МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

Київський національний університет імені Тараса Шевченка

Навчально-науковий інститут філології
Кафедра української мови та прикладної лінгвістики

Автоматичний сентимент-аналіз українськомовних текстів соцмережі Twitter

Кваліфікаційна робота
освітнього ступеня «бакалавр»
за спеціальністю 035 «Філологія»,
спеціалізацією 035.10 «Прикладна лінгвістика»,
галузі знань 03 «гуманітарні науки»
ОПП «Прикладна (комп'ютерна) лінгвістика та англійська мова»
студентки IV курсу
Валерії ХОЗІНОЇ

науковий керівник:
Микола КОСТІКОВ

«Допущено до захисту»
Протокол № 11 засідання кафедри
української мови та прикладної лінгвістики
ННІФ від 01.06.2023
Завідувач кафедри_____**Сергій Різник**

КИЇВ – 2023

ЗМІСТ

ВСТУП

Тема дослідження: автоматичний сентимент аналізів текстів соціальної мережі Twitter

Актуальність дослідження: створення системи автоматичного сентимент аналізу є актуальним через відсутність матеріалів українською мовою та екстенсивне використання користувачами мережі Інтернет та, зокрема, соціальних мереж.

Зі збільшенням популярності таких сайтів та соціальних мереж як Facebook, Twitter, Instagram, YouTube, TikTok, все більше з'являється платформ для висловлення думок та опіній, ведення дискусій тощо. Тепер, для того щоб дізнатися думку про товар, компанію, політичного кандидата тощо, зробити висновки та прийняти рішення користувач може не тільки поцікавитися думкою друзів та близьких, але й переглянути відгуки в Інтернеті. Для організацій ж відкрилася можливість зменшити свої витрати, відкинути соціологічні дослідження та опитування-фокус груп або доповнити їх оцінками користувачів Інтернету, адже ці дані знаходяться у відкритому доступі. Проте, через надлишок інформації, яка представлена в Інтернеті, та велику кількість різноманітних сайтів, що містять великі обсяги даних, дедалі важче стає розрізнити потенційно неприємну, небезпечну, некорисну інформацію, від безпечної чи корисної, або, принаймні, нешкідливої, нейтральної. Ручний аналіз текстів, що містяться на інтернет-ресурсах, потребує багато часу та ресурсів, та не завжди може призвести до бажаного результату. Для розв'язання таких задач і використовується автоматичний сентимент аналіз.

До того ж, можемо стверджувати, що українська мова є низькоресурсною. За даними порталу Statista українська мова посідає вісімнадцяте місце у рейтингу найбільш частотних мов, які використовуються для створення контенту в Інтернеті, а частка такого контенту складає 0.6%. У той час як англійська мова посідає перше місце, а кількість контенту в Інтернеті англійською мовою складає 58.8% [21], системи автоматичного сентимент аналізу англійської мови вже давно широко застосовуються, для української мови таких систем дуже мало і/або вони знаходяться у стані розробки. Тому важливо, опираючись на вже існуючі дослідження, запровадити можливі рішення для завдання автоматичного сентимент аналізу власне української мови.

Мета дослідження: Створення автоматичної системи визначення сентименту тексту, запропонованого користувачем. Мета передбачає виконання таких **завдань дослідження:**

1. визначити проблеми та завдання процесу автоматичного сентимент аналізу
2. дослідити методи та алгоритми автоматичного сентимент аналізу, зокрема текстів з Інтернету та соціальних мереж
3. створити базу даних текстів із соціальної мережі Twitter
4. побудувати алгоритм створення системи машинного навчання на основі дослідженої літератури
5. розробити програмне забезпечення системи автоматичного сентимент аналізу

Об’єктом дослідження є наповнення соціальної мережі Twitter, її базові основні одиниці – текстові твіти.

Предметом дослідження є автоматичне визначення текстів соцмережі Twitter.

Практичне значення роботи полягає в можливості використання розробленого програмного продукту для автоматичного визначення сентименту текстів, запропонованих користувачем, для забезпечення комфортного користування соціальною мережею.

Теоретичне значення роботи: отримані результати дослідження дозволяють розвивати та продовжувати роботу над створенням систем автоматичного сентимент аналізу української мови, що є важливим, оскільки українська мова є малоресурсною і не підтримується широковживаними мультилінгвальними системами.

Матеріал дослідження: тексти соціальної мережі Twitter.

РОЗДІЛ 1

ПРОБЛЕМИ ТА ЗАВДАННЯ АВТОМАТИЧНОГО СЕНТИМНТ АНАЛІЗУ

1. 1 Теоретичні засади автоматичного сентимент аналізу

Автоматичний сентимент аналіз тексту це одне із завдань NLP, або обробки природної мови (ОПМ), що передбачає використання методів ОПМ та машинного навчання для визначення тональності тексту, його емоційного відтінку, ставлення автора тексту до особи, явища чи реалії, що виступає суб'єктом цього тексту та класифікації аналізованого тексту.

Автоматичний сентимент аналіз тексту широко використовується для оцінки відгуків клієнтів у маркетингу, оцінки комунікації політичних кандидатів із виборцями у політичному дискурсі, загальній оцінці настроїв населення у соціології тощо [14].

Тексти можна класифікувати за сентиментом на позитивні, негативні та нейтральні. Текст із позитивним сентиментом, це той текст, який виражає позитивну/оптимістичну думку почуття чи емоцію. Такий текст може виражати емоції радості, захоплення, вдячності, задоволення, симпатії, впевненості любові. Негативний сентимент використовується на позначення тексту, що виражає песимістичну, негативну чи критичну думку. Такий текст може включати вираження гніву, відчаю, тривоги, розчарування, сорому, смутку, страху. Текст із нейтральним сентиментом не виражає чітко окреслених позитивних чи негативних емоцій та думок. До таких текстів належать констатації фактів, описи та вираження байдужості.

Сентимент у тексті може виражатися з допомогою слів з так званою "prior polarity", або попередньою полярністю, яка за визначає тональність слова

незалежно від контексту [17]. Такі слова, виходячи з їхніх значень та використання, зазвичай тісно асоціюються з певним сентиментом. До слів із попередньою полярністю можемо віднести слова на позначення емоцій відповідних сентиментів та похідні від них. Так, слова “любов”, “радість”, “щастя”, “успіх” вважаються позитивними, а “ненависть”, “сум”, “гнів”, “розчарування” - негативними. Проте, все ж при проведенні семантичного аналізу тексту, не можна відкидати контекст, що оточує слово та робити висновки про сентимент цілого тексту лише через наявність у ньому слова з певною попередньою полярністю. До прикладу, речення *“Цей вечір зробив його неймовірно щасливим!”* містить прикметник, утворений від “щастя” із попередньою позитивною полярністю, та з контексту ми можемо зрозуміти, що речення загалом є позитивним. А от речення “Він так і не став щасливим” вже має негативний тон, хоча містить той самий прикметний “щасливий”.

Не можна відкидати того факту, що певні слова та словосполучення викликають стійкі асоціації у носіїв мови, проте відкидання контексту грубо суперечить меті сентимент аналізу. Таким чином, при проведенні семантичного аналізу використовується ширший набір методів та підходів, які беруть до уваги особливості живої мови.

Опорними для визначення сентименту тексту є розділові та друкарські знаки, а також регістр тексту. Все більше користувачів Інтернету нехтують правилами правопису, оскільки вони не є ключовими для вираження думки. Можемо казати і про виникнення так званого Інтернет етикету, негласного набору правил, що визначає норми спілкування у соціальних мережах та онлайн-платформах.

У текстах з Інтернету, стандартні розділові знаки можуть мати додаткові смислорозрізнювальні функції. Розглянемо такі зміни детальніше. Три крапки

… зазвичай передають незавершену або перервану думку чи висловлювання, у текстах з Інтернету вони також можуть виражати емоції суму чи розчарування. Дужки, які зазвичай використовуються для виокремлення певної інформації, в Інтернет-текстах, якщо використані одноосібно, надають тексту позитивний чи негативний тон. Цю особливість різних типів знаків ми розглянемо детальніше далі.

Відсутність пунктуації, як і її надмірне вживання, може свідчити про високу емоційність автора. Щодо регістру тексту, то використання великих літер для написання усього тексту чи його частини, вважається підвищенням тону та може значно підсилювати емоції, виражені словесно, при чому це можуть бути як позитивні емоції (радість, збудження) так і негативні (гнів, ненависть тощо)

У контексті Інтернет-текстів, важливим показником тональності є емотикони, графічні зображення, що мають на меті передачу виразів обличчя за допомогою знаків пунктуації, цифр та літер, графічних зображень та анімованих зображень. Це збірне поняття, гіперонім, для різних типів графічних зображень, що можуть використовуватися у тексті. До них належать традиційні емотикони, зображення, що створенні за допомогою знаків пунктуації, цифр та літер та широко використовуються досі. Прикладами стандартних емотиконів є :) XD ;-).

Емотикони у ширшому розумінні також включають у себе емоджі, графічні символи, які спершу мали на меті передати вирази обличчя, а на даний момент поділяються кілька категорій із зображеннями тварин, об’єктів, їжі та напоїв символів та прапорів. Еможді є частиною Unicode, стандарту кодування символів, який кожному символу присвоює унікальне числове значення. На даний момент Unicode нараховує 3664 емоджі [41] Емоджі є найпоширенішим на даний момент різновидом емотиконів та підтримується більшістю

операційних систем та Інтернет-платформ. Так, операційна система підтримує IOS - більше ніж 3,600 емоджі [20], Windows 11 – 3952 [29], Android 14 – 3664 [26]. Соцмережа Twitter підтримує 3,245 емоджі, [https://twemoji.twitter.com/40]. До стандартних емоджі належать **heart** ❤, grinning face 😀, victory hand ✌.

Каомоджі, як і традиційні емотикони, створюються за допомогою комбінацій літер, цифр та пунктуаційних знаків, але вони складніші за традиційні емотикони та можуть передавати ширший спектр емоцій та дій (інтеракцій), до них належать привітання, обійми (⊃｡•́‿•̀｡)⊃, здивування w(° o °)w, підморгування, , кровотеча з носу тощо.

Можемо казати про попередню полярність емотиконів, адже частина цих зображень передає певні вирази обличчя та емоції, і викликає сильні асоціації у всіх користувачів.

Не можна оминути зміну тональності емотиконів в залежності від контексту, та зміну їхніх значень із часом. До прикладу емоджі 😷☠💀 окремо від контексту сприймаються негативно та мають відповідну тональність, проте в сучасному Інтернет-сленгу, такі емоджі часто вживаються на позначення сильного сміху та використовуються у відповідних ситуаціях.

Розбіжності в інтерпретуванні юнікод-символів було розглянуто у статті Міллер та інших [15]. Оскільки такі символи мають різне представлення в залежності від платформи, це може призвести до непорозуміння користувачів та втрати змісту, першочергово закладеного у повідомлення. У ході дослідження було використано 25 юнікод-символів, інтерпретовані 5 платформами та проведено опитування учасників, у якому ті мали описати значення наданих символів, визначити їхній сентимент та описати ситуацію, в якій вони вважають

доречним використання такого символу. У результаті було визначено, що існують значні розбіжності у семантичному значенні символів та їхньому сентименті при інтерпретації різними платформами, і в межах інтерпретування символу однією платформою його сприйняття користувачами може суттєво відрізнятися. Такі результати показують, що сентимент та семантичне значення юнікод-символів ґрунтується на перцепції кожного окремого користувача, а отже універсальне визначення їхніх значень та тональностей майже неможливе.

У дослідженнях автоматичного сентимент аналізу існують різні підходи до визначення тональності емотиконів та їхнього впливу на загальний сентимент тексту. При проведенні сентимент аналізу на основі лексем усі знаки, що не є словами відкидаються, тому такий аналіз не є повним для аналізу текстів з Інтернету. У дослідження Агравал та інших [4] що базувалося на tree kernel моделі, запропонований підхід створення анотованого вручну словника тональності емотиконів, який нараховував 170 входжень. Проте такий словник не є вичерпним та потребує постійного оновлення. До того ж, далі ми розглянемо зміну тональності емотиконів у різних контекстах та доцільність використання такого словника загалом.

Шиха та Айваз [16] досліджують вплив емоджі на процес інтелектуального аналізу тексту та сентимент аналізу, у статті проаналізовано випадки використання емоджі у позитивних та негативних текстах соцмережі Twitter для глибшого розуміння природи використання таких символів. Для проведення дослідження було створено лексикон емоджі, який містив 18 варіантів представлення 843 емоджі, підтримувані різними мовами програмування. Кожному входженню призначено мітками -1, 0, 1, на позначення його тону та вказано місце у загальному рейтингу за частотою вживання. Вхідні дані було відібрано на основі сентименту актуальної на момент написання статті теми, тема “The New Year’s Eve” (“Новорічна ніч”)

була обрана як позитивна, а тема “Istanbul Attack” (”Теракт у Стамбулі”) як негативна. Відповідно до наборів вхідних даних, було відібрано позитивні тексти без емоджі, позитивні тести з емоджі, негативні тексти без емоджі та негативні тексти з емоджі. За результатами дослідження, користувачі Twitter частіше використовують емоджі для вираження позитивних емоцій, при чому частка текстів з емоджі серед усіх позитивних текстів складає 19.38%, у той час як частка негативних текстів, що мітили хоча б один емоджі складала 2.8%. Врахування емоджі під час сентимент аналізу безумовно покращує загальні результати визначення сентименту тексту, проте, через перевагу у вхідних даних, для цього дослідження врахування емоджі під час сентимент аналізу більше впливає на визначення позитивного тексту аніж негативного.

При проведенні сентимент аналізу методом нейронних мереж, моделі можна навчити розпізнавати та асоціювати певні емоджі з конкретними емоціями. Такі моделі зазвичай використовують поєднання вхідних текстових даних та емоджі для визначення сентименту тексту. Наприклад, модель BERT включає токенізацію емоджі, але не містить їх у словнику, таким чином сентимент емоджі визначається лише за текстовим оточенням.

