

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Національний університет «Запорізька політехніка»

Інститут інформатики та радіoeлектроніки,
Факультет комп'ютерних наук і технологій
(повне найменування інституту, факультету)

Кафедра програмних засобів
(повне найменування кафедри)

Пояснювальна записка

до дипломного проєкту (роботи)

магістр

(ступінь вищої освіти)

на тему ДОСЛІДЖЕННЯ ТА ПРОГРАМНА РЕАЛІЗАЦІЯ МЕТОДІВ АНАЛІЗУ
ЕМОЦІЙНОЇ ТОНАЛЬНОСТІ ПОВІДОМЛЕНЬ У СОЦІАЛЬНИХ МЕРЕЖАХ
RESEARCH AND SOFTWARE IMPLEMENTATION OF METHODS FOR
ANALYZING THE EMOTIONAL TONE OF MESSAGES IN SOCIAL
NETWORKS

Виконав: студент 2 курсу, групи КНТ-110м
Спеціальності 121 Інженерія програмного
забезпечення

(код і найменування спеціальності)

Освітня програма (спеціалізація)
Інженерія програмного забезпечення

Ільченко М.М.

(прізвище та ініціали)

Керівник Гофман Є.О.

(прізвище та ініціали)

Рецензент Коцур М.І.

(прізвище та ініціали)

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Національний університет «Запорізька політехніка»
(повне найменування закладу вищої освіти)

Інститут, факультет ІРЕ, ФКНТ
Кафедра програмних засобів
Ступінь вищої освіти магістр
Спеціальність 121 Інженерія програмного забезпечення
(код і найменування)
Освітня програма (спеціалізація) Інженерія програмного забезпечення
(назва освітньої програми (спеціалізації))

ЗАТВЕРДЖУЮ

Завідувач кафедри ПЗ, д.т.н, проф.
С.О. Субботін
“ ” 2021 року

З А В Д А Н Н Я

НА ДИПЛОМНИЙ ПРОЄКТ (РОБОТУ) СТУДЕНТА(КИ)

Ільченку Миколі Миколайовичу
(прізвище, ім'я, по батькові)

1. Тема проєкту (роботи) Дослідження та програмна реалізація методів аналізу емоційної тональності повідомлень у соціальних мережах. Research and Software Implementation of Methods for Analyzing the Emotional Tone of Messages in Social Networks

керівник проєкту (роботи) Гофман Євгеній Олександрович, к.т.н.
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

затверджені наказом закладу вищої освіти від “05” листопада 2021 року № 435

2. Строк подання студентом проєкту (роботи) 5 грудня 2021 року

3. Вихідні дані до проєкту (роботи) рекомендована література, технічне завдання

4. Зміст розрахунково-пояснювальної записки (перелік питань, які потрібно розробити) 1. Аналіз проблемної області та постановка завдань дослідження. 2. Матеріали та методи. 3. Розробка архітектури програми. 4. Основні рішення щодо реалізації компонентів системи. 5. Експлуатація, тестування та експериментальне дослідження програми.

5. Перелік графічного матеріалу (з точним зазначенням обов'язкових креслень) Слайди презентації

6. Консультанти розділів проєкту (роботи)

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	прийняв виконане завдання
1-5 Основна частина	Гофман Є.О., доцент		
Нормоконтроль	Липовець М.В., асистент		

7. Дата видачі завдання “4” вересня 2021 року.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів дипломного проєкту (роботи)	Строк виконання етапів проєкту (роботи)	Примітка
1	Постановка завдання роботи.	1 тиждень	Завдання, ТЗ
2	Аналіз предметної області.	2-3 тижні	Розділ 1
3	Розробка та удосконалення методів, моделей й алгоритмів вирішення задачі.	4-5 тижні	Розділ 2
4	Розробка архітектури програми.	6 тиждень	Розділ 3
5	Розробка програми.	7-8 тижні	Розділ 4
6	Тестування та експериментальне дослідження програмного забезпечення.	9 тиждень	Розділ 5
7	Оформлення пояснювальної записки та документів до неї.	10-13 тиждень	Додатки
8	Нормоконтроль та рецензування.	14-15 тижні	
9	Захист роботи.	16 тиждень	

Студент(ка)

(підпис) Ільченко М.М.
(прізвище та ініціали)

Керівник проєкту (роботи)

(підпис) Гофман Є.О.
(прізвище та ініціали)

РЕФЕРАТ

Пояснювальна записка до дипломної кваліфікаційної роботи магістра:
94 с., 22 рис., 7 табл., 4 дод., 24 джерела.

ЕМОЦІЙНА ТОНАЛЬНІСТЬ, ДОСЛІДЖЕННЯ, АНАЛІЗ,
СОЦІАЛЬНІ МЕРЕЖІ, МАШИННЕ НАВЧАННЯ, ШТУЧНИЙ ІНТЕЛЕКТ.

Об'єкт дослідження – процес аналізу емоційної тональності повідомлень в соціальних мережах.

Предмет дослідження – методи аналізу емоційної тональності повідомлень в соціальних мережах.

Мета роботи – дослідити, проаналізувати та реалізувати методи аналізу соціальних мереж емоційної тональності повідомлень користувачів соціальних мереж.

Матеріали, методи та технічні засоби: структурне та об'єктно-орієнтоване програмування, технології машинного навчання з учителем, мова програмування Java, бібліотека OpenNLP, середовище розробки NetBeans, персональний комп'ютер під управлінням операційної системи Microsoft Windows 10.

Результати. Створено програмний застосунок, що реалізує методи аналізу емоційної тональності повідомлень в соціальних мережах.

Висновки. Розроблено програмний застосунок для аналізу емоційної тональності повідомлень в соціальних мережах.

Галузь використання – системи електронної комерції.

ABSTRACT

Explanatory note to the diploma qualifying work of the master: 94 pages, 22 figures, 7 tables, 4 appendixes, 24 sources.

EMOTIONAL TONE, RESEARCH, ANALYSIS, SOCIAL NETWORKS, MACHINE LEARNING, ARTIFICIAL INTELLIGENCE.

The object of research is the process of analyzing the emotional tone of messages in social networks.

The subject of the research is methods of analyzing the emotional tone of messages in social networks.

The aim of the work is to research, analyze and implement methods for analyzing the emotional tone of social media users ' messages.

Materials, methods and technical tools: structural and Object-Oriented Programming, machine learning technologies with a teacher, Java programming language, OpenNLP library, NetBeans development environment, personal computer running the Microsoft Windows 10 operating system.

Results. A software application has been created that implements methods for analyzing the emotional tone of messages in social networks.

Conclusions. A software application has been developed for analyzing the emotional tone of messages in social networks.

The field of use is e-commerce systems.

