Міністерство освіти і науки України
Тернопільський національний технічний університет імені Івана Пулюя

(повна назва факультету)

(повна назва кафедри)

КВАЛІФІКАЦІЙНА РОБОТА

на здобуття освітнього ступеня

(назва освітнього ступеня)

на тему: Алгоритмічне та програмне забезпечення систем автоматизованого оцінювання емоційного нахилу статей про Україну

Виконав(ла): студент(ка) 6 курсу, групи СІм-61

спеціальності 123 Комп'ютерна інженерія

(шифр і назва спеціальності)

	(підпис)	Кохан В.В.Б. (прізвище та ініціали)
Керівник	(підпис)	(прізвище та ініціали)
Нормоконтроль	(підпис)	(прізвище та ініціали)
Завідувач кафедри	(підпис)	(прізвище та ініціали)
Рецензент	(підпис)	(прізвище та ініціали)

Тернопіль
20_21_

Міністерство освіти і науки України
Тернопільський національний технічний університет імені Івана Пулюя

Факультет _____
(повна назва факультету)

Кафедра _____
(повна назва кафедри)

ЗАТВЕРДЖУЮ
Завідувач кафедри

_____ _____
(підпис) (прізвище та ініціали)

« » 20___ р.

З А В Д А Н Н Я
НА КВАЛІФІКАЦІЙНУ РОБОТУ

на здобуття освітнього ступеня _____
(назва освітнього ступеня)

за спеціальністю _____
(шифр і назва спеціальності)

студенту _____
(прізвище, ім'я, по батькові)

1. Тема роботи _____

_____

_____

_____

_____

Керівник роботи _____
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

Затверджені наказом ректора від «___» _____ 20___ року № _____

2. Термін подання студентом завершеної роботи _____

3. Вихідні дані до роботи _____

_____

_____

4. Зміст роботи (перелік питань, які потрібно розробити)

_____

_____

_____

_____

_____

_____

5. Перелік графічного матеріалу (з точним зазначенням обов'язкових креслень, слайдів)

_____

_____

_____

_____

_____

_____

6. Консультанти розділів роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв

7. Дата видачі завдання _____

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів роботи	Термін виконання етапів роботи	Примітка

Студент _____ _____
(підпис) (прізвище та ініціали)

Керівник роботи _____ _____
(підпис) (прізвище та ініціали)

АНОТАЦІЯ

Алгоритмічне та програмне забезпечення систем автоматизованого оцінювання емоційного нахилу статей про Україну // Кохан Василь Володимир Богданович // Тернопільський національний технічний університет імені Івана Пулюя, факультет комп'ютерно - інформаційних систем та програмної інженерії, кафедра комп'ютерних систем та мереж, група СІм-61 // Тернопіль, 2021 // с. - 70, рис. – 33, бібліогр. - 33.

Ключові слова: оцінка емоційного нахилу, Україна, Твіттер, твіт, АПІ, Твіттер АПІ, датасет.

Мета цієї роботи полягає у дослідженні та розробці алгоритму для оцінки емоційного нахилу текстів про Україну з соціальних мереж, дослідженні методів до проведення подібних оцінок, дослідженні способів збору даних для дослідження.

Для отримання інформації про поточну ситуацію та динаміку зміни оцінки емоційного нахилу використовуються актуальні дані зібрані безпосередньо у процесі написання роботи. Для швидшого опрацювання даних та скорочення процесу підготовки – використовуються словникові методи аналізу емоційного нахилу текстів, проте через складність оцінки об'єктивності та нижчі показники точності оцінки у роботі проводиться аналіз із використанням трьох різних словників.

Створено алгоритми збору, опрацювання та аналізу даних для оцінки динамки емоційного нахилу текстів про Україну в соціальній мережі Твіттер за останні 15 років.

ANNOTATION

Algorithms and software for automated sentiment analysis of articles about Ukraine // Kokhan Vasyl Volodymyr Bohdanovych // Ternopil Ivan Puluj National Technical University, Faculty of Computer Information Systems and Software Engineering, Department of Computer Systems and Networks, Group SIm-61 // Ternopil, 2021 // p. - 65, fig. - 33, bibliogr. - 33.

Key words: sentiment anaysis, Ukraine, Twitter, tweet, API, Twitter API, dataset.

The purpose of this work is to study and develop an algorithm for sentiment analysis of texts about Ukraine from social networks, the study of methods for conducting such analysis, the study of ways to collect data for research.

To obtain information about the current situation and the dynamics of changes in the sentiment analysis, up to date data collected directly in the process of writing the work. Vocabulary methods of sentiment analysis of texts are used to speed up data processing and shorten the preparation process, but due to the complexity of assessing the objectivity and lower accuracy of such assessment, the analysis is performed using three different dictionaries.

Algorithms for collecting, processing and analyzing data to analyse sentiment of texts about Ukraine on the social network Twitter over the past 15 years were developed at the time of writing this paper.

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ

Гідрація – процес завантаження повних доступних твітів через Твіттер АПІ маючи перелі їх унікальних ідентифікаторів.

Датафрейм – об’єкт Python бібліотеки pandas що представляє собою таблицю значень

Датасет – набір однотипних даних

Твіт – повідомлення у соціальній мережі мікроблогів Твіттер

Токен – слово з тексту

API (АПІ) – (Application Programming Interface), набір підпрограм та протоколів взаємодії що використовуються для розробки програмних комплексів та їх окремих частин.

ЗМІСТ

ВСТУП

Актуальність теми: ми живемо у час стрімкого руху та розвитку цифрових технологій, де щодня генеруються тисячі та мільйони нових одиниць даних. Провідним компаніям, які колись могли найняти людей для аналізу тенденцій ринку та оцінки ставлення суспільства до бренду, через великий обсяг даних зробити це дедалі складніше. Саме у такі моменти в пригоді стають комп'ютерні алгоритми оцінки емоційного нахилу текстів, які дозволяють зробити опрацювання великих обсягів даних значно швидшим.

Для дослідження емоційного нахилу текстів використовують алгоритми машинного навчання та статистичні методи. Такими алгоритми користуються такі провідні компанії як Google, Facebook, Amazon, Apple, Netflix для аналізу своїх статей, постів, фільмів чи продуктів у інтернет магазині, а дослідженнями у цій галузі зокрема займаються вчені X. J. Zeng, J. Bollen, L. Lee, E. Cambria з багатьох країн світу від Індії до Америки, від Китаю до Іспанії.

Соціальні мережі грають значну роль у повсякденному житті сучасних людей. Окремим аспектом їх впливу на користувачів – є формування думок на різні питання, зокрема, на ті, що стосуються конфлікту між Україною та Росією. Проте, ще не було проаналізовано динаміки зміни оцінки емоційного нахилу публікацій у соціальних мережах про Україну. Реалізація такого алгоритму та практичне його втілення у програмі дозволить зрозуміти як світова спільнота реагувала та продовжує реагувати на події в Україні, які з подій були найрезонанснішими.

Актуальність теми пояснюється відсутністю алгоритмів оцінок емоційного нахилу публікацій у соціальних мережах про Україну та необхідністю зрозуміти характер настрою (позитивний, негативний чи нейтральний), що існує у соціальних мережах, по відношенню до стану справ в Україні.

Мета і завдання дослідження: огляд динаміки рівня емоційного нахилу в текстах про Україну соціальної мережі Твіттер за останні 15 років. Розробка

алгоритму та програмного забезпечення для проведення автоматизованої оцінки статей з соціальної мережі мікроблогів Твіттер.

Для досягнення поставленої мети необхідно виконати наступні завдання:

- проаналізувати наявні алгоритми оцінки емоційного нахилу тексту, що використовують дані з соціальних мереж;
- зібрати тексти про Україну через програмне API;
- підготувати отримані дані до аналізу;
- написати алгоритм для аналізу та його програмну реалізацію;
- провести аналіз та сформувати результати.

Об'єкт дослідження: зміна емоційного нахилу текстів про Україну у соціальній мережі Твіттер за останні 15 років.

Предмет дослідження: алгоритм аналізу емоційного нахилу.

Методи дослідження: аналіз наявних алгоритмів оцінки емоційного нахилу текстів та їх програмних реалізацій; узагальнення проаналізованого матеріалу; програмування нової програмної реалізації алгоритму оцінки емоційного нахилу текстів; проведення експерименту над досліджуваними даними новостворенним алгоритмом; порівняння результатів різних підпрограм програми.

Наукова новизна одержаних результатів: вперше алгоритмами оцінювання емоційного нахилу текстів було проаналізовано твіти з тегом «#ukraine» створені на протязі 2006 – 2021 року

Практичне значення одержаних результатів: результати допомагають зрозуміти ставлення користувачів соціальної мережі мікроблогів Твіттер до подій в Україні, напрямки та тенденції його коливання.

Публікації: Результати дослідження було апробовано на X Міжнародній науково-технічній конференції молодих учених та студентів «Актуальні задачі сучасних технологій» [1] та IX Науково-технічній конференції «Інформаційні моделі, системи та технології» [2] у вигляді тез конференцій.

Структура роботи: робота складається з пояснювальної записки та графічної частини. Пояснювальна записка складається із вступу, чотирьох розділів, висновків, списку використаних джерел та додатків. Обсяг роботи:

пояснювальна записка – 62 аркушів формату А4, графічна частина – 8 аркушів формату А1.

РОЗДІЛ 1
АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ ТА ОГЛЯД АЛГОРИТМІВ ДЛЯ ВИЗНАЧЕННЯ ЕМОЦІЙНОГО НАХИЛУ ТЕКСТІВ

1.1. Оцінка емоційного нахилу текстів

Штучний інтелект – популярний на сьогодні напрям проведення багатьох досліджень, який дозволяє робити неймовірні речі з використанням можливостей сучасних комп’ютерних технологій. Виданнями, що спеціалізуються на штучному інтелекті [3], було визначено ключові завдання штучного інтелекту, такі як:

- міркування, вирішення проблем. Ціль – створити систему яка зможе вирішувати задачі з невизначеною або неповною інформацією, розв’язувати – головоломки та логічні задачі;
- представлення знань. Ціль – створити онтологію, систему представлення даних, яка зможе оперувати фактами та їх зв’язками та давати осмислені відповіді на запитання чи робити висновки про реальні факти;
- планування. Ціль – створити систему яка буде усвідомлювати стан справ у навколишньому світі та зможе робити передбачення, оцінювати наслідки свого впливу на систему, при взаємодії з іншими учасниками процесу;
- навчання. Машинне навчання – фундаментальна концепція досліджень штучного інтелекту з моменту заснування галузі [4]. Ціль – створити систему яка зможе накопичувати досвід та на його основ ставати кращою у тому що вона робить;
- обробка природної мови. Ціль – створити систему яка буде розуміти людську, природну мову та зможе використовувати її для отримання чи передачі інформації;
- сприйняття. Ціль – створити систему яка зможе використовувати інформацію з мікрофонів, камер, радарів та інших давачів для сприйняття аспектів навколишнього світу;
- рух і маніпуляції. Ціль – створити систему яка зможе оцінювати свою позицію в невідомому просторі та переміщатись по ньому;

- соціальний інтелект. Ціль – створити систему яка, за допомогою напрацювань для вирішення інших завдань штучного інтелекту, зможе розуміти та імітувати людські емоції та почуття;
- загальний інтелект. Ціль – створити систему яка зможе розв’язувати задачі різної складності як людина. Для виконання завдання загального інтелекту потрібні системи які добре справляються з виконанням усіх вище згаданих завдань, де система загального інтелекту буде виступати основою для застосування інших систем.

Кожне таке завдання, або комбінація завдань, дозволило окреслити великий перелік простіших, конкретніших цілей, рішення яких вже почали впливати на наше життя. Однією із таких цілей є оцінка емоційного нахилу (sentiment analysis), яка утворилось на перетині машинного навчання та обробки природної мови. Алгоритми оцінки емоційного нахилу використовуються для класифікації текстових даних за їх емоційним нахилом, об’єктивністю та ставленням автора тексту до об’єктів про які йде мова. Опрацьовані дані потім застосовуються компаніями та брендами для того, щоб зрозуміти як їхні користувачі та фанати сприймають рішення компанії з того чи іншого питання. Результати оцінки емоційного нахилу є свого роду компасом суспільної думки, який показує реакцію суспільства на діяльність компанії, при цьому не потребуючи безперервної роботи десятків чи сотень працівників, які будуть перечитувати усі можливі відгуки від користувачів.

Більшість сучасних систем використовують системи аналізу емоційного нахилу текстів для оцінки лише по одній шкалі значень, яка показує позитивного чи негативного нахилу є аналізований текст. Проте є дослідження які підтверджують можливість оцінки емоційного нахилу за декількома шкалами вимірів такими як тривожність, впевненість, доброзичливість, наполегливість, тощо [5, 6].

Самі шкали значень також можуть бути наступних видів:

- бінарні шкали;
- багатосмугові шкали.

Найпростішим з них – є бінарна шкала. Бінарні шкали можуть приймати лише 2 протилежні значення, як позитивне чи негативне. При використанні таких шкал складно добитися точно результату, адже дуже часто речення може мати слова як позитивного так і негативного емоційного забарвлення, а тому категорична оцінка тексту по бінарній шкалі буде неточною.

Складнішою і зазвичай більш поширеною є багатосмугова шкала значень. Такі шкали можуть приймати одне зі заданої кількості значень між крайніми точками максимуму та мінімуму. Результати оцінювання за багатосмуговими шкалами точніше передають емоційний нахил аналізованих текстів, бо показують міру наближення оцінки тексту до максимально позитивного чи негативного значення.

Існує декілька поширених методів автоматизованої оцінки емоційного нахилу:

- методи засновані на правилах і словниках;
- статистичні методи;
- комбіновані методи.

Методи засновані на правилах і словниках – дозволяють проаналізувати текст за допомогою попередньо складених словників тональностей та правил лінгвістичного аналізу [7]. Суть цього методу полягає у присвоєнні кожному слові значення зі словника, якщо воно є, а за загальну оцінку тексту приймають суму оцінок усіх слів. Хоча безпосередньо застосування цього методу є доволі простим – основна частина роботи припадає на складання словника з правильними вагами слів для досліджуваної галузі. Для прикладу, слово «великий» буде мати позитивне значення, якщо мова буде йти про обсяг пам’яті жорсткого диска і негативне, якщо мова буде йти про розміри телефона.

Статистичні методи через хороші результати в інших завданнях штучного інтелекту набирають дедалі більшої популярності. Сюди належать:

- машинне навчання з вчителем;
- машинне навчання без вчителя;
- підхід заснований на теоретико-графових моделях.

У першому випадку текст розбивають на токени, яким людина присвоює позитивні чи негативні значення. Використовуючи масив наданої інформації – система робить висновок про емоційний нахил наступних текстів цієї ж категорії. Точність такого підходу статистичного методу вища за точність словникового методу, але потребує великого обсягу даних для тренування моделі.

Підхід машинного навчання без вчителя полягає у статистичному аналізі частоти входження токенів у аналізований текст. Тут припускається що токени які часто зустрічаються у даному тексті та присутні у інших текстах вибірки – мають найбільший вплив на тональність текстів, а подальша оцінка тональності отримується з тональності обраних токенів та частоти їх появи.

Підхід теоретико-графових моделей припускає що різні слова мають різний вплив на емоційний нахил тексту, тому потребує створення спеціальних графів на основі досліджуваного тексту, які потім проходять процес ранжування вершин, класифікації знайдених слів і лише після того дозволяють отримати результат.

Суть останнього, комбінованого, методу аналізу тональності текстів полягає у поєднанні частин попередніх методів для виконання лише окремих етапів перевірки. Такий метод дозволяє отримати переваги вище описаних методів та підходів, нівелюючи їхні недоліки.

1.2. Аналіз інструментів та публікацій на тему аналізу емоційного нахилу

Публікації у соціальних мережах та статті новин опубліковані у онлайн виданнях – великий пласт інформації, доступної для кожного, яка постійно генерується. Експерти різних галузей використовують їх для дослідження загального настрою економіку в умовах карантину спричиненого COVID-19 [8], чи для передбачення вартостей акцій [9, 10] та ф’ючерсів [11] на біржі.

Із швидким ростом популярності аналізу емоційного нахилу та відносній простоті отримання перших результатів – з’явилось доволі багато компаній, які

почали займатися цим професійно. Унікальним у кожній компанії яка пропонує використання своїх алгоритмів оцінки емоційного нахилу є власне алгоритми, та можливості їх зміни під потреби конкретного користувача.

Для прикладу платформа MonkeyLearn - Text Analysis [12] дозволяє на основі класифікації декількох текстів провести аналіз цілої вибірки. На рисунку 1.1 зображено вікно навчання моделі для класифікації текстів.