У статті Делобелль та Берентдт [9] обговорюється використання емоджі у розмовних системах. Робота базується на мультилінгвальній моделі BERT bert-base-multilingual-cased, до токенайзеру були додані нові токени для 2740 емоджі, емоджі у кодуванні UTF-8 було перетворено у текстові псевдоніми для приведення усіх токенів до одного формату та для уникання втрати емоджі при токенізації. Точність роботи системи при урахуванні емодзі зросла до 17,8% у порівнянні з результатом 12,7% для моделі без урахування емоджі. У роботі досліджено варіанти модифікації системи, заснованої на BERT, запропоновано методи токенізції для підтримки емоджі, однак недостатня кількість високоякісних наборів даних, що містять емоджі обмежує можливості моделей

опрацювання текстів, і створення таких наборів даних значно збільшить можливості завданнь ОПМ.

Розглянемо кілька прикладів текстів соціальної мережі Twitter українською мовою та наочно розберемо проблему визначення їхніх сентиментів. До прикладу, твіт із текстом “НОВИЙ СЕЗОН НЕНСІ ДРЮ ЧЕРЕЗ 25 ДНІВ”[1] не містить розділових знаків та емотиконів, слів з негативною чи позитивною попередньою полярністю, повністю написаний заголовними буквами, але з контексту ми можемо зрозуміти, що при написанні цього тексту автор переживав/ла емоційне піднесення, стан схвилювання та радості. З іншого боку твіт “Я ХОЧУ НА ФУТБОЛ НА ДОНБАС АРЕНІ”[2] має такі самі характеристики, проте з історичного та соціокультурного контексту ми можемо зрозуміти, що він має більш негативну тональність.

Твіт “А уявіть, якби Петріоти були з травня минулого року 🚬” [3] не містить негативно забарвленої лексики, проте у соціальному контексті завдяки емотикону цигарки виражає емоції розчарування та суму. Такі ж емоції передає твіт з текстом “Нашо я домовилася про зустріч на 17 годину, якщо можна попрацювати…”[4] у цьому випадку відсутні емотикони, а емоцію передає три крапки.

Можемо спостерігати велику кількість комбінацій поєднань регістру, лексики та емотиконів для вираження цілого спектру емоцій та можливості надання тексту бажаної тональності. Отже можемо зробити висновок, що сентимент аналіз текстів із Інтернету, а особливо із соціальних мереж складна та багатоетапна задача, оскільки на тон тексту впливає не лише його лексичний

[1] https://twitter.com/whoisrumpel/status/1654600795172618250?s=20
[2] https://twitter.com/navgahb/status/1629447980766961672?s=20
[3] https://twitter.com/LyrYevhen/status/1655350311341400064?s=20
[4] https://twitter.com/borschanesa/status/1655547478790348810?s=20

склад, але й пунктуація, регістр та використання емотиконів, які можуть змінювати своє значення у різному лексичному оточенні та у відповідності до соціально-культурних та історичних контекстів.

1.2 Огляд літератури

Основними методами та алгоритмами для проведення автоматичного семантичного аналізу є метод на основі лексем, метод машинного навчання та комбінація різних методів.

Ґрунтовне порівняння методів сентимент аналізу проводять Кателлі та інші [6]. Дослідження має на меті порівняти та оцінити доцільність використання аналізу на основі лексем та на основі BERT, архітектури моделі глибинного навчання для задач обробки природної мови. Матеріал для дослідження складали 600 відгуків про 6 різних послуг: машини, смартфони, книги, фільми, готелі та відеоігри. До кожної з цих категорій було дібрано 50 позитивних та 50 негативних відгуків. Результати дослідження показують, що використання методу аналізу на основі лексем є менш продуктивними, доцільність їх використання полягає у невеликих наборах вхідних даних та обмежених обчислювальних можливостях. Використання методів глибинного навчання є більш доцільним, оскільки такі системи здатні підлаштовуватись під різні типи вхідних даних, та з їх допомогою у перспективі можливе розв’язання проблеми наявності різних сентиментів в одному тексті.

Шиганов та інші розглядають різні підходи до вирішення поставленої задачі, двома основними напрямками вони вважають аналіз на основі лексем та аналіз методами машинного навчання [2]. Для порівняння вони пропонують лексемно-орієнтований підхід, метод опорних векторів, дерева рішень, метод наївного класифікатора Баєса та нейронні мережі. За результатами дослідження

найкращий результат має метод нейронних мереж, оскільки він не потребує застосування словників, попередньої лінгвістичної обробки текстів, цей метод можна застосовувати до різних типів даних. Менш ефективними постають метод наївного класифікатора Баєса та аналіз на основі використання лексем, оскільки потребують застосування словників, можуть бути застосовані до різних типів даних, до того ж, метод на основі лексем потребує попередньої лінгвістичної обробки тексту.

Розглянемо детальніше роботи з автоматичного сентимент аналізу, які на меті мають дослідження окремих методів вирішення цієї задачі.

Н.П. Дарчук розглядає спосіб автоматичного визначення сентименту тексту українською мовою на основі правил, тобто на основі полярності лексем, які зустрічаються у тексті [1]. Для проведення такого аналізу було використано лексикон публіцистичної лексики української мови, що містив 40 тисяч лексем, відібраних за тональністю (дихотомія позитивний/негативний). Значення сентименту речення у системі визначалося як середнє арифметичне сентиментів слів, які до нього входять. Система, заснована на лексиконі певного стилю, відповідно показує кращі результати у виявленні сентименту текстів, що належать до цього, такі лексикони не є вичерпними для репрезентації мови та потребують постійно доповнення лексикою інших стилів та жанрів.

Автоматичний сентимент аналіз засобами нейронної мережі було розглянуто у статті Ялової та інших [3]. У роботі запропоновано використання двонаправленої архітектури нейронної мережі із довгою короткотривалою пам’яттю (BiLSTM) з додатковим шаром умовно випадкових полів (CRF). Проведення семантичного аналізу за таким алгоритмом передбачає етап набору вхідних даних, попередню обробку вхідних даних, розробка архітектури нейронної мережі, навчання та тестування навченої нейронної мережі та

оцінювання отриманих результатів. Проект було реалізовано як застосунок, користувачі якого могли вручну вводити обраний текст, чи обрати шлях до текстового файлу, що міститься на ПК, результатом роботи застосунку була відповідь щодо сентименту введеного тексту. Після оцінки якості класифікації сентиментів програмою, було визначено, що розроблена модель має високу точність при класифікації сентименту вхідного тексту, при чому найкраще був розпізнаний позитивний сентимент. Ця стаття підкреслює доцільність використання нейронних мереж при проведенні автоматичного сентимент аналізу, проте варто зазначити, що обсяг вхідних даних складав лише 100 текстів англійською мовою, а отже для покращення результатів сентимент аналізу треба значно збільшити розмір вхідних наборів даних.

У роботі Денга та інших [8] розглядається метод проведення сентимент аналізу засобами глибинного навчання, проведено компаративне дослідження отриманих результатів різних моделей та вхідних даних. Метою дослідження стало порівняння продуктивності трьох обраних моделей та покращення методів сентимент аналізу. Було обрано моделі ГНМ, глибинної нейронної мережі, ЗНМ, згорткової нейронної мережі та РНМ рекурсивної нейронної мережі. Також було обрано сім різних наборів маркованих вхідних даних з різних джерел з різною тематикою. Такими наборами стали Sentiment140, Tweets Airline, Tweets SemEval, IMDB Movie Reviews, IMDB Movie Reviews obtained from Stanford University, Cornell Movie Reviews, Book Reviews and Music Reviews. Окрім проведення порівняння продуктивності моделей, для попередньої обробки текстів було використано два методи, це TF-IDF, тобто вага слова у тексті та метод вкладання слів. У результаті дослідження виявлено, що найкращі результати показує комбінація рекурсивної нейронної мережі та метод вкладання слів, при порівнянні часу обробки та точністю результатів найбільш прийнятні результати демонструє згорткова нейронна мережа, а

глибинна нейронна мережа показує посередні результати у порівнянні часу обробки та точності результатів.

Окремо розглянемо дослідження, що мають на меті сентимент аналіз текстів із соцмережі Twitter.

Користь лінгвістичних ознак для проведення сентимент аналізу на основі текстів соцмережі Twitter розглянуто у роботі Колуломпіс та інших [13]. Вхідними даним було обрані три різних корпуси, це HASH, набір даних з гештегами, укладений на основі Edinburgh Twitter corpus, EMOT, набір даних емотиконів та ISIEVE, вручну анотований набір даних, створений компанією iSieve Corporation. Для класифікації експериментів було обрано ряд ознак, вони включали n-грами (уніграми та біграми), ознаки лексикону, тобто попередня полярність слів, частиномовні ознаки, ознаки мікроблогінгу, тобто наявність у тексті емотиконів різної полярності, абревіатур та підсилювачів (верхній регістр слів, повторення символів тощо). Внаслідок проведення експериментів, виявлено, що визначення частиномовної приналежності слів тексту негативно впливає на продуктивність аналізу, це може бути пов’язано як і з низькою якістю частиномовного тегування, так і з загальною відсутністю потреби такого тегування при проведенні сентимент аналізу текстів із Twitter. Найкращі результати дає використання n-грамів, визначення попередньої полярності слів та ознак мікроблогінгу, а от використання ознак мікроблогінгу для навчання даних на основі емоджі зменшується. Також підтвердив свою ефективність метод збору даних на основі гештегів.

Завдання сентимент аналізу текстів із Twitter за допомогою моделі BERT розглянуто у статті Чіорріні, Діамантіні та інших [7]. Метою роботи було дослідження використання двоспрямованих кодувальних представлень з трансформерів для проведення сентимент аналізу та розпізнавання емоцій. Для

проведення дослідження було обрано два варіанти моделі BERTBase: uncased, та що конвертує весь вхідний текст у нижній регістр, та cased, яка цей крок ігнорує. Продуктивність моделі при виконанні сентимент аналізу оцінювалася на тестовому наборі даних, запропонованому Го та ін., який складають вручну анотовані тексти соцмережі Twitter обсягом 430 входжень. Для оцінки результатів визначення емоції тексту було використано Tweet Emotion Intensity dataset, який складається з 6755 входжень, маркованих відповідно до вираженої емоції (гнів, страх, щастя, сум). Результати дослідження показують ефективність застосування BERT для обраних завдань, оскільки моделі досягають точності 0,92 та 0,90 для аналізу настроїв та розпізнавання емоцій відповідно.

Для створення власної системи автоматичного сентимент аналізу було обрано метод глибинного навчання за допомогою моделі BERT. У попередньому пункті було визначено, що контекст слова, фрази чи знаку значно впливає та тональність усього тексту, а отже підхід на основі лексем видається нам недоречним та дещо обмежуючим при аналізі текстів з Інтернету. BERT - це модель глибинного навчання, що використовує нейронні мережі для навчання. Така система опрацьовує контекстне значення слів та символів, до того ж вона є двонаправленою, тобто при аналізі враховує правий та лівий контекст слова. Вона не потребує попередньо створених лексиконів, а отже дозволяє проводити аналіз на текстах різних тем та з різних галузей. Також, BERT можна налаштувати для виконання конкретного аналізу (fine-tuning) та інших мов (окрім англійської), що є великою перевагою у нашому дослідженні.

РОЗДІЛ 2

СТВОРЕННЯ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ АВТОМАТИЧНОГО СЕНТИМЕНТ АНАЛІЗУ УКРАЇНОМОВНИХ ТВІТІВ

2. 1. Опис матеріалів

Матеріалом дослідження було обрано тексти соціальної мережі Twitter. Twitter - одна з найпопулярніших соціальних мереж для мікроблогінгу, користувачі якої можуть публікувати свої думки та опінії у режимі реального часу. Тексти, опубліковані у соціальній мережі Twitter називаються “твіти” та їхня довжина не має перевищувати 280 знаків [22], проте варто зазначити, що користувачі, які придбали підписку Twitter Blue, мають можливість публікувати тексти довжиною 10000 знаків [18].

Сентимент аналіз даних такого формату людиною-редактором не постає складним завданням, оскільки ми здатні розрізняти сарказм та іронію, підтекст та тон такого тексту, можемо краще розуміти культурні контексти та нюанси. Однак, при аналізі великого обсягу даних, цей процес потребує багато часу та людських ресурсів, а його результати можуть бути засновані на вподобаннях та упередженнях редактора.

Проблема автоматичної обробки таких текстів полягає у великій кількості одруків; широкому вживанні емотиконів та зображень, заголовних літер у сегменті тексту для передачі тону повідомлення; використанні користувачами неологізмів, які не містяться у лексиконах та словниках обсценної лексики.

З іншого боку, значну роль у визначенні сентименту тексту грає його жанр [1]. Так, для проведення дослідження було здобуто тексти у режимі реального часу, без застосування сегрегування текстів за тематикою, автором,

типом твіта (ретвіт, відповідь) чи іншими параметрами, єдиною умовою для відбору текстів була українська мова. Така вибірка є більш репрезентативною, оскільки містить тексти будь-яких жанрів, а отже не обмежена спеціальною лексикою (публіцистична тощо) та передає реальні процеси, що відбуваються в Інтернет-спільноті.

Безпосередньо для навчання моделі автоматичного сентимент аналізу було створена два набори даних – марковані та немарковані. База даних маркованих даних складала 1200 входжень, а база даних немаркованих даних – 500000. Обидва набори даних зберігаються у csv-файлах з відповідними назвами, labeled_data.csv для маркованих даних та 500k1.csv для немаркованих даних. Файл з маркованими даними має 4 стовпчики, де перший елемент ім’я користувача (User), другий – повний текст твіту (Tweet), тертій – мітка тональності тексту (label). Файл з немаркованим даним має 3 стовпчики, де перший елемент порядковий номер входження, другий – ім’я користувача (User), третій – повний текст твіту (Tweet), у цьому файлі відсутній стовпець label.