ЗМІСТ

	С.
Перелік скорочень та умовних познач.....	8
Вступ.....	9
1 Аналіз проблемної області та постановка завдань дослідження.....	11
1.1 Соціальні мережі, сфери та об'єм їх впливу	11
1.2 Аналіз тональності в соціальних мережах	17
1.3 Визначення області аналізу емоційної тональності повідомлень	24
1.4 Визначення полярності тексту.....	28
1.5 Методи автоматичного вилучення аспектів.....	31
1.6 Порівняння підходів машинного навчання для ШІ в рамках аналізу емоційної тональності повідомлень	35
1.7 Постановка завдання	36
1.8 Висновки за розділом 1	36
2 Матеріали та методи.....	37
2.1 Дослідження особливостей повідомлень в Twitter.....	38
2.2 Реалізація комплексу для визначення полярності повідомлень	39
2.3 Вибір методу класифікації	41
2.4 Вибір навчальної вибірки.....	43
2.5 Висновки за розділом 2	43
3 Розробка архітектури програми	44
3.1 Опис основних вимог до програми.....	44
3.2 Вибір мови програмування	44
3.3 Бібліотека OpenNLP.....	47
3.4 Вибір середовища розробки.....	48
3.5 Вимоги до технічних засобів	51
3.6 Висновки за розділом 3	51
4 Основні рішення щодо реалізації компонентів системи	52
4.1 Загальне представлення розробки.....	52
4.2 Взаємодія модулів програми	53

4.3	Об'єктно-орієнтовна розробка представлених модулів.....	54
4.4	Бібліотека LIBLINEAR.....	55
4.5	Експлуатаційні характеристики ПЗ	56
4.6	Висновки за розділом 4	56
5	Експлуатація, тестування та експериментальне дослідження програми..	57
5.1	Опис стратегії тестування	57
5.2	Результати тестування	57
5.3	Реалізація методу автоматичного вилучення аспектів	58
5.4	Вибір методу вилучення аспектів	59
5.5	Висновки за результатами тестування.....	60
	Висновки	61
	Перелік джерел посилання	62
	Додаток А Технічне завдання	65
	Додаток Б Опис програми	70
	Додаток В Текст програми	73
	Додаток Д Слайди презентації.....	87

ПЕРЕЛІК СКОРОЧЕНЬ ТА УМОВНИХ ПОЗНАК

JVM	– java virtual machine;
HIPAA	– health insurance portability and accountability act;
NLP	– natural language processing;
SMM	– social media marketing;
SMO	– social media optimization;
SMO	– метод опорних векторів;
ЗМІ	– засоби масової інформації;
ПЗ	– програмний засіб;
ПК	– персональний комп'ютер;
ОС	– операційна система;
ІІ	– штучний інтелект.

ВСТУП

З моменту своєї появи, соціальні мережі стали стрімко набирати популярність [1], [2]. Сьогодні мільйони людей діляться враженнями про навколишній світ зі своїми знайомими і друзями. Завдяки повсюдному поширенню бездротових мереж і мобільного інтернету, користувачі можуть відправляти повідомлення майже в будь-який зручний для них момент. Так, наприклад, журналісти ведуть текстові репортажі з місць подій в реальному часі, туристи пишуть про свої подорожі з будь-якої точки планети [1], [2].

Свій профіль в популярних соціальних мережах мають відомі компанії, організації, університети. Політики і громадські діячі, які бажають бути більш відкритими для суспільства, діляться своїми думками і цікавляться думкою людей про прийняті ними рішення [1], [2].

Соціальні мережі, таким чином, надають дослідникам можливості для детального аналізу думок користувачів. Соціальна мережа Твіттер виділяється серед інших малою тимчасовою затримкою між подією і появою думки про нього. Наприклад, дослідницький проект Pulse of the Nation створений для визначення настрою протягом дня у громадян США, які активно користуються соціальною мережею Твіттер. Творці проекту SportSense розробили алгоритми для визначення рівня схвильованості спортивних уболівальників за їхніми повідомленнями в Твіттер, що дозволяє їм відстежувати ключові моменти ігор Національної футбольної ліги США в реальному часі.

Для вирішення завдань, пов'язаних з виявленням і подальшим аналізом емоційно забарвленої лексики в тексті, використовуються методи, загальна назва яких – аналіз тональності тексту (Sentiment Analysis). Інша назва даної області українською мовою – «аналіз емоційного забарвлення тексту». Аналіз тональності тексту входить в область завдань комп'ютерної лінгвістики і є підзадачею отримання та обробки інформації (Information Retrieval) [1]-[6].

Завдання аналізу тональності повідомлень в соціальних мережах вирішується в рамках дослідницького роботи. Мета дипломної кваліфікаційної роботи магістра полягає в дослідженні методів та розробці програмних засобів (ПЗ) аналізу соціальних мереж.

1 АНАЛІЗ ПРОБЛЕМНОЇ ОБЛАСТІ ТА ПОСТАНОВКА ЗАВДАНЬ ДОСЛІДЖЕННЯ

Перший розділ буде присвячено детальному аналізу проблемної області пов'язаної з повідомленнями в соціальних мережах. Особливу увагу буде приділено використанню методів машинного навчання для аналізу емоційної тональності повідомлень. Під час роботи буде проведено дослідження основних термінів таких, соціальна мережа, повідомлення, емоційна тональність та забарвлення, методи машинного навчання.

1.1 Соціальні мережі, сфери та об'єм їх впливу

1.1.1 Визначення соціальних мереж

Соціальна мережа – це інтернет-платформа, яка дозволяє людям зі схожими інтересами збиратися і обмінюватися інформацією, фотографіями і відео [7]-[10]. Соціальні мережі стали невід'ємною частиною нашого життя. Людство використовує їх, щоб залишатися на зв'язку з друзями, отримувати новини або знаходити роботу. З кожним днем спостерігаємо все більше і більше нових користувачів [7]-[10]. Для маркетологів це означає, що існує величезний потенціал для залучення уваги до свого бренду широкої і зацікавленої аудиторії [7]-[10]. Приблизно 3,5 мільярда інтернет-користувачів у світі користуються соціальними мережами, і очікується, що ця цифра буде зростати, оскільки використання мобільних пристроїв і мобільних соціальних мереж все більше набирає обертів [7]-[10].

Напрямки в соціальних мережах дуже різноманітні: такі платформи, як Facebook, в значній мірі орієнтовані на обміни фотографіями або статусами між друзями. Інші соціальні мережі, такі як Tumblr або Twitter називаються мікроблогами. Деякі соціальні мережі фокусуються на спільнотах; інші виділяють і відображають користувальницький контент.

Завдяки постійній присутності в житті своїх користувачів, соціальні мережі мають вирішальний соціальний вплив [7]-[10].

1.1.2 Можливості соціальних медіа

В першу чергу соцмережі створювалися для спілкування. Вони дозволяють створювати особисті профілі, додавати інших користувачів в друзі або підписуватися на них, спілкуватися в публічних коментарях або приватних повідомленнях. Але спілкування з друзями не єдина можливість соціальних мереж [7-10].

Більшість соцмереж підтримує зображення, відеоролики та анімовані картинки, деякі також аудіозаписи. Текстовим і мультимедіа-контентом можна ділитися, пересилати його іншим або викладати публічно.

Ігри, музика, Серфінг розважального контенту – соціальні мережі пропонують широкі можливості для відпочинку і дозвілля. На цьому заснована Популярність безлічі спільнот з розважальним контентом на зразок смішних картинок або цікавих історій [7-10].

