НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»

Факультет інформатики та обчислювальної техніки

Кафедра інформаційних систем та технологій

Індивідуальний дослідницький проєкт

на здобуття ступеня бакалавра
за освітньо-професійною програмою «Інтегровані інформаційні системи»
спеціальності 126 «Інформаційні системи та технології»
на тему: «Інтелектуальна система аналізу тональності україномовних відгуків»

Виконала:
студентка IV курсу, групи ІА-81
Никифорова Олександра Олександрівна _____

Керівник:
доцент каф. ІСТ, к. т. н.
Шимкович Володимир Миколайович _____

Засвідчую, що у цьому проєкті немає запозичень з праць інших авторів без відповідних посилань.
Студентка _____

Київ – 2022 року

Національний технічний університет України

«Київський політехнічний інститут імені Ігоря Сікорського»

Факультет інформатики та обчислювальної техніки

Кафедра інформаційних систем та технологій

Рівень вищої освіти – перший (бакалаврський)

Спеціальність – 126 «Інформаційні системи та технології»

Освітньо-професійна програма «Інтегровані інформаційні системи»

ЗАВДАННЯ

на індивідуальний дослідницький проєкт студентці

Никифоровій Олександрі Олександрівній

1. Тема проєкту «Інтелектуальна система аналізу тональності україномовних відгуків», керівник проєкту Шимкович Володимир Миколайович, доцент каф. ІСТ, к. т. н.

2. Термін подання студенткою проєкту: 15 червня 2022 року

3. Вихідні дані до проєкту: система, що виконує аналіз тональності відгуків різними методами, на основі проведеного навчання і тестування з залученням власноруч створеної бази україномовних відгуків.

4. Зміст пояснювальної записки: огляд області дослідження, аналіз існуючих рішень, проведення попередньої підготовки текстових даних, проектування графічного представлення системи, програмна реалізація з подальшим порівнянням отриманих результатів.

5. Перелік графічного матеріалу (із зазначенням обов’язкових креслеників, плакатів, презентацій тощо): схема структурна, діаграма прецедентів, діаграма діяльності, діаграма послідовностей.

6. Дата видачі завдання 19 травня 2022 року

Календарний план

№ з/п	Назва етапів виконання дослідницького проєкту	Термін виконання етапів проєкту	Примітка
1	Огляд області дослідження	09.05.2022	
2	Аналіз існуючих рішень	16.05.2022	
3	Попередня підготовка текстових матеріалів	23.05.2022	
4	Розробка графічних матеріалів	30.05.2022	
5	Програмна реалізація системи	13.06.2022	
6	Оформлення документації дослідницького проекту	19.06.2022	

Студентка Олександра НИКИФОРОВА

Керівник Володимир ШИМКОВИЧ

АНОТАЦІЯ

Никифорова О.О. Інтелектуальна система аналізу тональності відгуків. КПІ ім. Ігоря Сікорського, Київ, 2022.

Проєкт містить 70 с. тексту, 20 рисунків, 6 таблиць, посилання на 20 літературних джерел та 4 конструкторських документа.

Ключові слова: аналіз тональності, сентимент-аналіз, україномовні відгуки, аналіз відгуків.

Об’єктом розробки є система, що визначає тональну забарвленість поданих відгуків.

Мета розробки – підвищення користувацького досвіду за допомогою аналізу відгуків на товари.

Робота включає аналіз теоретичної складової підходів та методів визначення тональності відгуків; існуючих застосунків, які визначають тональну складову коментарів і текстів; вибір і обґрунтування методів класифікації; формування бази україномовних відгуків; опис роботи обраних алгоритмів, що виконують розпізнавання тональності; а також порівняння отриманих результатів з подальшим визначенням кращого методу для класифікації відгуків.

Система виконує аналіз поданих відгуків, попередньо обробивши і представивши їх у векторному вигляді, визначаючи їх тональне забарвлення за допомогою обраних класифікаторів.

SUMMARY

Nykyforova O. Intelligent system of sentiment analysis of Ukrainian-language reviews. Igor Sikorsky KPI, Kyiv, 2022.

The project contains 70 pages. text, 20 figures, 6 tables, links to 20 literary sources and 4 design documents.

Keywords: sentiment analysis, review analysis, sentiment review analysis, Ukrainian-language reviews.

The object of development is the Ukrainian-language review sentiment analysis system.

The purpose of the development - increasing Customer Experience basing on analysis of product reviews.

Scope of the held work includes analysis and research of existing methods of sentiment analysis, which are applied to analysis reviews written in the Ukrainian language. In addition, current research is dedicated to the analysis of linguistic features distinctive to the Natural Language Processes of the Ukrainian language.

An implemented system should perform reviews preprocessing with the following definition of sentiment component.

Номер рядка	Формат	Позначення	Найменування	Кільк. аркушів	Номер екзем.	Примітка
1			Документація загальна			
2						
3			Знову розроблена			
4						
5	А4	ІА81.200БАК.003 ПЗ	Пояснювальна записка	70		
6	А3	ІА81.200БАК.003 Э1	Інтелектуальна система	1		
7			аналізу тональності			
8			україномовних відгуків			
9			Схема структурна			
10	А3	ІА81.200БАК.003 Д1	Інтелектуальна система	1		
11			аналізу тональності			
12			україномовних відгуків			
13			Діаграма прецедентів			
14	А3	ІА81.200БАК.003 Д2	Інтелектуальна система	1		
15			аналізу тональності			
16			україномовних відгуків			
17			Діаграма діяльності			
18	А3	ІА81.200БАК.003 Д3	Інтелектуальна система	1		
19			аналізу тональності			
20			україномовних відгуків			
21			Діаграма послідовностей			
22						
23						
24						
25						
26						
27						
28						

Зм.	Аркуш	№ докум.	Підпис	Дата	ІА81.200БАК.003 ТП		
Розроб.		Никифорова О.О.			Інтелектуальна система аналізу тональності україномовних відгуків Відомість проєкту	Літ.	Аркуш 1 / Аркушів 1
Керівн.		Шимкович В.М.				Т	
						КПІ ім. Ігоря Сікорського Група ІА-81	
Затв.							

Пояснювальна записка
до індивідуального дослідницького проєкту на тему: «Інтелектуальна система аналізу тональності україномовних відгуків»

Київ – 2022

ЗМІСТ

ПЕРЕЛІК СКОРОЧЕНЬ

CX (Customer Experience) – користувацький досвід

NLP (Natural Language Processing) – сфера дослідження природомовних процесів

БД – база даних

BOW – Bag of words

МОВ (англ.) – метод опорних векторів

НБК – наїіний Байєсівський класифікатор

ПНБК – поліноміальний Байєсівський класифікатор

CSV – Comma-separated values

ВСТУП

Безліч інтернет-ресурсів нині є найактуальнішими джерелами інформації у сучасному світі, особливо серед молодого покоління, що несуть відповідальність за величезну кількість суспільних рішень. Всесвітня мережа заповнена нескінченними обсягами різноманітної інформації стосовно будь-якої теми. Різні джерела надають вільний доступ до цієї інформації, а це означає, що будь-який користувач може черпати нові знання з ресурсів, які не обов’язково оперують достовірними та об’єктивними фактами. Коментарі в різноманітних соціальних мережах або спеціальних сайтах, створених для відгуків поширюють різні погляди людей на один і той же товар, послугу чи об'єкт. За відсутності можливості критичного оцінювання обставин за яких був написаний відгук, наслідком може стати неправильно сформована думка про товар або послугу, яка виникла не в завдяки результатам власних висновків, а через отриману інформацію з Інтернету.

З точки зору ведення бізнесу, для організацій та підприємств, які дбають про якість наданих товарів або послуг, відгуки мають виключно важливу роль у формуванні суспільної думки про обраний продукт. Наявність відгуків надає можливість відслідкувати суспільні думки та реакції на наданий товар. Завдяки цьому спрощуються наступні процеси: збір статистики про цільову аудиторію, оцінка переваг та недоліків послуги перед конкурентами та виявлення безлічі інших критеріїв, без урахування яких продукт приречений на провал. Для того, щоб дана концепція, враховуючи величезні обсяги даних, які з часом лише збільшуються, працювала як годинник, потрібно щонайменше цілий штат досвідчених аналітиків, SMM та PR менеджерів, які майстерно виконують свою роботу. Однак треба розуміти, що такий підхід є марнуванням людського ресурсу, адже в такому випадку команда спеціалістів буде виконувати рутинну роботу, замість розробки дієвих стратегій в інших галузях вдосконалення продукту. Тому значно актуальнішим рішенням даного питання в умовах сучасного світу з

динамічним розвитком інформаційних технологій – створення автоматизованої системи аналізу відгуків.

Метою даного дослідницького проєкту є розробка інтелектуальної системи з аналізу україномовних відгуків.

Після проведеного аналізу коментаря або відгука даний застосунок повинен визначати його емоційне забарвлення відповідно до заданих в програмі категорій. Таке рішення значно спростить процес аналізу, адже представляє собою автоматичне оброблення величезних обсягів обробки текстових даних. На сьогоднішній день відгуки можна написати на будь який товар або послугу: будь то заклад харчування, сервіс оренди квартир або конкретний продукт, який можна придбати. Тому даний застосунок не має конкретних обмежень на область про яку написано відгук.

Важливо зазначити, що для системи аналізу тональності відгуків буде створено відповідну базу даних, що означає обмежену кількість слів, яка згодом може збільшуватися. Беручи до уваги дане обмеження, було визначено, що програма буде оброблювати виключно україномовні текстові масиви. В даному випадку такі умови формують ключову перевагу розроблювальної системи, адже на сьогоднішній день визначається значна нестача матеріалів стосовно саме тонального аналізу текстів, що написані українською мовою.

Протягом виконання дослідницького проєкту було вирішено наступні задачі:

- Огляд теоретичних відомостей про обрану галузь дослідження;
- аналіз існуючих рішень;
- попередня підготовка текстових даних;
- проектування графічного представлення системи;
- програмна реалізація;
- порівняння обраних методів визначення тональної складової.

1 ОГЛЯД ОБЛАСТІ ДОСЛІДЖЕННЯ

1.1 Аналіз предметної області

Під тональністю ми визначаємо ставлення, що зазвичай носить емоційне забарвлення, автора тексту до конкретного об'єкту (або послуги, події, товару), виражене у відгуку. Емоційна складова, виражена на рівні лексеми або комунікативного фрагмента, називається лексичною тональністю (або лексичним сентиментом). Тональність всього тексту в цілому можна визначити як функцію (в найпростішому випадку суму) лексичних тональностей складових його одиниць (речень) і правил їх поєднання.

Аналіз тональності тексту (сентимент-аналіз, англ. Sentiment analysis, англ. Opinion mining) — клас методів контент-аналізу в комп'ютерній лінгвістиці, призначений для автоматизованого виявлення в текстах емоційно забарвленої лексики і емоційної оцінки авторів (думок) по відношенню до об'єктів, мова про які йде в тексті [1].

Результати такого аналізу відгуків можуть набувати значення вкрай позитивного впливу на ведення бізнесу вирішуючи задачу покращення клієнтського досвіду (Customer Experience або CX). В основі клієнтського досвіду лежать відчуття що виникають у результаті взаємодії користувача з брендом, що надає послугу. Тональний аналіз може бути корисним на наступних етапах:

– Відстеження настроїв користувачів. Звісно, таким способом стає можливим збір бінарно-налаштованих думок – позитивних та негативних. На основі цього формується статистичний звіт щодо виявлених недоліків та переваг і відбувається усунення проблеми та покращення існуючих сервісів відповідно. Більш розширений аналіз емоційних забарвлень надає можливість коректним чином сформувати політику бренду відносно набутої на момент аналізу аудиторії.

– Моніторинг конкуренції. Сентимент-аналіз можна використовувати не тільки для власного бізнесу, а й для детального розгляду зворотного зв’язку

компаній, які знаходяться в спільному галузевому діапазоні відносно бізнесу, з яким ведеться порівняння. Таким способом можливо превентивно уникнути невдалих рішень, що можуть призвести до неочікуваних втрат та наслідків, а також запровадити нові способи покращення сервісу.

– Уникнення кризових станів. Завдяки можливому відокремленню виключно негативних відгуків стає можливим локалізувати невдоволення користувача на різних етапах надання послуги (стан об'єкту, час очікування, якість сервісу, тощо).

Необхідність аналізу тональності, що визначає емоційну складову відгуків в наш час набуває важливості не тільки для організацій, що продають свої продукти на ринку послуг, а й для звичайних користувачів. Обмін думками та почуттями – позитивними та негативними, щодо наданого продукту або сервісу між клієнтами є звичайною людською нормою. Так, спираючись на досвід попередніх користувачів, людина може зробити висновки щодо того, чи підходить їй об'єкт описаний у коментарі, ступінь необхідності цього об'єкту та визначити персональні ризики.

Сентимент-аналіз може бути сформований як контекстуальна оцінка суб'єктивної інформації, що надається автором коментаря. Поняття відгук можна привести до формульного вигляду:

$$\langle f, m, o \rangle, \tag{1.1}$$

де f – особливість обраного продукту;

o – висловлене ставлення користувача до f;

m – детермінант, що використовується для формування експресивності o [2].

Зазвичай тональний аналіз прийнято розглядати як задачу обробки природньої мови (Natural Language Processing). Natural Language Processing (скорочено – NLP) – сфера застосування алгоритмів штучного інтелекту, що знаходиться на перетині машинного навчання та мовознавства.

В основі NLP полягає аналіз природньої розмовної мови, для якої притаманно використання скорочень і нехтування деякими морфологічним, фонетичними та пунктуаційними правилами, що є недопустимим для штучних та літературних мовних форм.

Слід зауважити, що однією з проблем аналізу тональності є коментарі з іронічним або саркастичним забарвленням. «Проплачені» рекламні відгуки, що бувають як, позитивними, так і негативними – якщо розміщені на сторінці послуги конкурента також завдають перешкод.

Визначення та відокремлення таких відгуків є майже непосильною задачею для комп'ютера через те, що нерідко навіть живий користувач зазнає значних труднощів з класифікацією подібних висловлювань. Тому такі випадки можна розглядати як виключення.

Ще однією складністю можна вважати аспект емоційного забарвлення конкретного слова. Час від часу трапляються випадки коли слово, яке прийнято вживати у негативному контексті набуває абсолютно позитивного значення і навпаки. Також на загальний тон речення, або навіть цілого повідомлення, часто впливають слова, що підсилюють емоційний відтінок слова, що стоїть після (наприклад, «значно», «достатньо», «менш», «більш-менш»). Цілком непередбачуваним характером наділені частини мови, що відповідають за заперечення (такі як «ніколи», «нінащо», «не», «ніхто», «ні» та багато інших), адже їх основна властивість – змінення значення тексту на протилежне.

Окремою проблемою можна виділити такі природомовні словоформи, як діалектизм та неологізми.

Діалектизмом прийнято вважати слова, що притаманні для використання на певній території, або у соціальному колі. Серед носіїв української мови більш поширеним є територіальний вид діалектизму.

Використання неологізмів у контексті неформальної мови набирає все більших обертів, особливо серед молоді, тому сьогодні все частіше можливо спостерігати вживання новостворених слів, значення яких ще недостатньо закріпилося у соціумі.

Вирішення подібних мовних питань потребує окремого розгляду кожного прецеденту та коригування існуючої бази даних опираючись на отримані результати мовознавцями, що детальним чином займаються дослідженням обробки природньої мови.

1.2 Класифікація тонального аналізу

В реаліях сучасності об’єктивна міра оцінювання тональних відтінків визначається дуальністю емоційного простору – існує забарвлення хороше і погане. Сентиментальний напрямок думки, що висловлена у відгуку, виражається особливістю *f,* яка визначає чи є коментар позитивним, негативним, або нейтральним [3].

Нейтральну оцінку прийнято відносити до стану «посередині» цих двох понять. Оскільки позитивний та негативний стан – цілком полярні значення, то нейтральність прирівнюється до відсутності будь-якого емоційного тону. Також під дане поняття деколи відносять текстові масиви, де надано недостатню тональну характеристику об’єкту аналізу, - їх можна вважати «невизначеними». Такі коментарі також містять в собі інформацію щодо продукту, на який написаний відгук, і, в першу чергу, впливають на кількісну складову коментарів. Для звичайного користувача це надзвичайно важливий показник, адже з точки зору психології, велика кількість відгуків на продукт або послугу викликає значно більше довіри, аніж сервіс на який відгуків немає зовсім.