2. 3. Здобування твітів українською мовою

Здобування даних відбувалося під час попереднього етапу дослідження автоматичного сентимент аналізу. Для виконання цього завдання було розглянуто 2 шляхи: 1. Використання бібліотеки Tweepy [39] 2. Використання бібліотеки snscrape [35]. Методом апробації було вирішено використати бібліотеку snscrape, оскільки використання бібліотеки Tweepy вимагає роботу через Акаунт розробника Twitter, що призводить до обмежень у кількості здобутих твітів. Використання бібліотеки snscrape, на момент виконання цього

етапу, дозволяла здобути необмежену кількість інформації без обмеження доступу. Для створення файлу з розширенням CSV було використано модуль pandas [30], призначений для роботи з файлами. Тип структури даних DataFrame дозволяє виконувати операції набагато простіше, ніж із модулем CSV. Дані мають структуру списку списків, у яких перший елемент – ім'я користувача (User), а другий – повний текст твіту (Tweet).

2. 2. Маркування вхідних даних

Загалом було марковано 2000 текстів, з яких було відібрано 400 текстів кожного сентименту, що в підсумку склало базу даних у 1200 входжень з однаковою кількістю позитивних, негативних та нейтральних текстів.

Оскільки основним методом проведення автоматичного аналізу було обрано глибинне навчання за допомогою моделі BERT, для реалізації програми нам було необхідно створити базу вхідних даних з мітками, що відповідають тональності кожного тексту. В ході роботи текстам присвоювалися мітки 0, 1, 2, де відповідно 0 – це негативний текст, 1 – нейтральний та 2 – позитивний.

Присвоєння тексту певної мітки залежало від ряду факторів: контекст повідомлення, лексичне наповнення, використання емотиконів та розділових знаків.

Для визначення сентименту тексту вручну важливо було визначити лексичне наповнення тексту. Так, опорними є слова з попередньо визначеною полярністю, які виражають емоції та оцінки. Зазвичай до таких слів належать прикметники, прислівники, дієслова, що вказують на позитивну чи негативну оцінку. Наявність таких слів частково вказувала про сентимент усього тексту.

Сюди ж ми включили обсценну лексику. У частині випадків її наявність вказувала на негативний тон усього тексту, проте узагальнення в таких випадках буде недоречним, оскільки обсценна лексика може передавати як негативні емоції (обурення, гнів, сум), так і позитивні (захоплення, радість).

До лексичного наповнення тексту відносимо також мову ворожнечі в цілому. У соціальних мережах мовою ворожнечі можемо вважати такий текст, що спрямований на групу людей, на основі певних характеристик, що притаманні цим людям, з метою їх принизити, розпалити ненависть чи закликати до насильства щодо цієї групи людей. Згідно з правилами спільноти Twitter щодо ненависницької поведінки: “Ви не маєте права пропагувати насильство проти інших людей або безпосередньо нападати чи погрожувати іншим людям на основі раси, етнічної приналежності, національного походження, сексуальної орієнтації, статі, гендерної ідентичності, релігійної приналежності, віку, інвалідності або серйозного захворювання» [25].

До цієї категорії ми заносили вживання лексики, що вважається слюрами, тобто образливі слова, що застосовуються до людей певних категорій за ознаками расової та ґендерної приналежності, сексуальної орієнтації, захворювання тощо; образи на основі політичних переконань; цькування, приниження, залякування окремих людей, що не основані на їх приналежності до певної маргіналізованої групи. Відповідно текстам, що містили такі висловлювання було надано оцінку 0 та визначено їх як негативні.

На загальний тон тексту впливали використані емотикони, розділові знаки та регістр тексту**.** Як і у випадку зі словами **з** попередньо визначеною полярністю, емотикони з попередньо визначеною полярністю допомагати при визначенні загального сентименту тексту, проте визначення тону тексту виключно на основі вжитих у ньому емотиконів було б недоречним. Зміни у

значенні емотиконів з часом та у різних контекстах ми попередньо розглянули у пункті *1.1 Теоретичні засади автоматичного сентимент аналізу*.

Як уже було зазначено, загальна база даних усіх маркованих текстів складала 2000 входжень, проте при відборі текстів до бази даних, яка застосовувалася при навчанні моделі, було помічено, що частка позитивних текстів значно менша, ніж частка негативних чи нейтральних текстів. Це обумовлено особливістю спілкування у соцмережах, яка полягає у перевазі негативних та нейтральних настроїв, при чому кількість негативних текстів постійно зростає. Згідно з даними дослідження Mention [19] у період з 2013 по 2017 рік соцмережі стали значно більш негативними, а відсоток негативних текстів досліджуваних даних англійською мовою зріс з 2.71% до 6.86%, при значенні частки позитивних текстів з 13.04% до 5.57% (при чому більша частина текстів визначена як нейтральні). У зв'язку з повномасштабним вторгненням російської федерації логічним є зростання негативних настроїв населення, значно зросла частка користувачів соцмереж, які відкрито висловлюють ненависть до ворога.

Варто зазначити, що маркування текстів вручну було проведено одним експертом, що навіть при наявності визначених критеріїв, безпосередньо може впливати на результати, оскільки таке маркування не є абсолютно об'єктивним. Контроль над упередженістю та її зменшення при маркуванні вхідних даних, було представлено у презентації проведення автоматичного сентимент аналізу на основі трансформеру BERT команди аналітиків банкової корпорації ING [33]. Так, вони почали з відбору великої кількості експертів-анотаторів з різних культур, команд та регіонів та випадково призначали людей для маркування вхідних даних. Для кожного коментаря було призначено три незалежних експерти-анотатори та потім проведено голосування для визначення оцінки текстів. До того ж, для постійного оновлення інструкцій, було обраховано Inter-

Annotator Agreement (IAA), метрику, що показує наскільки два експерти погоджуються між собою. Такий підхід майже повністю виключає упередженість експертів та значно покращує результати. Оскільки команда нашого дослідження обмежена однією людиною, використання такого підходу було неможливим і було прийнято рішення керуватися окресленими критеріями визначення сентименту вхідних текстів.

2. 4. Попередня обробка вхідних даних

Для створення програмного коду автоматичного укладання частотних словників було обрано мову програмування Python 3.10 [42], через її доступність, можливість доповнення та розширення коду. Перевагою Python є велика кількість вбудованих та зовнішніх бібліотек, що дозволяє створювати програми різних типів. Написання програми відбувалося в інтегрованому середовищі розробки PyCharm Community Edition версії 2020.2.2 [31].

Повний програмний код знаходиться за покликанням:

https://drive.google.com/drive/folders/1y5F1mV0--lj8PTAsvvhYHf33t1WlYTp8?usp=sharing

У кореневій папці міститься файли preprocessing.py, main.py, GUI.py, labeled_data.csv, 500k1.csv. Перші три файли Python: програмне забезпечення системи, написані мовою програмування Python.

Файл preprocessing.py забезпечує попередню обробку маркованих та немаркованих текстових даних. Текстовий опис алгоритму вміщено у додатку А. Файл main.py містить програмний код тренування попередньо навченої моделі на основі попередньо оброблених маркованих та немаркованих даних. Текстовий опис алгоритму вміщено у додатку Б. Файл GUI.py містить

програмний код графічного інтерфейсу користувача для автоматично аналізу сентименту введеного тексту. Текстовий опис алгоритму вміщено у додатку В. Файл labeled_data.csv містить базу даних маркових вхідних даних обсягом 1200 твітів. Фрагмент бази даних подано у Додатку Г**.** Файл 500k1.csv містить базу даних немаркових вхідних даних обсягом 500000 твітів. Фрагмент бази даних подано у Додатку Д**.** Бази даних вхідних текстів формуються у формі текстових файлів для представлення табличних даних із розширенням CSV.

Для створення функціональної системи автоматичного сентимент аналізу на основі текстів соцмережі Twitter роботу було розділено на низку задач:

1. Попередня обробка вхідних даних
2. Тренування попередньо навченої моделі на основі маркованих та немаркованих даних
3. Створення графічного інтерфейсу користувача

Попередня обробка тексту є важливим етапом сентимент аналізу тексту. Він створює основу для ефективної класифікації текстів, готує вхідні дані для подальшої обробки зменшуючи зашумленість тексту та залишаючи лише релевантні дані, що місять інформацію про сентимент тексту. Попередня обробка впливає на точність результатів та продуктивність створеної системи.

Дослідження Хамалта, Варми та Говардана [12] пояснює необхідність попередньої обробки текстів для визначення їхнього сентименту. У роботі запропоновано метод попередньої обробки вхідного тексту для проведення автоматичного сентимент аналізу текстів із соцмережі Twitter на основі SentiWordNet, де етап нормалізації тексту підпорядковується вимогам семантичного словника WordNet. У роботі використано такі завдання попередньої обробки тексту: видалення URL-адрес, видалення символів, що

повторюються більше ніж три рази у слові, так слово happyyyyy після обробки набувало стандартного вигляду happy, видалення питальних слів, видалення спеціальних символів, та видалення символів ретвіту RT, які були актуальними на момент проведення дослідження.

Вплив різних методів попередньої обробки тексту при проведенні сентимент аналізу методом опорних векторів досліджено у статті Хадді, Ліу та Ші [11]. У цій роботі порівняно результати роботи програми визначення сентименту відгуків на фільми для попередньо оброблених та необроблених вхідних даних. Для попередньої обробки було використано методи видалення пробілів та стоп-слів, зведення слів до основи, тобто стемінг, розшифрування абревіатур та обробка абревіатур. Такий підхід значно покращив точність роботи програми.

Порівняння методів попередньої обробки тексту для автоматичного сентимент аналізу текстів соцмережі Twitter надано у роботі Ангіані та інших [5]. Дослідження проводилося на наборі даних 2015 та 2016 років від SemEval. Для визначення сентименту текстів було обрано наївний баєсів класифікатор. Для визначення оптимального методу попереднього опрацювання текстів усі методи було розділено на кілька комбінацій: стандартна обробка, стандартна обробка та стемінг, стандартна обробка та видалення стоп-слів, стандартна обробка та опрацювання заперечень, тобто заміна усіх конструкцій з запереченням на слово not, стандартна обробка та обробка емотиконів, стандартна обробка та робота зі словником, поєднання усіх методів, поєднання усіх методів без використання словника та використання вхідного тексту без використання попередньої обробки. Стандартна обробка полягала у видаленні URL-адрес, гештегів та тегувань користувачів, видалення розділових знаків, приведення слів до словникових варіантів, перетворення емотиконів на текстові теги, конвертація усього тексту у нижній регістр та видалення зайвих пробілів.

За результатами дослідження найкращі результати показує комбінація стандартної обробки та стемінгу, а стандартна обробка та використання словника зовсім не впливають на результат.

Оскільки ми працюємо з попередньо навченою моделлю BERT, така модель використовує усю інформацію в реченні, включаючи розділові знаки та стоп-слова, використовуючи механізм багатоголової самоуваги [10], для проведення дослідження детальна попередня обробка вхідного тексту не є необхідною. Однак, підготовка вхідних маркованих та немаркованих даних все ще є важливим кроком. Оскільки вхідні дані було отримано з соцмережі Twitter, для такого тексту буде типовим використання URL-адрес та тегувань інших користувачів соцмережі. На цьому етапі ми видаляємо такі послідовності символів, оскільки вони не впливають на сентимент тексту. Така обробка також забезпечить видалення повторюваних твітів, які могли бути опубліковані ботами, та забезпечить унікальність вхідних даних. Додатково при обробці вхідних немаркованих даних було видалено входження, що містили кириличні символи, непритаманні українській абетці. Це було зроблено з метою уникання текстів, які бібліотека snscrape помилково вважала текстами, написаними українською мовою.

У проект імпортуємо бібліотеки Pandas для роботи з даними та re [32] для роботи з регулярними виразами.

Починаємо з підготовки маркованих даних, визначаємо шлях до вхідного та вихідного файлу, визначаємо цільовий стовпець “Tweet” для роботи з його вмістом. Зчитуємо вхідний csv-файл у форматі DataFrame за допомогою модуля Pandas.

Застосовуємо регулярний вираз, щоб видалити знаки, які не мають тональності та не впливають на проведення сентимент аналізу у цільовому стовпці. Оскільки вхідні дані було отримано з соцмережі Twitter, для такого тексту буде типовим використання URL-адрес та згадок інших користувачів соцмережі за допомогою символу @, ці дані не впливають на загальний сентимент тексту, а отже ми їх видаляємо. На цьому ж етапі видаляємо дублікати рядків та зберігаємо лише унікальні значення у стовпці “Tweet”. Врешті, зберігаємо оброблені дані у вихідний csv-файл labeled_data1.csv.

Виконуємо аналогічні кроки для файлу з немаркованими даними: застосовуємо регулярний вираз для видалення посилань та згадок у текстах цільового стовпця, у цьому ж стовпці видаляємо дублікати рядків. Для видалення текстів, написаних не українською мовою визначаємо кириличні символи, які не містяться в українському алфавіті та видаляємо усі рядки, що містять такі символи. Зберігаємо отримані дані у вихідний csv-файл unlabeled_500k. Після обробки кількість входжень у файлі з немаркованими даними складає 475182 входження.

2. 5. Навчання моделі автоматичного сентимент аналізу

У ході роботи було реалізовано програмний код навчання моделі для виконання автоматичного сентимент аналізу на основі текстів із соцмережі Twitter за допомогою архітектури BERT. При тренуванні моделі було реалізовано підхід навчання з учителем та самонавчання. Навчання з учителем здійснювалося за допомогою маркованих даних, де вхідні тексти мають відповідні мітки тону. На етапі самонавчання модель генерує передбачення для немаркованих даних, тобто генерує псевдомітки та псевдомарковані дані, та

об’єднує їх із маркованими даними для збільшення набору даних та покращення продуктивності.

За допомогою бібліотеки transformers [38], створеною Hugging Face [28] було імпортовано токенайзер та попередньо навчену модель BERT для класифікації послідовностей. У нашому випадку такою моделлю була обрана bert-base-multilingual-cased [27], оскільки вона тренована на даних статей з Wikipedia у тому числі української мови, також вона зберігаю інформацію про регістр тексту вхідних даних, що важливо для аналізу тексту з Інтернету, а особливо із соцмереж.

Для створення програмного коду навчання моделі у консольний проект імпортуємо необхідні модулі Pandas для роботи з даними, Torch [37] для глибинного навчання та Transformers для імплементації обраної попередньо навченої моделі BERT, додатково з бібліотеки Sklearn [34] імпортуємо модуль Sklearn.metrics для обрахування оцінки точності моделі.

