МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Харківський національний університет радіоелектроніки

Факультет Центр післядипломної освіти
(повна назва)

Кафедра Програмної інженерії
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)

Дослідження методів аналізу емоційного забарвлення тексту українською мовою
(тема)

Виконав:
Випускник 2 курсу, групи ІПЗздм-19-1
Рябишев О.В.
(прізвище,ініціали)

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

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

Керівник проф., д.т.н. Єрохін А.Л.
(посада,прізвище)

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

2021 р.

Харківський національний університет радіоелектроніки

Факультет Центр післядипломної освіти
(повна назва)

Кафедра Програмної інженерії
(повна назва)

Рівень вищої освіти другий (магістерський)

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

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

Освітня програма Інженерія програмного забезпечення
(повна назва)

ЗАТВЕРДЖУЮ:
Зав.кафедри _____
(підпис)
« 26 » березня 2021 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

студента Рябишева Олексія Вікторовича
(прізвище, ім'я, по батькові)

1. Тема роботи Дослідження методів аналізу емоційного забарвлення тексту українською мовою

затверджена наказом університету від 26.03.2021 № 34Стз

2. Термін подання роботи до екзаменаційної комісії 08 травня 2021р.
3. Вихідні дані до роботи алгоритми класифікації текстів, претренована модель нейронної мережі, пояснювальна записка. Використовувати ОС Windows, середовище об'єктно-орієнтованого проєктування Microsoft Visual Studio
4. Перелік питань, що потрібно опрацювати в роботі мета роботи, аналіз проблемної галузі і постановка задачі, огляд методів аналізу емоційного забарвлення тексту, існуючі рішення та бібліотеки
5. Перелік графічного матеріалу із зазначенням креслеників, схем, слайдів, ілюстрацій (п.5 включається до завдання за рішенням випускової кафедри)

мета дослідження, обґрунтування доцільності дослідження, підходи до аналізу емоційного забарвлення тексту, постановка задач дослідження, використані технології, формування набору даних українською мовою, порівняння результатів навчання для різних алгоритмів, метрики навченої моделі логістичної регресії, побудова моделі на основі BERT, порівняння отриманих результатів, клієнт-серверна програмна система, висновки

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

Найменування розділу	Консультант (посада, прізвище, ім'я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата
Спецчастина	проф., д.т.н. Єрохін А.Л.		

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1	Аналіз предметної галузі	25 лютого 2021р.	
2	Огляд існуючих методів та інструментів	05 березня 2021р.	
3	Реалізація програмного продукту	01 квітня 2021р.	
4	Підготовка пояснювальної записки	28 квітня 2021р..	
5	Спецчастина	30 квітня 2021р.	
6	Нормоконтроль	03 травня 2021р.	
7	Рецензування	04 травня 2021р.	
8	Підготовка презентації та доповіді	05 травня 2021р.	
9	Попередній захист	06 травня 2021р.	
10	Занесення роботи в електронний архів	07 травня 2021р.	
11	Допуск до захисту у зав. кафедри	08 травня 2021р.	

Дата видачі завдання 25 січня 2021р.

Студент _____
(підпис)

Керівник роботи _____ проф., д.т.н. Єрохін А.Л.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ / ABSTRACT

Кваліфікаційна робота магістра містить: 86 с., 9 рис., 8 табл., 29 джер.

ТОНАЛЬНІСТЬ ТЕКСТУ, СЕНТИМЕНТ-АНАЛІЗ, УКРАЇНСЬКА МОВА, МАШИННЕ НАВЧАННЯ.

Об’єктом дослідження є емоційне забарвлення (тональність) тексту природної мови.

Метою роботи є виявлення найбільш ефективних методів аналізу емоційного забарвлення тексту українською мовою та їх практична реалізація.

Методи розробки базуються на наступних технологіях та мовах програмування: C#, .NET Framework, ML.NET, Python, PyTorch, pandas.

У результаті роботи було сформовано датасет українською мовою, були побудовані та натреновані моделі для аналізу емоційного забарвлення тексту, розроблений вебдодаток для визначення емоційного забарвлення тексту українською мовою.

TEXT SENTIMENT, SENTIMENT ANALYSIS, UKRAINIAN LANGUAGE, MACHINE LEARNING.

Text sentiment of natural language is the object of the research.

The aim of the work is to identify the most effective sentiment analysis methods of the text in Ukrainian language and their practical implementation.

Development methods are based on the following technologies and programming languages: C #, .NET Framework, ML.NET, Python, PyTorch, pandas.

As a result of the work, the dataset in Ukrainian language was created, models for text sentiment analysis were built and trained, and the web application for determining the sentiment of the text in Ukrainian language was developed.

Я, Рябишев Олексій Вікторович, студент групи ІПЗздм-19-1, здобувач вищої освіти на другому (магістерському) рівні кафедри «Програмна інженерія»,

заявляю: моя кваліфікаційна робота на тему «Дослідження методів аналізу емоційного забарвлення тексту українською мовою», 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 ElAr KhNURE. Всі запозичення з друкованих та електронних джерел мають відповідні посилання.

Я ознайомлений 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

Аналіз емоційного забарвлення тексту – це задача обробки природної мови, яка сьогодні широко використовується в таких областях, як соціологія (збір даних із соціальних мереж про симпатії та антипатії людей), політологія (збір даних про політичні погляди певних соціальних груп), маркетинг (створення рейтингів товарів/компаній), медицина та психологія (виявлення ознак психічних захворювань або ознак депресії в повідомленнях користувачів, виявлення хуліганів у соціальних мережах) [1]. Проте яким би корисним не був такий інструмент, системи аналізу емоційного забарвлення тексту українською мовою ще немає. Це обумовлює актуальність теми даного дослідження.

Було впроваджено велику кількість проєктів для аналізу емоційного забарвлення тексту відгуків про готелі, банки, оглядів ресторанів, коментарів до фільмів, відгуків про товари, повідомлення про політичні події тощо. Велика кількість досліджень присвячена аналізу тональності повідомлень у мікроблогах. Ці дослідження використовують різні підходи до аналізу емоційного забарвлення тексту: підхід, заснований на словниках, підхід, що заснований на правилах та підхід з використанням машинного навчання.

Можна виділити дослідження В. Каспера (W. Kasper) і М. Вели (M. Vela) [2], К. Мойланена (K. Moilanen) та С. Пульмана (S. Pulman) [3] , Д. Кана (D. Kan) [4]. Дослідженнями методів машинного навчання займалися Джейкоб Девлін (Jacob Devlin) [5], Мінг-Вей Чанг (Ming-Wei Chang) [5], Крістіна Тоутанова (Kristina Toutanova) [7].

Серед робіт вітчизняних вчених можна виділити дослідження А. Єрохіна [8], З. Дудар [9], О. Турути [10] А. Нечипоренко [11], А. Бабія [12].

Мета дослідження – виявлення найбільш ефективних методів аналізу емоційного забарвлення тексту українською мовою та їх практична реалізація.

Для досягнення мети були вирішені наступні задачі:

– проаналізовано підходи до аналізу емоційного забарвлення;

– розглянуто способи класифікації текстів за емоційним забарвленням;

– проаналізовано існуюче програмне забезпечення для проведення аналізу емоційного забарвлення тексту;

– досліджені методи та інструменти аналізу емоційного забарвлення тексту з використанням машинного навчання та розглянуті існуючи датасети (набори даних), що можуть бути використані у машинному навчанні;

– сформовано набір даних (датасет) українською мовою, що містить відгуки клієнтів про мобільні додатки;

– проведено експериментальне дослідження, з розробкою відповідного програмного коду, щодо ефективності роботи різних алгоритмів машинного навчання для виконання задачі бінарної класифікації тексту українською мовою з використанням сформованого набору даних;

– використано попередньо навчену багатомовну модель машинного навчання BERT, проведено її доналаштування та адаптацію до задачі бінарної класифікації текстів українською мовою, враховуючи сформований набір даних;

– реалізована система аналізу емоційного забарвлення тексту українською мовою.

Об’єкт дослідження – емоційне забарвлення (тональність) тексту природної мови.

Предмет дослідження – процес автоматичного визначення емоційного забарвлення тексту, що заснований на алгоритмах машинного навчання.

Для аналізу предметної галузі використано такі теоретичні методи дослідження як аналіз та синтез. Для проведення практичної частини дослідження були застосовані такі емпіричні методи як експеримент, вимірювання та порівняння.

Результати дослідження можуть бути використані для порівняння підходів до аналізу тональності тексту, для вибору оптимального алгоритму аналізу емоційного забарвлення тексту українською мовою з використанням машинного навчання, для розробки систем з автоматичного визначення емоційного забарвлення тексту.

1 АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ ТА ПОСТАНОВКА ЗАДАЧ ДОСЛІДЖЕННЯ

1.1 Аналіз емоційного забарвлення тексту та його застосування

Аналіз емоційного забарвлення тексту (аналіз тональності тексту, сентимент-аналіз, англ. sentiment analysis, англ. opinion mining) призначений для виявленням в текстах емоційно забарвленої лексики та емоційної оцінки автора стосовно об’єктів, про які йдеться в тексті.

Емоційне забарвлення тексту є ставленням автора висловлювання до об’єкту реального світу, події, процесу чи їх властивостей), виражене в тексті.

Текстова інформація може бути розділена на два типи: факти та думки. Факти – це об’єктивні вирази про щось. Думки є суб’єктивними виразами, що описують почуття та оцінки [1].

Аналіз емоційного забарвлення тексту є складовою процесу обробки природної мови (natural language processing, NLP), основною метою якої є знаходження думок в тексті і виявлення їх властивостей. Ця задача може розглядатися як задача класифікації тексту.

При проведенні аналізу емоційного забарвлення може виникати низка труднощів, що можуть бути пов’язані з порядком слів у реченні, омонімічністю слів, орфографічними і синтаксичними помилками, оскільки це може змінювати зміст, а отже, і емоційне забарвлення.

За допомогою аналізу емоційного забарвлення текстова інформація може бути трансформована в структуровані дані громадської думки про товари, послуги, бренди, політику тощо. Ці дані можуть бути використані для комерційних застосувань, таких як маркетинговий аналіз, зв’язки з громадськістю, репутаційний менеджмент, обслуговування клієнтів та огляди товарів.

1.2 Способи класифікації тексту за емоційним забарвленням

Основним завданням при аналізі емоційного забарвлення є класифікація полярності тексту. В сучасних автоматизованих системах визначення емоційного забарвлення тексту найчастіше використовується класифікація за бінарною шкалою: чи є текст позитивним або негативним.

Недоліком цього підходу є те, що емоційну складову документа незавжди можна однозначно визначити, оскільки документ може містити ознаки як позитивної оцінки, так і негативної ознаки. Роботи в цій області включають в себе праці Терні (Peter Turney) [13] та Панга (Bo Pang) [14], які застосовували різні методи для виявлення полярності оглядів товарів та оглядів фільмів відповідно. Можна також класифікувати полярність документа за багатомірною шкалою, що спробували Панг і Снайдер (Benjamin Snyder) [15]. Ними була розширена класифікація кіновідгуків від оцінки «позитивний» або «негативний» в бік прогнозування рейтингу за 3-х або 4-бальною шкалою тоді як Снайдер здійснив поглиблений аналіз оглядів ресторанів, прогнозуючи рейтинги для різних аспектів даного ресторану, таких як їжа та атмосфера (за п'ятизірковою шкалою).

