

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ  
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ  
імені ІГОРЯ СІКОРСЬКОГО»  
Інститут прикладного системного аналізу  
Кафедра математичних методів системного аналізу

«На правах рукопису»  
УДК 004.9

«До захисту допущено»  
Завідувач кафедри  
\_\_\_\_\_ О.Л.Тимошук  
«\_\_» \_\_\_\_\_ 20\_\_ р.

**Магістерська дисертація**  
на здобуття ступеня магістра  
зі спеціальності 124 Системний аналіз  
на тему: Інформаційна система аналізу тональності новин  
на основі системного підходу

Виконав: студент II курсу, групи КА-61м  
Рись Артем Андрійович \_\_\_\_\_

Науковий керівник: професор кафедри  
математичних методів системного аналізу  
ІПСА КПІ ім. Ігоря Сікорського,  
д.т.н., проф. Данилов В.Я. \_\_\_\_\_

Рецензент: професор кафедри інформаційної  
безпеки ІПСА КПІ ім. Ігоря Сікорського,  
д.т.н., проф. Качинський А.Б. \_\_\_\_\_

Засвідчую, що у цій магістерській  
дисертації немає запозичень з праць  
інших авторів без відповідних посилань.  
Студент \_\_\_\_\_

Київ  
2018

## РЕФЕРАТ

Магістерська дисертація: 72 с., 5 рис., 23 табл., 2 додатки і 37 джерел.

Об'єкт дослідження – семантична орієнтація текстів новин.

Предмет дослідження – метод Наївного Байєса та згорткові нейронні мережі.

Мета роботи – дослідження семантичної орієнтації тексту, використовуючи різні підходи до побудови списку ознак та різні методи оцінки орієнтації (позитивної, негативної). Було проведено огляд існуючих моделей, що використовуються для побудови списку ознак та оцінки семантичної орієнтації, підібрані оптимальні.

Методи дослідження – нейронні мережі, методи обробки текстів.

Основний результат даного дослідження – це розробка алгоритмів класифікації тональності новин. Для цього вивчаються два методи. По-перше, це алгоритм Наївного Байєсу, який використовує репрезентативну групу для класифікації. Другий - це згорткова нейронна мережа, яка складається зі згорткових шарів, агрегувальних шарів, повноз'єднаних шарів та шарів нормалізації. Були проведені експерименти та дослідження ефективності двох різних алгоритмів, що виявляють позитивні та негативні окраси тексту. Крім того, необхідно визначити алгоритм, що дає кращі результати; ще важливо вивчити, як точність алгоритмів може впливати на попередню обробку даних, вибір ознак та даних.

Основними джерелами даних - є платформа Twitter, експертні дані оцінки впливу новин на економіку США, заголовки новин з Австралійського джерела новин ABC. Аналіз проводився з використанням мови програмування Python та таких бібліотек для аналізу даних, як pandas, sklearn тощо.

**СЕМАНТИЧНА ОРІЄНТАЦІЯ ТЕКСТУ, НАЇВНИЙ БАЙЄС, ЗГОРТКОВІ НЕЙРОННІ МЕРЕЖІ, ПОБУДОВА СПИСКУ ОЗНАК ТЕКСТУ, АЛГОРИТМИ КЛАСИФІКАЦІЇ**

## ABSTRACT

The theme title is «Information system for tonality analysis of the text».

Masterr's thesis: 71 p., 5 fig., 23 tabl., 2 appendices and 37 sources.

Object of research - semantic orientation of news.

Subject of research - the Naive Bayes method and the convolutional neural network.

The main goal - studying the semantic orientation of the text, using different approaches to constructing a list of features and different methods of assessing the orientation (positive, negative). A review of existing models used to build a list of features and evaluate semantic orientation was carried out, and optimal ones were selected.

Methods of investigation - neural networks, methods for text processing.

The main result of this study is the development of algorithms for the classification of news tones. For this purpose two methods are studied. First, it is the Naive Bayes algorithm, which uses a representative group for classification. The second one is a convolutional neural network, which consists of spinning layers, aggregation layers, full-blown layers and normalization layers. Experiments and studies of the effectiveness of two different algorithms that show positive and negative colors of the text were carried out. In addition, it is necessary to determine the algorithm that gives better results; It is still important to study how accuracy of algorithms can affect the pre-processing of data, the choice of attributes and data.

The main sources of data are the Twitter platform, the expert data on the impact of news on the US economy, news headlines from the Australian news source ABC. The analysis was conducted using the Python programming language and such libraries for data analysis like pandas, sklearn, etc.

SEMANTIC ORIENTATION OF THE TEXT, NAIVE BAEYS,  
CONVOLUTIONAL NEURAL NETWORKS, FEATURE EXTRACTION,  
CLASSIFICATION ALGORITHM

## ЗМІСТ

ПЕРЕЛІК СКОРОЧЕНЬ .....	7
ВСТУП.....	8
РОЗДІЛ 1 ОГЛЯД ЛІТЕРАТУРИ.....	10
1.1 Вступ.....	10
1.2 Огляд підходів.....	12
1.2.1 Підхід, заснований на лексиконі .....	12
1.2.1.1 Словниковий підхід.....	12
1.2.1.2 Корпусний підхід.....	14
1.2.2 Підхід, заснований на машинному навчанні .....	16
1.2.2.1 Навчання без вчителя.....	16
1.2.2.2 Навчання з учителем .....	18
Висновки до розділу .....	22
РОЗДІЛ 2 РОЗРОБКА МЕТОДІВ АНАЛІЗУ ТОНАЛЬНОСТІ НОВИН.....	23
2.1 Вступ .....	23
2.2 Розробка інформаційної системи.....	24
2.3 Збір та обробка даних .....	24
2.3 Розробка методів класифікації тексту .....	27
2.3.1 Вилучення ознак з тексту .....	27
2.3.2 Алгоритми класифікації.....	29
2.3.2.1 Алгоритм Наївного Байєса .....	29
2.3.2.2 Згорткова нейронна мережа .....	32
Висновки до розділу .....	36
РОЗДІЛ 3 АНАЛІЗ РЕЗУЛЬТАТІВ .....	38
3.1 Оцінка показників ефективності алгоритмів .....	38

3.2 Класифікатор Наївного Байєса .....	39
Висновки до розділу .....	41
<b>РОЗДІЛ 4 РОЗРОБКА СТАРТАП-ПРОЕКТУ ..</b>	<b>Ошибка! Закладка не определена.</b>
4.1 Інформаційна карта проекту .....	<b>Ошибка! Закладка не определена.</b>
4.2 Команда стартап-проекту.....	<b>Ошибка! Закладка не определена.</b>
4.3 Бізнес-модель Canvas проекту .....	<b>Ошибка! Закладка не определена.</b>
4.4 Аналіз ринкових можливостей запуску стартап-проекту..	<b>Ошибка! Закладка не определена.</b>
	<b>не определена.</b>
4.5 Розроблення ринкової стратегії проекту .	<b>Ошибка! Закладка не определена.</b>
4.6 Розроблення маркетингової програми стартап-проекту ....	<b>Ошибка! Закладка не определена.</b>
	<b>не определена.</b>
Висновки до розділу .....	<b>Ошибка! Закладка не определена.</b>
<b>ПЕРЕЛІК ПОСИЛАНЬ.....</b>	<b>Ошибка! Закладка не определена.</b>
<b>ДОДАТОК А ЛІСТИНГ ПРОГРАМИ .....</b>	<b>Ошибка! Закладка не определена.</b>
<b>ДОДАТОК Б ГРАФІЧНІ МАТЕРІАЛИ.....</b>	<b>Ошибка! Закладка не определена.</b>

## ПЕРЕЛІК СКОРОЧЕНЬ

AT - аналіз тональності

CNN - згорткова нейронна мережа

NB - Наївний Байєс

SMV – Support Vector Machine

ReLU - Rectified Linear Unit

## ВСТУП

Різниця між людьми та машинами полягає в тому, що люди мають здатність сформулювати особисті думки, і мрія штучного інтелекту полягає в тому, щоб машина поводитися як людина. Область комп'ютерної лінгвістики, яка аналізує думки, називається аналіз суспільної думки (opinion mining) або, як її також називають, латентним аналізом (sentiment analysis). Вивчення думки є частиною обробки природної мови, яка стосується аналізу думок про продукти, послуги та навіть людей. Грунтуючись на [1], аналіз настроїв та «opinion mining» в першу чергу зосереджуються на думках, що вказують на позитивні або негативні настрої. Але спочатку думки людей про певні події, продукти тощо повинні бути вилучені.

На сьогоднішній день пошук та вилучення думок стало простішим, оскільки люди діляться поглядами на різні теми через соціальні мережі, такі як Twitter, Facebook або вони залишають коментарі та огляди продуктів на спеціалізованих сайтах. Мікроблогінг - надзвичайно популярний спосіб обміну думками, і він щодня виробляє величезну кількість повідомлень. Отже, мікроблогінг може розглядатися як багате джерело надійних повідомлень, які можна збирати та використовувати для аналізу почуттів. Аналіз думок відіграє важливу роль у всіх галузях науки (політиці, економіці та суспільному житті). Наприклад, у маркетингу, якщо продавець знає про задоволення замовника певним товаром, він може краще оцінити попит на продукт. Те саме для політиків, вони будуть знати, чи підтримують їх люди чи ні.

Задача класифікації настроїв - це не нова наукова сфера. Проте основна увага досліджень полягала у аналізі великих документів, а не на мікроблогах, які сьогодні потрібні. Twitter є одним з прикладів платформи мікроблогів. «Твіт» (tweet) - це коротке повідомлення (максимум 140 символів), яке може містити думку або просто висловлювати деякі факти. Класифікація твітів є важким завданням, оскільки твіти можуть містити іронію, орфографічні помилки, смайлики, сленг, аббревіатури і можуть містити лише кілька слів.

