Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук
(повна назва)

Кафедра Програмної інженерії
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

Рівень вищої освіти другий (магістерський)

Дослідження методів аналізу емоційного забарвлення коментарів.
Алгоритми
(тема)

Виконав:
студент (ка) 2 курсу, групи ІПЗм-22-4

Фролов М. В.
(прізвище, ініціали)

Спеціальність 121 – Інженерія програмного забезпечення
(код і повна назва спеціальності)

Тип програми освітньо-наукова

Керівник доц. Валенда Н. А.
(посада, прізвище, ініціали)

Допускається до захисту
Зав. кафедри _____ З.В.Дудар
(підпис) (прізвище, ініціали)

2024 р.

Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук
Кафедра Програмної інженерії
Рівень вищої освіти другий (магістерський)
Спеціальність 121 – Інженерія програмного забезпечення
Тип програми освітньо-наукова програма
Освітня програма Інженерія програмного забезпечення

Курс 2 Группа ІПЗм-22-4 Семестр 2

ЗАТВЕРДЖУЮ:

Зав. кафедри _____

(підпис)

«___» _____ 2024 р.

ЗАВДАННЯ

НА КВАЛІФІКАЦІЙНУ РОБОТУ

Студенту Фролову Максиму Вячеславовичу

1. Тема роботи: Дослідження методів аналізу емоційного забарвлення коментарів. Алгоритми

Затверджена наказом по університету від 29.03 2024р. № 250 Ст.

2. Термін подання студентом роботи до екзаменаційної комісії 04.06.2024

3. Вихідні дані до роботи *алгоритми класифікації текстів, перетренована модель нейронної мережі, методи класифікації тональності тексту, мови програмування Python, технології Streamlit, середовища розробки PyCharm 2024.1.1*

4. Перелік питань, що потрібно опрацювати в роботі: *аналіз предметної галузі, опис проведених теоретичних досліджень, опис проведених експериментальних досліджень*

КАЛЕНДАРНИЙ ПЛАН

Номер	Назва етапів роботи	Термін виконання етапів роботи	Примітка
1	Видача завдання	29.04.2024	виконано
2	Аналіз предметної галузі	01.05.2024	виконано
3	Постановка задачі	02.05.2024	виконано
4	Експериментальні дослідження	02.05 – 20.05.24	виконано
5	Аналіз результатів експериментальних досліджень та розробка рекомендацій	20.05 – 22.05.24	виконано
6	Написання та оформлення статті та тез доповіді	20.05 – 23.05.24	виконано
7	Підготовка пояснювальної записки	01.05 – 26.05.24	виконано
8	Підготовка презентації та доповіді	26.05 – 2.05.24	виконано
9	Нормоконтроль	3.06 – 08.06.24	виконано
10	Рецензування	08.06 – 14.06.24	виконано
11	Занесення диплома в електронний архів	15.06.2024	виконано
12	Попередній захист	15.06.2024	виконано
13	Допуск до захисту у зав. кафедри	18.06.2024	виконано

Дата видачі завдання «29» березня 2024 р.

Студент гр. ІПЗм-22-4 _____ Фролов М. В.

(підпис) (прізвище, ініціали)

Керівник кваліфікаційної роботи _____ доц. Валенда Н. А.

(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ / ABSTRACT

Пояснювальна записка містить: 59 ст., 27 рис., 5 табл., 22 джерел.

МАШИННЕ НАВЧАННЯ, МЕТОДИ АНАЛІЗУ, МОДЕЛЬ, НАБІР ДАНИХ, ТОНАЛЬНІСТЬ ТЕКСТУ, PYTHON, STREAMLIT.

Об’єктом дослідження є емоційне забарвлення коментарів природної мови, які залишили користувачі у Твіттері.

Мета дослідження – ідентифікація найефективніших методів аналізу емоційного відтінку коментарів англійською мовою та реалізація алгоритму.

Метод рішення – методи NLP для визначення емоційного забарвлення, алгоритм логістичної регресії який використовуються для навчання моделі позначеним намірам і сутностям.

В результаті роботи було сформовано датасет англійською мовою, порівнянною різні моделі та алгоритми тональності тексту, створено програму, яка може отримати емоційне забарвлення тексту користувача. Додаток реалізовано з використанням бібліотеки обробки природної мови (NLP) і логістичної регресії, щоб витягувати наміри та сутності з введених користувачем даних. Програма створена з використанням Streamlit, бібліотеки Python для створення інтерактивних веб-додатків.

ANALYSIS METHODS, DATASET, MACHINE LEARNING, MODEL, STREAMLIT, TENTATION OF TEXT.

The object of the study is the emotional coloring of natural language comments.

The purpose of the research is to identify the most effective methods of analyzing the emotional tone of comments in English and to implement the algorithm.

The solution method is NLP methods for determining emotional coloring, a logistic regression algorithm that is used to teach a chatbot to marked intentions and entities.

As a result of the work, a dataset in English was formed, a chatbot was created that can understand and respond to user input based on intentions. The chatbot is built using a natural language processing (NLP) library and logistic regression to extract intent and essence from user input. The chatbot was built using Streamlit, a Python library for building interactive web applications.

Заява щодо самостійного виконання кваліфікаційної роботи та можливості її публікації в електронному архіві відкритого доступу ElArKhNURE.

Я, Фролов Максим Вячеславович, студент(ка) гр. ІПЗм-22-4, здобувач вищої освіти на другому (магістерському) рівні кафедри «Програмна інженерія», заявляю: моя кваліфікаційна робота на тему «Дослідження методів аналізу емоційного забарвлення коментарів. Алгоритми», що буде представлена в екзаменаційну комісію для публічного захисту, виконана самостійно, в ній не містяться елементи плагіату і вона може бути опублікована в електронному архіві відкритого доступу ElArKhNURE. Всі запозичення з друкованих та електронних джерел мають відповідні посилання.
Я ознайомлений(на) з діючим положенням «Про протидію академічному плагіату в ХНУРЕ», згідно з яким виявлення плагіату є підставою для відмови в допуску кваліфікаційної роботи до захисту та застосування дисциплінарних заходів.

ЗМІСТ

ВСТУП

У сучасному інформаційному суспільстві взаємодія з електронними сервісами та платформами найчастіше здійснюється через текстові комунікації, що ставить перед завданням ефективного розуміння та обробки емоційного забарвлення користувацьких коментарів. Зазначена проблематика надихнула дане дослідження на вивчення та розробку методів аналізу емоційного відтінку у текстових коментарях з використанням алгоритмів машинного навчання.

Аналіз емоційного забарвлення тексту – це процес визначення та класифікації емоцій, які виражені у текстовому висловлюванні. Цей аналіз спрямований на розуміння емоційного стану або настрою автора тексту. Такий підхід використовується в області обробки природної мови (Natural Language Processing, NLP) та машинного навчання для автоматизованого визначення сентименту тексту. Цей аналіз використовується у багатьох сферах таких як соціальні мережі та медіа, маркетинг і реклама, клінічна психологія [1], обслуговування клієнтів, оцінка відгуків та коментарів, політологія. Аналіз емоційного забарвлення тексту є потужним інструментом для різних сфер, де важливо розуміти емоції людей і взаємодіяти з ними.

Мета даного проекту полягає в пошуку найкращого алгоритму, який шукає емоційне забарвлення тексту та програмна реалізація цього алгоритму, який здатний розуміти та реагувати на користувацькі запити на основі їхніх намірів. Для досягнення цієї мети використовуються бібліотека обробки природної мови (Natural Language Processing, NLP) та алгоритм логістичної регресії. Процес тренування моделі здійснюється на позначених наборах намірів та сутностей, а сам інтерфейс програми реалізований за допомогою веб-фреймворку Streamlit, розробленого на мові програмування Python для створення інтерактивних веб-застосунків.

1 АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ

1.1 Аналіз предметної області

Аналіз тональності тексту – це процес визначення емоційного тону або відчуттів, які виражені в текстовому матеріалі. Цей аналіз спрямований на визначення того, чи текст має позитивний, негативний чи нейтральний характер. Він часто використовується в комп'ютерній лінгвістиці та обробці природної мови для розуміння емоційного стану або відгуку автора тексту.

1.1.1 Види класифікації

В сучасних системах автоматичного визначення емоційної оцінки тексту найчастіше використовується одномірний емотивний простір: позитив чи негатив (добре або погано). Однак відомі успішні випадки використання і багатовимірних просторів.

Основним завданням в аналізі тональності є класифікація полярності документа, тобто визначення, чи є виражена думка в документі або реченні позитивною, негативною або нейтральною. Більш розгорнуто, «поза полярності» класифікація тональності виражається, наприклад, такими емоційними станами, як «злий», «сумний» і «щасливий».

Класифікація за бінарною шкалою: полярність документа можна визначати за бінарною шкалою. У цьому випадку для визначення полярності документа використовується два класи оцінок: позитивна чи негативна. Одним із недоліків цього підходу є те, що емоційну складову документа не завжди можна однозначно визначити, тобто документ може містити як ознаки позитивної оцінки, так і негативної. Ранні роботи в цій сфері включають праці Терні та Панга, які застосовують різні методи розпізнавання полярності оглядів товару та відгуків про фільми відповідно. Це приклад роботи на рівні документа.

Класифікація за багатосмуговою шкалою: можна класифікувати полярність документа за багатосмуговою шкалою, що було зроблено Пангом і Снайдером (серед інших). Ними було розширене основне завдання класифікації кіновідгуків

від оцінки «позитивний або негативний» в бік прогнозування рейтингу за 3-х або 4-бальною шкалою. Водночас Снайдер провів поглиблений аналіз оглядів ресторанів, пророкуючи рейтинги різних властивостей, таких як їжа та атмосфера (за 5-бальною шкалою).

Системи шкалювання: іншим методом визначення тональності є використання систем шкалювання, за допомогою чого словами, зазвичай пов'язаними з негативними, нейтральними або позитивними тональностями, ставляться відповідно числа за шкалою від – 10 до 10 (від негативного до самого позитивного). Спочатку фрагмент неструктурованого тексту досліджується з допомогою інструментів та алгоритмів обробки природної мови, а потім виділені з цього тексту об'єкти та терміни аналізуються з метою розуміння значення цих слів.

Суб'єктивність/об'єктивність: інший дослідницький напрямок – це ідентифікація суб'єктивності/об'єктивність.

Це завдання зазвичай визначається як віднесення тексту в один із двох класів – суб'єктивний або об'єктивний. Ця проблема іноді може бути, ніж класифікація полярності: суб'єктивність слів і фраз може залежати від контексту, а об'єктивний документ може містити суб'єктивні речення.