Іншим способом визначення емоційного забарвлення є використання системи масштабування, за допомогою якої слова, які зазвичай асоціюються з негативним, нейтральним або позитивним настроєм, отримують відповідне число за шкалою від -10 до +10 (найбільш негативне до найбільш позитивного) або просто від 0 до позитивної верхньої межі, наприклад +4. Це дає можливість скоригувати настрої даного терміна щодо його середовища (як правило, на рівні речення). Коли фрагмент неструктурованого тексту аналізується, кожному поняттю у вказаному середовищі виставляється оцінка на основі того, як слова сентименту відносяться до поняття та пов’язаного з ним балу. Це дозволяє перейти до більш витонченого розуміння настрою, оскільки тепер можна регулювати значення настрою концепції щодо модифікацій, які можуть оточувати його. Слова, наприклад, які посилюють,

послаблюють або заперечують настрої, виражені концепцією, можуть вплинути на її оцінку.

У більшості методів класифікації нейтральний клас ігнорується, припускаючи, що нейтральні тексти лежать поблизу межі бінарного класифікатора. Проте є дослідники, які пропонують виділяти три категорії. Більше того, деякі класифікатори від введення нейтрального класу можуть покращити загальну точність класифікації. В принципі існує два способи роботи з нейтральним класом. Або алгоритм виконує спочатку ідентифікацію нейтральної мови, її фільтрування, а потім оцінку решти з точки зору позитивних і негативних настроїв, або він створює тристоронню класифікацію за один крок. Цей другий підхід часто включає оцінку розподілу ймовірностей за всіма категоріями (наприклад, наївні класифікатори Байєса). Чи використовувати і як використовувати нейтральний клас, залежить від природи даних: якщо дані чітко згруповані в нейтральну, негативну та позитивну мову, має сенс відфільтрувати нейтральну мову та зосередити увагу на полярності між позитивними та негативними настроями. Якщо, навпаки, дані в основному нейтральні з невеликими відхиленнями до позитивного та негативного ефекту, ця стратегія ускладнить чітке розрізнення двох полюсів.

1.3 Методи аналізу емоційного забарвлення тексту

Можна виділити наступні підходи до аналізу емоційного забарвлення тексту:

– підхід, заснований на словниках настроїв;

– підхід, заснований на правилах;

– машинне навчання.

Переваги та недоліки кожного з підходів вказані в таблиці 1.1.

Таблиця 1.1 – Переваги та недоліки підходів до аналізу емоційного забарвлення тексту

Підхід	Переваги	Недоліки
Підхід, заснований на словниках настроїв	Непотрібні розмічений корпус даних і процедура навчання.	Порівняно низька якість результатів. Потрібен потужний лінгвістичний ресурс для створення словників
Підхід, заснований на правилах	Відносно висока точність для певного домену	Трудомісткість, оскільки система вимагає великої кількості рукописних правил для ефективної роботи.
Підхід з використанням машинного навчання	Не потрібні словники настроїв. Висока точність класифікації	Необхідний коментований корпус даних.

Підхід, що заснований на словниках настроїв, використовує так звані словники настроїв. Словник настроїв – це список слів з їх значеннями настроїв. Значенням настрою може бути число (наприклад, 1–10, де 1 – негативне слово, а 10 – позитивне слово) або певна категорія (наприклад, позитивне чи негативне).

Відповідно до цього підходу кожному слову в огляді присвоюється значення настрою, вказане у словнику, а після цього обчислюється сентимент усього огляду. Загальні настрої обчислюються статистично. Зазвичай це не дає високої точності, що є першим недоліком такого підходу. Наступним недоліком є відсутність місця для глибокого аналізу текстового повідомлення. Однак такий тип статистичного підходу не потребує ані частини мовлення, ані синтаксичного аналізу мови, що має велике значення для мов, у яких відсутні такі інструменти обробки тексту.

Дуже часто у словнику настроїв перераховуються лише іменники, дієслова, прикметники та прислівники. Наприклад, для дослідження настроїв текстів російською мовою [16] був використаний словник настроїв, який містив лише найпоширеніші іменники, дієслова, прикметники та прислівники, зібрані зі статей у ЗМІ. Кожному слову було присвоєно свою частину мови та силу настрою (від 1 до 3). Прикметники та прислівники поділяються на позитивні, заперечні та посилювальні. Іменники поділяються на позитивні, негативні, потенційно позитивні та потенційно негативні (це слова, почуття яких спирається на оточуючі

слова; вони позитивні в позитивному оточенні та негативні в негативному оточенні). Дієслова поділяються на вісім категорій залежно від оточення та ролі, яку вони відіграють у реченні. Дієслова, що зв'язують, складають окрему категорію.

Усі слова в словнику настроїв зазвичай стосуються певного домену (наприклад, банки, ресторани, фільми тощо), оскільки набагато складніше здійснити аналіз настроїв для загального домену.

Оскільки ручне створення такого словника є надзвичайно трудомістким завданням, розроблено методи автоматичного створення словників настроїв на основі коментованих настроїв корпусів та онтологій, які посилаються на певний домен. Якщо домен словника не визначений, також використовуються загальнодоменні семантичні мережі. Наприклад, у роботі [16] використано словник настроїв, створений на основі WordNet 2.1. Серед доступних словників настроїв найбільш часто використовуваними є SentiWordNet та General Inquirer Lexicon. SentiWordNet охоплює сім мов і часто використовується в області аналізу настроїв. На жаль, він не підтримує українську мову. General Inquirer Lexicon може використовуватися лише для англійської мови.

Слід зазначити, що статистичний підхід до аналізу настроїв, який базується на словниках настроїв, досить простий у реалізації, оскільки йому потрібні лише словник настроїв та алгоритм обчислення середнього значення настрою фрагмента тексту. Також цей підхід не вимагає часткової мітки або синтаксичного аналізу мови. Недоліком цього підходу є низька якість результатів і відсутність можливості проводити глибший аналіз тексту (наприклад, визначити причину ставлення, визначити діапазон емоцій, які автор хотів озвучити, розрізнити фрагменти тексту з певним настроєм чи емоцією тощо).

Підхід до аналізу настроїв, що заснований на правилах, базується на наборі правил, які система використовує для визначення настрою фрагмента тексту. Більшість комерційних систем використовують цей підхід, незважаючи на те, що він надзвичайно трудомісткий, оскільки система вимагає великої кількості рукописних правил для ефективної роботи. Як і словники настроїв, майже завжди

правила стосуються певного домену (наприклад, готелів, ресторанів, банків, фільмів, музики тощо), що ускладнює їх використання для аналізу тексту інших тем. Тим не менше, хороша база правил змушує такий підхід працювати досить точно в певному домені.

Обов’язковим інструментом для реалізації цього підходу є словник настроїв. Він використовується для присвоєння значення настрою кожному слову в реченні (якщо таке значення існує).

Один із найпростіших алгоритмів, що базується на правилах, для обчислення настрою речення був представлений у роботі [2], яка описує обробку текстових повідомлень про книги, фільми та цифрові камери. У цьому дослідженні аналіз настрою складається з таких кроків:

– в кожному реченні шукаються слова-інвертори (якщо таке слово знайдено, почуття наступних трьох слів змінюється на протилежне);

– кількість позитивних та негативних слів підраховується окремо;

– шукається протилежне речення, і, якщо воно знайдено, кількість суб’єктивних слів у реченні ділиться на два;

– розраховується загальний настрій: кількість заперечних слів та результат попереднього етапу віднімається від кількості позитивних слів. Якщо загальний настрій перевищує нуль, фрагмент тексту позитивний; якщо воно менше нуля, фрагмент тексту вважається негативним; у випадку нульового результату текст нейтральний.

Цей метод не вимагає жодних конкретних інструментів попередньої обробки тексту, але результати такого аналізу буде кращим, ніж у підході, що заснований на словниках, оскільки в цьому випадку беруться до уваги слова-інвертори та протилежні речення.

Отже, підхід, що ґрунтується на правилах, до аналізу настроїв справді трудомісткий, оскільки для ефективної роботи аналізатора потрібна значна кількість написаних вручну правил, але при застосуванні у межах певного домену, такий підхід може забезпечити точність понад 90%.

Підхід з на основі машинного навчання став популярним протягом останніх кількох років. Реалізація аналізу емоційного забарвлення з використанням алгоритмів машинного навчання передбачає навчання класифікатора машинного навчання на сентиментативному кодуванні, а потім використання отриманої моделі для аналізу нових текстів.

Даний підхід вимагає коментованого корпусу даних (датасету). Точність аналізатора в значній мірі залежить від якості та величини коментованого корпусу.

Машинне навчання дозволяє застосовувати різні алгоритми класифікації для визначення емоційного забарвлення тексту.

Процес реалізації аналізу настроїв із використанням машинного навчання можна розділити на наступні етапи:

– коментування корпусу для класифікатора;

– подання кожного коментованого огляду у вигляді вектора ознак (об’єкти можуть бути представлені словами, n-грамами, оточуючими словами або навіть розділовими знаками; слова можуть або не бути перетворені на їх початкова форма);

– вибор алгоритму класифікації та навчання класифікатора.

Класифікатор перетворює текст в числове представлення, зазвичай вектор. Кожний компонент вектора є частотою слова або виразу в анотованому корпусі даних. Цей процес відомий як екстракція ознак або векторизація тексту.

Отже, алгоритми машинного навчання є ефективним інструментом для аналізу емоційного забарвлення тексту.

1.4 Метрики аналізу емоційного забарвлення тексту

Для визначення ефективності класифікатора і встановлення рівня точності моделі аналізу емоційного забарвлення часто використовується перехресна перевірка.

Перехресна перевірка включає поділ даних на тренувальні (80% даних) і тестові (20% даних), використання тренувальних даних для навчання моделі класифікатора та його перевірку під час тестування для отримання показників продуктивності. Процес повторюється кілька разів і обчислюється середнє значення для кожної метрики.

Основою перевірки є тестова вибірка, в якій встановлено відповідність між документами та їх класами.

При наявності тестової виборки перевіряється результат, наданий класифікатором для документів цієї тестової виборки, і далі виконується співвідношення рішення класифікатора з відомим правильним рішенням. Але для того, щоб приймати рішення щодо ефективності роботи алгоритму, необхідна чисельна метрика його якості.

У найпростішому випадку метрикою чисельної оцінки алгоритму може бути точність (accuracy), що являє собою частку документів, за якими класифікатор прийняв правильне рішення.

$$A = \frac{P}{N}, \tag{1.1}$$

де A – точність;

P – кількість документів за якими класифікатор прийняв правильне рішення;

N – розмір навчальної вибірки.

Проте, у цій метриці є одна особливість, яку необхідно враховувати. Вона надає всім документам однакову вагу, що може бути некоректно у разі, якщо розподіл документів у навчальній вибірці сильно зміщений в бік одного класу. В цьому випадку у класифікатора є більше інформації про цей класа і, відповідно, в рамках цього класу він буде прийматиме більш адекватні рішення. На практиці це призводить до того, що значення точності високе, але при цьому в рамках якогось конкретного класу класифікатор працює погано.

Рішення цієї проблеми полягає у використанні збалансованого набору даних (датасету).

Точність (precision) і повнота (recall) є метриками, що використовуються при оцінці алгоритмів вилучення інформації. Іноді вони використовуються самі по собі, іноді в якості базису для похідних метрик, таких як F-міра або R-Precision.