Компанії використовують сторінки в соцмережах як частину самопрезентації. Ведення соціальної мережі, або Social Media Marketing (SMM), має на увазі створення контенту, спілкування з користувачами, підтримку обраного іміджу, розповіді про новини фірми, просування. Засоби масової інформації (ЗМІ) застосовують соцмережі як ще один новинний канал: існують навіть спільноти, орієнтовані на новини [7]-[10].

1.1.3 Просування і реклама в соцмережах

Більшість сучасних соціальних мереж мають широкі рекламні можливості. Це зрозуміло: реклама-основне джерело доходу такого сайту, а велика кількість відомостей, які користувачі вказують про себе, дозволяє чітко таргетувати рекламу по цільовій аудиторії. Велика кількість

можливостей для контент-маркетингу та реклами перетворює соцмережі в потужний інструмент просування [7]-[10].

Правильна розкрутка в соціальних мережах збільшує охоплення, підвищує впізнаваність, конверсію і трафік, дозволяє заявити про себе аудиторії. Головне-вибрати майданчик, яка підходить для вашої сфери.

Просування в соцмережах зводиться до реклами і контент-маркетингу. Перша буває двох видів: явна і прихована. Явна-це оголошення в окремих блоках, банери, публікації з позначкою «рекламні». Прихована - так звана нативна реклама, органічно вписана в контент.

Рекламна кампанія в соцмережах може дуже гнучко налаштовуватися, але вона не єдине, що потрібно для успішного просування. Грамотне ведення спільноти або сторінки значить набагато більше [7]-[10].

1.1.4 Соціальні мережі як інструмент рекламної та PR-діяльності в Інтернеті

Соціальна мережа є онлайн-сервісом, сайтом або платформою, що служить для комунікації різних груп громадськості. У мережі абсолютно кожна людина може залишити про себе інформацію (вік, місце і рік навчання, робота хобі і т.д.), для полегшення пошуку цього облікового запису серед інших користувачів соціальної мережі. Варто відзначити, що соціальні мережі можуть являти собою як відкриті джерела отримання інформації та спілкування, так і закриті, де доступ отримує тільки обмежене число осіб. Але, кожна соціальна мережа володіє одними загальними ознаками – діє система спільнот та друзів. Згідно з дослідженнями різних аналітичних сайтів, в соціальних мережах складається кожен другий користувач Інтернету. На сьогоднішній день, малий і великий бізнес воліє в якості інструменту реклами своєї діяльності Інтернет-маркетинг [7]-[10]. Слід розглянути переваги просування компанії, її товарів і послуг за допомогою соціальних мереж. Серед них можна відзначити [7-10]:

- низьку вартість рекламної компанії (на відміну від традиційних рекламних носіїв, які в два рази дорожче): кожен користувач є реальним споживачем інформації [7]-[10];
- можливість широкого охоплення цільової аудиторії (крім того, аудиторія соціальних мереж щодня збільшується) [7]-[10];
- використання зворотного зв'язку між клієнтом і організацією, можливість негайного реагування на ті, чи інші повідомлення [7]-[10];
- олюднення бренду компанії в очах цільової аудиторії (в даному випадку реклама сприймається як рекомендація, порада або думка знайомих, через що рівень довіри підвищується) [7]-[10];
- компанія має можливість проводити моніторинг популярності своєї спільноти або акаунту серед своїх передплатників, бачити число нових учасників, відстежувати їх коментарі стосуються діяльності організації.

Всі ці дії необхідні для створення бази даних лояльних споживачів і їх переваг. Грамотно використовую такі моделі просування як SMM та Social Media Optimization (SMO) можна домогтися успіху серед користувачів соціальних мереж. Розглянемо кожну з них докладніше [7]-[10].

SMM – просування сайту, бренду або послуг компанії через спілкування з представниками цільової аудиторії за допомогою соціальних мереж. Такий підхід сприяє залученню користувачів на сайт компанії, не використовуючи пошукові системи. Говорячи про SMM необхідно відзначити, що цей метод вносить великий внесок в перспективу просування, так як у цільової аудиторії формується позитивне сприйняття бренду або товару [7]-[10].

SMO – включає в себе проведення внутрішніх технічних робіт, спрямованих на збільшення ефективності та оптимізації сайту з соціальними мережами. Всі ці дії спрямовані на інформаційну складову ресурсу, зміну інтерфейсу, що дає можливість синхронізувати сайт з різними соціальними мережами. Коштів для SMO-оптимізації не так вже й багато. Але їх достатньо щоб залучати нових користувачів і забезпечити необхідне

функціонування ресурсу. Далі розглянемо основні інструменти SMO просування – розкрутка корпоративних і особистих блогів з урахуванням можливості відкритого коментування – - ведення форумів різної тематики; - впровадження та використання в інтерфейсі кнопок оперативного додавання записів в соціальні мережі (репости). - використання таких форматів як: pdf-файли, відео-та аудіозаписи. Беручи до уваги всі перераховані вище засоби соціального просування в мережі Інтернет, з'являється можливість домогтися ефективної, налагодженої комунікації з цільовою аудиторією. Паралельно з цим, компанія вирішує безліч інших завдань: підвищує впізнаваність бренду, дозволяє виділитися серед конкурентів, підвищує рівень товарообігу і обсяг продажів. Багато сучасних дослідників схиляються до думки, що SMO і SMM стають більш значущими методами просування, ніж реклама. Тому що одними з головних переваг при використанні цих інструментів є оперативність, реклама має рекомендаційний характер, не стає нав'язливою і сприймається споживачем як рекомендація друзів або будь-яких фахівців [7]-[10].

1.1.5 Соціальні мережі, як інструмент охорони здоров'я

Медичні працівники можуть подумати, що соціальні медіа призначені більше для того, щоб ділитися тим, що ви приготували на сніданок, ніж про те, яку медичну процедуру ви виконуєте, – і вони мають рацію. Але це не означає, що соціальні мережі не є цінним маркетинговим інструментом для лікарів і постачальників медичних послуг.

Якщо все зроблено акуратно, соціальні мережі, включаючи Facebook, Instagram, Twitter, YouTube і LinkedIn, є потужним способом завоювати лояльність поточних пацієнтів і встановити зв'язок з майбутніми новими пацієнтами [7-10].

Тим не менш, багато лікарів і фахівців у галузі охорони здоров'я уникають цифрового маркетингу через соціальні мережі, побоюючись, що

вони будуть суперечити законам Health Insurance Portability and Accountability Act (HIPAA). Це законне занепокоєння – лише одна з унікальних проблем маркетингу через соціальні мережі в галузі охорони здоров'я [7]-[10].

Але, слідуючи кращим практикам професіоналів в області маркетингу, можна використовувати соціальні мережі для збільшення своєї бази пацієнтів. Можна уважно стежити за своєю репутацією в Інтернеті і активно ділитися повідомленнями, які відповідають цінностям вашої практики, при цьому захищаючи конфіденційність ваших пацієнтів.

HIPAA – не єдина проблема для маркетингу в соціальних мережах в сфері охорони здоров'я. Також важливо (а іноді і складно) підтримувати рівень професіоналізму, який часто може бути підірваний випадковим, іноді дурним характером соціальних мереж [7]-[10].

