляду буде такою: «"model NN produces VBZ excellent JJ alignment NN results NNS». Для цього речення триграма матиме вигляд «"excellent alignment results», дві біграми «produces excellent», «excellent alignment» і одна уніграма «excellent».

Той самий метод буде використано для пошуку шаблонів n-грамів на основі тегу прислівника RB. Отже, набір функцій, який надається як вхідні дані для програми аналізу настроїв для експериментів, включає:

- прикметники;
- прислівники;
- комбінацію прикметників + прислівник;
- n-грами.

В роботі основну приділено використанню функцій n-грамів вищого порядку для класифікації настрою цитування. Для n-грамів вищого порядку, ймовірність послідовності тегів POS обчислюється як добуток умовних ймовірностей її триграм і пентаграм. Отже, якщо послідовність тегів позначається як t_1, t_2, t_3, ... t_n і відповідну послідовність слів як w_1, w_2, w_3, ... wn, вище вказаний факт можна пояснити за допомогою рівнянь:

$$L_{\min} \le Len(pos_sequence) \le L_{max} \qquad (3.1)$$

$$P\left(t_i \mid w_i\right) = f\left(t_{i-4}, t_{i-3}, t_{i-2}, t_{i-1}\right) / f\left(t_{i-4}, t_{i-3}, t_{i-2}, t_{i-1}\right)$$

Це забезпечує перехід між тегами та допомагає фіксувати контекст цитованих речень у термінах n-грамів вищого порядку. Імовірності обчислюються за такою формулою в рівнянні:

$$P\left(t_i / t_{i-4}, t_{i-3}, t_{i-2}, t_{i-1}\right) = f\left(t_{i-4}, t_{i-3}, t_{i-2}, t_{i-1}\right) / f\left(t_{i-4}, t_{i-3}, t_{i-2}, t_{i-1}\right))$$

Для аналізу настроїв у цитованому запропонований метод не лише розглядає прикметники та прислівники окремо, але й оцінювала їх безперервні та послідовні слів у формі n-грам. Він також розрахував ступінь подібності кожного шаблону слова в даних цитування d до тексту цитати, як зазначено в рівнянні:

$$resp(dt) = \begin{cases} 1, \text{ ветор цитування містить шаблон як він є, у тому самому порядку;} \\ \alpha.n / N, \text{ n слів із N слів шаблону з″являються в тексті в правильному порядку;} \\ 0, \text{ якщо в тексті цитати немає слів шаблону.} \end{cases} \quad (3.3)$$

3.3 Скіп-грам

Узагальнення для вирішення проблеми розрідженості, пов’язаної з n-грамами, полягає у використанні пропуску грам шляхом пропуску деяких слів у послідовності через довільні пропуски. Скіп-грама має форму k-скіп-n-грама, яка зазвичай є n-грамою, що складається з кількох суміжних членів, а значення k визначає максимальну кількістьслова або терміни, які потрібно пропустити. У прикладі «запропонована аналогічна модель статистичного тегування», «запропонована аналогічна модель статистичного перекладу», «запропонована аналогічна модель статистичного вирівнювання» є прикладом 1-skip-5-gram (5-грамова пентаграма з 1 пропущеним слово, тобто 1-скіп-пентаграма). Враховуючи цитатне речення «The IBM models produce

excellent word alignments», результуюче представлення 1-скіп-2-грами: <The IBM>, <The models>, <IBM models >, <IBM produce>, <models produce>, <models excellent> , < produce excellent>, <produce word>, <excellent word>, <excellent alignments>, <word alignments>. Точніше, якщо між словами дозволено більше проміжків, це призведе до подальшого збільшення кількості вилучених пропусків грамів. Точніше, якщо між словами дозволено більше проміжків, це призведе до подальшого збільшення кількості вилучених пропусків грамів.

Оскільки skip-gram передбачає навколишні/ближчі слова (контекст) у векторному просторі, надаючи поточне слово, цей підхід було використано для отримання герундія прикметника, пов’язаного з іменниковими фразами, для визначення контексту та почуття, пов’язаних із іменниковою фразою . Отже, намір полягає в тому, щоб передбачити сусідні слова, враховуючи фразу центрального слова як основну вхідну інформацію. Для цього він розглянув skip-grams і навчив модель з різними контекстними розмірами вікон (2, 3, 5) для розгляду сусідніх слів перед, після та навколо іменникової фрази в корпусі цитат для обчислення вбудованих слів. Мета полягає в тому, щоб максимізувати вірогідність слова на основі інших слів у тій же послідовності. Скіп-грами використовувалися для розгляду прикладу слів, які існують далеко одне від одного, але розглядаючи їх разом, вони надають суттєву контекстну інформацію. Наприклад, розглядаючи слово «перевершити» як цільову роботу, контекст можна передбачити, це вказуватиме на те, що існує висока ймовірність того, що йогоконтекст: «The technique has really [...] other approaches». Однак, якщо розмір вікна збільшити, якість створеного вектора підвищиться, але витрати на обчислення також зростуть.

3.4 Класифікація настрою цитування на рівні документа

У модулі класифікації настроїв використовувався метод

неконтрольованого аналізу настроїв для визначення полярності настроїв тексту цитати на основі словника SentiWordNet 3.0. Після націлювання на весь профіль термінів цитованого речення, він витягнув лише ті терміни, які позначені як прикметники, прислівники та їх n-грами. Комп'ютерні лінгвісти припускають, що прикметники та прислівники здебільшого представляють почуття в реченні. Наприклад, якщо в цитованому реченні сказано «SVM outperformed other classifiers and demonstrated excellent results», тут присутність прикметника «excellent» свідчить про те, що автор високо оцінює цитовану роботу. Подібним чином наявність прислівників ще більше змінює сентиментальну спрямованість речення. Наприклад, цитатне речення «ambiguity classes have been shown to be successfully employed» виражає більш позитивне ставлення до цитованого дослідження. Оцінка настрою кожного токенізованого рядка, позначеного як прикметники, прислівники та їхні n-грами, обчислюється за допомогою лексичного аналізатора SentiWordNet. Частота появи кожної n-грами також агрегується для кожного з текстів цитат. Усі відповідні бали прикметників, прислівників та їхніх комбінацій агрегуються, щоб отримати загальну оцінку настрою кожного цитованого тексту. Потім було виконано класифікацію настроїв на рівні документа, щоб класифікувати весь текст цитат на «позитивні», «негативні» або «нейтральні» класи. Запропонований алгоритм на основі SentiWordNet, що включає різні кроки для класифікації настрою цитування на рівні документа, показано в лістингу 3.1.

Лістинг 3.1 - Запропонована модель класифікації настрою цитування

```
Модель: Класифікація настрою цитування
Вхідні дані: Набір цитат
Метод:
  For Для кожного контексту цитування
  if контекстуальні речення > 1 Тоді
  Розділіть контекст на речення
  Визначте перехідні сполучники
  End if
  Виконайте тегування POS
  Витяг шаблону POS-тегу (тобто прикметника та прислівника)
  End for
  For Для кожного прикметника та прислівника
```

```
  Виконайте Stemming
  Сформулюйте n-грами різних розмірів
  Обчисліть оцінку настрою кожного токенізованого рядка
  Агрегування загальної оцінки настрою кожного цитованого тексту
на основі n-грамів
  End for
  Ініціалізуйте початкові значення для аналізу настрою
  Введіть n-грами для аналізу настроїв
  Призначте полярність почуття кожному тексту цитати на основі
n-грамів
  Призначте полярність настрою контрастивних контекстів на
основі сили настрою
  Порівняйте результати різних n-грам
  Порівняйте результати з класифікованими вручну цитатами
  Порівняйте результати з інструментами Citations Classified by
Commercial
Результат: модель класифікації
```

Однак деякі речення з цитатами не вказують на полярність настроїв, тому що в них недостатньо чітких слів думки. Фрагмент визначення оцінки настрою та класифікації тексту цитати наведено в таблиці 3.2.