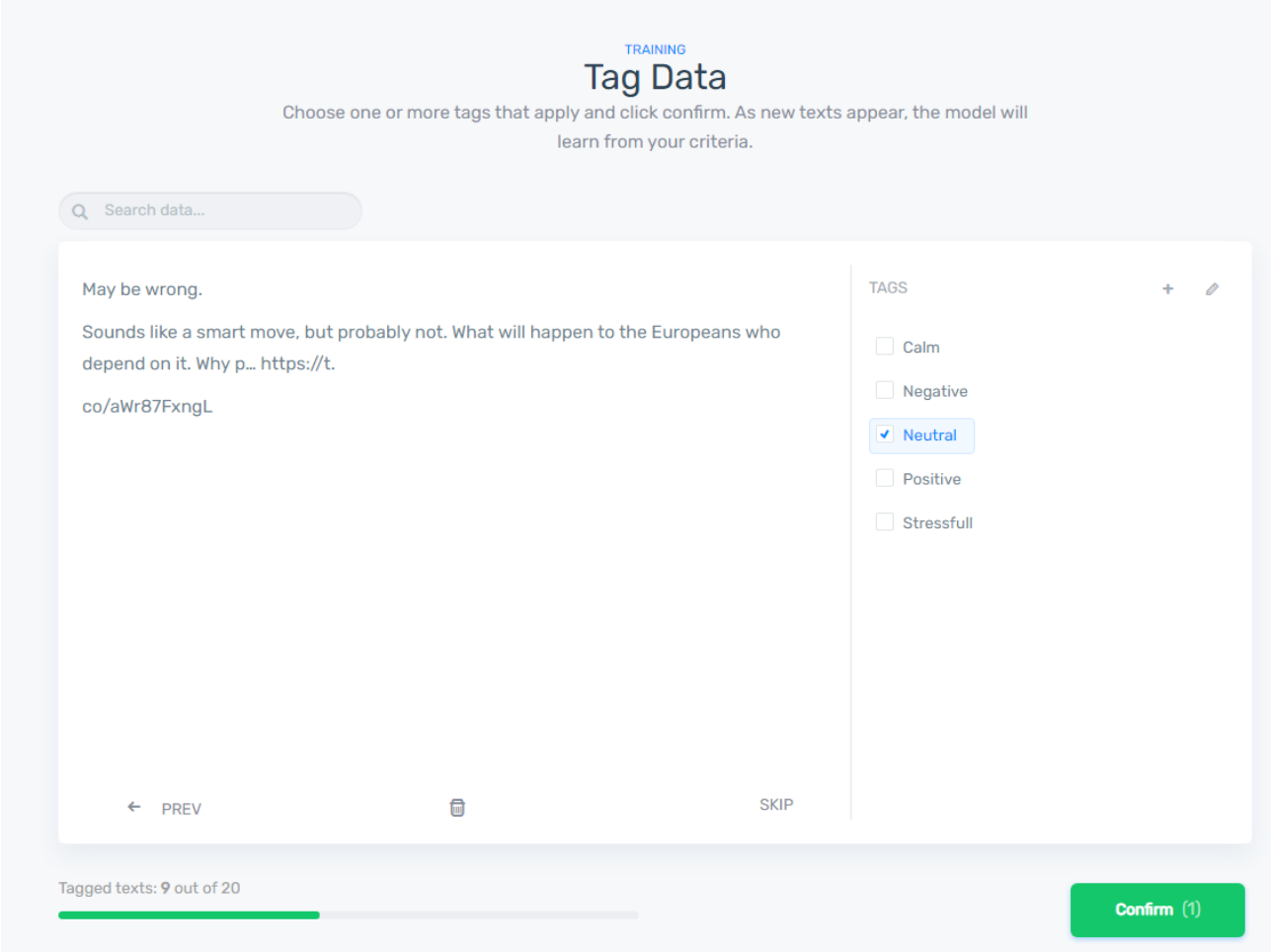


Рис 1.1. Класифікація текстів на платформі MonkeyLearn

Після проходження короткого процесу навчання моделі – вона вже може класифікувати нові тексти на наближену тему. На рисунку 1.2 зображено результат демонстраційного класифікації тексту, якого модель ще не бачила.

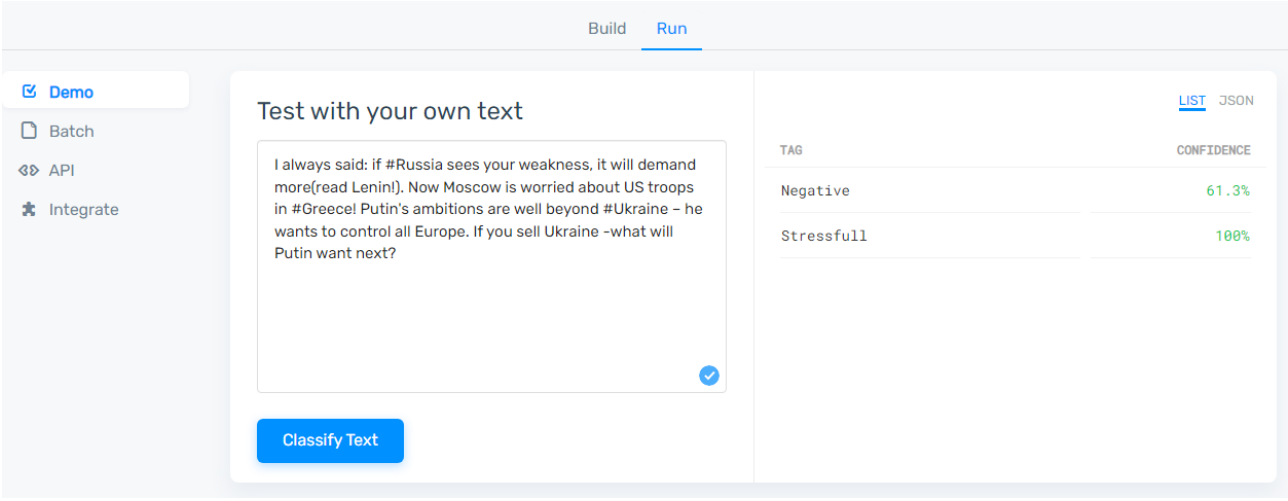


Рис 1.2. Результат демонстраційного класифікації тексту платформою MonkeyLearn

На цьому можливості безкоштовного плану платформи обмежуються. У платних планах – можна створювати робочі процеси, які по АПІ платформи будуть проводити аналіз та виводити результати на екран. Для прикладу в статті від платформи MonkeyLearn є приклад таких графіків, згенеровних для оцінки емоційного нахилу відгуків про компанію, ці графіки подано на рисунку 1.3.

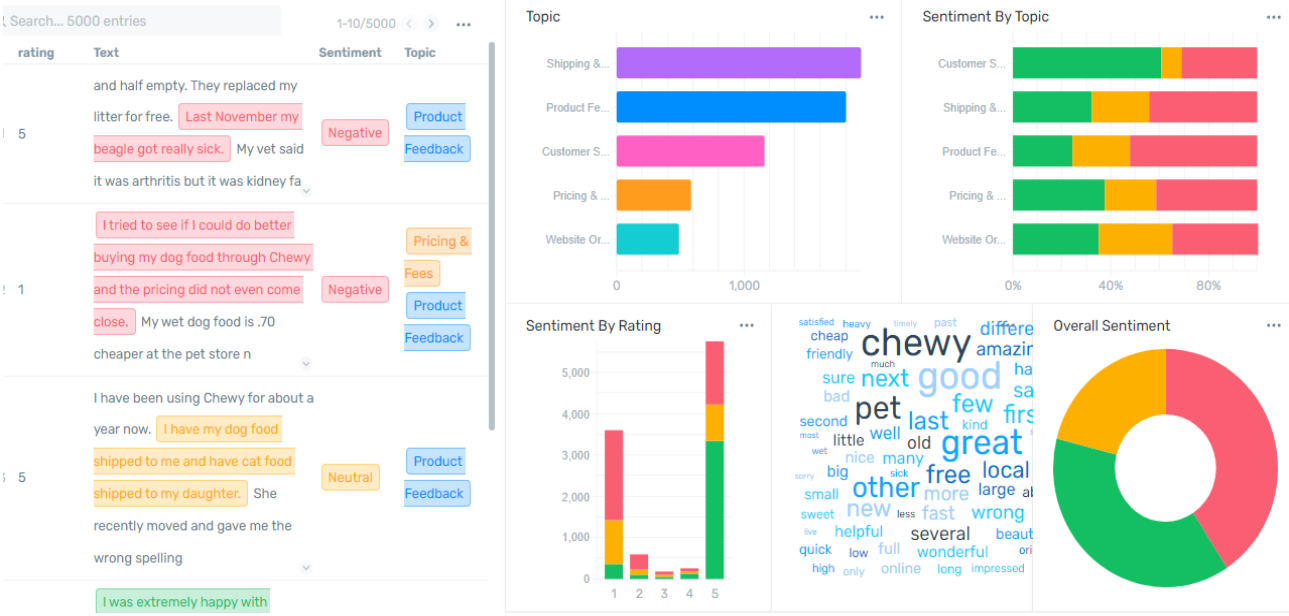


Рис 1.3. Графіки для аналітики відгуків про компанію на платформі MonkeyLearn

Іншим інструментом вартим згадування є ресурс Social Buzz [13]. Даний ресурс дозволяє шукати контент у соціальних мережах в режимі реального часу та надає глибокі аналітичні дані. Користувачі можуть шукати публічно опубліковану інформацію в Twitter, Google+, Facebook, Youtube, Instagram, Tumblr, Reddit, Flickr, Dailymotion і Vimeo без входу у ці системи. Хоча оцінка емоційного нахилу знайдених онлайн публікацій не є основним продуктом компанії – тут є узагальнений аналізатор зібраної інформації. На рисунку 1.4 показані результати аналізу емоційного нахилу знайдених публікацій по запиту «ukraine».

Рис 1.4. Результати аналізу емоційного нахилу знайдених публікацій по запиту «ukraine» інструментом Social Buzz

Окрім готових інструментів – різні аспекти впливу соціальних мереж та емоційний нахил опублікованих постів можна аналізувати самостійно, використовуючи бібліотеки для обробки природної мови. Хоча такий підхід до аналізу є складнішим, оскілки алгоритм чи модель потрібно розробляти самостійно, проте тут немає штучних обмежень, які встановлюються компаніями для монетизації свого інструменту.

По даній темі було проведено багато досліджень, зокрема праці які акцентують увагу на формуванні думок соціальними мережами «Opinion formation on social media: An empirical approach» [14]. У дослідженні було використано близько 6 мільйонів твітів створених 2.3 мільйонами авторів на протязі 2010 року. У процесі аналізу підготовлені дані було проаналізовано алгоритмом обмеженої симетричної невід'ємної матричної факторізації на основі лексики (CSNMF) [15]. Також було створено спрощену модель взаємодії користувачів у соціальних мережах. Результати показали що соціальні мережі можуть формувати та змінювати думку людини, проте цей процес відбувається повільно, а кількість агентів що змінюють свою думку спадає за степеневим розподілом. Позитивно на зміну думки чи її наближення до запропонованої у мережі впливає активність поширення запропонованої думки, кількість авторів які її поширюють та наявність однієї популярної думки. Симуляція підтвердила цей факт.

У 2018 також було опубліковано результати іншого дослідження на тему оцінки емоційного нахилу текстів «Attitudes Toward Feminism in Ukraine: A Sentiment Analysis of Tweets» [16]. Тут було проаналізовано емоційний нахил настроїв україномовних твітів про фемінізм. Для проведення дослідження було створено україномовні словниками для аналізу емоційного нахилу текстів та налаштовано алгоритм SentiStrength [17] для роботи з ними. Результати показали переважаючий негативний настрій проаналізованих твітів.

1.3. Підходи до збору даних для аналізу емоційного нахилу текстів

Для проведення аналізу потрібні досліджувальні дані. У даному випадку досліджуваними даними будуть виступати тексти у яких ведеться мова про Україну. Для дослідження інформації яка вже існує давно можна використовувати підготовлені датасети. Часто датасети відповідають змісту однієї таблиці бази даних і характеризують певну сутність чи об'єкт. Такі датасети створюються компаніями чи навчальними закладами для проведення

власних досліджень, а потім можуть бути поширені публічно. Оскільки машинне навчання, наука про дані та штучний інтелект використовуються у широкому діапазоні досліджень та експериментів – існує багато датасетів для багатьох із цих досліджень. Важливим критерієм датасетів є їхня збалансованість. Для навчання моделей чи проведення досліджень потрібно враховувати усі точки зору або, якщо це не можливо – потрібно враховувати як омога більшу кількість точок зору на проблему. Збалансований датасет – це такий у якому кількість данних з різних джерел чи різних точок зору однакова. Такий датасет дозволить отримати менш упереджені результати дослідження. Досягти ідеального балансу без втрати даних неможливо, але до такого балансу варто наближатись. Варто також додати що датасети можуть бути попередньо опрацьованими, та містити додаткову інформацію яка може спростити проведення дослідження. Для прикладу тексти можуть мати пораховані кількості слів, містити позначки джерел та років публікації текстів чи бути подані у очищеному вигляді, хоча зазвичай ситуація протилежна і публічні датасети потрібно очищати від неповних даних.

Оскільки компанія Твіттер забороняє поширювати тексти твітів, ідентифікуючі дані авторів та іншу чутливу інформацію у мережі публічно [18] –датасети твітів часто містять лише перелк унікальних ідентифікаторів кожного твіта, який підходить під обрану тему. Для отримання повного датасету потрібно провести так звану гідрацію [19].

Існує інший, платний спосіб отримати чи згенерувати готові, повні датасети, які будть містити всю необхідну інформацію для проведення дослідження. Компанія Твіттер, після покупки підписки на Historical PowerTrack API [20], дозволяє налаштувати процеси збору та формування датасетів за обраними критеріями. Цим платним API користуються відомі бренди для проведення своїх дослідженнь, компанії які надають послуги генерації датасетів, або дослідницькі установи які проводять багато досліджень на різні теми. Такі установи-посередники можуть запропонувати генерацію

датасету за доступнішою ціною ніж оплата підписки на платформі Твіттер [20], проте можуть додати свої обмеження на процес генерації.

Іншим підходом до збору чи генерації датасету для дослідження – є використання АПІ цільової платформи для збору даних [20]. API інтерфейс дозволяє різним частинам програми чи декількох програм, часто написаних різними розробниками чи групами розробників, якісно та передбачувано взаємодіяти між собою. Хоча АПІ використовується як у операційних системах, для взаємодії їх ключових сатин, так і у додатках для мобільних телефонів, для зображенян пікселів на екрані, – найчастіше використовуються веб АПІ, які дозволяють взаємодіяти з веб серверами. Твіттер – популярна платформа мікроблогів, у якій щодня залишають близько мільярду нових твітів, надає доступ до свого АПІ. Для отримання такого доступу потрібно бути користувачем твіттера, створити обліковий запис розробника та надіслати запит на отримання доступу до АПІ. У такому запиті у розробників запитують у яких цілях будуть використовуватись дані отримані через АПІ.

Станом на 2021 рік існує 2 версії та 3 рівні доступу до використання наданого Твіттер АПІ [20]. На рисунку 1.5 подано порівняльну таблицю пропонованих версій та рівнів доступу до Твіттер АПІ. З цього рисунку видно що найширший доступ має рівень АПІ для академічного дослідження.

	Essential	Elevated	Elevated+ (coming soon)	Academic Research
Getting access	Sign up	Apply for additional access within the developer portal	Need more? Sign up for our waitlist	Apply for additional access
Price	Free	Free		Free
Access to Twitter API v2	✔	✔		✔
Access to standard v1.1	✖	✔		✔
Access to premium v1.1	✖	✔		✔
Access to enterprise	✖	✔		✔
Project limits	1 Project	1 Project		1 Project
App limits	1 App per Project	3 Apps per Project		1 App per Project
Tweet caps	Retrieve up to 500k Tweets per month	Retrieve up to 2 million Tweets per month		Retrieve up to 10 million Tweets per month
Filtered stream rule limit	5 rules	25 rules		1000 rules
Filtered stream POST rules rate limit	25 requests per 15 minutes	50 requests per 15 minutes		100 requests per 15 minutes
Access to full-archive search Tweets	✖	✖		✔
Access to full-archive Tweet counts	✖	✖		✔
Access to advanced filter operators	✖	✖		✔
Manage a team in the developer portal	✖	✔		✔
Access to the Ads API	✖	✔ (Requires additional application)		✔

Рис 1.5. Порівняльна таблиця пропонованих версій та рівнів доступу до Твіттер АПІ

Доступ до Твіттер АПІ для академічного дослідження надається лише після проходження верифікації командою підтримки Твіттер. На рисунку 1.6 представлено перший крок подачі запиту на отримання доступу до Твіттер АПІ для академічного дослідження. Коли доступ до АПІ надано – можна починати процес збору даних звертаючись по них до сервера, як описано у документації.

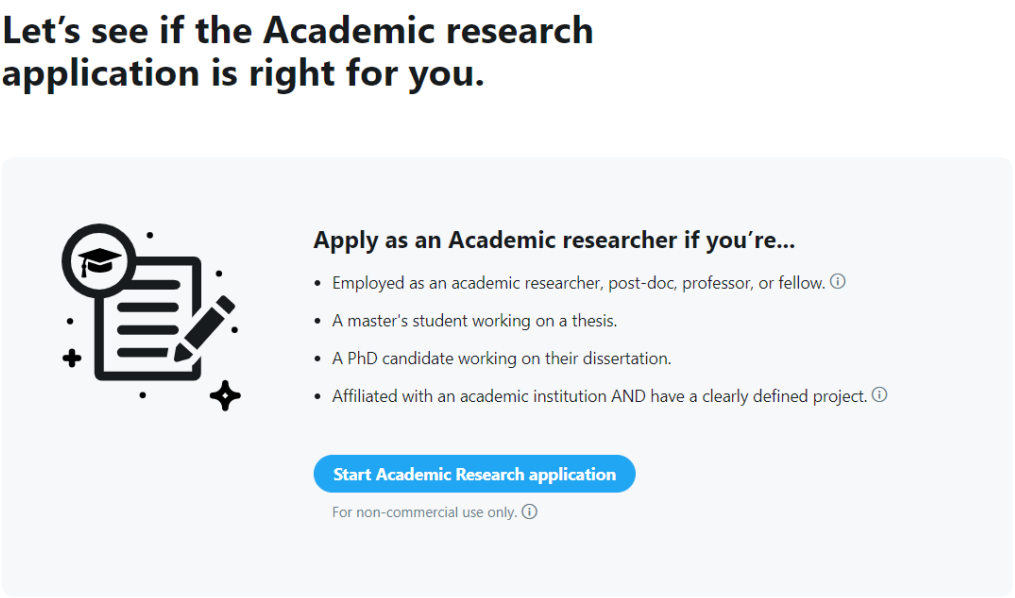


Рис 1.6. Перший крок подачі запиту на отримання доступу до твіттер АПІ для академічного дослідження

Третій підхід збору даних – веб скрапінг. Такий підхід вимагає певних вмінь та знань про те як працюють вебсайти. Основна суть веб скрапінгу – перетворення інформації представленої у вигляді веб сторінок, призначених для перегляду людиною, у інші форми представлення цієї інформації. Прцес веб скрапінгу можна поділити на наступні кроки:

- пошук веб сторінок з потрібною інформацією;
- пошук корисної інформації на сторінці;
- огляд структури сторінки;
- написання алгоритму для витягнення даних.

Написання алгоритму витягнення даних умовно ділиться на наступні етапи:

- запит до сторінки з інформацією;
- витягнення потрібної інформації з отриманої сторінки;
- збереження витягнутої інформації;
- перехід до наступного запиту.