Система зберігає інформація про те, скільки разів за документами заданого класу прийняте вірне і скільки разів невірне рішення:

$$Precision_a = \frac{TP}{TP+FP} \tag{1.2}$$

$$Precision_b = \frac{TN}{TN+FN} \tag{1.3}$$

$$Recall_a = \frac{TP}{TP+FN} \tag{1.4}$$

$$Recall_b = \frac{TN}{TN+FP} \tag{1.5}$$

де $Precision_a$ – точність позитивний рішень;

$Precision_b$ – точність негативних рішень;

$Recall_a$ – повнота позитивних рішень;

$Recall_b$ – повнота негативних рішень;

TP – істинно позитивні рішення;

TN – істинно негативні рішення;

FP – помилково позитивні рішення;

FN – помилково негативні рішення.

Точність (precision) системи в межах класу – це частка документів, що істинно належать даному класу, щодо всіх документів які система віднесла до цього класу. Повнота системи – це частка знайдених класифікатором документів, що належать класу, щодо всіх документів цього класу в тестовій вибірці [25].

Зрозуміло, що чим вище точність і повнота, тим краще. Але в реальному житті максимальна точність і повнота недосяжні одночасно і доводиться шукати

якийсь баланс. Тому хотілося б мати якусь метрику, яка об'єднувала б у собі інформацію про точність та повноту алгоритму. Саме такою метрикою є F-міра.

F-міра – це гармонійне середнє між точністю і повнотою. Вона прагне до нуля, якщо точність або повнота прагне до нуля.

$$F = 2\frac{Precision \times Recall}{Precision + Recall} \quad (1.6)$$

де *Precision* – точність;

Recall – повнота.

Дана формула надає однакову вагу точності і повноти, тому F-міра буде падати однаково при зменшенні і точності і повноти. Можливо розрахувати F-міру надавши різну вагу точності і повноти, якщо ви свідомо віддаєте пріоритет однієї з цих метрик при розробці алгоритму.

$$F_\beta = (1 + \beta^2)\frac{Precision \times Recall}{\beta^2 Precision + Recall} \quad (1.7)$$

β приймає значення в діапазоні $0 < \beta < 1$, якщо точність має пріоритет, а при $\beta > 1$ пріоритет віддається повноті. При $\beta = 1$формула зводиться до попередньої і ми отримуємо збалансована F-міру (також її називають F1-мірою) .

1.5 Програмне забезпечення для аналізу емоційного забарвлення тексту

Технології обробки природної мови активно розвиваються. Розроблена низка програмного забезпечення у сфері обробки природної мови, в тому числі для аналізу емоційного забарвлення тексту.

У таблиці 1.2 наведена порівняльна характеристика основних програм для аналізу емоційного забарвлення тексту.

Таблиця 1.2 – Порівняльна характеристика програм для аналізу тональності тексту

Назва	Метод	Мова	Ліцензія	Платформа
Sentiment140	Машинне навчання	Англійська, іспанська	Комерційна	Вебсервіс
Eureka Engine	Машинне навчання	Російська	Комерційна	Вебсервіс
TextBlob	Машинне навчання	Англійська	MIT	Бібліотека Python
RCO	Правила	Російська	Комерційна	Windows
DictaScope	Правила	Російська	Комерційна	Windows
Pattern	Правила	Англійська, іспанська, німецька, французька	BSD	Бібліотека Python
MonkeyLearn	Правила	Англійська	Комерційна	Вебсервіс
Lexalytics	Машинне навчання	Англійська	Комерційна	Вебсервіс
Brandwatch	Машинне навчання	Англійська	Комерційна	Вебсервіс
Brand24	Правила	Англійська	Комерційна	Вебсервіс

З таблиці 1.2 видно, що більшість програмного забезпечення є комерційним. Також більша частина існуючих програмних рішень розроблена для англійської та інших мов. Більшість з них націлені на моніторинг думок про товари, послуги та бренди, а решта аналізують дані лише окремими реченнями. Тому необхідність дослідження за темою даної роботи визначається тим, що поки не існує доступних автоматизованих систем оцінки емоційного забарвлення текстів українською мовою.

1.6 Постановка задач дослідження

Метою роботи є виявлення найбільш ефективних методів аналізу емоційного забарвлення тексту українською мовою та впровадження цих методів.

Враховуючи сучасні можливості машинного навчання, було прийнято рішення досліджувати методи аналізу емоційного забарвлення тексту за допомогою машинного навчання.

Для досягнення мети необхідно виконати наступні задачі:

– дослідити методи та інструменти аналізу емоційного забарвлення тексту з використанням машинного навчання та розглянути існуючи датасети (набори даних), що можуть бути використані у машинному навчанні;

– сформувати набір даних (датасет) українською мовою для домену відгуків клієнтів;

– провести експериментальне дослідження, з розробкою відповідного програмного коду, щодо ефективності роботи різних алгоритмів машинного навчання для виконання задачі бінарної класифікації тексту українською мовою з використанням сформованого набору даних;

– побудувати модель класифікатора на основі попередньо навченої багатомовної моделі машинного навчання BERT, адаптовану до виконання задачі бінарної класифікації текстів українською мовою, враховуючи сформований набір даних;

– реалізувати клієнт-серверну програмну систему для аналізу емоційного забарвлення тексту українською мовою.

Прийнято рішення реалізувати вебдодаток у якості клієнтської частини клієнт-серверної програмної системи.

2 МЕТОДИ ТА ІНСТРУМЕНТИ АНАЛІЗУ ЕМОЦІЙНОГО ЗАБАРВЛЕННЯ ТЕКСТУ ЗА ДОПОМОГОЮ МАШИННОГО НАВЧАННЯ

2.1 Дослідження наборів даних (датасетів)

Важливою частиною для виконання аналізу емоційного забарвлення тексту є робота з набором даних. Оскільки одні і ті ж слова і словоформи можуть бути використані з різною тональністю, бажано, щоб дані були віднесені до певного домену, наприклад відгуки покупців про товар, коментарі політичних новин тощо.

Пошук та аналіз датасетів показав, що для української мови у вільному доступі не існує великого датасету, що підходив би для виконання аналізу емоційного забарвлення текстів. Датасети представлені здебільшого англійською мовою. Серед англомовних датасетів можна виділити наступні.

Набір даних, що містить кілька мільйонів відгуків клієнтів про продукти Amazon.

Набір даних від IMDB, що складається з 5331 позитивних і 5331 негативних оброблених відгуків про фільми [18].

Набір даних з відгуками про продукти харчування – цей набір даних складається з приблизно 500 тис. відгуків про продукти харчування від Amazon. Цей набір даних містить інформацію про продукт і користувача, рейтинги та звичайну текстову версію кожного огляду.

Відгуки про авіакомпанії на Kaggle: цей набір складається з приблизно 15 тис. позначених твітів (позитивний, нейтральний і негативний) про авіакомпанії.

Для української мови складений тональний словник української мови, що містить 3442 слів української мови, які мають не нейтральну тональність (-2, -1, 1, 2). В словнику слова приведені до базової граматичної форми, а також прислівники замінені на спільнокореневі прикметники. Дані експертних оцінок надані: Олександром Маріковським та В’ячеславом Тихоновим. Розширений словник підготовлений: Сергій Шеховцов, Олесь Петрів, Дмитро Чаплинський, Всеволод Дьомкін [19].

Тональний словник української мови містить тільки слова, але не речення. Більш того, слова цього словника не мають прив'язки до певного домену, що робить даний словник не оптимальним для використання при вирішенні задачі з виявлення емоційного забарвлення тексту.

З огляду із наведеного вище, постає задача збору датасету українською мовою, що містив би речення, що відносяться до конкретного домену.

2.2 API для аналізу емоційного забарвлення тексту

Існує багато інструментів для аналізу емоційного забарвлення тексту, які можна використовувати через API.

Python є однією з провідних мов програмування для науки про дані. Вона має розвинену спільноту і великий набір варіантів для реалізації моделей обробки природної мови..

Scikit-learn – один з найбільш широко використовуваних пакетів Python для машинного навчання. Він дозволяє виконувати безліч операцій і надає безліч алгоритмів. Scikit-learn також пропонує відмінну документацію про свої класи, методи та функції, а також опис використовуваних алгоритмів.

NLTK – це провідна платформа для побудови програм з використанням Python для роботи з даними природної мови. Вона надає прості у використанні інтерфейси для понад 50 корпусів та лексичних ресурсів, таких як WordNet, а також набір бібліотек обробки тексту для класифікації, токенізації, стемінгу, тегування, синтаксичного аналізу та семантичних міркувань, обгортки для бібліотек NLP [20].

TensorFlow – це відкрита платформа для машинного навчання. Він має всеосяжну, гнучку екосистему інструментів, бібліотек та ресурсів спільноти, що дозволяє дослідникам впроваджувати найсучасніші технології машинного навчання, а розробникам – легко створювати та розгортати додатки, що працюють з використанням машинного навчання.

PyTorch – бібліотека для машинного навчання з є відкритим вихідним кодом, що використовуються для додатків таких спрямувань, як комп’ютерний зір і обробка природної мови.

Transformers – бібліотека для обробки природних мов для всіх моделей машинного навчання, за підтримки таких бібліотек, як Flair , Asteroid , ESPnet , Pyannote та багато інших.

Для порівняння різних алгоритмів класифікаторів і для створення моделі може бути використана бібіліотека ML.NET, що створена для вирішення задач машинного навчання в екосистемі .NET. ML.NET має вражаючи результати. Так, використовуючи набір даних огляду Amazon розміром 9 ГБ, ML.NET навчив модель аналізу настроїв із 95% точністю [21].

Для кожної задачі машинного навчання, в тому числі і для задачі бінарної класифікації, існує кілька можливих алгоритмів навчання. Вибір конкретного алгоритму визначається проблемою, яка вирішується, характеристиками даних, а також доступними обчислювальними ресурсами і ресурсами зберігання. Важливо відзначити, що навчання моделі машинного навчання – це ітеративний процес. Може знадобитися спробувати кілька алгоритмів, щоб визначити найкращий з них.

Алгоритми працюють на базі ознак. Ознаки – це числові значення, що обчислюються на основі вхідних даних. Вони є оптимальним вхідними даними для алгоритмів машинного навчання. Ви перетворювати необроблені вхідні дані в ознаки, використовуючи одне або кілька перетворень даних . Наприклад, текстові дані перетворюються в набір з числа слів і числа сполучень слів. Після вилучення ознак з необроблених даних за допомогою перетворень даних вони вважаються певними ознаками.

Алгоритм – це математичний опис, що використовується для створення моделі. Різні алгоритми дають моделі з різними характеристиками.

Розглянемо алгоритми, що запропоновані бібілотекою ML.NET.

У ML.NET один алгоритм можна застосувати до різних завдань. Наприклад, стохастичний подвійний покоординатно підйом можна використовувати для

бінарної класифікації, багатокласової класифікації та регресії. Різниця полягає в інтерпретації вихідних даних алгоритму для зіставлення із завданням.

Для кожного поєднання алгоритму і завдання ML.NET надає компонент, який виконує алгоритм навчання і здійснює інтерпретацію. Такі компоненти називаються навчальними алгоритмами. Наприклад, SdcaRegressionTrainer використовує алгоритм StochasticDualCoordinatedAscent , що застосовується до задачі регресії.

Лінійні алгоритми створюють модель, яка обчислює оцінки на базі лінійного поєднання вхідних даних і набору вагових коефіцієнтів. Вагові коефіцієнти – це параметри моделі, які оцінюються під час навчання.