Визначаємо клас SentimentDataset**,** який відповідає за обробку і кодування вхідних даних за допомогою встановленого токенайзеру BERT, та застосовує padding, тобто скорочення тексту, якщо він перевищує встановлену довжину, та truncation, відповідно доповнення тексту до необхідної довжини, для уніфікації довжини вхідних даних.

За допомогою функції pd.read_csv бібліотеки pandas завантажуємо csv-файли з маркованими та немаркованими даними, задаємо необхідні гіперпараметри: max_length (максимальна довжина послідовності), batch_size (розмір партії), num_epochs (кількість епох) та learning_rate (темп навчання).

З попередньо навченої моделі bert-base-multilingual-cased за допомогою бібліотеки transformers завантажуємо токенайзер BERT. Створюємо два

екземпляри класу SentimentDataset labeled_dataset для маркованих даних та unlabeled_dataset для немаркованих даних. Завантажуємо модель BERT для класифікації послідовностей з попередньо навченої моделі bert-base-multilingual-cased, налаштовуємо модель на роботу з трьома мітками, відповідно до кількості аналізованих сентиментів. Налаштовуємо оптимізатор AdamW для оптимізації параметрів моделі із заданим темпом навчання.

Далі ініціюємо тренувальний цикл із самонавчанням. Переводимо модель у режим навчання model.train(). Навчання виконується у циклі epoch in range(num_epochs) протягом попередньо визначеної кількості епох, у нашому випадку десять. Спершу виконується тренування на маркованих даних. У кожній ітерації циклу обробляється пакет (batch) маркованих даних, з якого витягуються вхідні ідентифікатори, тобто токенізований текст, attention masks та істинні мітки, тобто мітки надані текстам анотатором та переносяться на пристрій. Перед обчисленням втрат градієнти оптимізатора обнуляється. Пакет маркованих даних подається у модель BERT, яка обчислює втрати.

Для пакету немаркованих даних витягуються вхідні ідентифікатори та attention masks та переносяться на пристрій. Оскільки цей набор даних не має істинних міток, генеруються псевдомітки. Модель робить передбачення щодо міток для немаркованих даних та присвоює їх заданому набору даних.

Потім, марковані та псевдомарковані дані об’єднуються, створюючи один набір даних. Оновлюються параметри оптимізатора, для покращення продуктивності та зменшення втрат.

Після завершення циклу навчання, виконуємо оцінку точності моделі на маркованих даних. Переводимо модель у режим оцінювання model.eval(). Для

маркованих даних робляться передбачення та пізніше порівнюється з істинними мітками за допомогою функції accuracy_score.

Зберігаємо навчену модель та токенізатор за допомогою методу save_pretrained.

Отже, творений програмний код завантажує марковані та немарковані дані, налаштовує попередньо навчену модель BERT для класифікації сентименту тексту із використанням маркованих та псевдомаркованих даних (процес fine-tuning), обчислює оцінку точності моделі та зберігає навчену модель та токенізатор.

2. 5 Створення графічного інтерфейсу користувача

Наступним етапом створення програми автоматичного сентимент аналізу стало створення графічного інтерфейсу користувача (GUI). Програмний код створює GUI додаток, який приймає на вхід введений користувачем текст, та на основі попередньо навченої моделі визначає сентимент заданого тексту.

Стандартною бібліотекою для реалізації програмного коду користувацького інтерфейсу є Tkinter [36]. Перевагами цієї бібліотеки є інтуїтивний синтаксис, Tkinter простий у розумінні та опануванні, це вбудована бібліотека Python, а отже не потребує додаткових завантажень. Однак, одним із найбільш обговорюваних недоліків Tkinter є його застарілий вигляд.

Для створення нашого користувацького інтерфейсу було імпортовано бібліотеку customtkinter [24], це сучасна та налаштовувана UI-бібліотека, що заснована на Tkinter. Оскільки ми вже знайомі з методами створення графічного інтерфейсу користувача на основі Tkinter, такий вибір покращить загальний вигляд програми, зберігаючи простоту роботи зі стандартною Tkinter. Візуальні

елементи customtkinter створюються та використовуються як стандартні елементи Tkinter та можуть бути використані у поєднанні з елементами Tkinter в одному проекті. Цю бібліотеку було обрано через витриманий та сучасний стиль елементів, легку адаптацію елементів інтерфейсу до зовнішнього вигляду системи користувача.

Отже, для створення програми з графічним користувацьким інтерфейсом спершу імпортуємо до проекту необхідні бібліотеки та модулі. Для створення та роботи з графічним інтерфейсом імпортуємо customtkinter та tkinter, для роботи з моделлю BERT імпортуємо бібліотеки torch та transformers.

Далі створюємо вікно GUI, визначаємо розміри вікна та його заголовок. Встановлюємо зовнішній вигляд вікна, задаємо темну тему з блакитним кольором. Створюємо функцію perform_sentiment_analysis(), яка виконує сентимент аналіз тексту, введеного користувачем у текстове поле. Далі текст токенізується та і подається на вхід попередньо навченій моделі. Результати аналізу виводяться у вигляді повідомлення у спливаючому вікні. Визначаємо шлях до збереженої моделі та токенайзеру, який використовувався під час навчання моделі. Створюємо графічні елементи вікна та запускаємо основний цикл програми.

2.6 Демонстрація роботи системи

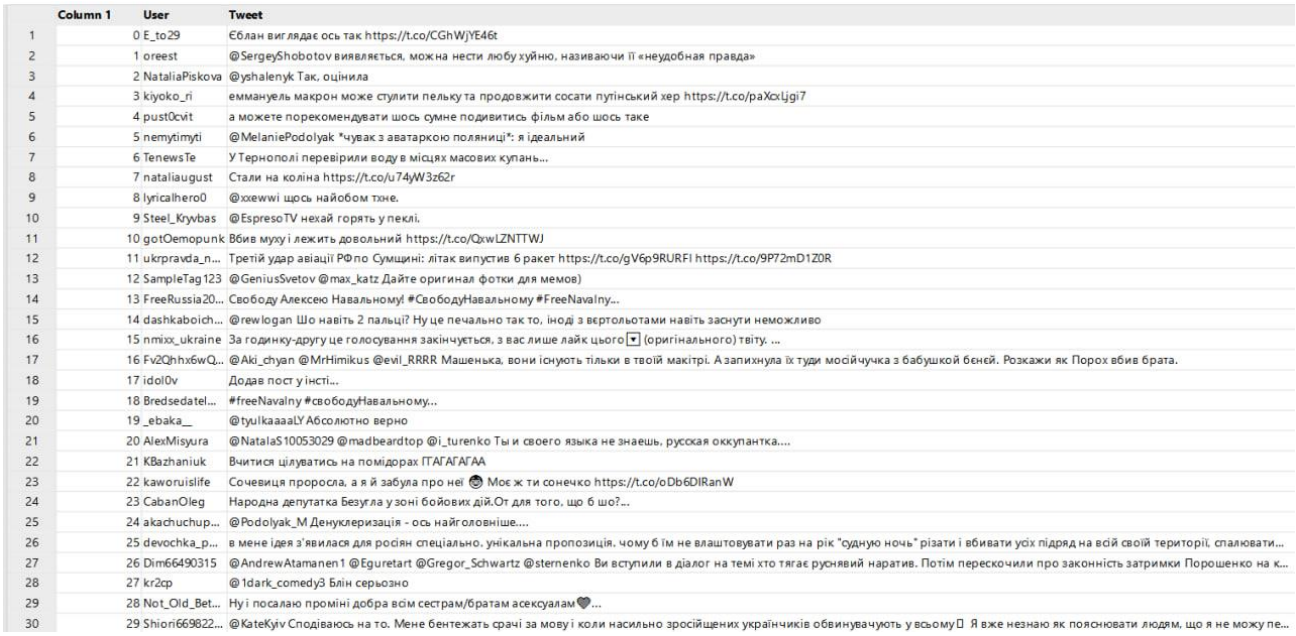
Першим етапом створення системи автоматичного сентимент аналізу була обробка вхідних даних. На цьому етапі ми видалили із вхідних текстів усі символи та послідовності, які не впливають на сентимент тексту та можуть зменшувати продуктивність роботи системи. Представлення csv-файлів було забезпечено програмою CSViewer [23]

	User	Tweet	label
1	__agusha__di__	ранок!! гарних вам вихідних	2
2	__asyya__	Вам реально так сподобався мій малюнок? Безмежно дякую всім котикам, за такі прекрасні слова. Моя мотивація малювати зараз просто злетить до небес Люблю вас, м...	2
3	__Mabelel__	@HochuDeruniv Ваааау це крутезно! Надіюсь це буде гарне і круте місце, удачі!!	2
4	_Anhelina_a	@Robin_z_Fajnogo Доброго ранку, Руслан.	2
5	_Irunka1408	Дізналась, що чувак просто утилізує батарейки з одноразок і робить з них павери для ЗСУ. Це геніально. https://t.co/NjYWMp4pVJ	2
6	_kolizhanka	@yosei_sm яка краса!!	2
7	_kotuk_	@Bulba_Men Ну я і раніше їх знав і чув, але вчора конкретно послухал багато пісень. ...	2
8	10nyk	От би було класно в якомусь спецпроєкті закроссоверити канал "Обличчя Незалежності" @DarkaHirna з каналом "На пошуки грамоти" @cabbage_sad ...	2
9	1705Belka	Вчора дві години слухала лекції з палеобіології від університету Альберти, потім годину дивилась на ютубі підбірку нових динозаврів Jurassic World Evolution 2 у 4K, і вважаю, що це іде...	2
10	1femmefatale_	@jjhopiie ЖІНКОООО, ти неймовірна! ...	2
11	4Ed'tCuQMCtkZmef	Роки 4чи5 тому,коли кацапи пропонували Польщі розділити з ними Україну, хтось з польських політиків тихо сказав:"Краще ми з Україною поділимо вас"...Ну,тоді я поґеґала і вспокоїл...	2
12	5Bykodir	@mirvla Морган майже такий як сьогодні)	2
13	74Fraza	Гумор від наших "І чому їм не подобається NLAW? Постріл, і "сусіди" за лічені секунди "розігріваються". Що в них горить, що вони тут... А там могли б тушити". Так зрозумів, що це про "...	2
14	7kitLea	Від усієї душі бажаю цього кожному з українців! https://t.co/rd95wnftjM	2
15	ada_devil_	Ранку, сонечки ...	2
16	AdasssB77	@MirovichMedia ситуація в Україні відображена на голові в Бориса Дж.))) https://t.co/Ov3dJ0ceWI	2
17	advoka_tesa	@326840nk Тримайтеся, Наталю!	2
18	AFormusyak	Нарешті норм погода у Львові. Свіжо. 3бс	2
19	Ailee_hyong	@tenmybb Сексиии	2
20	aitch6ar	@tataserpeny Шліть іще як буде настрій мої повідомлення відкритті для хороших людей	2
21	aitch6ar	@tataserpeny Дякую	2
22	AjzadaSerikova	@fluffynotwitch богиня	2
23	Akina_Hitomi	@VodkaForNat @MiraidaShinomia Клас виглядає так знайомо	2
24	alexalex6169	@Gadzhega @kaliha Це була відсилка до справжнього персонажу https://t.co/kvxFI5HB3I	2
25	Alexand39995840	@frankbuld Ранку, кицюня	2
26	Alexand39995840	@irinaromashkin1 Раночку, Іруня	2
27	alexmushak	@ohiAnnablya @uspenovka19632 виживеш, переможеш, виховаєш дітей!	2
28	alina_bondarnk	@lady_shtopor ЦІ ЛЮДИ	2
29	anastasiias_k	О, гарна ідея!...	2
30	AntiX_in_jeans	@kvikksolv В тебе є кіт - то вже причина прокидатися)...	2
31	anxietypanianna	@lublucoffeok Начебто трохи виспалася. Ще нічого не планувала, тому піду поки що поп'ю чаю та поїм....	2
32	anywayandme	@freiya_twt мені мама досить часто пропонує випити разом ахвхв	2
33	ar7102000	Доброго ранку,ми з України! https://t.co/DCmkyU9qif	2
34	Arrii_aa	@olyavug Найкращий варіант	2

Малюнок 1 **Представлення бази даних маркованих текстів до обробки**

	User	Tweet	label
1	__agusha__di__	ранок!! гарних вам вихідних	2
2	__asyya__	Вам реально так сподобався мій малюнок? Безмежно дякую всім котикам, за такі прекрасні слова. Моя мотивація малювати зараз просто злетить до небес Люблю вас, м...	2
3	__Mabelel__	Ваааау це крутезно! Надіюсь це буде гарне і круте місце, удачі!!	2
4	_Anhelina_a	Доброго ранку, Руслан.	2
5	_Irunka1408	Дізналась, що чувак просто утилізує батарейки з одноразок і робить з них павери для ЗСУ. Це геніально.	2
6	_kolizhanka	яка краса!!	2
7	_kotuk_	Ну я і раніше їх знав і чув, але вчора конкретно послухал багато пісень. ...	2
8	10nyk	От би було класно в якомусь спецпроєкті закроссоверити канал "Обличчя Незалежності" з каналом "На пошуки грамоти" ...	2
9	1705Belka	Вчора дві години слухала лекції з палеобіології від університету Альберти, потім годину дивилась на ютубі підбірку нових динозаврів Jurassic World Evolution 2 у 4K, і вважаю, що це іде...	2
10	1femmefatale_	ЖІНКОООО, ти неймовірна! ...	2
11	4Ed'tCuQMCtkZmef	Роки 4чи5 тому,коли кацапи пропонували Польщі розділити з ними Україну, хтось з польських політиків тихо сказав:"Краще ми з Україною поділимо вас"...Ну,тоді я поґеґала і вспокоїл...	2
12	5Bykodir	Морган майже такий як сьогодні)	2
13	74Fraza	Гумор від наших "І чому їм не подобається NLAW? Постріл, і "сусіди" за лічені секунди "розігріваються". Що в них горить, що вони тут... А там могли б тушити". Так зрозумів, що це про "...	2
14	7kitLea	Від усієї душі бажаю цього кожному з українців!	2
15	ada_devil_	Ранку, сонечки ...	2
16	AdasssB77	ситуація в Україні відображена на голові в Бориса Дж.)))	2
17	advoka_tesa	Тримайтеся, Наталю!	2
18	AFormusyak	Нарешті норм погода у Львові. Свіжо. 3бс	2
19	Ailee_hyong	Сексиии	2
20	aitch6ar	Шліть іще як буде настрій мої повідомлення відкритті для хороших людей	2
21	aitch6ar	Дякую	2
22	AjzadaSerikova	богиня	2
23	Akina_Hitomi	Клас виглядає так знайомо	2
24	alexalex6169	Це була відсилка до справжнього персонажу	2
25	Alexand39995840	Ранку, кицюня	2
26	Alexand39995840	Раночку, Іруня	2
27	alexmushak	виживеш, переможеш, виховаєш дітей!	2
28	alina_bondarnk	ЦІ ЛЮДИ	2
29	anastasiias_k	О, гарна ідея!...	2
30	AntiX_in_jeans	В тебе є кіт - то вже причина прокидатися)...	2
31	anxietypanianna	Начебто трохи виспалася. Ще нічого не планувала, тому піду поки що поп'ю чаю та поїм....	2
32	anywayandme	мені мама досить часто пропонує випити разом ахвхв	2
33	ar7102000	Доброго ранку,ми з України!	2
34	Arrii_aa	Найкращий варіант	2