Важливо зазначити, що нині все більша кількість дослідників висловлює думку про те, що обробка нейтрально-забарвленої інформації прямим чином впливає на точність результатів тонального аналізу, тому ігнорування таких коментарів не є допустимим. Однак для тонального оцінювання такі коментарі не несуть жодного інформаційного сенсу, тому їх прийнято видаляти на одному з етапів аналізу [4].

Базовий сентимент-аналіз прийнято класифікувати наступним чином:

– Класифікація за бінарною шкалою. В такій моделі класифікації використовується лише одномірна емоційна оцінка: позитивна або негативна.

– Класифікація за багатосмуговою шкалою. Бінарна шкала набуває додаткових значень. В даному випадку шкала {позитивно; негативно} розмежовується на проміжні уточнюючі емоціональні значення між абсолютно позитивним та абсолютним негативним. Наприклад, оцінку тона можна представити у наступному вигляді: «дуже позитивно», «позитивно», «нейтрально», «негативно», «дуже негативно».

– Класифікація методом систем шкалювання. Сутність даної моделі полягає у представленні оцінки в межах чисельної послідовності, наприклад, від -10 до 10 де найменше число відповідає найгіршій оцінці, а найбільше – найкращій відповідно [1]. На сьогоднішній час більш доцільним є використання діапазону {0; +5}, що демонструють сучасні системи коментарів та відгуків.

Окремим напрямом дослідження аналізу сентиментальної складової є визначення об’єктивності або суб’єктивності інформації, що написана у відгуку. Ця область передбачає детальний розгляд настроїв, що виражаються у коментарях, тому рамки однополярності не завжди є доцільними для використання, адже в даному випадку тональний відтінок, що набуває коментар, залежить від контексту використаних слів та частин мови.

За способом ідентифікації суб’єктивної/об’єктивної складової можна виділити наступні класифікації:

– Класифікація на основі особливості/аспекту. Такій моделі притаманний більш детальний аналіз, який описує відношення різноманітних функцій, що висловлені в коментарі. Таким чином, думка, висловлена у відгуку має наступний вигляд [3]:

$$(o_j,\ f_{jk}, oo_{ijkl}, h_i, t_l), \qquad (1.2)$$

де o_j – об’єкт відгуку;

f_{jk} – функція, яку виконує об’єкт o_j;

oo_{ikjl} – тональний напрямок думки;

h_i – власник думки;

t_l – час, в який думка була виражена h_i.

Така модель вимагає визначення всіх сутностей, аспектів, полярності думки та особливостей описаних у відгуку.

– Класифікація за визначенням емоцій. Поняття «емоція» було розглянуто під різними кутами та в різних наукових сферах, таких як філософії, психології, біології, соціології, біохімії тощо. Наразі виділено чотири базових типи емоцій, а саме: радість, смуток, злість та страх, які відносяться до ключових галузей впливу на людину: заохочення (радість), осуд (сум) та стрес (страх та злість). Відносно цих галузей можна провести аналогію до трьох основних кольорів (жовтий, червоний, і синій) в тому сенсі, що при комбінації базових емоцій в різних пропорціях можливо отримати безліч другорядних (наприклад, любов, задоволення, розчарування, здивування та інші) [5]. При використанні даної моделі класифікації емоції визначаються на суб’єктивний розсуд дослідника.

Класифікація за допомогою шкали інтенсивності. В деяких випадках доцільно визначати тональну забарвленість коментаря за допомогою розмежування інтенсивності емоцій, що відчуває власник думки. Шкала інтенсивності використовує різні ступені насиченості певної емоції (наприклад, «чудово» та «блискуче» або «неприємно» і «жахливо»). Така класифікація відрізняється від моделі з багатосмуговою шкалою тим, що у даному випадку можна опустити полярність значень «позитивно» та «негативно» та може прирівнюватись до суб’єктивної градації будь-яких емоційних станів в залежності від поставленої задачі.

Емоційна складова є виключно суб’єктивним поняттям, адже кожна людина проживає різний досвід і відчуття в межах однієї і тієї самої ситуації. Відповідно, і результат процесу визначення емоцій та ступінь їх інтенсивності залежить від

висновків фахівця, що проводить аналіз, спираючись виключно на призму власного сприйняття – це є цілковитою нормою. Доцільно навести приклад відгуку стосовно позиції меню в кав'ярні: «Кава тут – з глузду можна з'їхати!». Словосполучення «з'їхати з глузду» зазвичай носить негативне забарвлення, але в подібних випадках через однозначно оцінити настрій коментаря є досить складною задачею через двояку забарвленість тону.

Реалізація будь-якої моделі класифікації можлива як і ручним способом – коли тональний аналіз проводиться фахівцями з області обробки природньої мови, так і за допомогою різноманітних методів машинного навчання.

1.3 Методи класифікації тонального аналізу

Серед основних методів, що використовуються для класифікації можна виділити наступні:

- словниковий метод;
- метод на основі теоретико-графових моделей;
- машинне навчання з вчителем/без вчителя.

Сутність методу, що заснований на правилах і словниках полягає у пошуку та виділенню з тексту окремих слів, що містять певне тональне забарвлення, відповідно до попередньо складеної бази даних (словника). Також в даному методі можливо використання певних правил, завдяки яким стає можливим урахування контексту (наприклад, підсумовування значення тональностей кожного слова, що заздалегідь було внесене до словника). За результатами підсумків можна обчислити загальну тональність тексту.

Складність такого методу головним чином обумовлена труднощами процесу складання самих баз даних. Точність методу напряму залежить від того, наскільки ретельно був утворений словник та наскільки докладно був урахований контекст. Також перешкод завдає наявність специфічної для кожної галузі лексики, тому

для полегшення процесу складання словника можна розбити на певні предметні області.

В основі теоретико-графового методу лежить припущення, що деякі слова мають вищий бо нижчий вплив на результати аналізу тональності. Дослідження сентимент-аналізу даним методом проходить наступним чином: спочатку на основі оброблювального тексту будується граф; вершини, що в даному випаду є словами, ранжуються (тобто розміщуються за ступенем важливості); виконується класифікація вибраних слів відносно попередньо створеного словника, де кожне слово має певну тональну оцінку («позитивна», «нейтральна» або «негативна»); на основі попередніх кроків проводиться обчислення результату.

Щоб отримати кінцевий результат, необхідно знайти позитивну і негативну складову тексту: обчислюється сума всіх тональних термінів з урахуванням їх ваги. Далі загальна оцінка T визначається відношенням позитивної та негативної оцінок – P та N відповідно, за формулою:

$$T = P/N. \tag{1.3}$$

Нейтральною оцінку можна вважати, якщо отримане значення дорівнює або наближене до 1. Якщо значення більше, то оцінка вважається позитивною, у зворотному випадку, коли отримана відповідь менша одиниці, – негативною.

Машинне навчання використовується для розробки автоматичної моделі визначення тонального аналізу тексту і може бути декількох видів:

– навчання без вчителя;
– навчання з учителем.

Навчання без вчителя у контексті сенимент-аналізу передбачає собою визначення найбільш вагомими слова, що найчастіше вживаються в конкретному тексті, а також можуть зустрічатися в деяких інших текстах з вибірки. На основі цього вибраним словам надається певна тональність, за допомогою якої можна зробити висновок про весь текст.

Найбільш дієвим способом розробки автоматичної системи тонального аналізу є підхід машинного навчання з вчителем, адже результати зазвичай мають вищу точність ніж при навчанні без вчителя. В такому випадку машинний класифікатор виконує навчання на текстах, для яких попередньо була визначена тональність, і на основі отриманих «знань» виконує подальшу обробку з класифікацією тексту.

1.4 Векторизація тексту

Як правило, машинна обробка представляє собою роботу з числами. Векторизація тексту – методологія обробки природньої мови, яка зумовлює собою приведення тексту або окремих слів та речень як числових векторів, завдяки чому задача класифікації приводиться до задачі розпізнавання образів. Для сентимент-аналізу можливе використання наступних способів:

– Bag-Of-Words (BOW, укр. «торба слів») є найбільш поширеним способом векторизації тексту для задач класифікації. «Торба слів» надає кожному слову бінарного значення і зберігає інформацію щодо кількості слів у тексті, не беручи до уваги лексичні зв’язки між словами, а також вагу цих слів.

– N-Grams (укр. n-грами) – техніка векторизації, що представляє ряд слів як послідовність елементів та передбачає побудову словника з урахуванням комбінації слів. На практиці більш вживаними є біграми – послідовність з двух слів, та триграми – з трьох, адже використання послідовностей з більшою кількістю слів робить результат менш ефективним, через недостачу інформаційних об’ємів у навчальній вибірці.

– Word2Vec – метод, який визначає тональність тексту з урахуванням контексту слів, зменшуючи при цьому об’єми даних. Розширенням є модель Doc2Vec, який враховує не тільки контекст, а й порядок слів.

Найбільш оптимальною для вирішення задачі, поставленої у даній дослідницькій роботі, є BOW-модель. Хоч вона не передбачає настільки

глибокого аналізу даних, як інші описані методи, функціонал даної моделі повним чином покриває визначені завдання.

1.5 Алгоритми класифікації сентимент-аналізу

Наразі існує достатня різноманітність алгоритмів класифікації в галузі тонального аналізу текстів, які передбачають використання для машинного навчання як з вчителем, так і без.

1.5.1 Дерево рішень

Дерево рішень (англ. decision tree) – логічний алгоритм класифікації, який, відповідно до своєї назви, використовує деревоподібну структуру для представлення зв'язків між логічними умовами. Алгоритм починає роботу з «кореневого» вузла, від якого йде розгалуження на певні варіанти розвитку подій та закінчує її відповідно до рішень, які були прогнозовані «листям» дерева.

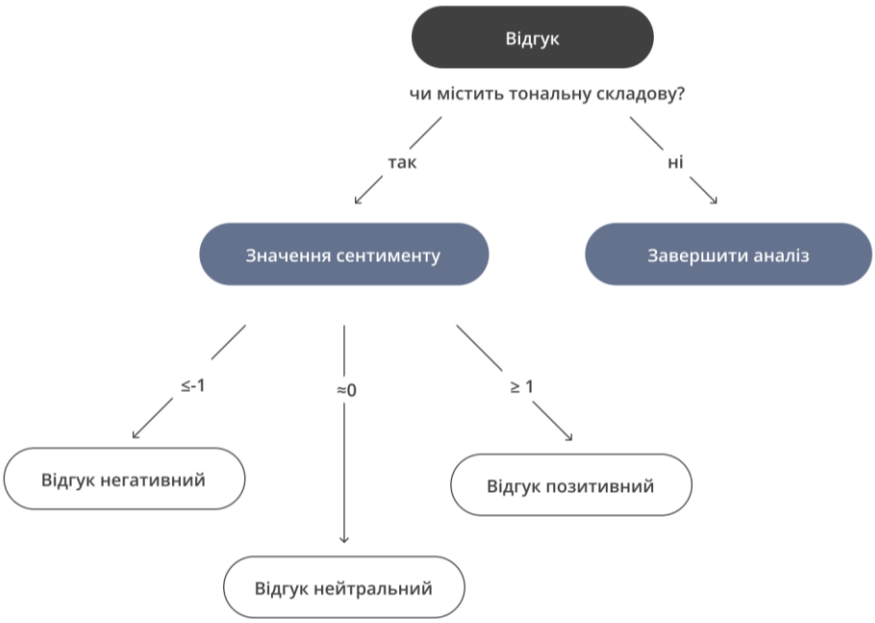


Рисунок 1.1 – Дерево рішень. Приклад роботи алгоритму

Найбільш розповсюдженим прикладом, що заснований на побудові дерева рішень, є C4.5, що виконує побудову з використанням концепції інформаційної ентропії спираючись на набір навчальних даних, що представлені наступною формулою:

$$S = s_1, s_2, \dots s_n, \tag{1.4}$$

де s_i – зразок, що складається p-мірного вектору

$$(x_{1,i}, x_{2,i}, \dots x_{p,i}), \tag{1.5}$$

в якому x_j – певні ознаки зразків s_i [6].

Даний алгоритм працює наступним чином: на кожному вузлі дерева обирається найбільш ефективний для розбиття на підмножини атрибут даних. Ефективність розбиття визначається інформаційний прибуток (тобто різниця ентропії), і, атрибут, в якому інформаційний прибуток набуває найбільшого значення, вибирається для розгалуження і подальшого прийняття рішення. Алгоритм продовжує розбиття на підмножини до моменту досягнення обраних результатів.

1.5.2 Метод опорних векторів

Метод опорних векторів (англ. Support Vector Machine або SVM) є одним з найпоширеніших методів машинного навчання, який можна використати для вирішення задачі класифікації. В контексті тонального аналізу даний метод заснований на ідеї попереднього приведення документів до векторного вигляду і подальшого представлення у векторному просторі, після чого виконується побудова гіперплощини для розділу відомих класів [7].

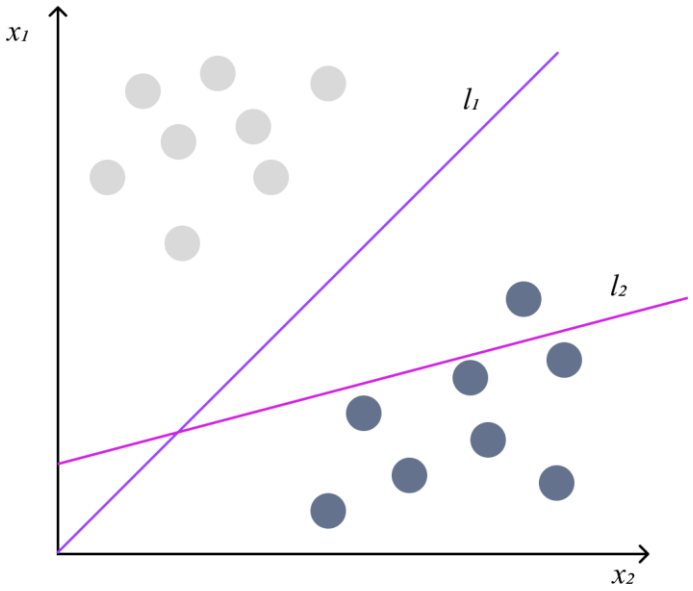


Рисунок 1.2 – Графічне представлення методу опорних векторів. Зображені гіперплощини l_1 та l_2 поділяють елементи на класи

Головною задачею методу в якості класифікатора є знаходження гіперплощини, яка виконає ділення на класи найбільш оптимальним чином. На прикладі задачі бінарної класифікації, можна позначити певний набір даних як $(x_1, y_1), \dots (x_n, y_n)$, що знаходяться в певному просторі R^n, де $y_i = \{-1; 1\}$, відповідно до значення якого відбувається розподілення x_i до певного класу. Для вирішення даної задачі, можна побудувати площину у вигляді рівняння:

$$w \cdot x - b = 0 \tag{1.6}$$

В даному випадку w – вектор, який виконує роль перпендикуляру до гіперплощини, яка розділяє простір. Також враховується параметр $\frac{b}{\|w\|}$, що визначає відстань від гіперплощини до початка координат.

Метод опорних векторів передбачає собою максимального збільшення в відстані (від англ. margin) між об'єктами класифікації та гіперплощиною. Тому, для отримання максимізації відстані, ваги w і b налаштовуються відповідно:

$$w \cdot x - b = 1, \tag{1.7}$$

де все, що знаходиться на або вище поділки відноситься до класу 1, а також

$$w \cdot x - b = -1, \tag{1.8}$$

в такому випадку все, що знаходиться на або нижче поділки відноситься до класу -1.