Інколи, замість написання власного алгоритму та програми для веб скрапінгу – достатньо використати наявні на ринку готові продукти. На

рисунку 1.7 представлено процес налаштування веб скрапінгу із використанням засобів платформи ParseHub [21]. Під час веб скрапінгу веб сайти сприйматимуть коректно сформовані запити, як запити від користувачів, які переглядають сайт через браузер. Це може мати свої негативні та позитивні ефекти на процес збору даних. Для прикладу сайт може мати обмеження на перегляд певної кількості статей на день що ускладнить збір даних. Також існують сайти що використовують динамічну генерацію сторінок, яка відбувається у браузері, через javascript запити. Інформацію з таких сайтів неможливо зібрати без аналізу самих запитів, що потребує значно глибших знань у побудові веб платформ.

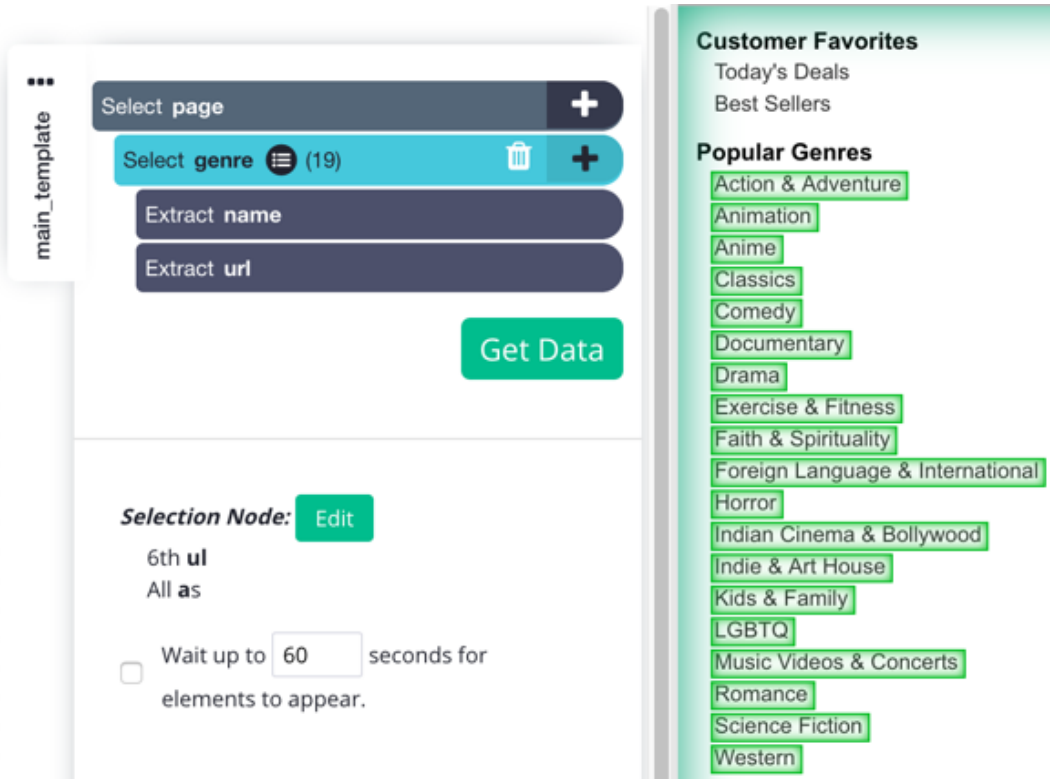


Рис 1.7. Налаштування веб скрапінгу за засобами платформи ParseHub

1.4. Висновки розділу 1

У даному розділі було встановлено, що оцінка емоційного нахилу текстів є важливим напрямом проведення досліджень у галузі штучного інтелекту.

Серед оглянутих інструментів та публікацій по даній темі було вирішено розробити свій алгоритм для оцінки емоційного нахилу текстів, через відсутність штучних обмежень у обробці, подачі даних та представлені отриманих результатів аналізу.

Для отримання вхідних даних для дослідження було вирішено використати публічне АПІ надане компанією Твіттер.

Було обговорено підходи до самого процесу аналізу емоційного нахилу текстів. Серед словникового підходу та підходу на основі алгоритмів машинного навчання, для проведення аналізу – було обрано перший підхід. Недоліком цього підходу може бути нижча точність у порівнянні з іншим підходом, проте важливою перевагою словникового підходу є простота його реалізації, а також низькі затрати часу та обчислювальних ресурсів для попереднього навчання моделі.

РОЗДІЛ 2
РОЗРОБКА АЛГОРИТМУ ОЦІНКИ ЕМОЦІЙНОГО НАХИЛУ СТАТТЕЙ НОВИН

2.1. Процес отримання доступу до Твіттер АПІ

Як вже було обговорено раніше основні етапи проведення аналізу емоційного нахилу тексту такі:

- пошук даних;
- збір даних;
- аналіз даних.

Як джерело даних було обрано платформу мікроблогів Твіттер. Далі необхідно визначити спосіб збору інформації за платформи та спосіб аналізу зібраної інформації.

Для отримання доступу до Твіттер АПІ в першу чергу потрібно бути користувачем Твіттер. Пройти короткий процес реєстрації можна зайшовши на офіційну сторінку платформи за посиланням twitter.com [22]. Якщо вхід зробити з території України – користувача зустріне україномовний екран реєстрації чи входу в існуючий акаунт, представлений на рис. 2.1.

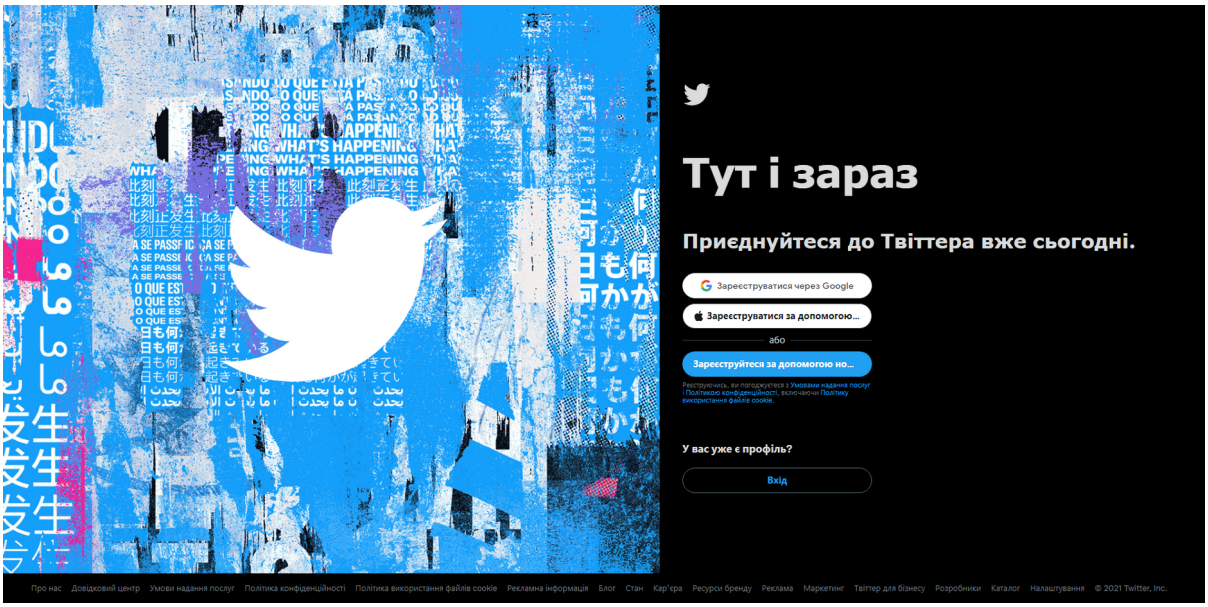


Рис 2.1. Вікно реєстрації та входу на платформу Твіттер

Після завершення налаштувань профілю – можна переходити до створення облікового запису розробника та подачі запиту на отримання доступу до АПІ для академічного дослідження. Для створення обікового запису розробника необхідно перейти за посиланням developer.twitter.com [23] та увійти до платформи використовуючи звичайний обліковий запис Твіттер. Якщо все виконано – перед користувачем відкриється екран представлений на рис. 2.2, де після натискання кнопки «Developer Portal» – користувача перенаправить у кабінет розробника. Дана кнопка розташована у верхній правій частині екрану, лівіше від зображення користувача.

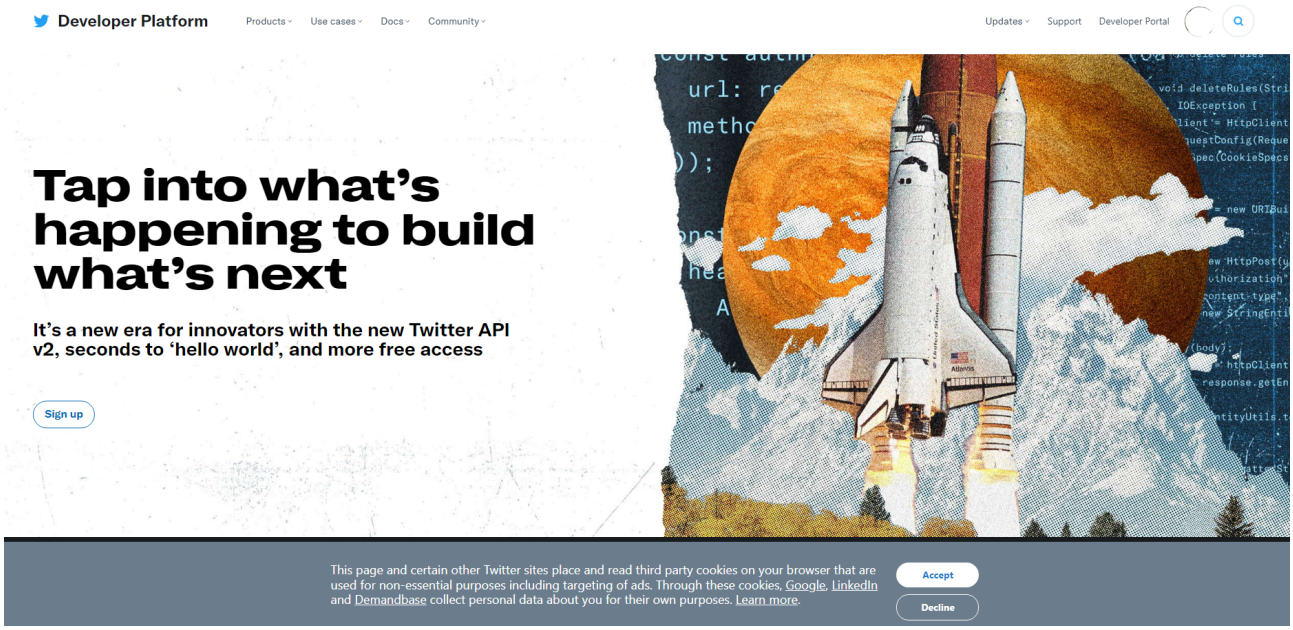


Рис 2.2. Вітальний екран входу на платформу Твіттер для розробників

При першому вході до кабінету розробника потрібно буде відповісти на декілька запитань та підтвердити свої облікові дані. Перевірка цих даних відбувається автоматично, тому після завершення усіх запитань – доступ до базового (Essential) рівня АПІ буде надано автоматично.

Для отримання вищого рівня доступу, а саме рівня доступу до АПІ для проведення академічного дослідження – потрібно детально розповісти про дослідження, як у ньому будуть використовуватись та опрацьовуватись дані, а також форму подачі результатів дослідження. Ці дані потрібні для забезпечення

дотримання політики конфіденційності платформи. Процес перевірки наданої інформації проводиться людьми, тому перевірка може зайняти певний час. Якщо наданих даних буде не достатньо – працівники компанії звернуться за уточненнями чи доповненнями через електронну пошту, на яку було зареєстровано обліковий запис. Хоча такий процес отримання доступу до АПІ може бути складним та довготривалим – він того вартий беручи до уваги різницю у обмеженнях на використання АПІ. На рис. 2.3 представлені узагальнені порівняння трьох рівнів доступу до АПІ, які можна отримати безкоштовно.

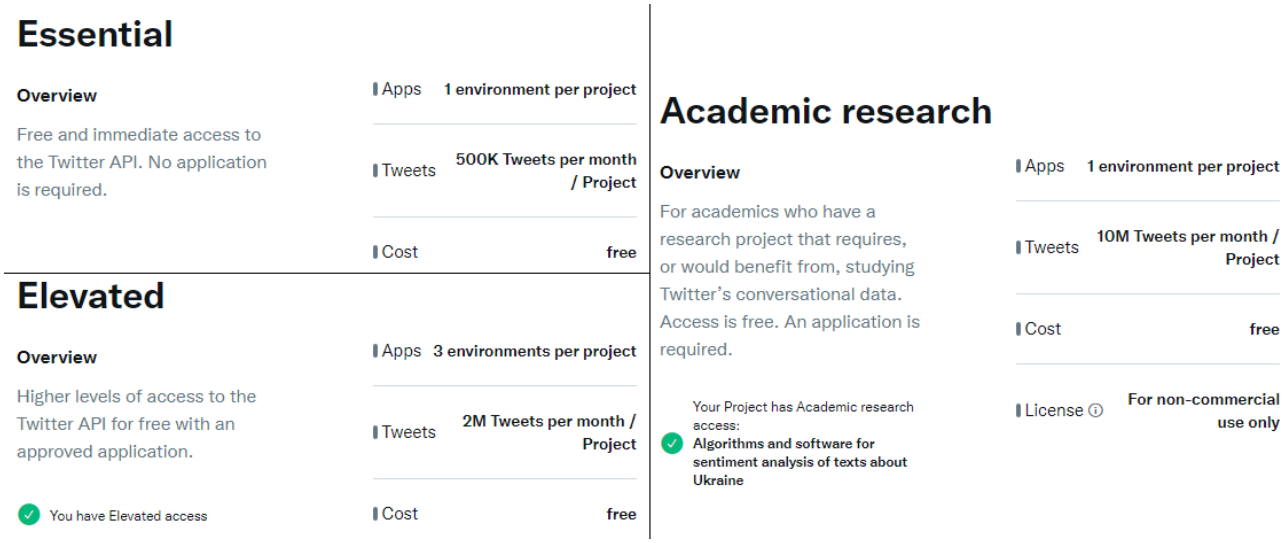


Рис 2.3. Узагальненні порівняння трьох безкоштовних рівнів доступу до Твіттер АПІ

З цього рисунку видно що найвагоміша різниця полягає у значенні Tweets, яке вказує яку максимально можливу кількість твітів можна отримати через АПІ за один місяць. Для базового рівня це значення становить 500 тисяч твітів у місяць, для підвищеного рівня – 2 мільйони, а для рівня академічного дослідження – 10 мільйонів. Окрім даних узагальнених відмінностей між рівнями – вищі рівні надають доступ до більшої кількості типів звернень, зокрема до пошуку по всьому архіву в рівні для академічного дослідження.

2.2. Звернення до Твіттер АПІ

Після отримання доступу до Твіттер АПІ для розробників – необхідно створити ключі та токени, які будуть застосовуватись для формування запитів. На рис. 2.4 представлено екран налаштування проекту на панелі керування для розробників, де до проекту потрібно додати програмні додатки.