Модель більш докладного аналізу називається аналізом на основі функції/аспекту. Ця модель посилається на ухвалу думок або настроїв, виражених різними функціями або аспектами сутностей, наприклад, у стільникового телефона, цифрової камери або банку. Властивість/аспект – це атрибут або компонент сутності, досліджуваної на тональність, наприклад, екран мобільного телефона або ж якість зйомки камери. Ця проблема вимагає вирішення ряду завдань, наприклад, ідентифікація актуальних сутностей, витяг їхніх функцій, аспектів і визначення, є думка, що висловлена по кожній функції/аспекту, позитивною, негативною або нейтральною.

1.1.2 Методи класифікації тональності тексту

Методи класифікації тональності в текстах використовуються для визначення емоційного забарвлення висловлювань чи текстового матеріалу. Основна мета –

визначити, чи висловлювання має позитивний, негативний чи нейтральний відтінок. Існує кілька методів класифікації тональності тексту, які використовуються в області обробки природної мови та машинного навчання. Деякі з найпоширеніших методів надані у таблиці 1.1.

Таблиця 1.1 – Методи класифікації тональності тексту

Метод	Опис	Переваги	Недоліки
Словниковий метод	Використовує попередньо складені словники з позитивними та негативними словами	Простий у реалізації. Ефективний для коротких текстів	Залежність від точності словників. Може недостатньо ураховувати контекст
Машинне навчання	Використовує алгоритми машинного навчання, такі як класифікація на основі наївного Баєса, SVM, нейронні мережі тощо	Можливість використання для різноманітних типів текстів	Вимагає великої кількості навчальних даних. Витрати часу та ресурсів на навчання
Глибинне навчання	Використовує нейронні мережі глибинного навчання, такі як RNN, LSTM, BERT	Здатність розуміти складніше значення слів та контексту. Висока точність	Вимагає потужних обчислювальних ресурсів. Потребує великої кількості даних для навчання
Правила та експертні системи	Використовує ручно визначені правила чи експертні системи для оцінки тональності тексту	Здатність враховувати специфічні правила та контекст	Може бути неефективним у випадках складних мовних конструкцій або нестандартних виразів

Кінець таблиці 1.1

Метод	Опис	Переваги	Недоліки
Комбіновані підходи	Використовують комбінацію різних методів для покращення точності класифікації	Здатність поєднувати переваги різних методів	Складність в розробці та оптимізації. Вимагає уважного налаштування параметрів для досягнення оптимальних результатів

Для оцінки тональності тексту на Twitter можна використовувати методи машинного навчання. Ці методи можуть бути особливо ефективними, враховуючи особливості Twitter, такі як короткі повідомлення, велика кількість сленгу, абревіатур та емотиконів. Наприклад, можна навчити модель машинного навчання на велику кількість твітів із відомою тональністю. Потім, використовуючи цю модель, можна визначати загальний тон нових твітів, аналізуючи їхній текст і виявляючи закономірності, характерні для позитивних, нейтральних та негативних повідомлень.

1.1.3 Методи обробки природної мови

Обробка природної мови – це керований штучним інтелектом процес, завдяки якому мова людського введення стає доступною для програмного забезпечення. Найпопулярніші методи, які використовує обробку природної мови (NLP) для отримання даних з тексту надані у таблиці 1.2 [2].

Таблиця 1.2 – Методи обробки природної мови

Назва методу	Опис	Переваги	Недоліки
Sentiment Analysis (Аналіз тональності)	Визначення емоційного тону тексту виражає позитивні, негативні або нейтральні почуття	Допомагає в розумінні відгуків, аналізі громадської думки, та управлінні репутацією	Може бути менш точним при розумінні складних конструкцій та саркастичного мовлення
Named Entity Recognition (Визначення іменованих сутностей)	Виділення та класифікація іменованих сутностей, таких як особи, місця та інші, у тексті	Допомагає в структуруванні інформації та розумінні ключових об'єктів у тексті	Може виявити непоодинокі помилки при визначенні контексту та подвійній інтерпретації
Summarization (Стиснення тексту)	Створення короткого огляду або сжатого варіанту тексту, зберігаючи його основний зміст	Зменшує обсяг інформації, полегшуючи сприйняття та аналіз	Може втратити деяку деталі або контекст при стисненні тексту
Topic Modeling (Моделювання тем)	Визначення головних тем або категорій, що представлені в тексті	Допомагає в структуруванні великих обсягів інформації та розумінні основних напрямків в тексті	Вимагає попередньої обробки та налаштування параметрів для точних результатів

Кінець таблиці 1.2

Назва методу	Опис	Переваги	Недоліки
Text Classification (Класифікація тексту)	Розподіл текстів за певними категоріями або класами на основі їх змісту	Допомагає в підсумовуванні тексту та зберіганні ключових понять	Може враховувати не всі контекстуальні взаємозв'язки
Keyword Extraction (Видобування ключових слів)	Виділення найважливіших слів або фраз у тексті, які найкраще визначають його зміст	Допомагає в підсумовуванні тексту та зберіганні ключових понять	Може враховувати не всі контекстуальні взаємозв'язки

Для визначення емоційного забарвлення коментарів у дослідженні буде використовуватися перший NLP метод, а саме Sentiment Analysis, бо саме він Визначає емоційного тону тексту, тобто виявлення, чи текст виражає позитивні, негативні або нейтральні почуття.

1.2 Аналіз існуючих рішень

Існує багато програмного забезпечення для аналізу емоційного забарвлення тексту, яке може бути використане для визначення емоційної тону в текстових даних. Нижче представлено таблицю 1.3 у якій є кілька програм та бібліотек, які широко використовуються у цій області [3].

Таблиця 1.3 – Порівняльна характеристика програмного забезпечення

Назва програми	Метод аналізу	Мова	Платформа
VADER	Лексичний та синтаксичний аналіз	Англійська	Бібліотека (Python)
TextBlob	Машинне навчання	Англійська	Бібліотека (Python)
IBM Watson Natural Language Understanding API	Глибоке навчання	Декілька мов (включаючи англійську)	Вебсервіс
Google Cloud Natural Language API	Глибоке навчання	Декілька мов (включаючи англійську)	Вебсервіс
Microsoft Azure Text Analytics API	Машинне навчання	Декілька мов (включаючи англійську)	Вебсервіс
NLTK (Natural Language Toolkit)	Лексичний та синтаксичний аналіз	Багато мов	Бібліотека (Python)
spaCy	Статистичні та глибокі моделі	Багато мов	Бібліотека (Python)
SentiWordNet	Емоційний словник	Багато мов	Бібліотека (Python)

Кожна програма має свої переваги та недоліки. Наприклад, готові вебсервіси, такі як Google Cloud та IBM Watson, надають зручний спосіб використання, але можуть бути платними та обмеженими у функціональності для безкоштовних користувачів. Бібліотеки Python, такі як TextBlob та spaCy, дають більше гнучкості, але можуть бути менш точними в порівнянні з глибокими моделями. Дослідження методів аналізу емоційного забарвлення коментарів є актуальною з кількох причин:

- зростання важливості емоційного аналізу: з великим обсягом текстової інформації в інтернеті, зростає значення аналізу емоційного забарвлення коментарів для платформ, бізнесів та громадськості;
- розвиток технологій: швидкі та точні алгоритми для аналізу емоційного тону стають доступнішими завдяки розвитку технологій машинного навчання та обробки природної мови;
- специфічні вимоги платформ: різні платформи можуть мати свої унікальні вимоги до аналізу емоцій, і дослідження методів дозволить підібрати оптимальний підхід для кожної з них;
- практичні застосування: розробка ефективних алгоритмів аналізу емоційного забарвлення може мати практичні застосування в сферах відгуків користувачів, соціальних мереж, аналізу відгуків про товари та послуги.

1.3 Постановка задачі

1.3.1 Мета дослідження

Дослідити та порівняти різні інструменти аналізу емоційного забарвлення текстових коментарів, розробити програмну систему для семантичного аналізу коментарів та їх класифікації на позитивні, негативні та нейтральні емоції.

1.3.2 Завдання дослідження

а) літературний огляд:

- провести аналіз існуючих методів аналізу емоційного забарвлення текстів та їхніх застосувань у відгуках та коментарях;
- визначити ключові аспекти та відмінності між різними алгоритмами.

б) вибір та підготовка даних:

- дані попередньо оброблені і підготовлені для подальшого аналізу.

в) порівняння наявних інструментів для семантичного аналізу:

- реалізувати різні базові перевірки аналізу емоційного забарвлення, використовуючи лексичні, синтаксичні та машинно-навчальні підходи

наявних інструментів.

г) оцінка ефективності:

- порівняти результати роботи за допомогою метрик точності, відтворюваності на тестовому наборі даних.

д) валідація на реальних даних:

- провести тестування на реальних відгуках з відомих веб-платформ та соціальних мереж.

1.3.3 Очікувані результати

Очікується, що дослідження приведе до розробки знайдення найкращого інструменту аналізу емоційного забарвлення коментарів, здатних працювати з англійською мовою та ефективно виявляти емоційний стан авторів. Результати можуть бути використані для покращення систем аналізу відгуків та взаємодії з користувачами в різних сферах.

2 ОПИС ПРОВЕДЕНИХ ТЕОРЕТИЧНИХ ДОСЛІДЖЕНЬ

2.1 Аналіз метрик

За останні роки зростає популярність аналізу емоційного забарвлення тексту як інструменту для розуміння та вимірювання емоційного стану людей у великому обсязі текстових даних. Аналіз емоційного забарвлення тексту знаходить широке застосування в галузі відгуків користувачів, соціальних мереж, моніторингу громадської думки, а також в інших сферах.

Однак, важливо визначити, як оцінити ефективність методів аналізу емоційного забарвлення та визначити їхню точність та застосовність у різних умовах. Для цього використання метрик є необхідним етапом у вивченні та порівнянні різних алгоритмів та моделей.

Існую п’ять метрик аналізу емоційного забарвлення тексту [4].

Перша це точність (Accuracy): точність є мірою для оцінки ефективності класифікаторів. Він визначається як відношення правильно визначених спостережень до загальної кількості спостережень. Чим вища точність, тим кращі результати. Його значення коливається від 0 (найгірший) до 1 (найкращий). Рівняння (2.1) показує обчислення точності.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}, \tag{2.1}$$

де Accuracy – точність;

TP – істинно позитивні рішення;

TN – істинно негативні рішення;

FP – помилково позитивні рішення;

FN – помилково негативні рішення.

Друга це точність класифікації (Precision): це також відоме як позитивне прогностичне значення (PPV) і визначається як співвідношення загальної кількості правильних позитивних прогнозів і всіх позитивних прогнозів. Найкращим значенням специфічності є 1, а найгіршим значенням є 0. Рівняння (2.2) показує обчислення точності класифікації.

$$Precision = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad (2.2)$$

де Precision – точність класифікації;

TP – істинно позитивні рішення;

FP – помилково позитивні рішення.