Також, об’єкти які лежать на певній гіперплощині, або найближче до неї розташовані, називаються опорними векторами, саме від яких пішла назва алгоритму. Графічний приклад роботи методу опорних векторів представлений на рисунку 1.5.2.2

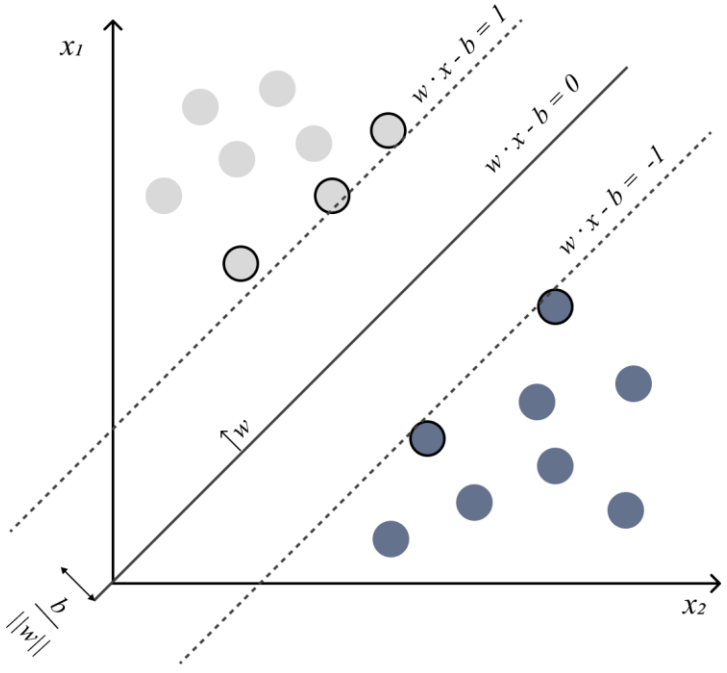


Рисунок 1.3 – Метод опорних векторів. Процес максимізації відстані між гіперплощиною і об’єктами, що розподілені на два класи

1.5.3 Класифікатор Байєса

Наївний класифікатор Байєса (англ. Naive Bayes classifier) є досить простим і широко розповсюдженим сімейством алгоритмів машинного навчання з вчителем для виконання задачі класифікації. Дані алгоритми засновані на теоремі Байеса з чітким (наївним) представленням незалежних припущень. Через свою простоту, алгоритми наївного Байєса є дуже швидкими та легко розширювальними у використанні. Всі алгоритми наївної байєсівської класифікації беруть на себе припущення, що значення кожної особливості об'єкту є строго незалежним від іншої особливості того ж самого об'єкту. Для прикладу можна представити абстрактний відгук на кімнату готелю: «Номер чистий і світлий, хоч і досить невеликий.», де ознаки «чистий», «світлий» та «невеликий» не залежать одна від одної за припущенням, що лежить в основі всього сімейства байесівських класифікаторів.

Вірогідність за теоремою Байєса визначається наступною формулою (на прикладі покупки товарів):

$$P\left(\frac{A}{B}\right) = P\left(\frac{B}{A}\right) \cdot \frac{P(A)}{P(B)}, \tag{1.9}$$

Де $P\left(\frac{A}{B}\right)$ – вірогідність придбання товару А після покупки товару Б;

$P\left(\frac{B}{A}\right)$ – ймовірність покупки товару Б після придбання товару А;

$P(A)$ – вірогідність придбання товару А;

$P(B)$ – ймовірність покупки товару Б.

Розрахування вірогідності зустрічі певного слова в документі відбувається за формулою знаходження середнього арифметичного значення кожного слова w відносно наданого класу c:

$$P(w|c) = \frac{\text{слово}_{wc}}{\text{слово}_c} \qquad (1.10)$$

Слід зазначити, що трапляються випадки, коли певні слова не зустрічаються в окремих відгуках, тобто кількість появ n-го слова дорівнює нулю. Такі випадки можуть негативним чином впливати на результати аналізу, а іноді навіть спотворювати їх. Тому доцільно до попередньої формули використати згладжування Лапласа – коефіцієнт α, що набуває вкрай невеликих значень, завдяки якому невживані слова не будуть засмічувати навчальну вибірку.

$$P(w|c) = \frac{\text{слово}_{wc} + \alpha}{\text{слово}_c + |v| + 1};\ \alpha = 0{,}001 \qquad (1.11)$$

В зазначеній вище формулі значення v представляє масив всіх слів, що містяться в складеному словнику [8].

Гаусівський класифікатор наївного Байєса (англ. Gaussian naïve Bayes) прийнято використовувати у випадках, коли існує скінченна множина прогнозуємих значень особливостей. Також даний метод припускає, що всі зазначені особливості розміщаються відповідно до розподілу Гауса, що також відомий як нормальний розподіл. Якщо існує припущення, що ймовірність певної особливості є гаусовою, то в такому випадку формула вірогідності набуває наступного значення:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{\left(x_i - \mu_y\right)^2}{2\sigma_y^2}\right). \qquad (1.12)$$

Поліноміальний або мультиноміальний наївний байєсівський класифікатор (англ. Multinomial naïve Bayes), як правило, в більшості випадків прийнято використовувати при роботі з класифікацією тексту – коли певні особливості відповідають частоті вживання слів у документі або їх кількості. В даному

випадку слова у векторному вигляді представлені як згенерована мультиноміальним розподілом частота конкретних подій.

Для того, щоб отримати коректні результати, слід уникати «недопереповнення» (англ. underflow), а також надати стоп-словам нульового значення для згладжування результатц. З урахуванням цих критеріїв, формула, що описує даний алгоритм, має наступне представлення:

$$\Pr(c) = \log \pi_c + \prod_{w=1}^{|v|} f_w \log(t_w \Pr(w|c)). \qquad (1.13)$$

Особдивість методу Бернуллі полягає у представлені вхідних даних, що подаються як певні особливості, у вигляді незалежних булевих функцій (тобто бінарних змінних). Дана модель також часто використовується для обробки тексту, однак в даному випадку замість частотних виразів використовуються двійкові. Булевий вираз можна позначити як x_i, а словник як v, і вивести наступний вираз [9]:

$$\mathrm{P}(x_i|c_i) = \prod_{t=1}^{|v|}(B_{it}P(w_t|c_i) + (1 - B_{it})(1 - P(w_t|c_i))), \qquad (1.14)$$

де вимір t вектору для документу d записаний як B_{it}

1.5.4 Латентно-семантичний аналіз ймовірності

Латентно-семантичний аналіз ймовірності або pLSA (англ. Probabilistic latent semantic analysis) – метод аналізу данних способом кореляції декількох видів даних, який використовується в машинному навчанні. Стосовно аналізу тональності тексту прийнято використовувати розширення алгоритму, а саме S-PLSA (англ. Sentiment probabilistic latent semantic analysis), який відрізняється саме тим, що фокусується на тональній складовій, а не на абстрактному критерію.

Оскільки даний алгоритм є статистичним, то, на основі аналізу сентименту певних текстових даних, стає можливим прогнозування майбутніх продаж товару [10]. Однак, S-PLSA не працює з BOW-моделлю.

Висновки до розділу

В даному розділі було проведено аналіз теоретичних відомостей про обрану для розгляду тему, після чого було визначено напрямок дослідження. Спираючись на проведений аналіз лінгвістичних особливостей української мови, були визначені певні мовні засади, що можуть виникнути при подальшому дослідженні україномовних відгуків, а саме: вживання неологізмів та діалектизмів, а також наявність «несправжніх» відгуків, які, відповідно, не несуть інформаційного сенсу для тонального аналізу.

Був проведений аналіз існуючих рішень щодо класифікації та розрізнення емоційних станів, що можуть бути виражені у відгуках. Було розглянуто різноманітні підходи та методи класифікації тексту в тональному аналізі, на основі чого було обрано модель представлення тексту як «торбу слів».

Після проведення детального аналізу роботи різноманітних класифікаторів, було визначено, що з обраною моделлю найбільш доцільно використовувати метод опорних векторів і алгоритм класифікації наївного Байєса, а саме його поліноміальний різновид.

2 АНАЛІЗ ІСНУЮЧИХ РІШЕНЬ

Наразі на ринку IT-послуг вже існує велика кількість способів та застосунки для проведення аналізу тональності тексту – існують навіть інструменти з відкритим кодом, а деякі проводять аналіз не лише сентименту, а й тону в цілому. Більшість застосунків головним чином заточені під аналіз англомовних текстів, оскільки англійська мова є однією з найпоширеніших в усьому світі. Іноді трапляються застосунки для роботи з текстом на інших, також широко розповсюджених мовах, втім для груп мов, що засновані на кирилиці, тема тонального аналізу є недостатньо оглянутою.

На даному етапі слід відокремити поняття аналізу тональності (англ. sentiment analysis) та аналізу тону (англ. tone analysis), адже через особливості перекладу дані поняття нерідко можна сплутати. В сентимент-аналізі закладена ідея бінарної класифікації оброблювальних текстових даних – «позитивна» та «негативна», в той час як аналіз тону охоплює більш широкий спектр емоцій.

Для детального аналізу були обрані такі інструменти сентимент-аналізу як «Grammarly Tone Detector», «MonkeyLearn Sentiment Analyzer», «IBM Watson Tone Analyzer», а також було окремо розглянуто систему рейтингу товарів. Для тестування перелічених інструментів було обрано реальні відгуки різних, незалежних один від одного, користувачів зі збереженням орфографії та пунктуації, вживаних автором.

2.1 Система рейтингу товарів

Наразі на більшості ресурсів, які передбачають можливість написання відгуків на певний продукт або послугу, надається можливість додатково поставити оцінку товару. Зазвичай даний інструмент представлений у вигляді шкали від мінімальної до максимальної оцінки товару – зазвичай від 1 до 5, але в

деяких випадках трапляються різні варіації. Приклад подібних систем представлений на рисунку 2.1.1.

Рисунок 2.1 – Системи рейтингу товару, впроваджені на різних сайтах

Подібного роду системи не мають нічого спільного з аналізом сентименту або тону, адже спосіб їх реалізації а також закладена мета несуть абсолютно інший сенс. В більшості випадків, окрім можливості додати оцінку до свого відгука, користувачу може поставити її окремо, якщо немає бажання або можливості описати свої враження текстовим чином. В такому випадку для тонального аналізу подібні речі не несуть ніякого сенсу.

Однак, слід зазначити, що, з іншого боку, комбінація оцінки та коментаря можуть значно полегшити процес збору емоційно забарвлених відгуків для формування бази даних для навчальної вибірки, оскільки коментарі вже певним чином поділені на «позитивні» та «негативні» за вбудованою на сайті шкалою. Детально цей процес буде розглянутий в наступних розділах, коли буде сформовано навчальний словник.

2.2 Grammarly Tone Detector

Застосунок Grammarly [11], створений трьома українськими розробники, головним чином призначений, згідно своєї назви, для перевірки правиотності написаних англомовних текстів – а саме перевірки з точки зору граматики, коректності вживання деяких слів, пунктуації, орфографії тощо. Основна задача програми – допомогти користувачу написати чіткий, цікавий і грамотний текст відповідно до складених обставин: твір, робочий лист, резюме та багато інших задач. Система дозволяє аналізувати текст не тільки всередині застосунку, а й в багатьох інших програмах – браузерах, месенджерах та текстових редакторах.

В 2021 році компанія представила інструмент для аналізу тону – Grammarly Tone Analyzer, який дає можливість користувачу відстежити які саме емоції вкиликає написаний ним текст для того, щоб бути впевненим, що одержувач зрозуміє повідомлення правильним чином.

Спектр емоцій, які визначає застосунок наразі є дуже широким і включає в себе як основні емоції, такі як задоволення, смуток і здивування, так і більш глибокі та складні для ідентифікації, а саме: наполегливість, захоплення, звинувачування та інші. Повний список емоцій представлений на рисунку 2.1.1.

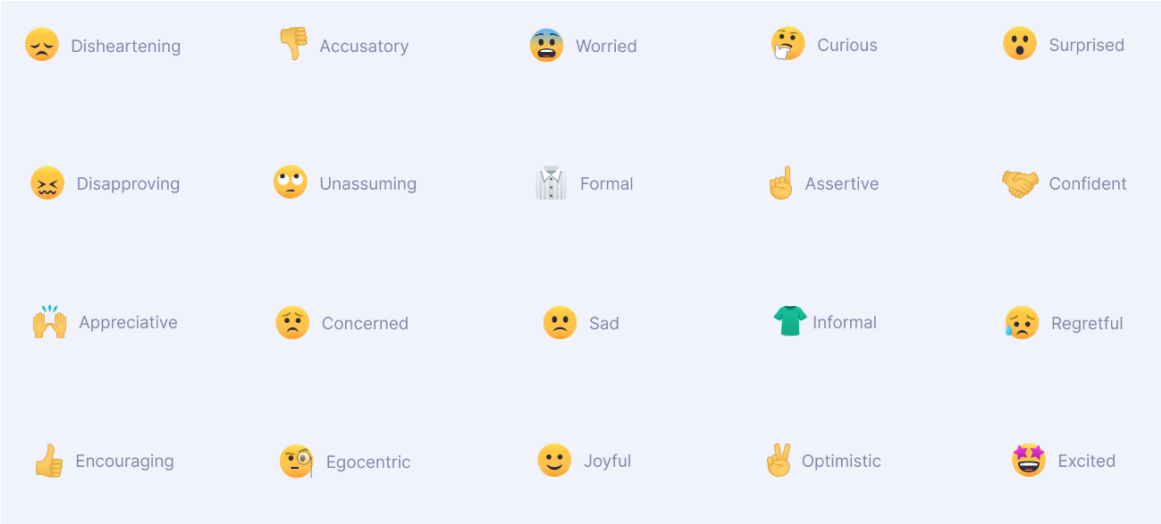


Рисунок 2.2 – Повний емотивний перелік застосунку Grammarly Tone Detector, який становить 20 емоційних станів

Даний інструмент є вбудованим до основного функціоналу програми, тому, як зазначалося раніше, аналіз тону можливий при написанні тексту будь-якому сайті та в більшості застосунків, які передбачають набір тексту.

Оскільки застосунок в першу чергу створений для роботи з власноруч написаними текстами, для подальшого дослідження якості отриманих результатів тональності, вже існуючі відгуки будуть взяті з деяких англомовних сайтів.

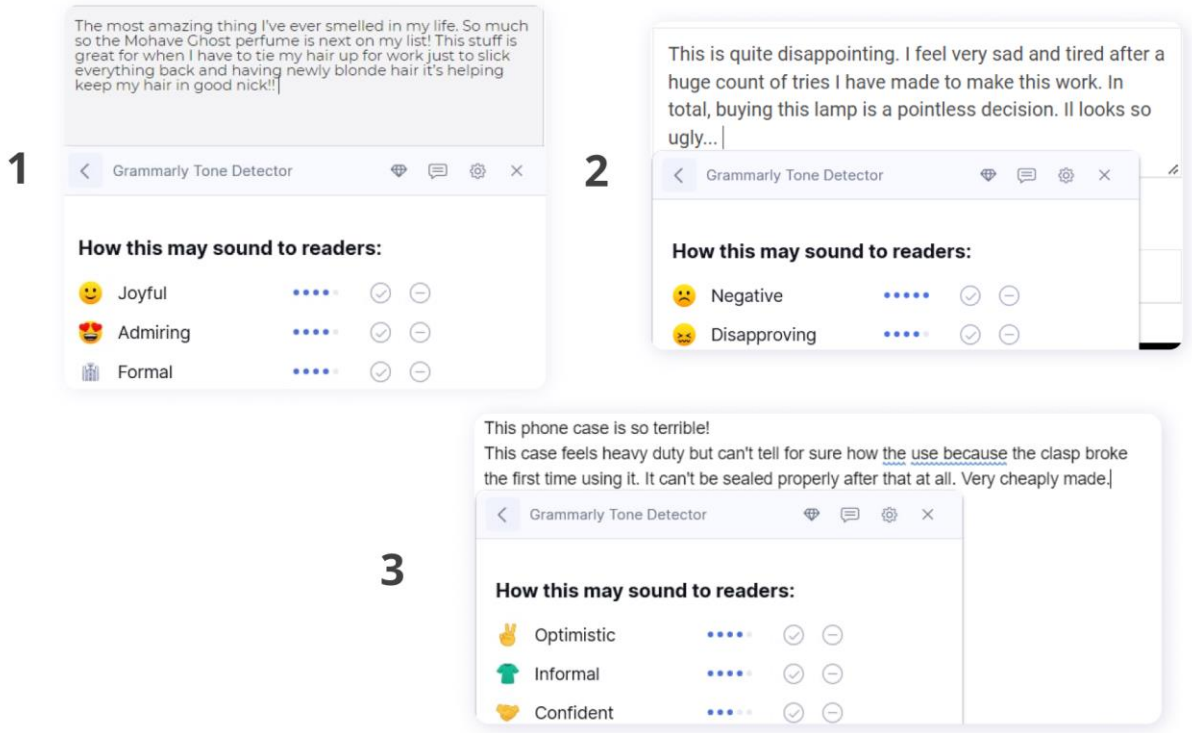


Рисунок 2.3 – Деякі відгуки, проаналізовані інструментом Grammarly Tone Detector. В першому відгуку йде мова про позитивні враження від продукту, а в інших двох – про негативні

В першому відгуку йде мова про цілковите задоволення продуктом – користувач, що написав відгук знаходиться у повному захваті. Програма визначила в даному коментарі висловлені емоції задоволення та захвату, що виражені формальним чином. Дане визначення є цілком точним та повним.