Один із способів забезпечити професіоналізм – це створити веб-сайт, до якого поточні та потенційні пацієнти можуть отримати доступ, щоб отримати точну та важливу інформацію про послуги та процедури, які надає ваша практика. Переконайтеся, що ваш веб-сайт адаптується до мобільних пристроїв, оскільки багато людей виконують пошук з мобільних пристроїв.

Лікарю можна позиціонувати себе як експерта у своїй галузі, але тільки якщо він завжди публікує доречний контент. Ось кілька прикладів того, як можна використовувати соціальні мережі для просування своєї практики [7]-[10].

Ділитися новими послугами. Раніше можна було надсилати прес-реліз до місцевої газети. Сьогодні більшість людей отримують новини через соціальні мережі. Всякий раз, коли лікарі розширюють спектр послуг, наймаючи нового медичного працівника або купуючи нове обладнання, вони можуть оголосити про це в своїх облікових записах в соціальних мережах.

Соціальні мережі дають можливість безпосередньо спілкуватися з клієнтами. Якщо у них є питання про нові послуги, можна порадити їм

зв'язатися з вашим офіс-менеджером або фахівцем з маркетингу за допомогою прямого повідомлення.

Повідомляти про нові правила. Точно так само необхідно оголосити про нову політику в соціальних мережах. Наприклад, у лікаря може бути новий протокол, введений в дію в результаті пандемії коронавірусу. Заздалегідь повідомляючи своїм пацієнтам про новий зал очікування та процедури маскування, лікар забезпечує покращене обслуговування клієнтів, якого вони заслуговують [7]-[10].

Заохочення пацієнтів до участі. Не всі знають, що потрібно робити, щоб активно впливати на стан свого здоров'я. Тут професіонали охорони здоров'я можуть надати відмінні послуги. Слід розробити план, що написати в блозі, що наповнений порадами, пов'язаними зі здоров'ям, які мають відношення спеціальності лікаря.

Керування своєю репутацією в Інтернеті. Ще одна проблема, з якою медичні працівники стикаються в соціальних мережах, – це побоювання, що виникне негативна реакція. Реальність така, що завжди знайдуться незадоволені люди. Професійно обробивши негативні відгуки, можна заспокоїти пацієнта, який скаржиться, а також отримати нові.

Соціальні мережі можуть працювати на лікарів. Цифровий маркетинг – це ефективний спосіб зв'язатися з поточними пацієнтами та зв'язатися з потенційними новими пацієнтами. Публікуючи освітню та важливу інформацію на сайтах компанії в соціальних мережах, можна збільшити відвідуваність свого сайту. Також можна почати рекламну кампанію в Facebook, щоб клініку помітили [7]-[10].

1.2 Аналіз тональності в соціальних мережах

З кожним роком аналіз тональності тексту стає все більш актуальним завданням як з теоретичної, так і з практичної точки зору. Це, в основному,

пов'язано з розвитком Інтернету і змінами в форматі спілкування між людьми [1]-[6].

Поряд зі збільшенням кількості інтернет-користувачів, збільшується і обсяг генерованого ними контенту. Люди розміщують повідомлення на форумах, пишуть в блогах, коментують товари на сторінках інтернет-магазинів і висловлюють свою думку в соціальних мережах [1]-[6].

Весь цей контент несе в собі значну кількість інформації, яку можна і потрібно використовувати [7]-[10]. Щоб краще розібратися що це таке, давайте почнемо з самого початку. Аналіз тональності – це область інтелектуального аналізу тексту, призначена для автоматичного вилучення суб'єктивних думок з тексту. Як дисципліна, аналіз тональності знаходиться на стику пошуку даних і комп'ютерної лінгвістики; він не обмежується дослідженням текстового контексту і більш зосереджений на його тоні [1]-[6].

Аналіз тональності тексту дозволяє виділити емоційно забарвлений словниковий запас і емоційне ставлення авторів до об'єктів, описаних в тексті. У сучасному світі на наш вибір часто впливають думки інших людей: ми читаємо відгуки про продукти, перш ніж замовляти їх в інтернет-магазині, ми вивчаємо думки інших людей, перш ніж голосувати на виборах конкретного кандидата, ми ретельно вибираємо університет для вступу і навіть музей або ресторан, який ми збираємося відвідати [1]-[6].

Аналіз тональності може допомогти маркетологам, соціологам і багатьом іншим фахівцям. Крім того, ці дані важливі для багатьох компаній, не залежно від галузі, оскільки дозволяє їм дізнатися думку своїх клієнтів. Будь то думка про інновації, зроблені в продуктах або послугах, або про останні новини на вашому сайті, або оцінка користувачами товарів в інтернет-магазині [1]-[6].

1.2.1 Використання аналізу тональності повідомлень в соціальних мережах для електронної комерції

Оцінка іміджу бренда. Щоб правильно інтерпретувати думку покупців про бренд, необхідно зрозуміти соціальні настрої [5]-[10].

Багато клієнтів, як правило, відмовляються від послуг компанії з поганою репутацією бренду. Наприклад, якщо фірма відома як ресторан, в якому подають зіпсовану їжу, ніхто не захоче самостійно перевіряти, чи дійсно це так – люди просто туди не прийдуть. Або, наприклад, якщо це є компанія з доставки, і про неї говорять, що зазвичай доставляється тільки половина замовлення, навряд чи хтось вибере організацію в якості першої в списку кращих служб доставки.

Інструменти для аналізу тональності дозволять побачити імідж бренду в соціальних мережах. Більш того, професійне програмне забезпечення допоможе глибше зануритися в причини і витоки такого іміджу і зрозуміти його основну причину [5]-[10].

Відшукати клієнтські інсайти. Насправді, цей пункт включає в себе безліч аспектів, які важливі для задоволення очікувань ваших клієнтів. Інструменти для аналізу тональності в соціальних медіа нададуть важливу інформацію про ставлення клієнтів до продуктів і послуг під різними кутами.

Завдяки комплексному аналізу тональності можна помітити негативний сплеск раніше, ніж він зможе привести до незворотних наслідків.

Основна ідея полягає в тому, що все це можна виявити за допомогою належного програмного забезпечення; воно може надати безліч інсайтів, які більше ніде не знайти. Інструменти аналізу тональності в соціальних мережах звільняють від рутинних завдань, оптимізують робочий процес і дозволяють приділяти увагу дійсно важливим речам [5-10].

Управління клієнтським досвідом. Кілька років тому клієнти в основному використовували різні форуми, веб-сайти компаній та опитування, щоб висловити свою думку про товар або послугу. У міру того

як соціальні мережі набирали обертів, саме вони стали основним джерелом відгуків клієнтів з кількох причин. Перш за все, клієнти, які залишають відгуки в соціальних мережах, вимагають негайної уваги компанії; по-друге, вони хочуть попередити друзів і колег про використання продукту або, навпаки, спонукати їх до цього; по-третє, люди хочуть привернути увагу громадськості до того чи іншого питання. Незважаючи на причину, будь-яка публічна зворотній зв'язок матиме певні наслідки для вашої компанії, залежно від контексту [5]-[10].

Професійне програмне забезпечення для аналізу тональності допоможе вам ефективно управляти якістю обслуговування клієнтів і всіма супутніми показниками, такими як клієнтська лояльність, ступінь задоволеності клієнтів і т. д.