Третя це повнота (Recall): показує відсоток елементів позитивного класу, які були правильно класифіковані відносно всіх елементів позитивного класу. Рівняння (2.3) показує обчислення повноти.

$$Recall = \frac{\text{TP}}{\text{TP} + FN}, \quad (2.3)$$

де Recall – повнота;

TP – істинно позитивні рішення;

FN – помилково негативні рішення.

Четверта це F1-мера (F1-Score): є гармонічним середнім між точністю (precision) та повнотою (recall), і дозволяє отримати збалансовану оцінку ефективності моделі. Рівняння (2.4) показує обчислення F-міри.

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}, \quad (2.4)$$

де Precision – точність класифікації;

Recall – повнота.

П’ята це вагові коефіцієнти при обчисленні F1-мери: щоб врахувати різну вагу точності і повноти, можна використовувати вагові коефіцієнти при обчисленні F1-мери. Наприклад, якщо нам важливіше уникнути помилкових позитивів (збільшити точність), можна встановити більший ваговий коефіцієнт для точності у формулі F1-мери. Рівняння (2.5) показує обчислення F-міри з вагомими коефіцієнтами.

$$F1_{weighted} = \frac{(1+\beta^2)(Precision * Recall)}{(\beta^2 * Precision + Recall)}, \quad (2.5)$$

де β – це ваговий коефіцієнт, який контролює баланс між точністю і повнотою.

Значення параметра β розташовані в межах від 0 до 1, якщо надається перевага точності, і від 1 вище, якщо надається перевага повноті. При $\beta = 1$ формула стає ідентичною попередній, і ми отримуємо збалансовану F-міру, яку також відомо як F1-міру.

2.2 Дослідження наборів даних

Дослідження наборів даних для аналізу тональності тексту є ключовим етапом у розробці моделей для визначення сентименту в текстах. Виконання аналізу емоційної тональності тексту вимагає відповідної обробки набору даних, який є ключовим компонентом. Так як певні й ті самі слова чи фрази можуть мати різний емоційний відтінок, бажано, щоб певна вибірка була віднесена до певного контексту, наприклад, відгуки клієнтів про придбані товари, коментарі до політичних новин тощо. Існує багато наборів даних для аналізу тональності тексту, і їх вибір може залежати від конкретних потреб та завдань проекту. Нижче наведено деякі популярні датасети, які часто використовуються для аналізу тональності:

- IMDb Reviews Dataset: даний набір містить відгуки про фільми з IMDb, і вони анотовані на позитивні та негативні відгуки. Це дозволяє тренувати моделі на аналіз сентименту в кіноіндустрії;
- Amazon Customer Reviews (Amazon Product Reviews): набір даних від Amazon, який включає відгуки користувачів про різні товари. Містить анотації для класифікації на позитивні та негативні відгуки;
- Twitter Sentiment Analysis Dataset: деякі датасети складаються з твітів з Twitter, анотованих на сентимент. Це може бути корисно для аналізу вражень або реакцій на події в реальному часі;
- Yelp Reviews Dataset: набір даних із відгуками користувачів на платформі Yelp. Анотований на позитивні та негативні відгуки;
- Stanford Sentiment Treebank (SST): це дерев'яний банк анотованих рецензій для фільмів та продуктів. Відгуки відзначені за ступенем сентименту;
- Kaggle Datasets: платформа Kaggle має численні конкурси та набори даних, включаючи ті, які стосуються аналізу тональності тексту. Можна знайти

різні за тематикою та обсягом датасети;

– Multi-Domain Sentiment Dataset (MDS): набір даних, який об'єднує відгуки з кількох доменів, таких як електроніка, книги, одяг і т.д.

Для отримання даних, буде використано Twitter Sentiment Analysis Dataset [5]. Twitter є однією з найпопулярніших платформ соціальних мереж на сьогоднішній день з 330 мільйонами активних користувачів щомісяця. Люди висловлюють свою думку про своє повсякденне життя, різні соціальні/національні/міжнародні проблеми тощо. Вони діляться своїми поглядами в межах 140 символів тексту, а іноді також діляться аудіо/відеофайлами. Публікації називаються твітами, і вони загальнодоступні. Інші люди можуть лайкати публікації, коментувати їх або робити ретвіти. Люди можуть стежити один за одним або дружити один з одним [6].

Для дослідження методів аналізу емоційного забарвлення коментарів буде використано два набора даних у контексті навчання системи штучного інтелекту (ШІ) для розпізнавання емоцій:

– натренований датасет: це набір даних, на якому вже була проведена процедура тренування моделі. Модель вивчила закономірності у цьому наборі даних та налаштована для розпізнавання емоцій на основі зразків з цього тренувального датасету;

– тестовий датасет: це інший набір даних, який передається системі ШІ для тестування її здатності розпізнавання емоцій. Тестування допомагає визначити, наскільки добре модель може застосовувати свої знання, отримані в результаті тренування, до нових.

2.3 Інструменти для аналізу емоційної тональності тексту

Інструменти для аналізу емоційної тональності тексту мають ряд важливих застосувань в сучасному аналізі даних та природної мови. Існує кілька інструментів та бібліотек, які можна використовувати для аналізу емоційної тональності тексту.

Python має потужні бібліотеки для реалізації моделей машинного навчання, а також інструменти для обробки природної мови. Завдяки бібліотекам, таким як

Pandas, Neattext, Scikit-learn, Natural Language Toolkit, можна реалізувати та навчати моделі для аналізу емоційного тону.

Pandas – це бібліотека для обробки та аналізу даних в середовищі Python. Вона надає структури даних, такі як DataFrame, що роблять роботу з табличними даними зручною та ефективною. Pandas також дозволяє виконувати різноманітні операції над даними, включаючи фільтрацію, групування, агрегацію та багато інших.

Neattext – це бібліотека для обробки тексту в Python, призначена для видалення або обробки непотрібної або некорисної інформації в текстових даних. Вона містить інструменти для чистки тексту від зайвих символів, англійських абревіатур, стоп-слів тощо. Neattext може бути корисною при передобробці тексту перед аналізом чи витягуванням інформації.

Scikit-learn – це бібліотека для машинного навчання в Python. Вона містить інструменти для класифікації, регресії, кластеризації та інших задач машинного навчання. Scikit-learn надає імплементації різних алгоритмів, а також інструменти для підготовки та обробки даних. Завдяки простоті використання, вона є популярним вибором для початківців та досвідчених дослідників у галузі машинного навчання.

Natural Language Toolkit (NLTK) – це бібліотека для обробки природної мови (NLP) у мові програмування Python. NLTK надає легкий спосіб використання та реалізації різноманітних завдань, пов'язаних із здійсненням обробки текстів та лінгвістичного аналізу.

Логістична регресія часто використовується для завдань аналізу тональності тексту, де метою є визначення тональності текстового фрагмента (позитивний, негативний чи нейтральний). Використання логістичної регресія включає в себе:

а) підготовка даних

- зібрати або підготувати набір даних, що містить текст та мітки тональності (наприклад, позитивний, негативний);
- виконати передобробку даних, таку як токенізація, видалення стоп-слів, лематизація та інші операції.

б) означення ознак

- перетворити текстові дані на числовий вектор, наприклад, за допомогою методу мішка слів (Bag of Words) або TF-IDF (Term Frequency-Inverse Document Frequency);
- означити ці вектори як ознаки для логістичної регресії.

в) розділення даних

- розділити дані на тренувальний та тестовий набори для оцінки ефективності моделі.

г) навчання логістичної регресії

- використати тренувальний набір для навчання логістичної регресії;
- знайти оптимальні ваги для ознак та функцію активації.

д) оцінка моделі

- використовувати тестовий набір для оцінки ефективності моделі;
- оцінка може включати в себе точність, виклики, точність та інші метрики.

е) використання моделі

- використовувати навчену модель для аналізу тональності нових текстових фрагментів.

Помимо розробки моделей класифікатора з використанням різних алгоритмів, важливо розглянути можливість використання моделей нейронних мереж, які вже були навчені раніше.

2.4 Попередньо навчені моделі

Попередньо навчена модель – це модель машинного навчання, яка була навчена на великому наборі даних перед використанням у конкретній задачі чи додатку. Це ефективний спосіб використовувати знання, набуте моделлю в одному контексті, для вирішення схожих задач в іншому контексті безпосередньо.

Основні переваги попередньо навчених моделей:

- попередньо навчені моделі часто навчаються на великих та репрезентативних наборах даних, тому вони мають широкий спектр знань;

- використання попередньо навчених моделей може значно зменшити час та ресурси, які потрібні для навчання моделі від початку;
- попередньо навчені моделі можуть бути використані в різних областях та завданнях, що полегшує їх використання в різних проектах.

Трансферне навчання – це підхід в машинному навчанні, який використовує знання, набуте в одному завданні чи домені, для поліпшення ефективності моделі у новому завданні чи домені. В основі трансферного навчання лежить ідея, що знання, набуте в одному контексті, може бути корисним або передбачуваним в іншому контексті.

Основні принципи трансферного навчання:

- знання передбачуване відповідністю завдань: якщо в двох завданнях чи доменах існують деякі схожості, то знання, набуте в одному з них, може бути корисним для іншого;
- використання попередньо навчених моделей: попередньо навчені моделі, які вже мають знання з великих наборів даних, можуть слугувати вихідною точкою для подальшого навчання на обмеженому обсязі даних для конкретного завдання;
- адаптація ваг: зазвичай ваги моделі адаптуються для врахування нового набору даних, але деякі частини моделі можуть залишитися незмінними.

Трансферне навчання можна розділити на кілька основних типів:

- від одного завдання до одного завдання: знання передається від одного завдання до іншого завдання в тому ж домені;
- від багатьох завдань до одного завдання: знання, набуте від кількох завдань, використовується для поліпшення ефективності в одному конкретному завданні;
- від одного домену до іншого домену: знання передається від одного домену (наприклад, зображення в природі) до іншого домену (наприклад, зображення в медицині).

Для аналізу емоційної тональності тексту можна використовувати універсальні моделі обробки природної мови, які використовуються в різних

сферах, таких як машинний переклад, системи відповідей на запитання, аналіз настроїв тощо. Є безліч прикладів попередньо навчених моделей, які вдало використовуються у різних галузях:

- BERT (Bidirectional Encoder Representations from Transformers) є моделлю для розуміння природної мови, побудованою на основі трансформаторної архітектури. Вона навчалася на великому корпусі текстів та показує вражаючі результати у завданнях, таких як розпізнавання сутностей, класифікація тексту та інші;
- ImageNet Models: моделі, як ResNet, Inception та EfficientNet, були навчені на великому наборі зображень ImageNet для завдань класифікації зображень. Ці моделі стали стандартом в галузі комп'ютерного зору;
- GPT (Generative Pre-trained Transformer) серія: моделі, такі як GPT-3, вивчають генерацію тексту та розуміння мови. Вони здатні вирішувати широкий спектр завдань, включаючи відповіді на питання та інше;
- YOLO (You Only Look Once): є моделлю для реального часу розпізнавання об'єктів на зображеннях та відео. Вона навчалася на великих наборах даних для розпізнавання об'єктів;
- OpenAI Codex: це модель, яка вивчила програмування на багатьох мовах програмування. Вона може генерувати код для вказаних завдань;
- VGG (Visual Geometry Group) серія: моделі VGG були однією з перших глибоких архітектур для класифікації зображень та використовуються як стандарт у деяких завданнях комп'ютерного зору.