Основні підходи, які можуть бути використані для аналізу почуттів - машинне навчання [1], [2], [3], [4], [5] та підхід, заснований на лексиконі [1], [2], [6], [7], [8]. Підхід машинного навчання використовує набір даних для класифікатора навчання, який буде додатково застосований для визначення настроїв певного тексту. Метод, заснований на лексиконі, використовує семантичну орієнтацію слова або фраз, щоб визначити, чи є текст позитивним чи негативним.

Платформа Twitter буде використовуватися як джерело думок. Ця робота зосереджена на аналізі різних методів, які будуть використовуватися для навчання та тестування мережі. Більш того, головним інтересом нашого дослідження є пошук ефективних алгоритмів, які можуть застосовуватися для цілей класифікації твітів. У цій роботі алгоритми машинного навчання будуть застосовуватися для класифікації почуттів. Важливо відзначити, що контрольоване навчання є найбільш прикладною технікою для класифікації почуттів. Точніше, основна увага в цій темі полягає в алгоритмі Наївного Байєса та згортковим нейронним мережам як класифікаційним алгоритмам.



## РОЗДІЛ 1 ОГЛЯД ЛІТЕРАТУРИ

### 1.1 Вступ

Аналіз тональності тексту це не нова задача, вона була вивчена ще в 90х роках. Проте, на початку 21 століття, аналіз тональності зацікавив вчених завдяки своєму значенню в різних наукових сферах, а також в АТ було багато не досліджених наукових питань [1]. Більш того, широка наявність даних з висловлюванням суспільної думки, підштовхнула дослідження в цій галузі на новий етап. З тих пір аналіз тональності став швидко розвиватися. За словами Бінг Лю в [1]: “аналіз сентиментальності, який також називають пошуком соціальної думки, полягає у вивченні аналізу думок людей, почуттів, оцінок, поглядів, емоцій до таких суб’єктів, як продукти, послуги, організації, приватні особи, проблеми, події, теми та їх атрибути.”

Іншими словами, аналіз настроїв стосується обробки тексту з виразом суспільної думки, щоб витягувати та класифікувати думки з певного документа. Тональність почуттів зазвичай виражається в позитивній або негативній думці (бінарна класифікація [9], [10]). Однак це може бути багатокласова класифікація [11], [12], [13], отже, настрої можуть мати нейтральну мітку або навіть розширену варіацію міток як дуже позитивну, позитивну, нейтральну, негативну, дуже негативну, також мітки можуть бути пов’язані з емоціями, такими як сум, гнів, страх або щастя.

Аналіз тональності - це область, що розвивається, що викликає зацікавленість людей та особливо організацій, оскільки АТ може бути використаний для прийняття рішень. Люди більше не обмежені запитувати думки друзів про конкретний товар або послугу, вони можуть вільно знайти таку інформацію в Інтернеті. Крім того, організації можуть заощаджувати час та кошти, уникаючи проведення опитувань, замість того, вони можуть зосередити увагу на обробці думок, які можна вільно отримати з Інтернету. Тим не менше, потрібно пам’ятати,

що джерела, які містять дані з суспільними думками, є іноді зашумленими, тому важливо витягувати зміст з цієї інформації, щоб використовувати їх далі.

АТ використовує різні методи та підходи для вирішення цього складного завдання:

а) рівень документа [14]. На цьому рівні основним завданням є визначення думки всього документа (думка повинна бути щодо однієї теми);

б) рівень речення [10]. Тут кожне речення розглядається як короткий документ, який може бути суб'єктивним або об'єктивним. Суб'єктивне (виражене) висловлювання виражає почуття;

в) рівень аспектів (рівень ознак) [15]. Дозволяє отримувати думки щодо аспектів суб'єктів.

Методи класифікації аналізу настроїв переважно поділяються на методи машинного навчання та лексичні підходи [2] (рис. 1.1). Більш детальні пояснення цих методів будуть наведені в наступному підрозділі.

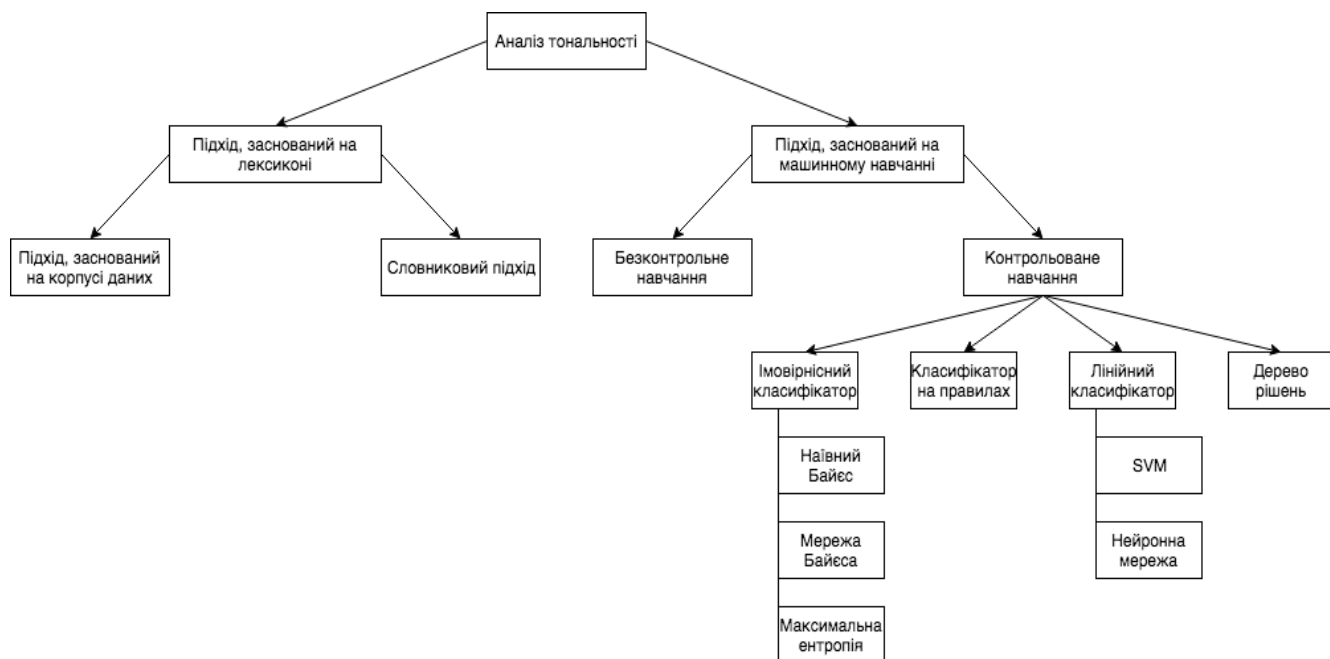


Рисунок 1.1 – Підходи до аналізу тональності

## 1.2 Огляд підходів

Ця глава присвячена огляду підходів, з використанням лексикону та машинного навчання.

### 1.2.1 Підхід, заснований на лексиконі

Перший метод, який можна використовувати для тонального аналізу, - це метод, оснований на лексиконі. Він використовує лексикон, який складається з умов з відповідними оцінками почуттів до кожного терміну. Термін може бути пов'язаний з одним словом, фразою або ідіомом [16]. Почуття визначається виходячи з наявності або відсутності термінів у лексиконі. Підхід, оснований на лексиконі, включає в себе підхід, що базується на корпусі, та словниковий підхід, який обговорюється далі.

#### 1.2.1.1 Словниковий підхід

Основна ідея словникового підходу полягає в тому, щоб використовувати лексичні бази даних з думками, щоб витягувати почуття з документів. Виходячи з [1], [12], множина розмічених сентиментальних слів (наприклад, добре, погано) з їх полярністю збирається вручну. На початку цей початковий набір не повинен бути великим, достатньо 30 думок [7]. Наступним кроком є використання полярних слів, щоб збагатити набір, шукаючи відповідні синоніми на антоніми в лексичній базі даних. Прикладами таких баз даних є WordNet [17], HowNet [18], SentiWordNet [19], SenticNet [19], MPQA [19] та ін. Процедура пошуку ітеративна. На кожній ітерації алгоритм приймає оновлений набір слів (розширений набір) і робить пошук

знову, доку не буде нових слів для включення. Зрештою, сукупність сентиментальних слів може бути переглянута з метою видалення помилок.

Ху и Лю [7] зосередили свої дослідження на класифікації відгуків клієнтів, а саме, вони вилучали характеристику продукту, які містять почуття, а потім класифікували пропозиції на основі цих особливостей, в результаті чого було складено резюме оглядів продукту. Наприклад, якщо огляд стосується камери, автори знайшли такі особливості, як якість зображення та розмір камери, класифікація проводилася на позитивних та негативних оглядах камери. Щоб призначити позитивний чи негативний тег для пропозиції, спочатку дослідники видали полярні слова для кожного огляду. У цьому випадку були використані прикметники. Прогноз базувався на полярності прикметника, який мав таку саму полярність, як її синоніми і протилежність полярності його антонімів. Полярні слова використовувались для пошуку їх синонімів на антонімів з відомими орієнтаціями в WordNet. Завдяки цьому, було виявлено полярність слів, що з'являється в огляді. Метод описаний у [7], показав хороші результати, середня точність становила 84%. Отже, даний метод може бути ефективним для прогнозування семантичних орієнтирів прикметника та полярності речення.