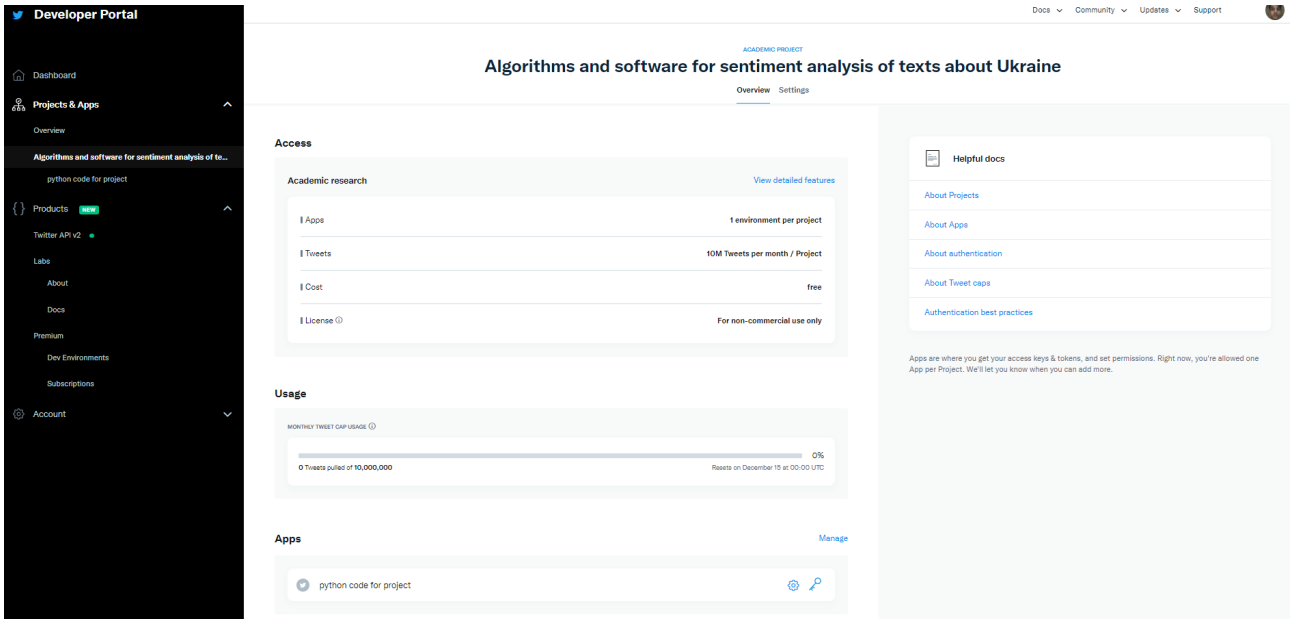


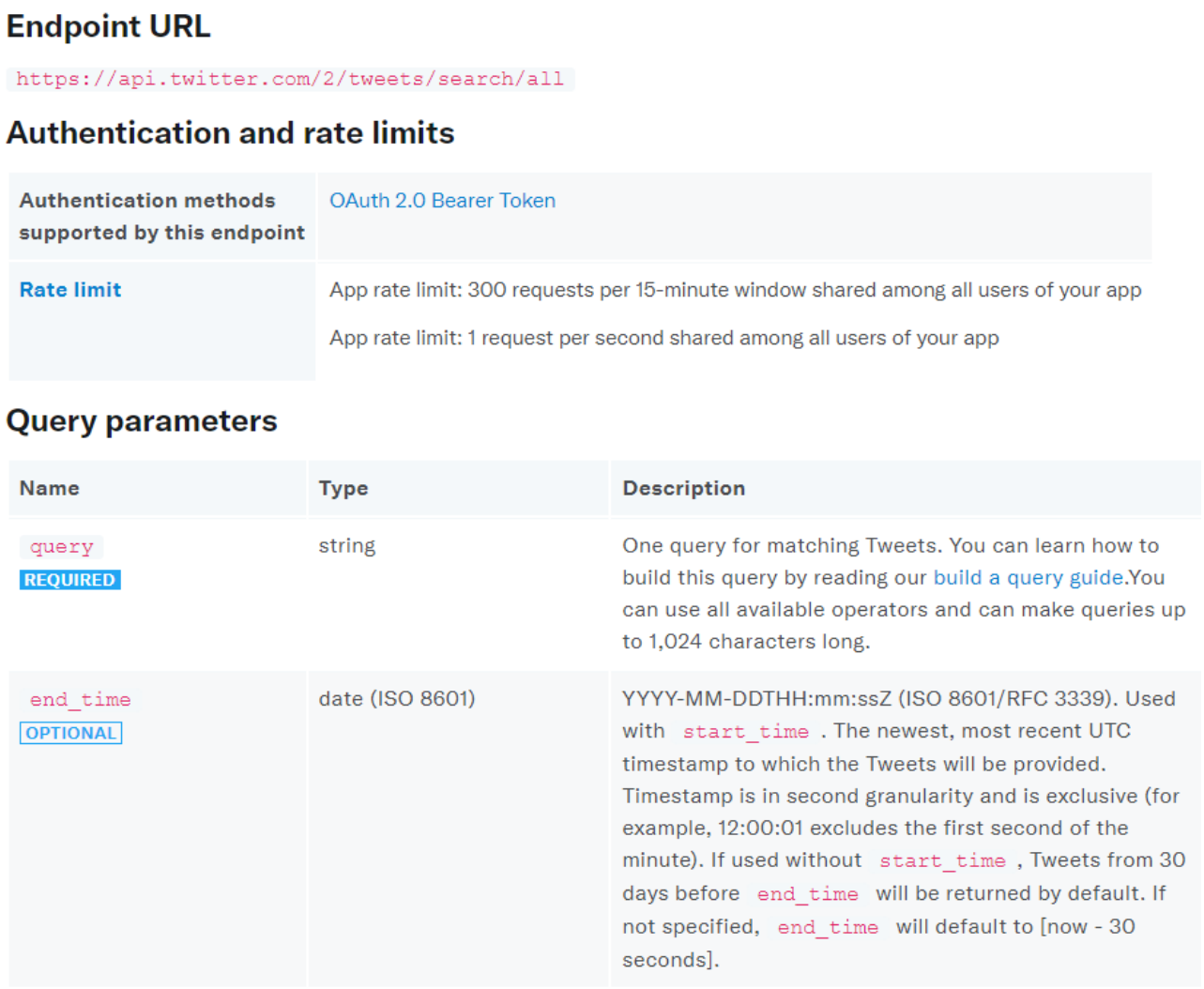
Рис 2.4. Екран налаштування проекту на панелі керування для розробників

Тут після створення програмного додатку можна згенерувати наступні ключі потрібні для роботи з Твіттер АПІ:

- API Key (Consumer Key);
- API Secret (Consumer Secret);
- Bearer Token;
- Access Token;
- Access Secret.

Далі необхідно знайти шлях звернення до АПІ для отримання бажаних даних. Для отримання твітів з повного архівного пошуку потрібно робити запити на АПІ шлях /2/tweets/search/all з параметрами вказаними у описі цього АПІ шляху. Опис цього АПІ шляху можна знайти за посиланням [24]. Також

сторінка опису показує аутентифікацію по якому із попередньо згенерованих ключів потрібно проходити для успішного використання API шляху. На рис. 2.5 показано частину опису API шляху пошуку твітів по всьому архіву.



Endpoint URL

`https://api.twitter.com/2/tweets/search/all`

Authentication and rate limits

Authentication methods supported by this endpoint	OAuth 2.0 Bearer Token
Rate limit	App rate limit: 300 requests per 15-minute window shared among all users of your app App rate limit: 1 request per second shared among all users of your app

Query parameters

Name	Type	Description
`query` REQUIRED	string	One query for matching Tweets. You can learn how to build this query by reading our build a query guide.You can use all available operators and can make queries up to 1,024 characters long.
`end_time` OPTIONAL	date (ISO 8601)	YYYY-MM-DDTHH:mm:ssZ (ISO 8601/RFC 3339). Used with `start_time`. The newest, most recent UTC timestamp to which the Tweets will be provided. Timestamp is in second granularity and is exclusive (for example, 12:00:01 excludes the first second of the minute). If used without `start_time`, Tweets from 30 days before `end_time` will be returned by default. If not specified, `end_time` will default to [now - 30 seconds].

Рис 2.5. Частина сторінки опису API шляху пошуку твітів по всьому архіву

Для формування запиту можна використати сторінку Twitter API v2 calls [23], яка дозволить у інтерактивному режимі створити коректний запит. На рис. 2.6 наведено приклад використання цієї сторінки для генерації запиту, який потім можна надіслати, як показано на рис. 2.7.

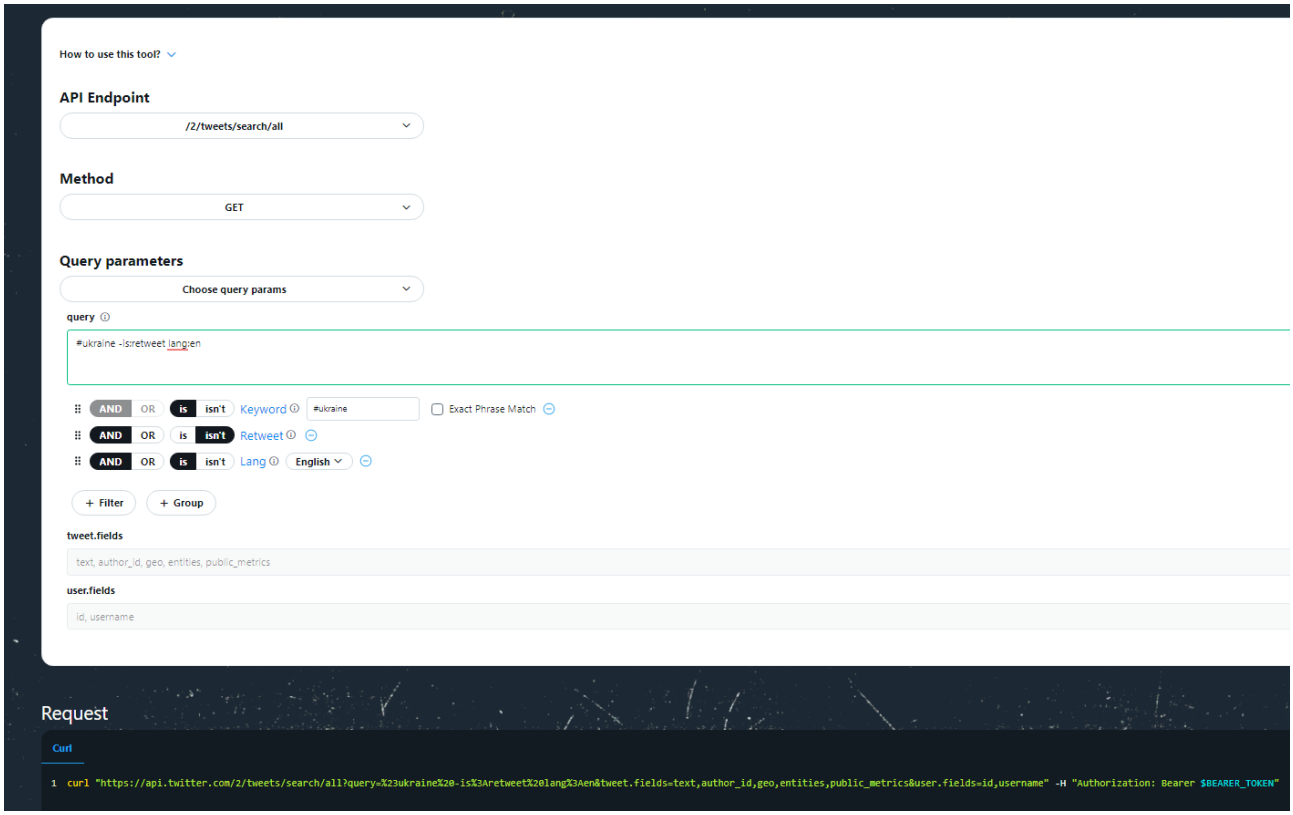


Рис 2.6. Приклад генерації запиту інструментом Twitter API v2 calls

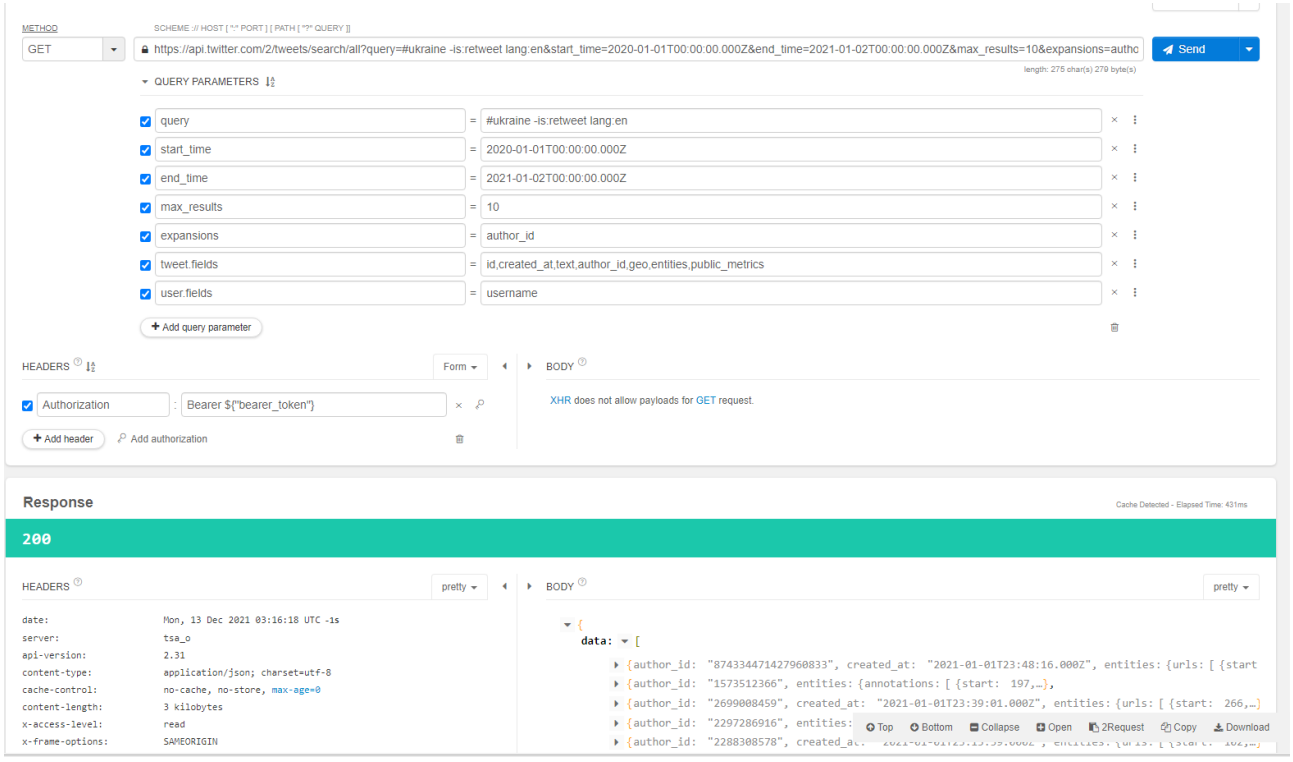


Рис 2.7. Виконання запиту до Твіттер АПІ

Такий спосіб надсилання запитів підходить для тестування та поодиноких запитів. У випадку коли потрібно надсилати багато запитів – у нагоді стають бібліотеки для роботи із запитами чи конкретним АПІ. Для мови програмування Python є написана бібліотека для роботи з Твіттер АПІ під назвою tweepy. Розробники цієї бібліотеки створили та підтримують у актуальному стані об'єкти та класи для звернення до різних шляхів на Твіттер АПІ. На рис. 2.8 представлено код на мові Python для надсилання запиту аналогічного тому що зображений на рис. 2.7, з використанням бібліотеки tweepy.

```python
from tweepy import Client

import twitter_credentials

client = Client(bearer_token=twitter_credentials.BEARER_TOKEN)
tweets = client.search_all_tweets(query='#ukraine -is:retweet lang:en',
                                  start_time='2020-01-01T00:00:00.000Z',
                                  end_time='2021-01-02T00:00:00.000Z',
                                  max_results=2,
                                  expansions='author_id',
                                  tweet_fields='id,created_at,text,author_id,geo,entities,public_metrics',
                                  user_fields='username')
```

Рис 2.8. Виконання запиту до Твіттер АПІ з використанням бібліотеки tweepy

Інший поширений варіант звернення до Твіттер АПІ – інструмент командного рядка та Python бібліотека twarc. У випадку його використання як бібліотеки – потрібно написати Python код подібний до коду для використання бібліотеки tweepy, його представлено на рис. 2.9 а. На рис. 2.9 б представлено команду для використання twarc як інструменту командного рядка, з виведенням отриманих даних у файл.

```python
import twitter_credentials
import datetime
from twarc.client2 import Twarc2

t = Twarc2(bearer_token=twitter_credentials.BEARER_TOKEN)
start_time = datetime.datetime(2009, 1, 1, 0, 0, 0, 0, datetime.timezone.utc)
end_time = datetime.datetime(2010, 1, 2, 0, 0, 0, 0, datetime.timezone.utc)

search_results = t.search_all(query="#ukraine -is:retweet lang:en",
                              start_time=start_time,
                              end_time=end_time,
                              expansions='author_id',
                              tweet_fields='id,created_at,text,author_id,geo,entities,public_metrics',
                              user_fields='username'
                              )
```

А

```
C:\Users\vasyl>twarc2 search --start-time '2009-01-01' --end-time '2010-01-02' --archive --max-results 2  --expansions 'author_id' --tweet-fields 'id,created_at,
text,author_id,geo,entities,public_metrics,source,lang' --user-fields 'id,name,username,location,public_metrics' '#ukraine -is:retweet lang:en' allTweets.jsonl
```

Б

Рис 2.9. Виконання запиту до Твіттер АПІ а – бібліотекою twarc, б – інструментом командного рядка twarc

Варто зауважити що при використанні twarc як інструменту командного рядка – усі обмеження по використанні АПІ, такі як максимальна кількість запитів у секунду, максимальна кількість отриманих твітів за запит, обробляються автоматично. Це означає що користувачу не потрібно писати додаткових інструкцій для забезпечення цілісності отримуваних даних та обробляти помилки.

2.3. Аналіз даних

Для отримання точної оцінки емоційного нахилу тексту потрібно створювати модель машинного навчання та тренувати її на попередньо підготовлених даних. Хоча галузь постійно розвивається та стає доступнішою, обсяги даних потрібні для навчання, перевірки коректності роботи та самого проведення аналізу можуть сягати декількох десятків мільйонів одиниць, а підготовчі процеси можуть займати роки.

Простішим способом перевірки порівняння на основі попередньо складених словників полярності слів та словосполучень. Важливою частиною такого підходу є таблиця з інформацією про характеристики кожного слова. У

2012 році, лінгвісти Tom De Smedt та Walter Daelemans створили бібліотеку під назвою Pattern, яка зокрема мала таку таблицю. Пізніше на основі Pattern була створена популярніша бібліотека TextBlob, яка також має більше можливостей[26][27]. Для розрахунку емоційного нахилу речення Ssent використовується формула $\mathrm{Ssent} = \sum \mathrm{wisent},$ де wisent – значення емоційного нахилу слова взятого з таблиці. Інша бібліотека побудована на тому ж принципі роботи і додатково скорегована для опрацювання текстів з соціальних мереж – VADER (Valence Aware Dictionary and sEntiment Reasoner) [28]. Для коригування бібліотеки – у її словник було додано скорочення та сленг, який можна часто зустріти на просторах соціальних мереж. Третя популярна бібліотека яка дозволяє зробити оцінку емоційного нахилу тексту – SentiWordNet [29].

2.4. Висновки розділу 2

У розділі описано основні етапи проведення аналізу емоційного нахилу тексту. Описано порядок подачі запиту на отримання рівня доступу до Твіттер АПІ для академічного дослідження та продемонстровано спосіб формування АПІ запиту.

Розглянуто декілька підходів до збору інформації з Твіттер АПІ, та бібліотек для проведення аналізу емоційного нахилу тексту методом, що базується на словниках та правилах.

РОЗДІЛ 3
ПРАКТИЧНА РЕАЛІЗАЦІЯ АЛГОРИТМУ ТА ДОСЛІДЖЕННЯ АНАЛІЗУ ЕМОЦІЙНОГО НАХИЛУ ТЕКСТУ

3.1. Процес збору даних

Беручи до уваги поширеність та зростаючу популярність оцінки емоційних нахилів текстів та відсутність публікацій на тему оцінки емоційного нахилу текстів про Україну, було вирішено написати програмну реалізацію такого алгоритму.

Для проведення дослідження було вирішено використати метод заснований на словниках та правилах, який буде реалізовуватись усіма описаними бібліотеками.

Збір інформації з Твіттер АПІ було вирішено зробити за допомогою інструменту командного рядка twarc, через його надійність та простоту використання у порівнянні з іншими варіантами.

Оскільки компанія Твіттер встановила обмеження на використання свого АПІ – потрібно зібрати репрезентативну вибірку даних не виходячи за вказані ліміти. На момент збору даних та написання програмної реалізації алгоритму запит на отримання рівня доступу до Твіттер АПІ для розробників було надано. Завдяки цьому обмеження на кількість твітів, які можна завантажити протягом місяця виросла до 10 мільйонів.