Для аналізу емоційної тональності тексту буде використовувано модель машинного навчання з використанням логістичної регресії.

3 ОПИС ПРОВЕДЕНИХ ЕКСПЕРИМЕНТАЛЬНИХ ДОСЛІДЖЕНЬ

3.1 Формування набору даних

Набір даних (датасет) грає важливу роль у машинному навчанні. У ході дослідження було сформовано датасет коментарів користувачів Твіттера.

Для генерування датасету використовувалися наступні технології:

- мова програмування Python;
- бібліотека pandas використовується для обробки та аналізу даних у Python. Вона дозволяє легко завантажувати і зчитувати дані з різних джерел, обробляти їх (наприклад, видаляти зайві символи або приводити текст до нижнього регістру), аналізувати дані, включаючи групування та агрегування, підготовляти дані для моделей машинного навчання, а також візуалізувати дані у вигляді графіків та діаграм (через інтеграцію з іншими бібліотеками, такими як matplotlib та seaborn);
- бібліотека neattext призначена для полегшення обробки текстових даних. Вона забезпечує інструменти для очищення тексту, видалення зайвих символів, стоп-слів, URL-адрес, емодзі та інших небажаних елементів. Це корисно для підготовки текстових даних перед їх аналізом або використанням у моделях машинного навчання;
- бібліотека contractions використовується для розширення скорочень у тексті англійською мовою. Вона автоматично замінює скорочення, такі як «can't» на «cannot» або «I'm» на «I am», що полегшує подальшу обробку тексту та аналіз. Це особливо корисно при підготовці даних для моделей машинного навчання та обробки природної мови (NLP).

Загалом було отримано 35528 коментарів, які мають наступні емоції (див. рис 3.1).

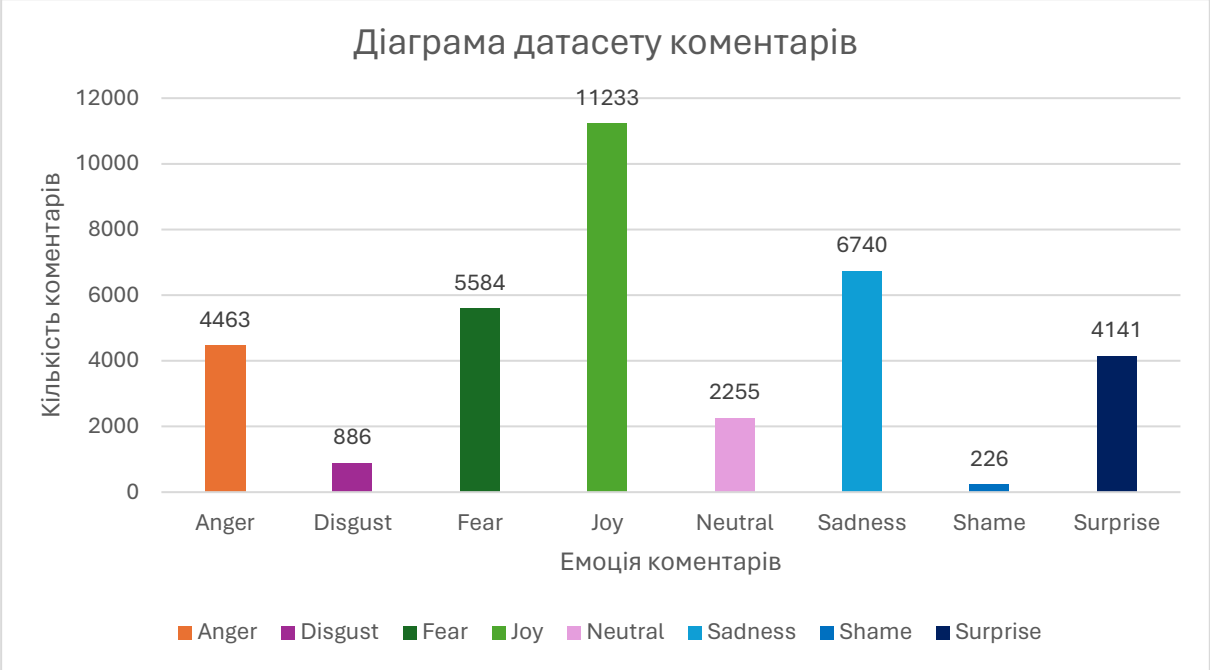

Рисунок 3.1 – Датасет коментарів Твіттеру (виконано самостійно)

Для задачі багатокласової класифікації тональності тексту коментарі були розділені на три категорії: негативні, позитивні та нейтральні. Для відмітки цих категорій був введений додатковий стовбець, що має одне з трьох можливих значень: -1 (для текстів з негативною емоційною складовою), 0 (для нейтральних текстів) та 1 (для текстів з позитивною емоційною складовою).

Коментарі з емоцією «anger», «disgust», «fear», «sadness», «shame» були позначені як негативні, коментарі з емоцією «neutral», «surprise» як нейтральні та коментарі з емоцією «joy» як позитивні.

На рисунку 3.2 можна побачити діаграму яка показує розподіл коментарів за трьома класами: негативні, позитивні та нейтральні. На підставі цієї діаграми можна зробити такі висновки, що спостерігається значна невідповідність кількості коментарів між класами, тобто класовий дисбаланс. Для цього можна використовувати параметр class_weight='balanced' у Scikit-learn для автоматичного зважування класів під час навчання моделі. Та для покращення якості моделі класифікації, можливо, знадобиться методи oversampling (наприклад, SMOTE) або undersampling [7].

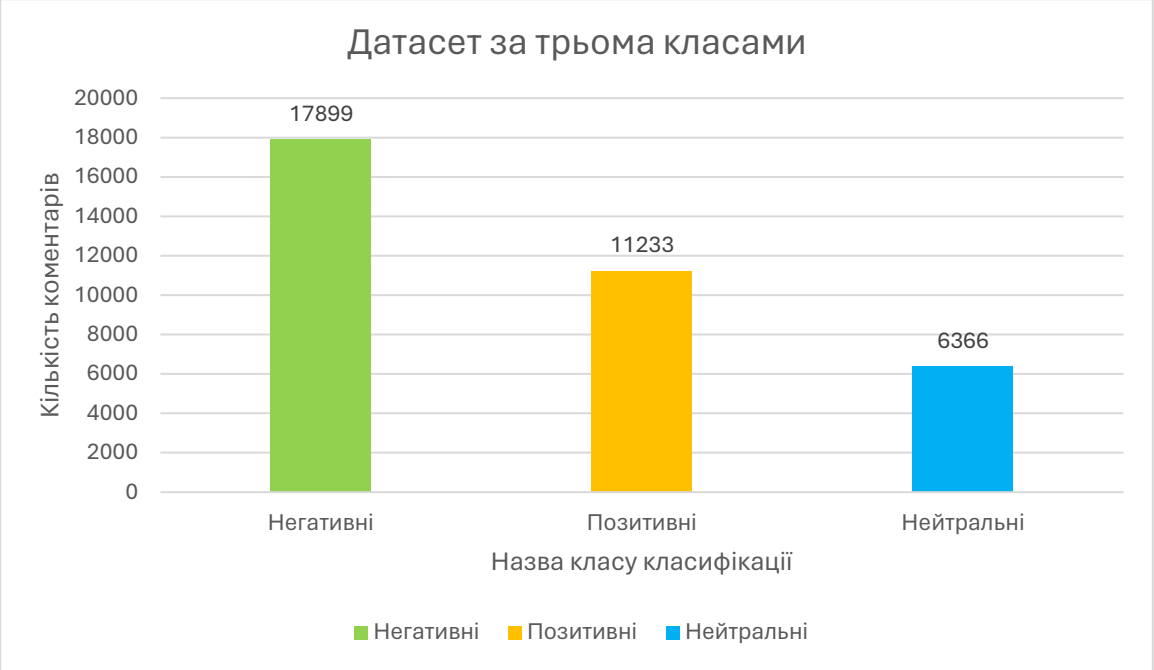

Рисунок 3.2 – Датасет багатокласової класифікації (виконано самостійно)

Для отримання найкращих результатів навчання багатокласової класифікації навчальні дані мають бути збалансовані (тобто число позитивних, нейтральних та негативних навчальних даних має бути однаковим).

Отриманий датасет був поділений на дві частини: 80% для тренування моделі та 20% для її тестування. Ці набори даних використовувалися для створення моделей класифікації в рамках цього дослідження.

3.2 Оптимальний алгоритм аналізу та побудова класифікатора

Був проведений експеримент з визначенням найкращого алгоритму для розв'язання задачі багатокласової класифікації тексту, використовуючи бібліотеку Scikit-learn. Дослідження проводилося на прикладі коментарів користувачів у Twitter.

Багатокласова класифікація в контексті тональності тексту означає розділення текстів на кілька класів в залежності від їхньої емоційної забарвленості. У такій класифікації можуть бути такі основні класи: позитивний, негативний та нейтральний. Позитивний клас відображає текст, який містить позитивні емоції або висловлює позитивне ставлення до чого-небудь. Негативний клас відповідає

текстам з негативними емоціями або відображає негативне ставлення. Нейтральний клас застосовується до текстів, що не містять яскраво виражених позитивних або негативних емоцій і є відносно об'єктивними. Такий підхід дозволяє більш точно аналізувати тон текстів та виявляти їхні емоційні відтінки.

У дослідженні було використано Bag of Words (BoW) – це один із найпростіших і найпоширеніших методів для представлення текстових даних у вигляді числових векторів. Основна ідея BoW полягає в наступному: текст представляється як мішок слів, де важлива частота появи слів, але не їхній порядок [8].

Основні кроки методу BoW:

- токенізація: розбити текст на окремі слова або токени;
- створення словника: створити список усіх унікальних слів у тексті;
- створення векторів: подати кожен текст як вектор, довжина якого дорівнює розміру словника. Вектор містить кількість входжень кожного слова зі словника у текст.

Переваги BoW:

- простота реалізації та інтерпретації;
- ефективний для невеликих та середніх текстових даних.

Недоліки BoW:

- ігнорування порядку слів: BoW не враховує синтаксичних та семантичних відносин між словами;
- висока розмірність: для великих текстових корпусів розмір словника може бути дуже великим, що призводить до векторних розріджених уявлень;
- обмежене уловлювання контексту: BoW не враховує контекст, у якому використовуються слова, що може знижувати точність більш складних завдань.