Таблиця 3.2 - Оцінка настрою ознак біграми

ID	БІГРАМИ	SENT VAL	ТИП
4	the same	0	Нейтральний
4	same cut	0	Нейтральний
4	a reasonable	0.45	Позитивний
4	reasonable starting	0.45	Позитивний
4	subsequent research	0.30	Позитивний
10	neglected notable	-0.49	Негативний

Оцінки SentiWordNet обчислюються для позитивних і негативних фраз, знайдених у тексті цитати, і використовуються для визначення орієнтації настрою шляхом класифікації тексту цитати до класу з найвищим балом. У таблиці 3.2 представлено порівняння балів настрою та орієнтації на настрої на основі різних ознак n-грамів. Аналіз результатів демонструє чітку тенденцію покращення та зміцнення показників настрою з n-грамами вищого

порядку порівняно з n-грамами нижчого порядку. З таблиці 3.2 також можна побачити модель зміни полярності настрою від об’єктивного класу до позитивного або негативного. У випадку, якщо є протиріччя між текстом цитати, що складається з більш ніж одного речення, де одне речення цитується як позитивне, а наступне речення або речення цитуються як негативні або нейтральні, абзац спочатку сегментується на підречення. З огляду на контекстне цитування речень та їх синтаксичну структуру визначено роль кожного слова в реченнях. Сегментація виконується просто за допомогою розділових знаків, знаків питання, ком і крапок разом із зібраними вручну ключовими словами, наприклад, «because», «and», «but», «however» і «since» тощо. Для сполучників переходу, головним фокусом є «but» та «however» через їхню більшу частоту появи.

Далі йде екстракція словосполучень, що містять почуття (прикметники, прислівники тощо). Потім розрахунок і агрегування сили настрою полярних слів у кожному реченні. Щоб включити сценарій зміни настрою, речення з the вища сила почуття вважається остаточною полярністю контексту цитування.

Таблиця 3.3 - Порівняння класифікації настрою n-грам

ID	Оцінка (n=1)	Тип (n=1)	Оцінка (n=2)	Тип (n=2)	Оцінка (n=3)	Тип (n=3)	Оцінка (n=5)	Тип (n=5)
1	0,37	П	0,73	П	1.1	П	1.83	П
2	0	О	0	О	0	О	0	О
4	0,44	П	1.18	П	1.63	П	2.51	П
10	0	О	-0,49	Н	-0,98	Н	-1,96	Н
11	0	О	0	О	0,41	П	0,41	П
17	0	О	0	О	-0,49	Н	-1,96	Н

У разі контрастивного переходу та заперечень, наприклад, якщо позитивно цитоване речення супроводжується негативно цитованим

реченням, загальна орієнтація цитатного речення вважатиметься негативною, оскільки тут цитуючий автор має намір приховати сентиментальну орієнтацію. Через те, що заперечення є найпопулярнішою структурою, яка призводить до зміни полярності. Розглядаючи такий сценарій, «This is classical technique. However, results produced are inefficient». Тим не менш, головною метою тут є створення підходу для включення сценаріїв зміни настроїв. «Without», «not», «lack», «never» і «so on» є деякими з відповідних тригерних слів, які розглядаються в таких сценаріях.

3.5 Виявлення аспекту

Наступним кроком у аспектному аналізі настрою цитування є визначення аспектів для аналізу. Основною метою є виділення різних аспектів із цитованих речень, які мають відношення до цитованої роботи. Аспекти - це атрибути або особливості рецензованої дослідницької роботи. Наприклад, ефективний підхід, алгоритм, техніка, надлишковий метод і так далі є різними аспектами дослідницької роботи. Речення з кількома цитатами формують загальне ставлення дослідника до цитованого дослідження, але часто містять різні почуття. Для виявлення аспектів ця техніка зосереджена на тегах NN, NNS, NNP, JJ, DT, VBG POS, оскільки вони представляють іменники, іменники у множині, власні іменники в однині, прикметники, визначники та дієслова-герундії. Одне цитатне речення може містити кілька аспектів цитованої наукової статті. Наприклад, розглядаючи речення цитати, «The approach used can not be generalized, but the corpus is comprehensive».

У цьому одному реченні автор, який цитує, розповідає про два аспекти цитованої наукової статті «approach» і «corpus». У такому сценарії фрази, багаті на почуття, тобто дієслова, прикметники, прислівники тощо, пов’язані з словосполученнями, виділяються за допомогою відношень залежності між ними.

Ключовим моментом тут є фрази з іменниками, які частіше

зустрічаються разом із конкретними фразами мають більшу тенденцію бути справжніми аспектами. Кожен аспект цитатного тексту з’являється в конкретному контексті та має свій настрій і семантику. На аспект не впливають усі контекстуальні фрази. Для кожного виявленого аспекту цитування існує зовсім інший список ключових слів. Наприклад, у цитованому реченні для аспекту «approach» контекстне слово «comprehensive» не таке важливе, як «generalized». Оскільки мітки полярності настроїв потрібні для кожного ідентифікованого аспекту, потрібно зробити сегментацію для цитат, щоб мати справу з кожним із аспектів. І натомість розглянути ці сегменти окремо для виявлення полярності. У наведеному вище прикладі речення цитати можна розбити на пару {Citation Sentence, Aspect}: CiteSent, {Asp1, Asp2} :- {CiteSent, Asp1}, {CiteSent, Asp2}. Для кожної з пар аспектів цитування розглядається слово, яке найкраще описує аспект. Потім розглядається вікно з n слів навколо терміна аспекту, який є сегментом уваги для виявлення цього аспекту. Речення, що цитується може містити різні речення. Аспект вважається явним, якщо він з’являється в цитованому реченні, тоді, якщо він не з’являється, але його значення можуть матися на увазі, тоді він вважається неявним. У наведеному вище прикладі цитування «approach» і «corpus» є явними аспектами, тоді як визначення імпліцитного значення або аспекту речення, наприклад, «using NLP tagger is not a piece of cake», дуже складно. У цій дипломній роботі я зосереджую увагу на явних аспектах цитованої наукової роботи. Потім розглядається вікно з n слів навколо терміна аспекту, який є сегментом уваги для виявлення цього аспекту. Цитатне речення може містити імпліцитні або експліцитні речення. Аспект вважається явним, якщо він з’являється в цитованому реченні, тоді як, якщо він не з’являється, але його значення можуть матися на увазі, тоді він вважається неявним. У наведеному вище прикладі цитування «підхід» і «корпус» є явними аспектами, тоді як визначення імпліцитного значення або аспекту речення, наприклад, «використання НЛП-тегера — це не шматок пирога», дуже

складно. У цій дипломній роботі я зосереджую увагу на явних аспектах цитованої наукової роботи. Потім розглядається вікно з n слів навколо терміна аспекту, який є сегментом уваги для виявлення цього аспекту.