Малюнок 2 **Представлення бази даних маркованих текстів після обробки**



	Column 1	User	Tweet
1		0 E_to29	Єблан виглядає ось так https://t.co/CGhWjYE46t
2		1 oreest	@SergeyShobotov виявляється, можна нести любу хуйню, називаючи її «неудобная правда»
3		2 NataliaPiskova	@yshalenyk Так, оцінила
4		3 kiyoko_ri	еммануель макрон може стулити пельку та продовжити сосати путінський хер https://t.co/paXcxLjgi7
5		4 pust0cvit	а можете порекомендувати шось сумне подивитись фільм або шось таке
6		5 nemytimyti	@MelaniePodolyak *чувак з аватаркою поляниці*: я ідеальний
7		6 TenewsTe	У Тернополі перевірили воду в місцях масових купань...
8		7 nataliaugust	Стали на коліна https://t.co/u74yW3z62r
9		8 lyricalhero0	@xxewwi щось найобом тхне.
10		9 Steel_Kryvbas	@EspresoTV нехай горять у пеклі.
11		10 gotOemopunk	Вбив муху і лежить довольний https://t.co/QxwLZNTTWJ
12		11 ukrpravda_n...	Третій удар авіації РФпо Сумщині: літак випустив 6 ракет https://t.co/gV6p9RURFI https://t.co/9P72mD1Z0R
13		12 SampleTag123	@GeniusSvetov @max_katz Дайте оригинал фотки для мемов)
14		13 FreeRussia20...	Свободу Алексею Навальному! #СвободуНавальному #FreeNavalny...
15		14 dashkaboich...	@rewlogan Шо навіть 2 пальці? Ну це печально так то, іноді з вертольотами навіть заснути неможливо
16		15 nmixx_ukraine	За годинку-другу це голосування закінчується, з вас лише лайк цього ▾ (оригінального) твіту. ...
17		16 Fv2Qhhx6wQ...	@Aki_chyan @MrHimikus @evil_RRRR Машенька, вони існують тільки в твоїй макітрі. А запихнула їх туди мосійчучка з бабушкой бєней. Розкажи як Порох вбив брата.
18		17 idol0v	Додав пост у інсті...
19		18 Bredsedatel...	#freeNavalny #свободуНавальному...
20		19 _ebaka__	@tyulkaaaaLY Абсолютно верно
21		20 AlexMisyura	@NatalaS10053029 @madbeardtop @i_turenko Ты и своего языка не знаешь, русская оккупантка....
22		21 KBazhaniuk	Вчитися цілуватись на помідорах ГГАГАГАГАА
23		22 kaworuislife	Сочевиця проросла, а я й забула про неї 🧒 Моє ж ти сонечко https://t.co/oDb6DIRanW
24		23 CabanOleg	Народна депутатка Безугла у зоні бойових дій.От для того, що б шо?...
25		24 akachuchup...	@Podolyak_M Денуклеризація - ось найголовніше....
26		25 devochka_p...	в мене ідея з'явилася для росіян спеціально. унікальна пропозиція. чому б їм не влаштовувати раз на рік "судную ночь" різати і вбивати усіх підряд на всій своїй території, спалювати...
27		26 Dim66490315	@AndrewAtamanen1 @Eguretart @Gregor_Schwartz @sternenko Ви вступили в діалог на темі хто тягає руснявий наратив. Потім перескочили про законність затримки Порошенко на к...
28		27 kr2cp	@1dark_comedy3 Блін серьозно
29		28 Not_Old_Bet...	Ну і посилаю проміні добра всім сестрам/братам асексуалам🖤...
30		29 Shiori669822...	@KateKyiv Сподіваюсь на то. Мене бентежать срачі за мову і коли насильно зросійщених українчиків обвинувачують у всьому Я вже незнаю як пояснювати людям, що я не можу пе...

Малюнок 3 **Представлення бази даних немаркованих текстів до обробки**

	User	Tweet
1	E_to29	Єблан виглядає ось так
2	oreest	виявляється, можна нести любу хуйню, називаючи її «неудобная правда»
3	NataliaPiskova	Так, оцінила
4	kiyoko_ri	емманюель макрон може стулити пельку та продовжити сосати путінський хер
5	pust0cvit	а можете порекомендувати шось сумне подивитись фільм або шось таке
6	nemytimyti	*чувак з аватаркою поляниці*: я ідеальний
7	TenewsTe	У Тернополі перевірили воду в місцях масових купань...
8	nataliaugust	Стали на коліна
9	lyricalhero0	щось найобом тхне.
10	Steel_Kryvbas	нехай горять у пеклі.
11	gotOemopunk	Вбив муху і лежить довольний
12	ukrpravda_n...	Третій удар авіації РФ по Сумщині: літак випустив 6 ракет
13	SampleTag123	Дайте оригинал фотки для мемов)
14	dashkaboich...	Шо навіть 2 пальці? Ну це печально так то, іноді з вєртольотами навіть заснути неможливо
15	nmixx_ukraine	За годинку-другу це голосування закінчується, з вас лише лайк цього ▼ (оригінального) твіту. ...
16	Fv2Qhhx6wQ...	Машенька, вони існують тільки в твоїй макітрі. А запихнула їх туди мосійчучка з бабушкой бєнєй. Розкажи як Порох вбив брата.
17	idol0v	Додав пост у інсті...
18	Bredsedatel...	#freeNavalny #свободуНавальному...
19	_ebaka__	Абсолютно верно
20	KBazhaniuk	Вчитися цілуватись на помідорах ГТАГАГАГАА
21	kaworuislife	Сочевиця проросла, а я й забула про неї 🤓 Моє ж ти сонечко
22	CabanOleg	Народна депутатка Безугла у зоні бойових дій.От для того, що б шо?...
23	akachuchup...	Денуклеризація - ось найголовніше....
24	devochka_p...	в мене ідея з'явилася для росіян спеціально. унікальна пропозиція. чому б їм не влаштовувати раз на рік "судную ночь" різати і вбивати усіх підряд...
25	Dim66490315	Ви вступили в діалог на темі хто тягає руснявий наратив. Потім перескочили про законність затримки Порошенко на кордоні. Потім знову пере...
26	kr2cp	Блін серьозно
27	Not_Old_Bet...	Ну і посилаю проміні добра всім сестрам/братам асексуалам🖤...
28	Shiori669822...	Сподіваюсь на то. Мене бентежать срачі за мову і коли насильно зросійщених українчиків обвинувачують у всьому □ Я вже незнаю як пояснюват...
29	Not_Old_Bet...	По ітогу ми перевели тему, бо вони на щастя помітили що мені не хочеться за це говорити, і вибачились за то...
30	Not_Old_Bet...	На що мені почали казати що ну може просто тому що ви мало загалом через карантин бачитесь, чи просто ви разом не живете та й от ще не хо...
31	Ilijak5	Та похуй, кацапи нахуй разом з пушкіном, чия баба спала біля ефіопа. Пушкін своїм ефіопським приладом втокмачував офранцуженим русскім д...
32	xxewwi	в 2015 ще хоч "Фінес та Ферб" крутили, Ніндзяго, Вінкс, а зараз параша якась
33	Not_Old_Bet...	чи може я просто ще не влюблялась/не найшла ТУ САМУ людину...
34	Not_Old_Bet...	Ну і дрібне доповнення...

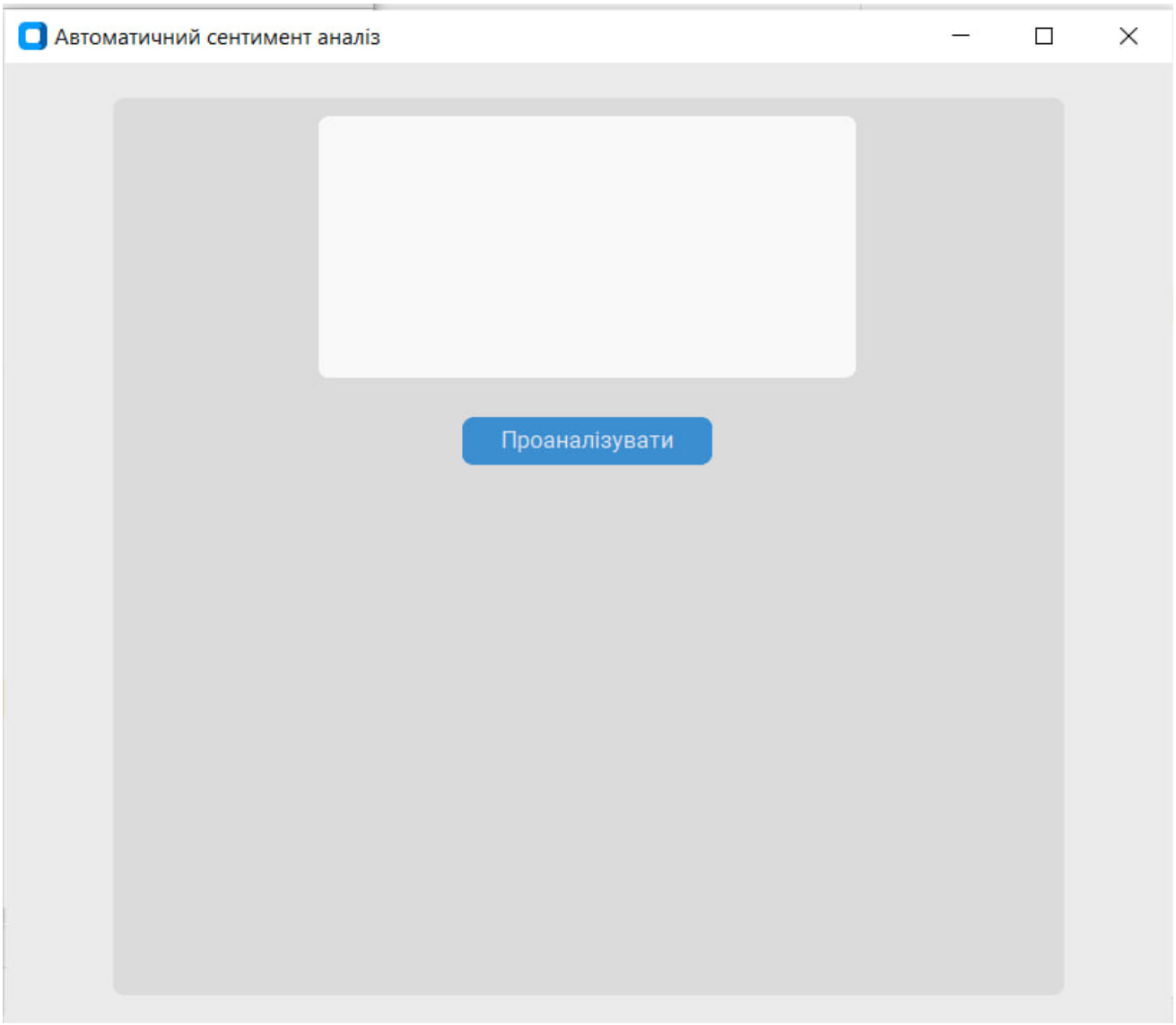
Малюнок 4 **Представлення бази даних немаркованих текстів після обробки**

Наступним етапом створення системи автоматичного сентимент аналізу стало тренування моделі. На ***Малюнку 5*** наведено результати тренування моделі у вигляді кількості виконаних епох навчання та виведені у консоль результати обчислення точності моделі.