BoW є базовим методом для перетворення текстових даних на числові вектори, які можуть бути використані в моделях машинного навчання. Незважаючи на свої обмеження, BoW часто є відправною точкою для складніших моделей обробки природної мови.

Для здійснення експерименту використовувалися алгоритми, що містяться у бібліотеці Scikit-learn:

– логістична регресія: LogisticRegression;
– метод опорних векторів (SVM): SVC (для класифікації) або SVR (для регресії);
– метод k-найближчих сусідів (KNN): KNeighborsClassifier;
– наївний байєсовський класифікатор: GaussianNB (для нормального розподілу ознак) або MultinomialNB (для мультиноміального розподілу ознак);
– випадковий ліс: RandomForestClassifier;
– градієнтний бустинг: GradientBoostingClassifier;
– дерева прийняття рішень: DecisionTreeClassifier.

Як показує таблиця 3.1, модель LogisticRegression, яка базується на алгоритмі логістичної регресії, демонструє найкращі результати у задачі багатокласової класифікації тексту [9-11].

Таблиця 3.1 – Порівняння алгоритмів бібліотеки Scikit-learn

Algorithm	Accuracy	Precision (Avg)	Recall (Avg)	F1-Score (Avg)
LogisticRegression	94%	0.94	0.94	0.94
KNN	92%	0.92	0.92	0.92
SVM	92.67%	0.91	0.93	0.92
Decision Tree	90%	0.91	0.91	0.91
Random Forest	92%	0.92	0.92	0.92
Gaussian Naive Bayes	91%	0.90	0.90	0.90
Gradient Boosting Class	93%	0.92	0.93	0.93

Логістична регресія – це метод машинного навчання, який використовується для вирішення завдань бінарної класифікації. Однак його можна модифікувати для роботи з багатокласовими задачами класифікації, включаючи класифікацію тексту. Один із підходів до вирішення таких задач – це метод «Один проти Всіх» (One vs

Rest, OvR), також відомий як «Один проти Всіх інших» або «One vs All» (OvA) [12].

У випадку з логістичною регресією та підходом OvR для багатокласової класифікації тексту ми створюємо окрему модель для кожного класу. Наприклад, якщо у нас є 3 класи, ми навчаємо 3 моделі логістичної регресії. Кожна модель навчається на даних, де цільовий клас вважається позитивним, а всі інші класи вважаються від'ємними.

Процес навчання та передбачення виглядає наступним чином:

– навчання: для кожного класу k навчаємо модель логістичної регресії на даних, де цільовий клас є класом k, а всі інші класи об'єднуються в один від'ємний клас. Таким чином, у нас буде k моделей для k класів;
– передбачення: щоб зробити передбачення для нового тексту, ми подаємо його на вхід кожній з k моделей. Кожна модель дає оцінку належності тексту до свого класу. Потім ми обираємо клас з найбільшою оцінкою як передбачений клас для даного тексту.

Цей підхід добре підходить для багатокласової класифікації тексту, особливо якщо класи в задачі несбалансовані або якщо у нас немає великого обсягу даних. Логістична регресія відносно проста в розумінні та реалізації, що робить її популярним вибором для багатьох застосувань класифікації тексту.

Математичну логістичну регресію можна описати так [13-15].

Позначимо дані як (x_1, x_2, ..., x_n), де (x_n) – це ознаки тексту, (k) – кількість унікальних класів. Ми маємо k моделей для класифікації, де кожна модель використовує логістичну регресію для передбачення ймовірності належності тексту до конкретного класу.

Функція логістичної регресії для класу k може бути виражена так:

$$P(y = k|x) = \frac{1}{1+ e^{-z_k}}, \quad (3.1)$$

де z_k – лінійна комбінація вхідних ознак та параметрів моделі для класу k:

$$z_k = \theta_{k0} + \theta_{k1} * x_1 + \theta_{k2} * x_2 + \theta_{k0} + \cdots + \theta_{kn} * x_n, \quad (3.2)$$

де $\theta_{k0}, \theta_{k1}, \ldots, \theta_{kn}$ – це параметри моделі для класу k, які потрібно навчити з

даних.

При класифікації нового тексту ми визначаємо ймовірність для кожного класу за допомогою відповідних моделей логістичної регресії, а потім обираємо клас з найбільшою ймовірністю як передбачений клас для тексту.

Отже, розширена логістична регресія для багатокласової класифікації тексту використовує принцип «Один проти Всіх» для створення декількох моделей класифікації, що дозволяє розв'язувати завдання класифікації з більшою кількістю класів.

Модель, створена за допомогою алгоритму логістичної регресії, а також результати її метрик представлені на рисунку 3.3.

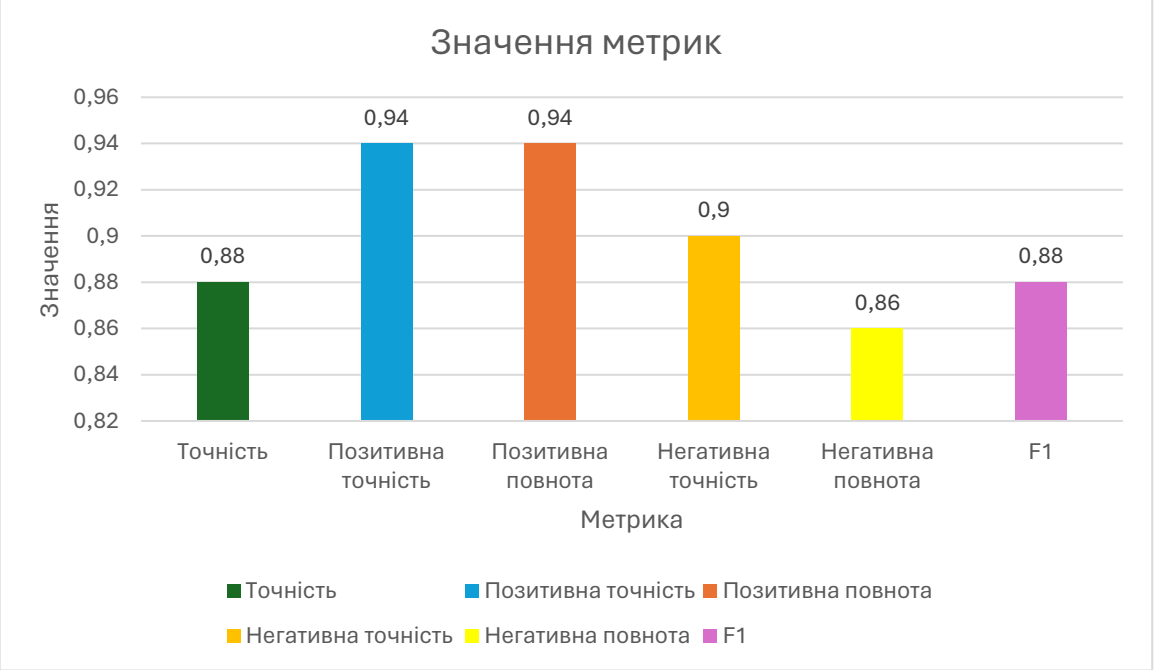

Рисунок 3.3 – Значення метрик логістичної регресії (виконано самостійно)

Отже, враховуючи важливість метрик, можна зазначити, що модель класифікатора, побудована за допомогою алгоритму логістичної регресії, демонструє високу якість.

3.3 Використання багатомовної BERT-моделі

Під час дослідження використовувалася передова багатомовна модель BERT для вирішення завдання бінарної класифікації текстів англійською мовою. Процес включав такі етапи:

– попередня обробка текстових даних для підготовки їх до використання з моделлю BERT та створення набору даних;
– використання навчання для розробки класифікатора емоційного відтінку;
– оцінка ефективності моделі на тестових даних.

BERT (Bidirectional Encoder Representations from Transformers) означає двонапрямлене кодування представлення від трансформерів. На відміну від моделей, які обробляють вхідний текст послідовно (зліва направо або справа наліво), кодер трансформеру аналізує всю послідовність слів одразу. Ця можливість дозволяє моделі аналізувати контекст слова, враховуючи його оточення як зліва, так і справа [16].

BERT був навчений шляхом маскування 15% токенів для їхнього відновлення та передбачення наступного речення. Зібраний датасет був використаний для тренування моделі BERT з метою бінарної класифікації тексту. Бібліотека Transformers надає різноманітні моделі Transformer, включаючи BERT, із підтримкою TensorFlow та PyTorch, а також включає токенізатори. Модель bert-base-multilingual-uncased використовувалась у цьому дослідженні та була завантажена за допомогою бібліотеки Transformers.

У процесі дослідження були виконані такі кроки:

– коментарі були поділені на токени;
– були додані спеціальні токени;
– визначена оптимальна довжина речення.

BERT оптимально працює з послідовностями фіксованої довжини. У даному дослідженні застосовувалася проста стратегія вибору максимальної довжини. Для цього зберігалася довжина символів у кожному відгуку, і визначалася їхня оптимальна довжина (див. рис. 3.4).

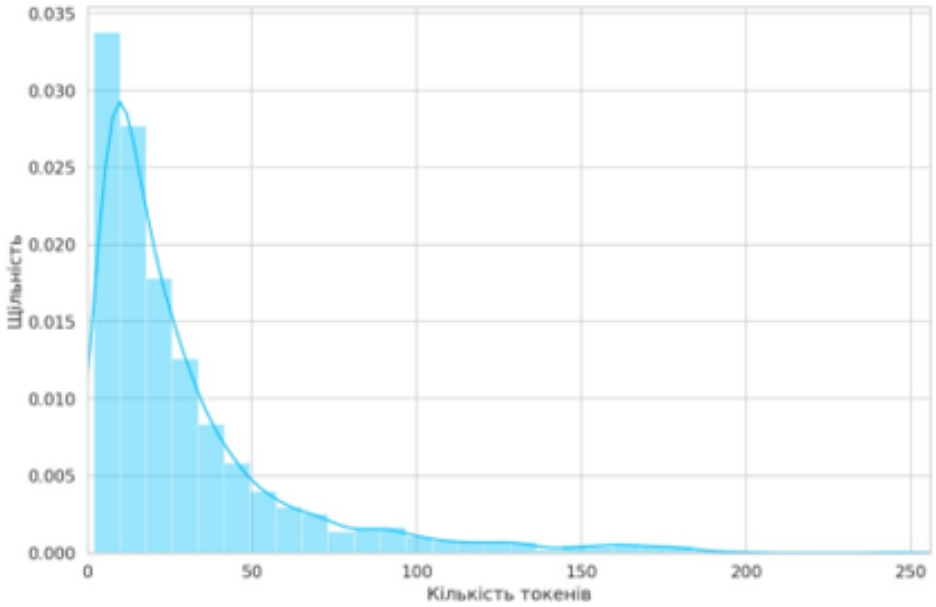

Рисунок 3.4 – Розподілення довжини токенів (виконано самостійно)

Згідно з рис. 3.4, більшість відгуків мають менше 150 токенів. Для налагодження моделі встановлено довжину 200 токенів. Після цього було побудовано нейронну мережу та проведено навчання на тренувальному наборі даних, перевіривши роботу моделі на тестових даних. Щоб повторити процедуру навчання з документації BERT, було використано оптимізатор AdamW, наданий Hugging Face, який компенсує втрату ваги. Для моделі потрібно вказати функцію втрат та оптимізатор для навчання. Оскільки це завдання бінарної класифікації, і модель буде відображати ймовірності, була використана функція втрат CrossEntropyLoss.