Кім та Хові [20] досліджували полярність тексту та його власника щодо даної теми. Автори дослідження застосували декілька класифікаторів. Перший класифікатор був застосований до кожного слова в реченні, щоб отримати його полярність. Другий класифікатор визначав полярність всього речення, вираженого власником думок. Крім того, автори запровадили використання малого початкового списку слів вже розмічених слів аналогічним чином [7] (прикметник та дієслова). Цей список слів був розширений, шукаючи відповідні синоніми та антоніми в WordNet. Автори згадували, що деякі синоніми/антоніми мали нейтральну або навіть протилежну спрямованість, що робить їх неприйнятними для використання. Крім того, дослідники наголошували на необхідності визначення сили позитивності та негативності слів, які дозволять усунути неоднозначні слова. Автори визначили чотири різні регіони в реченні, які є близькими до власника думки та можуть містити настрій речення. Для визначення сентиментальності

думки автори розробили три моделі. Перша модель приймала припущення, що при зустрічі одного негативного слова інші негативні вже не впливають. Друга та третя моделі - це гармонійне та геометричне значення сильних почуттів у конкретному регіоні відповідно. Після проведення експериментів було зроблено висновок, що найкращі результати отримані за допомогою першої моделі та регіону, яка починається від власника думки до кінця речення.

У роботі [21] автори розробили метод, який використовував три різних словника (традиційно використовується лише один) для отримання синонімів та антонімів. Після цього розгорнутий лексикон використовувався для класифікації твітів. Автори сказали, що їх запропонована методика дозволила класифікувати твіти, на що традиційний словниковий метод був нездатний. Тим не менше, запропонований підхід має кілька недоліків. Основна проблема полягає в тому, що збір такої кількості синонімів та антонімів вимагає багато часу. Крім того, зазвичай словники містять формальні слова, але твіти повні неформальної лексики.

Як правило, головним недоліком словникового підходу є нездатність виявляти сентиментальні слова, пов'язані з якоюсь специфічною тематикою.

#### 1.2.1.2 Корпусний підхід

В [1] Бінг Лю зазначає, що корпусний підхід може застосовуватися у двох випадках. Перший - ідентифікація оціночних слів та їх популярності в корпусі, використовуючи заданий набір слів. Другий підхід полягає в тому, щоб побудувати новий лексикон у межах певного домену з іншої лексики, використовуючи корпус домену. Результати показують, що навіть якщо слова, що виражають думки, залежать від домену, може статися, що одне і те ж слово буде мати протилежну орієнтацію залежно від контексту.

Дослідження, проведенні Nazivassiloglou та McKeown [8], відомі в літературі. Автори запропонували метод, який витягує семантичну спрямованість з'єднаних прикметників із корпусу. Методика базується на використанні текстуальних слів та прикметників, які виражають думки. Спеціальні лінгвістичні правила

застосовуються до тексту, щоб виявляти важливі слова з відповідними орієнтаціями. Автори припускають, що прикметники мають однакову полярність, якщо вони пов'язані словом «and». А слово «but» пов'язує слова з протилежними семантичними орієнтаціями. Додатково використовуються словосполучення «or», «either-or», «neither-nor». Іноді ці правила не можуть бути застосовані. Тому автори передбачають полярність об'єднаних прикметників для перевірки того, чи є полярності однакові чи ні, для цього використовується модель логістичної регресії. Після етапу прогнозування отримується граф, який забезпечує зв'язок між прикметниками. Тоді кластеризація виконується на графіку, щоб розділити прикметники на позитивні та негативні підмножини. В кінці кінців, автори змогли досягти точності в 90%.

Як згадувалося вище, одне і те ж сентиментальне слово може мати різну семантичну спрямованість залежно від контексту. Ding та співавтори [22] запропонували метод пошуку орієнтації настроїв, переданих рецензентами. Автори підкреслювали, що деякі прикметники (в основному квантори, такі як довгі, короткі тощо) залежать від контексту і можуть змінювати полярність. Дослідники вказують на зміст слів з їх аспектами у реченні, щоб визначити полярність ознаки продукту. Список прикметників та прислівників взято із [7] і доповнюється авторами дієсловами та іменниками. Більш того, вони анотували більш 1000 ідіом, що містять чітко виражені настрої. Після того, як лексикон готовий, вони визначають оцінку полярності для кожної ознаки в тексті. Щоб отримати оцінку для всього тексту, вони підсумовують всі оцінки за допомогою запропонованої функції оцінки, яка дає кращі результати, ніж простий підсумок, який був використаний в [7]. Крім того, автори запропонували декілька лінгвістичних правил для обробки заперечень та пропозицій, які містять зв'язок «but». Крім того, в документі представлений цілісний підхід до вирішення проблеми виявлення полярності контекстно-незалежних сентиментальних слів.

## 1.2.2 Підхід, заснований на машинному навчанні

Другий спосіб, який можна використати для аналізу тональності - це машинне навчання, яке включає в себе методи навчання з учителем та навчання без учителя.

### 1.2.2.1 Навчання без вчителя

Цей метод використовує не позначені набори даних, щоб виявити структуру та знайти схожі патерни з вхідних даних. Навчання без вчителя зазвичай використовується, коли збір надійного анотованого набору даних є складною задачею, але не позначених - простою. При такому підході не викликає жодних труднощів при завантаженні нових даних, які навіть не з цієї тематики. Turney [3] використовує даний метод для класифікації відгуків. Кожен з відгуків класифікується як рекомендований або не рекомендований. Автор витягує фрази які складаються з двох слів на основі шаблонів тегів. Шаблони розроблені таким чином, що вони повинні захоплювати сентиментальні фрази. Кожна фраза це комбінація прикметника/прислівника та дієслова/іменника (загалом пропонується 5 моделей). Для того того, щоб вирішити, які фрази треба завантажувати, до документа застосовується ідентифікатор частини мовлення POS (part-of-speech). Важливо зазначити, що фраза витягується, якщо два слова підпадають під одну з запропонованих моделей. Наступним кроком є розрахунок семантичної орієнтації отриманих фраз з огляду. Автор застосовує алгоритм Pointwise Mutual Information та Information Retrieval algorithm (PMI-IR) для пошуку семантичної орієнтації. PMI вимірює семантичну подібність між двома термінами. Фраза, яка відповідає моделям, вважається першим терміном, а словосполучення вважається другим терміном. Слова, такі як «Excellent» та «roog» розглядаються як довідкові слова, оскільки природньо оцінювати огляд як «roog», коли він отримує одну зірку і «excellent», якщо огляд отримує п'ять зірок. Семантична орієнтація фрази

визначається як різниця між PMI фрази «excellent» та PMI фрази «poor». Семантична орієнтація позитивна, якщо фраза має сильніший зв'язок з довідковим словом «excellent» і негативним, якщо асоціація є сильнішою зі словом «poor». Для розрахунку PMI необхідно визначити ймовірності співпадіння відповідних термінів. Останній крок полягає в тому, щоб визначити настрій для всього огляду (рекомендовано чи не рекомендовано). Якщо середня семантична орієнтація є позитивною, огляд визначається як рекомендований, і не рекомендується в іншому випадку. Як повідомлялося, середня точність склала 74%.

Rothfels та Tibshirani [23] застосували метод навчання без вчителя для класифікації рецензій на фільми. Автори адаптували метод, запропонований Загібаловим та Carroll [24] для класифікація китайського тексту. Ідея методу полягає в тому, щоб використовувати позитивні слова, які можна отримати з документа. Такі сентиментальні слова (прислівники) є попередніми запереченнями або можуть бути й без заперечення (найбільш поширений випадок). Наявність початкового зерна Забігалова та Carroll збагатили перелік позитивних слів, застосовуючи ітераційну класифікацію. Натхнені даним підходом автори статті склали початковий набір слів. Текст документа, що підлягав класифікації, був розділений на зони, кожна з яких відповідає фрагменту тексту, розташованому між пунктуаційними символами. Потім вже була виконана класифікація кожної зони. Семантична орієнтація всього тексту визначається переважанням позитивних або негативних зон у документі. А саме, якщо позитивні зони зустрічаються частіше, ніж негативні зони, то документ вважається позитивним, інакше негативним. Автори також розширили список початкових слів. Вони намагалися використати бі-, три-, чотири-грами в якості початкових слів. Проте бі- та три-грами не змогли зберегти змісту фраз. Використання чотири-грам дало незадовільні результати. Автори зробили другу спробу, але вже на цей раз використовували семантично значущі прикметники в якості початкових слів. Тим не менше, підвищення точності не було досягнуто. Дослідники також намагалися змінити метод підрахунку на k-середніх кластеризацію, але отримані результати також не показали суттєвого покращення.



### 1.2.2.2 Навчання з учителем

Методи навчання з учителем передбачають наявність розмічених навчальних даних, які використовуються для навчального процесу. Зазвичай модель «bag-of-words» [4] використовується для представлення документа в якості вектору ознак

$$d = (w_1, w_2, \dots, w_i, \dots, w_N) \quad (1.1)$$

де  $N$  це набір унікальних термів в тренувальному наборі даних та  $w_i$  це вага  $i$ -ої ознаки. Для перетворення навчального набору в векторну форму, потрібно створити словник з  $N$  унікальних слів. Далі, будь-яка з моделей може бути використана для подубови вектору ознак:

а) бінарна модель ознак.  $w_i$  визначається як 1, якщо ознака присутня в документі, інакше – 0;

б) частота ознак (TF) - визначає кількість разів, яку термін зустрічається в документі;

в) TF-IDF (IDF - зворотня частота в документі). IDF - вимірює важливість ознаки (TF припускає, що всі ознаки однаково важливі);

г) тест Хі-квадрат буде описаний в наступній главі.