У цій дипломній роботі я зосереджую увагу на явних аспектах цитованої наукової роботи. Використовується лінгвістичний синтаксичний аналізатор Go tagger для виконання тегів POS та аналізу речень цитат. Подібно до попередніх досліджень, де розглядають іменники та іменникові словосполучення як аспекти або ознаки, даний метод також витягує іменникові фрази з цитованих речень, які можуть стати аспектами-кандидатами. Кожне цитатне речення аналізується, а термін-кандидат аспекту використовується для його представлення на основі ключових слів або фраз-підказок, які вказані для його опису. На наступному кроці він ідентифікував POS-тег поточної фрази разом із послідовними POS-тегами до та після поточної фрази. Він виявив різні послідовні візерунки тегів POS різної довжини, які складалися з біграми, триграми та пентаграми up-till. Часті послідовні шаблони використовуються для ідентифікації фраз як аспектів.

Шаблони аспектів-кандидатів можна розділити на три різні категорії:

- іменник-прикметник: відповідає поєднанню герундія іменника та прикметникау якому прикметник описує іменник, наприклад, фрагмент цитати «модель було вдосконалено»;

- іменник-дієслово: відповідає комбінаціям герундія іменника та дієслова, у яких дієслово описує іменник, наприклад, цитатне речення «we refined the corpus»;

- дієслово-прислівник: поєднання герундія дієслова та прислівника означає, що дієслово описується прислівником, наприклад, «техніка використовується широко».

Детальне розуміння моделей аспектів представлено в таблиці 3.4.

Таблиця 3.4 - Витягнуті шаблони фраз

Шаблон	Перший термін	Другий термін	Третій термін	Четвертий термін
Шаблон 1	JJ	NN/NNS/NNP	–	–
Шаблон 2	JJ	NN/NNS/NNP	NN/NNS/NNP	–
Шаблон 3	RB/RBR/RBS	JJ	–	
Шаблон 4	RB/RBR/RBS	JJ/RB/RBR/RBS	NN/NNS/NNP	–
Шаблон 5	RB/RBR/RBS	ВБН/ВБД	–	
Шаблон 6	RB/RBR/RBS	RB/RBR/RBS	JJ	
Шаблон 7	VBN/VBD	NN/NNS/NNP	–	
Шаблон 8	VBN/VBD	RB/RBR/RBS	–	
Шаблон 9	VBZ/VBN/VBD	JJ	NN/NNS/NNP	
Шаблон 10	JJ	JJ	NN/NNS/NNP	
Шаблон 11	RB/RBR/RBS	RB/RBR/RBS	JJ	NN/NNS/NNP
Шаблон 12	VBN/VBD	RB/RBR/RBS	RB/RBR/RBS	

Термін аспекту кваліфікується на основі наступних умов:

- він має бути іменником;

- частота появи результуючого аспектного шаблону, що містить кандидатний аспектний термін, має бути помітно високою в ядрі цитувань; якщо виділений іменник не існує в частому шаблоні POS-тегу, він не розглядається як аспект;

- виділені іменникові фрази порівнюються зі словником синонімів, створеним на основі (Hern´andez-A´lvarez та ін., 2016); іменники, яких немає в словнику ігноруються;

- рівень участі кандидата аспектного терміну високий на основі тесту хі-квадрат.

Головна основа цього підходу полягає в тому, що іменники, які мають багато думок рецензентів, швидше за все, розглядатимуться як важливі та відмінні аспекти порівняно з тими, які не містять частих думок. Оскільки аспект цитованої роботи може бути виражений за допомогою різних слів, таких як Corpus, Corpora, Dataset або Data автором, який цитує, створено вектор аспектів.

4 РЕЗУЛЬТАТИ ЕКСПЕРЕМЕНТАЛЬНИХ ДОСЛІДЖЕНЬ

4.1 Опис експериментального набору даних

Експерименти проводяться на двох різних наборах даних, що належать до домену інформатики та біоінформатики. Перший набір є золотим стандартом, що складається з 8736 речень цитат, витягнутих із 310 наукових статей із мережі антології ACL [1]. Це анотований вручну набір даних, з якого 829 посилань є позитивними, 280 є негативними та 7627 є нейтральними. Другий набір даних отримано з матеріалів клінічних випробувань, які складаються з думокречення цитат, взяті з 285 довільно вибраних статей [15]. Кілька цитат із посилань вважаються однією цитатою, оскільки вони представляють однакову полярність настроїв. Набір даних складається з 4182 речень цитат, взятих із розділу обговорення документів про клінічні випробування, які стверджують відтворюваність біомедичних досліджень. Він містить більшість нейтральних зразків, тобто 3172, разом із 702 позитивними та 308 негативними цитатами. Таблиця 4.1 представляє описову статистику для обох наборів даних.

Таблиця 4.1 - Описова наборів даних

	Вибрані документи	Цитувані речень	Позитивні цитати	Негативні цитати	Нейтральні цитати
Набір даних 1	310	8736	829 (9,5%)	280 (3,2%)	7627 (87,3%)
Набір даних 2	285	4182	702 (17%)	308 (7,36%)	3172 (75,9%)

Це показує розподіл класів у наборах, який є дуже незбалансованим: 87% і 76% є нейтральними цитатами, і лише близько 14% і 24% є суб’єктивними для набору даних 1 і 2 відповідно. Обидва набори даних були

використані в дослідженнях у сфері аналізу настроїв цитування, які встановили, що більшість цитат є нейтральними за своєю природою. Це обмеження та добре встановлений факт у сфері аналізу цитат, оскільки більшість цитат цитуються як нейтральні. У цьому сценарії екземпляри з основного класу (речення з нейтральним цитуванням) були викинуті, тобто (недостатня вибірка), оскільки для кожного класу була однакова кількість маркованих зразків перед навчанням класифікаторів. Цей підхід допоміг підвищити чутливість класифікатора для класу меншин.

В роботі використовуються почуття, аспект, IMRaD і розширений контекст до набору Athar et al. який складається з 310 статей антології ACL. Оригінальний набір даних складається з 8736 цитат, з яких 829 позначені як позитивні, тоді як 280 позначені тегами негативних настроїв. Оскільки основною одиницею інтересу є цитоване речення та його релевантне розташування в цитованій дослідницькій статті. Було зібрано PDF-версії документів, до яких належать речення з цитатами, вилучив метадані (формат XML) і розділи/розташування, у яких з'являється речення з цитатами. Для цього експериментального дослідження в першу чергу було зосереджено на реченнях цитування, що містять настрої (p – позитивне, n – негативне), а також випадково вибраних об'єктивні речення. Було розширено обсяг цитованих речень і вручну додано контекст цитати для кожного з цитованих речень.

Базовий набір складається приблизно з 2109 контекстів цитування, що відповідає приблизно 24% вихідного набору даних. Мета полягала в тому, щоб вивчити поведінку та вплив контексту цитування, у якому з'являється цитата, на полярність настроїв цитованих речень. У наявному наборі даних немає інформації щодо риторичної структури, де існує цитатне речення. Таким чином, існуючий набір даних було розширено для визначення розташування кожного речення в дослідницькій статті. Наступним завданням була категоризація розділів на основі структури з чотирьох розділів: вступ, метод, результат та обговорення.