Другий коментар виражає повне незадоволення і розчарування товаром, яке підкреслене декілька разів. В даному випадку програма також дала правильне визначення – відгук є негативним і в ньому транслюється розчарування.

В третьому випадку незадоволення не є настільки вираженим, тому програма ідентифікувала висловлені емоції дуже розпливчато – оптимістично, впевнено та інформативно. Останні два емоційних забарвлення можливо ототожнювати з відгуком, однак коментар є негативним, тому оптимістичний настрій в цьому випадку визначений неправильно.

Після детальної перевірки багатьох відгуків текстів було визначено, що в більшості випадків застосунок правильно визначає емоційне забарвлення коментарів, однак якщо настрій написаного відгука не прослідковується досить очевидним чином, можуть виникнути проблеми з ідентифікацією тону. Вочевидь, так трапляється через те, що, по-перше, відгуки не завжди написані граматично правильно і з урахуванням правил щодо структури формулювання, а, по-друге, задача застосунку орієнтована на визначення настрою іншого користувача після прочитання тексту, а не думки клієнта щодо товару.

2.3 MonkeyLearn Sentiment Analyzer

Розглянемо детальніше сервіс MonkeyLearn [12] – це платформа без коду для різноманітних способів обробки текстових даних. Застосунок дозволяє бізнесу проводити аналіз цілого спектру важливих показників: ступіть задоволення покупців, коректність роботи служби підтримки, настрої у відгуках. Сервіс проводить детальну аналітику по багатьом критеріям, серед яких не останнє місце займає тональний аналіз, що розмежовує категорії відгуків на різні сфери.

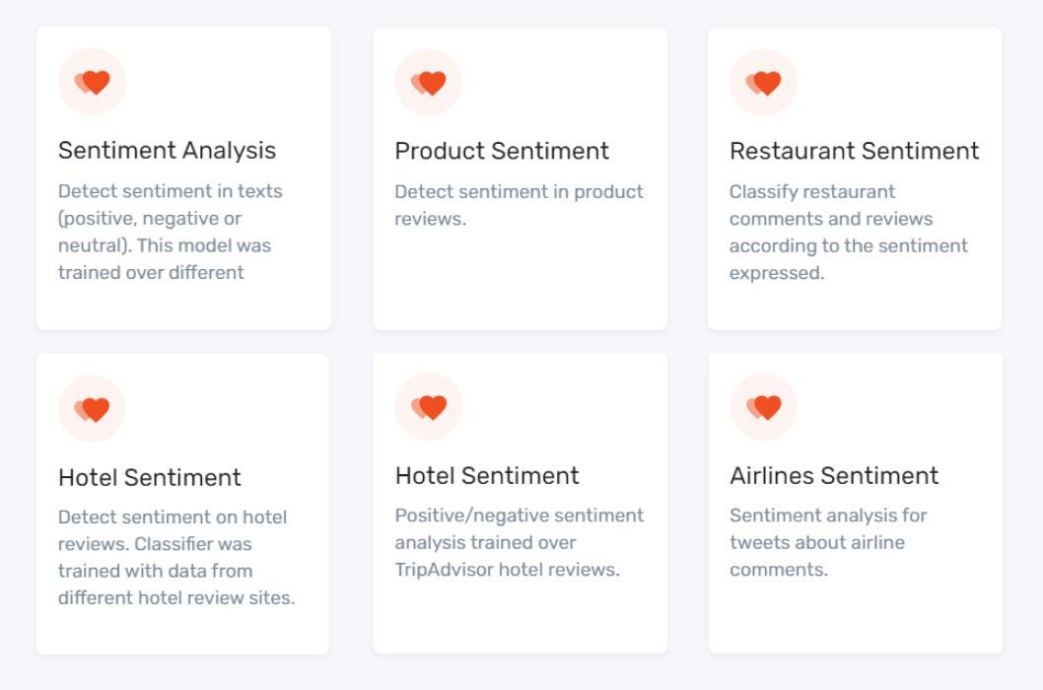


Рисунок 2.4 – Повний список предметних областей для тонального аналізу

На рисунку 2.4 зображені всі сфери, що доступні для сентимент аналізу. Серед них: загальний тональний аналіз; аналіз тональності відгуків на товари; аналіз відгуків на авіакомпанії, заснований на коментарях з твіттеру; тональний аналіз коментарів та відгуків на заклади харчування. Також є два способи аналізу тональності відгуків на готелі: один на основі коментарів, взятих з всесвітньо відомого сайта про відпочинок TripAdvisor; а інший базується на навчальній вибірці, що сформована з відгуків, які взяті з багатьох різних сайтів.

Для проведення аналізу будуть взяті інструменти для аналізу відгуків на товари, заклади харчування і готелі на базі даних декількох сайтів.

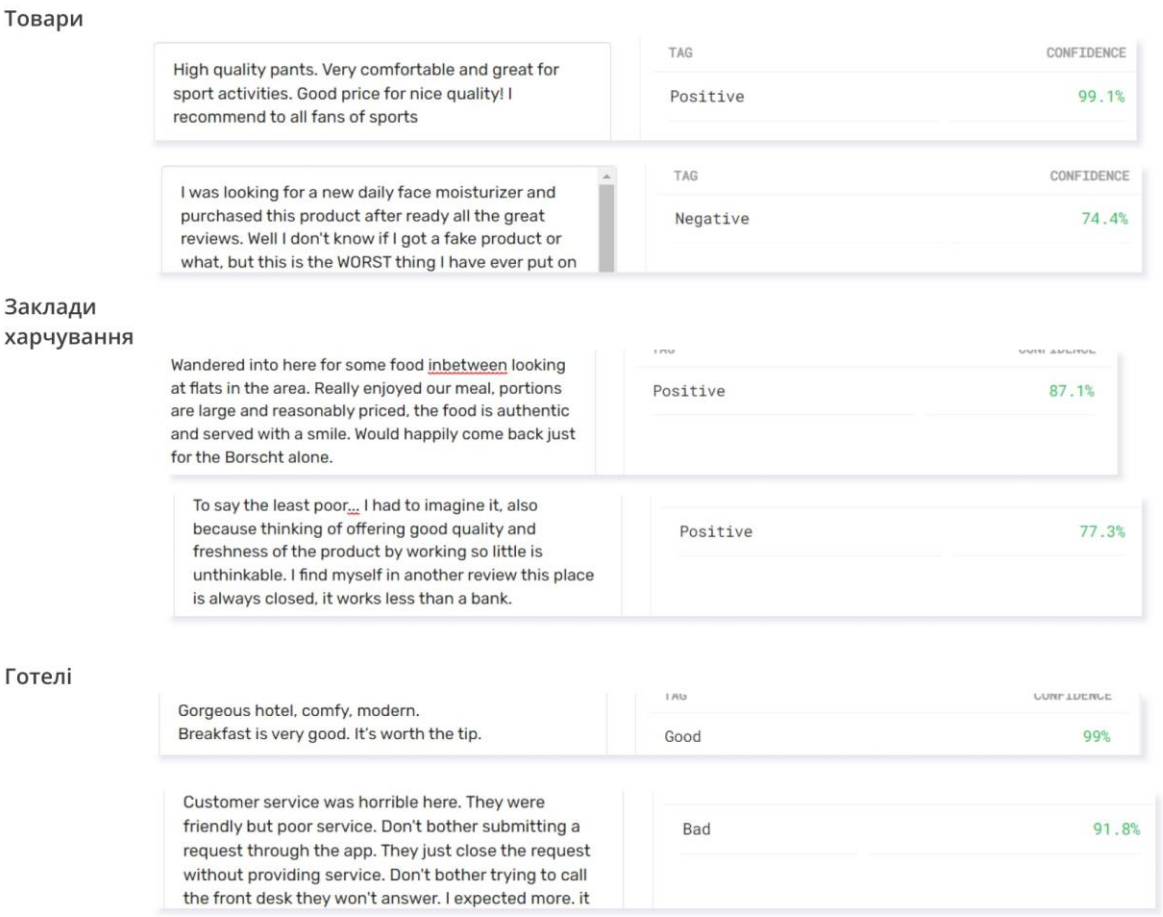


Рисунок 2.5 – Результати аналізу тональності відгуків різними інструментами сервісу MonkeyLearn

Частину результатів після проведення аналізу тональності зображено на рисунку 2.5. Оскільки сервіс передбачає саме сентимент-аналіз, то і визначення емотивної складової є бінарним – «позитивним» або «негативним», де додатково зображена ступіть вираження сентименту у відсотковому вигляді. В більшості випадків, тональність відгука була визначена досить точно, втім, інколи трапляються помилкові результати. Прикладом є другий відгук на заклад харчування, зображений на рисунку 2.5, який є цілком негативним і оригінальному вигляді містить оцінку 1, однак програма визначила його на 77.3% позитивним.

Також даний сервіс надає можливість використовувати власну тренувальну вибірку, але ця функція доступна лише для англомовних БД.

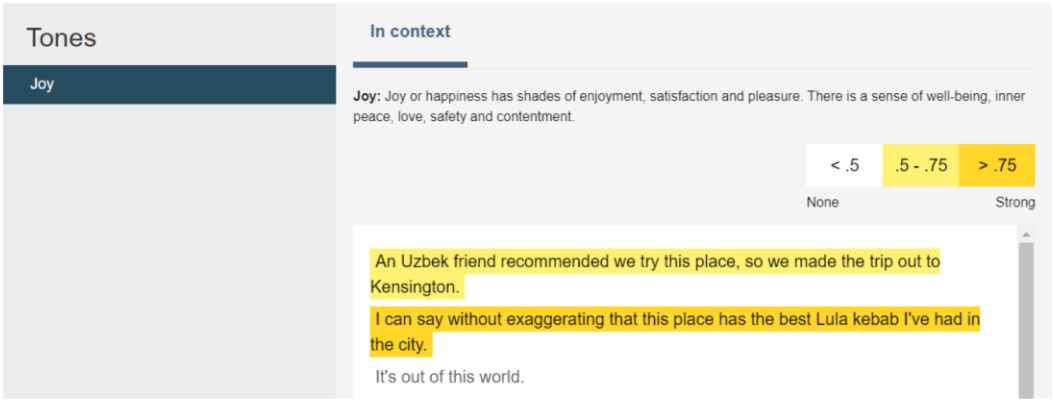
2.4 IBM Watson Tone Analyzer

Наступним досліджуваним інструментом буде сервіс Watson Tone Analyzer [13], який розробила компанія IBM. За допомогою даного застосунку також можливо здійснити аналіз тону текстів, твітів, електронних листів, а також відгуків – вони винесені в окрему категорію, і, окрім аналізу англомовних відгуків, доступна функція обробки відгуків, що написані французькою мовою. В системі представлені сім емоцій, які досить сильно розрізнені за забарвленням: з них три очевидно негативні (страх, сум і злість), одна позитивна (задоволення), і три окремі емоції, які однаково поєднуються з попередніми – впевненість, сумнів та аналітичність.

Для аналізу були взяті відгуки на товари, а також на заклади харчування з різноманітних англомовних сайтів. Отримані результати можна визначити як досить неоднозначні через те, що в бінарно визначених коментарях «позитивних» та «негативних» можна отримати визначення досить широкого спектру емоцій, в рамках тих, що визначає система. Під час проведення аналізу обраних відгуків нерідко траплялись випадки, коли система визначала протилежно-забарвлені емотивні стани у тексті. Причиною є особливий метод, що реалізований у застосунку для визначення тону поданих текстів та відгуків.

Приклади отриманих результатів зображені на рисунку 2.6, де в першому випадку програма визначила емоцію задоволення з різним ступенем інтенсивності в позитивному коментарі: а у другому, негативному – сум, впевненість, сумління та аналітичність, які теж шкалюються за визначеною в програмі інтенсивністю. За результатами, які містять декілька одночасно виражених емоцій не завжди можливо ідентифікувати загальний настрій коментатора щодо описаного товару, тому потрібно додатково аналізувати загальний контекст коментарю – даний процес зазвичай вимагає додаткового людського ресурсу.

1



2

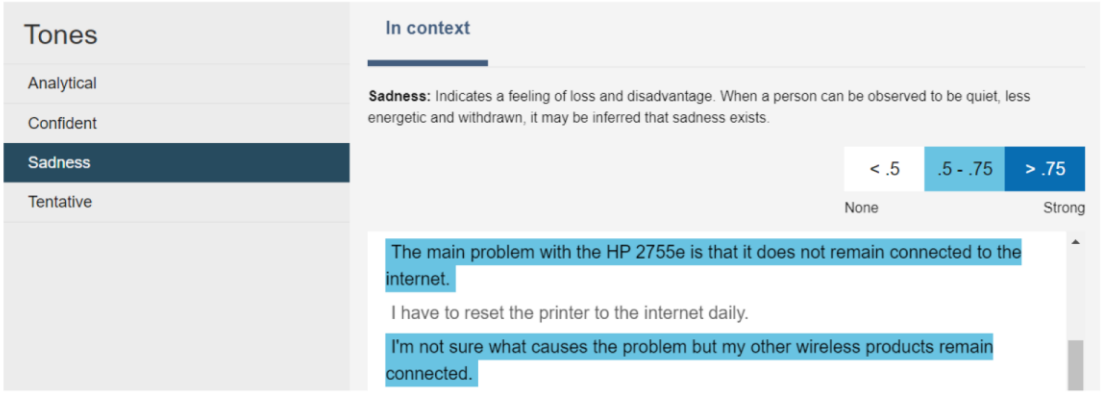


Рисунок 2.6 – Результати аналізу тону, проведеного застосунком Watson Tone Analyzer

Дана система демонструє унікальний підхід, який цілком відрізняється від рішень, реалізованих в попередніх застосунках. Виведені результати відображають окремі речення, в яких присутнє вираження певного емоційного стану. Також відображається більш детальне тлумачення та ступінь вираження обраної емоції за системою шкалювання: від відсутнього емотивного забарвлення до дуже сильного.

Через те, що застосунок окремо аналізує кожне речення на емотивну складову, трапляються інциденти, коли в одному коментарі трапляються цілком протилежні емоції. Подібний випадок продемонстрований на рисунку 2.4.2, де зображені результати проведення тонального аналізу завідомо негативного коментаря.

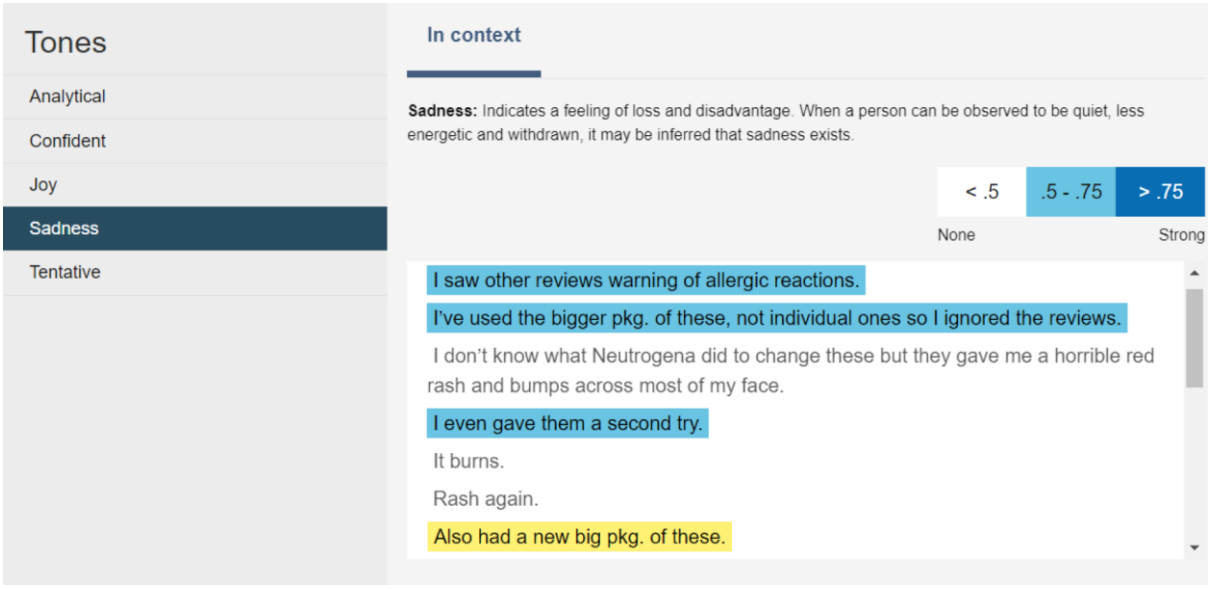


Рисунок 2.7 – Результати аналізу тону негативного відгуку, де виявлене вираження протилежно забарвлених емоцій: задоволення і смутку

Застосунок досить чітко і дуже детально визначає висловлені емоції, що містяться в певних реченнях – хибних результатів не спостерігалося. Однак, визначити суцільний настрій відгуку через дану особливість не є можливим.