Спершу для завантаження твітів було вирішено використати слово «ukraine», а також задати параметр мови та виключити з вибірки ретвіти, що описується пошуковим запитом представленим на рис. 3.1. За діапазон часу для пошуку було вирішено обрати рівно 15 років, з моменту запуску платформи та публікації першого твіта 21 березня 2006 року.

Для оцінки об’єму вибірки існує окремий АПІ шлях /2/tweets/counts/all, який дозволяє отримати цифру, скільки твітів задовольняє умови вибірки. На

рис. 3.2 показано команду для надсилання запиту інструментом командного рядка twarc та результат виконання цієї команди.

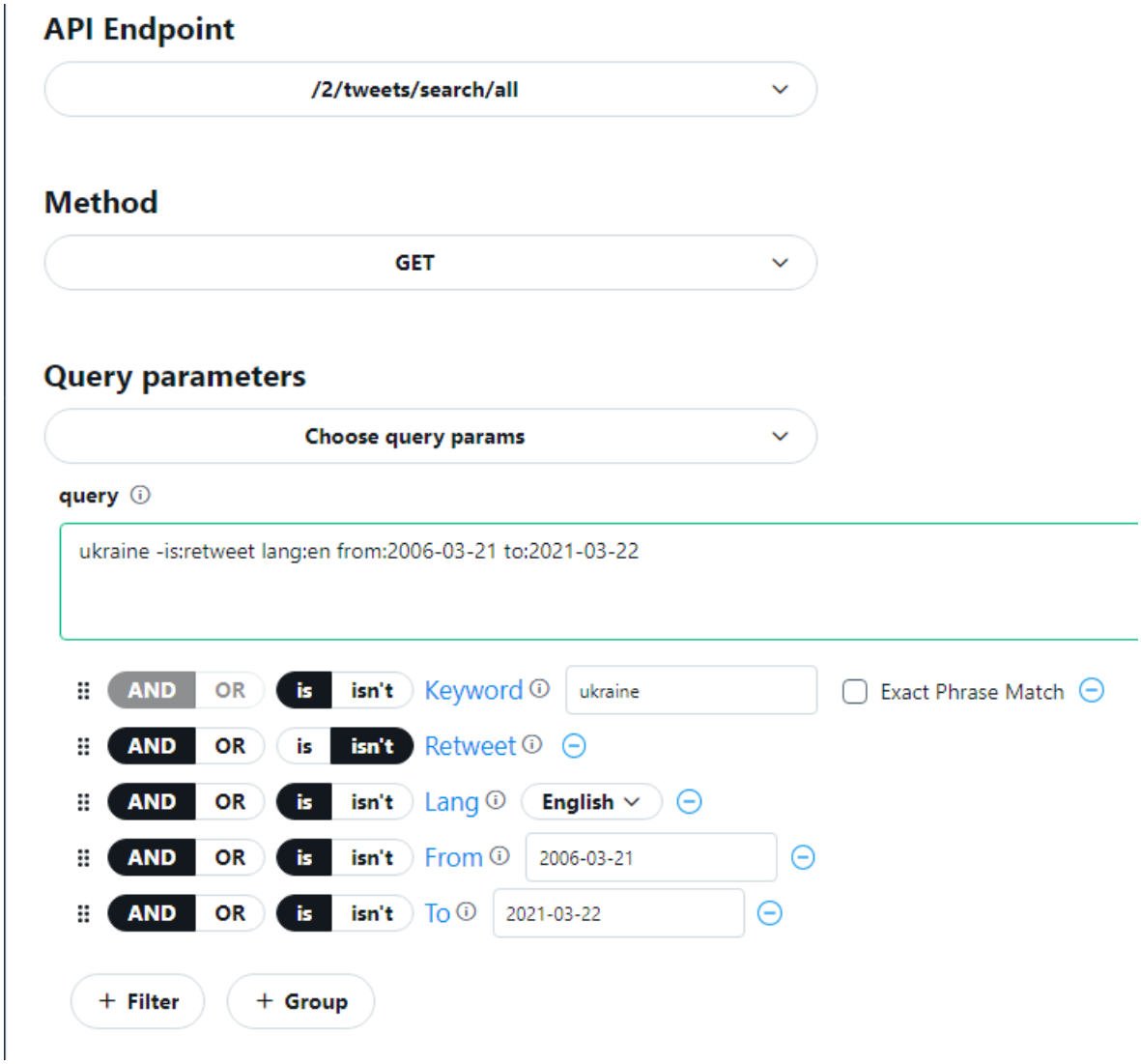


Рис 3.1. Згенерований запит для завантаження твітів

```
Total Tweets: 29,348,314


Aborted!
PS D:\F\Crea\Code\Python\Masters> twarc2 counts --start-time 2006-03-21 --end-time 2021-03-22
 --archive --granularity day --text --hide-progress 'ukraine -is:retweet lang:en'
```

Рис 3.2. Запит для оцінки об’ємів вибірки

З рисунку видно що кількість твтів по такому запиту більша за 29 мільйонів. Це однозначно виходить за рамки обмежень встановлених

компанією, тому запиту було змінено на використання хештегу «#ukraine». Оцінка обсягу такої вибірки показала що є всього 4,2 мільйони твітів, що її задовольняють. Блок-схему описаного вище процесу формування запиту для збору даних наведено у Додатку Б.

Далі залишалось зробити запит на завантаження твітів та очікувати його завершення. Команда для запуску процесу завантаження зображена на рис. 3.3. Близько через 7 годин у папці з проектом був файл з усіма потрібними твітами

```
PS D:\F\Crea\Code\Python\Masters> twarc2 search --start-time '2006-03-21T00:00:00' --end-time '2021-03-22T
00:00:00' --archive --max-results      --expansions 'author_id' --tweet-fields 'id,created_at,text,author_
id,geo,entities,public_metrics,source,lang' --user-fields 'id,name,username,location,public_metrics' '#ukr
aine -is:retweet lang:en' allTweetsAll.jsonl
```

Рис 3.3. Команда для завантаження всієї вибірки

3.2. Підготовка даних до аналізу

Зібрані інструментом командного рядка twarc дані представлені у вигляді JSON рядків, тобто на кожному наступному рядку – знаходиться наступний JSON об’єкт. Структуру цього об’єкту представлено на рис. 3.4.

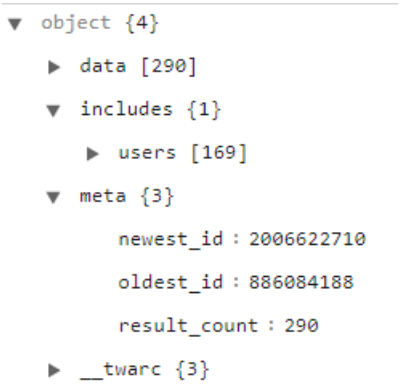


Рис 3.4. Структура Json об’єкту отриманого через Твіттер АПІ

Для опрацювання з такого Json об’єкту, потрібно перебрати усі твіти що знаходяться по шляху $.data та перенести потрібні поля у інший масив. Новостворений масив після цього можна опрацьовувати, або зберегти у файл

використовуючи методи бібліотеки pickle. Код для формування масиву та збереження його на диск на комп'ютері подано на рисунку 3.5.

Наступним кроком потрібно створити датафрейм на основі поданих твітів. Датафрейми – найпоширеніший вигляд представлення даних для роботи з алгоритмами штучного інтелекту та машинного навчання.

```
import json
import pickle

tweets = []
users = {}
count = 0
with open("allTweets06-21.jsonl") as file:
    for line in file:
        json_line = json.loads(line)
        for user in json_line['includes']['users']:
            if user['id'] not in users:
                users[user['id']] = user
        for json_tweet in json_line['data']:
            location = ''
            if users[json_tweet['author_id']] is not None and 'location' in users[json_tweet['author_id']]:
                location = users[json_tweet['author_id']]['location']
            tweet = {
                'text': json_tweet['text'],
                'source': json_tweet['source'] if 'source' in json_tweet and json_tweet['source'] is not None else '',
                'created_at': json_tweet['created_at'],
                'user_location': location
            }
            tweets.append(tweet)
        count = count + 1
        if count % 100 == 0:
            print('Count of processed rows:' + str(count))
            print('Count of tweets:' + str(len(tweets)))

pickle.dump(tweets, open("tweets.pkl", "wb"))
print('Count of processed rows:' + str(count))
print('Count of tweets:' + str(len(tweets)))
```

Рис 3.5. Код для приведення завантажених твітів до робочого вигляду

Після створення датафрейму – потрібно очистити текст що буде аналізуватись від зайвих символів та термінів, які не несуть емоційного змісту а лише збільшують обсяг тексту та розмір файлу. До таких символів належать згадування інших користувачів, хештеги, посилання на інші сайти, розділові знаки, цифри, емоджі та інші.

Для проведення процесу очистки даних було використано бібліотеку neattext, яка містить багато корисних функцій для роботи з текстовими даними. Код та результат його роботи представлено на рис. 3.6.

```
import neattext.functions as nfx
import re

re_pattern = '(' + nfx.USER_HANDLES_REGEX.pattern + ')|(' + nfx.HASTAG_REGEX.pattern + ')|(' + nfx.URL_PATTERN.pattern +
df['clean_text'] = df['text'].progress_apply(nfx.remove_custom_pattern, term_pattern=re.compile(re_pattern))

df['clean_text'] = df['clean_text'].progress_apply(nfx.fix_contractions)
df['clean_text'] = df['clean_text'].progress_apply(nfx.remove_non_ascii)
df['clean_text'] = df['clean_text'].progress_apply(nfx.replace_bad_quotes)
df['clean_text'] = df['clean_text'].progress_apply(nfx.remove_stopwords)
df['clean_text'] = df['clean_text'].progress_apply(nfx.remove_puncts, most_common=False)
df['clean_text'] = df['clean_text'].progress_apply(nfx.remove_multiple_spaces)

pickle.dump(df, open("pickles/df_clean.pkl", "wb"))

df.head()
```

	text	source	created_at	user_location	clean_text
0	#Ukraine Dollar Bonds Jump as #IMF Says Aid Ta...	iOS	2014-03-21T23:59:41.000Z	London New York Ukraine	dollar bonds jump says aid talks making progress
1	#Kiev Mar 22 01:30 Temperature 9C no or few cl...	update weather tokyo	2014-03-21T23:59:18.000Z	Kiev, Ukraine	mar temperature c clouds wind sw kmh humidity
2	"@JeffFortenberry: #Ukraine - why should we ca...	Twitter for Windows Phone	2014-03-21T23:57:47.000Z	Lincoln Nebraska	care president handling well u credit
3	#Ukraine Putin has thumb his nose at the world...	TweetCaster for Android	2014-03-21T23:57:31.000Z		putin thumb nose world it starting cold war ne...
4	Why #Russia Is So Worried About #Ukraine http:...	Twitter for Websites	2014-03-21T23:57:06.000Z	Toronto Ontario Canada	worried

Рис 3.6. Код для очисти аналізованого тексту та результат виконання цього коду

Наступним кроком проводитися токенізація тексту. Токенізація – це процес розбиття тексту на окремі елементи, зазвичай слова, які називаються токенами. Ці токени будуть далі використовуватись у аналізі емоційного нахилу тексту, тому для більшої точності аналізу важливо щоб попередні кроки підготовки не залишили зайвих даних.

Коли усі токени згенеровано – потрібно додати позначки яку частину мови представляє кожен токен. Код для токенізації та додавання позначок частин мови подано на рис. 3.7. Цей процес зазвичай займає порівняно більше часу ніж інші підготовчі процеси. Обумовлено це необхідністю перевірки кожного слова, що залишилось після очистки, а тому кількість необхідних дій зростає в рази.

```
from nltk import pos_tag
from nltk import word_tokenize
from nltk.corpus import wordnet

pos_dict = {'J': wordnet.ADJ, 'V': wordnet.VERB, 'N': wordnet.NOUN, 'R': wordnet.ADV}


def token_pos(text):
    tags = pos_tag(word_tokenize(text))
    newlist = []
    for word, tag in tags:
        newlist.append(tuple([word, pos_dict.get(tag[0])]))
    return newlist


file = open('pickles/df_clean.pkl', 'rb')
df = pickle.load(file)
file.close()

df['pos'] = df['clean_text'].progress_apply(token_pos)
df.head()

pickle.dump(df, open("pickles/df_clean_pos.pkl", "wb"))
```

	text	source	created_at	user_location	clean_text	pos
0	#Ukraine Dollar Bonds Jump as #IMF Says Aid Ta...	iOS	2014-03-21T23:59:41.000Z	London New York Ukraine	dollar bonds jump says aid talks making progress	[(dollar, n), (bonds, n), (jump, v), (says, v)...
1	#Kiev Mar 22 01:30 Temperature 9C no or few cl...	update weather tokyo	2014-03-21T23:59:18.000Z	Kiev, Ukraine	mar temperature c clouds wind sw kmh humidity	[(mar, n), (temperature, n), (c, n), (clouds, ...
2	"@JeffFortenberry: #Ukraine - why should we ca...	Twitter for Windows Phone	2014-03-21T23:57:47.000Z	Lincoln Nebraska	care president handling well u credit	[(care, n), (president, n), (handling, v), (we...
3	#Ukraine Putin has thumb his nose at the world...	TweetCaster for Android	2014-03-21T23:57:31.000Z		putin thumb nose world it starting cold war ne...	[(putin, n), (thumb, n), (nose, a), (world, n)...
4	Why #Russia Is So Worried About #Ukraine http:...	Twitter for Websites	2014-03-21T23:57:06.000Z	Toronto Ontario Canada	worried	[(worried, a)]

Рис 3.7. Код для токенізації та встановлення міток відповідності частинам мови.

Останній процес підготовки даних – отримання коренів та лем позначених слів. Для виконання цього завдання також є бібліотеки які пришвидшують ці процеси, проте варто розуміти різницю між цими процесами. Процес отримання коренів (stemming) – це алгоритм який обрізає суфікси та префікси для знаходження кореня слова. Далі ці кореневі частини слова використовуються для аналізу, проте часто до таких обрізаних слів не вдається знайти відповідні слова з словників емоційного нахилу.

Процес отримання лем опирається на морфологічний аналіз слова та знаходить базову його форму. Точність такого пошуку також сильно залежить від словника по якому здійснюється пошук лем. Після знаходження лем аналіз також проводиться через пошук цих слів у словниках емоційного нахилу текстів. Код для знаходження лем з використанням бібліотеки nltk та результат його виконання подано на рис. 3.8.

```
from nltk.stem import WordNetLemmatizer

wordnet_lemmatizer = WordNetLemmatizer()


def lemmatize(pos_data):
    lemma_rew = " "
    for word, pos in pos_data:
        if not pos:
            lemma = word
            lemma_rew = lemma_rew + " " + lemma
        else:
            lemma = wordnet_lemmatizer.lemmatize(word, pos=pos)
            lemma_rew = lemma_rew + " " + lemma
    return lemma_rew


df['lemma'] = df['pos'].progress_apply(lemmatize)
df.head()
```

```
  0%|          | 0/4291631 [00:00<?, ?it/s]
```

	text	source	created_at	user_location	clean_text	pos	lemma
0	#Ukraine Dollar Bonds Jump as #IMF Says Aid Ta...	iOS	2014-03-21T23:59:41.000Z	London New York Ukraine	dollar bonds jump says aid talks making progress	[(dollar, n), (bonds, n), (jump, v), (says, v)...	dollar bond jump say aid talk make progress
1	#Kiev Mar 22 01:30 Temperature 9C no or few cl...	update weather tokyo	2014-03-21T23:59:18.000Z	Kiev, Ukraine	mar temperature c clouds wind sw kmh humidity	[(mar, n), (temperature, n), (c, n), (clouds, ...	mar temperature c cloud wind sw kmh humidity
2	"@JeffFortenberry: #Ukraine - why should we ca...	Twitter for Windows Phone	2014-03-21T23:57:47.000Z	Lincoln Nebraska	care president handling well u credit	[(care, n), (president, n), (handling, v), (we...	care president handle well u credit
3	#Ukraine Putin has thumb his nose at the world...	TweetCaster for Android	2014-03-21T23:57:31.000Z		putin thumb nose world it starting cold war ne...	[(putin, n), (thumb, n), (nose, a), (world, n)...	putin thumb nose world it start cold war nee...
4	Why #Russia Is So Worried About #Ukraine http:...	Twitter for Websites	2014-03-21T23:57:06.000Z	Toronto Ontario Canada	worried	[(worried, a)]	worried

Рис 3.8. Код для пошуку лем та результат його виконання

На цьому моменті підготовча частина опрацювання даних завершується і починається сам процес аналізу.

3.3. Оцінювання емоційного нахилу текстів

Оскільки проведення оцінки емоційного нахилу текстів проводиться за словниковим методом та не вимагає додаткових затрат часу на навчання моделі – було вирішено провести оцінку з використанням усіх бібліотек згаданих у розділі 2.3. Блок-схема алгоритму проведення оцінки подана у Додатку В.