Контексти цитування також були анотовані розділами статті («Вступ», «Огляд літератури», «Методологія», «Результати», «Обговорення», «Висновок» та інші), у яких вони з'являлися. Мітки та назви розділів дозволили вирівняти дослідницькі статті в нашому наборі даних. Для цього заголовки розділів, у яких з'явилися релевантні цитати, були проаналізовані та позначені тегами на основі регулярних виразів, що відповідають кожному типу розділу, з урахуванням можливих варіантів заголовків розділів. Понад 98% цитованих речень у корпусі стосуються цих чотирьох розділів, але не обов'язково втака ж послідовність. У цьому конкретному дослідженні я не враховував порядок розділів у дослідницькій статті. Цитатне речення, відповідні контекстуальні речення разом із полярністю настрою, виглядом розділу та релевантністю витягуються та зберігаються в базі даних Oracle. Приклад анотації зображено на рисунку 4.1 з контекстом цитування, його міткою класу разом із файлом XML для довідкової інформації.

Turney (2008) has recently proposed a simpler SVM-based algorithm for analogical classification called PairClass. While it does not adopt a set-based or distributional model of relational similarity. We have noted above that PairClass implicitly uses a feature representation which does not produce compatible results.

citationText	Turney (2008) has recently proposed a simpler SVM-based algorithm for analogical classification called PairClass.	P
postSentence1	While it does not adopt a set-based or distributional model of relational similarity.	N
postSentence2	We have noted above that PairClass implicitly uses a feature representation which does not produce compatible results.	N

```
- <annotatedPaper id="1">
  - <paper title="Using lexical and relational similarity to classify semantic relations " year="2009" authors="Diarmuid O Seaghdha, Ann Copestake">
    - <citationContext>
      - <citationText>
          Turney (2008) has recently proposed a simpler SVM-based algorithm for analogical classification called PairClass.
          <section> Literature Review </section>
          <cite polarity="p">Turney (2008)</cite>
        </citationText>
      - <postSentence1>
          While it does not adopt a set-based or distributional model of relational similarity.
          <cite polarity="n">Turney (2008)</cite>
        </postSentence1>
      - <postSentence2>
          We have noted above that PairClass implicitly uses a feature representation which does not produce compatible results.
          <cite polarity="n">Turney (2008)</cite>
        </postSentence2>
      </citationContext>
    </paper>
  </annotatedPaper>
```

Рисунок 4.1 - Приклад контексту цитування

4.2 Аналіз настрою цитування

У цьому розділі представлено результати щодо аналізу настрою цитування, яке пов’язане з впливом прикметника вищого порядку, прислівника та їх комбінації n-грамних словосполучень на класифікацію настрою цитування. Щоб відповісти на це запитання, було використано попередньо позначений і класифікований вручну набір для експериментів, що буде служити базовою лінією для порівняння та оцінки результатів. Крім того, також було позначино і класифіковано набір даних, що складається з тексту цитат, використовуючи два інструменти комерційного аналізу настроїв SEMANTRIA та THEYSAY. Порівняв результати класифікації за нашою методикою з класифікованим вручну текстом і з комерційними інструментами. Рисунки 4.2, 4.3 і 4.4 представляють відсоток позитивних цитат розмічених комерційних інструментів. Результати класифікації, виконаної комерційними інструментами, відрізняються, що встановлює, що ці інструменти реалізовані на основі різних внутрішніх алгоритмів. Різні відсотки позитивних, негативних та нейтральних цитат і описують, що обидва інструменти справляються з негативністю та нейтральністю по-різному. За даними SEMANTRIA, серед цих 8730 цитованих текстів 2264 мають клас позитивних настроїв, 1266 мають негативну орієнтацію, а решта позначені як нейтральні. Тоді як, згідно з THEYSAY, цей розподіл становить 4802 позитивних, 2130 негативних і 1799 нейтральних. У наборі даних, позначених вручну, більшість цитат, позначених як нейтральні або об’єктивні, тобто 87%, тоді як відсоток нейтральності чи об’єктивності значно менший у комерційних інструментах, які становлять 59,5% за SEMANTRIA та 20,5% за THEYSAY відповідно. Класифікація, виконана за допомогою анотації вручну, сильно спотворена в бік об’єктивності, і той самий шаблон зображено в SEMANTRIA. Це показує, що SEMANTRIA демонструє чітку тенденцію призначення об’єктивних позначок набору

даних цитування, у порівнянні з THEYSAY, який демонструє упереджену поведінку щодо позитивного класу. Порівнюючи тенденцію міток класів, створених комерційними інструментами, з базовим набором даних, анотованим вручну, встановлено, що результати, отримані SEMANTRIA, ближчі до еталонних, що стосується всіх міток класів. Наявність прикметників і прислівників у тексті цитати використовується для класифікації його як позитивного, негативного та нейтрального. В експерименті,Я видобув мішок слів, що складаються з прикметників і прислівників на основі n-грам. Для цієї мети я не тільки розглянув уніграми (n = 1) і біграми (n = 2), але також використовував використання триграм (n = 3) і пентаграм (n = 5). Найбільш конкретною причиною використання n-грам вищого порядку є те, що більшість сучасних підходів стверджує, що вони можуть відігравати більш значну роль у виявленні настроїв, ніж n-грами нижчого порядку. Таким чином, у цьому дослідженні я досліджував, як різні пари слів високого порядку поводяться у визначенні орієнтації настроїв у наборі даних цитування. Я не лише розглядав прикметники та прислівники окремо, щоб визначити полярність настрою тексту цитати, але також оцінював їх за різними комбінаціями. Використання n-грам високого порядку як вхідної функції призвело до високовимірного простору ознак (19 648 вимірів уніграми, 55 796 вимірів триграм і 87 599 вимірів пентаграм) у нашому сценарії. Витягнуті пакети слів сортуються відповідно до частоти їх появи в наборі даних, щоб вибрати найважливіші з них. Основною причиною цього було виявлення частих термінів у науковій літературі для виявлення настроїв. У таблиці 4.1 наведено перелік найуживаніших уніграм, біграм, триграм і пентаграм прикметників і сполучень прислівників, за якими формується думка.

Тут представлено результати класифікації настрою цитування за допомогою використання прикметника, прислівника та їх комбінацій за допомогою n-грам високого порядку на рівні документа. Тут не тільки порівняв результати, отримані за допомогою запропонованого підходу, з

анотованими вручну набором даних, але й зіставив результати з класифікованими цитатами за допомогою двох комерційних інструментів. У таблиці 4.2 наведено значення метрик оцінки (точність, відкликання та швидкість розпізнавання) та проведено порівняння з анотаціями вручну та результатами класифікації з анотаціями, виконаними комерційними інструментами.