Лінійні алгоритми добре підходять для ознак, які є лінійно сепарабельними.

Перед навчанням за допомогою лінійного алгоритму потрібно нормалізувати ознаки. Це не дозволяє одній ознаці чинити більший вплив на результат порівняно з іншими ознаками.

У загальному випадку лінійні алгоритми є масштабованими і швидкими, а також не вимагають великих витрат на навчання і прогнозування. Вони масштабуються за кількістю ознак і за розміром набору даних для навчання.

Лінійні алгоритми роблять кілька проходів за даними для навчання. Якщо набір даних вміщується в пам’ять, то додавання контрольної точки кеша в конвеєр ML.NET перед додаванням навчального алгоритму прискорить навчання [22].

Лінійні навчальні алгоритми наведені в таблиці 2.1.

Таблиця 2.1 – Лінійні навчальні алгоритми

Алгоритм	Властивості
Усереднений персептрон	Добре підходить для класифікації тексту.
Стохастичний подвійний покоординатний підйом	Не має потреби в налаштуванні для забезпечення високої продуктивності.
L-BFGS	Використовується при великій кількості ознак. Створює статистику навчання логістичної регресії, але масштабується не так добре, як AveragedPerceptronTrainer.
Посимвольний стохастичний градієнтний спуск	Найшвидший лінійний навчальний алгоритм бінарної класифікації. Добре масштабується за кількістю процесорів.

Алгоритми дерева прийняття рішень створюють модель, яка містить ряд рішень: по суті, блок-схему для значень даних.

Для використання цього типу алгоритму не потрібні лінійно масштабовані ознаки. Крім того, ознаки не потрібно нормалізувати, оскільки окремі значення у векторі ознак використовуються незалежно в процесі прийняття рішень.

Алгоритми дерева прийняття рішень зазвичай дуже точні.

За винятком узагальнених адитивних моделей (GAM), моделі дерева можуть бути недостатню пояснювальними, коли число ознак велике.

Алгоритми дерева прийняття рішень використовують більше ресурсів і гірше масштабуються в порівнянні з лінійними алгоритмами. Вони добре підходять для наборів даних, що містяться в пам’яті.

Розширені дерева прийняття рішень – це сукупність невеликих дерев, де кожне дерево оцінює вхідні дані і передає результат наступному дереву для уточнення оцінки. Тобто кожне наступне дерево покращує результат попереднього.

Таблиця 2.2 – Навчальні алгоритми дерев прийняття рішень

Алгоритм	Властивості
Машина слабкого градієнтного бустінга	Найшвидший і точний з навчальних алгоритмів дерев бінарної класифікації. Високі можливості налаштування.
Швидке дерево	Використовується для даних зображення з певними ознаками. Стійкий до незбалансованих даних. Високі можливості налаштування.
Швидкий ліс	Дуже добре підходить для даних з високим рівнем шуму.
Узагальнена аддитивна модель (GAM)	Підходить для задач, де добре справляються алгоритми дерева, якщо пояснюваність є пріоритетним завданням.

Також ML.NET надає можливість використання алгоритму факторизації матриці, що підходить для розріджених категоріальних даних з великими наборами даних.

Окрім побудови моделей класифікатора з використанням різних алгоритмів слід розглянути можливість використання попередньо навчених моделей нейронних мереж.

2.3 Попередньо навчені моделі нейронних мереж

Попередньо навчена модель – це модель, яка раніше була навчена на великому наборі даних. Така модель може бути використана без доналаштування або може бути застосоване трансферне навчання для виконання іншої задачі.

Трансферне навчання передбачає повторне використання попередньо навченої моделі для нової проблеми. У процесі трансферного навчання машина використовує знання, отримані з попередньої задачі.

Трансферне навчання може навчити глибокі нейронні мережі із порівняно невеликими даними. Це дуже корисно в галузі науки про дані, оскільки більшість реальних проблем, як правило, не мають великих прокоментованих корпусів даних для підготовки таких складних моделей.

Для аналізу емоційного забарвлення тексту можуть бути використані багатоцільові моделі для обробки природної мови. Ці моделі застосовуються для вирішення різних завдань, таких як машинний переклад, системи відповіді на запитання, чат-боти, аналіз настроїв тощо. Основний компонент цих багатоцільових моделей обробки природних мов – концепція мовного моделювання [23].

Найвідомішими є наступні попередньо навчені моделі:

– ULMFiT – це метод трансферного навчання, який можна застосувати до завдань NLP. Модель попередньо навчена на тексті Вікіпедії [23]. Ця модель свого часу покращила масштаби глибокого навчання в процесі обробки природної мови, зробивши можливим підготовку моделей для різних завдань за значно менший час, ніж раніше;

– ELMo – це тип глибокого контекстуалізованого подання слів, що моделює як складні характеристики вживання слів (наприклад, синтаксис та семантика), так і те, як ці вживання змінюються залежно від мовного контексту (тобто модель багатозначності);

– BERT (Bidirectional Encoder Representations from Transformers) –попередньо навчена модель від Google, яка містить результати виконання багатьох задач обробки природної мови, в тому числі аналізу емоційного забарвлення тексту [27].

У ході проведення даного дослідження буде використана модель BERT.

BERT означає двоспрямоване кодування представлення від трансформерів [28] та на відміну від моделей спрямованості, які послідовно зчитують введення тексту (зліва направо або справа наліво), кодер Transformer зчитує всю послідовність слів відразу. Тому він вважається двонаправленим, хоча було б точніше сказати, що він неспрямований. Ця характеристика дозволяє моделі вивчати контекст слова на основі всіх його оточень (ліворуч і праворуч від слова).

BERT-модель була попередньо навчена на великому багатомовному корпусі даних, включаючи українську мову.

В даний час BERT використовується в Google для оптимізації інтерпретації пошукових запитів користувачів, для задач генерації мови таких як: відповідь на запитання, абстрактне узагальнення, генерація розмовної відповіді. Також BERT використовується для вирішення задач на розуміння природної мови, таких як: визначення неоднозначночті сенсу слова, класифікація тональності тексту. BERT є відкритим кодом і доступний до використання.

3 ПРАКТИЧНА РЕАЛІЗАЦІЯ МЕТОДІВ АНАЛІЗУ ЕМОЦІЙНОГО ЗАБАРВЛЕННЯ ТЕКСТУ

3.1 Формування набору даних (датасету) українською мовою

Набір даних (датасет) грає важливу роль у машинному навчанні. У ході дослідження було сформовано датасет відгуків користувачів про мобільні додатки в категорії «Покупки» з платформи Google Play.

Для генерування датасету використовувалися наступні технології:

– мова програмування Python;

– бібліотека google-play-scraper для збору відгуків про мобільні додатки з платформи Google Play;

– бібліотека pandas для здійснення маніпуляцій з набором даних – представлення датасету у вигляді таблиці та формування файлу з даними, що готовий для завантаження;

– бібліотека sklearn для розбиття датасету на набір для тренування та набір для тестування.

Для отримання відгуків була використана бібіліотека google-play-scraper для мови Python. API цієї бібліотеки приймає ідентифікатор додатка, відгуки для якого треба отримати, двобуквенний код мови, в якому потрібно завантажити сторінку додатка та двобуквенний код країни, який використовується для отримання додатків (потрібно, коли програма доступна лише в деяких країнах).

Метою було отримати збалансований набір даних – з однаковою кількістю позитивних та негативних відгуків та з репрезентативними відгуками для кожного додатка.

Для фільтрування оцінки відгуків було використано опцію бібліотеки google-play-scraper. Відгуки були відсортовано за їхнею корисністю – це ті відгуки, які Google Play вважає найважливішими. Також була отримана підмножина найновіших відгуків (відсортовано за датою додавання).

Із вмісту кожного відгуку було видалено смайлики, символи транспорту, прапори та інші символи, що не є текстом. Також з датасету було видалено відгуки, що не містять тексту.

Загалом було отримано 10 276 відгуків, які мають наступні оцінки (рис. 3.1).

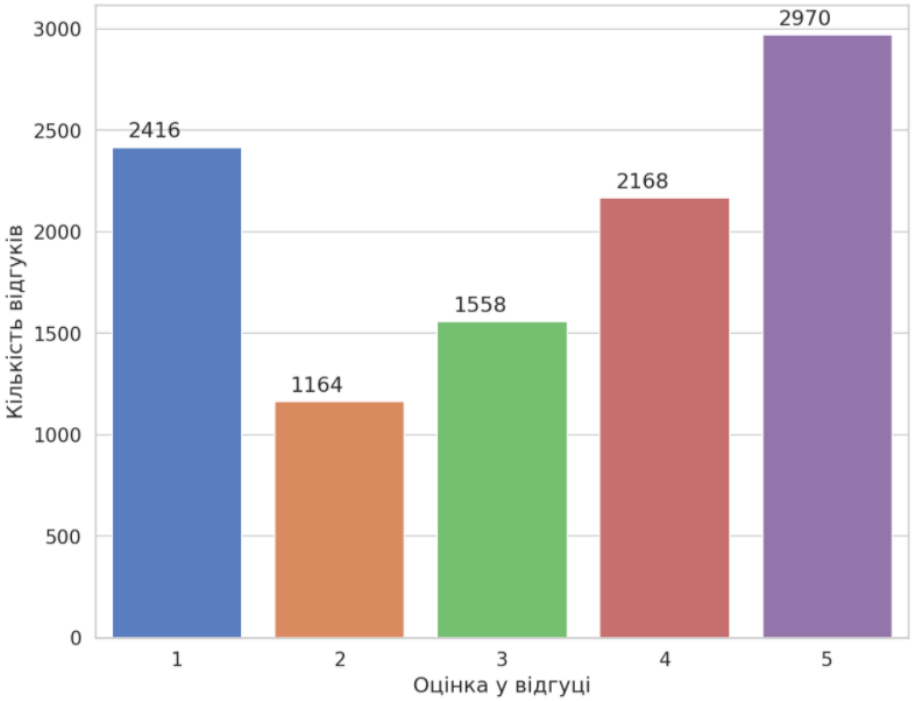

Рисунок 3.1 – Зібрані відгуки про мобільні додатки

Для задачі бінарної класифікації тональності тексту відгуки було розділено на дві категорії: ті, що мають негативну тональність, та ті, що мають позитивну тональність. Для позначення відповідних категорій було введено додатковий стовбець «sentimentScore», що має одне з двох можливих значень – 0 (текст з негативним емоційним забарвленням) або 1 (текст з позитивним емоційним забарвленням).

Відгуки з оцінкою 4 та 5 за п’ятибальною шкалою були позначені як позитивні, відгуку з оцінкою 1-3 – як негативні.

З рис. 3.2 видно, що було отримано збалансований датасет (з рівною кількістю текстів з позитивною та негативною тональністю).