Після того, як дані були представлені як вектор, він може бути використаний класифікатором для навчання та отримання міток. Досить багато методів може бути використано для тренування класифікатора.

Найбільш простий та поширений метод, який використовується для класифікації тексту, це метод Наївного Байєсу [4], [25], [26], [27]. Ця модель базується на теоремі Баєйса з припущенням, що всі ознаки є незалежними. Наївний Байєсовий класифікатор визначає ймовірність того, що документ належить до конкретного класу. Переваги Байєсового класифікатора це - простота реалізації, швидкий процес навчання та доволі гарні результати [10], [26], [27]. Проте,

«наївне» припущення може викликати проблему, тому що в реальних обставинах ознаки є залежними одна від одної.

Згідно з [25] «the idea behind Maximum Entropy models is that one should prefer the most uniform models that satisfy a given constraint». Ймовірність того, що документ належить до конкретного класу [25], [26] розраховується наступним чином:

$$P(c|d, \lambda) = \frac{\exp[\sum_i \lambda_i f_i(c, d)]}{\sum_{c'} \exp[\sum_i \lambda_i f_i(c', d)]} \quad (1.2)$$

де,  $c$  - це клас,  $d$  - це документ, який потрібно класифікувати,  $\lambda$  - це вага  $i$ -ого класифікаційного індикатора  $f_i$ . Класифікатор максимальної ентропії не припускає незалежність ознак. Однак, цей класифікатор теоретично може дати кращі результати, ніж класифікатор Наївного Байєса, але цей алгоритм є більш складним для реалізації та процес навчання є довшим.

Інший підхід до класифікації - заснований на правилах. Ідея, яка лежить в основі, визначити набір правил, які будуть згенеровані експертами на основі аналізу області. Цей метод може показати гарні результати, коли використовує дуже широкий набір правил. Однак, створення такої кількості правил займає дуже багато часу. Підхід, заснований на правилах, був використаний Chikersal та іншими [28]. Вони запропонували правила, які залежали від використання емоцій та сентиментальних слів в твітах. Крім цього, автори [28] використали Support Vector Machine (SVM) класифікатор. Вони використовували лінійне ядро та L1-регуляризацию; також вони застосували такі підходи, як n-грами, POS-теги, словесні n-грами на декількох різних лексиконах. Головна ідея їх підходу це скомбінувати два різні методи для підвищення точності та зниження відхилень. Кожен твіт, який мав нейтральну характеристику завдяки SVM класифікатору, далі був проаналізований класифікатором, який використовує правила, для визначення

кінцевої оцінки. Роблячи висновок, можна сказати, що підхід, який використовує правила, може покращити прогнози, які були дані SVM класифікатором.

SVM класифікатор також використовувався в роботах [25], [26], [27]. Модель SVM є представленням зразків як точок у просторі, відображених таким чином, що зразки окремих категорій розділені чистою прогалиною, яка є щонайширшою. Нові зразки тоді відображаються до цього ж простору, й робиться передбачення про їхню належність до категорії на основі того, на який бік прогалини вони потрапляють.

Інакшим способом розв'язання проблеми з класифікації текстів є підхід з використанням нейронних мереж (NN). Штучна нейронна мережа дотримується принципів біологічної нейронної мережі. Передбачається, що нейронна мережа може вирішити проблеми таким же чином, як вони можуть бути вирішені людьми. NN - це сукупність взаємопов'язаних нейронів. Як правило, NN має декілька шарів. Нейронна мережа здатна вчитися, регулюючи вагу нейронів. Розглянемо наступні типи нейронних мереж, які використовувались для класифікації тексту:

а) згорткові нейронні мережі (CNN). CNN використовують різновид багат шарових перцептронів, розроблених так, щоби вимагати використання мінімального обсягу попередньої обробки. Вони відомі також як інваріативні відносно зсуву або просторово інваріативні штучні нейронні мережі, виходячи з їхньої архітектури спільних ваг та характеристик інваріативності відносно паралельного перенесення. Згорткові мережі було натхнено біологічними процесами, в яких схему з'єднання нейронів натхнено організацією зорової кори тварин. Крім того, Kim [5] використовував CNN для класифікації фраз. Він класифікував речення на позитивні та негативні, а також на більш детальні класи, а також визначив, чи є пропозиція суб'єктивною або об'єктивною думкою, і класифікує речення на 6 категорій. Для цього використовувалися різні набори тестових даних. Як і в попередніх дослідженнях [5], використовувався один шар CNN. Було повідомлено, що модель показала хороші результати і "pre-trained vectors are 'universal' feature extractors that can be utilized for various classification tasks";

б) рекурентна нейронна мережа (RNN). RNN - це мережа зі зворотним зв'язком, що дозволяє зберігати інформацію про попередній момент часу. У цьому типі мережі вихід, який було обчислено на попередньому кроці, використовується для обчислення наступного. Потім вихід порівнюється з даними тесту і оцінюється коефіцієнт помилки, заснований на тому, які ваги кориговані, що робить процес навчання більш точнішим. RNN корисний для прогнозування наступного слова в реченні [29]. Ця властивість дозволяє краще зрозуміти речення, зафіксувавши контекст кожного слова на основі попередніх. Pengfei Liu та інші [30] застосували RNN для класифікації тексту з багатозадачним навчанням. У якості завдань вони обрали наступне: класифікацію 5 класів, бінарну класифікацію, класифікацію речення суб'єктивно та об'єктивно та бінарну класифікацію на рівня документа. У статті [30] автори представили 3 архітектури обміну інформацією для моделювання текстової послідовності. Перша архітектура використовує загальний шар для всіх завдань. Друга - використовує різні шари для різних завдань. Остання модель передбачає присвоєння певного завдання до першого рівня, але також має загальний шар для всіх завдань. Після проведення експериментів автори порівнювали отримані результати і дійшли висновку, що на певному завданні вони досягли кращих результатів, протилежних найсучаснішим вихідним рівням.

Дерева рішень є ще одним способом виконання класифікації. Дерево рішень [31] - це класифікатор, який представлений як ієрархічне розбиття простору даних. Структура дерева містить 2 типи вузлів: листовий вузол (містить значення цільового атрибута, тобто позитивна або негативна мітка в задачі бінарної класифікації) та вузол вирішення (містить умову для одного з атрибутів розділу пробілів). Розбиття простору даних здійснюється рекурсивно.

## Висновки до розділу

У цьому розділі наведено короткі пояснення алгоритмів, які можна застосовувати для завдання класифікації. Крім того, розглядаються підходи, які використовувались дослідниками для класифікації тексту. Більш конкретно, розглядаються різноманітні методи лексики та машинного навчання. Показано, що підходи навчання з учителем, показує хороші результати при розгляді класифікації настроїв. Класифікатор Наївного Байєса добре працює над класичним текстом, незважаючи на свою простоту. Крім того, CNN також ефективна для обробки природних мов, це дозволяє значно зменшити кількість параметрів для навчання та отримати високу оцінку класифікації. Виходячи з цих факторів, методи навчання з учителем, такі як Наївний Байєс та CNN, також використовується для аналізу настроїв на твітах.

## РОЗДІЛ 2 РОЗРОБКА МЕТОДІВ АНАЛІЗУ ТОНАЛЬНОСТІ НОВИН

### 2.1 Вступ

На сьогоднішній день ми не можемо уявити собі наше життя без доступу до Всесвітньої павутини, кожен використовує Інтернет для різних цілей, тобто для пошуку деяких відомостей або публікації чогось. Інформація може бути легко опублікована користувачами в блогах, форумах, соціальних мережах, зворотній зв'язок може бути залишений на певних веб-сторінках. Є багато сайтів, які пропонують огляди продуктів. Наприклад, Amazon - електронний магазин, де клієнти можуть опублікувати свої відгуки про продукти, а також шукати відгуки, щоб прийняти рішення щодо придбання продукту. Іншим цікавим і корисним джерелом думок є TripAdvisor. TripAdvisor - це веб-сайт, який надає десятки виважених відомостей про готелі, ресторани, рейси, що дуже корисно для мандрівників. Twitter - це ще один спосіб обміну думками. Інформація з таких джерел використовується не тільки клієнтами, але також життєво важливо для різних організацій. Велика кількість наявних даних свідчить про необхідність створення автоматизованої системи для пошуку та класифікації думок. Задача класифікації тексту - це не нова область вивчення; проте, в основному проводилися дослідження на коротких текстах та оглядів фільмах. Стосовно Twitter, повідомлення відрізняються від відгуків по їх довжині (140 символів) та спеціальними символами, наприклад, @, # або RT. Крім того, спосіб спілкування неформальний, що призводить до використання сленгу та ідіом, помилок в написанні.

## 2.2 Розробка інформаційної системи

Була розроблена інформаційна система аналізу тональності новин, яка складається з наступних частин (рис. 2.1):

- попередня обробка даних;
- застосування методів Наївного Байєса та згорткових нейронних мереж;
- порівняння та аналіз результатів.