Таблиця 4.2 - Топ-20 прикметників

Most Frequent Adjective Уніграми (n=1)	High, good, successful, effective, popular, efficient, important, interesting, robust, difficult, considerable, easy, competitive, suitable, poor, central, appealing, straightforward, satisfactory, dominant
Most Frequent Adjective Біграми (n=2)	such as, statistical machine, natural language, good performance, significant improvements, human judgments, minimum error, maximum entropy, automatic evaluation, the same, human evaluation, more accurate, in recent, current state, Recent work, more sophisticated, very successful, be effective, most notable
Most Frequent Adjective триграми (n=3)	such as the, minimum error rate, statistical machine translation, with human judgments, in recent years, phrase based statistical, from many heterogeneous, In recent years, Introduction In recent, the past few, to be effective, past few years, successful in recent, current state of, greedy search algorithm,various NLP tasks, techniques such as
Most Frequent Adjective пентаграми (n=5)	proven increasingly successful in recent, increasingly successful in recent years, major developments in statistical approaches, current state of the art, notable examples of unsupervised polarity, received much attention in recent, many advances in recent years, such as part of speech, past five years important research, Introduction In recent years statistical

Таблиця 4.3 - Precision, Recall and Accurac

	Manual			SEMANTRIA			THEYSAY		
Шаблони	P	R	F-score	П	P	F-бал	П	P	F-бал
Adj. (1-g)	70,54	40	55.27	75.5	29.17	52.34	44,45	80	62.23
Adj. + Adv. (1-g)	75,54	55,56	65,55	94.11	55,56	74,84	53,35	78.17	65,76
Adj. (2-g)	75	50	62.5	70	54.17	62.09	55,56	76,92	66.24
Adj. + Adv. (2-g)	88,89	80	84,45	85.5	92.30	88.9	58.06	85,71	71,89
Adj. (3-g)	65,55	50	57,78	96	58.33	77.17	70,54	80	75,27
Adj. + Adv. (3-g)	88,89	72,72	80,81	96	68,57	82,29	59,38	82,60	70,99
Adj. (5-g)	75,54	60	67,77	94.11	66,67	80,39	54.17	81.26	67,72
Adj. + Adv. (5-g)	90	81,82	85,91	96.15	96.15	96.15	55,88	82,60	69,24

Рисунок 4.2 - Відсоток міток класів вручну класифікованого тексту цитат

Рисунок 4.3 - Відсоток міток класів, створених SEMANTRIA

Рисунок 4.4 - Відсоток міток класів, створених THEYSAY

На рисунку 4.5, 4.6, 4.7 і 4.8 було побудовано частоту термінів, пов’язаних із текстом цитати, включаючи їх варіації.