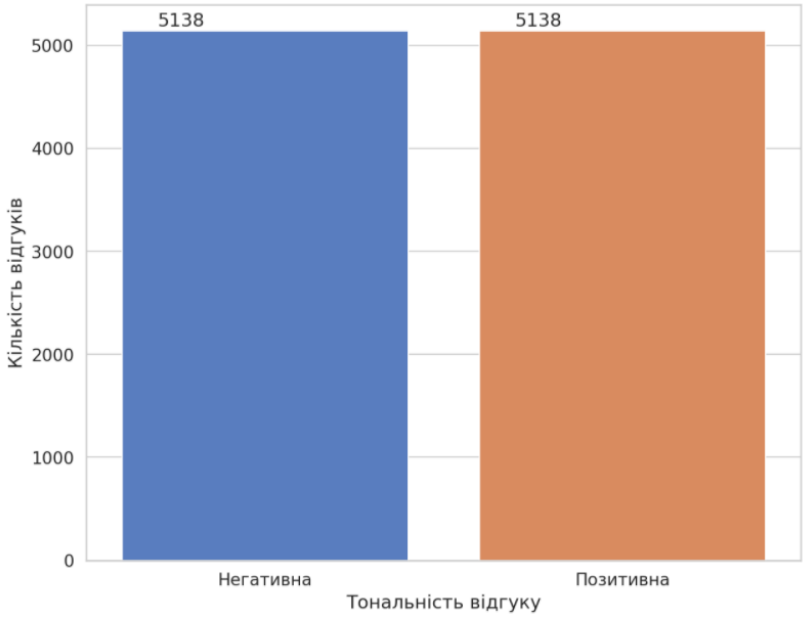

Рисунок 3.2 – Кількість відгуків в датасеті для задачі бінарної класифікації

Для отримання найкращих результатів навчання бінарної класифікації навчальні дані мають бути збалансовані (тобто число позитивних і негативних навчальних даних має бути однаковим). Відсутні значення необхідно обробити до навчання [26].

Отриманий датасет був розділений на два набори: набір даних для тренування моделі та набір даних для тестування моделі (80% та 20% відповідно) та використаний для побудови моделей класифікатора в даному дослідженні.

3.2 Проведення експериментального дослідження з виявлення оптимального алгоритму аналізу емоційного забарвлення тексту українською мовою та побудова моделі класифікатора

За допомогою бібліотеки ML.NET проведено експеримент з виявлення оптимального алгоритму для виконання задачі бінарної класифікації тексту українською мовою на прикладі відгуків користувачів про мобільні додатки.

Бінарна класифікація – задача контрольованого машинного навчання, яка прогнозує розподіл елементів даних за двома класами (категоріям). На вхід алгоритму класифікації подається набір прикладів з мітками, кожна з яких представляє собою ціле число 0 або 1. Результатом роботи алгоритму бінарної класифікації є класифікатор, який вміє прогнозувати клас для нових екземплярів без мітки.

Для проведення експерименту використовувалися наступні алгоритми, для яких присутній API бібліотеки ML.NET [24]:

– алгоритм лінійної бінарної класифікації з використанням середнього перцептрона (AveragedPerceptronTrainer);

– алгоритм бінарної логістичної регресії за допомогою стохастичного методу подвійних координат (SdcaLogisticRegressionBinaryTrainer);

– алгоритм на основі використання стохастичного градієнтного спуску (SymbolicSgdLogisticRegressionBinaryTrainer);

– алгоритм на основі лінійної моделі логістичної регресії (LbfgsLogisticRegressionBinaryTrainer);

– алгоритми на основі бінарних дерев прийняття рішень (FastTreeBinaryTrainer, FastForestBinaryTrainer, LightGbmBinaryTrainer);

– алгоритм бінарної класифікації з узагальненими адитивними моделями (GamBinaryTrainer);

– алгоритм для прогнозування цільового об’єкту за допомогою моделі лінійної бінарної класифікації, що навчена методом опорних векторів SVM (LinearSvmTrainer).

Для отримання найкращих результатів навчання бінарної класифікації навчальні дані мають бути збалансовані, тобто число позитивних і негативних навчальних даних має бути однаковим. Відсутні (пусті) значення необхідно обробити до навчання.

Для проведення експерименту використаний зібраний нами датасет, який відповідає наведеним вище критеріям.

Дані були розділені на дві категорії: датасет для тренування моделі (80% даних) та датасет для тестування (20% даних).

Схема роботи класифікатора наведені на рис. 3.3:

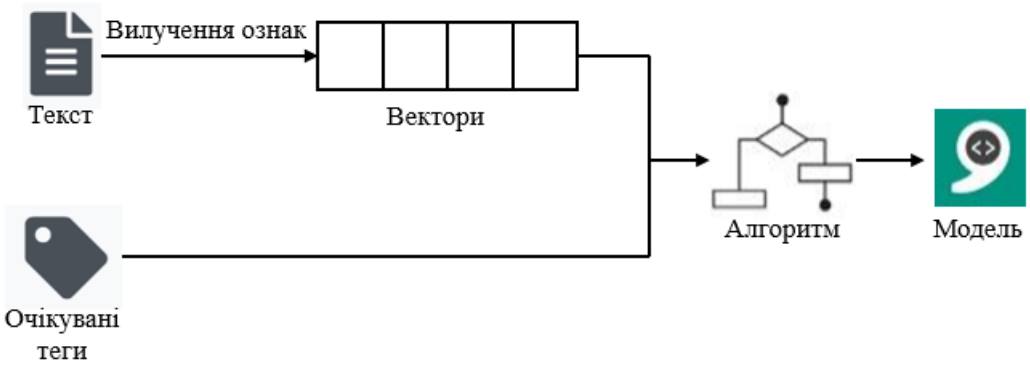

Рисунок 3.3 – Схема роботи класифікатора

По-перше, класифікатор здійснює вилучення ознак: метод використовується для перетворення кожного тексту в числове представлення у вигляді вектора. Далі на вхід алгоритму надаються пари зі створених векторів та набору тегів. На виході утворюється моделі класифікації.

Після навчання з достатньою кількістю навчальних зразків, модель машинного навчання може починати робити точні прогнози. Цей же векторизатор використовується для перетворення нових текстів на набори векторів, які можна подавати до моделі класифікації для отримання прогнозів. Тобто на вхід отриманій моделі надається набір векторів, а на вихід повертається спрогнозовані теги.

Результати експерименту (найкращі результати для кожного алгоритму) наведено в таблиці 3.1.

Як видно з таблиці 3.1, найкращі результати для задачі бінарної класифікації тексту українською мовою має модель SdcaLogisticRegressionBinary, в основу якої покладено алгоритм логістичної регресії з використанням стохастичного методу здвоєнних координат.

Таблиця 3.1 – Результати виконання задачі бінарної класифікації тексту українською мовою для різних алгоритмів бібліотеки ML.NET

Клас моделі бібліотеки ML.NET	Точність (accuracy)	AUC (площа під кривою)	AUPRC (площа під кривою «точність-повнота»)	F1-метрика
SdcaLogisticRegressionBinary	0.8806	0.9380	0.9366	0.8748
AveragedPerceptronBinary	0.8795	0.9344	0.9244	0.8713
SymbolicSgdLogisticRegression Binary	0.8702	0.9162	0.9108	0.8617
FastTreeBinary	0.8801	0.9374	0.9422	0.8704
SgdCalibratedBinary	0.8731	0.9295	0.9370	0.8703
LightGbmBinary	0.8681	0.9340	0.9222	0.8560
LinearSvmBinary	0.8445	0.9125	0.8979	0.8290
LbfgsLogisticRegressionBinary	0.8527	0.9130	0.9236	0.8544
FastForestBinary	0.7715	0.8482	0.8506	0.7306

Цей алгоритм заснований на методі стохастичного подвійного координатного підйому (SDCA), найсучаснішій методиці оптимізації опуклих цільових функцій.

Конвергенція забезпечується шляхом періодичного забезпечення синхронізації між первинними та подвійними змінними в окремому потоці. Також пропонується кілька варіантів функцій збитків (наприклад, логістична втрата). Залежно від використовуваних втрат, навчена модель може бути, наприклад, машиною з підтримкою вектора або логістичною регресією. Метод SDCA поєднує в собі кілька найкращих властивостей, таких як можливість проведення потокового навчання (без розміщення всього набору даних у пам’яті), досягаючи розумного результату за допомогою декількох сканувань всього набору даних.

SDCA – це стохастичний та потоковий алгоритм оптимізації. Результат залежить від порядку навчальних даних. В умовах сильноопуклої оптимізації оптимальне рішення є унікальним, і тому всі зрештою досягають того самого місця. Навіть у не сильно опуклих випадках, ви отримаєте однаково хороші рішення.

Цей алгоритм використовує емпіричну мінімізацію ризику (тобто ERM) для формулювання проблеми оптимізації, побудованої на зібраних даних. Емпіричний ризик зазвичай вимірюється шляхом застосування функції збитків до прогнозів моделі щодо зібраних точок даних. Якщо навчальні дані не містять достатньо точок

даних (наприклад, для навчання лінійної моделі в n-вимірному просторі, нам потрібно принаймні n точок даних), може статися «перенавчання» , так що модель, вироблена ERM, добре описує дані навчання, але може не спрогнозувати правильні результати в небачених подіях. Регуляризація є загальноприйнятою методикою полегшення такого явища шляхом зміни цієї величини (як правило, вимірюється функцією норми) параметрів моделі.

Математично логістична регресія може бути описана наступним чином [25].

Нехай є деяка випадкова величина y, що може набувати лише двох значень (0 або 1). Нехай ця величина залежить від деякої множини змінних $x_1, \dots x_n$, на основі значень яких потрібно обчислити ймовірність прийняття того чи іншого значення залежної змінної.

Отже, нехай об’єкти задаються n числовими ознаками $f_i : X \to R,\ j = 1 \dots n$ і простір ознакових описів в такому випадку $X = R^n$. Нехай Y – кінцева множина міток класів і задана навчальна вибірка пар «об’єкт-відповідь» в такому разі:

$$X^m = \{(x_1, y_1), \dots, (x_m, y_m)\} \tag{3.1}$$

Розглянемо випадок двох класів: $Y = \{-1, +1\}$. У логістичної регресії будується лінійний алгоритм класифікації $a: X \to Y$ виду:

$$a(x, \omega) = sign\left(\sum_{j=1}^{n} \omega_j f_j(x) - \omega_0\right) = sign\langle x, \omega\rangle \tag{3.2}$$

де ω_j – вага j-ої ознаки;

ω_0 – поріг прийняття рішення;

$\omega = (\omega_0, \dots, \omega_{n)}$ – вектор ваг;

$\langle x, \omega\rangle$ – скалярний добуток ознакового опису об'єкта на вектор ваг.

Передбачається, що штучно введена нульова ознака: $f_0(x) = -1$.

Завдання навчання лінійного класифікатора полягає в тому, щоб за вибіркою X^m налаштувати вектор ваг ω. У логістичній регресії для цього вирішується завдання мінімізації емпіричного ризику з функцією втрат спеціального виду:

$$Q(\omega) = \sum_{i=1}^{m} \ln(1 + \exp(-y_j \langle x_j, \omega \rangle)) \to min_{\omega} \quad (3.3)$$

Після того, як рішення ω знайдено, стає можливим не тільки обчислювати класифікацію $a(x) = sign\langle x, \omega \rangle$ для довільного об'єкта x, але й оцінювати апостеріорні ймовірності його приналежності класам:

$$\mathbb{P}\{y|x\} = \sigma(y\langle x, \omega \rangle),\ y \in Y, \quad (3.4)$$

де $\sigma(z) = \frac{1}{1+e^{-z}}$ – сігмоїдна функція.