Першим було проведено аналіз з використанням методів бібліотеки TextBlob. На рис. 3.9 представлено код проведення аналізу, та декілька рядків результуючого датафрейму з новими стовпцями polarity та analysis, які показують числове та якісне значення емоційного нахилу тексту кожного твіта. Як описувалось раніше дана бібліотека використовує словник зі значеннями полярності текстів. Вигляд такого словника подано на рис. 3.10.

```python
from textblob import TextBlob


# function to calculate subjectivity
def getSubjectivity(review):
    return TextBlob(review).sentiment.subjectivity


# function to calculate polarity
def getPolarity(review):
    return TextBlob(review).sentiment.polarity


# function to analyze the reviews
def analysis(score):
    if score < 0:
        return 'Negative'
    elif score == 0:
        return 'Neutral'
    else:
        return 'Positive'


fin_data = pd.DataFrame(df[['text', 'lemma']])
# fin_data['Subjectivity'] = fin_data['Lemma'].apply(getSubjectivity)
fin_data['polarity'] = fin_data['lemma'].progress_apply(getPolarity)
fin_data['analysis'] = fin_data['polarity'].progress_apply(analysis)
fin_data.head()
```

	text	lemma	polarity	analysis
0	#Ukraine Dollar Bonds Jump as #IMF Says Aid Ta...	dollar bond jump say aid talk make progress	0.0	Neutral
1	#Kiev Mar 22 01:30 Temperature 9C no or few cl...	mar temperature c cloud wind sw kmh humidity	0.0	Neutral
2	"@JeffFortenberry: #Ukraine - why should we ca...	care president handle well u credit	0.0	Neutral
3	#Ukraine Putin has thumb his nose at the world...	putin thumb nose world it start cold war nee...	-0.6	Negative
4	Why #Russia Is So Worried About #Ukraine http:...	worried	0.0	Neutral

Рис 3.9. Код для оцінки емоційного нахилу тексту з використанням бібліотеки TextBlob

```
<word form="loose" wordnet_id="a-02232112" pos="JJ" sense="(of textures) full of small openings or gaps" polarity="0.0" subjectivity="0.0" intensity="1.0" confidence="0.9" />
<word form="loose" wordnet_id="a-02404421" pos="JJ" sense="not tense or taut" polarity="0.0" subjectivity="0.0" intensity="1.0" confidence="0.9" />
<word form="losers" wordnet_id="" pos="NNS" sense="" polarity="-0.2" subjectivity="0.2" intensity="1.0" reliability="0.9" />
<word form="loses" wordnet_id="" pos="VB" polarity="-0.3" subjectivity="0.1" intensity="1.0" confidence="0.9" />
<word form="loud" cornetto_synset_id="n_a-512029" wordnet_id="a-01452593" pos="JJ" sense="characterized by or producing sound of great volume or intensity" polarity="0.1" subjecti
<word form="lousy" wordnet_id="a-422374" pos="JJ" sense="infested with lice" polarity="-0.5" subjectivity="0.5" intensity="1.0" reliability="0.9" />
<word form="lovable" wordnet_id="a-01459422" pos="JJ" sense="having characteristics that attract love or affection" polarity="0.5" subjectivity="0.5" intensity="1.0" confidence="0
<word form="love" wordnet_id="v-1775164" pos="VB" sense="have a great affection or liking for" polarity="0.5" subjectivity="0.6" intensity="1.0" confidence="0.9" />
<word form="love" wordnet_id="v-1775535" pos="VB" sense="be enamored or in love with" polarity="0.5" subjectivity="0.6" intensity="1.0" confidence="0.9" />
<word form="love" wordnet_id="v-1828736" pos="VB" sense="get pleasure from" polarity="0.5" subjectivity="0.6" intensity="1.0" confidence="0.9" />
<word form="loved" cornetto_synset_id="n_a-506898" wordnet_id="a-01461822" pos="JJ" sense="held dear" polarity="0.7" subjectivity="0.8" intensity="1.0" confidence="0.9" />
<word form="lovely" cornetto_synset_id="d_a-9388" wordnet_id="a-00219809" pos="JJ" sense="appealing to the emotions as well as the eye" polarity="0.5" subjectivity="1.0" intensity
```

Рис 3.10 – Частина словника що використовується бібліотекою TextBlob

Кожне значення у словнику має наступні важливі для оцінки емоційного нахилу текстів параметри:

- позначка частини мови (pos);
- опис значення у якому використовується слово (sense);
- значення емоційного нахилу (polarity);
- значення суб’єктивності (subjectivity);
- значення інтенсивності (intensity);
- значення впевненості (confidence).

Використовуючи значення зі словника, розташування слів у тексті та інші параметри – бібліотека дає висновок про полярність тексту.

На рис. 3.11 представлена інші частини програмної реалізації алгоритму, для аналізу тексту з використанням vaderSentiment бібліотеки. Підхід що використовується цією бібліотекою дуже подібний до попереднього, проте окрім звичайного словника англомовних слів тут додатково описані сленгові слова, слова підсилювачі, які не несуть емоційного нахилу, а підсилюють інші слова, та слова для заміни емоджі. На рис. 3.12 зображено декілька рядків словника емоджі та текстового словника.

```
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer

analyzer = SentimentIntensityAnalyzer()


# function to calculate vader sentiment
def vadersentimentanalysis(review):
    vs = analyzer.polarity_scores(review)
    return vs['compound']


# function to analyse
def vader_analysis(compound):
    if compound >= 0.5:
        return 'Positive'
    elif compound <= -0.5:
        return 'Negative'
    else:
        return 'Neutral'


fin_data['vader sentiment'] = fin_data['lemma'].progress_apply(vadersentimentanalysis)
fin_data['vader analysis'] = fin_data['vader sentiment'].progress_apply(vader_analysis)
fin_data.head()
```

	text	lemma	polarity	analysis	vader sentiment	vader analysis
0	#Ukraine Dollar Bonds Jump as #IMF Says Aid Ta...	dollar bond jump say aid talk make progress	0.0	Neutral	0.4215	Neutral
1	#Kiev Mar 22 01:30 Temperature 9C no or few cl...	mar temperature c cloud wind sw kmh humidity	0.0	Neutral	0.0000	Neutral
2	"@JeffFortenberry: #Ukraine - why should we ca...	care president handle well u credit	0.0	Neutral	0.7845	Positive
3	#Ukraine Putin has thumb his nose at the world...	putin thumb nose world it start cold war nee...	-0.6	Negative	-0.7269	Negative
4	Why #Russia Is So Worried About #Ukraine http:...	worried	0.0	Neutral	-0.2960	Neutral

Рис 3.11. Код для оцінки емоційного нахилу тексту з використанням бібліотеки vaderSentiment

```
😁 grinning squinting face
😉 winking face
😊 smiling face with smiling eyes
😋 face savoring food
😎 smiling face with sunglasses
😍 smiling face with heart-eyes
😘 face blowing a kiss
🥰 smiling face with 3 hearts
😗 kissing face
😙 kissing face with smiling eyes
```

```
4894 offline -0.5    0.92195 [0, 0, 0, 0, -3, 0, -1, 0, -1, 0]
4895 ok  1.2 0.4 [1, 2, 1, 1, 1, 1, 2, 1, 1, 1]
4896 okay    0.9 0.53852 [1, 1, 0, 0, 1, 1, 1, 2, 1, 1]
4897 okays   2.1 1.13578 [1, 1, 1, 4, 3, 2, 2, 1, 2, 4]
4898 ominous -1.4    1.49666 [-3, -2, -1, -2, -2, -1, -1, 1, 1, -4]
4899 once-in-a-lifetime  1.8 1.4 [4, 2, 1, 0, 1, 1, 4, 3, 2, 0]
4900 openness    1.4 0.8 [2, 1, 1, 2, 2, 1, 1, 1, 3, 0]
4901 opportune   1.7 0.78102 [2, 2, 0, 1, 2, 3, 2, 2, 1, 2]
4902 opportunely 1.5 1.0247  [1, 1, 4, 1, 2, 1, 1, 2, 2, 0]
4903 opportuneness   1.2 1.249   [0, 1, 2, 2, 2, 2, 2, 2, -2, 1]
```

Рис 3.12. Частини емоджі та текстового словників що використовуються бібліотекою vaderSentiment

Для проведення останнього аналізу було використано бібліотеку nltk, з словником SentiWordNet. Бібліотека nltk – це платформа для створення програм які працюють з природними мовами. Зокрема бібліотека надає прості у використанні інтерфейси для більше 50 словників, зокрема і WordNet словнику. Код для оцінки емоційного нахилу текстів засобами бібліотеки nltk з використанням SentiWordNet словника представлено на рис. 3.13. На рис. 3.14 подано декілька рядків самого SentiWordNet словника.

```
from nltk.corpus import sentiwordnet as swn


def sentiwordnetanalysis(pos_data):
    sentiment = 0
    tokens_count = 0
    for word, pos in pos_data:
        if not pos:
            continue
        lemma = wordnet_lemmatizer.lemmatize(word, pos=pos)
        if not lemma:
            continue
        synsets = wordnet.synsets(lemma, pos=pos)
        if not synsets:
            continue
            # Take the first sense, the most common
        synset = synsets[0]
        swn_synset = swn.senti_synset(synset.name())
        sentiment += swn_synset.pos_score() - swn_synset.neg_score()
        tokens_count += 1
        # print(swn_synset.pos_score(),swn_synset.neg_score(),swn_synset.obj_score())
        if not tokens_count:
            return 0
        if sentiment > 0:
            return "Positive"
        if sentiment == 0:
            return "Neutral"
        else:
            return "Negative"


fin_data['SWN analysis'] = df['pos'].progress_apply(sentiwordnetanalysis)
fin_data.head()
```

```
  0%|          | 0/4291631 [00:00<?, ?it/s]
```

	text	lemma	polarity	analysis	vader sentiment	vader analysis	SWN analysis
0	#Ukraine Dollar Bonds Jump as #IMF Says Aid Ta...	dollar bond jump say aid talk make progress	0.0	Neutral	0.4215	Neutral	Neutral
1	#Kiev Mar 22 01:30 Temperature 9C no or few cl...	mar temperature c cloud wind sw kmh humidity	0.0	Neutral	0.0000	Neutral	Neutral
2	"@JeffFortenberry: #Ukraine - why should we ca...	care president handle well u credit	0.0	Neutral	0.7845	Positive	Neutral
3	#Ukraine Putin has thumb his nose at the world...	putin thumb nose world it start cold war nee...	-0.6	Negative	-0.7269	Negative	Neutral
4	Why #Russia Is So Worried About #Ukraine http:...	worried	0.0	Neutral	-0.2960	Neutral	Negative

Рис 3.13. Код для оцінки емоційного нахилу тексту з використанням SentiWordNet словнка

```
a   00025728   0   0   alkaline#1 alkalic#1    relating to or containing an alkali; having a pH greater than 7; "alkaline soils derived from
a   00026051   0.125   0.375   alkalescent#1 alcalescent#1 tending to become alkaline; slightly alkaline
a   00026168   0.25    0   basic#4 of or denoting or of the nature of or containing a base
a   00026294   0.125   0.125   base-forming#1  yielding a base in aqueous solution
a   00026388   0.25    0.125   saltlike#1  resembling a compound formed by replacing hydrogen in an acid by a metal
a   00026515   0.5 0   amphoteric#1 amphiprotic#1  having characteristics of both an acid and a base and capable of reacting as either
a   00026706   0   0.25    acid-loving#1   thriving in a relatively acidic environment (especially of plants requiring a pH well below 7)
a   00026895   0   0   aciduric#1 acidophilous#1 acidophilic#1 especially of some bacteria; growing well in an acid medium
a   00027074   0   0.25    alkaline-loving#1   thriving in a relatively alkaline environment; (especially of plants requiring a pH above 7)
```

Рис 3.14. Декілька рядків SentiWordNet словника

Після завершення аналізу усіма переліченими алгоритмами та зібравши статистичні дані – було побудовано графіки. На рис. 3.15 зображено відсоткове

відношення між нейтральними, позитивними та негативними потсами за версіями кожного із алгоритмів. З цього рисунку можна зробити висновки що найбільше негативного емоційного нахилу знайшов підхід VADER, а найбільше позитивного емоційного нахилу – TextBlob. Усі три алгоритми оцінили що більше 58% тівтів, близько 2,5 мільйонів, були нейтральними. Як було сказано рніше особливістю бібліотеки vaderSentiment є використання словників з сленгом. Висока ймовірність того що саме через це загальна оцінка вибірки алгоритмом VADER особливо відрізняється від двох інших.

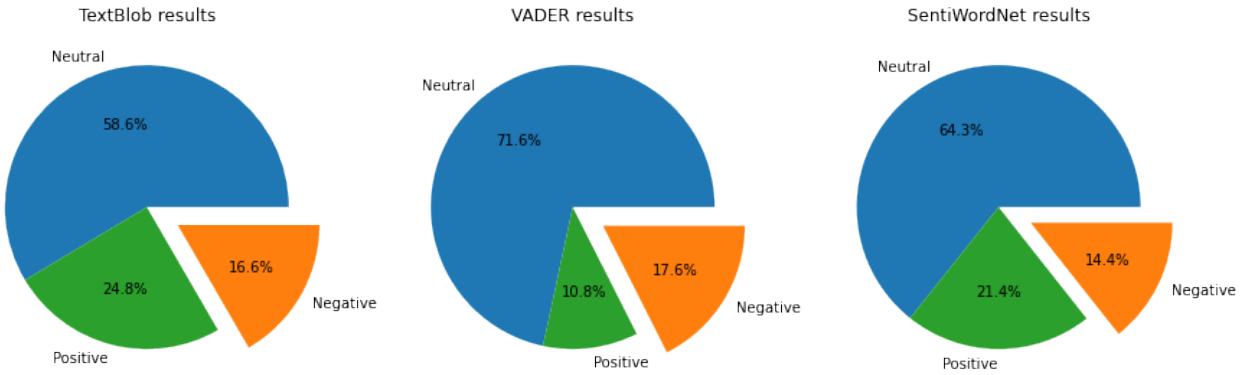


Рис 3.15. Оцінка емоційного нахилу усього датасету трьома алгоритмами

Узагальнена характеристика датасету не покаже розподілу емоційного нахилу тексту по роках. Для графічного представлення динаміки зміни емоційного нахилу було створено графік, що подано на рис. 3.16.

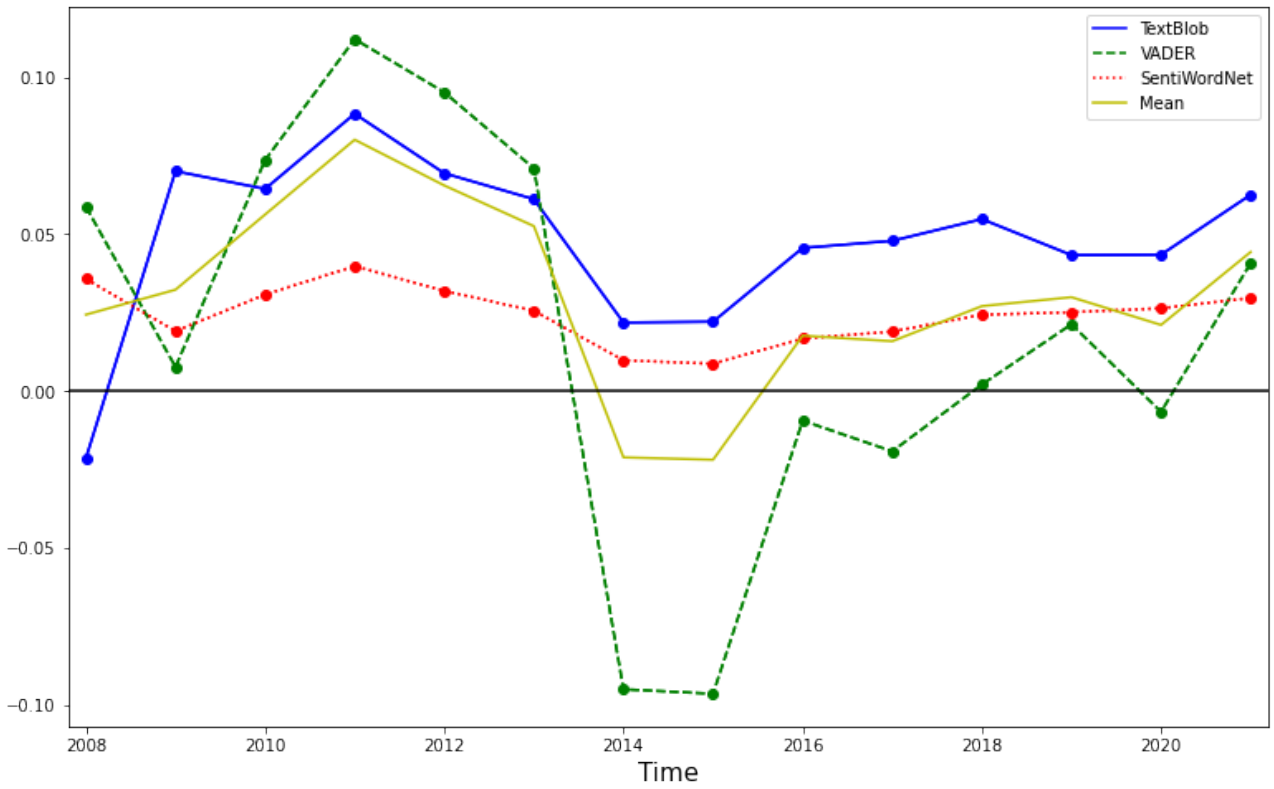


Рис 3.16. Динаміка зміни емоційного нахилу твітів про Україну

З цього графіку по перше видно що твітів з хештегом «#ukraine» на протязі 2006 – 2007 років не було і графік починається з позначки часу 2008 року. Інше цікаве спостереження – алгоритм з використанням словника SentiWordNet у кожному з проаналізованих років показував позитивний результат емоційного нахилу. Також, з графіку видно що найбільш негативними були 2014 – 2015 роки, що можна трактувати як світове занепокоєння на тривогу щодо подій на майдані, АТО та інших подій які трапились у ці роки.

Також було створено графік, що подано на рис. 3.17, для оцінки розподілу твітів про Україну по роках. З даного графіку видно, що починаючи з 2008 року твіти почали з’являтись і до 2013 року росли рівномірно. Причиною цього росту стали ріст популярності платформи та доступності інтернету. У 2014 році видно високий зріст активності від 150 тисяч твітів у 2013 до 1,8 мільйона твітів у 2014. Такі дані підкреслюють масовість та зацікавленість у темі майдану, АТО та подій 2014 – 2015 років.