Для того, щоб сформулювати кінцеві висновки про кожен з розглянутих у попередніх підрозділах застосунків (крім системи рейтингу відгуків), було складено таблицю 2.1. з порівняльною характеристикою.

Таблиця 2.1. – Порівняльна характеристика інструментів тонального аналізу

Назва системи	Grammarly Tone Detector	MonkeyLearn Sentiment Analyzer	IBM Watson Tone Analyzer
Спектр емоцій	Застосунок дозволяє визначити 20 емоційних станів	Система визначає лише позитивну або негативну складову відгуку	Аналіз тексту передбачає вилучення 7 емоційних станів

Робота з відгуками	Аналіз доступний лише у представленні відгука як звичайний текст	Система спеціалізована на роботі з відгуками	Можливий аналіз на двух мовах
Глибина аналізу	В більшості випадків визначається декілька емоцій, є шкала яка показує силу емоції, що виражена у тексті	Результати аналізу представлені у відсотковому вигляді, що демонструє інтенсивність тональної складової	Система проводить окремий аналіз кожного речення, після чого визначає його емоційне забарвлення
Частота хибних результатів	Трапляються у межах норми	Трапляються у межах норми	Не спостерігались через особливий підхід програми
Доступність застосунку	Передбачає преміум-підписку, однак аналіз тону є вбудованим до стандартного функціоналу інструментом	Надається пробний період, який передбачає аналіз першої 1000 відгуків безкоштовний	Застосунок є повністю безкоштовним і з відкритим кодом
Доступні мови	Англійська	Англійська	Англійська; Французька для відгуків

Всі розглянуті інструменти час від часу можуть видавати спірні та неоднозначні результати аналізу – таке відбувається через динамічний аспект мови (тобто вираження думок користувачем не завжди відповідає прийнятим нормам, також можуть буди граматичні помилки) і статичну сформовану базу даних. Таке явище є цілковитою нормою.

Висновки до розділу

Кожна з вищерозглянутих систем демонструє свій унікальний підхід до вирішення задачі з визначення тонального аналізу коментарів. Система Grammarly Tone Detector заснована на ідеї покращення структури написаних користувачем текстів, тому в результатах аналізу тону зображені емоції, що можуть виникнути у читача після прочитання тексту. З такої точки зору, емоційна складова є достатньо вірно визначеною, однак подібний підхід не є оптимальним для аналізу відгуків, через вірогідність хибного результату саме в контексті думки про товар або послугу. Інструмент IBM Watson Tone Analyzer дуже цікавим чином працює як і з відгуком, так і з текстом в цілому, окремо визначаючи емоційну складову кожного речення. Результати є досить чіткими і правильними у більшості випадків. Система MonkeyLearn Sentiment Analyzer спеціалізується на взаємодії з потенціальними клієнтами бізнесу, тому аналізу відгуків приділяється достатньо уваги – це можна спостерігати за кількістю різних інструментів для їх аналізу. Даний застосунок демонструє в більшості гарні і достовірні результати.

Всі розглянуті інструменти час від часу можуть видавати спірні та неоднозначні результати аналізу – таке відбувається через динамічний аспект мови (тобто вираження думок користувачем не завжди відповідає прийнятим нормам, також можуть буди граматичні помилки) і статичну сформовану базу даних. Таке явище є цілковитою нормою.

Доцільними будуть висновки щодо глибини емотивних станів, що будуть доступні у розроблювальній системи. При ознайомленні з відгуками на певний

товар, користувача цікавить питання чи коштує цей товар уваги та покупки. Даний критерій визначається відношенням позитивних та негативних відгуків. Для бізнесу також важливо відслідковувати настрої щодо своїх продуктів, що здійснюється за таким самим принципом. Тому, проведення аналізу тону з ідентифікацією тону не є доречним для відгуків (окрім специфічних випадків таких як рецензія на фільм) – це може створити додатковий інформаційний шум і плутанину. В даному випадку більш раціональною буде ідея реалізації тонального аналізу.

Важливо зауважити, що попри всіх переваг цих застосунків, є дуже важлива особливість – це відсутність можливості аналізувати тексти, що написані мовами кириличного алфавіту, в тому числі українською.

Система рейтингу коментарів не має відношення до тонального аналізу, але завдяки існуючим коментарям з оцінкою було обрано відгуки для аналізу, а також буде сформовано навчальну вибірку у наступному розділі.

3 ПОПЕРЕДНЯ ПІДГОТОВКА ТЕКСТОВИХ ДАНИХ

3.1 Формування бази даних

Формування бази даних є однією з найважливіших задач при розроблюванні системи для сентимент-аналізу, адже його результати залежать безпосередньо від того, наскільки докладно і належно була складена вибірка. Тому дане питання повинне підлягати кропіткому та детальному вирішенню

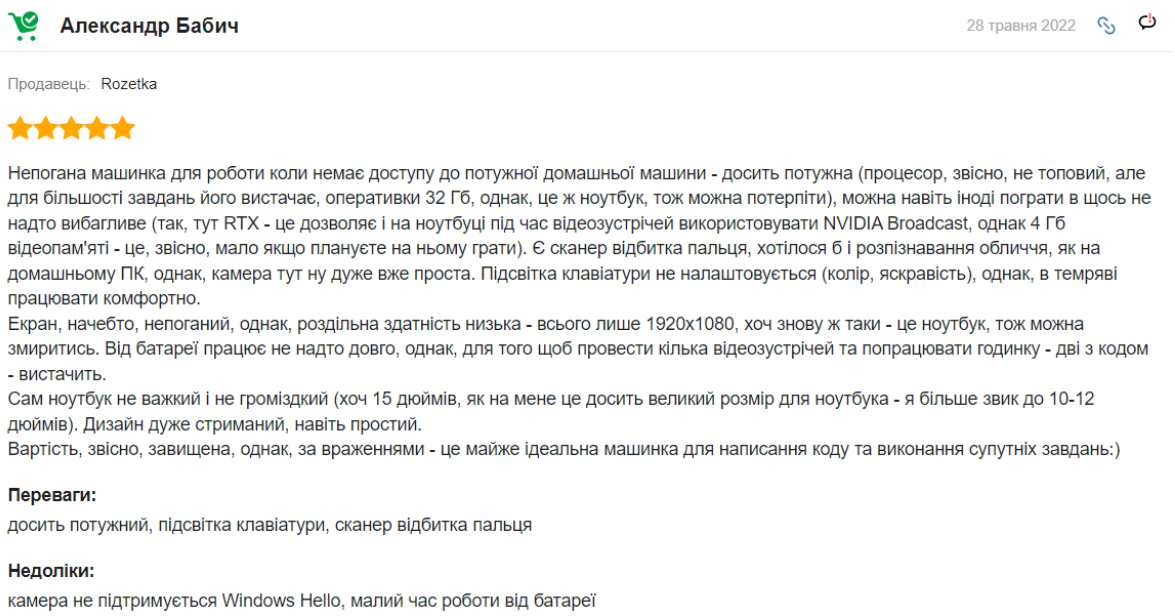
У ході роботи було вирішено брати відгуки для формування бази даних з сайту українського торгівельного майданчику Rozetka[14]. Раніше даний онлайн-магазин спеціалізувався виключно на продажі різної техніки – як і побутової, так і малогабаритної. Однак за останні роки сайт значно розширив свій асортимент, і наразі для покупки доступний дуже широкий вибір товарів починаючи зі звичних для кожної людини речей (продукти харчування або одяг) і закінчуючи товарами для специфічних галузей (наприклад, деталі для автомобільної промисловості або професійні фарби для митців). Завдяки такому величезному асортименту продуктів стає можливим створити досить універсальну вибірку відгуків для навчання та тестування нейронної мережі. Крім того, даний інтернет-ресурс є дуже відомим серед українців, тому на неймовірну кількість товарів вже написано сотні тисяч відгуків реальними користувачами.

На сайті також присутня система рейтингу товарів, яка більш детально була розглянута у розділі 2.1. При написанні відгука надається можливість поставити оцінку за шкалою від 1 до 5. Загальна оцінка товару визначається середнім арифметичним від кількості всіх відгуків з точністю до десятих. Більш детальне тлумачення оцінок зображене на таблиці 3.1.

Таблиця 3.1 – Система оцінки товару на сайті Rozetka

Кількість балів	Значення оцінки
★☆☆☆☆	Поганий
★★☆☆☆	Так собі
★★★☆☆	Нормальний
★★★★☆	Добрий
★★★★★	Чудовий

Також сайт надає можливість окремо продублювати переваги та недоліки товару для підвищення точності рейтингу товару. Приклад наповненого відгуку представлено на рисунку 3.1.



Александр Бабич 28 травня 2022

Продавець: Rozetka

★★★★★

Непогана машинка для роботи коли немає доступу до потужної домашньої машини - досить потужна (процесор, звісно, не топовий, але для більшості завдань його вистачає, оперативки 32 Гб, однак, це ж ноутбук, тож можна потерпіти), можна навіть іноді пограти в щось не надто вибагливе (так, тут RTX - це дозволяє і на ноутбуці під час відеозустрічей використовувати NVIDIA Broadcast, однак 4 Гб відеопам'яті - це, звісно, мало якщо плануєте на ньому грати). Є сканер відбитка пальця, хотілося б і розпізнавання обличчя, як на домашньому ПК, однак, камера тут ну дуже вже проста. Підсвітка клавіатури не налаштовується (колір, яскравість), однак, в темряві працювати комфортно.
Екран, начебто, непоганий, однак, роздільна здатність низька - всього лише 1920x1080, хоч знову ж таки - це ноутбук, тож можна змиритись. Від батареї працює не надто довго, однак, для того щоб провести кілька відеозустрічей та попрацювати годинку - дві з кодом - вистачить.
Сам ноутбук не важкий і не громіздкий (хоч 15 дюймів, як на мене це досить великий розмір для ноутбука - я більше звик до 10-12 дюймів). Дизайн дуже стриманий, навіть простий.
Вартість, звісно, завищена, однак, за враженнями - це майже ідеальна машинка для написання коду та виконання супутніх завдань:)

Переваги:

досить потужний, підсвітка клавіатури, сканер відбитка пальця

Недоліки:

камера не підтримується Windows Hello, малий час роботи від батареї

Рисунок 3.1 – Приклад детального відгуку з заповненими параметрами

Для наповнення вибірки «позитивними» коментарями будуть вибрані відгуки з оцінкою 4-5, а для «негативної» частини будуть відібрані коментарі з рейтингом 1-2. Середня оцінка становить 3 бали і містить значення «нормальний», тобто ні хороший, ні поганий. Такі відгуки будуть мати маркер «нейтральний». Також, час від часу трапляються поодинокі випадки, коли коментарі з подібним змістом не мають оцінки. Вони також будуть віднесені до нейтральних відгуків.

Прикладом гарно складеної бази для тонального аналізу є вибірка від компанії Amazon – одного з найбільших торгівельних майданчиків світового рівня. Вибірки від Amazon сформовані дуже кропітким чином, і, безумовно, є найвдалішим прикладом для подальшого формування інших баз даних з відгуками. На кожну категорію складена окрема база даних, яка містить від 10,261 до 8,898,041 відгуків [15].

reviewerID	asin	reviewerName	helpful/0	helpful/1	reviewText	overall	summary	unixReviewTime	reviewTime
A2IBPI20UZIR0U	1384719342	cassandra tu "'	0	0	Not much to write about here,	5	good	1393545600	02 28, 2014
A14VAT5EAX3D9S	1384719342	Jake	13	14	The product does exactly as it	5	Jake	1363392000	03 16, 2013
A195EZSQDW3E21	1384719342	Rick Bennette '	1	1	The primary job of this device	5	It Does The Job W	1377648000	08 28, 2013
A2C00NNG1ZQQG2	1384719342	RustyBill "Sunc	0	0	Nice windscreen protects my M	5	GOOD WINDSCREE	1392336000	02 14, 2014
A94QU4C90B1AX	1384719342	SEAN MASLANI	0	0	This pop filter is great. It looks	5	No more pops wh	1392940800	02 21, 2014
A2A039TZMZHH9Y	B00004Y2UT	Bill Lewey "ble	0	0	So good that I bought another	5	The Best Cable	1356048000	12 21, 2012
A1UPZM995ZAH90	B00004Y2UT	Brian	0	0	I have used monster cables for	5	Monster Standard	1390089600	01 19, 2014
AJNFQI3YR6XJ5	B00004Y2UT	Fender Guy "Ri	0	0	I now use this cable to run fro	3	Didn't fit my 1996	1353024000	11 16, 2012
A3M1PLEYNDEYO8	B00004Y2UT	G. Thomas "To	0	0	Perfect for my Epiphone Shera	5	Great cable	1215302400	07 6, 2008
AMNTZU1YQN1TH	B00004Y2UT	Kurt Robair	0	0	Monster makes the best cable	5	Best Instrument C	1389139200	01 8, 2014
A2NYK9KWFMJV4Y	B00004Y2UT	Mike Tarrani "J	6	6	Monster makes a wide array o	5	One of the best in	1334793600	04 19, 2012
A35QFQI0M46LWO	B00005ML71	Christopher C	0	0	I got it to have it if I needed it.	4	It works great but	1398124800	04 22, 2014
A2NIT6BKW11XJQ	B00005ML71	Jai	0	0	If you are not use to using a la	3	HAS TO GET USE T	1384646400	11 17, 2013
A1C0O09LOLVI39	B00005ML71	Michael	0	0	I love it, I used this for my Yam	5	awesome	1371340800	06 16, 2013
A17SLR18TUMULM	B00005ML71	Straydogger	0	0	I bought this to use in my hom	5	It works!	1356912000	12 31, 2012

Рисунок 3.2 – Фрагменті вибірки відгуків на товари від Amazon з категорії «Музичні інструменти»

Всі вибірки від компанії сформовані виключно англійською мовою та подаються у форматі json. На рисунку 3.1.2 зображений фрагмент такої вибірки, що містить наступні поля:

– reviewerID – відображує ID коментаря;

– asin – ID товару;
– reviewerName – ім’я коментатора;
– helpful – оцінка корисності відгука;
– reviewText – текст відгука;
– overall – оцінка товару за системою рейтингу;
– summary – висновок з відгуку;
– unixReviewTime – час написання коментаря (за unix);
– reviewTime – дата написання коментаря (за ДД/ММ/РРРР);

Оскільки подібні вибірки використовуюся не тільки виключно для визначення тональності відгука, а й для збору даних про продукт, база даних має досить розширену форму представлення даних. У рамках даного дослідницького проєкту буде доцільним знехтувати деякими полями.

Розроблювана база даних буде сформована у форматі CSV (англ. Comma-Separated Values) – це формат текстових даних, що представляється у вигляді таблиці; та міститиме виключно україномовні відгуки. На сайті Rozetka дуже часто трапляються відгуки написані російською, тому, при умові недостатньої кількості відгуків для вибірки, допустимим буде переклад таких відгуків на українську мову зі збереженням особливостей написання, і подальшого занесення до вибірки.