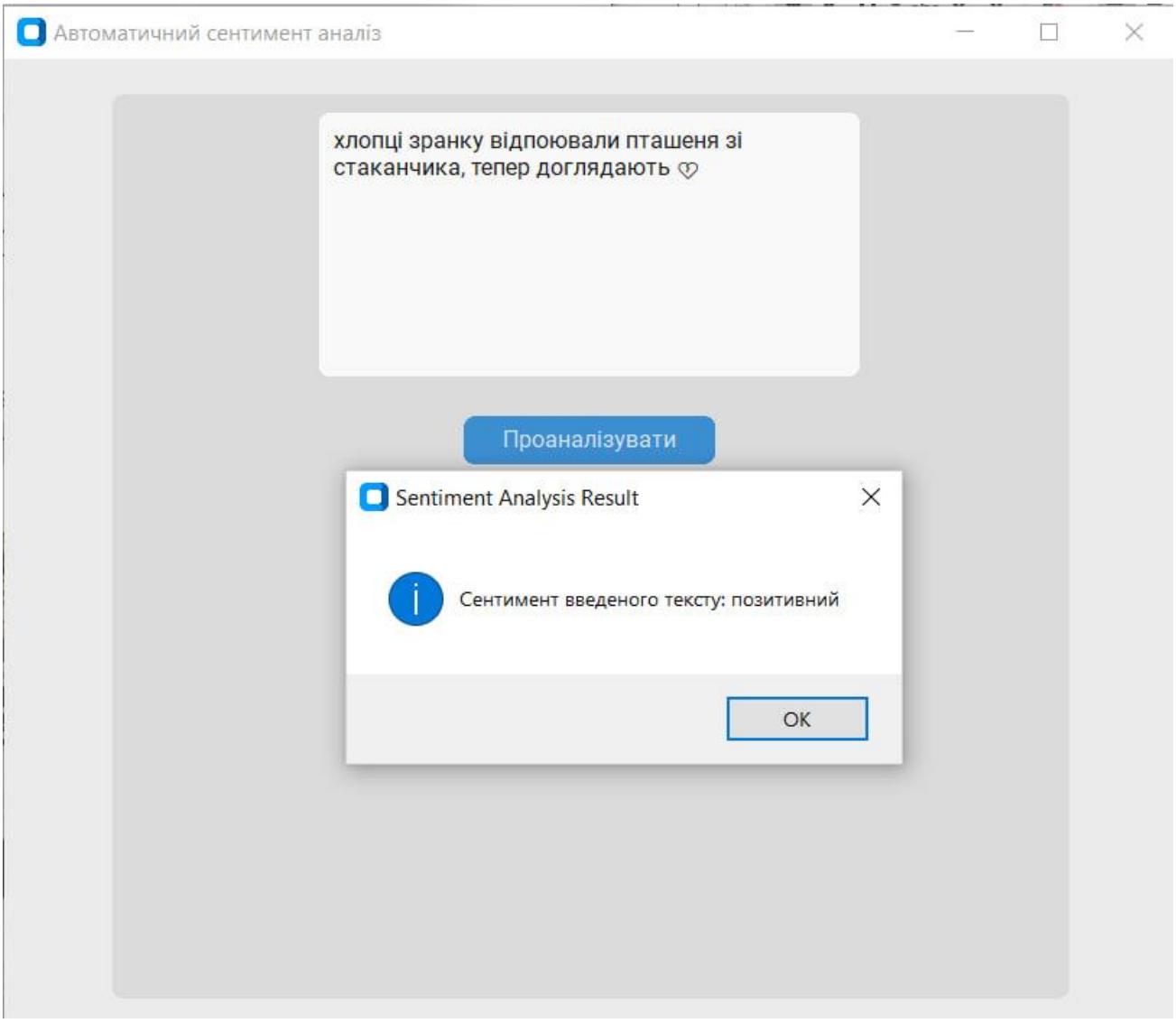
```
Epoch 1/10
Epoch 2/10
Epoch 3/10
Epoch 4/10
Epoch 5/10
Epoch 6/10
Epoch 7/10
Epoch 8/10
Epoch 9/10
Epoch 10/10
Accuracy: 0.998330550918197

Process finished with exit code 0
```

Малюнок 5 **Представлення результатів тренування моделі**

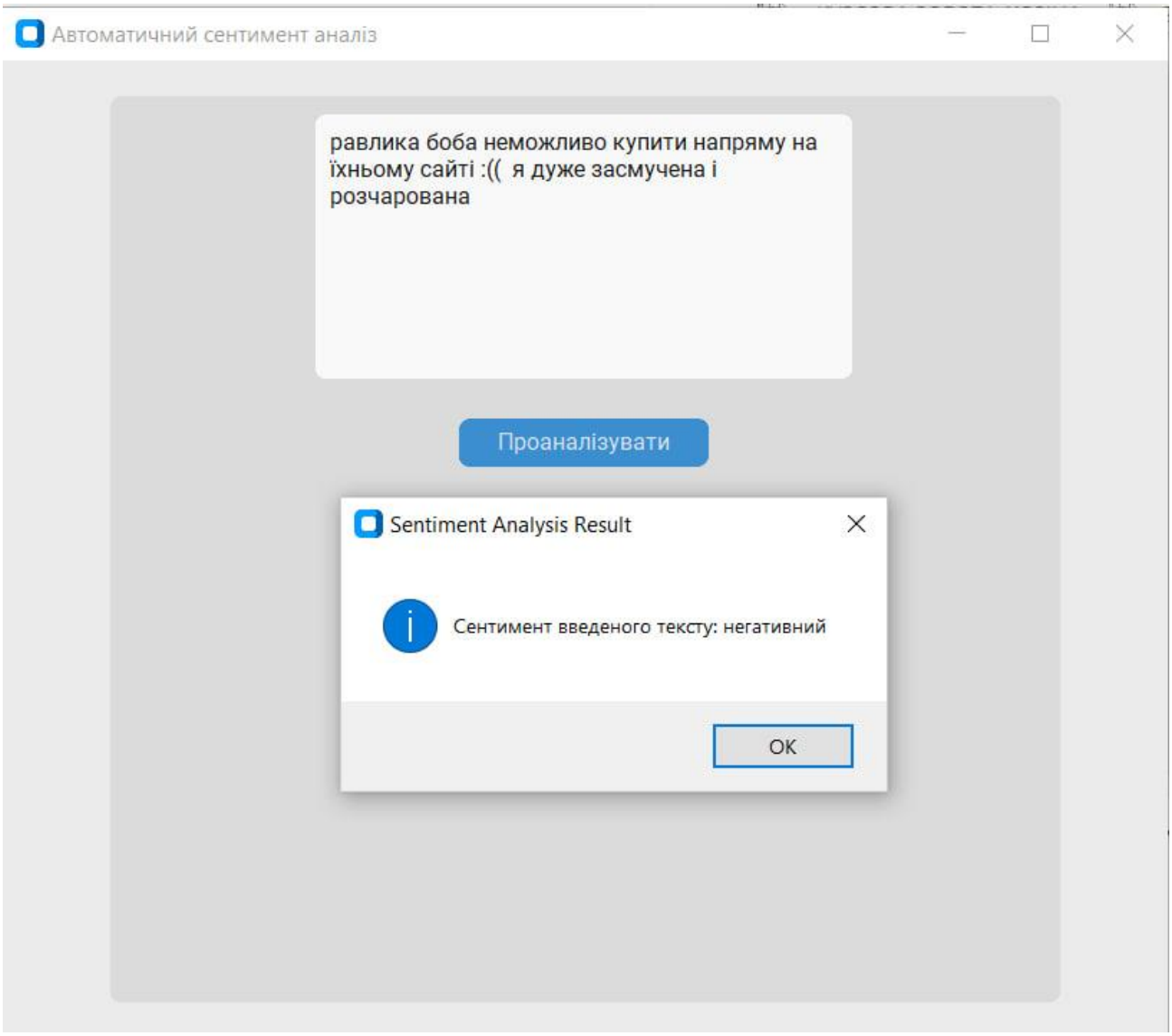
Заключним етапом створення системи автоматичного сентимент аналізу було створення графічного інтерфейсу користувача. На ***Малюнку 6*** наведено створений графічний інтерфейс, який під'єднано до навченої моделі та результати автоматичного сентимент аналізу текстів, введених користувачем. Інтерфейс являє собою вікно із текстовим полем та кнопкою «Проаналізувати». Після введення користувачем тексту та натискання кнопки результат аналізу виводиться на екран у вигляді спливаючого вікна, що містить повідомлення про сентимент вхідного тексту.



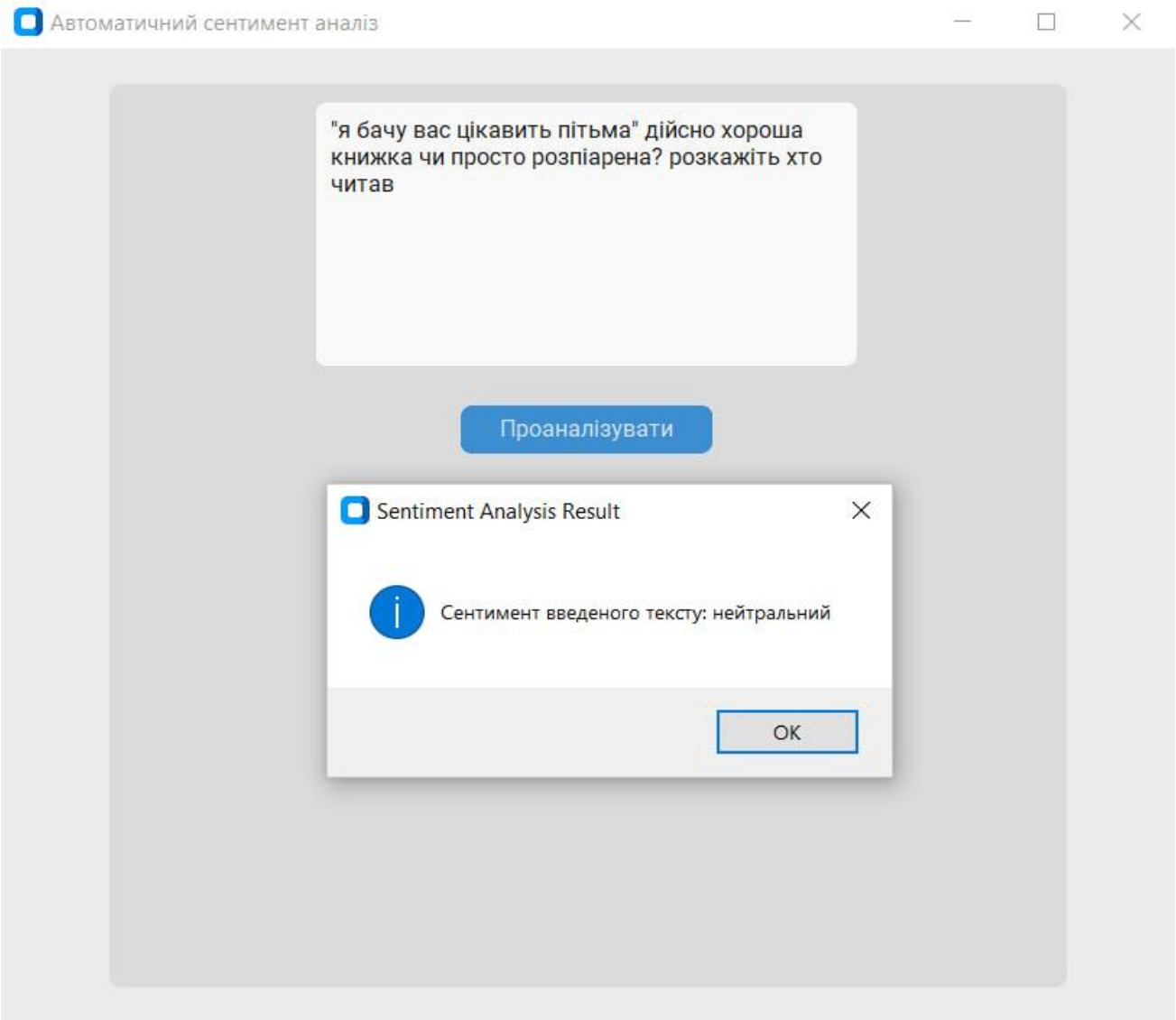
Малюнок 6 **Представлення графічного інтерфейсу користувача до введення тексту**



Малюнок 7 **Представлення графічного інтерфейсу користувача при аналізі позитивного тексту**

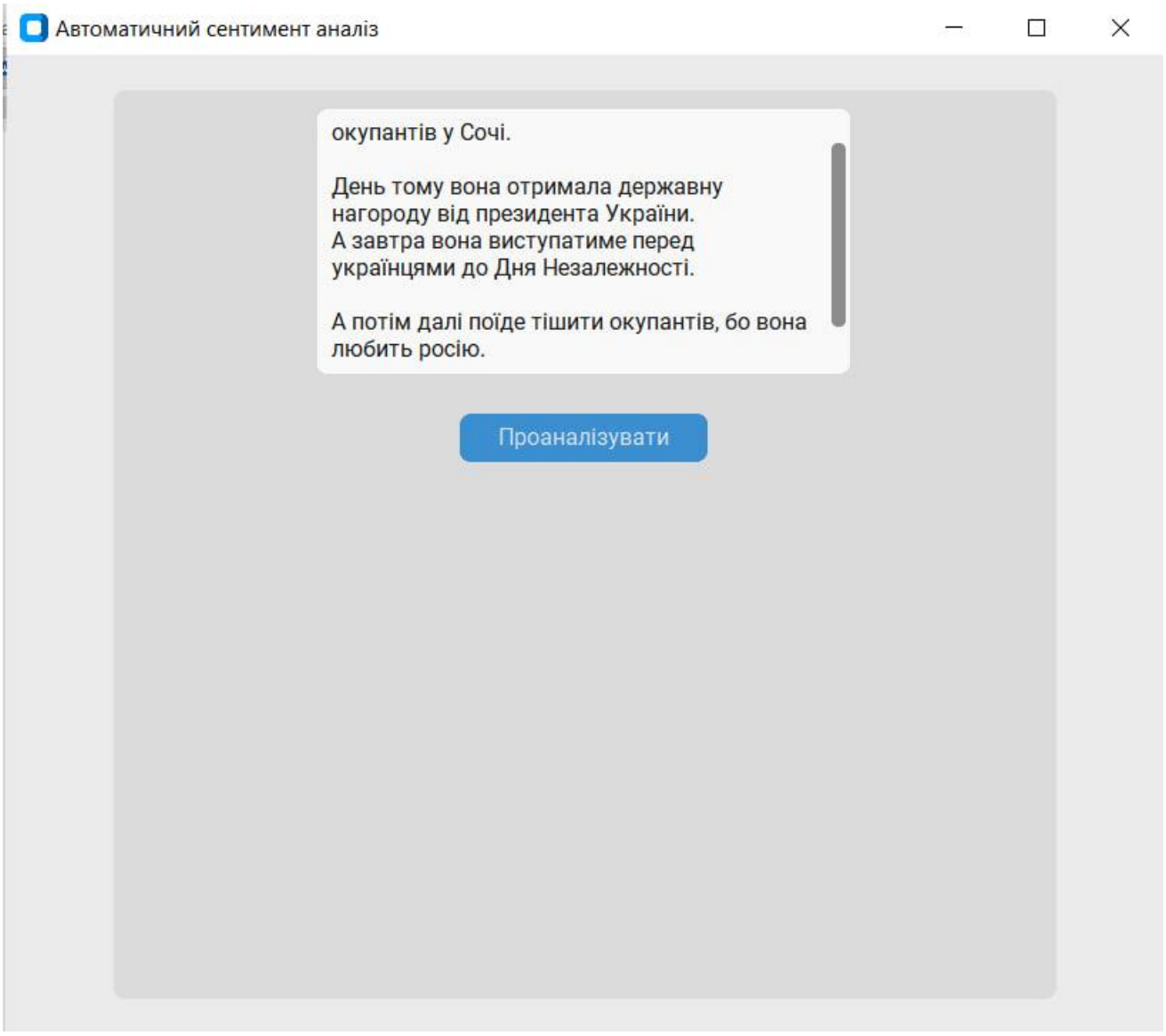


Малюнок 8 **Представлення графічного інтерфейсу користувача при аналізі негативного тексту**



Малюнок 9 **Представлення графічного інтерфейсу користувача при аналізі нейтрального тексту**

При введенні вхідний текст автоматично загортається, а якщо введений текст перевищує розміри вікна, автоматично створюється смуга прокручування.