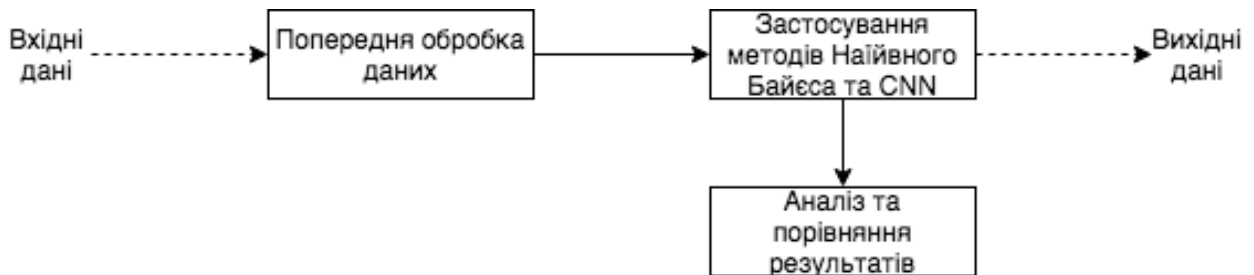


Рисунок 2.1 – Інформаційна система

Етап попередньої обробки описаний в розділі 2.3, застосування методів описуються в розділі 2.4, аналізу та порівнянню результатів присвячений розділ 3.

## 2.3 Збір та обробка даних

У даній дисертації було використано декілька наборів даних:

а) заголовки новин з Австралійського джерела новин ABC (Australian Broadcasting Corp.) за 15 років. Кількість строк у файлі - 1103665. Це файл формату csv (comma separated value), має 2 колонки - дата публікації (у форматі ууууММдд - рік-місяць-день) та текст заголовку новини. Набір даних був завантажений з сайту: <https://www.kaggle.com/hiteshp/what-is-with-news-headlines/data>;

б) набір даних з Твіттеру. Немає однієї єдиної тематики, яка б об'єднувала усі ці твіти. Але великим бонусом цього набору є те, що кожний твіт помічений у відповідності до категорії:

- empty (не має категорії/пустий твіт, 827 рядків);

- sadness (сум, 5165 рядків);
- enthusiasm (ентузіазм, 759 рядків);
- neutral (нейтральний твіт, 8638 рядків);
- worry (переживання, 8459 рядків);
- surprise (сюрприз, 2187 рядків);
- love (любов, 3842 рядків);
- fun (веселощі, 1776 рядків);
- hate (ненависть, 1323 рядків);
- happiness (щастя, 5209 рядків);
- boredom (нудьга, 179 рядків);
- relief (полегшення, 1529 рядків);
- anger (злість, 110 рядків).

Не кожна категорія з переліку вищу підходить для обробки тональності. Для використання даного набору даних в оцінці тональності тексту на 2 класи (позитивний, негативний) було зроблено припущення, що категорії, такі як «love», «fun», «happiness» відносяться до позитивного класу, а «hate», «anger», «worry» до негативного. Загалом, кількість позитивно помічених твітів склала майже 11000, негативно помічених - майже 10000.

Окрім колонки з категорією тональності цей набір має 3 колонки: `tweet_id` (унікальний номер твіту), `author` (користувач Twitter, який створив цей твіт), `content` (текст самого твіту). Цей файл також має формат csv та містить 40000 рядків.

Обробка даних з платформи Twitter є критичною в рамках часу роботи моделей та точності класифікаторів, оскільки зашумлені дані можуть сповільнювати процес навчання та зменшувати точність систему.

Наступні кроки повинні бути виконані для обробки твітів:

- видалення веб-посилань, часто твіти містять веб-посилання, щоб ділитися деякою додатковою інформацією. Зміст посилань не аналізується, тому адреса сама по собі не надає ніякої корисної інформації, і її усунення може зменшити розмір об'єкта;



- видалення імен користувачів, інший користувач може бути згаданий автором повідомлення у твіті, використовуючи символ «@», а потім ім'я користувача. Ці дані також не несуть жодної релевантної інформації, тому були видалені з твіту;

- видалення хеш-тегів. Хеш-тег зображується за допомогою символу «#» і використовується перед словом, що представляє назву теми. У даній роботі не стоїть завдання класифікувати теми твітів, тому вони також видаляються;

- видалення ретвітів і дублікатів, retweet - твіт, який написаний одним користувачем, а потім копіюється та публікується іншим користувачем. Retweet містить аббревіатуру «RT». Такі дані видаляються, оскільки можуть додати зайві ваги на якісь слова;

- стиснення «витягнутих» слів. Зазвичай слова подовжують, щоб виразити почуття, наприклад «Ні, тааааааааааа», радість від того, що зустрів приятеля. Від користувача залежить, скільки букв він буде повторювати, зазвичай, це не є фіксованим числом. Тобто, в одному твіті буде написано «ahh», а в іншому - «ahhh». Для спрощення можна вважати, що це одне й те саме слово і нормалізувати їх обох до «ahh»;

- видалення стоп-слів. Стоп-слова є надзвичайно частими частинами речень, які вважаються марними для сприйняття їх у якості ознак, тобто «the», «for», «her», «a» тощо;

- приведення усіх слів до нижнього регістру, це є необхідно тому, що слова «sad» та «SaD» є однаковиими, також цей крок гарантує узгодженість у наборі ознак.

Посилання на набір даних - <https://www.figure-eight.com/data-for-everyone/>, розділ - «Sentiment Analysis: Emotion in Text»;

в) фрагменти новин зі статей, які були оброблені експертами та помічені, чи були актуальні вони актуальні для економіки США, і якщо так, то яка тональність у цієї статті. Тональність була оцінена за 9-бальною шкалою (від 1 до 9, де 1 - найбільш негативна). Файл формату csv, містить наступні колонки: `_unit_id`, `_golden`, `_unit_state`, `_trusted_judgments`, `_last_judgment_at`, `positivy`, `positivity:confidence`, `relevance`, `relevance:confidence`, `articleid`, `date`, `headline`,

positivity\_gold, relevance\_gold, **text**. Жирним виділені ті колонки, які будуть використовуватися в аналізі: positivity, headline, text. Колонка positivity містить наступні значення експертів: 2, 3, 4, 5, 6, 7, 8, 9, NaN. Для тренування та тестування текстом з позитивним окрасом буде вважатися той, у якого positivity більший або рівний 5, відповідно, негативний - менше 5. Ті рядки, для яких значення колонки positivity рівне NaN не будуть аналізуватися. В підсумку, остаточно кількість рядків для аналізу рівна 1420. Посилання на набір даних - <https://www.figure-eight.com/data-for-everyone/>, розділ - «Economic News Article Tone and Relevance»;

г) експерти переглянули заголовки новин статті та короткий виділений фрагмент виразу або двох із супровідної статті. Далі вони вирішили, чи впливає вирок на стан здоров'я економіки США, потім оцінювали показник у масштабі 1-9, де 1 - негативний та 9 - позитивний. Файл формату csv, містить наступні колонки: \_unit\_id, \_golden, \_unit\_state, \_trusted\_judgments, \_last\_judgment\_at, **positivity**, positivity:confidence, relevance, relevance:confidence, orig\_golden, articleid, date, **headline**, lineid, **next\_sentence**, positivity\_gold, **previous\_sentence**, relevance\_gold, **text**. Жирним виділені ті колонки, які будуть використовуватися в аналізі: positivity, headline, next\_sentence, previous\_sentence, text. Колонка positivity містить усі можливі значення, від 1 до 9 та NaN. Для тренування та тестування текстом з позитивним окрасом буде вважатися той, у якого positivity більший або рівний 5, відповідно, негативний - менше 5. Ті рядки, для яких значення колонки positivity рівне NaN не будуть аналізуватися. В підсумку, остаточно кількість рядків для аналізу рівна 2899. Посилання на набір даних - <https://www.figure-eight.com/data-for-everyone/>, розділ - «U.S. economic performance based on news articles».

## 2.4 Розробка методів класифікації тексту

### 2.4.1 Вилучення ознак з тексту

Після того, як обробка даних була завершена, потрібно витягнути слова-ознаки та використати їх для тренування класифікаторів. У даній роботі було застосовано декілька методів, які дозволяють отримувати слова-ознаки з тексту.

Був досліджений алгоритм вибору ознак на основі моделі Хі-квадрат [2], [32] для моделі Наївного Байєса.  $\chi^2$  - це статистична перевірка, яка вимірює незалежність між класом та ознакою. Вона оцінює значення кореляції між класом та ознакою.  $\chi^2$  можна розрахувати, використовуючи наступну формулу [32]:

$$\chi^2(D, t, c) = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} \frac{(N_{e_t e_c} - E_{e_t e_c})^2}{E_{e_t e_c}} \quad (2.1)$$

де  $D$  - це тренувальний набір даних,

$N$  - спостережувана частота, а  $E$  - очікувана частота,

$e_t = 1$  якщо документ містить слово  $t$ ,

$e_t = 0$  якщо документ не містить слово  $t$ ,

$e_c = 1$  якщо документ знаходиться в класі  $c$ ,

$e_c = 0$  якщо документ не знаходиться в класі  $c$ .

Перепишемо формулу вище:

$$\chi^2(D, t, c) = \frac{(N_{00} + N_{01} + N_{10} + N_{11}) * (N_{11} N_{00} - N_{10} N_{01})^2}{(N_{11} + N_{01}) * (N_{11} + N_{10}) * (N_{10} + N_{00}) * (N_{01} + N_{00})} \quad (2.2)$$

де  $N$  - це кількість даних у тренувальному наборі, де  $N$  розраховується за формулою:  $N = N_{00} + N_{01} + N_{10} + N_{11}$ ,

$N_{00}$  - кількість речень, які не містять ознаку ( $e_t = 0$ ) та не в класі ( $e_c = 0$ ),

$N_{11}$  - кількість речень, які мають співпадіння ознаки та класу,

$N_{10}$  - кількість речень, які містять ознаку, але знаходяться не в класі,

$N_{01}$  - кількість речень, які знаходяться в класі, але не містять ознаку.