Аналіз. конкурентів. На ринку, усі конкуренти приділяють багато часу розробці нових маркетингових стратегій, випуску нових продуктів, обмірковуванню інноваційних маркетингових кампаній. І з усіма цими ініціативами у всіх конкурентів є свої злети і падіння. Незважаючи на всі зусилля, складно передбачити реакцію і поведінку аудиторії у відповідь на маркетингові активності. Тому дуже часто найкраще рішення – це спостерігати, як інші щось зробили (якщо вони це зробили), і що вийшло в результаті [5]-[10].

Візуальна аналітика. Аналіз візуального контенту – це порівняно новий напрямок в моніторингу соціальних мереж. Тепер сучасне програмне забезпечення здатне виявляти логотип компанії на зображенні, що може стати в нагоді для різних цілей: аналіз споживчих ситуацій, оцінка реакції аудиторії на певну рекламу, вимірювання охоплення та інші [5]-[10].

Найбільш професійні системи йдуть ще далі завдяки своїй здатності ідентифікувати не тільки логотип, але і визначати сцену, об'єкти, людей, рід занять (якщо це відображено на зображенні), виконувані дії і так далі.

Розпізнавання трендів. Програмне забезпечення для аналізу соцмереж на базі штучного інтелекту (ШІ) допоможе визначити актуальні теми, які

стосуються вашого бренду, послуги, продукту або навіть ваших конкурентів. Таке ПЗ охоплює контекст навколо необхідної теми і надає вам всю необхідну додаткову інформацію: стаття авторів, їх місце розташування, домінуючий тип розміщуваного контенту і основне джерело [5]-[10].

Персональний Асистент. Комплексні системи моніторингу соціальних мереж повинні працювати на основі ШІ, а розумний особистий помічник – одна з кращих функцій, яку вони можуть запропонувати на додаток до аналітики тональності. Особистий помічник може не тільки провести аналіз, але і вирішити, що слід проаналізувати, як це зробити, порівняти результати з цілями вашої організації і видати зрозумілий результат. Оснащений можливостями відстеження тональностей, персональний асистент надасть вам більш точну інформацію і повідомить вас про найбільш значущі події в заданій темі [5]-[10].

1.2.2 Використання аналізу тональності повідомлень в соціальних мережах для визначення соціальних процесів та прийняття рішень в охороні здоров'я на прикладі епідеміологічних досліджень

Соціальні мережі розробили платформи, що сприяють комунікації між людськими спільнотами і допомагають їм в обміні ідеями, інформацією, знаннями та іншими даними за допомогою форм електронної комунікації. Платформи соціальних мереж вважаються однією з найбільш швидкозростаючих інформаційних систем для суспільного застосування. Соціальні платформи вважаються глобальним центром зберігання баз даних, оскільки люди проводять на них багато часу, використовуючи їх Додатки. Деякі з найбільш часто використовуваних додатків для соціальних мереж у світі – це Facebook, Twitter та Instagram. Наприклад, історія Twitter почалася в 2006 році, як один з науково-дослідних проектів Компанії з Сан-Франциско для внутрішнього застосування – Odeo. Автор ідеї, Джек Дорсі, мав намір створити сервіс, який дозволив би йому вільно обмінюватися з друзями

короткими повідомленнями. Успіх до Twitter прийшов вже в 2007 році на фестивалі South by Southwest) Під час заходу (за добу) кількість користувачів Twitter зросла з 20 000 до 60 000, а в травні 2008 року Twitter мав 1 мільйон зареєстрованих користувачів [5]-[10].

На сьогоднішній день, сервіс мікроблогів Twitter – це система, яка дозволяє користувачам відсилати текстові короткі замітки (не більше 140 символів), застосовуючи SMS, веб-інтерфейс, миттєві повідомлення або інші програми-клієнти. Відмітна особливість Twitter – це публічна доступність повідомлень, які на ньому розміщені. Twitter на заході став не тільки одним з популярних способів спілкування, але і проник в лексику і образ думок [5]-[10].

Таким чином, враховуючи час, який користувачі проводять в подібних соціальних мережах, від декількох годин на тиждень до щоденного використання, ці додатки безсумнівно впливають на поведінку людей. Незважаючи на великий обсяг даних, представлених на цих платформах соціальних мереж, їх зміст може мати суперечливі ефекти, які варіюються від негативного до позитивного психологічного впливу на життя людей. Люди, захоплені соціальними мережами, швидше за все, поділяться думками та ідеями на цих платформах. Згодом дуже цінно перетворити ці думки і повідомлення в активи [5]-[10].

Твіт або повідомлення в Facebook може бути виявлений, якщо було поставлено мільйон лайків (умовних виразів схвалення матеріалу або користувача, що виражаються натисненням однієї кнопки) і зроблено велику кількість ре-твітів (повторюваних текстових повідомлень), але така масова взаємодія з повідомленням не відображає його важливості або емоцій користувачів, які беруть участь в публікації, через багатьох факторів, таких як характер повідомлень, що включає заперечення і іронію; щастя і смуток; гнів; занепокоєння, здивування, відроза або плутанина; і величезна кількість твітів. Проте широкомасштабне визначення людських емоцій у соціальних мережах має важливе значення для міжнародного суспільного впливу,

ділових рішень та розробки політики. Аналіз тональності текстів у цьому контексті став корисним інструментом оцінки громадської думки. Аналіз тональності текстів сприяє розумінню людських емоцій, оскільки він може трактувати поведінку людей, які взаємодіють з додатками соціальних мереж. Крім того, ці мультимедійні програми використовуються в різних областях, включаючи здоров'я, бізнес, туризм, освіта, таких як аналіз думок і надання людям можливості вільно висловлювати свої емоції. Таким чином, аналіз тональності текстів в соціальних мережах є потужним інструментом для розуміння найбільш важливих світових тенденцій і подій. Особливо коли такою подією стає епідемія [5]-[10].

Адже негайне і ефективне спілкування необхідно для зменшення невизначеності, оскільки воно дозволяє отримати чітке уявлення про ситуацію і знати, що робити, щоб впоратися з ситуацією. Це також підвищує впевненість людей в тому, що керівництво справляється з ситуацією [5-10].

З тих пір, як в грудні 2019 року було опубліковано перший заголовок про спалах коронавірусного захворювання 2019 року (COVID-19), соціальні мережі стали живильним середовищем для поширення інформації про новий коронавірус. Інформація, що представляє собою і правду і брехню, була опублікована на різних платформах соціальних мереж, випередивши поширення хвороби. Пандемія в соціальних мережах передувала пандемії хвороби, викликаючи широкий спектр емоцій. Хоча світ і раніше був свідком пандемій, всі вони були в епоху до появи соціальних мереж. І вплив соціальних мереж під час такої безпрецедентної пандемічної кризи ще належить визначити. Адже в різних дослідницьких статтях вказувалося, що багато спалахів і пандемії можна було б оперативно контролювати, якби експерти враховували дані соціальних мереж. Отже, аналіз тональності текстів при вивченні пандемій, таких як COVID-19, може відігравати важливу роль. Адже COVID-19 залишається спірною глобальною темою в соціальних медіа [5]-[10].