Для більш зручної роботи з даними при складенні вибірки буде використана програма ModernCSV, яка виділяється оптимізованим під редагування інтерфейсом, а також швидкістю роботи порівняно зі звичайними текстовими редакторами. Інтерфейс програми буде продемонстрований на наступних рисунках.

rozetkaReviewDatabase.csv * ×

	0	1	2	3	4	5	6
0	reviewID	productCategory	authorName	reviewScore	reviewContent	sentimentScore	reviewDate

Рисунок 3.3 – Вибрані поля даних про відгук

Де reviewID – це ID відгука;

productCategory – це категорія товару, на який написаний вігук;

authorName – ім’я автору відгука;

reviewScore – оцінка товару, що визначив автор;

reviewContent – вміст відгука;

sentimentScore – оцінка тональності відгука;

reviewDate – дата написання відгука.

Поле «reviewID» містить ідентифікатор відгука, який відображається у індивідуальному посиланні на коментар.

Для заповнення поля «productCategory» були вибрані 10 категорій товарів зі списку представленого на сайті, серед них: ноутбуки та комп’ютери; смартфони, ТВ і електроніка; зоотовари; краса та здоров’я; алкогольні напої та продукти; товари для геймерів; спорт і захоплення; сантехніка та ремонт; товари для бізнесу та послуги; а також товари для дому. Така різноманітність категорій сприяє збільшенню критерію універсальності використання системи. Важливо зауважити, що «Алкогольні напої та продукти» – це назва категорії від сайту Rozetka, втім для вибірки будуть взяті відгуки на товари, які не підлягають акцизному маркуванню.

Поля «authorName», «reviewScore», «reviewContent», та «reviewDate» містять інформацію взяту безпосередньо з відгуку і не підлягають редагуванню.

Виключенням може бути переклад коментаря з російської мови на українську. Якщо у змістовному відгуку відсутня оцінка від автора коментаря, тоді значення «reviewScore» визначається на суб’єктивний розсуд.

Значення поля «sentimentScore» визначається на основі оцінки «reviewScore» та набуває значень {-1; 0; 1}, де -1 означає «негативний» відгук, 0 – «нейтральний» відгук, а 1 – це «позитивний» відгук.

Фрагмент утвореної вибірки зображений на рисунку 3.2.4, де продемонстровано по одному коментарю з кожної обраної категорії у якості прикладу.

reviewID	productCategory	authorName	reviewScore	reviewContent	sentimentScore	reviewDate
51664343	Комп'ютери та ноутбуки	Олександр Шкодич	5	Забрав сьогодні тільки налаштував, поки все супер. Через пару місяців напишем повторно. Тільки питання одне, невеличке, не можу підключити навушники хоча вони і справні	1	26.11.21
52574568	Смартфони, ТВ і електроніка	Пинкевич Денис	1	Пішли тріщини на дужці через погане качество пластмаса в користуванні були обережні як вернути	-1	28.03.22
52475694	Зоотовари	Світлана	4	Собаки їдять із задоволенням. Корм брали зі знижкою і це його відповідна ціна, на мою думку. За дорожче не брали б його.	1	15.02.22
51708557	Краса та здоров'я	Мирослава Зайцева	3	Туш жахлива. Осипається. Радує лише те,що купувала не за повну вартість.	0	30.11.21
52182569	Продукти харчування	Артем Логвиненко	1	Без смаку і слабкий аромат .для нормального смаку на 150г сиплю 5 чайних ложок	-1	15.01.22
52607886	Товари для геймерів	Олег	5	Отримав крісло, прийшло досить швидко. Легко збирається, в порівнянні зі старим, як жигуль і мерседес:) Недоліків поки що не виявлено, хіба не зовсім звично на такому сидіти, вже як виявляться якісь недоліки - обновлю відгук	1	14.04.22
47788810	Спорт і захоплення	Максим ААА	2	Замовив цей м'яч у розмірі 75 (маркування на упаковці відповідне). Надувши його, виявилося, що виробник "надув" і мене - м'яч виявився меншого розміру (віддав низькорослим лепреконам)Не певен, що це провина розетки, але не рекомендую цього виробника.	-1	08.10.20
52763934	Сантехніка і ремонт	Руслан Капитонов	5	Напевно, найкращий варіант для гідроізоляції ванної кімнати.	1	02.06.22
52706459	Товари для бізнесу та послуги	Виктор Радецький	5	Непоганий та недорогий сейф. Приємно здивувало те, що доставили на слідуючий день після замовлення. Чудово підходить для зберігання 1 одиниці зброї.	1	18.05.22
49045246	Товари для дому	Ірина Ковальова	2	Мені не сподобались. Тягнуться вони то добре, але не тримаються і сковзають. Розміри кришок не дуже зручні, їх ніби і багато, але підібрати під свої розміри тарілок важко, на одні налізає, але триється не герметично, для інших злазить.	-1	22.01.22

Рисунок 3.4 – Фрагмент утвореної бази відгуків

Готова вибірка становить 650 відгуків, де на кожну категорію було виділено по 65 коментарів. Тексти відбирались так, щоб кількість «позитивних» та «негативних» відгуків була приблизно однаковою. Через те, що «нейтральні» коментарі трапляються значно рідше, ніж тонально-забарвлені, їх відсоток у базі значно менший. Тому, кількість «позитивних» та «негативних» відгуків становить 570 штук з приблизно однаковим відношенням, в той час як кількість «нейтральних» коментарів становить 80.

3.2 Попередня обробка тексту

Для того, щоб нейронна мережа працювала правильним чином, потрібно провести приведення тексту до більш зрозумілого для неї вигляду. Різноманітні знаки пунктуації, дужки, цифри, та різниця регістрів у процесі навчання не несуть ніякого сенсу, через те що кожний символ має різне значення. Наприклад, слово «чудовий» написане з великої і маленької літери нейронна мережа буде

сприймати як абсолютно різні за значенням. Через це в процесі навчання формуються різноманітні «варіації» одного і того ж слова, через що дуже часто виникають хибні результати.

Перед першим етапом обробки тексту здійснюється видалення нейтральних відгуків – це здійснюється досить простим чином: виконується перевірка на значення тональної оцінки, і якщо вона дорівнює нулю – відгук видаляється.

3.2.1 Видалення пунктуаційних та розділових знаків

Головний сенс використання знаків пунктуації – це правильно сформувати думку і розставити сенсові наголоси в тексті. Відгуки, які мають лише суцільний текст, зазвичай є складними для сприйняття користувачами, і потребують значно більше часу для того, щоб зрозуміти сенс.

Згідно правил пунктуації, всі знаки розділяються пробілом лише після його вживання у реченні, тому слово, після якого стоїть кома, крапка або знак питання будуть визначатися системою як слова з різним значенням.

Вживання розділових знаків також майже у всіх випадках не мають ніякого впливу на тональну складову відгуку. Тому доцільним буде твердження, що всі перелічені знаки підлягають видаленню.

Єдиним виключенням є апостроф. Цей знак зустрічається однаково часто як в англійській мові, так і в українській. Однак, в першому випадку апостроф вживається лише для скорочення або для визначення приналежності. В українській мові апостроф є невід’ємною частиною великої кількості слів і його видалення не є доречним.

3.2.2 Видалення цифр

В деяких випадках використання цифр може сприяти на тональну складову тексту, але в процесі створення вибірки відгуків у розділі 3.1, було визначено, що майже всі випадки вживання цифр в будь-якій формі не несуть навантаження з

боку сенсу. Тому всі цифри будуть видалені, в тому числі і словесне їх представлення – воно буде занесено до складу стоп-слів в наступному розділі.

Слово «нуль» можна охарактеризувати як окремий випадок через те, що окрім числового значення, в багатьох мовах воно має тлумачення «ніякий», «пустий», «не містить нічого хорошого», що може виражати негативну складову.

3.2.3 Приведення літер до єдиного регістру

З правил орфографії відомо, що існує достатня кількість випадків, коли необхідно починати написання слів з великої літери: імена, назви міст, початок нового речення тощо. Деякі користувачі через наплив емоцій можуть зловживати написанням слів суцільним верхнім регістром. При машинному навчанні перелічені аспекти не визначають тональну складову, а лиш створюють безліч словоформ, що в масштабах навчальних вибірок може призвести до неочікуваних результатів.

Для запобігання подібних оказій достатньо привести всі слова до єдиного регістру, тобто провести всі літери верхнього регістру до нижнього.

3.2.4 Токенізація тексту

Процес сегментації, тобто розбиття тексту на окремі складові, називається токенізацією. Набір текстових даних можливо поділити на певні частини – речення або слова. Оскільки процес підготовки тексту до тонального аналізу у розроблювальній системи передбачає видалення знаків пунктуації та роздільників, відгуки будуть розбиті на окремі слова – «токени». Такий метод представлення тексту надає можливість нейронній мережі «зрозуміти», що реальна людина в обличчі користувача сприймає текст окремими фрагментами – словами, а не потоком літер.

В розроблювальній системі «токени» представлятимуть собою уніграми та біграми для випадків з часткою «не». Завдяки попередньому видаленню зайвих даних, всі сегментовані слова визначаються як змістовні.

Для більшого розуміння було сформовано таблицю 3, де зображено порівняння вхідного та вихідного відгуку зі сформованої вибірки (з урахуванням попередніх описаних процесів обробки).

Таблиця 3.2 – Порівняння відгуку та тексту, розбитого на «токени»

Вхідний текст	Вихідний текст
Гарні навушники,насиченість звуку добре.Заряд тримає чудово. З недоліків хочу підкреслити те,що іноді є перебої в звучанні треку. Це єдиний недолік.	‘гарні’ ‘навушники’ ‘насиченість’ ‘звуку’ ‘добре’ ‘заряд’ ‘тримає’ ‘чудово’ ‘з’ ‘недоліків’ ‘хочу’ ‘підкреслити’ ‘те’ ‘що’ ‘іноді’ ‘є’ ‘перебої’ ‘в’ ‘звучанні’ ‘треку’ ‘це’ ‘єдиний’ ‘недолік’

3.2.5 Обробка стоп-слів

Під поняттям «стоп-слова» в контексті сентимент-аналізу мається на увазі визначення слів, що не несуть ніякої тональної складової. Окрім слів, в яких відсутнє емотивне забарвлення, до стоп-слів прийнято відносити також службові частини мови: прийменник, частку та сполучник; а також вигук. Перелічені слова формують українську мову в звичному для користувача вигляді, адже якщо вилучити ці частини мови з повсякденного життя, люди перестануть розуміти один одного. Втім для нейронної мережі подібні слова лише сприяють швидкому засміченню навчального процесу – їх необхідно видаляти. До стоп-слів належать також чисельники і окремі літери.

Для розробки системи тонального аналізу відгуків буде взята найбільша вибірка україномовних стоп-слів Ukrainian Stopwords [16]. Вибірка представлена у вільному доступі на веб-сервісі GitHub і налічує 1983 стоп-слова, які розміщені у алфавітному порядку.

3.2.6 Лемматизація та стемінг

Для подальшого «очищення» навчальної вибірки від зайвих за змістом слів, як правило, потрібно привести розрізнені словоформи до єдиної «інфінітивної» – за це відповідають процеси лемматизації або стемінгу. Дані поняття є досить спорідненими за своїм сенсом, однак все ж мають деякі фундаментальні розрізнення. Обидва процеси передбачають виділення «початкової» форми слова, але вирішують дану задачу різними підходами.

Лемматизація, згідно своєї назви, перебачає процес вилучення зі слова «лемми» тобто початкової форми слова, що представляється у родовому відмінку однини або, у випадку з дієсловом, як інфінітив.

Стемінг приводить слово до початкової форми завдяки визначенню та подальшому видаленню закінчень зі слова.

Для підкреслення різниці між цими поняттями, доречно буде привести приклад з англійської мови, в якій присутні модальні форми дієслів: наприклад зв'язка слів «begin-began-began» має значення «почати» у різному періоді часу та мають різне написання і звучання, однак початкова форма слова (в даному прикладі «лемма») – це «begin». Саме до такої форми приводиться слово у процесі лемматизації. У випадку використання стемінгу видаляються закінчення і слово матиме форму «beg».

Оскільки в українській мові відсутнє поняття модальних дієслів, а також використання інших специфічних словоформ, що містять одне значення, але передбачають різне написання, є сенс пропустити етап лемматизації, і оброблювати слова лише з підходом стемінгу.

Таблиця 3.3 – Перелік закінчень, притаманних для української мови

Ч.м. / Відмінок	Іменник	Прикметник	Дієслово
Н. в.	-о, -е, -а, -я, -ович, -йович,	-ий, -е, -а, -і, -ій, -я, -ї, -лиций, -лице, -лиця, -лиці	-е, -є, -у, -ю, -ать, -ять, -ив, -ав, -ячи, -ати, -яти, -али, -ме, -вши, -ши, -учи, -сь -уть, -ють, -си, -ся,
Р. в.	-а, -я, -у, -ю, -ях, -ах, -ух, -юх	-ого, -ої, -их, -(ь)ого, -(ь)ої, -іх, -їх, -лицього, -лицьої, -лицих	Дієприкметник
Д. в.	-ові, -еві,	-ому, -ій, -(ь)ому, -їй	-им, -их, -ою, -йми, -ий,
З. в.	-	-у, -ю, -лицю	
О. в.	-ею, -єю, -ою, -ом, -ам, -ям, -ем, -єм	-им, -ою, -ими, -ім, -їм, -іми, -їми, -лицим(и),	Дієприслівник -ив, -ивши(сь)
М. в.	-	-ому, -их, -(ь)ому, -лицьому, -лицій,	
К. в	-о, -е, -є	-	

Україномовні закінчення [17], що характерні для частин мови, які не підлягають під визначення стоп-слів, записані у таблицю 4, де іменники і прикметники підлягають загальному відмінюванню. Через те, що для дієслова, дієприкметника та дієприслівника притаманне окреме поняття дієвідмінювання, ці частини мови винесені в окремий стовпець.

Останнім етапом обробки тексту є процес його представлення у вигляді векторних символів – тобто векторизації. Способи приведення тексту до

векторного вигляду було детально розглянуті у розділі 1.4, де було вибрано BOW-модель. Представлення відгуків як «торбою слів» передбачає підрахунок кількості разів, коли вживається кожне слово, що присутнє у вибірці. Після цього визначається які слова притаманні для «позитивних» та «негативних» відгуків.

Після проведення всіх етапів обробки тексту програма створює новий документ у форматі CSV, який містить поля «processedReviewContent», і «sentimentScore», де передбачається відображення обробленого тексту відгуку і тональної оцінки, що набуває значення -1 або 1 (без «нейтральних» відгуків).

Висновки до розділу

В даному розділу був розглянутий процес створення бази відгуків для подальшого навчання та тестування нейронної мережі. Утворена база відгуків налічує 650 коментарів, більша частина яких є тонально забарвленими, а інші – нейтральними. Відгуки були відібрані з 10 категорій: ноутбуки та комп’ютери; смартфони, ТВ і електроніка; зоотовари; краса та здоров’я; алкогольні напої та продукти; товари для геймерів; спорт і захоплення; сантехніка та ремонт; товари для бізнесу та послуги; а також товари для дому. Сформована вибірка містить 7 полів, які визначають ідентифікатор відгуку та автора, оцінку товару, вміст відгуку, тональну оцінку і дату написання коментаря.

Детально було розглянуто процеси обробки тексту та їх послідовність. Було визначено, які складові відгуку підлягають виділенню. Для етапу стемінгу було створено спеціальну таблицю, яка визначає список україномовних закінчень для таких частин мови як іменник, прикметник, дієслово, дієприкметник та дієприслівник.

Після обробки створюється новий файл, в якому міститься лише тональна оцінка та текст обробленого коментаря.

4 РОЗРОБКА СХЕМИ СТРУКТУРНОЇ

Проектування структурної схеми перш за все має мету наочно продемонструвати процес роботи алгоритму розроблювальної системи, а також взаємодію між її елементами (тобто головними блоками). Розроблена схема зображує логічний та послідовний зв’язок між блоками системи.

Роздроблена структурна схема для системи тонального аналізу відгуків зображена на кресленику ІА81.200БАК.003 Э1. На даній схемі зображена взаємодія модулів машинного навчання з модулем визначення тональності відгуку.

Спочатку взяті з сайту Rozetka відгуки були сформовані в спеціальну базу даних, після чого одна частина відводиться для навчання нейронної мережі, а інша – для тестування. Як правило, для забезпечення точних результатів в подальшому, на етап навчання виділяється значно більша частина вибірки, ніж на тестування. Класичним співвідношенням в галузі машинного навчання є поділ вибірки наступним чином: 80% складає навчальна вибірка, а 20% становить тестова.