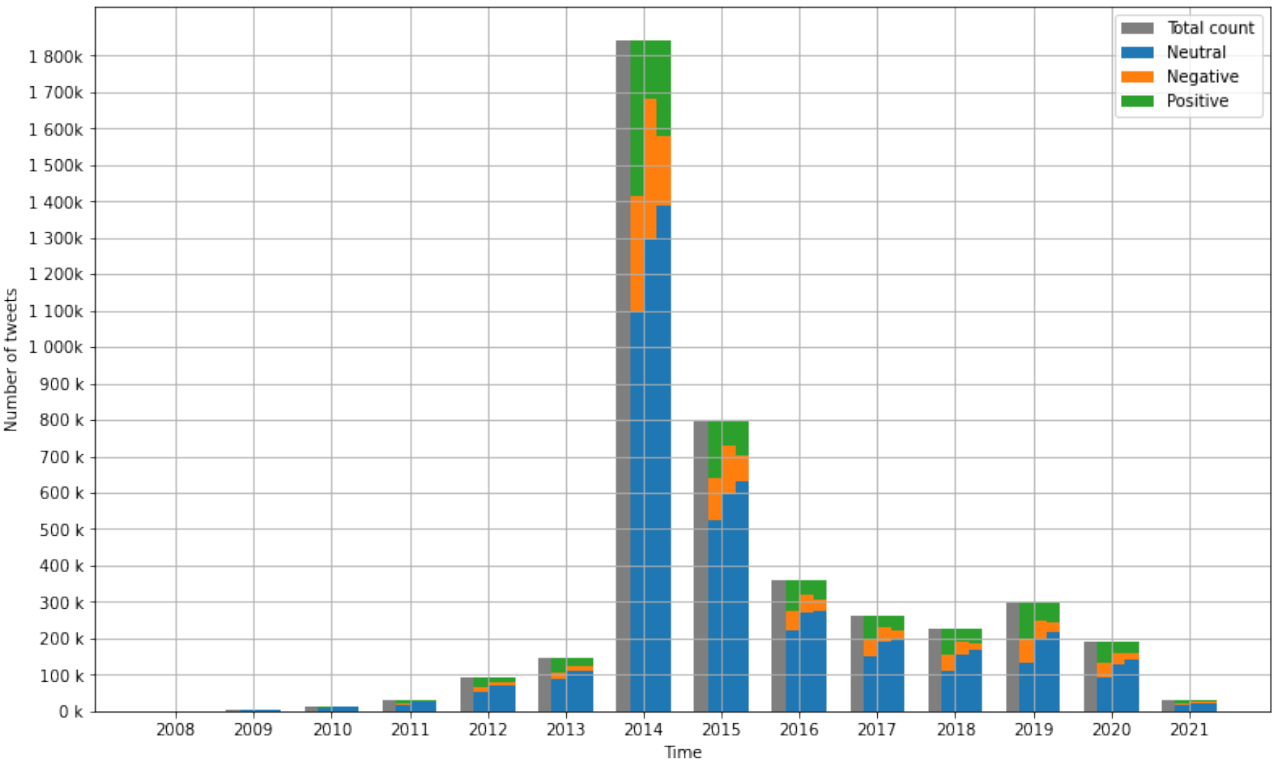


Рис 3.17. Розподіл твітів по роках

3.4. Висновки розділу 3

У даному розділі було описано процес формування та коригування пошукового запиту у рамках підготовку до процесу збору даних та сам процес.

Описано процес очистки даних для аналізу, а саме: очистка від зайвих символів та стоп слів, розбиття на токени, маркування тегами частин мови та виокремлення базових слів.

Проведено оцінку емоційного нахилу текстів трьома словниковими алгоритмами з використанням бібліотек TextBlob, vaderSentiment та nltk зі словником SentiWordNet. На основі результатів оцінювання було згенеровано та описано ряд графіків.

РОЗДІЛ 4
ОХОРОНА ПРАЦІ ТА БЕЗПЕКА В НАДЗВИЧАЙНИХ СИТУАЦІЯХ

4.1. Охорона праці

Охорона праці – це система правових, соціально-економічних, організаційно-технічних, санітарно-гігієнічних і лікувально-профілактичних заходів та засобів, спрямованих на збереження життя, здоров'я і працездатності людини у процесі трудової діяльності [30].

Сьогодні, через зростаючий попит на експертів комп’ютерних спеціальностей, який тягне за собою зростання пропозиції – важливо пам’ятати про безпеку життєдіяльності та охорону праці при роботі з комп’ютерами та іншими ЕОМ. Саме задля цього законодавством України було сформовано та чітко врегульовано норми та вимоги до використання комп’ютерної техніки на підприємстві. Беручи до уваги особливості розробки та застосування програмної реалізації алгоритму для оцінки емоційного нахилу текстів – основні документи з охорони праці є такими ж як загальні норми, правила поведінки та правила техніки безпеки при роботі з ПК. Основними документами які описують правила та норми поведінки при роботі з комп’ютерами та ЕОМ є:

– Закон України «Про охорону праці» введений в дію Постановою Верховної Ради № 2695-XII від 14.10.92, ВВР, 1992, № 49, ст.669 [30].

– Державні санітарні правила і норми роботи з візуальними дисплейними терміналами електронно-обчислювальних машин ДСанПІН 3.3.2.007-98 затверджені постановою Головного державного санітарного лікаря України № 7 від 10 грудня 1998 року [31].

– Вимоги щодо безпеки та захисту здоров’я працівників під час роботи з екранними пристроями зареєстровані в Міністерстві юстицій України 25 квітня 2018р. № 508/31960 [32].

Згідно з ДСанПІН 3.3.2.007-98 визначаються наступні вимоги до виробничих приміщень для експлуатації комп'ютерів:

– Заборонено розміщати робочі місця з персональними комп'ютерами у підвальних приміщеннях чи цокольних поверхах.

– Площа одного робочого місця повинна мати не менше 6 квадратних метрів, а об'єм не менше 20 кубічних метрів.

– У приміщенні має бути природнє та штучне освітлення відповідно до ДБН В.2.5-28:2018 Природне і штучне освітлення.

– Виробничі приміщення для роботи з екранами не повинні межувати з приміщеннями, в яких рівні шуму і вібрації перевищують допустимі значення за СН 2.2.42.1.8.562-96, ДНАОП 0.03-3.12-84, ГР 2411-81. Приміщення мають бути оснащені аптечками першої медичної допомоги.

Згідно Вимог щодо безпеки та захисту здоров'я працівників під час роботи з екранними пристроями, вимоги безпеки до робочих місць працівників з екранними пристроями включають наступне:

– Мікроклімат виробничих приміщень з робочими місцями працівників з екранними пристроями має підтримуватись на постійному рівні та відповідати вимогам Санітарних норм мікроклімату виробничих приміщень ДСН 3.3.6.042-99, затверджених постановою Головного державного санітарного лікаря України від 01 грудня 1999 року № 42 [33].

– Освітлення робочого місця працівника з екранними пристроями має створювати відповідний контраст між екраном і навколишнім середовищем та відповідати вимогам ДСанПІН 3.3.2.007-98 [31].

– Робочий стіл або робоча поверхня повинні бути достатнього розміру та мати поверхню з низькою відбивною здатністю, допускати гнучкість під час розміщення екрана, клавіатури, документів і відповідного устаткування.

– Робоче крісло має бути стійким і дозволяти працівнику з екранними пристроями легко рухатися та займати зручне положення. Сидіння має регулюватися по висоті, спинка сидіння - як по висоті, так і по нахилу.

Для забезпечення психічного та нервово-емоційного здоров'я користувачів комп'ютерів згідно з ДСанПІН 3.3.2.007-98 у залежності від професійної групи після кожної години чи 2 годин безперервної роботи потрібно призначати перерви на 10-15 хв. Також біля приміщень з комп'ютерами мають бути обладнані побутові приміщення для відпочинку під час роботи, кімната психологічного розвантаження. В кімнаті психологічного розвантаження слід передбачити встановлення пристроїв для приготування й роздачі тонізуючих напоїв, а також місця для занять фізичною культурою. Вимоги для допоміжних приміщень повинні відповідати ДБН В.2.2-28-2010. Якщо виробничі обставини не дозволяють застосувати регламентовані перерви, тривалість безперервної роботи за комп'ютером не повинна перевищувати 4 години.

За умов дотримання усіх вище згаданих та наведених у документах правил та норм – робота за комп'ютером під час застосування алгоритму оцінки емоційного нахилу статей буде безпечною.

4.2. Забезпечення безпеки життєдіяльності при роботі з ПК

Одним із основних документів які описують правила та норми поведінки при роботі з комп'ютерами та ЕОМ є ДСанПІН 3.3.2.007-98 затверджені постановою Головного державного санітарного лікаря України № 7 від 10 грудня 1998 року.

У документі зустрічаються наступні скорочення:

- ВДТ - візуальні дисплейні термінали;
- ЕОМ - електронно-обчислювальні машини;
- ПЕОМ - персональні ЕОМ.

Документ визначає наступні професійні групи, які за характером трудової діяльності можуть проводити до 8-ми годин працюючи з ЕОМ:

- розробники програм (інженери-програмісти);
- оператори електронно-обчислювальних машин;
- оператори комп'ютерного набору.

Встановлюються такі внутрішньозмінні режими праці та відпочинку при роботі з ЕОМ при 8-годинній денній робочій зміні в залежності від характеру праці:

- для розробників програм із застосуванням ЕОМ, слід призначати регламентовану перерву для відпочинку тривалістю 15 хвилин через кожну годину роботи за ВДТ;
- для операторів із застосування ЕОМ, слід призначати регламентовані перерви для відпочинку тривалістю 15 хвилин через кожні дві години;
- для операторів комп'ютерного набора слід призначати регламентовані перерви для відпочинку тривалістю 10 хвилин після кожною години роботи за ВДТ.

У всіх випадках, коли виробничі обставини не дозволяють застосувати регламентовані перерви, тривалість безперервної роботи з ВДТ не повинна перевищувати 4 години.

З метою зменшення негативного впливу монотонності є доцільним застосовувати чергування операцій усвідомленого тексту і числових даних (зміна змісту роботи). Чередування вводу даних та редагування текстів [31].

Працюючі з ВДТ ЕОМ і ПЕОМ підлягають обов'язковим медичним оглядам: попереднім - при влаштуванні на роботу і періодичним - протягом трудової діяльності відповідно до наказу МЗ України N 45 від 31.03.94 р.

Протипоказання з боку органів зору:

- гострота зору з корекцією не нижча ніж 0,5 на одному оці і 0,2 - на другому;
- рефракція: міопія вище 6,0 Д, гіперметропія вище 4,0 Д, астигматизм (будь якого виду) вище 3,0 Д;
- відсутності бінокулярного зору;
- лагофтальм;

- хронічні захворювання переденього відрізку очей;
- захворювання зорового нерва і сітки;
- глаукома [31].

Приміщення для роботи з ВДТ повинні мати природне та штучне освітлення відповідно до СНиП II-4-79 [31].

Звукоізоляція огороджувальних конструкцій приміщень з ВДТ має забезпечувати параметри шуму, що відповідають вимогам СН 3223-85, ГОСТ 12.1.003-83, ГОСТ 12.1.012-90 (дод.1).

Приміщення для роботи з ВДТ мають бути обладнані системами опалення, кондиціонування повітря, або припливно-витяжною вентиляцією відповідно до СНиП 2.04.05-91. Нормовані параметри мікроклімату, іонного складу повітря, вмісту шкідливих речовин мають відповідати вимогам СН 4088-86, СН 2152-80, ГОСТ 12.1.005-88, ГОСТ 12.1.007-76 та (дод.2,3).

Віконні прорізи приміщень для роботи з ВДТ мають бути обладнані регульованими пристроями (жалюзі, завіски, зовнішні козирки)[31].

Приміщення з ВДТ мають бути оснащені аптечками першої медичної допомоги [29].

У виробничих приміщеннях на робочих місцях з ВДТ мають забезпечуватись оптимальні значення параметрів мікроклімату: температури, відносної вологості й рухливості повітря (ГОСТ 12.1.005-88, СН 4088-86) [31].

Система загального освітлення має становити суцільні або преривчасті лінії світильників, розташовані збоку від робочих місць (переважно ліворуч), паралельно лінії зору працюючих [31].

Іонізуючі електромагнітні випромінювання на відстані 0,05 м від екрана до корпуса відеотермінала при будь-яких положеннях регулювальних пристроїв не повинна перевищувати 7,74 х 10 в ст.-12 А/кг, що відповідає еквівалентній дозі 0,1 мбер/год (100 мкР/год) НРБУ N 58.

Правила та норми поведінки при роботі з ПК чітко описують норми допустимого шуму, вібрацій, електромагнітних та іонізуючих випромінювань, освітленості, вентильованості та часу безперервної роботи, яких необхідно

дотримуватись для забезпечення безпеки життєдіяльності при роботі за ПК. Також важливо змінювати тип роботи за рівнем концентрації для забезпечення хорошого самопочуття та психічного здоров’я. Потрібно приділяти медоглядам та особливо звертати увагу на здоров’я очей, адже саме вони зазнають найбільшої шкоди під час довготривалої взаємодії з ПК.

ВИСНОВКИ

У процесі написання кваліфікаційної роботи магістра було розроблено алгоритм аналізу емоційного нахилу тексту та проаналізовано публікації доступні публічно на платформі мікроблогів Твіттер, створені на проміжуку між 2006 та 2021 роками з ключовим тегом «#ukraine». Було отримано наступні результати:

1. У процесі огляду наявних алгоритмів оцінки емоційного нахилу текстів було обгрунтовано актуальність теми отримання такої оцінки про Україну;
2. Проаналізовано наявні алгоритми оцінки емоційного нахилу тексту що використовують дані з соціальних мереж;
3. Для оцінки було обрано використати словниковий метод оцінки емоційного нахилу тексту;
4. Досліджено алгоритм роботи з Твіттер АПІ та описано поширені інструменти для роботи з АПІ. Засобами інструменту twarc було створено датасет з 4,33 мільйонів твітів що відповідали критеріям пошуку
5. Отриманий датасет було очищено та приведено до потрібного для аналізу вигляду;
6. Складено алгоритм для оцінки емоційного нахилу текстів словниковим методом з використанням бібліотек TextBlob, vaderSentiment, nltk та написано його програмну реалізацію;
7. На основі отриманих даних було побудовано та описано графіки зміни емоційного нахилу аналізованих даних.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Тиш Є.В., Кохан В.В.Б. Формування суспільної думки в соціальних мереж на прикладі мережі Twitter. Актуальні задачі сучасних технологій: збірник тез доповідей X міжнародної науково-практичної конференції Молодих учених та студентів, (Тернопіль, 24–25 листопада 2021 р.). Міністерство освіти і науки України, Тернопільський національний технічний університет імені Івана Пулюя [та ін.]. Тернопіль: ФОП Паляниця В. А., 2021. Т. 1. С. 127.
2. Кохан В.В.Б, Тиш Є.В. Методи оцінювання емоційного нахилу текстів засобами штучного інтелекту. Матеріали IX науково-технічної конференції «Інформаційні моделі, системи та технології» Тернопільського національного технічного університету імені Івана Пулюя, (Тернопіль, 8 – 9 грудня 2021 р.). Тернопіль: Тернопільський національний технічний університет імені Івана Пулюя, 2021. С. 112.
3. Artificial Intelligence: A Modern Approach, 4th US ed. Artificial Intelligence: A Modern Approach, 4th US ed. URL: http://aima.cs.berkeley.edu (дата звернення: 15.11.2021).
4. Turing A. M. I.–COMPUTING MACHINERY AND INTELLIGENCE. Mind. 1950. URL: https://academic.oup.com/mind/article/LIX/236/433/986238 (дата звернення: 15.11.2021).
5. Bollen J., Mao H. Twitter mood predicts the stock market. arXiv.org e-Print archive. URL: https://arxiv.org/PS_cache/arxiv/pdf/1010/1010.3003v1.pdf (дата звернення: 15.11.2021).
6. Pang B., Lee L. Opinion mining and sentiment analysis. Now the essence of knowledge. 2008. URL: https://www.cs.cornell.edu/home/llee/omsa/omsa.pdf (дата звернення: 15.11.2021).
7. Cambria E. Affective Computing and Sentiment Analysis. IEEE Intelligent Systems. 2016. Т. 31, № 2. P. 102–107. URL: https://ieeexplore.ieee.org/document/7435182 (дата звернення: 15.11.2021).

8. Can news help measure economic sentiment? An application in COVID-19 times / P. Aguilar et al. Economics Letters. 2021. P. 109730. URL: https://doi.org/10.1016/j.econlet.2021.109730 (дата звернення: 15.11.2021).

9. Kaya M. Stock price prediction using financial news articles. 2010 2nd IEEE International Conference on Information and Financial Engineering. 2010. URL: https://doi.org/10.26782/jmcms.spl.10/2020.06.00048 (дата звернення: 15.11.2021).

10. Yadava R., Kumarb A. V., Kumarc A. News-based supervised sentiment analysis for prediction of futures buying behaviour. IIMB Management Review. 2019. URL: https://www.sciencedirect.com/science/article/pii/S0970389619301569 (дата звернення: 15.11.2021).

11. Nam K., Seong N. Financial news-based stock movement prediction using causality analysis of influence in the Korean stock market. Decision Support Systems. 2019. T. 117. C. 100–112. URL: https://www.sciencedirect.com/science/article/abs/pii/S0167923618301957 (дата звернення: 15.11.2021).

12. MonkeyLearn - Text Analysis. MonkeyLearn. URL: https://monkeylearn.com/ (дата звернення: 15.11.2021).

13. Social Searcher - Free Social Media Search Engine. Social Searcher. URL: https://www.social-searcher.com (дата звернення: 15.11.2021).

14. Xiong F., Liu Y. Opinion formation on social media: An empirical approach. Chaos: An Interdisciplinary Journal of Nonlinear Science. 2014. T. 24, № 1. URL: https://aip.scitation.org/doi/full/10.1063/1.4866011 (дата звернення: 15.11.2021).

15. Peng W. Generate adjective sentiment dictionary for social media sentiment analysis using constrained nonnegative matrix factorization. in Proceedings of the International AAAI Conference on Weblogs and Social Media. 2011. URL: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewPaper/2723 (дата звернення: 15.11.2021).

16. Levchenko O., Dilai M. Attitudes Toward Feminism in Ukraine: A Sentiment Analysis of Tweets. Advances in Intelligent Systems and Computing III. CSIT 2018. Advances in Intelligent Systems and Computing. 2019. T. 871. URL: https://link.springer.com/chapter/10.1007/978-3-030-01069-0_9 (дата звернення: 15.11.2021).

17. SentiStrength - sentiment strength detection in short texts - sentiment analysis, opinion mining. SentiStrength. URL: http://sentistrength.wlv.ac.uk/#About (дата звернення: 15.11.2021).

18. Developer Agreement and Policy – Twitter Developers | Twitter Developer Platform. Developer Agreement and Policy. URL: https://developer.twitter.com/en/developer-terms/agreement-and-policy. (дата звернення: 15.11.2021).