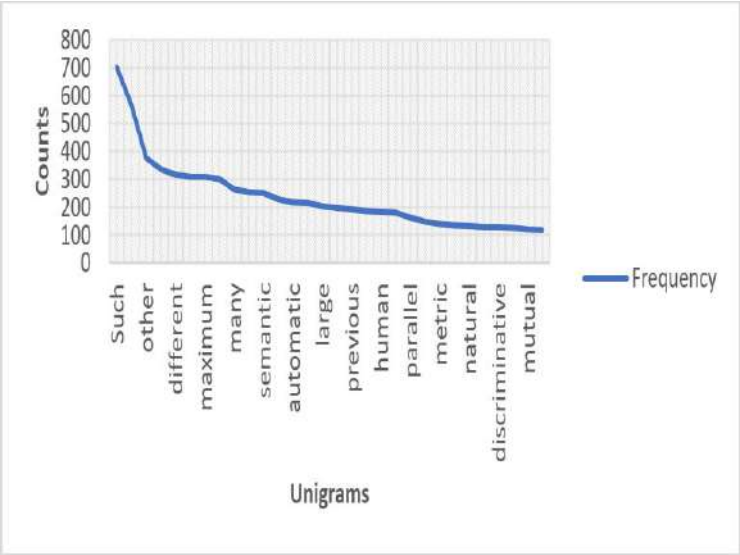

Рисунок 4.5 - Частотний графік уніграм прикметників

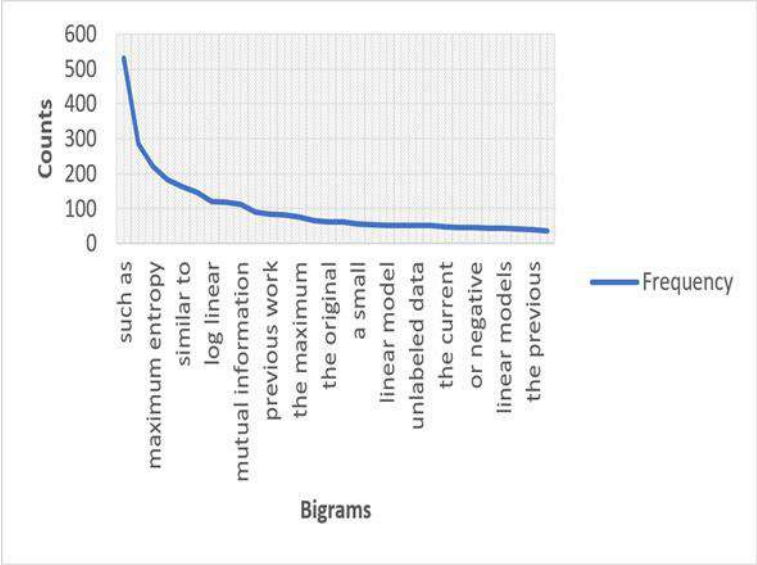

Рисунок 4.6 - Частотний графік біграм прикметників

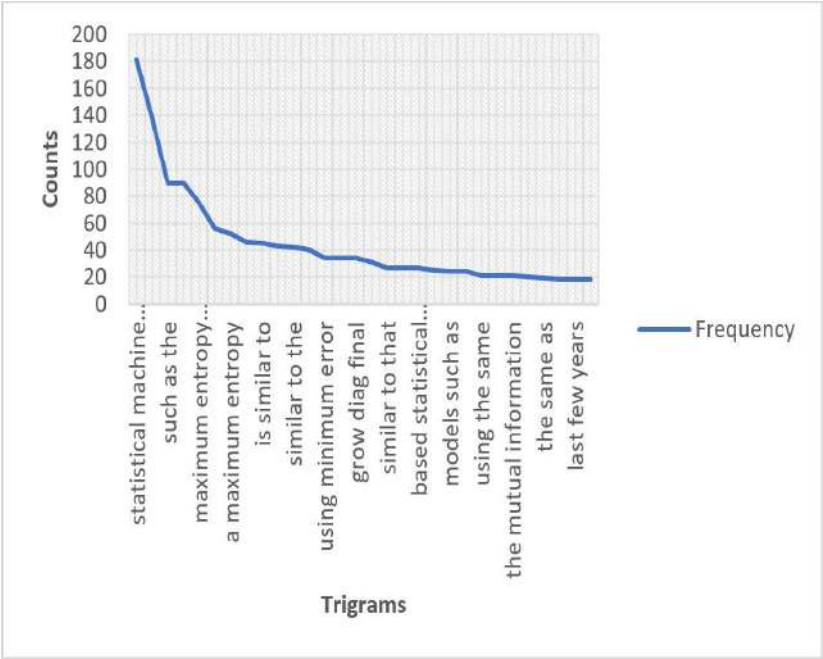

Рисунок 4.7 - Частотний графік триграм прикметників

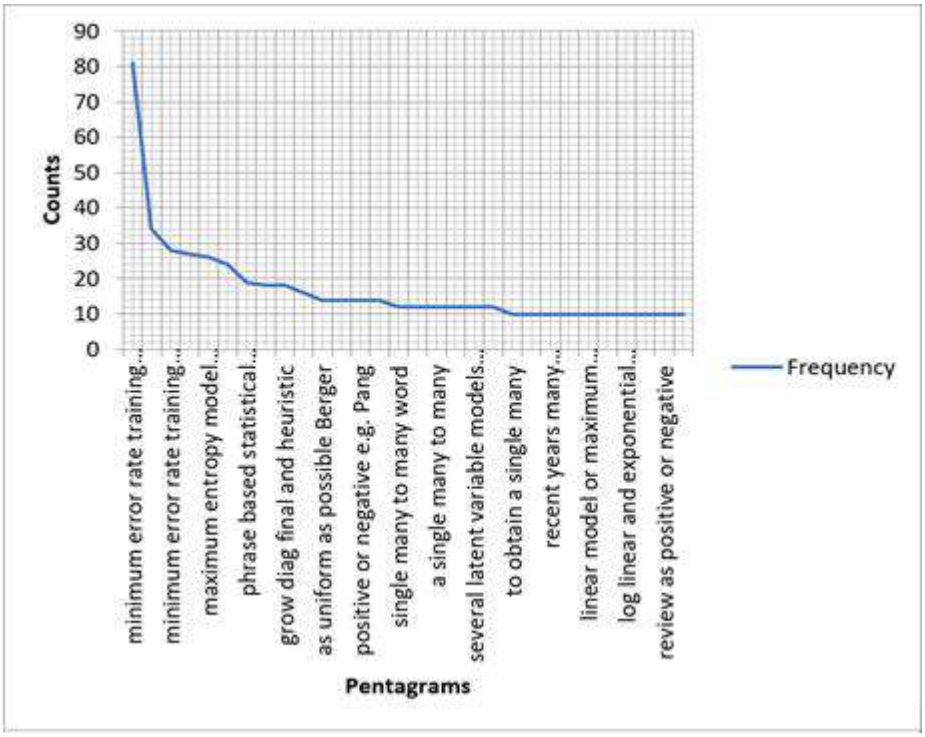

Рисунок 4.8 - Частотний графік пентаграм прикметників

Можна проаналізувати, що значення точність, отримана за допомогою n-грамів вищого порядку, краща, ніж значення,

отримане за допомогою n-грамм нижчого порядку. Тому, коли проводився аналіз, використовуючи «одне слово тобто, уніграма», «подвійне слово тобто, біграма», «потрійне слово тобто, триграма», «п’ять слів тобто, пентаграма», прикметники, прислівники та їх поєднання. Значення точності, отримане за допомогою комбінацій прикметників і прислівників, порівняно краще, ніж значення, отримане за допомогою окремого прикметника та прислівників. Триграма і пента-gram перевершують n-грами нижчого порядку, що встановлює використання n-грам вищого порядку як відповідну характеристику для корпусу цитат. Тепер обговоримо точність висновків результатів класифікації. Коли я вибираю n-грами як ознаки, семантична інформація частково втрачається або нехтується більш конкретно у випадку BOW, тобто уніграм. Результати запропонованого дослідження показують точність понад 80% порівняно з анотованим вручну корпусом із використанням комбінацій прикметників і прислівників із n-грамою вищого порядку, як показано на рисунку 4.9.

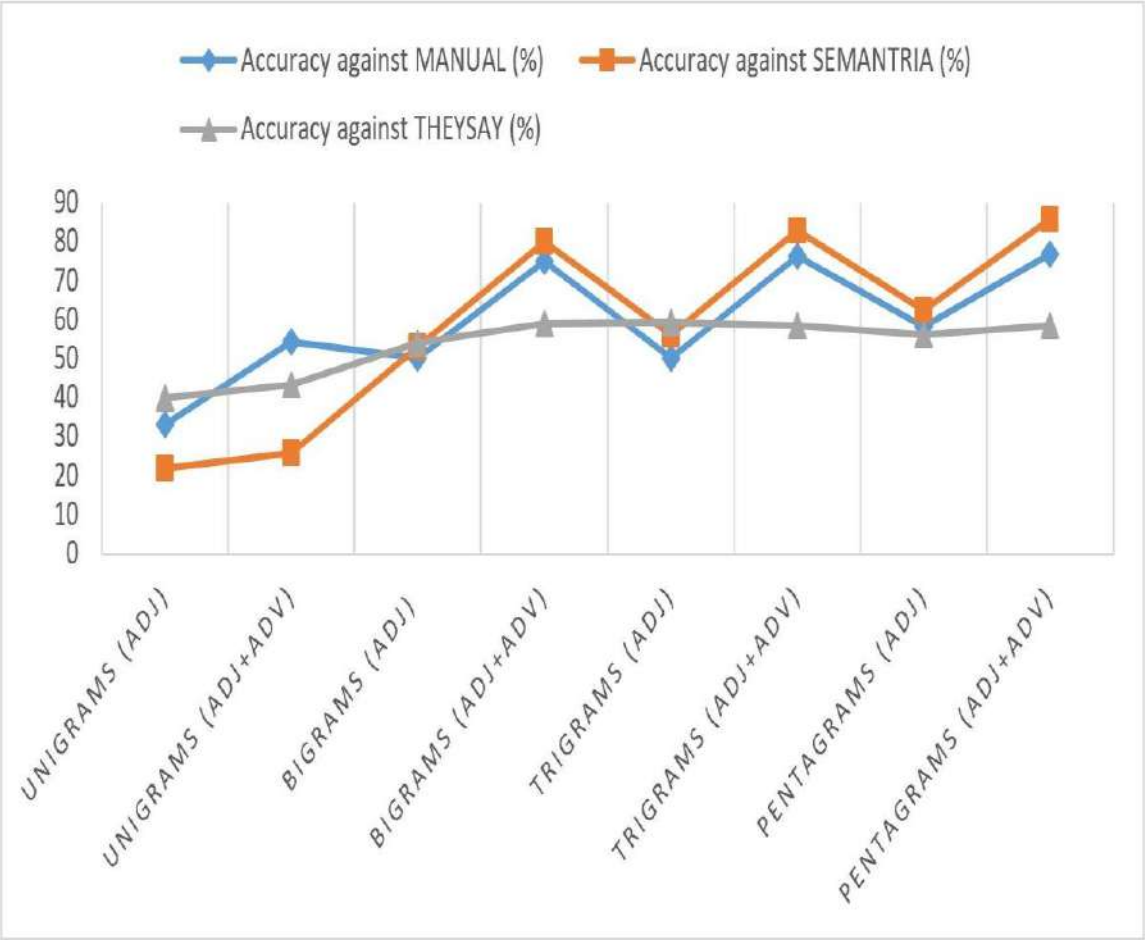

Рисунок 4.9 - Порівняння точності з базовою лінією

З результатів можна помітити, що зі збільшенням значення n точність класифікації підвищилася більш конкретно, коли прикметники та прислівники використовуються в комбінації. Це додатково підтверджує, що комбінації прикметників і прислівників із n-грамою вищого порядку є більш точними та детермінованими виразами, ніж n-грами нижчого порядку. З точністю вище, 90% для комбінацій прикметників і прислівників, SEMANTRIA створила найточніші прогнози, дотримуючись тієї самої лінії тренду, яка зображена на фоні вручну. Це показує, що ці функції (комбінації прикметників і прислівників із n-грамою високого порядку) дають кращі результати порівняно з уніграмами та біграмами. Аналіз результатів показав, що точність n-грам нижчого порядку (n = 1) залишилася незмінною для окремих ознак (прикметників і прислівників) і обох у комбінаторному