Автори моделі BERT рекомендують кілька налаштувань для оптимізації процесу навчання:

– розмір партії (батчу): 16 або 32;

– швидкість навчання (метод оптимізації Адам): 5e-5, 3e-5, 2e-5;

– кількість епох (ітерацій): від 2 до 10.

Збільшення розміру партії значно прискорює час навчання, але може призвести до зниження точності моделі. Після завершення навчання оцінюються втрати та точність моделі на 20% випадкових зразків з набору даних для перевірки. Для цього використовується функція, яка отримує на вхід два масиви: тестові

відгуки та відповідні маркування позитивного/негативного відгука. Результатом роботи цієї функції є пара чисел: відсоток втрат (loss, чим менше, тим краще) та точність (див. рис. 3.5).

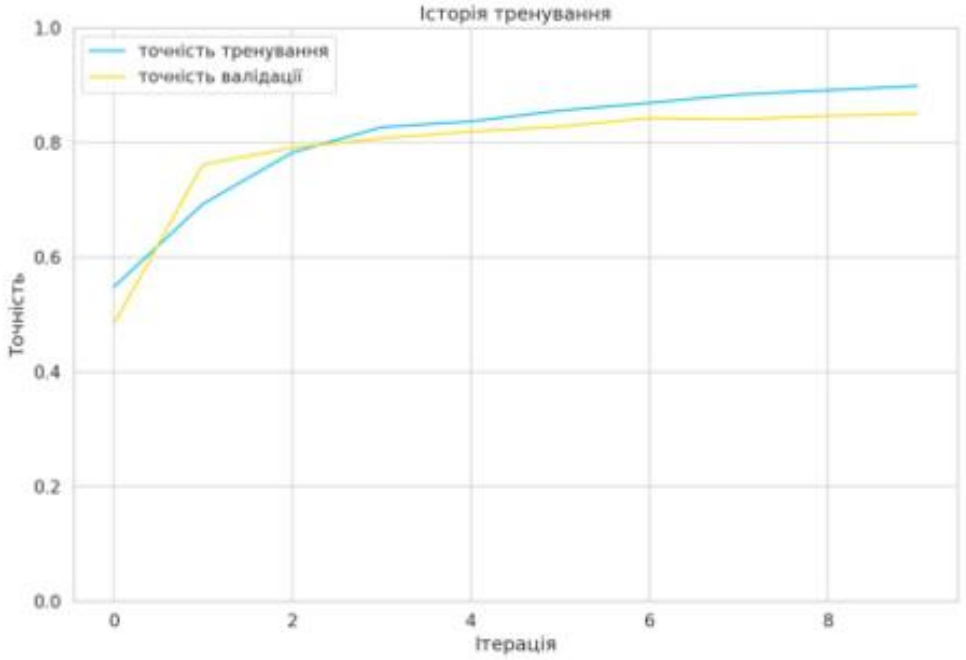

Рисунок 3.5 – Історія тренування моделі BERT (виконано самостійно)

Модель була оцінена після проведення перехресного тестування. Дані процесу оцінки представлені на рисунку 3.6.

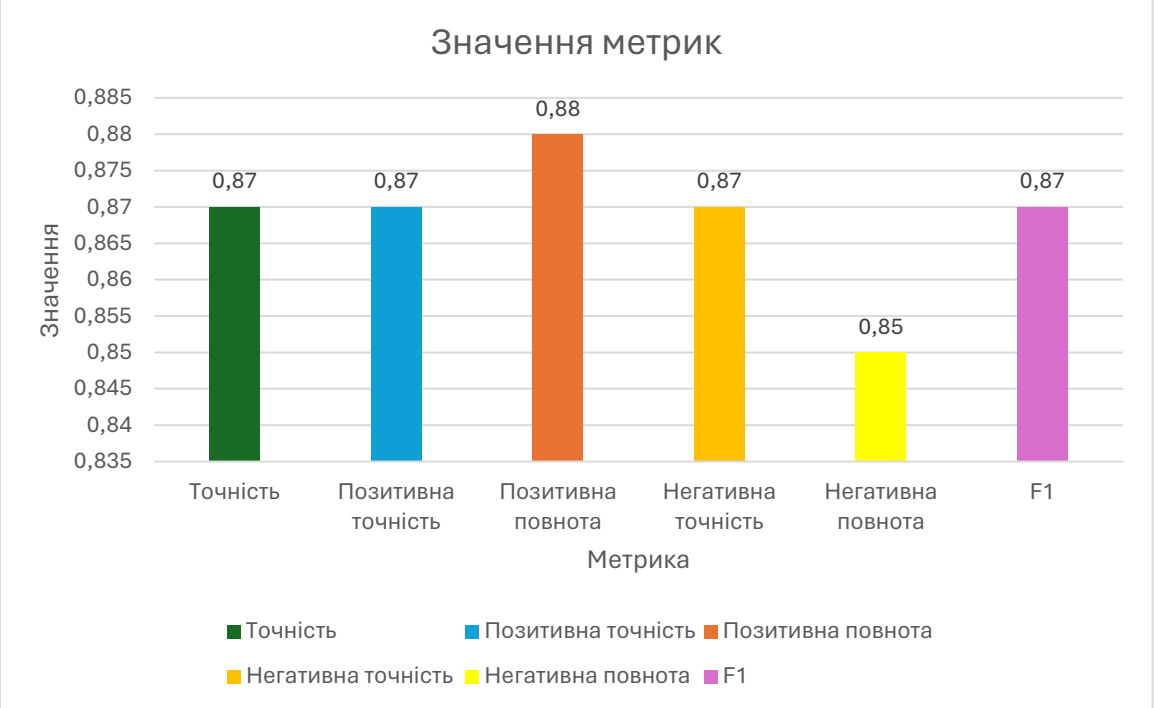

Рисунок 3.6 – Значення метрик BERT моделі (виконано самостійно)

Результати метрик після тренування та тестування моделі свідчать про те, що дотренована багатомовна модель BERT ефективна для бінарної класифікації тексту. Вона відзначається високою повнотою і точністю у виявленні текстів з різними емоційними відтінками, як позитивними, нейтральними, так і негативними.

3.4 Порівняння реалізацій моделей класифікатора

Для визначення найбільш ефективної моделі класифікації ми порівнювали показники ключових метрик для розроблених моделей логістичної регресії і BERT у таблиці 3.2.

Таблиця 3.2 – Порівняння показників ефективності моделі логістичної регресії та моделі на основі BERT

Метрика	Отримане значення		Порівняння
	Модель логістичної регресії	BERT-модель	
Точність	0.88	0.87	Отримані дані свідчать, що як класифікатори моделей є ефективними, проте класифікатор у моделі логістичної регресії проявляє трошки вищу ефективність порівняно з класифікатором у моделі BERT
Позитивна точність	0.94	0.87	Модель логістичної регресії ефективніше розпізнає тексти з позитивною тональністю, ніж модель BERT

Кінець таблиці 3.2

Метрика	Отримане значення		Порівняння
	Модель логістичної регресії	BERT-модель	
Позитивна повнота	0.94	0.88	Модель логістичної регресії більш точно встановлює текст із позитивною спрямованістю, ніж модель BERT
Негативна точність	0.90	0.87	Модель логістичної регресії ефективніше розпізнає текст із вираженою негативною тональністю, ніж модель BERT
Негативна повнота	0.86	0.85	Класифікатор моделі логістичної регресії більше уточнено визначає текст із негативною емоційною окраскою, ніж класифікатор моделі BERT
F1	0.88	0.87	Метрика оцінює точність класифікатора. Обидві моделі демонструють високу точність, але у цьому конкретному випадку модель логістичної регресії є точнішою ніж BERT-модель

Після проведення тренування та тестування моделей нашим зібраним датасетом ми отримали результати метрик для моделі логістичної регресії та для моделі BERT, що свідчать про високу якість обох моделей у виявленні емоційного

відтінку тексту. Однак модель логістичної регресії демонструє більшу точність та ефективність у розпізнаванні текстів з позитивним, нейтральним та негативним забарвленням, що робить її більш якісною для виконання завдання багатокласової класифікації текстів.

3.5 Розробка програмного забезпечення

Перегляд існуючих програм для аналізу емоційного забарвлення тексту підкреслив необхідність створення автоматизованої системи для такого аналізу. Вирішено розробити веб-додаток, який буде доступний для всіх без необхідності автентифікації чи авторизації. Для цього використовується PyCharm 2024.1.1 для реалізації Web API з використанням Python, а також Streamlit для клієнтської частини.

Python – це високорівнева, інтерпретована мова програмування з динамічним типом даних. Вона має простий і зрозумілий синтаксис, що робить її дуже популярною серед початківців і досвідчених програмістів. Python використовується для розробки різноманітних програм, від веб-додатків до штучного інтелекту. Він також має велику кількість бібліотек і фреймворків, що робить його потужним і гнучким інструментом для різних завдань.

Python має численні переваги:

- синтаксис Python легкий для вивчення і зрозуміння, що робить його ідеальним вибором для початківців;
- має активну спільноту користувачів, яка розвивається швидко і надає підтримку та рішення проблем;
- має велику кількість сторонніх бібліотек і фреймворків для різних задач, що дозволяє вам швидко розробляти програми з мінімальними зусиллями;
- підтримується на багатьох платформах, таких як Windows, macOS та Linux, що робить його універсальним інструментом для розробки;
- може легко інтегруватися з іншими мовами програмування, що робить його ідеальним вибором для розробки складних програм;
- має вбудовані функції для роботи зі списками, словниками, рядками,

файлами і багатьма іншими типами даних, що спрощує роботу з програмами.

PyCharm – це інтегроване середовище розробки (IDE) для мови програмування Python. Воно має багато корисних функцій, таких як підказки під час написання коду, автодоповнення, інтеграція з системами контролю версій, вбудована підтримка віртуальних середовищ і інструменти для відлагодження коду. PyCharm допомагає програмістам писати код швидше і ефективніше.

Для реалізації клієнтської частини додатку було використано бібліотеку Streamlit. Вона призначена для створення веб-застосунків за допомогою простого і зрозумілого синтаксису. Вона дозволяє швидко та легко створювати інтерактивні програми для візуалізації даних, машинного навчання та інших цілей.