Малюнок 10 **Представлення смуги прокручування тексту**

ВИСНОВКИ

Проведене дослідження ґрунтується на методах комп’ютерної лінгвістики і демонструє процес створення системи машинного навчання автоматичного сентимент аналізу на основі маркованих та немаркованих текстів із соцмережі Twitter.

У процесі дослідження було визначено проблеми та завдання процесу автоматичного сентимент аналізу, досліджено підходи вирішення задачі автоматичного сентимент аналізу, обрано оптимальний підхід для проведення власного дослідження, створено базу даних маркованих та немаркованих текстів із соцмережі Twitter українською мовою, на основі якої створено програмне забезпечення системи машинного навчання автоматичного сентимент аналізу.

Створена система базується на вхідних маркованих даних, а отже безпосередньо залежить від якості їхнього анотування, було визначено, що для покращення результатів роботи системи необхідна наявність кількох експертів-анотаторів з різних соціальних груп та культур для забезпечення неупередженої оцінки сентименту вхідного тексту. На якість результатів безпосередньо впливають об’єми вхідних даних та посередньо впливають обчислювальні потужності пристрою на якому проводиться навчання моделі, оскільки опрацювання більших об’ємів даних на пристрої зі слабким процесором може зайняти більше часу, а невелика кількість оперативної пам’яті може пристрою може призвести до збою та передчасної зупинки роботи програми, що і

відбувалося під час проведення дослідження до встановлення відповідних гіперпараметрів моделі.

При анотуванні вхідних маркованих даних виникали певні проблеми з визначенням сентименту тексту, оскільки частина з них була відповідями на повідомлення інших користувачів, а отже частково вирвана із контексту, тому для вирішення цієї задачі було обрано входження, сентимент яких був беззаперечно зрозумілий. Так само були опрацьовані тексти, які потенційно могли містити іронію чи сарказм, оскільки наміри користувачів, які публікували такі тексти не були до кінця прозорі, такі входження також відкидалися.

Створена система демонструє вирішення завдання автоматичного сентимент аналізу текстів із соцмережі Twitter українською мовою. Результати роботи системи є задовільними, з обрахованою точністю 0.99, проте з огляду на специфіку вхідних даних та особливостей їх маркування, можемо стверджувати, що система потребує постійного доповнення неупереджено анотованими текстами та оновлення текстами, що містять Інтернет-сленг, який постійно поповнюється.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Дарчук, Н., 2019. Лінгвістичні засади автоматичного сентимент-аналізу українськомовного тексту. Science and education a new dimension, 189, pp.10-13.
2. Шингалов, Д.В., Мелешко, Є.В., Минайленко, Р.М. and Резніченко, В.А., 2017. Методи автоматичного аналізу тональності контенту у соціальних мережах для виявлення інформаційно-психологічних впливів.
3. Ялова, К., Яшина, К., Говорущенко, Т. and Тарасюк, О., 2021. Сентимент аналіз засобами нейронної мережі. Математичне моделювання, (1 (44)), pp.30-37.
4. Agarwal, A., Xie, B., Vovsha, I., Rambow, O. and Passonneau, R.J., 2011, June. Sentiment analysis of twitter data. In Proceedings of the workshop on language in social media (LSM 2011) (pp. 30-38).
5. Angiani, G., Ferrari, L., Fontanini, T., Fornacciari, P., Iotti, E., Magliani, F. and Manicardi, S., 2016, September. A Comparison between Preprocessing Techniques for Sentiment Analysis in Twitter. In KDWeb.
6. Catelli, R., Pelosi, S. and Esposito, M., 2022. Lexicon-based vs. Bert-based sentiment analysis: A comparative study in Italian. Electronics, 11(3), p.374.
7. Chiorrini, A., Diamantini, C., Mircoli, A. and Potena, D., 2021, March. Emotion and sentiment analysis of tweets using BERT. In EDBT/ICDT Workshops (Vol. 3).
8. Dang, N.C., Moreno-García, M.N. and De la Prieta, F., 2020. Sentiment analysis based on deep learning: A comparative study. Electronics, 9(3), p.483.

9. Delobelle, P. and Berendt, B., 2019. Time to take emoji seriously: They vastly improve casual conversational models. arXiv preprint arXiv:1910.13793.

10. Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

11. Haddi, E., Liu, X. and Shi, Y., 2013. The role of text pre-processing in sentiment analysis. Procedia computer science, 17, pp.26-32.

12. Hemalatha, I., Varma, G.S. and Govardhan, A., 2012. Preprocessing the informal text for efficient sentiment analysis. International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), 1(2), pp.58-61.

13. Kouloumpis, E., Wilson, T. and Moore, J., 2011. Twitter sentiment analysis: The good the bad and the omg!. In Proceedings of the international AAAI conference on web and social media (Vol. 5, No. 1, pp. 538-541).

14. Liu, B., 2012. Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), pp.1-167.

15. Miller, H., Thebault-Spieker, J., Chang, S., Johnson, I., Terveen, L. and Hecht, B., 2016. “Blissfully happy” or “ready tofight”: Varying interpretations of emoji. In Proceedings of the international AAAI conference on web and social media (Vol. 10, No. 1, pp. 259-268).

16. Shiha, M. and Ayvaz, S., 2017. The effects of emoji in sentiment analysis. Int. J. Comput. Electr. Eng.(IJCEE.), 9(1), pp.360-369.

17. Taboada, M., Brooke, J., Tofiloski, M., Voll, K. and Stede, M., 2011. Lexicon-based methods for sentiment analysis. Computational linguistics, 37(2), pp.267-307.

ЕЛЕКТРОННІ ДЖЕРЕЛА

18. About Twitter Blue [Електронний ресурс]. — Режим доступу: https://help.twitter.com/en/using-twitter/twitter-blue

19. Analysis of 11 Billion Mentions: Social Media is More Negative Than Ever [Електронний ресурс]. — Режим доступу: https://mention.com/en/blog/social-media-mentions-analysis

20. Apple's iOS 16.4 is out now. These are all the new emojis available on your iPhone [Електронний ресурс]. — Режим доступу: https://eu.usatoday.com/story/tech/2023/03/28/ios-update-emoji-apple/11555381002/

21. Common languages used for web content 2023, by share of websites [Електронний ресурс]. — Режим доступу: https://www.statista.com/statistics/262946/most-common-languages-on-the-internet/

22. Counting characters when composing Tweets [Електронний ресурс]. — Режим доступу: https://developer.twitter.com/en/docs/counting-characters

23. CSViewer [Електронний ресурс]. — Режим доступу: https://csviewer.com

24. customtkinter [Електронний ресурс]. — Режим доступу: https://github.com/TomSchimansky/CustomTkinter

25. Hateful conduct policy [Електронний ресурс]. — Режим доступу: https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

26. How Many Emojis are There in Total? (Emoji Count) [Електронний ресурс]. — Режим доступу: https://www.webnots.com/how-many-emojis-are-there-in-total/

27. bert-base-multilingual-cased [Електронний ресурс]. — Режим доступу: https://huggingface.co/bert-base-multilingual-cased

28. HuggingFace [Електронний ресурс]. — Режим доступу: https://huggingface.co

29. Microsoft Windows 11 November 2021 Update [Електронний ресурс]. — Режим доступу: https://emojipedia.org/microsoft/windows-11-november-2021-update

30. pandas [Електронний ресурс]. — Режим доступу: https://pandas.pydata.org

31. Pycharm [Електронний ресурс]. — Режим доступу: https://www.jetbrains.com/ru-ru/pycharm/

32. re — Regular expression operations [Електронний ресурс]. — Режим доступу: https://docs.python.org/3/library/re.html

33. Sentiment Analysis Model Using BERT-transformer with Monitoring in Production and Continuous Training [Електронний ресурс]. — Режим доступу: https://cnvrg.io/mlcon-1/mojtaba-farmanbar-ing/

34. sklearn [Електронний ресурс]. — Режим доступу: https://scikit-learn.org/stable

35. snscrape [Електронний ресурс]. — Режим доступу: https://github.com/JustAnotherArchivist/snscrape

36. Tkinter [Електронний ресурс]. — Режим доступу: https://docs.python.org/3/library/tkinter.html

37. torch [Електронний ресурс]. — Режим доступу: https://pypi.org/project/torch/

38. transformers [Електронний ресурс]. — Режим доступу: https://pypi.org/project/transformers/

39. Tweepy [Електронний ресурс]. — Режим доступу: https://docs.tweepy.org/en/stable/

40. Twemoji [Електронний ресурс]. — Режим доступу: https://twemoji.twitter.com/

41. Unicode [Електронний ресурс]. — Режим доступу: https://home.unicode.org

42. Welcome to Python.org [Електронний ресурс]. — Режим доступу: https://www.python.org/

ДОДАТКИ

Додаток А. Текстовий опис алгоритму функцій програми попередньої обробки вхідних даних

1. Імпортувати необхідні бібліотеки
2. Визначити вхідний файл з маркованими даними
3. Визначити вихідний файл для оброблених маркованих даних
4. Визначити цільову колонку
5. Зчитати вхідний файл з маркованими даними
6. Застосувати регулярний вираз та видалити URL-посилання та теги користувачів
7. Видалити повторювані рядки
8. Зберегти вихідний файл з обробленими маркованими даними
9. Визначити вхідний файл з немаркованими даними
10. Визначити вихідний файл для оброблених немаркованих даних
11. Визначити цільову колонку
12. Визначити специфічні символи
13. Зчитати вхідний файл з немаркованими даними
14. Застосувати регулярний вираз та видалити URL-посилання та теги користувачів
15. Видалити першу колонку
16. Видалити рядки, які містять специфічні символи
17. Видалити повторювані рядки
18. Зберегти вихідний файл з обробленими немаркованими даними

Додаток Б. Текстовий опис алгоритму функцій програми навчання моделі автоматичного сентимент аналізу

1. Імпортувати необхідні бібліотеки
2. Перевірити чи доступний графічний процесор з підтримкою CUDA, якщо так - перевести пристрій у режим CUDA, якщо ні – у режим CPU
3. Визначити клас *SentimentDataset* для створення власного набору даних для тренування
   3.1. ***Визначаємо конструктор класу, який ініціалізує об'єкт класу із заданими парметрами***
   3.2. Визначити метод *__len__*
   3.3. Визначити метод *__getitem__*
4. У змінній *labeled_data_path* визначити шлях до файлу з обробленими маркованим даними
5. У змінній *labeled_data* зберегти завантажені дані
6. У змінній *unlabeled_data_path* визначити шлях до файлу з обробленими немаркованим даними
7. У змінній *unlabeled_data* зберегти завантажені дані
8. Визначити гіперпараметр максимальна довжина послідовності для токенізації
9. Визначити гіперпарметр розмір пакету
10. Визначити гіперпараметр кількість епох для тренування моделі
11. Визначити параметр темп навчання
1. Змінній *tokeinizer* присвоїти завантажений токенайзер
12. Створити екземпляр класу *SentimentDataset labeled_dataset* для маркованих даних
13. Створити об'єкт *DataLoader* для маркованих даних
14. Створити екземпляр класу *unlabeled_dataset* для немаркованих даних

15. Створити об’єкт *DataLoader* для немаркованих даних
16. Змінній model присвоїти завантажену модель для класифікації послідовностей, встановити кількість міток - 3
17. Перемістити модель на вказаний пристрій
18. Ініціалізувати оптимайзер AdamW з вказаним темпом навчання
19. Перевести модель у режим навчання
20. Розпочати цикл за епохами тренування
    - 20.1. Вивести у консоль номер поточної епохи із загальної кількості
    - 20.2. Розпочати цикл з маркованих і немаркованих даних з відповідних об’єктів *DataLoader*
    - 20.3. Отримати тензор *input_ids* маркованого пакету даних та перемістити на пристрій
    - 20.4. Отримати тензор *attention_mask* маркованого пакету даних та перемістити на пристрій
    - 20.5. Отримати тензор *label* маркованого пакету даних та перемістити на пристрій
    - 20.6. Обнулити градієнти оптимізатора
    - 20.7. Подати пакет маркованих даних до моделі
    - 20.8. Отримати значення втрати
    - 20.9. Обчислити градієнти методом зворотного поширення
    - 20.10. Оновити параметри моделі
    - 20.11. Отримати тензор *input_ids* немаркованого пакету даних та перемістити на пристрій
    - 20.12. Отримати тензор *attention_mask* немаркованого пакету даних та перемістити на пристрій
      - 20.12.1. Встановити контекст без обчислення градієнтів
      - 20.12.2. Подати пакет немаркованих даних до моделі
      - 20.12.3. Обчислити псевдомітки

20.12.4. Конвертувати псевдомітки в масив NumPy

20.12.5. Отримати індекси немаркованих пакетів даних та перемістити на пристрій

20.13. Створити копію немаркованих даних з обраними псевдомітками

20.14. Оновити стовпець *label* з мітками в скопійованих немаркованих даних

20.15. Об’єднати марковані та псевдомарковані дані в один датасет

20.16. Створити новий датасет з об’єднаними даними

20.17. Створити об’єкт *DataLoader* для об’єднаних даних

20.18. Оновити параметри моделі, виконати оптимізацію методом зворотного поширення

21. Перевести модель у режим оцінювання

22. Створити порожній список для прогнозованих міток

23. Створити порожній список для істинних міток

24. ?