порядку. Результати також засвідчили кращу точність середніх позитивних цитат порівняно з негативними та нейтральними. Крім того, об'єктивні цитати були неправильно класифіковані як позитивні або негативні. Можлива причина цього полягає в тому, що важко передбачити нейтралітет і заперечення. Аналіз результатів також показав, що використання n-грам вищого порядку може вирішити проблему композиційності (розуміння складного виразу через значення складових виразів). Ще раз варто зазначити, що точність де порівняно з THEYSAY стабільно нижчий порівняно з SEMANTRIA та ручною анотацією всіх функцій. Ця демонстрація в основному через упереджену прогнозну поведінку THEYSAY щодо позитивного класу. Однак для лінії тенденції точності THEYSAY також очевидно, що поєднання прикметника та прислівника має високий рівень точності порівняно з ізольованими ознаками. Це також підтверджує і посилює точність результатів класифікації. Хоча SEMANTRIA не може вважатися найкращим інструментом для аналізу настроїв, однак він продемонстрував постійну поведінку для різних наборів даних, наприклад, даних опитувань охорони здоров’я (Georgiou et al. 2015) і багатомовних наборів даних (Tandon and Jain 2012). THEYSAY виконала неправильну класифікацію великих пояснювальних речень. Оскільки середній розмір пропозиції експериментального набору даних становив 35 слів, це змусило THEYSAY зробити більшість помилкових прогнозів. На рисунку 4.10 для комбінацій прикметників і прислівників і n-грам високого порядку, запропонована методика досягла середньої точності приблизно 90% щодо ручної класифікації та 96,15% щодо SEMANTRIA відповідно.

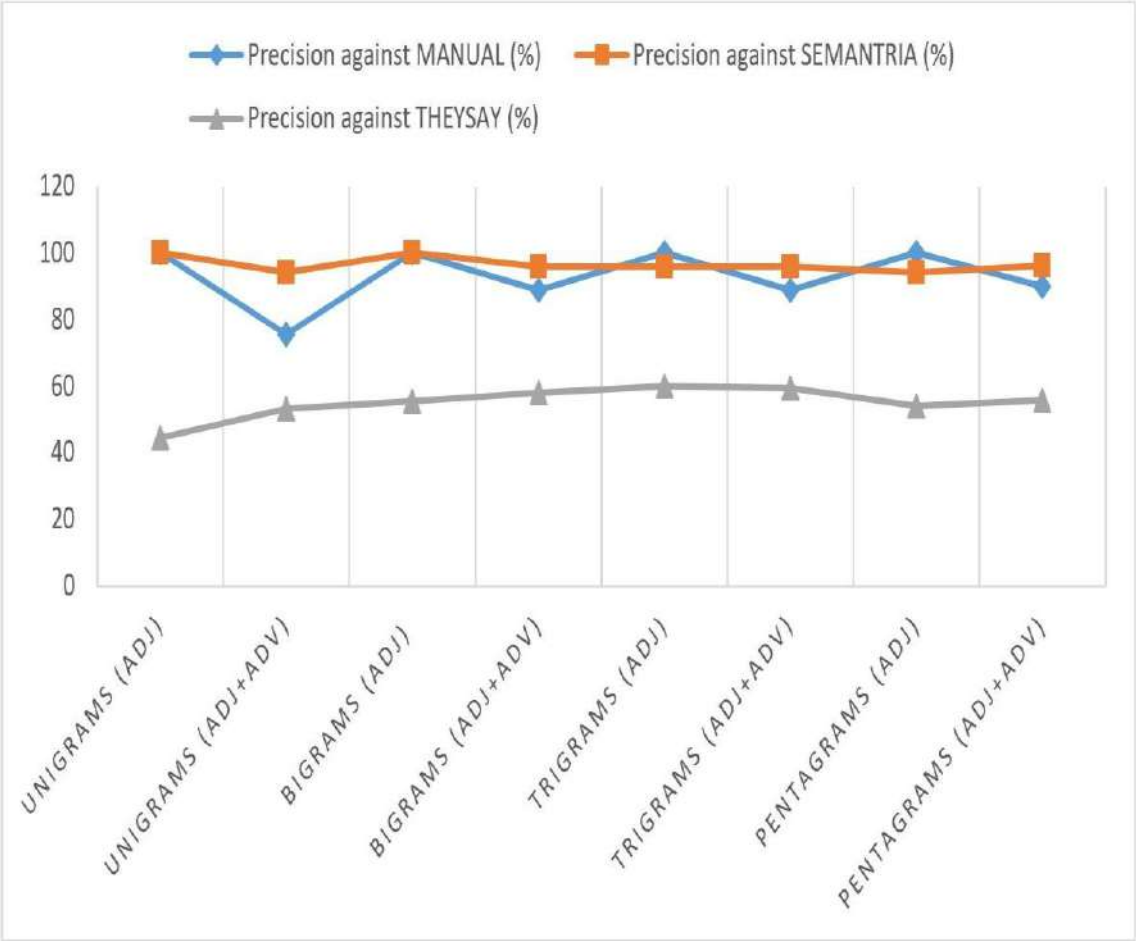

Рисунок 4.10 - Порівняння точності з базовою лінією

Таким чином, ефективність запропонованої методики знаходиться в значному діапазоні в порівнянні з результатами ручної класифікації та анотації з SEMANTRIA. Варто також зазначити, що запропонована нами методика значно покращила точність порівняно з пригадуванням. Точність і відкликання порівняно з THEYSAY незмінно низькі для всіх функцій. Високу точність і значення запам’ятовування можна спостерігати на рисунку 4.10 4.11 як для окремої особи, так і для комбінацій прикметників і прислівників у порівнянні з SEMANTRIA, що додатково забезпечує надійність згаданого інструменту. Комбінації прикметників і прислівників на основі n-грамів високого порядку є кращим вибором для прогнозування настрою на рівні документа для великих речень. Крім того, щоб перевірити обчислювальне навантаження, Я досліджував вплив розміру корпусів

цитувань на ефективність запропонованої системи.