Про плюси і мінуси соціальних мереж спілкувалося в громадській охороні здоров'я і медицині. Під час пандемії COVID-19 соціальні мережі використовувалися в якості платформ установами охорони здоров'я, урядовими організаціями, лікарнями та медичними журналами для своєчасного поширення інформації. Соціальні мережі також використовувалися лікарями, які лікували пацієнтів з COVID-19, щоб оперативно інформувати медичну спільноту про свій досвід [5]-[10].

На жаль, соціальні мережі також були каналом для розповсюдження чуток. Першорядна соціальна платформа Twitter для ефективної комунікації, з з'являються кожні 45 мс. твітами про COVID-19, була змушена видалити ряд призначених для користувача акаунтів на підставі поширення дезінформації про пандемію, що зміцнювало сприйняття людьми невизначеностей, викликаних пандемією [5]-[10].

1.3 Визначення області аналізу емоційної тональності повідомлень

В огляді розглядаються завдання, що виникають при аналізі тональності текстів і термінологія, в якій ці завдання описуються. Розглянуто методи визначення полярності та вилучення аспектів з емоційно забарвлених текстів [11]-[15].

1.3.1 Аналіз емоційного забарвлення тексту

Термін «аналіз емоційного забарвлення тексту» використовується в даній роботі в якості перекладу оригінального терміна «sentiment analysis», який часто так само перекладається, як аналіз тональності тексту.

Згідно з визначенням, даним: аналіз емоційного забарвлення тексту – завдання автоматичного аналізу думок і емоційно забарвленої лексики, виражених в тексті [11]-[15].

В аналізі тональності тексту вважається, що текстова інформація в мережі Інтернет ділиться на два класи: факти і думки [11]-[15]. Ключовим поняттям є визначення думки [11]-[15].

Думки діляться на два типи [11]-[15]:

- проста думка;
- порівняння.

1.3.2 Думка

Проста думка містить висловлювання автора про один об'єкт [11]-[15]. Воно може бути висловлено прямо: мене приємно здивувало якість збірки меблів, або неявно: після курсу терапії моє здоров'я зміцнилося. В обох випадках проста думка зазвичай має емоційне забарвлення-позитивну або негативну [11]-[15].

В аналізі тональності тексту для думки першого типу дається формальне визначення: простою думкою називається кортеж з п'яти елементів (entity, feature, sentiment value, holder, time) [11]-[15]. Де entity-об'єкт, про аспект (feature) якого автор (holder) висловив думку в момент часу (time) [11]-[15]. Виділяються 3 види емоцій (sentiment_value): позитивна, негативна і нейтральна. Під нейтральною мається на увазі, що текст не містить емоційної складової [11]-[15].

Об'єктом є людина, організація, подія, товар або тема обговорення. Тому в різних публікаціях об'єкт так само називається object або topic. Часто об'єкт можна представити у вигляді ієрархічного дерева компонент і підкомпонент (рис. 1.1). З кожною компонентою пов'язаний набір атрибутів. У наведеному вище визначенні думки аспект має на увазі під собою і компоненти, і їх атрибути. Окремим випадком аспекту є сам об'єкт [11]-[15].

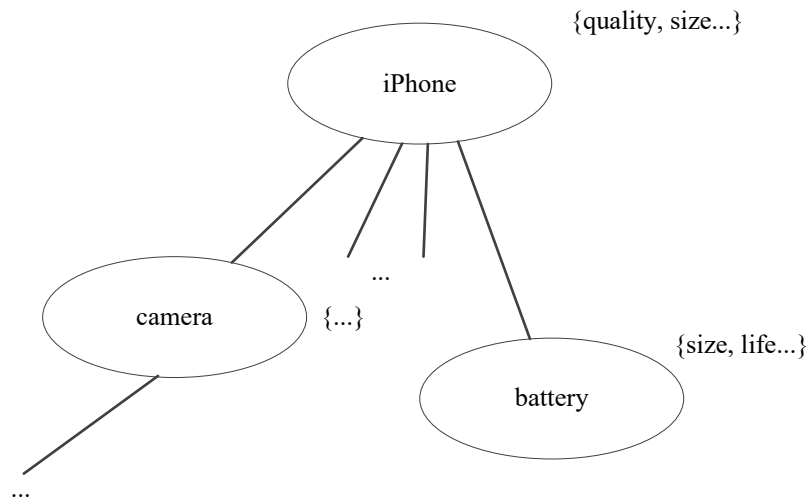


Рисунок 1.1 – Об'єкт володіє компонентами і атрибутами пов'язаними з ними

Аспект об'єкта має кілька назв: *feature*, *aspect*, *facet*. Замість «автора думки» (ортюп) часто використовується термін «джерело думки» (ортюп). Ці відмінності в термінології існують через специфічність текстів аналізованих в роботах дослідників і не змінюють суті визначення думки [11]-[15].

Наприклад, в повідомленні (рис. 1.2):

- «Twitter for iPhone» – об'єкт;
- «новий інтерфейс» – аспект об'єкта;
- «@mgerasimov» – володар думки,
- «11 Dec» – дата відправки повідомлення,
- «відверто розчарував» та «☹» – частини повідомлення, з яких випливає, що думка про аспект об'єкта негативно.

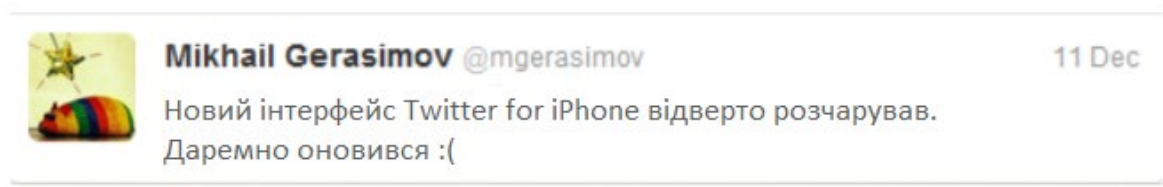


Рисунок 1.2 – Приклад емоційно забарвленого повідомлення, що містить думку першого типу

1.3.3 Порівняння

Другий тип думок-порівняння – можна розділити на три види [11]-[15]:

- порівняння аспектів об'єктів на користь одного;
- прирівнювання аспектів різних об'єктів;
- перевага одного об'єкта над іншими [11]-[15].

Порівняння першого типу мають вигляд «аспект об'єкта 1 перевершує в чомусь аспект об'єкта 2», наприклад: «екран Galaxy Tab зроблений більш якісно, ніж у конкурентів» [11]-[15]. Другий тип виражає схожість аспектів різних об'єктів, наприклад: «і Android, і iOS однаково зручні для розробки застосунків під них». Прикладом третього типу може послужити пропозиція «апарат від Canon виявився найдешевшим в тому магазині» [11]-[15].

Думка другого типу визначається як кортеж [11]-[15]:

$$(E_1, E_2, A, po, holder, time), \quad (1.1)$$

де E_1, E_2 – множина порівнюваних об'єктів по аспекту A ;

po – множина об'єктів, які автор ($holder$) вважав за краще;

$time$ – момент часу в який думка була опублікована.