24.1. Отримати тензор *input_ids* пакету даних та перемістити на пристрій

24.2. Отримати тензор *attention_mask* пакету даних та перемістити на пристрій

24.3. Отримати тензор *label* пакету даних та перемістити на пристрій

24.4. Подати пакет даних до моделі

24.5. Отримати логіти

24.6. Знайти індекси найімовірніших класів

24.7. Додати прогнозовані мітки у список *eval_predictions*, конвертувати дані у масив NumPy

24.8. Додати істині мітки у список *eval_labels*, конвертувати дані у масив NumPy

25. Обчислити точність за допомогою функції *accuracy_score*

26. Вивести результати обчислень на екран

27. Зберегти натреновану модель у директорії sentiment_model
28. Зберегти токенайзер у директорії sentiment_model

Додаток В. Текстовий опис алгоритму функцій програми створення графічного інтерфейсу користувача

2. Імпортувати необхідні бібліотеки
3. Створити вікно графічного інтерфейсу
4. Задати параметри вікна графічного інтерфейсу 650x550
5. Задати заголовок вікна графічного інтерфейсу «Автоматичний сентимент аналіз»
6. Встановити темний режим зовнішнього вигляду
7. Встановити блакитну тему
8. *Функція sentiment_analysis:*
   - 8.1. Задати змінну *text,* зчитати текст текстового поля від першого до останнього символу у змінній
   - 8.2. Токенізувати текст за допомогою функціїї *tokenizer* за вказаними параметрами
     - 8.2.1. Визначити вхідний текст
     - 8.2.2. Доповнити токенізований текст до максимальної довжини
     - 8.2.3. Усікти токенізований текст, якщо він перевищує максимальну довжину
     - 8.2.4. Встановити максимальну довжину послідовності
     - 8.2.5. Токенізований текст повернути назад у вигляді PyTorch тензорів
   - 8.3. Не зберігати градієнти
     - 8.3.1. Отримати токенізований текст як вхідні значення для моделі
     - 8.3.2. З вихідних значень обрати клас з найвищою вірогідністю, отримати і зберегти отримане числове значення

8.4. Створити словник *sentiment_labels* з відповідним мітками ентименту для числових значень

8.5. Присвоїти вхідному тексту сентимент

8.6. У спливаючому вікні вивести повідомлення про сентимент вхідного тексту

9. Визначити шлях до збереженої моделі

10. Змінній *model* присвоїти збережену модель

11. Змінній *tokeinizer* присвоїти збережений токенайзер за вказаним шляхом

12. Створити віджет фрейм у вікні інтерфейсу

13. Налаштувати параметри розміщення віджету фрейм

14. Створити віджет текстового поля, визначити розміри віджету та перенесення тексту

15. Налаштувати параметри розміщення віджету текстове поле

16. Створити віджет кнопки, задати текст віджету, виконувати функцію *sentiment_analysis* при натисканні

17. Налаштувати параметри розміщення віджету кнопка

18. Запустити основний цикл програми

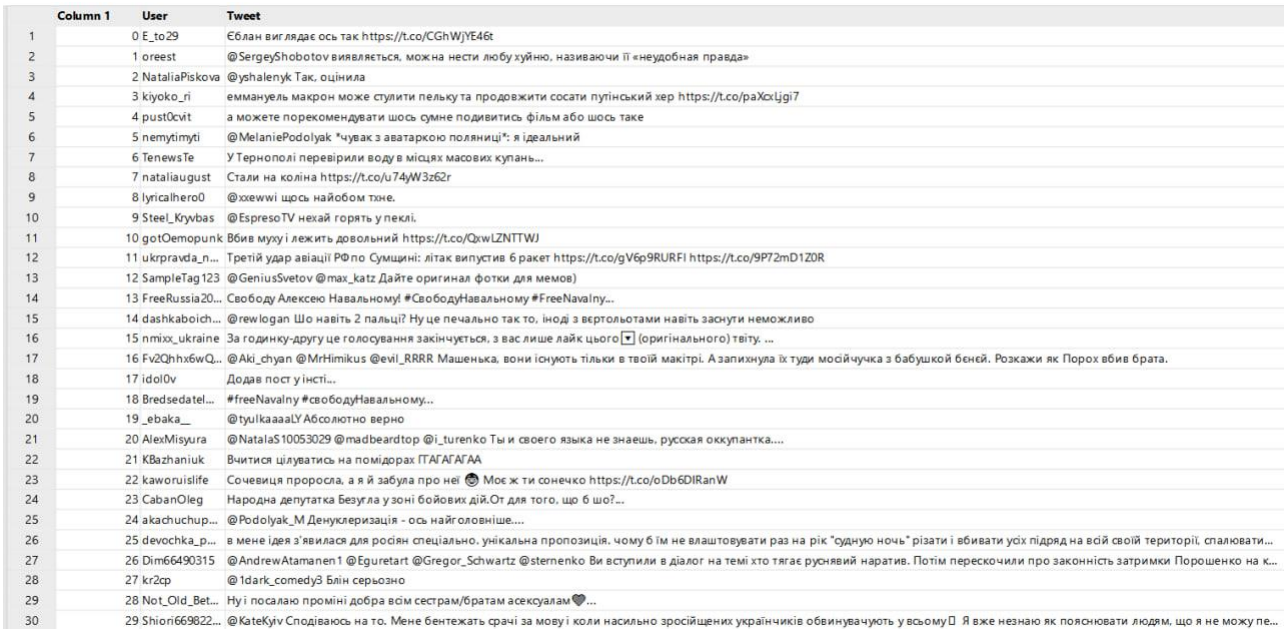
Додаток Г. Фрагмент бази даних маркованих текстів (файлу labeled_data.csv)

У додатку представлено скріншот вигляду csv-файлу labeled_data. Файл містить марковані дані до попередньої обробки. Перша колонка – ім’я користувача, друга – вміст твіту, третя – надана оцінка.

	A	B	C
1	User	Tweet	label
2	__agusha__di__	ранок!! гарних вам вихідних 😬♡ https://t.co/LX6QWXAkEE	2
3	__asyya__	🥺 Вам реально так сподобався мій малюнок? □ Безмежно дякую всім котикам, за такі прекрасні слова💖. Моя мотивація малювати зараз просто злетить до небес 🕳️ ❤️ Люблю вас, мої кошенята, всіх цьомаю 😊~💖 https://t.co/5K1ssT6jkp https://t.co/QESF29t4aD	2
4	__Mabelel__	@HochuDeruniv Baaaay це крутезно! Надіюсь це буде гарне і круте місце, удачі!!	2
5	_Anhelina_a	@Robin_z_Fajnogo Доброго ранку, Руслан. 🫖🧀 Спокійного дня нам усім 😊	2
6	_Irunka1408	Дізналась, що чувак просто утилізує батарейки з одноразок і робить з них павери для ЗСУ. Це геніально. https://t.co/NjYWMp4pVJ	2
7	_kolizhanka	@yosei_sm яка краса!!	2
8	_kotuk_	@Bulba_Men Ну я і раніше їх знав і чув, але вчора конкретно послухал багато пісень. Є ще львівський гурт Joryj Кюс, вони просто бомба. 😻	2
9	10nyk	От би було класно в якомусь спецпроєкті закроссоверити канал "Обличчя Незалежності" @DarkaHirna з каналом "На пошуки грамоти" @cabbage_sad 🤩 https://t.co/HyeMOCHGKk	2
10	1705Belka	Вчора дві години слухала лекції з палеобіології від університету Альберти, потім годину дивилась на ютубі підбірку нових динозаврів Jurassic World Evolution 2 у 4К, і вважаю, що це ідеальний вечір 😊🦕 @Paleo_Daddy, для повного щастя не вистачає лекції від Вас, чекаю 😐	2
11	1femmefatale_	@jjhopiie ЖІНКОООО, ти неймовірна!❤️ Обов'язково пиши, якщо потрібно допомогти, ДУЖЕ тобою пишаюся😬	2
12	4EdYCuQMCtkZmef	Роки 4чи5 тому,коли кацапи пропонували Польщі розділити з ними Україну, хтось з польських політиків тихо сказав:"Краще ми з Україною поділимо вас"...Ну,тоді я поґеґала і вспокоїлася...22-й рік,Польща стирає кордон з Україною аби допомагати їбашити русню...Мінє нравіцца😁...	2
13	5Bykodir	@mirvla Морган майже такий як сьогодні)	2
14	74Fraza	Гумор від наших "І чому їм не подобається NLAW? Постріл, і "сусіди" за лічені секунди "розігріваються". Що в них горить, що вони тут... А там могли б тушити" . Так зрозумів, що це про "трактористів" в яких вдома лісові пожежі?	2
15	7kitLea	Від усієї душі бажаю цього кожному з українців! https://t.co/rd95wnftjM	2
16	ada_devil_	Ранку, сонечки 🌞🚗 Вдалого та спокійного вам дня 😬❤️❤️❤️	2
17	AdasssB77	@MirovichMedia ситуація в Україні відображена на голові в Бориса Дж.))) https://t.co/Ov3dJ0ceWl	2
18	advoka_tesa	@326840nk Тримайтеся, Наталю!	2
19	AFormusyak	Нарешті норм погода у Львові. Свіжо. 3бс	2
20	Ailee_hyong	@tenmybb Сексиии	2
21	aitch6ar	@tataserpeny Шліть іще як буде настрій мої повідомлення відкритті для хороших людей	2
22	aitch6ar	@tataserpeny Дякую	2
23	AjzadaSerikova	@fluffynotwitch богиня	2
24	Akina_Hitomi	@VodkaForNat @MiraidaShinomia Клас виглядає так знайомо 😱	2
25	alexalex6169	@Gadzhega @kaliha Це була відсилка до справжнього персонажу😂 https://t.co/kvxFI5HB3I	2
26	Alexand39995840	@frankbuld Ранку, кицюня🫖💙💛	2
27	Alexand39995840	@irinaromashkin1 Раночку, Іруня🫖💙💛	2
28	alexmushak	@ohiAnnablya @uspenovka19632 виживеш, переможеш, виховаєш дітей!	2
29	alina_bondarnk	@lady_shtopor ЦІ ЛЮДИ ❤️❤️❤️	2

Додаток Д. Фрагмент бази даних немаркованих текстів (файлу 500k1.csv)

У цьому файлі містять немарковані дані до попередньої обробки. Перша колонка – ім’я користувача, друга – вміст твіту.



	Column 1	User	Tweet
1	0	E_to29	Єблан виглядає ось так https://t.co/CGhWjYE46t
2	1	oreest	@SergeyShobotov виявляється, можна нести любу хуйню, називаючи її «неудобная правда»
3	2	NataliaPiskova	@yshalenyk Так, оцінила
4	3	kiyoko_ri	еммануель макрон може стулити пельку та продовжити сосати путінський хер https://t.co/paXcxLjgi7
5	4	pust0cvit	а можете порекомендувати шось сумне подивитись фільм або шось таке
6	5	nemytimyti	@MelaniePodolyak "чувак з аватаркою поляниці": я ідеальний
7	6	TenewsTe	У Тернополі перевірили воду в місцях масових купань...
8	7	nataliaugust	Стали на коліна https://t.co/u74yW3z62r
9	8	lyricalhero0	@xxewwi щось найобом тхне.
10	9	Steel_Kryvbas	@EspresoTV нехай горять у пеклі.
11	10	gotOemopunk	Вбив муху і лежить довольний https://t.co/QxwLZNTTWJ
12	11	ukrpravda_n...	Третій удар авіації РФ по Сумщині: літак випустив 6 ракет https://t.co/gV6p9RURFI https://t.co/9P72mD1Z0R
13	12	SampleTag123	@GeniusSvetov @max_katz Дайте оригинал фотки для мемов)
14	13	FreeRussia20...	Свободу Алексею Навальному! #СвободуНавальному #FreeNavalny...
15	14	dashkaboich...	@rewlogan Шо навіть 2 пальці? Ну це печально так то, іноді з вертольотами навіть заснути неможливо
16	15	nmixx_ukraine	За годинку-другу це голосування закінчується, з вас лише лайк цього ▾ (оригінального) твіту. ...
17	16	Fv2Qhhx6wQ...	@Aki_chyan @MrHimikus @evil_RRRR Машенька, вони існують тільки в твоїй макітрі. А запихнула їх туди мосійчучка з бабушкой бенеї. Розкажи як Порох вбив брата.
18	17	idol0v	Додав пост у інсті...
19	18	Bredsedatel...	#freeNavalny #свободуНавальному...
20	19	_ebaka__	@tyulkaaaaLY Абсолютно верно
21	20	AlexMisyura	@NataliaS10053029 @madbeardtop @i_turenko Ты и своего языка не знаешь, русская оккупантка....
22	21	KBazhaniuk	Вчитися цілуватись на помідорах ІТАГАГАГАА
23	22	kaworuislife	Сочевиця проросла, а я й забула про неї 🧐 Мое ж ти сонечко https://t.co/oDb6DIRanW
24	23	CabanOleg	Народна депутатка Безугла у зоні бойових дій.От для того, що б шо?...
25	24	akachuchup...	@Podolyak_M Денуклеризація - ось найголовніше....
26	25	devochka_p...	в мене ідея з'явилася для росіян спеціально, унікальна пропозиція, чому б їм не влаштовувати раз на рік "судную ночь" різати і вбивати усіх підряд на всій своїй території, спалювати...
27	26	Dim66490315	@AndrewAtamanen1 @Eguretart @Gregor_Schwartz @sternenko Ви вступили в діалог на темі хто тягає руснявий наратив. Потім перескочили про законність затримки Порошенко на к...
28	27	kr2cp	@1dark_comedy3 Блін серьозно
29	28	Not_Old_Bet...	Ну і посилаю проміні добра всім сестрам/братам асексуалам💜...
30	29	Shiori669822...	@KateKyiv Сподіваюсь на то. Мене бентежать срачі за мову і коли насильно зросійщених українчиків обвинувачують у всьому Я вже незнаю як пояснювати людям, що я не можу пе...