Процес навчання передбачає, що, після видалення нейтрально забарвлених відгуків, подана вибірка проходить всі етапи попередньої обробки тексту, які описані в розділі 3.2. Після видалення стоп-слів, роздільних знаків, цифр та закінчень, утворені тексти у векторному заносяться у новостворену вибірку оброблених відгуків, яка містить оброблений текст відгуку та тональну оцінку. Векторизовані слова подаються до модулю тонального аналізу.

Модуль тестування також передбачає процеси видалення незмістовних символів та слів, а також нейтральних відгуків; процеси токенізації, стемінгу та векторизації. Після цього нейронна мережа виконує бінарну класифікацію поданих на аналіз відгуків за допомогою модулю, в якому виконується тональний аналіз.

В процесі тонального аналізу дані з обох вибірок подаються на одну з двох створених нейронних мереж. Одна з них визначає тональне забарвлення відгука

на основі Байєсівського класифікатора, а друга використовує метод опорних векторів.

Робота та схема класифікації обраних методів детально описана в розділі 1.5.2 відповідно. За допомогою одного з двух алгоритмів класифікації виконується визначення тональної складової відгука. На основі визначеного сентименту формується результат проведеного аналізу відгука – він може бути лише «позитивним» або «негативним».

Якість та точність проведеного тонального аналізу безпосередньо залежить від складеної навчальної та тестової вибірки та правильно визначених оцінок тональностей, тому база відгуків була створена ретельним чином, спираючись на відчуття користувачів, які виникли від купівлі певного товару, а також від оцінок за системою рейтингу.

5 РОЗРОБКА ДІАГРАМ ДЛЯ ОБРАНОЇ СИСТЕМИ

Після проведення детального аналізу існуючих моделей і алгоритмів у галузі тонального аналізу, а також після утворення бази відгуків з подальшим розглядом етапів попередньої обробки тексту, стає можливим приступити до етапу створення діаграм, що демонструють роботу розроблюваної системи з боку різних аспектів.

5.1 Діаграма прецедентів

Головна ідея розробки діаграм прецедентів (англ. use case diagram) полягає в тому, щоб відобразити операції, що виконує розроблювальна система в процесі роботи, тобто, згідно до назви, варіанти використання, досить концептуальним способом. Як правило, діаграма варіантів використання передбачає зображення дій системи з розподіленням задач для різних дійових осіб – тобто акторів. Зображені на діаграмі актори мають індивідуальний перелік дій, які виконуються у рамках обраної системи. Актором може бути користувач або сутність (в тому числі система або підсистема), яка виконує взаємодію з певними прецедентами. Вихідна діаграма також демонструє сформульовані вимогу до функціональної поведінки обраної системи.

На кресленику ІА81.200БАК.003 Д1 зображена побудована діаграма прецедентів для проектованої системи, яка зображує перелік дій трьох акторів: розробника користувача і безпосередньо самої системи тонального аналізу відгуків. Розробник може створити власну базу текстів на основі відгуків, що включає збір текстів з україномовних ресурсів, які передбачають написання відгуків на певний товар. Користувач може надати відгук для визначення тональності, після чого отримати результат.

Система з аналізу тональності видаляє нейтрально-забарвлені відгуки, після чого виконує попередню обробку текстів, що включає наступні етапи: видалення незмістовних знаків (цифр, пунктуаційних та розділових знаків), зведення літер

до одного регістру (нижнього), токенізацію слів, видалення стоп-слів та стемінг. Потім представляє оброблені відгуки у вигляді векторних слів і виконує класифікацію за способами ПНБК (Поліноміальний наївний Байєсівський класифікатор) та МОВ (Метод опорних векторів). Система визначає результат тонального аналізу, після чого надсилає його до користувача. Користувач, в свою чергу отримує визначений результат.

5.2 Діаграма діяльності

Діаграма діяльності (англ. activity diagram) досить важлива для зображення поведінки, зосереджена на вирішенні задачі зображення логічного алгоритму реалізації дій, що виконує розроблювана система. Діаграма діяльності демонструє динаміку елементів системи, активні процеси системи у часі, прогрес дій, від початку до самого кінця роботи програми. Проектування такого роду діаграм допомагає розставити наголоси на умовах та потоках дій, що виникають в системі.

Кресленик ІА81.200БАК.003 Д2 демонструє загальний поток активностей розроблювальної системи тонального аналізу, і передбачає дії розробника та системи. На початку розробник збирає і формує базу відгуків, після чого надсилає зібрані дані до системи. Після цього система виконує перевірку на наявність відгуків з нейтральною тональною оцінкою. Далі проходить процес обробки тексту, вилученого з відгуків, і процес його векторизації. Слова, представлені у векторному вигляді проходять класифікацію за одним із обраних методів і визначається значення тональної оцінки – відгук є негативним при набуванні оцінки -1, якщо значення становить 1, коментар визначається як позитивний. Визначений результат надсилається користувачу, після його отримання система завершує свою роботу.

5.3 Діаграма послідовностей

Ще одним різновидом графічного представлення різноманітних систем є діаграми послідовностей. Головою ідеєю подібних діаграм є проілюструвати послідовність взаємодії об’єктів або суб’єктів в рамках заданого сценарію. Взаємодія зображується як послідовний обмін певними «повідомленнями» між процесами системи. Кожен процес є скінченним, що спеціальним чином відображається на «лінії життя» кожного з елементів системи.

Спроектована діаграма продемонстрована на кресленику ІА81.200БАК.003 ДЗ, де показана послідовність дій розробника, користувача, та модулів обробки і сентимент-аналізу відгуків. Після надсилання розробника бази відгуків активується модуль тонального аналізу, який здійснює перевірку на наявність нейтральних відгуків, після чого за наявності видаляє їх. Вибірка з вилученими нейтрально-позначеними відгуками надсилається на модуль обробки тексту, де відгуки проходять попередню обробку для представлення у більш зрозумілому для нейромережі вигляді. Після векторизації, оброблені відгуки знов надсилаються до модуля сентимент-аналізу, де проходить навчання і тестування нейронної мережі, після чого визначається результат тональності, який надсилається до розробника. Користувач може надіслати відгук на модуль обробки тексту, після чого векторизований відгук надсилається на модуль визначення тональності і користувач отримує результат.

Висновки до розділу

В даному розділі було проведено проектування різноманітних діаграм, за допомогою яких становить можливим графічне представлення процесу роботи розроблювальної системи. Діаграми прецедентів та діяльності демонструють дії, які доступні для розробника, користувача і системи в рамках проекту. В той час як діаграма послідовностей демонструє лінію життя процесів кожної сутності.

6 ПРОГРАМНА РЕАЛІЗАЦІЯ СИСТЕМИ

6.1 Вибір інструментів та технологій

Розроблювальна система передбачає написання програмної частини мовою Python. Такий вибір зумовлений досить гарним пристосуванням мови до роботи з машинним навчанням та нейромережами. І хоча машинне навчання та штучний інтелект передбачають залучення вкрай складних алгоритмів, синтаксис Python дозволяє облегшити процес реалізації.

Середовищем для розробки було визначене IDE PyCharm 2022.3.2, що виділяється достатньо широким списком інструментів, таких як, наприклад, засоби аналізу коду або графічне відлагодження. Серед переваг також можна виділити надання можливості комфортного рефакторінга, коректне відображення всіх встановлених пакетів та бібліотек, а також інтерфейс, який є цілком налагодженим під користувача.

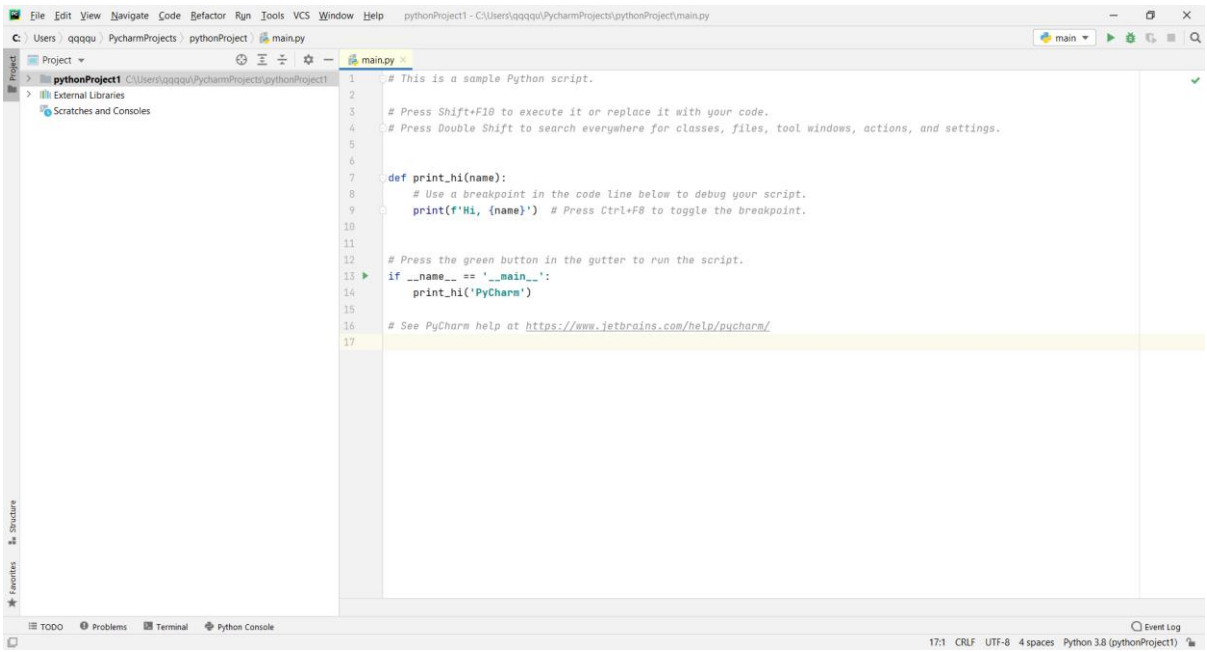


Рисунок 6.1 – Інтерфейс середовища PyCharm

Для реалізації машинного навчання була використана бібліотека scikit-learn, що в рамках даного проекту надає можливість векторного представлення слів моделлю «торба слів», а також застосувати метод опорних векторів та класифікатор наївного Байєса задля забезпечення проведення сентимент-аналізу.

Для того, щоб надати програмі можливість оперувати сформованою базою відгуків, яка має формат CSV, була залучена бібліотека pandas.

6.2 Структура програми

Для забезпечення чіткого визначення тональності, необхідно провести попередню обробку бази відгуків. Модуль підготовки реалізований у якості програми, головною задачею якої є послідовне виконання всіх описаних в розділі 3.2. процесів попередньої обробки відгуків, серед яких: приведення до нижнього регістру, видалення цифр та пунктуаційно-розділових знаків, що занесені в окремий список; приведення до нижнього регістру; токенізація; видалення стоп-слів; і стемінг. Єдиним виключенням, як було зазначено раніше, є апостроф, який оброблюється алгоритмом, що передбачає перевірку наявності знаку на початку або кінці слова і видаляє його. До програми залучені стандартні бібліотеки re, що дозволяє оперувати безпосередньо процесом стемінгу, а також csv, що дозволяє програмі залучити файли з відповідним форматом до роботи.

Також до вбудованого функціоналу середовища відносяться інструменти шифрування вхідних та вихідних даних: на вхід подаються незашифровані дані, які на виході кодуються за стандартом UTF-8, що є достатньо росповсюдженим, а також відрізняється компактним представленням зашифрованих даних. Програма автоматично кодує вхідні документи і дешифрує закодовані.

```
punctuation_list = "!\"#$%&()*+, ./:;<=>?@[\]^_`{|}~—-«»"
current_string = ''

with open('rozetkaReviewDatabase.csv', 'w', newline='') as csvfile:
    comment_writer = csv.writer(csvfile, delimiter=',', quotechar='"', quoting=csv.QUOTE_MINIMAL)
    comment_writer.writerow(["sentimentScore", "reviewContent"])
    with open('review-utf-8.csv', encoding='utf-8', newline='') as csvfile:
        comment_reader = csv.reader(csvfile, delimiter=';')
        for row in comment_reader:
            if (row[0] == "1"):
                continue
            current_string = row[6].lower()
            current_string = re.sub(r'\d+', ' ', current_string)
            current_string = current_string.translate(str.maketrans(punctuation_list, ' ' * len(punctuation_list)))
            current_string_tokenized = current_string.split()
            current_string_tokenized = sw.removeStopWords(current_string_tokenized)
            for word in current_string_tokenized:
                while word.startswith("'"):
                    word = word[1:]
                while word.endswith("'"):
                    word = word[:-1]
                current_string_tokenized = we.modifyEnding(current_string_tokenized)
            current_string = ' '.join(current_string_tokenized)
            comment_writer.writerow([row[0], current_string])
            print(current_string_tokenized)
```

Рисунок 6.2 – Лістинг коду попередньої обробки відгуків

Окремо був створений модуль, що реалізовує процес видалення стоп-слів видалення притаманних для української мови стоп-слів, що описані відповідному розділі 3.2.5. Спочатку був створений список ukrainianStopwordsList(), куди були занесені всі відповідні до існуючої бази слова. Важливо зазначити, що з вибірки була виключена заперечна частка «не», адже для визначення тональне слово має виключну роль. Для процесу самого видалення була визначена функція removeStopWords(), яка циклічно перевіряє поданий на вхід список, і видаляє стоп-слово при умові його ідентифікації.

```
def removeStopWords(string_tokenized):
    for word in ukrainianStopwordsList:
        while word in string_tokenized: string_tokenized.remove(word)
    return string_tokenized
```

Рисунок 6.2.2 – Лістинг коду для видалення стоп-слів

Бібліотека re, що призначена для роботи з регулярними виразами, надає можливість проводити операції з визначеними закінченнями решти частин мови, які не підлягають під визначення стоп-слів, та буде використана безпосередньо для реалізації модуля для стемінгу. Для кожної частини мови визначена функція, що видаляє закінчення слів о перуючи регулярними виразами. Слід зазначити, що поділ закінчень за приналежністю до іменника, дієприкметника, дієслова, прикметника та дієслова є виключно умовним і проведеним з метою покращення розуміння та уникнення плутанини при визначенні.

```
#іменник
noun = ("о", "е", "а", "я", "ович", "йович", "у", "ю", "ах", "ух",
        "юх", "ові", "еві", "єві", "ею", "єю", "ою", "ом", "ам",
        "ям", "ем", "єм", "є")


def checkNoun(list_ex):
    i = -1
    for word in list_ex:
        i = i + 1
        list_ex[i] = re.sub(
            r"(?:о|е|а|і|я|ович|йович|у|ю|ах|ух|юх|ові|еві|єві|ею|єю|ою|ом|ам|ям|ем|єм|є)$",
            "", word)
    return list_ex
```

Рисунок 6.3 – Визначення та видалення закінчень для іменника

На рисунку 6.3 зображена функція перевірки слова на наявність визначеного закінчення згідно таблиці 4 на прикладі іменника . Для решти частин мови було визначено аналогічні функції. Також закінчення "-ати", "-яти", "-ти", "-вши", "-ши", "-учи", "-ячи", "-ючи" були визначені як приналежні до загальної форми слова, тому вони не підлягають видаленню.

На вхід модулю обробки подається база відгуків, сформована у розділі 3.1, що містить 650 коментарів. Після того, як виконано вилучення незмістовних складових з відгуку, а слова розбиті на «токени», на виході отримуємо вихідний документ CSV-формату. Файл два типи даних – а саме «reviewContent», що

містить текст обробленого відгуку; і «sentimentScore», де занесений бал, що визначає тональну оцінку.

1	топ грош купув назад літа нарікань не ма
-1	перест працювати навушник декільк днів
1	навушник зручн зігріва вух не шумл використовуютьс дистанційного навчанн прослуховуванн віде юту
1	улюблен продукц
1	дуж крут монітор зайв не потрібн функц чудов підходить робот
-1	повернув звук дешев навушник кейс дешев китайськ іграшк годин почал боліт вух не звук а форм навушник
1	взяв систем задоволен диск диск встав компютер не бач заход диспетчер диск форматувати побачил систем випадк буд аналогічн проблем
-1	пішл тріщин дужц поган качеств пластмас користуванн був обережн вернут
-1	мож сказат розчарован наушник шумл прост код звук неробить

Рисунок 6.4 – Фрагмент вихідного файлу

На рисунку 6.4 зображений фрагмент вихідного файлу, який демонструє приклади вигляду відгуків після проходження всіх етапів обробки. Для людського сприйняття такий вигляд тексту виглядає як суцільна нісенітниця, але для нейронної мережі подібне представлення сприяє кращому процесу навчання через те, що вилучені незмістовна лексика, а також слова приведені до єдиного вигляду. Загальна кількість коментарів у файлі становить 570 штук через те, що нейтрально забарвлені відгуки були видалені.