Побудовану модель з використанням алгоритму логістичної регресії було дотреновано і отримані наступні значення метрик (таблиця 3.2):

Таблиця 3.2 – Метрики побудованої моделі класифікатора з використанням алгоритму логістичної регресії SDCA

Метрика	Опис метрики	Отримане значення	Еталонне значення
Точність (accuracy)	Частка правильних прогнозів за допомогою перевірочного набору даних. Це співвідношення числа правильно вгаданих і загального числа прикладів вхідних даних. Ця метрика працює добре, якщо існує аналогічна кількість вибірок, що належать кожному класу.	0.8820	Чим ближче до 1, тим ефективнішим вважається класифікатор. Точне значення 1 говорить про проблеми (зазвичай це витік міток і цілей, перенавчання або тестування за допомогою навчальних даних). Якщо тестові дані не збалансовані (більшість примірників відноситься до одного з класів), набір даних малий або оцінка підходить до значення 0 або 1, то точність не відображає фактичну ефективність класифікатора і вам потрібно перевірити додаткові метрики.
AUC (area under curve)	Площа під кривою оцінює площу під кривою, створеної підсумовуванням частот істинно позитивних результатів і помилково позитивних результатів.	0.9371	Чим ближче до 1, тим більша якість моделі. Для того, щоб модель була допустима, її значення повинно бути більше 0.5. Модель зі значенням AUC, що не перевищує 0.5, вважається непридатною.

Кінець таблиці 3.2

Метрика	Опис метрики	Отримане значення	Еталонне значення
AUPRC (area under precision recall curve)	Площа під кривою "точність - повнота" : зручна міра успішного прогнозу, коли класи розрізняються (вкрай нерівномірно розподілені набори даних).	0.9413	Чим ближче до 1, тим більш точні результати повертає класифікатор. Високий рівень оцінки, близький до 1, показує, що класифікатор повертає точні результати (висока точність), а також повертає більшу частину всіх позитивних результатів (високий рівень повноти).
Позитивна точність (positive precision)	Частка документів, що дійсно належать до класу документів з позитивною тональністю, щодо всіх документів які система віднесла до цього класу.	0.9000	Чим ближче до 1, тим якісніше працює класифікатор з виявлення позитивних документів. Ця метрика застосовується разом з іншою метрикою щодо вилучення інформації – позитивною повнотою.
Позитивна повнота (positive recall)	Частка правильно визначених класифікатором документів, що належать до класу позитивних прогнозів, щодо всіх документів цього класу в тестовій вибірці.	0.8600	Чим ближче до 1, тим більш повно класифікатор визначає позитивні документи.
Негативна точність (negative precision)	Частка документів, що дійсно належать до класу документів з негативною тональністю, щодо всіх документів які система віднесла до цього класу.	0.8700	Чим ближче до 1, тим якісніше працює класифікатор з виявлення негативних документів. Ця метрика застосовується разом з іншою метрикою щодо вилучення інформації – негативною повнотою.
Негативна повнота (negative recall)	Частка правильно визначених класифікатором документів, що належать до класу негативних прогнозів, щодо всіх документів цього класу в тестовій вибірці.	0.9009	Чим ближче до 1, тим більш повно класифікатор визначає негативні документи.
F1	Показник F1 також називається збалансованою F-оцінкою або F-мірою. Це середнє гармонійне значення точності і повноти. Показник F1 корисний в тому випадку, якщо необхідно знайти баланс між точністю і повнотою.	0.8807	Чим ближче до 1, тим більш точним є класифікатор.

Отже, враховуючи значення метрик, можна зробити висновок, що побудована модель класифікатора з використанням алгоритму логістичної регресії є якісною.

3.3 Використання багатомовної BERT-моделі для виявлення емоційного забарвлення тексту українською мовою

У ході дослідження було використано претреновану багатомовну модель BERT для задачі бінарної класифікації тексту українською мовою. Для цього було виконано наступні кроки:

– попередня обробка текстових даних для моделі BERT та побудова набору даних;

– використання навчання для побудови класифікатора емоційного забарвлення;

– оцінка моделі за тестовими даними.

BERT (Bidirectional Encoder Representations from Transformers) означає двоспрямоване кодування представлення від трансформерів [28].

На відміну від моделей спрямованості, які послідовно зчитують введення тексту (зліва направо або справа наліво), кодер Transformer зчитує всю послідовність слів відразу. Тому він вважається двонаправленим, хоча було б точніше сказати, що він неспрямований. Ця характеристика дозволяє моделі вивчати контекст слова на основі всіх його оточень (ліворуч і праворуч від слова).

BERT навчали, маскуючи 15% токенів з метою відгадати їх. Додатковою метою було передбачити наступне речення [27].

Зібраний в ході цього дослідження датасет був використаний для тренування моделі BERT для задачі бінарної класифікації тексту.

Бібліотека Transformers надає широкий спектр моделей Transformer (включаючи BERT). Вона працює з TensorFlow та PyTorch. Сюди також входять токенізатори.

В якості моделі використано bert-base-multilingual-uncased модель. Сама модель та токенизатор завантажено за допомогою бібліотеки Transformers.

Модель bert-base-multilingual-uncased є попередньо натренованою моделлю на 102 мовах з найбільшою Вікіпедією, включаючи українську.

У ході дослідження виконано наступні етапи:

– речення поділені на токени;

– додано спеціальні токени;

– обрано довжину речення.

BERT працює з послідовностями фіксованої довжини. У дослідженні було використано просту стратегію для вибору максимальної довжини. Збережено довжину символів кожного огляду і оцінена їхня довжина.

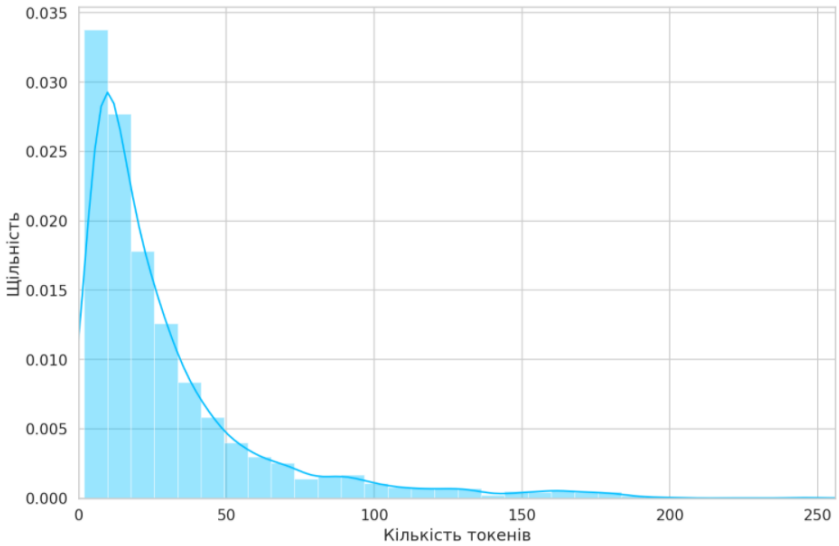

Рисунок 3.4 – Розподілення довжини токенів

Як видно з рис. 3.4, більшість відгуків містять менше 150 токенів. Для якісного налаштування моделі було встановлена довжина, що дорівнює 200 токенам.

Далі було побудовано модель нейронної мережі, проведено навчання нейронної мережі на тренувальному наборі даних та перевірено результат роботи моделі на тестових даних.

Щоб відтворити навчальну процедуру з документації BERT, було використано оптимізатор AdamW, наданий Hugging Face. Він виправляє зменшення ваги.

Для моделі необхідно вказати функцію втрат і оптимізатор для навчання.

Оскільки розв’язувана задача є прикладом бінарної класифікації та модель буде показувати ймовірність, то була використана функція втрат CrossEntropyLoss (пер. «Перехресна ентропія»).

Автори BERT мають кілька рекомендацій щодо налаштування моделі [28]:

– розмір партії (батчу): 16, 32;

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

– кількість епох (ітерацій): 2, 3, 4,…, 10.

Збільшення розміру партії значно скорочує час навчання, але дає меншу точність.

Після навчання вимірюються втрати і точність моделі шляхом перевірки на 20% зразків з перевірочного набору даних. Для перевірки моделі використовується функція, що отримує на вхід два масиви – тестові відгуки та відповідні маркування позитивного/негативного відгука. Результатом виконання цієї функції є пара чисел: відсоток втрат (loss; чим нижче це число, тим менше хибних прогнозів зробила нейронна мережа) та точність accuracy (рис. 3.5).

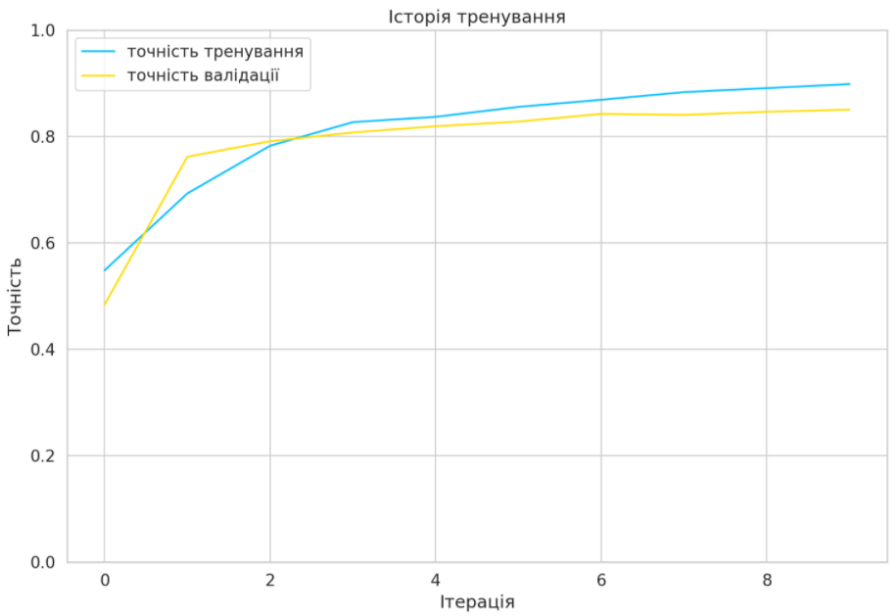

Рисунок 3.5 – Історія тренування моделі BERT на датасеті українською мовою

Після перехресного тестування оцінено модель. Результати оцінки наведені в таблиці 3.3.

Таблиця 3.3 – Метрики налаштованої та дотренованої моделі BERT для задачі бінарної класифікації тексту

Метрика	Опис метрики	Отримане значення	Еталонне значення
Точність (accuracy)	Частка правильних прогнозів за допомогою перевірочного набору даних. Це співвідношення числа правильно вгаданих і загального числа прикладів вхідних даних. Ця метрика працює добре, якщо існує аналогічна кількість вибірок, що належать кожному класу.	0.87	Чим ближче до 1, тим ефективнішим вважається класифікатор. Точне значення 1 говорить про проблеми (зазвичай це витік міток і цілей, перенавчання або тестування за допомогою навчальних даних). Якщо тестові дані не збалансовані (більшість примірників відноситься до одного з класів), набір даних малий або оцінка підходить до значення 0 або 1, то точність не відображає фактичну ефективність класифікатора і вам потрібно перевірити додаткові метрики.
Позитивна точність (positive precision)	Частка документів, що дійсно належать до класу документів з позитивною тональністю, щодо всіх документів які система віднесла до цього класу.	0.87	Чим ближче до 1, тим якісніше працює класифікатор з виявлення позитивних документів. Ця метрика застосовується разом з іншою метрикою щодо вилучення інформації – позитивною повнотою.
Позитивна повнота (positive recall)	Частка правильно визначених класифікатором документів, що належать до класу позитивних прогнозів, щодо всіх документів цього класу в тестовій вибірці.	0.88	Чим ближче до 1, тим більш повно класифікатор визначає позитивні документи.
Негативна точність (negative precision)	Частка документів, що дійсно належать до класу документів з негативною тональністю, щодо всіх документів які система віднесла до цього класу.	0.87	Чим ближче до 1, тим якісніше працює класифікатор з виявлення негативних документів. Ця метрика застосовується разом з іншою метрикою щодо вилучення інформації – негативною повнотою.