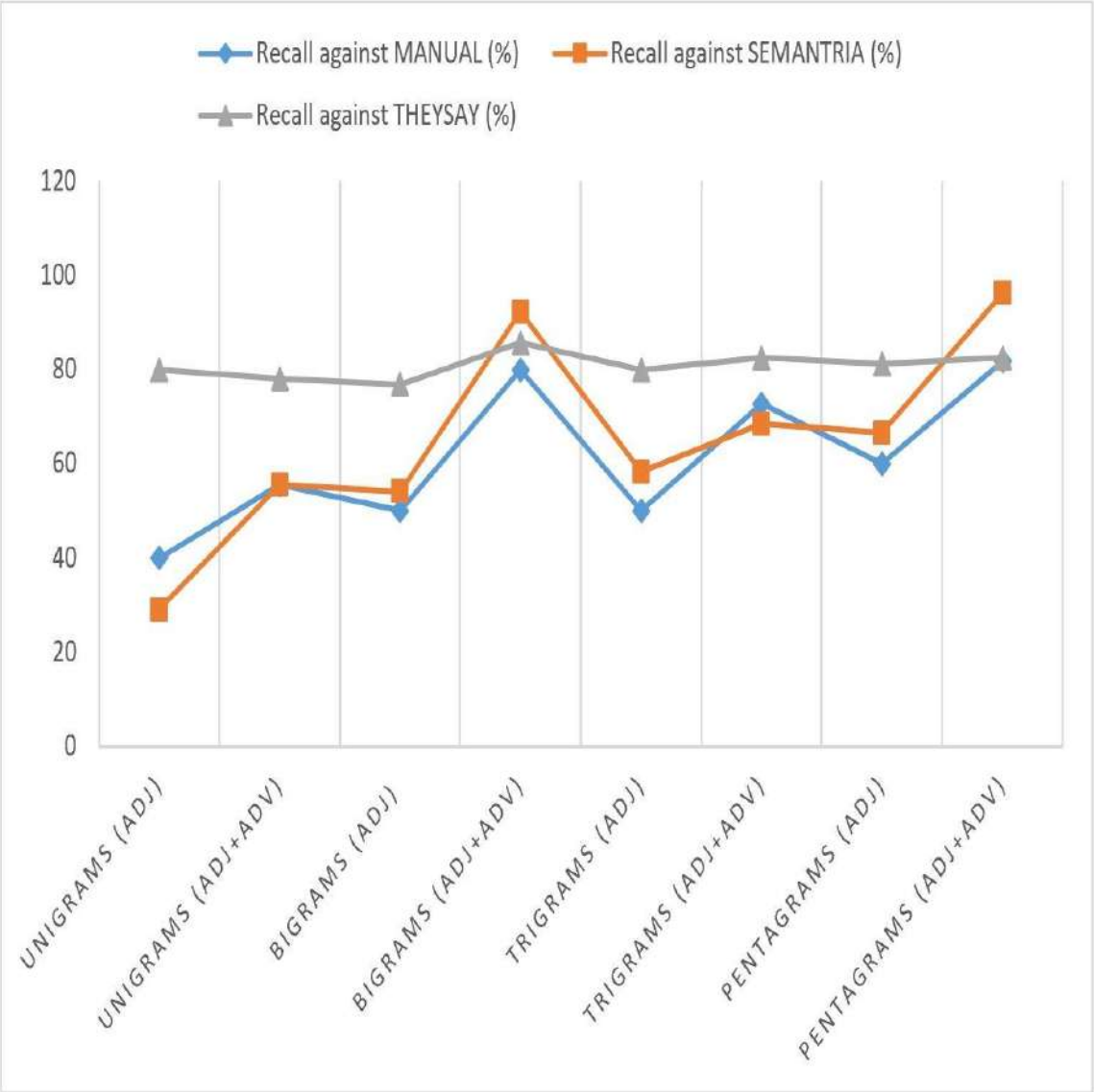

Рисунок 4.11 - Порівняння Recall з базовим рівнем

Результати представлені на рисунку 4.11. Як видно з графіка, зі збільшенням розміру екземплярів у вибірці відбувається покращення продуктивності системи. Однак у певний момент часу покращення не може бути досягнуто простим збільшенням розміру навчального набору даних, коли розмір корпусу достатньо великий.

ВИСНОВКИ

В кваліфікаційній роботі було досліджено ефективність використання прикметників, прислівників та їхніх комбінацій для класифікації настрою цитованого тексту на рівні документа за допомогою золотого стандарту набору настрою цитування.

Після цього було досліджено ефективність застосування комерційних інструментів до набору цитат для виявлення настроїв. Аналіз результатів виявив, що n-грами високого порядку (n = 5) для комбінацій прикметників і прислівників відіграють основну роль у покращенні точності класифікації настрою.

Результати дослідження показали точність понад 80% порівняно з анотованим вручну набором даних із використанням комбінацій прикметників і прислівників із n-грамою вищого порядку. Порівняльний аналіз результатів із сучасним дослідженням, заснованим на використанні n-грамів різної довжини, показав, що уніграма, біграма та триграма показали кращі результати у комбінації порівняно з уніграмою та уніграмою плюс біграма.

Експериментальні результати також визначили, що поточні інструменти аналізу настроїв, точніше THEYSAY, недостатньо ефективні для точного визначення настроїв, оскільки більшість цитат і триграма показали кращий результат у комбінації порівняно з уніграмою та уніграмою плюс біграма.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. A. Athar, “Sentiment analysis of citations using sentence structure-based features,” in Proceedings of the ACL 2011 student session. Association for Computational Linguistics, 2011, pp. 81–87.

2. M. Hern´andez and J. M. G´omez, “Sentiment, polarity and function analysis in bibliometrics: a review,” Natural Language Processing and Cognitive Science, p. 149, 2015.

3. S. Stamou, N. Mpouloumpasis, and L. Kozanidis, “Deriving the impact of scientifi c publications by mining citation opinion terms.” JDIM, vol. 7, no. 5, pp. 283–289, 2009

4. J. Oh, H. Chang, J. A. Kim, M. Choi, Z. Park, Y. Cho, and E.-g. Lee, “Citation analysis for biomedical and health sciences journals published in korea,” Healthcare informatics research, vol. 23, no. 3, pp. 218–225, 2017.

5. N. Tandon and A. Jain, “Citation context sentiment analysis for structured summarization of research papers,” in 35th German conference on artificial intelligence. Citeseer, 2012, pp. 24–27.

6. R. Klavans and K. W. Boyack, “Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge?” Journal of the Association for Information Science and Technology, vol. 68, no. 4, pp. 984–998, 2017.

7. B. H. Butt, M. Rafi, A. Jamal, R. S. U. Rehman, S. M. Z. Alam, and M. B. Alam, “Classification of research citations (crc),” arXiv preprint arXiv:1506.08966, 2015.

8. P. Kazi, M. Patwardhan, and P. Joglekar, “Towards a new perspective on context based citation index of research articles,” Scientometrics, vol. 107, no. 1, pp. 103–121, 2016.

9. X. Zhu, P. Turney, D. Lemire, and A. Vellino, “Measuring academic influence: Not all citations are equal,” Journal of the Association for Information

Science and Technology, vol. 66, no. 2, pp. 408–427, 2015.

10. C. Jochim and H. Sch¨utze, “Towards a generic and flexible citation classifier based on a faceted classification scheme,” in Proceedings of COLING 2012, 2012, pp. 1343–1358.

11. X. Wan and F. Liu, “Are all literature citations equally important? automatic citation strength estimation and its applications,” Journal of the Association for Information Science and Technology, vol. 65, no. 9, pp. 1929–1938, 2014.

12. S. D. Tembhurnikar and N. N. Patil, “Topic detection using bngram method and sentiment analysis on twitter dataset,” in 2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO)(Trends and Future Directions). IEEE, 2015, pp. 1–6.

13. H. Alam, A. Kumar, T. Werner, and M. Vyas, “Are cited references meaningful? measuring semantic relatedness in citation analysis.” BIRNDL@ SIGIR (1), vol. 1888, pp. 113–118, 2017.

14. B. Yu, “Automated citation sentiment analysis: What can we learn from biomedical researchers,” Proceedings of the American Society for Information Science and Technology, vol. 50, no. 1, pp. 1–9, 2013.

15. . Xu, Y. Zhang, Y. Wu, J. Wang, X. Dong, and H. Xu, “Citation sentiment analysis in clinical trial papers,” in AMIA annual symposium proceedings, vol. 2015. American Medical Informatics Association, 2015, p. 1334.

16. C. Dong and U. Sch¨afer, “Ensemble-style self-training on citation classification,” in Proceedings of 5th international joint conference on natural language processing, 2011, pp. 623–631

17. Рисенко О. С. et al. "Аналіз методів класифікації тональності текстів." Збірник наукових праць ΛΌГОΣ (2020): 36-41.