19. Where to get Twitter data for academic research • Social Feed Manager. Social Feed Manager. URL: https://gwu-libraries.github.io/sfm-ui/posts/2017-09-14-twitter-data (дата звернення: 15.11.2021).

20. Getting Started with the Twitter API | Docs | Twitter Developer Platform. Twitter Developer Platform. URL: https://developer.twitter.com/en/docs/twitter-api/getting-started/about-twitter-api#Access (дата звернення: 15.11.2021).

21. ParseHub | Free web scraping - The most powerful web scraper. ParseHub. URL: https://www.parsehub.com (дата звернення: 15.11.2021).

22. Твіттер. Все, що актуально. Твіттер. URL: https://twitter.com (дата звернення: 15.11.2021).

23. Use Cases, Tutorials, & Documentation | Twitter Developer Platform. Twitter Developer Platform. URL: https://developer.twitter.com/en (дата звернення: 15.11.2021).

24. GET /2/tweets/search/all | Docs | Twitter Developer Platform. Twitter Developer Platform. URL: https://developer.twitter.com/en/docs/twitter-api/tweets/search/api-reference/get-tweets-search-all (дата звернення: 15.11.2021).

25. Twitter API Tools. URL: https://developer.twitter.com/apitools/api (дата звернення: 15.11.2021).

26. Zhai S. NLP With Python: Build a Haiku Machine in 50 Lines Of Code. Medium. URL: https://betterprogramming.pub/nlp-with-python-build-a-haiku-machine-in-50-lines-of-code-6c7b6de959e3 (дата звернення: 15.11.2021).

27. Smedt T. D., Daelemans W. Pattern for Python. Journal of Machine Learning Research 13. 2012. URL: https://libraries.universityofcalifornia.edu/groups/files/about/desmedt12a.pdf (дата звернення: 15.11.2021).

28. Hutto C. J., Gilbert E. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). 2014. URL: http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf (дата звернення: 15.11.2021).

29. Baccianella S., Esuli A., Sebastiani F. SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. Istituto di Scienza e Tecnologie dell'Informazione. URL: http://nmis.isti.cnr.it/sebastiani/Publications/LREC10.pdf (дата звернення: 15.11.2021).

30. Про охорону праці. Офіційний вебпортал парламенту України. URL: https://zakon.rada.gov.ua/laws/show/2694-12#Text (дата звернення: 15.11.2021).

31. Державні санітарні правила і норми роботи з візуальними дисплейними терміналами електронно-обчислювальних машин. Офіційний вебпортал парламенту України. URL: https://zakon.rada.gov.ua/rada/show/v0007282-98#Text (дата звернення: 15.11.2021).

32. Про затвердження Вимог щодо безпеки та захисту здоров'я працівників під час роботи з екранними пристроями. Офіційний вебпортал парламенту України. URL: https://zakon.rada.gov.ua/laws/show/z0508-18#Text (дата звернення: 15.11.2021).

33. Санітарні норми мікроклімату виробничих приміщень ДСН 3.3.6.042-99. Офіційний вебпортал парламенту України. URL: https://zakon.rada.gov.ua/rada/show/va042282-99#Text (дата звернення: 15.11.2021).

ДОДАТОК А

Тези конференції

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Тернопільський національний технічний університет імені Івана Пулюя (Україна)
Університет імені П'єра і Марії Кюрі (Франція)
Маріборський університет (Словенія)
Технічний університет у Кошице (Словаччина)
Вільнюський технічний університет ім. Гедимінаса (Литва)
Білоруський національний технічний університет (Республіка Білорусь)
Міжнародний університет цивільної авіації (Марокко)
Наукове товариство ім. Т.Шевченка

АКТУАЛЬНІ ЗАДАЧІ СУЧАСНИХ ТЕХНОЛОГІЙ

Збірник
тез доповідей
Том І

X Міжнародної науково-практичної конференції молодих учених та студентів
24-25 листопада 2021 року

УКРАЇНА
ТЕРНОПІЛЬ – 2021

УДК 004.77-042.3:316.4

Є.В. Тиш, канд. техн. наук
В.В.Б. Кохан
Тернопільський національний технічний університет імені Івана Пулюя, Україна

ФОРМУВАННЯ СУСПІЛЬНОЇ ДУМКИ В СОЦІАЛЬНИХ МЕРЕЖ НА ПРИКЛАДІ МЕРЕЖІ TWITTER

Ie.V. Tysh
V.V.B. Kokhan

FORMATION OF PUBLIC OPINION IN SOCIAL NETWORKS ON THE EXAMPLE OF THE TWITTER NETWORK

Сьогодні складно уявити своє життя без гаджетів та інтернету. Сучасні технології значно спростили процеси роботи та побуту людей 21 століття. Соцмережі, які прийшли на заміну листуванню – також стали невід'ємною частиною побуту. Соціальні мережі, окрім очевидного спілкування з друзями, дозволили поширювати свої думки чи ідеї в маси та впливати на погляди людей які не знайомі між собою чи навіть можуть жити на іншому кінці світу.

Твіттер це соціальна мережа мікроблогів, яка дає змогу користувачам надсилати короткі текстові повідомлення до 280 символів, використовуючи SMS, служби миттєвих повідомлень і сторонні програми-клієнти. Створений у 2006 році, твіттер незабаром завоював популярність у всьому світі. Станом на 1 січня 2011 року сервіс нараховував понад 200 млн користувачів. За даними опублікованими у звіті про доходи компанії за 3 квартал 2021 року [1] середня кількість активних користувачів, яким можна показувати рекламу, у дні підзвітного періоду складала 211 мільйонів акаунтів. За даними аналітичного вебсайту «Datareportal» [2] опублікованими у листопаді 2021 року в Україні знаходиться 817.6 тисяч потенційних користувачів, яким можна показувати рекламу, які складають 2.2% від загальної кількості користувачів Твіттер віком старше 13 років. Загальний показник потенційних переглядачів реклами на платформі становить 463 мільйони користувачів.

Окрім висловлення своєї особистої думки, на платформі Твіттер можна створювати облікові записи організацій чи офіційних установ і ділитися новинами чи планами на подальшу роботу з читачами. У залежності від кількості підписників та займаної у суспільстві посади твіти можуть сильно впливати на економіку та інші сфери життя. Для прикладу твіт генерального директора та архітектора продуктів компанії Tesla, Inc., Ілона Маска [3] знизив вартість компанії Tesla, Inc. майже на 100 доларів США за даними американського бізнес-медіа Inc. [4]. Такі коливання на фондовому ринку за лічені хвилини завдали численних втрати інвесторам компанії Tesla, Inc. Також в інтернеті можна зустріти статті на тему спекуляцій Ілона Маска з криптовалютами Bitcoin та Dogecoin.

У статті [5], опублікованій у 2014 році, описано емпіричне дослідження формування думок у соціальній мережі Твіттер. У дослідженні було використано близько 6 мільйонів твітів створених 2.3 мільйонами авторів на протязі 2010 року. Результати показали що соціальні мережі можуть формувати та змінювати думку людини, проте цей процес відбувається повільно, а кількість агентів що змінюють свою думку спадає за степеневим розподілом. Позитивно на зміну думки чи її наближення до запропонованої у мережі впливає активність поширення запропонованої думки, кількість авторів які її поширюють та наявність однієї популярної думки. Додатково було проведено симуляцію яка імітувала поведінку людей у соціальній мережі та підтвердила факт що взаємодії через соціальні мережі формують думки людей.

З дослідження [6] опублікованого у 2020 році випливає висновок що особливо активно діляться своїми повідомленнями та думками самовпевнені, екстравертивні та недобросовісні агенти, яким не важлива репутація та думка інших. У статті досліджували поведінку поширення повідомлень користувачами у соцмережах. Для проведення

експерименту було симульовано агентів з 3 характеристиками, кожна з яких могла приймати одне значення із діапазону між 0 та 1. Агенти обмінювались повідомленнями у 6 різних типах мереж.

Підсумовуючи результати наведених досліджень та згаданих статей – варто пам'ятати про цифрову безпеку та грамотність. Для боротьби з дезінформацією необхідно перевіряти сумнівні твердження опубліковані у мережі та шукати декілька точок зору для того щоб отримати повну картину подій. Беручи до уваги напружені стосунки України з Росією – Україні варто посилювати свій вплив та присутність у соціальних мережах для зміцнення морального духу та збільшення міжнародної підтримки України.

Література:

1. Microsoft Word - Q3-21 Earnings Release Final_10.25.21__702pm.docx [Електронний ресурс] – Режим доступу до ресурсу: https://s22.q4cdn.com/826641620/files/doc_financials/2021/q3/Final-Q3'21-earnings-release.pdf.
2. The Latest Twitter Stats: Everything You Need to Know — DataReportal – Global Digital Insights [Електронний ресурс] – Режим доступу до ресурсу: https://datareportal.com/essential-twitter-stats.
3. Elon Musk on Twitter: "Tesla stock price is too high imo" / Twitter [Електронний ресурс] – Режим доступу до ресурсу: https://twitter.com/elonmusk/status/1256239815256797184.
4. Elon Musk's Tweets Move Markets. This Time, Downward | Inc.com [Електронний ресурс] – Режим доступу до ресурсу: https://www.inc.com/don-reisinger/elon-musks-tweets-move-markets-this-time-downward.html.
5. Xiong F. Opinion formation on social media: An empirical approach: Chaos: An Interdisciplinary Journal of Nonlinear Science: Vol 24, No 1 [Електронний ресурс] / F. Xiong, Y. Liu // Chaos: An Interdisciplinary Journal of Nonlinear Science. – 2014. – Режим доступу до ресурсу: https://aip.scitation.org/doi/full/10.1063/1.4866011.
6. Opinion Formation on the Internet: The Influence of Personality, Network Structure, and Content on Sharing Messages Online [Електронний ресурс] / L.Burbach, P. Halbach, M. Ziefle, A. Calero Valdez // Frontiers in Artificial Intelligence. – 2020. – Режим доступу до ресурсу: https://doi.org/10.3389/frai.2020.00045.

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ТЕРНОПІЛЬСЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
ІМЕНІ ІВАНА ПУЛЮЯ

М А Т Е Р І А Л И

IX НАУКОВО-ТЕХНІЧНОЇ КОНФЕРЕНЦІЇ

«ІНФОРМАЦІЙНІ МОДЕЛІ, СИСТЕМИ ТА ТЕХНОЛОГІЇ»

8–9 грудня 2021 року

ТЕРНОПІЛЬ
2021

УДК 004.8+004.02:[004.91+004.93]
В.В. Кохан, Є.В. Тиш, канд. техн. наук
(Тернопільський національний технічний університет імені Івана Пулюя, Україна)

МЕТОДИ ОЦІНЮВАННЯ ЕМОЦІЙНОГО НАХИЛУ ТЕКСТІВ ЗАСОБАМИ ШТУЧНОГО ІНТЕЛЕКТУ

UDC 004.8+004.02:[004.91+004.93]
V.V. Kokhan, Ie.V. Tysh

METHODS OF EVALUATION OF SENTIMENT ANALYSIS OF TEXTS BY MEANS OF ARTIFICIAL INTELLIGENCE

Штучний інтелект – популярний на сьогодні напрям проведення багатьох досліджень, який дозволяє робити неймовірні речі з використанням можливостей сучасних комп'ютерних технологій. До досліджень в цій галузі залучають спеціалістів усіх галузей знань у залежності від очікуваних результатів.

Виданнями, що спеціалізуються на штучному інтелекті [1], було визначено ключові завдання штучного інтелекту, такі як: комп'ютерний зір, машинне навчання, обробка природної мови та інші. Кожне таке завдання, або комбінація завдань, дозволило окреслити великий перелік простіших, конкретніших цілей, рішення яких вже почали впливати на наше життя.

Для прикладу оцінка емоційного нахилу (sentiment analysis) використовує напрацювання машинного навчання та обробки природної мови для класифікації текстових даних за їх емоційним нахилом та об'єктивністю. Опрацьовані дані потім застосовуються компаніями та брендами для того, щоб зрозуміти як їхні користувачі та фанати сприймають рішення компанії з того чи іншого питання, або навіть для виявлення проблем, про які компанія могла не здогадуватись. Така інформація є свого роду компасом суспільної думки, який показує реакцію суспільства на діяльність компанії, при цьому не потребуючи десятків чи сотень працівників, які будуть перечитувати усі відгуки від користувачів.

Існує декілька методів автоматизованої оцінки емоційного нахилу. Методи засновані на правилах і словниках, статистичні та комбіновані методи. Методи засновані на правилах і словниках – дозволяють проаналізувати текст за допомогою попередньо складених словників та правил лінгвістичного аналізу [2]. Суть цього методу полягає у присвоєнні кожному слові значення зі словника, якщо воно є, а за загальну оцінку тексту приймають суму оцінок усіх слів. Хоча безпосередньо застосування цього методу є доволі простим – основна частина роботи припадає на складання словника з правильними вагами слів для досліджуваної галузі. Для прикладу, слово «великий» буде мати позитивне значення, якщо мова буде йти про обсяг пам'яті жорсткого диска і негативне, якщо мова буде йти про розміри телефона.

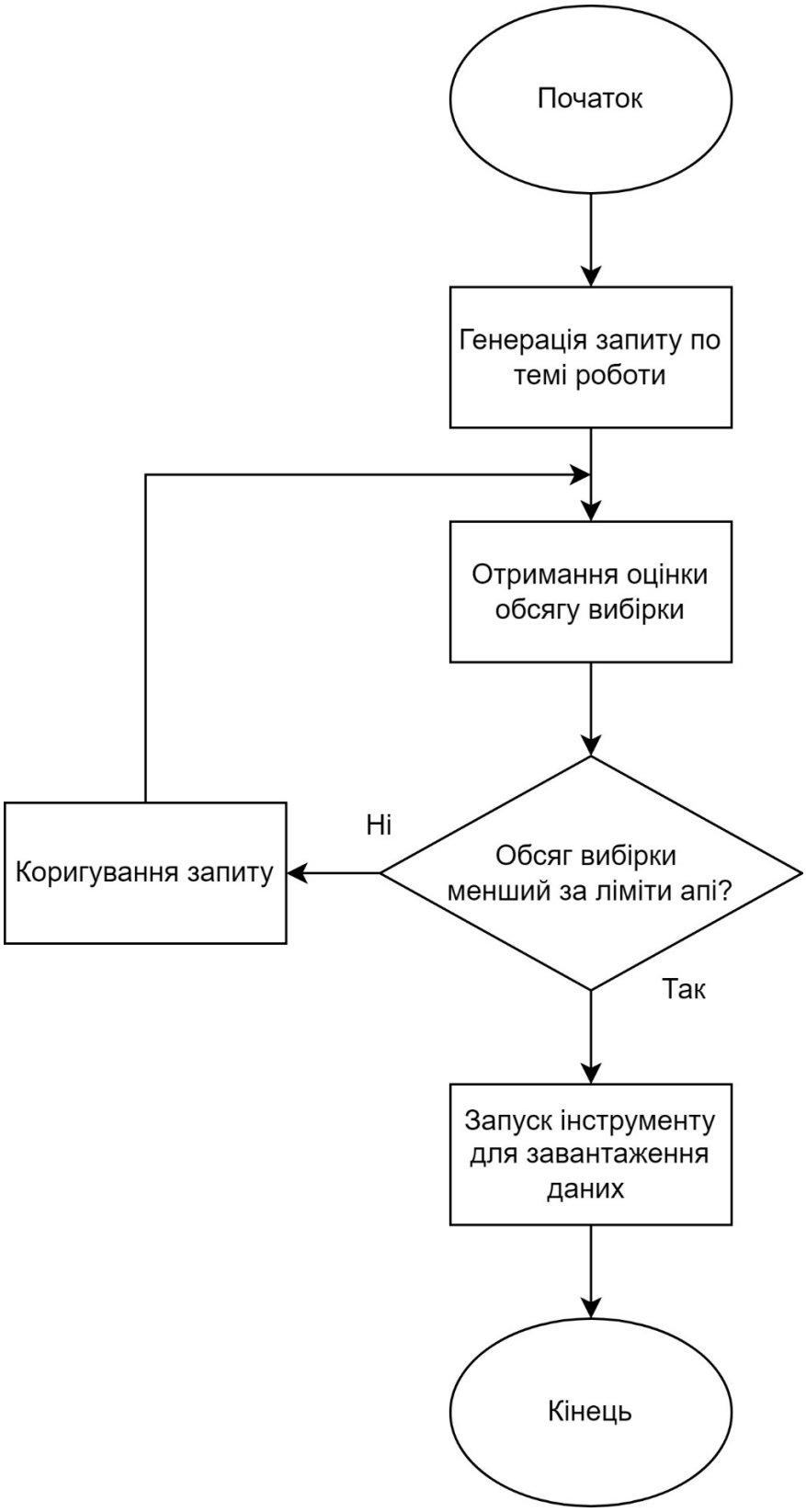
Статистичні методи через хороші результати в інших завданнях штучного інтелекту набирають дедалі більшої популярності. Сюди належать машинне навчання без вчителя, та метод заснований на теоретико-графових моделях. У першому випадку текст розбивають на ключові терміни, які людина позначає як позитивні чи негативні, маючи таку інформацію – система робить висновок про емоційний нахил всього тексту. Точність такого методу вища, але потребує великого обсягу даних для тренування моделі. Другий метод припускає що різні слова мають різний вплив на емоційний нахил тексту, тому потребує створення спеціальних графів досліджуваного тексту, які потім проходять процес ранжування вершин, класифікації знайдених слів і лише після того дозволяють отримати результат.

Література.

1. Russell S. Artificial Intelligence: A Modern Approach, 4th US ed. [Електронний ресурс] / S. Russell, P. Norvig – Режим доступу до ресурсу: http://aima.cs.berkeley.edu/.
2. Cambria E. Affective Computing and Sentiment Analysis [Електронний ресурс] / Erik Cambria // IEEE Intelligent Systems. – 2016. – Режим доступу до ресурсу: https:// ieeexplore.ieee.org/document/7435182.

ДОДАТОК Б

Блок-схема алгоритму коригування запиту для завантаження даних



ДОДАТОК В

Блок-схема алгоритму проведення оцінки емоційного нахилу текстів