Кінець таблиці 3.3

Метрика	Опис метрики	Отримане значення	Еталонне значення
Негативна повнота (negative recall)	Частка правильно визначених класифікатором документів, що належать до класу негативних прогнозів, щодо всіх документів цього класу в тестовій вибірці.	0.85	Чим ближче до 1, тим більш повно класифікатор визначає негативні документи.
F1	Показник F1 також називається збалансованою F-оцінкою або F-мірою. Це середнє гармонійне значення точності і повноти. Показник F1 корисний в тому випадку, якщо необхідно знайти баланс між точністю і повнотою.	0.87	Чим ближче до 1, тим більш точним є класифікатор.

Отримані значення метрик після тренування та тестування моделі дозволяють зробити висновки, що дотренована багатомовна модель BERT є ефективною для задачі бінарної класифікації тексту українською мовою та характеризується високою повнотою і точністю з виявлення як тексту з позитивним забарвленням, так і тексту з негативним забарвленням.

3.5 Порівняння реалізацій моделей класифікатора

Для виявлення найякіснішої моделі класифікатора порівняємо значення ключових метрик для побудованих моделей логістичної регресії та BERT-моделі (таблиця 3.4).

Таблиця 3.4 – Порівняння значень метрик для моделі логістичної регресії та BERT-моделі

Метрика	Отримане значення		Порівняння
	Модель логістичної регресії	BERT-модель	
Точність (accuracy)	0.88	0.87	Отримані значення свідчать про те, що класифікатори обох моделей є ефективними, але класифікатор моделі логістичної регресії є трохи ефективнішим за класифікатор BERT-моделі.
Позитивна точність (positive precision)	0.94	0.87	Класифікатор моделі логістичної регресії якісніше працює з виявлення тексту з позитивною тональністю, ніж класифікатор моделі BERT.
Позитивна повнота (positive recall)	0.94	0.88	Класифікатор моделі логістичної регресії більш повно визначає текст з позитивною тональністю, ніж класифікатор моделі BERT.
Негативна точність (negative precision)	0.90	0.87	Класифікатор моделі логістичної регресії якісніше працює з виявлення тексту з негативною тональністю, ніж класифікатор моделі BERT.
Негативна повнота (negative recall)	0.86	0.85	Класифікатор моделі логістичної регресії більш повно визначає текст з негативною тональністю, ніж класифікатор моделі BERT.
F1	0.88	0.87	Метрика характеризує точність класифікатора. Обидві моделі мають високу точність, але модель логістичної регресії в даному випадку більш точна за BERT-модель.

Після тренування та тестування моделей з використанням зібранного нами датасету отримані значення метрик для моделі логістичної регресії та для BERT-моделі свідчать про високу якість обох моделей для виявлення емоційного забарвлення тексту українською мовою.

Проте класифікатор моделі логістичної регресії характеризується більшою точністю, ефективністю з виявлення текстів з позитивною та негативною тональністю і може вважатися більш якісним для виконання задачі бінарної класифікації текстів українською мовою.

3.4 Розробка програмного забезпечення для аналізу емоційного забарвлення тексту українською мовою

Огляд існуючого програмного забезпечення для здійснення аналізу емоційного забарвлення тексту показав актуальність створення системи автоматизованого аналізу емоційного забарвлення тексту українською мовою. Було вирішено розробити вебдодаток для проведення аналізу емоційного забарвлення тексту українською мовою. Додаток є у вільному доступі і не потребує автентифікації та авторизації користувача.

Для розробки програмного забезпечення було обрано: для імплементації Web API – платформу .NET (а саме фреймворк .NET 5) з використанням мови програмування C#; для імплементації клієнської частини – фреймворк Angular з використанням мови програмування TypeScript.

C# – це сучасна, об’єктно-орієнтована мова програмування загального призначення, розроблена корпорацією Microsoft. C# розроблений для Common Language Infrastructure (CLI), яка складається з виконуваного коду та середовища виконання, що дозволяє використовувати різні мови високого рівня на різних комп’ютерних платформах та архітектурах.

Мова програмування C# має наступні переваги:

– об’єктно-орієнтована мова;

– автоматичний збір сміття (об’єктів пам’яті, що вже не використовуються програмою);

– крос-платформеність;

– зворотна сумісність версій.

Фреймворк .NET 5 є найновішою версією фреймворку і містить останні вдосконалення і запровадження. В цілому, починаючи з версії фремворку ASP.NET Core 2.1, є можливість розробляти кросплатформені вебдодатки, що можуть бути розгорнуті на Windows, на Linux, на MacOS та в контейнерах. Завдяки впровадженню вебсервера Kestrel, ASP.NET є одним з найшвидших доступних

вебфреймворків. Новий вебсервер Kestrel є легким і швидким та використовує асинхронні моделі програмування.

Для реалізації клієнтської частини додатку було обрано фреймворк Angular та мову програмування TypeScript. Angular – це платформа та фреймворк для побудови односторінкових клієнтських програм (single page applications, SPA) за допомогою HTML та TypeScript.

TypeScript є обгорткою над мовою програмування JavaScript та має такі переваги, як:

– підтримка класів та модулів;

– статична перевірка типів;

– підтримка функцій ES6;

– подібність синтаксису до інших мов (Java, C#).

Для визначення емоційного забарвлення тексту в якості класифікатора було використано модель, що має найвищі показники ефективності та якості при дослідженні, опис якого наведено вище в даній роботі, а саме – модель логістичної регресії.

Бібліотека ML.NET дозволяє використати цю модель для класифікації тексту під час виконання програми, реалізованої на платформі .NET.

Користувач має можливість ввести текст, емоційне забарвлення якого має бути проаналізовано. Цей аналіз розглядається як задача бінарної класифікації, тобто текст буде віднесенний до одного з двох класів: «позитивний» чи «негативний».

Класифікація здійснюється з використанням моделі, побудованої в ході даного дослідження на основі згенерованого датасету текстів українською мовою для домену відгуків про мобільні додатки.

Інтерфейс розробленого додатку наведено на рис. 3.6.

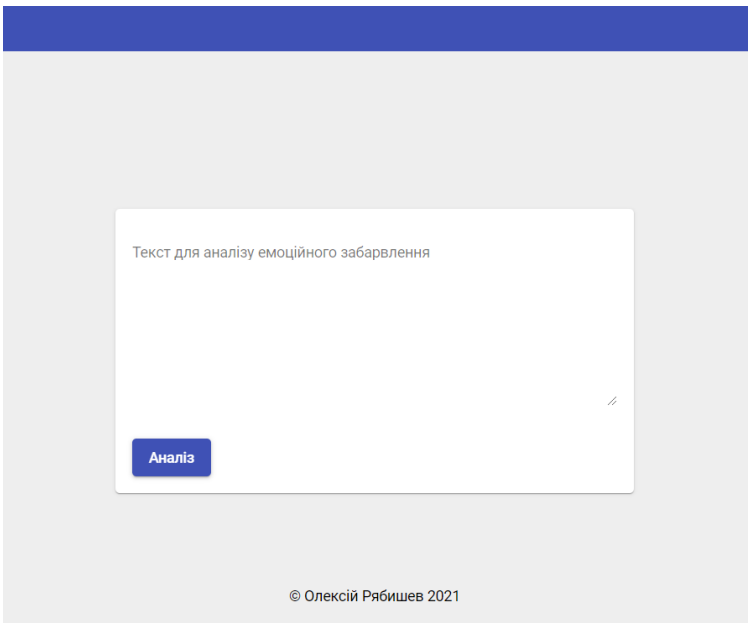

Рисунок 3.6 – Інтерфейс користувача вебдодатку

Побудовану модель було збережено в форматі zip і її може бути завантажено бібліотекою ML.NET у ході виконання програми для визначення емоційного забарвлення тексту.

Схема роботи класифікатора наведена на рис. 3.7.

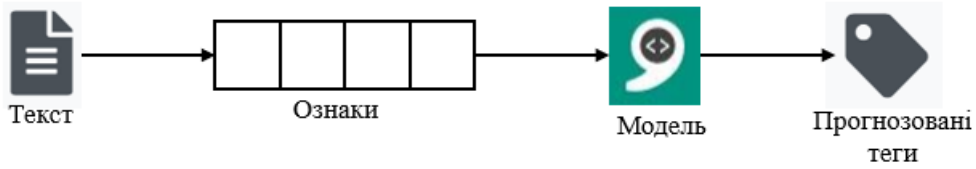

Рисунок 3.7 – Схема надання прогнозів класифікатором

Після навчання з достатньою кількістю навчальних зразків, отримана модель машинного навчання може починати робити прогнози. Векторизатор використовується для перетворення нових текстів на набори векторів, які можна подавати до моделі класифікації для отримання прогнозів. Тобто на вхід отриманій моделі надається набір векторів, а на вихід повертається спрогнозовані теги.

Діаграма процесу визначення емоційного забарвлення тексту через розробелену прорамну систему наведена на рис. 3.8.

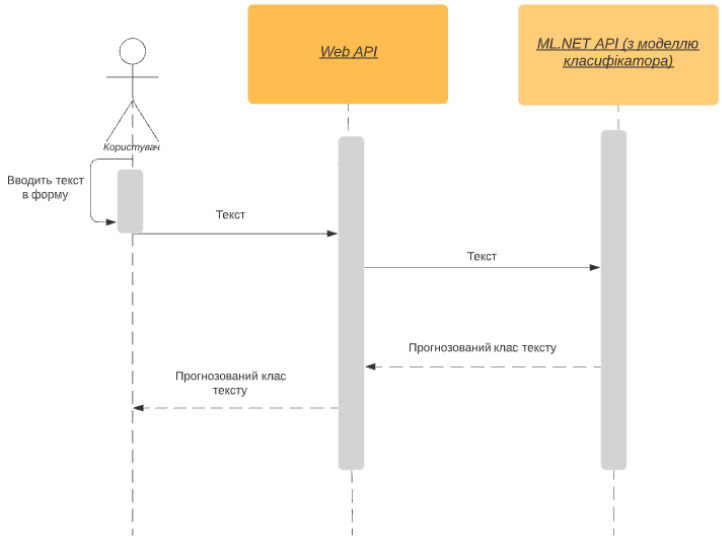

Рисунок 3.8 – Діаграма послідовності аналізу емоційного забарвлення тексту

В залежності від того, до якого класу був віднесений текст класифікатором, користувач отримує відповідне повідомлення про те, чи є текст позитивним чи негативним (рис. 3.9).