Наприклад для повідомлення (рис. 1.3) кортеж думок виглядає так:
 $(\{Sony\ Xperia\}, \{Samsung\ Galaxy\}, \{загальний\}, \{Sony\ Xperia\}, @wakeupwhy, 10May)$



Рисунок 1.3 – Приклад повідомлення, що містить думку другого типу

На відміну від кортежу, що визначає думку першого типу, кортеж думки другого типу не містить оцінку емоцій автора [11]-[15].

1.3.4 Суб'єктивність

В аналізі тональності тексту часто зустрічається термін, пов'язаний з поняттям думки-суб'єктивність [11]-[15]. В дається визначення об'єктивного і суб'єктивного пропозицій [11]-[15]:

Об'єктивна пропозиція виражає фактичну інформацію про що-небудь, тоді як суб'єктивна пропозиція виражає чийсь особисті почуття і припущення.

Пропозиції, що містять думку, зазвичай є суб'єктивними, тому аналіз тексту на наявність суб'єктивної інформації часто є підзадачею визначення полярності тексту [11]-[15].

1.3.5 Завдання

У комп'ютерній лінгвістиці текст природною мовою вважається неструктурованою інформацією. Давши визначення думки, ми визначаємо те, як повинна бути структурована ця інформація [11]-[15]. У завданнях, об'єднаних терміном «аналіз емоційного забарвлення тексту», визначається те, яким чином з тексту природною мовою витягується, аналізується і структурується інформація [11]-[15].

Отже, аналіз тональності тексту включає в себе наступні основні завдання [11]-[15]:

- визначення полярності тексту;
- витяг аспектів з емоційно забарвленого тексту.

1.4 Визначення полярності тексту

Завдання визначення полярності тексту формулюється наступним чином: «визначити, емоційне забарвлення тексту позитивне чи негативне?» Визначення полярності тексту зазвичай розглядається на двох рівнях:

- на рівні документа;

– на рівні пропозиції.

1.4.1 Визначення полярності документа

Завдання полягає у визначенні полярності документа в цілому. Причому, текст документа може одночасно містити пропозиції, як з негативною, так і з позитивною емоційним забарвленням.

На рівні документа робиться припущення про те, що документ містить думку про один об'єкт. Огляди товарів і послуг на спеціалізованих ресурсах мережі Інтернет задовольняють цьому припущенню [11]-[15].

Завданню визначення полярності документа присвячено багато ранніх робіт з аналізу тональності тексту. У більшості сучасних систем аналізу думок, це завдання не розглядається, тому що результат - полярність документа в середньому – вважається неінформативною оцінкою думки, вираженої в документі [11]-[15].

1.4.2 Визначення полярності речень

Завдання визначення полярності пропозицій вирішується як в якості підзадачі аналізу тональності документа, так і в якості самостійного завдання, наприклад, при аналізі коротких повідомлень і коментарів в соціальних мережах [11]-[15].

Часто визначення полярності пропозиції відбувається в два етапи. Спочатку проводиться аналіз пропозиції на суб'єктивність. Якщо пропозиція містить інформацію суб'єктивного характеру, то, швидше за все, в ньому виражена думка. Далі у суб'єктивного пропозиції визначається полярність. В іншому випадку пропозиція містить фактичну інформацію і далі не розглядається.

Методи визначення полярності окремих пропозицій призначені для більш точного аналізу думки в тексті: спільно з методами вилучення аспектів

можна провести детальний аналіз думки автора по всіх аспектах товару, порушеним в тексті [11]-[15].

1.4.3 Визначення полярності як завдання класифікації

Часто на спеціалізованих ресурсах користувацька думка про товар супроводжується його оцінкою за заданою шкалою. Тому в більшості досліджень визначення полярності розглядається як завдання класифікації тексту.

Класами можуть бути оцінки, поставлені автором огляду, а ознаки витягуються з тексту думки: часто в якості ознак вибирають n-грами слів і їх частини мови. Для вирішення завдання класифікації використовуються різні методи машинного навчання.

Наприклад, кожне слово виступає бінарною ознакою - чи присутнє воно в розглянутому думці, за умови, що його поява незалежно від інших слів. В якості класів використовується оцінка фільму за п'ятибальною шкалою. У роботі порівнюються популярні алгоритми навчання з учителем: метод опорних векторів і наївний байєсівський класифікатор [11]-[15].

1.4.4 Словникові методи у визначенні полярності

Так само існують методи, засновані на використанні словника емоційно забарвлених слів і словника символів, що позначають емоції. У словникових методах кожне слово має вагу, що характеризує його емоційну забарвленість. Часто словники складаються за допомогою сторонніх інструментів, таких як WordNet, потім терміни словника зважуються, наприклад, вручну.

У дослідженнях, заснованих на словникових методах, полярність тексту визначається різними способами. У [11]-[15] текст володіє трьома незалежними оцінками: позитивною, негативною і суб'єктивною.

Полярність документа складається з полярностей його пропозицій. Визначення полярності пропозиції полягає в обчисленні нормалізованих сум ваг термінів в тексті: кожен термін має позитивний, негативний і суб'єктивний ваги за шкалою цілих чисел від 0 до 1.

У [11]-[15] короткі текстові повідомлення незалежно оцінюються за негативною і позитивною шкалою: від -5 до -1 і від 1 до 5 відповідно. Кожен термін словника має дві ваги, визначені цим же способом. Позитивна вага текстового повідомлення дорівнює вазі терміна з максимальною позитивною оцінкою. Якщо текст не містить термінів, що входять до словника, йому присвоюється мінімальна позитивна вага. Аналогічним чином визначається негативна вага повідомлення.

Основною проблемою словникових методів вважається процес складання словника: щоб отримати метод, що класифікує документ з високою точністю, терміни словника повинні мати вагу, адекватний предметної області документа. Наприклад, слово «великий» по відношенню до обсягу пам'яті жорсткого диска є позитивною характеристикою, але негативною по відношенню до розміру мобільного телефону [11]-[15].

1.5 Методи автоматичного вилучення аспектів

Вищеописані завдання зачіпають проблему визначення полярності думки, але не проблему ідентифікації об'єктів (інакше кажучи, аспектів, згідно з наведеним визначенням думки), про які воно висловлено.

На сьогоднішній день даній задачі вилучення аспектів з емоційно забарвленого тексту присвячено менше число робіт, ніж іншим завданням входять в аналіз тональності тексту.

Дослідження, розглянуті в даній дипломній роботі, присвячені витяганню аспектів з онлайн-оглядів товарів і послуг. У подібному огляді користувач залишає думку про різні аспекти заздалегідь заданого об'єкта.

При вирішенні завдання вилучення аспектів популярні методи навчання без вчителя і статистичні методи. Для цих методів не потрібні розмічені тренувальні дані, які на сьогоднішній день відсутні у вільному доступі, на відміну від, наприклад, лінгвістичних корпусів, призначених для вирішення класичних завдань комп'ютерної лінгвістики: визначення частин мови, лемматизації і т.д. [11]-[15].

1.5.1 Методи навчання без вчителя

Навчання без вчителя – розділ машинного навчання, в якому вивчаються завдання виявлення прихованих закономірностей і взаємозв'язків між об'єктами з нерозміченої вибірки даних. Навчання без вчителя можна протиставити методи навчання з учителем, коли для кожного об'єкта з навчальної вибірки задана правильна відповідь, і потрібно знайти залежність між відповідями і об'єктами [11]-[15].