Велике значення  $\chi^2$  значить, що 2 події не є незалежними (у нашому випадку, ознака та клас є залежними), це значить, що нуль-гіпотеза про незалежність

повинна бути відхилена. Якщо події незалежні - ознака має бути врахована. Однак, значення результату  $\chi^2$  дозволяє вибрати найбільш інформативну ознаку для тренування класифікатора.

Другий експеримент проводився за допомогою згорткової нейронної мережі, CNN використовує фільтри (ядра), які відіграють роль детекторів ознак. Використовуючи початковий набір даних, необхідно сформувати словниковий запас, де кожне слово індексується (індекс - це ціле число, яке знаходиться в діапазоні від 0 до розміру словника). Після створення словника, перші шари CNN представлені як малорозмірні вектори. А саме, кожен документ обробляється як послідовність слів  $s = [s_1, s_2, \dots, s_n]$ , де кожне слово має свій індекс, який вказує на позицію цього слова в словнику  $v$ . Речення різної довжини були пронормовані, підбиванням їх до максимальної довжини речення. В цілому, кожне речення перетворюється у вектор, і весь вхідний текст представляється як матриця.

Для вибору інформативних ознак з початкового набору даних і переходу до вищого рівня, необхідно використовувати операції згортки та об'єднання. Вони будуть більш детально описані у наступній главі.

## 2.4.2 Алгоритми класифікації

### 2.4.2.1 Алгоритм Наївного Байєса

Підхід з використанням класифікатора Наївного Байєса показав свою ефективність та простоту у класифікації тональності [10], [26], [27]. Це ймовірнісний підхід з використанням алгоритма Байєса, який дозволяє обраховувати ймовірність того, що ознака належить до якогось класу:

$$P(\text{label}|\text{features}) = \frac{P(\text{label}) * P(\text{features}|\text{label})}{P(\text{features})} \quad (2.3)$$

де  $P(label|features)$  це апостеріорна ймовірність того, що ознака належить до класу (позитивного чи негативного),

$P(label)$  - це ймовірність обраного класу,

$P(features|label)$  - це умовна ймовірність того, що конкретна ознака з'явиться у даному класі,

$P(features)$  - це ймовірність обраної ознаки.

Зробимо «наївне» припущення про те, що ознаки є незалежними одна від одної. Це дає можливість написати наступне:

$$\begin{aligned} P(features|label) &\approx P(f_1|label) * P(f_2|label) * \dots * P(f_n|label) = \\ &= \prod_{i=1}^n P(f_i|label) \end{aligned} \quad (2.4)$$

$$P(label|features) = \frac{P(label) * \prod_{i=1}^n P(f_i|label)}{P(features)} \quad (2.5)$$

де  $f_i$  - це конкретна ознака.

Не дивлячись на те, що модель Наївного Байєса робить припущення про незалежність відносно позицій слів, вона показує досить непогані результати на реальних даних.

Основна ціль класифікатора це визначити клас, до якого належить ознака. Для цього, нам не цікаво знаходження самої ймовірності, але найбільш ймовірний клас повинен бути визначений. Класифікатор Наївного Байєса використовує оцінку максимальної апостеріорної ймовірності для визначення найбільш відповідного класу  $label_{map}$  [32]:

$$label_{map} = \underset{label \in L}{\operatorname{argmax}} \left[ \frac{\hat{P}(label) * \prod_{i=1}^n \hat{P}(f_i | label)}{\hat{P}(features)} \right] \quad (2.6)$$

Знаменник можна не включати, тому що для позитивного класу він дорівнює тому самому, що й для негативного. Отже, вираз для обчислення  $label_{map}$  можна переписати як:

$$label_{map} = \underset{label \in L}{\operatorname{argmax}} [\hat{P}(label) * \prod_{i=1}^n \hat{P}(f_i | label)] \quad (2.7)$$

$P$  помічено як  $\hat{P}$ , тому що істинні значення відповідних параметрів будуть оцінюватися з навчального набору даних [32].

Багато умовних ймовірностей множаться у рівняння вище. На підставі [32] останній вираз може призвести до проблеми зникнення порядку, яку можна уникнути, якщо використовувати логарифмічну властивість, а саме  $\log(xy) = \log(x) + \log(y)$ . Тепер множення ймовірностей представлено як сума логарифмів, тому рівняння вище можна переписати наступним чином:

$$label_{map} = \underset{label \in L}{\operatorname{argmax}} [\log \hat{P}(label) + \sum_{i=1}^n \log \hat{P}(f_i | label)] \quad (2.8)$$

Як було згадано, підрахунок  $P(label)$  та  $P(f_i | label)$  проводиться на тренувальному наборі даних. Ймовірність  $\hat{P}(label)$  може бути обчислена як:

$$\hat{P} = \frac{N_{label}}{N} \quad (2.9)$$

де  $N_{label}$  - кількість ознак, яка належить до конкретного класу, та  $N$  - загальна кількість ознак у тренувальному наборі даних.

Умовна ймовірність  $\hat{P}(f_i|label)$  обчислюється:

$$\hat{P}(f_i|label) = \frac{F_{ilabel}}{\sum_{i' \in V} F_{i' label}} \quad (2.10)$$

де  $F_{ilabel}$  - кількість разів, скільки  $i$ -та ознака зустрілася у тренувальному наборі даних у конкретному класі, включаючи повторення ознак, та  $V$  - це словник усіх унікальних ознак у конкретному класі.

Важливо зауважити, що на етапі класифікації може статися так, що класифікатор зустрів нову ознаку, якої не було у тренувальній вибірці, отже, є невідомою для класифікатора. У цьому випадку, класифікатор пропустить цю ознаку. Тим самим, остаточна формула для визначення кращого класу, використовуючи Наївний Байесовий класифікатор може бути переписана таким чином:

$$label_{map} = argmax_{label \in L} [\log \frac{N_{label}}{N} + \sum_{i=1}^n \log \frac{F_{ilabel}}{\sum_{i' \in V} F_{i' label}}] \quad (2.11)$$

#### 2.4.2.2 Згорткова нейронна мережа

Припускаємо, що перетворення вхідного тексту було виконано і тепер він представлений у вигляді високорозмірного вектора. Наступним кроком є застосування згорткових, нелінійних операцій та максимізаційного агрегування для вилучення ознак із вхідних даних. Нижче, можна знайти пояснення, як відбувається вилучення на різних рівнях CNN.

Згорткові шари. Згортковий шар дозволяє отримати шаблони, які часто використовуються в даних [5] (він знаходить області, які мають вирішальне значення для вибору ознак) (рис 2.1). Більш конкретно, для того, щоб отримати нову ознаку, необхідно застосувати операцію згортки. З цією метою,  $n$  слів з речення згортаються з ваговими фільтрами  $w$  для отримання карти ознак. Розмір

фільтра відповідає кількості слів, які перекриваються (при цьому робочий розмір фільтра - 3, 4, 5). Ваги фільтра ініціалізуються спочатку випадково, а потім регулюються під час навчання. Ця функція може бути математично представлена як [5]:

$$c_i = f(w * s_{i:i+n-1} + b) \quad (2.12)$$

де  $w$  - це вектор вагів,  $s_{i:i+n-1}$  - ковзне вікно,  $b \in R$  - вектор упередження,  $f$  - нелінійна функція.

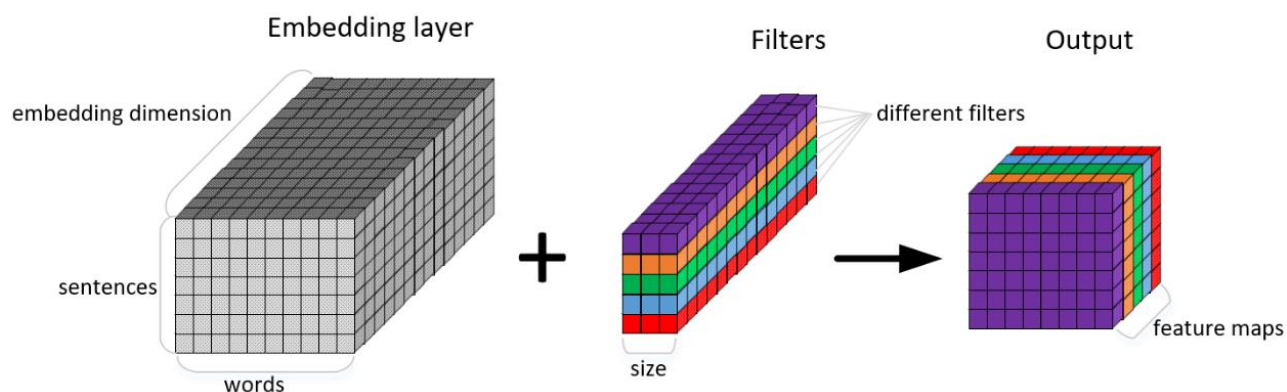


Рисунок 2.1 – Операція згортки

Фільтр застосовується до кожної послідовності слів у реченні, що відповідає розміру фільтра  $\{s_1:n, s_2:n+1, \dots, s_{s-n+1}:s\}$  для створення карти ознак [5]:

$$c(w) = [c_1, c_2, \dots, c_{s-n+1}] \quad (2.13)$$

ReLU (Rectified Linear Unit) - береться у якості нелінійної функції та дуже інтенсивно використовується дослідниками [35], [5], [33], і застосовується після шару згортки. Всі негативні значення на карті ознак перетворені на 0, щоб гарантувати, що карти ознак є позитивними (рис. 2.2) [5]. Визначається як:

$$f(x) = \max(0, x) \quad (2.14)$$



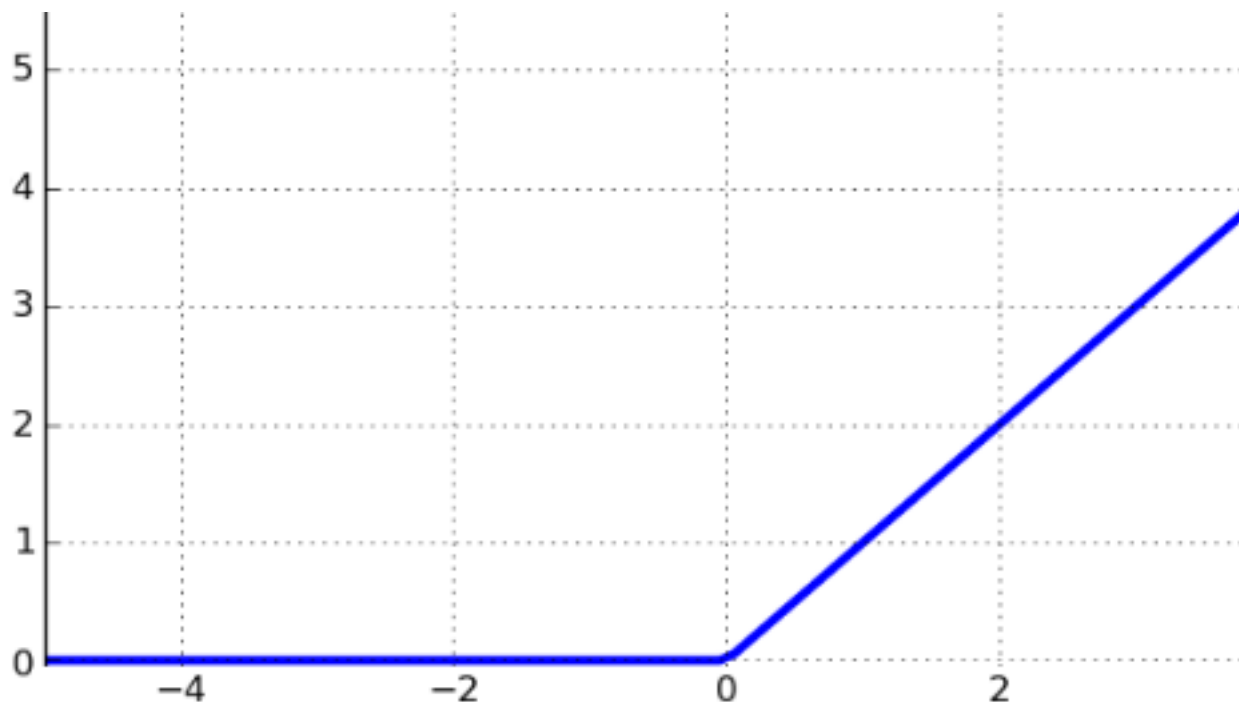


Рисунок 2.2 – Функція активації ReLU

Агрегувальні шари. Після того, як ReLU був використаний на шарі згортки, він генерує вхідну інформацію для агрегувального шару. У даній дисертації була застосована операція максимізаційного агрегування, яка дозволила зменшити розмір карти ознак та одночасно зберегти найбільш релевантні ознаки:

$$\hat{c} = \max(c(w)) \quad (2.15)$$

У цій дисертації було використано 128 фільтрів, що породжувало 128 карт ознак. Після максимізаційного агрегування виходи передаються повністю підключеному шару, де вони об'єднуються в один вектор ознак. Використовуючи останній рівень softmax, виводиться розподіл ймовірності по двох класах (позитивному чи негативному) [5]:

$$p(y = j|x) = \frac{e^{x^T w_j + b_j}}{\sum_{k=1}^K e^{x^T w_k + b_k}} \quad (2.16)$$

де  $x$  - вихід передостанніх згорток та агрегування, представлених у вигляді щільного вектора,  $w_k$  - вектор вагів  $k$ -ого класу, та  $b_k$  - зміщення  $k$ -ого класу.

Функція відключення - це спосіб запобігання переповнення мережі. Навчання, як правило, виконується за допомогою стохастичного градієнтного спуску шляхом випадкового вибору деяких зразків з набору даних. Відключення передбачає, що лише на етапі навчання вилучається деяка частина нейронів (показник відсіву встановлений до 0.5), що запобігає коадаптації нейронів і призводить до навчання більш надійних ознак та робить модель більш узагальнюючою. Вихід після застосування відсіву відображається у вигляді:

$$y = w(zr) + b \quad (2.17)$$

де передостанній шар  $z = [\hat{c}_1, \dots, \hat{c}_m]$ ,  $r$  - вектор, який містить 0 та 1.

У загальному випадку, відсів пришвидшує процес навчання.

Модель CNN навчається, щоб мінімізувати функцію перехресної ентропії, яка може бути записана наступним чином:

$$H(p, q) = -\sum_x p(x) \log q(x) \quad (2.18)$$

де  $p(x)$  - ймовірність істинної ймовірності (правильної відповіді),  $q(x)$  - оцінена ймовірність.

Тренування CNN передбачає коригування параметрів мережі. Цей процес налаштування називається методом зворотного поширення помилки. Для розрахунку градієнта функції помилки відносно ваги фільтра застосовується зворотне поширення. Алгоритм Адама [34] - це стохастичний градієнтний алгоритм спуску, використовується для оптимізації параметрів CNN (оновлення ваг).

Модель згорткової нейронної мережі зображена на рисунку 2.3. Вихідні розміри, вироблені після кожного шару, наведені на малюнку нижче, де партія відповідає розміру пакету і дорівнює 64. Параметр «len» відповідає максимальній довжині послідовності в наборі даних; затемнення позначає розмірність і становить 128; параметр «filter\_size» становить 3, 4, 5 відповідно (зображені 3-ома кольорами); параметр «num\_filters» - 128 та відповідає кількості фільтрів. Розмір кроку дорівнює 1 (зміна фільтра на крок).

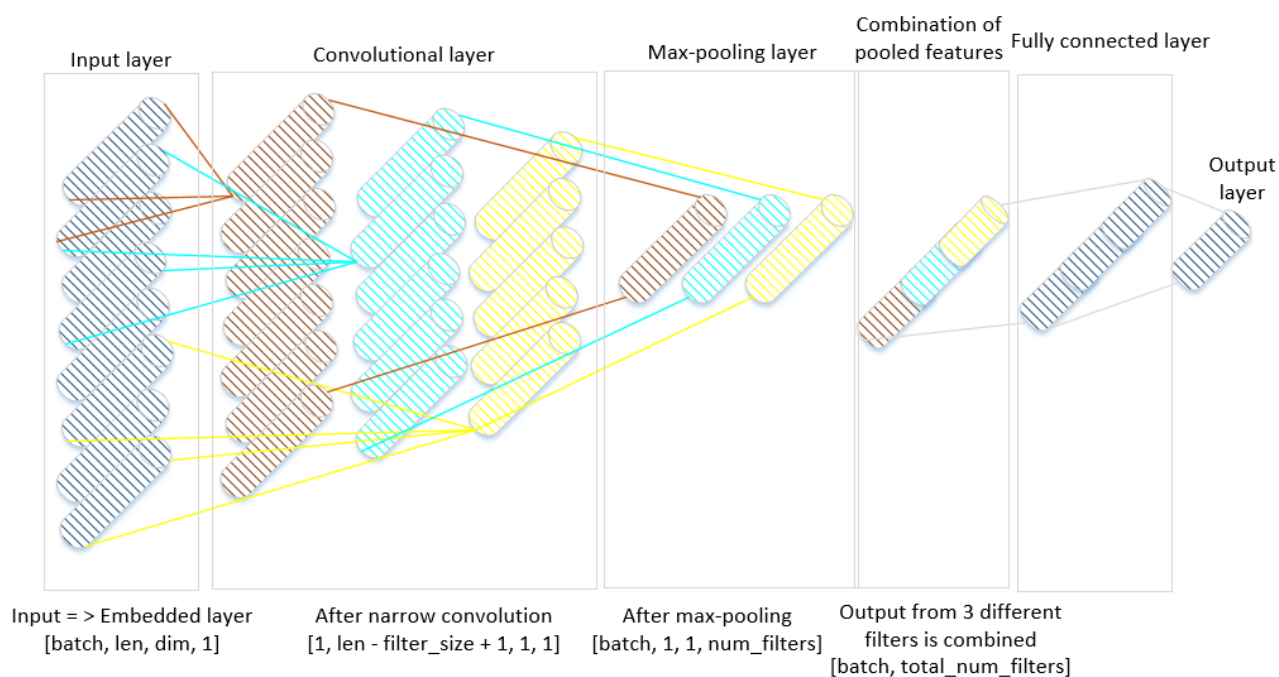


Рисунок 2.3 – Архітектура CNN

### Висновки до розділу

Підхід Наївного Байєса досить часто використовується дослідниками завдяки своїй простоті та високій продуктивності, незважаючи на припущення про незалежність ознак. Крім того, в останні десятиліття нейронними мережами стали цікавитися високопоставлені вчені, особливо після того, як Krizhevsky та інші [36] представили результати, отримані ними для завдання розпізнавання зображень та

відео. Подальші дослідження показали, що CNN також застосовується до завдань природньої мови та може ефективно класифікувати текст [5], [37].

Основна частина цього розділу ілюструє деталі алгоритмів, що використовуються для класифікації тексту, а саме Наївний Байєс та згортова нейронна мережа. Крім того, важливість попередньої обробки описана в поточному розділі, підход до вибору ознак також наведений у цьому розділі.