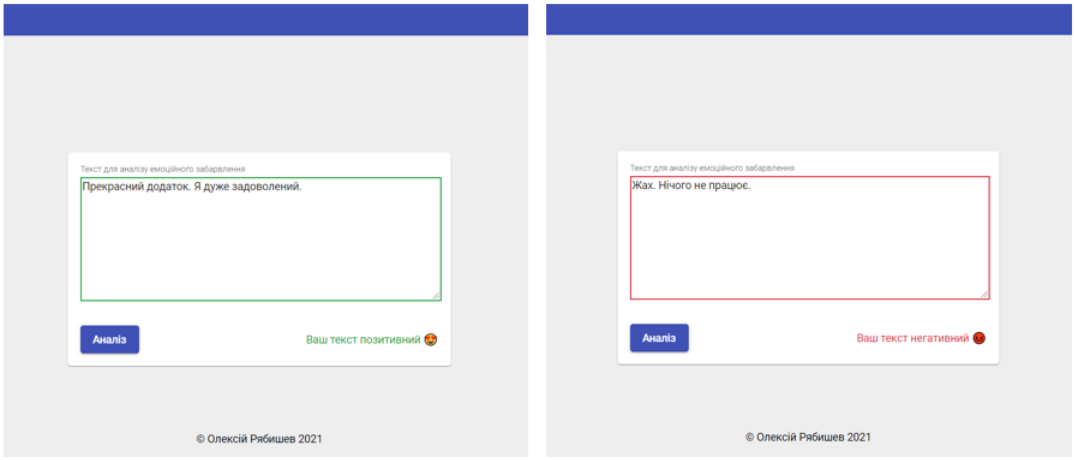

Рисунок 3.9 – Результати аналізу емоційного забарвлення тексту

Розроблена система може вільно використовуватися для визначення емоційного забарвлення тексту, проте найбільш високі результати вона демонструватиме для визначення емоційного забарвленн тексту для домену відгуків клієнтів, оскільки класифікатор, що використовується в даній системі, був отриманий з використанням датасету відгуків клієнтів про мобільні додатки.

ВИСНОВКИ

Аналіз емоційного забарвлення тексту є частиною процесу обробки природної мови, що займається виявленням в текстах емоційно забарвленої лексики та емоційної оцінки автора стосовно об’єктів, про які йдеться в тексті.

Емоційне забарвлення тексту – це ставлення автора висловлювання до об’єкту реального світу, події, процесу чи їх властивостей, про які йдеться в тексті.

Аналіз емоційного забарвлення тексту допомагає бізнесу зрозуміти соціальні настрої щодо бренду, продукту або послуги. Він може бути використаний для проведення досліджень ринку, аналітики продуктів, а також для підтримки користувачів та моніторингу соціальних медіа.

У ході дослідження було розглянуто існуючи підходи до аналізу емоційного забарвлення тексту, способи класифікації тексту за емоційним забарвленням, проаналізовано існуючі системи для аналізу емоційного забарвлення тексту. Були досліджені методи та інструменти аналізу емоційного забарвлення тексту з використанням машинного навчання, розглянуті набори даних (датасети), що можуть бути використані у машинному навчанні.

У ході роботи було реалізовано програмне забезпечення для формування датасету (набору даних) українською мовою, сформований такий датасет на базі відгуків клієнтів про мобільні додатки. Цей датасет може бути використаний для подальших досліджень в області емоційного забарвлення тексту та для побудови систем аналізу емоційного забарвлення тексту з використанням машинного навчання.

Було проведено експериментальне дослідження, з розробкою відповідного програмного коду, щодо ефективності роботи різних алгоритмів машинного навчання для виконання задачі бінарної класифікації тексту українською мовою з використанням сформованого набору даних. Отримані результати були проаналізовані та виявлений найефективніший алгоритм для задачі бінарної

класифікації тексту українською мовою. Також був розроблений відповідний класифікатор на базі виявленого алгоритму.

У ході дослідження, окрім іншого, було використано попередньо навчену багатомовну модель машинного навчання BERT, проведено її доналаштування та адаптацію до задачі бінарної класифікації, враховуючи сформований набір даних. Була виявлена ступінь ефективності використання моделі BERT для виконання задачі бінарної класифікації тексту українською мовою.

Серед іншого, у ході роботи була реалізована система аналізу емоційного забарвлення тексту українською мовою, яка (система) являє собою вебдодаток. Дана система виконує аналіз емоційного забарвлення тексту шляхом розв’язання задачі бінарної класифікації, використовуючи класифікатор, який був побудований на основі оптимального алгоритму, що був визначений в ході дослідження. Ця система може бути вільно використана для аналізу емоційного забарвлення тексту українською мовою, але, в першу чергу, вона може бути використана для домену відгуків клієнтів, оскільки класифікатор системи був побудований з використанням датасету (набору даних) відгуків клієнтів про мобільні додатки.

Результати дослідження подані до опублікування в науковому журналі «Біоніка інтелекту» (див. Додаток Г).

Напрямками подальших досліджень є: розширення оцінки емоційного забарвлення (перехід від бінарної до множинної класифікації) та побудова класифікатора, що зможе визначити емоційно нейтральні тексти; розширення датасету (збір більшого обсягу даних), що дозволить підвіщити якість прогнозів моделей, які побудовані з використанням машинного навчання; формування датасету та побудова класифікаторів для інших доменів; вдосконалення системи аналізу емоційного забарвлення за допомогою впровадження можливості аналізувати набори текстів одночасно (наприклад, всі відгуки під описом товару); реалізація інших підходів (відмінних від машинного навчання), таких як підхід, що заснований на правилах, для аналізу емоційного забарвлення тексту українською мовою.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Lerman K, Gilder A, Dredze M, Pereira F. Reading the markets: forecasting public opinion of political candidates by news analysis. In: Proceedings of the 22nd international conference on computational. – Linguistics 1, 2008. – С. 473–480.

2. Kasper W. Sentiment Analysis for Hotel Reviews / Walter Kasper, Mihaela Vela. – Proceedings of the Computational Linguistics-Applications Conference. – Jachranka, Poland: Polskie Towarzystwo Informatyczne, Katowice, 10/2011. – С. 45–52.

3. Moilanen K. Multi-entity Sentiment Scoring / Karo Moilanen, Stephen Pulman. – Proceedings of Recent Advances in Natural Language Processing (RANLP 2009). – Borovets, Bulgaria, September 14–16 2019. – С. 258–263.

4. Kan D. Rule-based approach to sentiment analysis at ROMIP 2011. URL: http://www.dialog-21.ru/digests/dialog2012/materials/pdf/Kan.pdf (дата звернення: 07.03.2021).

5. Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Cornell University. URL: https://arxiv.org/abs/1810.04805 (дата звернення: 08.03.2021).

6. Jonathan Herzig, Peter Shaw, Ming-Wei Chang, Kelvin Guu, Panupong Pasupat, Yuan Zhang. Unlocking Compositional Generalization in Pre-trained Models Using Intermediate Representations // Cornell University. URL: https://arxiv.org/abs/2104.07478 (дата звернення: 09.04.2021).

7. Peter Shaw, Ming-Wei Chang, Panupong Pasupat, Kristina Toutanova. Compositional Generalization and Natural Language Variation: Can a Semantic Parsing Approach Handle Both? // Cornell University. URL: https://arxiv.org/abs/2010.12725 (дата звернення: 10.04.2021).

8. Yerokhin, A., Nechyporenko A., Babii A., Turuta A. Usage of F-transform to finding informative parameters of rhinomanometric signals / Proc. of the International Conference on Computer Sciences and Information Technologies, IEEE, Lviv, Ukraine, 2015, 14-17 September. – P.129-132.

9. З. В. Дударь, Д. Е. Шуклин. Семантическая нейронная сеть, как формальный язык описания и обработки смысла текстов на естественном языке.- Радиоэлектроника и информатика.№3(12)- 2000.- С.10-25.

10. Yerokhin, A.L., Babii, A.S., Nechyporenko, A.S., Turuta, O.P. / A Lars-Based Method of the Construction of a Fuzzy Regression Model for the Selection of Significant Features // Cybernetics and Systems Analysis. №4, 2016. - P. 167–173. DOI: 10.1007/s10559-016-9867-

11. Andriy Yerokhin, Alina Nechyporenko, Andrii Babii, Oleksii Turuta, Ihor Mahdalina. Usage of Phase Space Diagram to Finding Significant Features of Rhinomanometric Signals // Computer Science & Information Technologies (CSIT'2016), 6-10 Sept. 2016, Lviv, Ukraine. – P. 70 – 72. DOI: 10.1109/STC-CSIT.2016.7589871.

12. Andriy Yerokhin, Valerii Semenets, Alina Nechyporenko, Andrii Babii, Oleksii Turuta. F-transform 3D Point Cloud Filtering Algorithm // Proc. of the 2th IEEE International Conference on Data Stream Mining & Processing. 21-25 August 2018, Lviv, Ukraine. - P.524-527. DOI: 10.1109/DSMP.2018.8478581

13. Turney P. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews // Proceedings of the Association for Computational Linguistics. – С. 417–424.

14. Pang B., Lee L., Vaithyanathan S. Thumbs up? Sentiment Classification using Machine Learning Techniques // Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). – С. 79–86.

15. Snyder B., Barzilay R. Multiple Aspect Ranking using the Good Grief Algorithm // Proceedings of the Joint Human Language Technology/North American Chapter of the ACL Conference (HLT-NAACL). – С. 300–307.

16. Пазельская А. Метод определения эмоций в текстах на русском языке / А. Г. Пазельская, А. Н. Соловьев // Компьютерная лингвистика и интеллектуальные технологии: сб. научных статей / Вып. 10 (17). – М.: Изд-во РГГУ, 2018. – С. 510–522.

17. Denecke K. Using SentiWordNet for Multilingual Sentiment Analysis / Kerstin Denecke. – ICDE Workshops. – 2018. – С. 507-512.

18. IMDB Review Dataset. URL: https://www.kaggle.com/utathya/imdb-review-dataset (дата звернення: 05.04.2021).

19. Тональний словник української мови // GitHub. URL: https://github.com/lang-uk/tone-dict-uk (дата звернення: 04.03.2021).

20. Python NLTK Demos for Natural Language Text Processing. URL: http://text-processing.com/demo/sentiment (дата звернення: 17.03.2021).

21. What is ML.NET and how does it work? URL: https://docs.microsoft.com/ru-ru/dotnet/machine-learning/how-does-mldotnet-work (Дата звернення: 05.04 2021).

22. How to choose an ML.NET algorithm. URL: https://docs.microsoft.com/ru-ru/dotnet/machine-learning/how-to-choose-an-ml-net-algorithm (Дата звернення: 07.04.2021).

23. Transfer Learning / University of WISCONSIN. URL: http://pages.cs.wisc.edu/~shavlik/abstracts/torrey.handbook09.abstract.html (Дата звернення: 10.04.2021).

24. Machine learning tasks in ML.NET. URL: https://docs.microsoft.com/ru-ru/dotnet/machine-learning/resources/tasks (дата звернення: 07.04.2021).

25. Khan A, Baharudin B, Lee LH, Khan K. A review of machine learning algorithms for text-documents classification. – J Adv Inf Technol 1, 2010. – С. 4–20.

26. A General Approach to Preprocessing Text Data // Machine Learning, Data Science, Big Data, Analytics, AI. URL: https://www.kdnuggets.com/2017/12/generalapproach-preprocessing-text-data.html (дата звернення: 05.03.2021).

27. BERT Explained: State of the art language model for NLP / Towards Data Science, URL: https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlpf8b21a9b6270 (дата звернення: 15.04.2021).

28. The Stanford Natural Language Processing Group. URL: https://nlp.stanford.edu/IRbook/html/htmledition/support-vector-machines-and-machine-learning-on-documents-1.html (дата звернення: 05.03.2021).

29. Frank Millstein, Python Machine Learning: Introduction To Machine Learning With Python, Kindle Edition, 2018 – 134 с.