Реалізація визначення тональної забарвленості відгука способом бінарної класифікації реалізує модуль сентимент-аналізу відгуків. Вирішення задачі векторного представлення слів також передбачене в даному модулі.

```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix
from sklearn.naive_bayes import MultinomialNB
from sklearn.svm import SVC

input_data = pd.read_csv("processed_reviews_utf-8.csv", header=0, delimiter=",", quoting=3)
num_rev_train = input_data.review.size

reviews_train = []
for i in range(num_rev_train):
    reviews_train.append(input_data.review[i])

vectorizer = CountVectorizer(analyzer="word", preprocessor=None, max_features=5000)
input_data_features = vectorizer.fit_transform(reviews_train)
input_data_features = input_data_features.toarray()
X = input_data_features
y = input_data.sentiment

print(X)
print(y)

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.20, random_state=109)
```

Рисунок 6.5 – Лістинг коду сентимент-аналізу

Функція CountVectorizer() надає можливість представити дані у вигляді вектору. Векторизація передбачає визначення змінних вхідних векторів ознак x та вихідних значень класів y. Відповідно до визначеної BOW-моделі, після виконання процесу векторизації, кожен змінна x представляється у вигляді матриці, що демонструє визначену кількість слів у відгуку. В той же час змінна y визначає порядковий номер відгуку і його тональну оцінку.

Функції MultinomialNB() та SVC() надають можливість провести класифікацію обраними методами поліноміальної класифікації наївного Байєса та методом опорних векторів відповідно.

```
naive = MultinomialNB()
nbclassifier = naive.fit(X_train, y_train)
y_pred_nbc = nbclassifier.predict(X_test)
print("MultinomialNB results")
print("\nAccuracy score: ", accuracy_score(y_test, y_pred_nbc))
print("\nConfusion matrix:\n", confusion_matrix(y_test, y_pred_nbc))
print("\nClassification report:\n", classification_report(y_test, y_pred_nbc))

y_test_array = y_test.to_numpy()

plt.plot(y_test_array, 'g')
plt.plot(y_pred_nbc, 'r--')
plt.show()

svc = SVC(kernel='linear')
svclassifier = svc.fit(X_train, y_train)
y_pred_svc = svclassifier.predict(X_test)
print("SVM results")
print("\nAccuracy score:", accuracy_score(y_test, y_pred_svc))
print("\nConfusion matrix:\n", confusion_matrix(y_test, y_pred_svc))
print("\nClassification report:\n", classification_report(y_test, y_pred_svc, zero_division=0))

plt.plot(y_test_array, 'g')
plt.plot(y_pred_svc, 'r--')
plt.show()
```

Рисунок 6.6 – Лістинг виведення результатів обраними методами

Було визначено, що тестова вибірка складає 114 відгуків, в той час як навчальна містить 456 штук. Загальна кількість «позитивних» відгуків у вибірці складає 63 коментарі, в той час як кількість «негативних» становить 51.

Для коректного формування результатів було застосовано наступні функції: confusion_matrix(), яка дозволяє вивести матрицю, що показує кількість правильно і неправильно визначених відгуків для класу «позитивний» та «негативний»; accuracy_score(), що визначає оцінку точності аналізу; а також classification_report(), яка надає звіт з проведеної класифікації.

```
MultinomialNB results

Accuracy score:  0.90350877193

Confusion matrix:
  [[59 4]
  [7 44]]

Classification report:
               precision    recall   f1-score    support

           1       0.93      0.89       0.90         63
          -1       0.83      0.91       0.86         51

    accuracy                           0.91        114
   macro avg       0.87      0.89       0.88        114
weighted avg       0.88      0.89       0.97        114
```

Рисунок 6.7 – Результати аналізу за ПНБК

На рисунку 6.7 зображені результати проведеної класифікації методом наївного Байєса. З матриці помилок стає відомо, що 59 з 63 позитивних відгуків були визначені правильно. Кількість правильно класифікованих відгуків становить 44 з 51. Даний класифікатор демонструє досить непогані результати визначення тональності відгуків.

```
SVM results

Accuracy score:  0.85087719298

Confusion matrix:
  [[56 7]
 [10 41]]

Classification report:
               precision    recall   f1-score    support

           1       0.88      0.84       0.84         63
          -1       0.80      0.85       0.82         51

    accuracy                           0.86        114
   macro avg       0.83      0.84       0.83        114
weighted avg       0.84      0.84       0.94        114
```

Рисунок 6.8 – Результати аналізу методом опорних векторів

З рисунку 6.8 видно, що результат тонального аналізу методом опорних векторів показав нижчу точність класифікації, а ніж попередній класифікатор. На отриманій матриці помилок відображено, що 57 з 63 відгуків були визначені позитивно, в той час як число правильно визначених негативних відгуків становить 41.

Більш детальний аналіз, що передбачає окремий розгляд кожного аспекту звіту з класифікації, буде розглянутий у наступному розділі, де буде проведене порівняння отриманих результатів.

6.3 Порівняльний аналіз отриманих результатів

Для того, щоб провести порівняльний аналіз отриманих обраними класифікаторами результатів, необхідно попередньо розглянути значення кожного аспекту отриманих результатів. Критеріями визначення коректності результатів проведеної класифікації є наступні чинники:

– Accuracy score – визначає загальну оцінку точності проведеного аналізу від 0 до 1, тому чим більше отриманий результат наближається до значення 1, тим коректніше була виконана класифікація. Визначається як відношення загальної кількості правильно визначених відгуків до їх загальної кількості;

– Confusion matrix – матриця помилок, що показує кількість вірно та невірно визначених результатів. У випадку розроблювальної системи набуває наступного вигляду:

$$\begin{matrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{matrix}, \tag{6.1}$$

де x_{11} та x_{22} – зображують число правильно визначені результати;

x_{21} та x_{12} – відображають кількість хибно визначених зразків.

– Precision [18] – оцінка точності визначення відгуків відповідно класу. Розрахунок відбувається за наступною формулою:

$$\frac{TP}{TP + FP}, \tag{6.2}$$

де TP (True Positives) – кількість правильно визначених зразків певного класу;
FP (False Positives) – кількість хибно визначених зразків того ж класу [14].

– Recall – відклик, що демонструє кількість правильно прогнозованих екземплярів відносно наступної пропорції:

$$\frac{TP}{TP + FN}, \tag{6.3}$$

де FN (False Negatives) – кількість хибно визначених «негативних» відгуків, іншими словами, кількість відгуків що не є негативними.

– F1-score [19] – метрик, що визначає загальну точність проведеної класифікації відповідно кожного класу, методом виведення середнього арифметичного враховуючи попередні параметри:

$$2 \cdot \frac{precision \cdot recall}{precision + recall}. \tag{6.4}$$

– Support – показує загальну кількість екземплярів в кожному класі.

– Accuracy [20] – визначає точність проведеної класифікації як зважене середнє f1-мір кожного класу.

– Macro avg – середнє арифметичне кожного класу, що вираховує середні значення кожного параметру.

– Weighted avg – вираховується для кожного метрика відношення його кожного значення до пропорційного значення загальної кількості відгуків.

Після того, як було визначено значення кожного параметру визначення точності класифікації відгуків, стає можливим утворити порівняльну характеристику двох обраних методів.

Таблиця 6.1 – Кількісне порівняння результатів

Метод/ Клас відгуків	НБК		МОВ		Support
	К-сть правильно визначених	К-сть хибно визначених	К-сть правильно визначених	К-сть хибно визначених	
Позитивні	59	4	56	7	63
Негативні	7	44	10	41	51

Таблиця 6.1 демонструє порівняння отриманих значень, що були надані за допомогою матриці помилок. Обидва класифікатори показали досить якісні результати визначення тональності тестових відгуків. Однак, при порівнянні, стає очевидним, що класифікатор наївного Байєса визначає тональну забарвленість відгуків краще, аніж метод опорних векторів. З таблиці видно, що обидва методи класифікують позитивні відгуки краще, ніж негативні. Можливо скласти припущення, що вираження негативної думки передбачає залучення значно більш широкого спектру лексики, що потребує подальшого розширення навчальної вибірки.

Таблиця 6.2 – Порівняння отриманих значень зі звіту з класифікації

Класифікатор/ Параметр	НБК			МОВ		
	Позитивні відгуки	Негативні відгуки		Позитивні відгуки	Негативні відгуки	
Precision	0.93	0.83		0.88	0.80	
Recall	0.89	0.91		0.84	0.85	
F-1 score	0.90	0.86		0.88	0.82	
Accuracy (f1-score)	0.91			0.86		
	Precision	Recall	F1	Precision	Recall	F1
Macro avg	0.87	0.89	0.89	0.83	0.84	0.83
Weighted avg	0.88	0.89	0.97	0.84	0.84	0.94
Accuracy score	0.90350877193			0.85087719298		

Таблиця 6, що зображує підраховані результати сентимент аналізу за обраними параметрами підтверджує висновок про те, що наївний Байєсівський класифікатор визначає тональну забарвленість відгуків з більшою точністю, ніж метод опорних векторів, про що свідчить показник загальної оцінки точності класифікації. Оскільки оцінки обох методів значно наближені до одиниці – це означає, що процес навчання був проведений вдало.

Через те, що максимальна оцінка – 1, то в даному випадку важливо враховувати точність аж до десятитисячних. Нейромережа з НБК показала результат, точність якого становить 0.9035, в той час як точність МОВ набуває значення 0.8508.

Висновки до розділу

В даному розділі була виконана програмна розробка системи з сентимент-аналізу україномовних відгуків. Були наведені основні принципи реалізації підготовки тексту до визначення тональної складової, а також подальша розробка модулю класифікації з використанням обраних методів.

Спираючись на отримані результати, доцільно зробити висновок, що доцільність використання НБК у рамках роботи з відгуками значно вища, аніж використання МОВ.

ВИСНОВКИ

Визначення тональної складової відгуків є нелегкою задачею через численну кількість лінгвістичних особливостей, притаманних як літературній, так і розмовній українській мові. Протягом виконання кожного етапу розробки даного проекту було проведено детальний огляд теоретичних та практичних рішень.

Завдяки проведеному аналізу існуючих рішень було визначено, що в рамках дослідження тональної складової відгуків більш доцільним буде проведення бінарної класифікації – тобто визначення «позитивної» або «негативної» забарвленості відгуків, спираючись на отримані висновки після огляду застосунків, які окрім тонального аспекту визначають ще і емотивну складову текстів. Через виключну суб'єктивність та розрізненість емоційних станів, більш широкий спектр класифікації може викликати труднощі для сприйняття інформації користувачем.

На основі системи рейтингу товарів з сайту Rozetka було складено базу відгуків, в якій заздалегідь визначений сентимент, спираючись на реальні оцінки користувачів і їх відчуття, що виникають при використанні певного продукту – завдяки цьому вирішується питання щодо мовних особливостей, які в тій чи іншій мірі притаманні природомовним процесам. Створена база відгуків була подана для подальшого навчання і тестування нейромереж, що використовують поліноміальний класифікатор Байєса і метод опорних векторів відповідно.

Отримані результати обох класифікаторів є достатньо задовільними, адже цілковита більшість відгуків була класифікована правильним чином. Нейромережа, що використовує класифікатор наївного Байєса, виконала тональне визначення поданих відгуків незначно, однак краще, ніж метод опорних векторів. З отриманих результатів також можна визначити, що для вираження позитивних емоцій в більшості випадків використовується більш очевидна лексика, тому обидві нейромережі визначили вірно більше «позитивні» зразки, що подавалися на тестування. Для експресії негативних думок користувачам зазвичай

притаманний більш широкий діапазон емотивного лексикона, який значним чином залежить від контексту та проблематики, з якою зіткнувся користувач. Засновуючись на даних припущеннях стає можливим зробити висновок щодо подальшого коригування існуючої бази даних. Для забезпечення підвищеної кількості вірно визначених результатів «негативної» складової, сформована база відгуків потребує додаткових прикладів відгуків, в яких користувачі виражають своє невдоволення.

В подальшому планується провести розширення вже існуючої бази відгуків, задля забезпечення однаково вірних результатів для обох класів. Також подальша робота передбачає розробку застосунку зі зручним користувацьким інтерфейсом на основі нейромережі, що виконує класифікацію саме поліноміальним Байєсівським класифікатором через свою підвищену ефективність в порівнянні з методом опорних векторів.

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Wikipedia. Аналіз тональності тексту. URL: https://uk.wikipedia.org/wiki/%D0%90%D0%BD%D0%B0%D0%BB%D1%96%D0%B7_%D1%82%D0%BE%D0%BD%D0%B0%D0%BB%D1%8C%D0%BD%D0%BE%D1%81%D1%82%D1%96_%D1%82%D0%B5%D0%BA%D1%81%D1%82%D1%83
2. Ahmad Kamal. Subjectivity Classification using Machine Learning Techniques for Mining Feature-Opinion Pairs from Web Opinion Sources. URL: https://arxiv.org/ftp/arxiv/papers/1312/1312.6962.pdf
3. Bing Liu. Sentiment analysis and subjectivity. URL: https://www.researchgate.net/publication/228667268_Sentiment_analysis_and_subjectivity
4. Jonathan Schler, Moshe Koppel. The Importance of Neutral Examples for Learning Sentiment. – Bar Ilan University URL: https://www.researchgate.net/publication/220541948_The_Importance_of_Neutral_Examples_for_Learning_Sentiment
5. Simeng Gu, Fushun Wang, Nitesh P Patel, James A Bourgeois, Jason H Huang. A Model for Basic Emotions Using Observations of Behavior in Drosophila. URL: https://pubmed.ncbi.nlm.nih.gov/31068849/#:~:text=These%20core%20affects%20are%20analogous,as%20love%20and%20aesthetic%20emotion
6. Wikipedia. C4.5 algorithm. URL: https://en.wikipedia.org/wiki/C4.5_algorithm
7. Wikipedia. Support-vector machine. URL: https://en.wikipedia.org/wiki/Support-vector_machine
8. Muhammad Abbas, Kamran Ali, Saleem Memon, Abdul Jamali. Multinomial Naive Bayes Classification Model for Sentiment Analysis. URL: https://www.researchgate.net/publication/334451164_Multinomial_Naive_Bayes_Classification_Model_for_Sentiment_Analysis

9. Andrew McCallum, Kamal Nigam. A Comparison of Event Models for Naive Bayes Text Classification. URL: http://www.kamalnigam.com/papers/multinomial-aaaiws98.pdf
10. Seth Okyere-Dankwa, Ebenezer Komla Gavua. Big Data Sentiment Analysis based on PLSA and its Application. URL: https://www.researchgate.net/publication/328031435_Big_Data_Sentiment_Analysis_based_on_PLSA_and_its_Application
11. Grammarly. URL: https://www.grammarly.com/
12. MonkeyLearn. URL: https://monkeylearn.com/
13. IMB Watson Tone Analyzer. URL: https://cloud.ibm.com/apidocs/tone-analyzer
14. Інтернет-магазин Rozetka. URL: https://rozetka.com.ua/ua/
15. Julian McAuley. Amazon product data. URL: http://jmcauley.ucsd.edu/data/amazon/
16. Serhii Kupriienko. Ukrainian Stopwords. URL: https://github.com/skupriienko/Ukrainian-Stopwords/blob/master/stopwords_ua.txt Ліцензія: https://github.com/skupriienko/Ukrainian-Stopwords/blob/master/LICENSE.txt
17. Марія Блажко, Олександр Авраменко Українська мова та література с. 49-103
18. David M W Powers. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation.
19. Wikipedia. F-Score. URL: https://en.wikipedia.org/wiki/F-score
20. Kenneth Leung. Micro, Macro & Weighted Averages of F1 Score, Clearly Explained. URL: https://towardsdatascience.com/micro-macro-weighted-averages-of-f1-score-clearly-explained-b603420b292f