## РОЗДІЛ 3 АНАЛІЗ РЕЗУЛЬТАТІВ

### 3.1 Оцінка показників ефективності алгоритмів

Ефективність алгоритмів класифікації зазвичай оцінюється на основі таких показників, як чутливість (precision), специфічність (recall), оцінка  $F_1$  та точність (accuracy). Більш того, дуже важливо врахувати обчислювальні ресурси, необхідні алгоритму для побудови класифікатора та його використання.

Розглянемо показники, які були використані для розрахунку чутливості, специфічності, оцінки  $F_1$  та точності. Матриця помилок містить розрахунковий і фактичний розподіл класів (табл. 3.1). Кожний стовпчик відповідає реальному класу, і кожен рядок відповідає приблизному класу речення.

Таблиця 3.1 – Матриця помилок для бінарного класифікатора

Клас, отриманий результат класифікатора	Реальний клас	
	+	-
+	TP	FP
-	FN	TN

TP - кількість істинно позитивних: речення є дійсно позитивним та було оцінено як позитивне,

TN - кількість істинно негативних: речення є дійсно негативним та було оцінено як негативне,

FP - кількість хибно позитивних: речення є дійсно негативне та було оцінено як позитивне,

FN - кількість хибно негативних: речення є дійсно позитивне та було оцінено як негативне.

Точність дає розуміння пропорції вірних відповідей, які були дані класифікатором, оцінюється за наступною формулою:

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3.1)$$

Чутливість може бути обрахована за наступною формулою:

$$precision = \frac{TP}{TP+FP} \quad (3.2)$$

Чутливість показує, наскільки багато позитивних відповідей від класифікатора є вірними. Чим більша чутливість, тим менше кількість хибних класифікацій. Проте, чутливість не показує, чи усі вірні відповіді були отримані класифікатором. Для того, щоб це зрозуміти, треба брати до уваги ще один показник - специфічність, який обраховується за формулою:

$$recall = \frac{TP}{TP+FN} \quad (3.3)$$

Специфічність показує можливість класифікатора вгадати якомога більше правильних відповідей.

Чим більше чутливість та специфічність, тим краще. Проте, одночасне досягання високих результатів з чутливості та специфічності майже неможливо у реальних задачах, тому був винайдений ще один показник,  $F_1$  - гармонічне середнє між чутливістю та специфічністю:

$$F_1 = \frac{2*precision*recall}{precision+recall} \quad (3.4)$$

### 3.2 Класифікатор Наївного Байєса

Класифікатор Наївного Байєса був навчений та протестований на наборі даних з Twitter, який містить твіти на загальні теми. Весь експеримент проводився

з використанням різної кількості слів для тренування класифікатора, а саме  $n$  слів, які мають найвищий бал  $b$ , були подані в класифікатор. Ця оцінка була розрахована з використанням тесту  $\chi^2$ , для цього було знайдено частотний розподіл всіх слів у наборі даних, а умовна частота визначається, щоб підрахувати, скільки разів слово з'явилося в позитивному реченні, і скільки разів в негативному. Результати експерименту можна побачити на рисунку нижче (рис. 3.1).

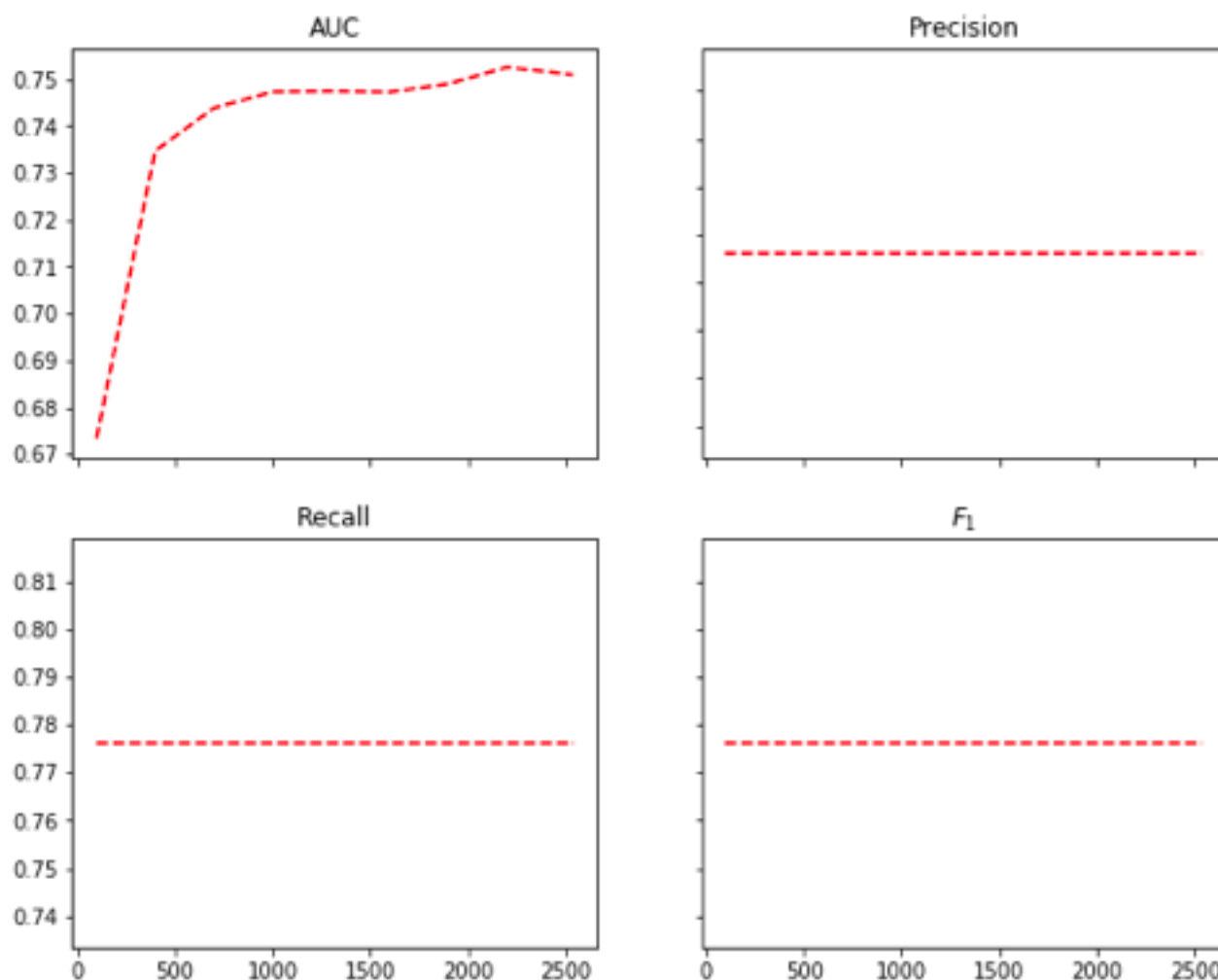


Рисунок 3.1 – Оцінки метрик класифікатора Наївного Байєса на даних з Twitter, по осі  $x$  - кількість слів за даними  $\chi^2$  тесту

Як можна побачити з графіків, найбільше значення метрики AUC досягає при кількості слів в 2200 (саме значення дорівнює 0.7525). Дивними можуть здатися графіки чутливості, специфічності та оцінки  $F_1$ . Але, насправді, немає нічого дивного, адже, наприклад, чутливість розраховується як середнє значення між

результатом чутливості, отриманим для негативного та позитивного класу. Тобто, якщо для кількості слів в 400 маємо значення чутливості для негативного класу - 0,79, для позитивного - 0,71, тобто отримуємо середнє значення в 0,75. А для кількості слів, наприклад, 1600, маємо значення чутливості для негативного класу - 0,77, для позитивного - 0,73, що у середньому дає таке ж значення в 0,75.

Висновки до розділу



## ВИСНОВКИ ПО РОБОТІ ТА ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

Створена інформаційна система аналізу тональності текстів, яка включає в себе етапи збору даних, попередньої їх обробки, використання методів Наївного Байєсу та згорткової нейронної мережі, аналіз та порівняння отриманих результатів. Аналізуючи проведену роботу та дослідження, можна зробити висновок, що поставлені цілі були успішно досягнуті. У дисертації досліджуються алгоритми Наївного Байєса та згорткових нейронних мереж для класифікації почуттів новин. Згідно з оглядом літератури, було встановлено, що більшість підходів до аналізу настроїв покладаються на методи машинного навчання з учителем. Тому було вирішено досліджувати саме підхід Наївного Байєса та згорткові нейронні мережі, оскільки ці методи є у тренді в дослідників, і вони дають суттєві результати. Отже, було проведено аналіз обох алгоритмів та оцінено їх ефективність. Класифікаційна модель була підготовлена на наборі даних з фрагментами новин, які потенційно впливають на економіку США, та звичайних твітах на загальну тематику. Це було зроблено, щоб зрозуміти, чи класифікації настроїв залежить від тематики текстів чи ні. Крім того, детально обговорюється значення етапу попередньої обробки даних.

Майбутня робота передбачає вивчення інших підходів до попередньої обробки текстів, оскільки вони повинні бути більш ретельно відфільтровані для досягнення більшої точності. Існує декілька сценаріїв, якими можна піти у наступних роботах:

- корегування орфографічних помилок у тексті;
- аналіз даних, які містять емоції, наприклад, «:)» (радість), «:(» (сум) для більш точної оцінки семантичної орієнтації текстів;
- заміна аббревіатур повними словами, адже дуже часто використовується в американському тексті;

- було б цікаво додати нейтральний клас і перевірити ефективність класифікатора, проте в цьому випадку набори даних для навчання та тестування повинні включати нейтральні зразки для подання їх моделі та їх оцінки.