Для аналізу емоційного відтінку тексту використовувалася модель логістичної регресії, яка показала найвищу ефективність та якість в цьому дослідженні, про яке було згадано раніше у цій роботі.

Бібліотека Scikit-learn надає можливість використовувати дану модель для класифікації тексту у процесі виконання програми, розробленої на платформі PyCharm.

Було використано клас LogisticRegression бібліотеки Scikit-learn для реалізації логістичної регресії, а саме методи цього класу:

- fit(X, y): цей метод використовується для навчання моделі на навчальних даних. Він приймає як аргументи матрицю ознак X і вектор міток класів y, навчає модель логістичної регресії цих даних і налаштовує параметри моделі в такий спосіб, щоб мінімізувати обрану функцію втрат;
- predict(X): цей метод використовується для передбачення позначок класів на нових даних. Після того, як модель навчена за допомогою методу fit(), ви можете передати нові дані X метод predict(), щоб отримати передбачені мітки класів;
- predict_proba(X): цей метод використовується для отримання ймовірностей приналежності до кожного класу нових даних. Повертає ймовірності, передбачені моделлю, у вигляді матриці, де рядки відповідають зразкам X,

а стовпці – класам;

– score (X, y): цей метод використовується для оцінки якості моделі на тестових даних. Він приймає матрицю ознак X і вектор міток класів y робить прогнози на даних X і порівнює їх з істинними мітками y, повертаючи точність моделі;
– pipe_lr.classes_ – це атрибут, який використовується для отримання списку унікальних класів, навчених моделлю логістичної регресії усередині конвеєра (pipeline) у бібліотеці scikit-learn.

Користувач може ввести текст, який буде проаналізовано на емоційне забарвлення. Цей аналіз виконується як завдання багатокласової класифікації, де текст призначається до одного з трьох класів: «позитивний», «нейтральний» або «негативний».

Класифікація здійснюється за допомогою моделі, яка була створена під час цього дослідження на основі зібраного датасету текстів англійською мовою з соціальної мережі Твіттер. Інтерфейс розробленого додатку показаний на рисунку 3.7.

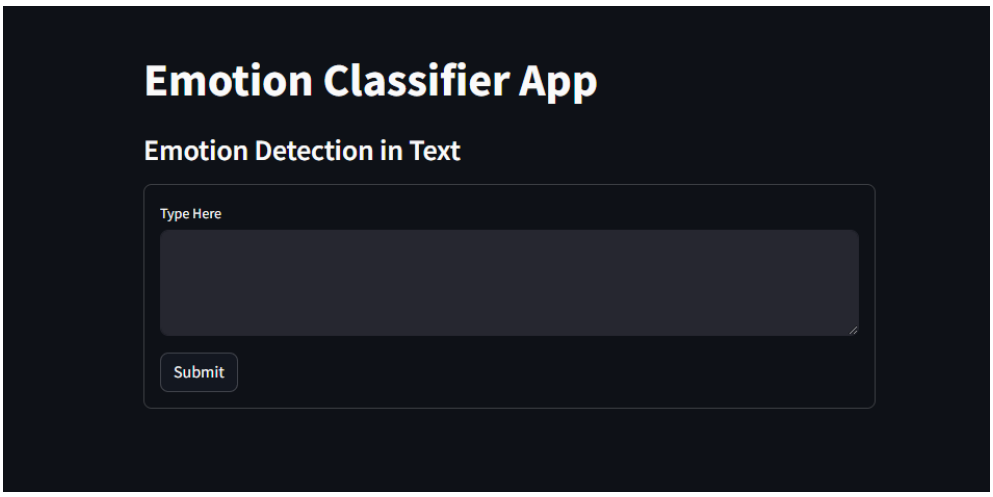

Рисунок 3.7 – Інтерфейс веб-додатку

Діаграма, що показує процес визначення емоційного відтінку тексту за допомогою розробленої програмної системи, зображена на рис. 3.8.

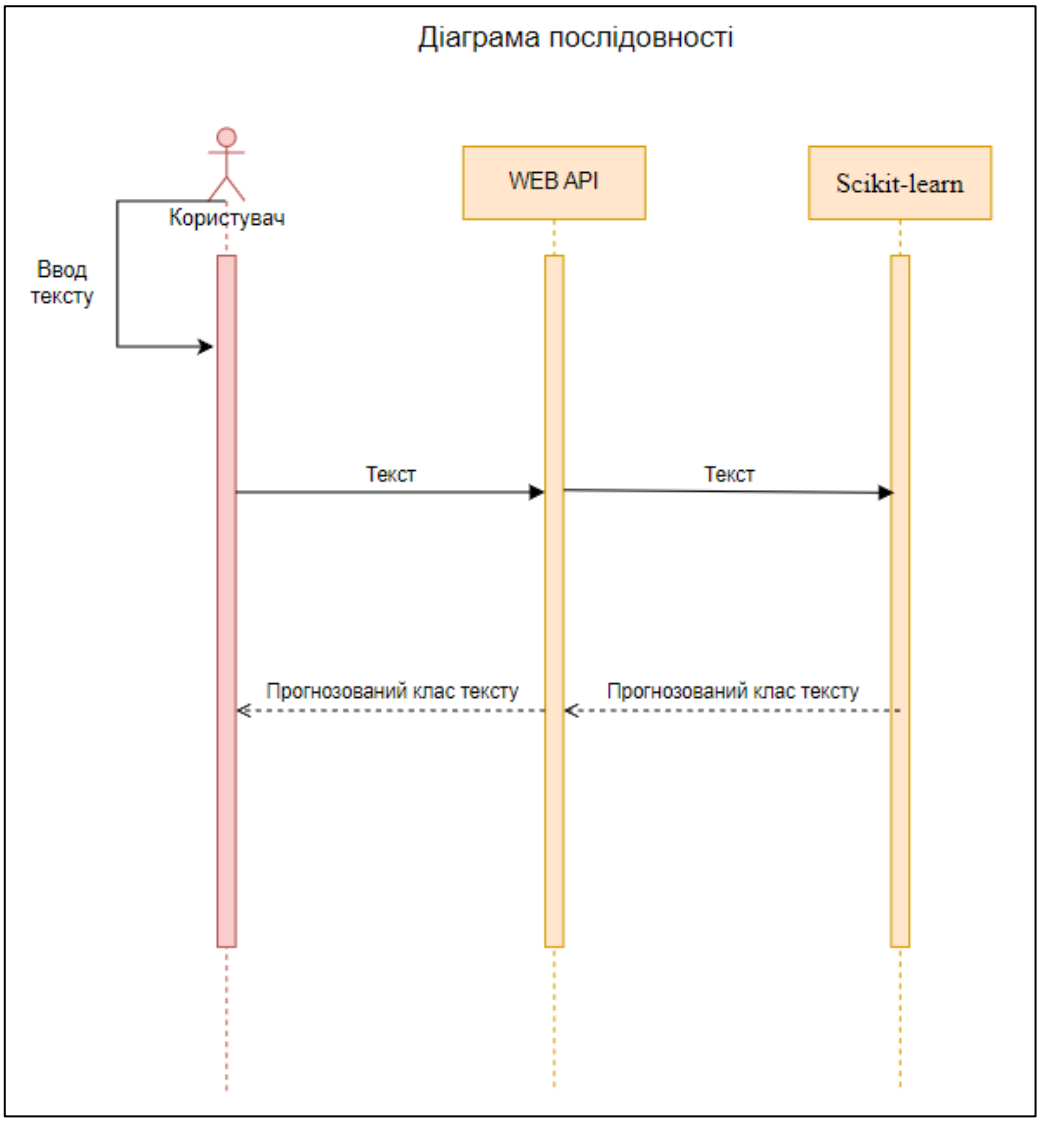

Рисунок 3.8 – Діаграма послідовності аналізу емоційного відтінку тексту (виконано самостійно)

Залежно від класу, до якого був віднесений текст класифікатором, користувач отримує відповідне повідомлення про те, чи має текст позитивне, нейтральне або негативне значення, також він побачить впевненість емоції тексту та діаграму імовірності прогнозу у вигляді трьох стовпчиків, а саме три класи класифікації тексту. Наприклад відображення позитивного забарвлення тексту можна побачити на рисунку 3.9.

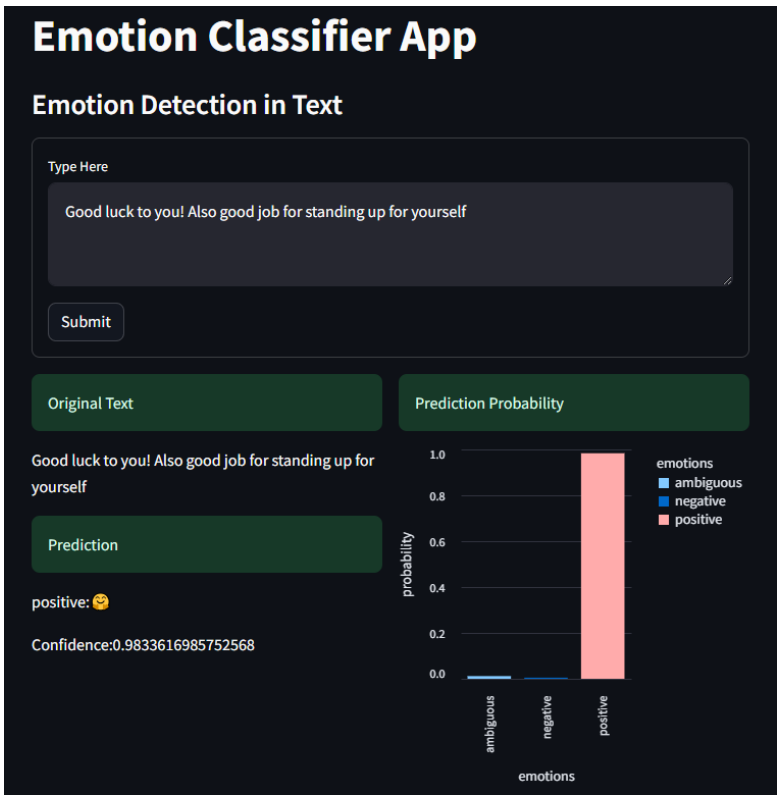

Рисунок 3.9 – Результати аналізу текстового вмісту з позитивними емоціями

Приклад відображення нейтрального забарвлення тексту можна побачити на рисунку 3.10.

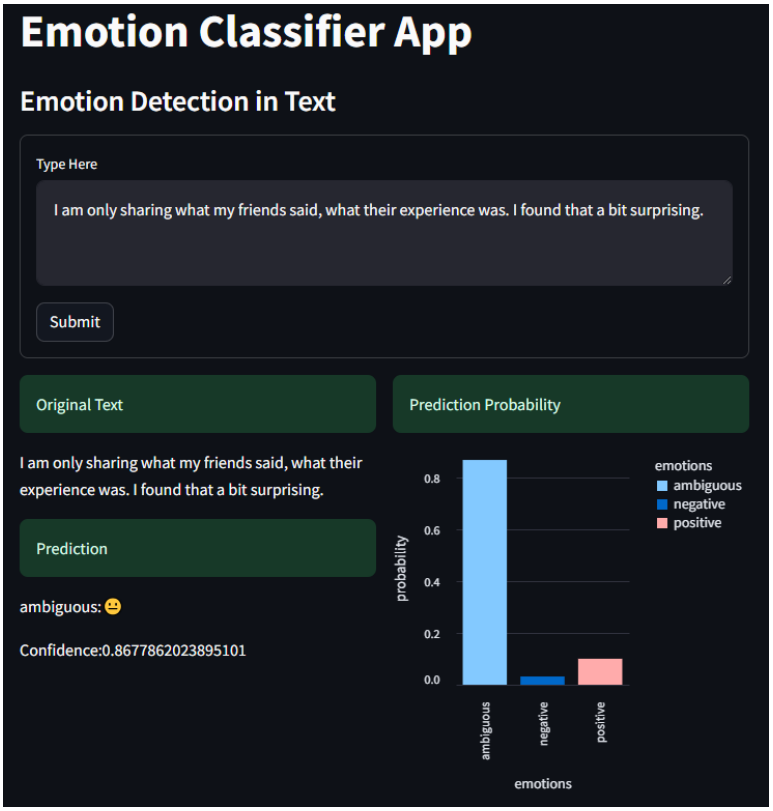

Рисунок 3.10 – Результати аналізу текстового вмісту з нейтральними емоціями

Останній приклад відображає негативну емоцію в тексті (див. рис. 3.11).

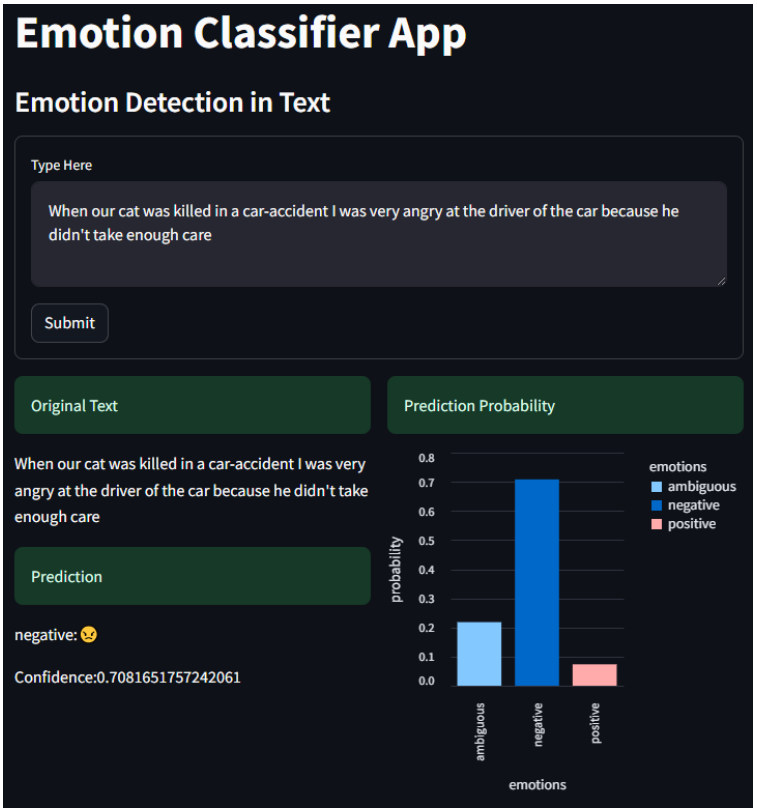

Рисунок 3.11 – Результати аналізу текстового вмісту з негативними емоціями

Система була розроблена для оцінки емоційного відтінку текстів і може бути використана у різних сферах, але найбільш ефективні результати вона показує в аналізі емоційного забарвлення текстів у коментарях на Twitter. Це пояснюється тим, що класифікатор, який використовується в цій системі, був навчений на основі даних з коментарів Twitter.

ВИСНОВКИ

Емоційне забарвлення є однією з найважливіших характеристик письмового тексту або мовлення, що відіграє важливу роль у сучасному аналізі текстових сутностей [17].

Проведення аналізу емоційного відтінку тексту сприяє розумінню соціальних настроїв щодо бренду, продукту або послуги в сфері бізнесу. Цей підхід може бути використаний для ринкового дослідження, аналітики продуктів, а також для підтримки користувачів та моніторингу соціальних медіа. досліджень та технологій в промисловості.

У ході дослідження було розглянуті існуючі підходи до аналізу емоційного відтінку тексту, методи класифікації текстів за їх емоційним забарвленням та проаналізовані існуючі системи для цього аналізу. Були досліджені методи та інструменти аналізу емоційного відтінку тексту з використанням машинного навчання, а також розглянуті набори даних (датасети), які можуть бути використані у машинному навчанні.

Під час роботи було створено програмне забезпечення для створення набору даних англійською мовою, яке було сформовано на основі коментарів користувачів у Twitter. Цей набір даних може бути використаний для подальших досліджень у сфері аналізу емоційного відтінку тексту.

Проведено експериментальне дослідження ефективності різних алгоритмів машинного навчання для виконання завдання багатокласової класифікації англійського тексту [18-22]. У цьому дослідженні був розроблений додаток, інтерфейс додатка реалізований за допомогою веб-фреймворку Streamlit, розробленого на мові програмування Python для створення інтерактивних веб-застосунків.

Під час дослідження була створена модель машинного навчання, яка навчається на витягнутих ознаках, щоб передбачити емоції, виражені в текстових даних.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Analysis Of Texts Emotional Content In Multidimensional Space URL: https://www.researchgate.net/publication/255566883_Analysis_of_texts'_emotional_content_in_a_multidimensional_space (дата звернення: 22.11.2023).

2. Natural Language Processing (NLP): 7 Key Techniques URL: https://monkeylearn.com/blog/natural-language-processing-techniques/ (дата звернення: 22.11.2023).

3. Text Analytics Tools for Natural Language Processing URL: https://www.repustate.com/blog/top-8-text-analytics-tools/ (дата звернення: 22.11.2023).

4. International Journal of Computational Intelligence Systems URL: https://link.springer.com/article/10.1007/s44196-023-00234-5#Equ2 (дата звернення: 28.11.2023).

5. Twitter Sentiment Analysis Dataset URL: https://public.tableau.com/app/profile/saif2522/viz/EmoInt-web2/EmotionIntensityDashboard (дата звернення: 30.11.2023).

6. Emotion and sentiment analysis from Twitter text URL: https://www.sciencedirect.com/science/article/abs/pii/S1877750318311037 (дата звернення: 30.11.2023).

7. API Reference UTL: https://scikit-learn.org/stable/modules/generated/sklearn.utils.class_weight.compute_class_weight.html (дата звернення: 26.05.2024).

8. An Introduction to Bag of Words (BoW) | What is Bag of Words? URL: https://www.mygreatlearning.com/blog/bag-of-words/ (дата звернення: 26.05.2024).

9. Classifier comparison URL: https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html (дата звернення: 26.05.2024).

10. Iris Dataset Classification with Python: A Tutorial URL: https://www.pycodemates.com/2022/05/iris-dataset-classification-with-python.html (дата звернення: 27.05.2024).

11. Iris Dataset Classification using Support Vector Machine, Random Forest, and Gradient Boosting Classifier URL: https://www.embedded-robotics.com/iris-dataset-classification/#google_vignette (дата звернення: 28.05.2024).

12. One-vs-Rest and One-vs-One for Multi-Class Classification URL: https://machinelearningmastery.com/one-vs-rest-and-one-vs-one-for-multi-class-classification/ (дата звернення: 28.05.2024).

13. Logistic Regression (Classification) – Mathematical intuition URL: https://parisrohan.medium.com/logistic-regression-classification-mathematical-intuition-5e462324598c (дата звернення: 28.05.2024).

14. Logistic Regression from Scratch: Multi classification with OneVsAll URL: https://medium.com/analytics-vidhya/logistic-regression-from-scratch-multi-classification-with-onevsall-d5c2acf0c37c (дата звернення: 29.05.2024).

15. Multi-class Classification – One-vs-All & One-vs-One URL: https://towardsdatascience.com/multi-class-classification-one-vs-all-one-vs-one-94daed32a87b (дата звернення: 29.05.2024).

16. Mastering BERT: A Comprehensive Guide from Beginner to Advanced in Natural Language Processing (NLP) URL: https://medium.com/@shaikhrayyan123/a-comprehensive-guide-to-understanding-bert-from-beginners-to-advanced-2379699e2b51 (дата звернення: 29.05.2024).

17. Фролов М.В., Валенда Н.А., Оцінка ефективності методів аналізу емоційного забарвлення коментарів // 28-й Міжнародний молодіжний форум «Радіоелектроніка та молодь у XXI столітті», Харків, ХНУРЕ, 2023. С.355-358.

18. Задача аналізу тональності тексту Шуляк С.М, Валенда Н.А. Topical issues of the development of modern science // Abstracts of the 9th International scientific and practical conference. Sofia, Bulgaria: ACCENT, 2020. с. 951-956.

19. Filatov V. O., Yerokhin A. L., Zolotukhin O. V., Kudryavtseva M. S. Hybrid simulation models for complex decision-making problems with partial uncertainty. Information Extraction and Processing. 2022, 50(126), 78-86. DOI: https://doi.org/10.15407/vidbir2022.50.078

20. Dmytro Panchenko, Daniil Maksymenko, Olena Turuta, Andriy Yerokhin,

Yana Daniiel, Oleksii Turuta . Evaluation and Analysis of the NLP Model Zoo for Ukrainian Text Classification // Communications in Computer and Information Science, 2022, 1698 CCIS, pp. 109–123. DOI: 10.1007/978-3-031-20834-8_6

21. Daniil Maksymenko, Nataliia Saichyshyna, Oleksii Turuta, Olena Turuta, Andriy Yerokhin, Andrii Babii. Improving the Machine Translation Model in Specific Domains for the Ukrainian Language // International Scientific and Technical Conference on Computer Sciences and Information Technologies, 2022, 2022-November, pp. 123–129. DOI: 10.1109/CSIT56902.2022.10000529

22. Extension Multi30K: Multimodal Dataset for Integrated Vision and Language Research in Ukrainian. Nataliia Saichyshyna, Daniil Maksymenko, Oleksii Turuta, Andriy Yerokhin, Andrii Babii, Olena Turuta / EACL 2023 - 2nd Ukrainian Natural Language Processing Workshop, UNLP 2023 - Proceedings of the Workshop, 2023, P.P. 54–61 / DOI: 10.18653/v1/2023.unlp-1.7