Розглянемо один алгоритм з класу методів навчання без вчителя в застосуванні до задачі вилучення аспектів. В основі методів поширення (bootstrapping) лежить наступна ідея: за допомогою невеликої множини вручну визначених прикладів (seed examples) деякого класу ітеративно автоматично витягувати схожі текстові одиниці, поступово нарощуючи цю множину.

У дослідженні [11]-[15] кандидатами в терміни можуть бути тільки n -грами розміром від 1 до 3 слів, що містять тільки іменники, прикметники, дієслова і прислівники. Близькість n -грами і термінів (аспектів) з множини U_i визначена за допомогою RlogF метрики [11]:

$$RlogF(ng) = \log freq(ng, U_i) \cdot \frac{freq(ng U_i)}{freq(ng)}, \quad (1.2)$$

де $freq(ngU_i)$ – частота спільної зустрічальності n -грами ng , і термінів з U в рамках фрагмента тексту, що складається з фіксованого числа слів.

1.5.2 Статистичні методи

Завдання вилучення аспектів можна розглядати, як завдання вилучення термінів, часто вживаних авторами думок.

У [11]-[15] дослідники припускають, що термінами, що описують аспекти, можуть бути поодинокі іменники і словосполучення містять іменник, часто зустрічаються в думках про об'єкти одного і того ж типу [11]-[15]. З усіх n -грам слів, що задовольняють цій вимозі, виділяються ті, чия частота в корпусі більше одного відсотка [11]-[15].

Виділені n -грами, що складаються з двох і більше слів, проходять перевірку на компактність. Якщо n -грама компактна як мінімум в двох реченнях, то вона потрапляє в список аспектів [11]-[15].

Компактність визначається наступним чином [11]-[15]:

- нехай f – n -грама з n слів, s – пропозиція, що містить всі слова з f (можливо розташовані не поспіль);
- якщо відстань між будь-якими двома словами, суміжними в f в реченні s становить не більше ніж три слова, то f компактна в даному конкретному реченні [11]-[15].

Терміни, що складаються з одного слова, також проходять статистичний тест на чистоту. Відшукуються всі пропозиції, що містять термін [11]-[15]. Серед знайдених пропозицій підраховуються пропозиції, що не містять пройшли тест на компактність n -грами, в які входить цей термін. Якщо число таких пропозицій вище деякого експериментально визначеного порогу, то термін потрапляє в список аспектів [11]-[15].

Схожий статистичний метод виявлення термінів-аспектів, що складаються з двох і більше слів, використовується в роботі [11]-[15] і має назву C-value. Для всіх n -грами, що містять в собі тільки певні частини мови,

що входять в деяку безліч документів, обчислюється їх значимість, визначена формулою (1.3) [11]-[15].

$$C-value(t) = \begin{cases} \log_2(len(term)) \cdot freq(term), & \text{якщо } |e_terms| = 0 \\ \log_2(len(term)) \cdot \left(freq(term) - \frac{1}{|e_terms|} \sum_{elder \in e_terms} freq(elder) \right), & \text{інакше} \end{cases}, \quad (1.3)$$

де $term$ – n -грама;

e_terms – множина, що складається з усіх n -грам старшого порядку, що містять $term$;

$|e_terms|$ – потужність множини;

$elder$ – елемент цієї множини;

$len(term)$ – довжина терміна в символах.

Розглянемо приклад, що ілюструє роботу C-value методу [11]-[15]. Нехай в корпусі думок про стільникових телефонах біграма «retina display» зустрічається 8 разів, що містять її триграми «great retina display» і «retina display worse» зустрічаються 3 і 2 рази відповідно. Тоді згідно з формулою (1.3) [11]-[15]:

$$\begin{aligned} C-value(retina\ display) &= \log(18) \cdot \left(8 - \frac{1}{2}(2 + 3) \right) \approx 20 \\ C-value(great\ retina\ display) &= \log(18) \cdot 3 \approx 13 \\ C-value(retina\ display\ worse) &= \log(18) \cdot 2 \approx 8 \end{aligned}$$

Якщо експериментально встановлений поріг C-value для даного корпусу дорівнює 15, то тільки n -грама «retina display» потрапить в безліч термінів-аспектів [11]-[15].

1.6 Порівняння підходів машинного навчання для ШІ в рамках аналізу емоційної тональності повідомлень

Для вирішення завдання визначення полярності пропозицій і коротких повідомлень ефективні як алгоритми навчання з учителем, так і методи, засновані на словниках [11]-[15].

Проблемою навчання з учителем є складання тренувального корпусу з прикладами з предметної області, в якій буде використовуватися класифікатор [11]-[15]. Однак схожою проблемою володіють і словникові методи: ваги термінів словника, складеного для однієї предметної області, можуть виявитися неадекватними для іншої [11]-[15].

Порівняємо різні групи методів машинного навчання та словникових методів, а результати представимо у вигляді таблиці – табл. 1.1.

Таблиця 1.1 – Порівняння методів

	Навчання з учителем	Навчання без учителя	Словникові методи
База прикладів для навчання моделі	Потребує значної бази	Не залежить від вхідних даних	Потребує значної бази
Продуктивність методів	Високий рівень	Високий рівень	Високий рівень
Коло практичних завдань для методів	Високий рівень	Середній рівень	Низький рівень
Подальший розвиток	Перспективно	Перспективно	Не перспективно

Методи машинного навчання є більш перспективною галуззю для використання ШІ в задачах аналізу емоційної тональності повідомлень в соціальних мережах. Методи навчання з учителем так само достатньо жорстко прив'язані до бази прикладів, проте коло їх практичного використання значно ширше, а перспективи їх використання спостерігаються вже сьогодні.

1.7 Постановка завдання

Основним завданням, поставленим в даній роботі, є реалізація методу аналізу емоційно забарвлених повідомлень в соціальній мережі Твіттер. Рішення даного завдання можна розділити на наступні етапи:

- дослідження особливостей повідомлень в соціальній мережі Твіттер;
- реалізація методу визначення полярності з урахуванням цих особливостей і тестування його ефективності;
- реалізація методу вилучення аспектів з урахуванням цих особливостей і тестування його ефективності.

Завдання вилучення аспектів часто вирішується за допомогою методів навчання без вчителя і статистичними методами. Для збільшення ефективності цих методів використовуються лінгвістичні і частотні фільтри, що дозволяють відсівати слова, що не мають відношення до аспектів.

1.8 Висновки за розділом 1

Під час роботи над першим розділом було розглянуто проблемну область аналіз емоційної тональності. Було з'ясовано, що більш надійними та актуальними є використання методів штучного інтелекту, для збільшення практичних завдань технології. Використання таких підходів дозволить підвищити якість роботи методу, що буде розроблено та реалізовано чого його ПЗ.

2 МАТЕРІАЛИ ТА МЕТОДИ

В першому розділі було проаналізовано сучасне становище проблемної області аналізу емоційної тональності повідомлень в соціальних мережах. Було проаналізовано сучасні практики області.

У цьому розділі на малюнку 4 показаний основний поетапний процес аналізу тональності текстів.

На першому етапі проводиться робота з платформами соціальних мереж: на цьому етапі користувач вибирає джерело, з якого хоче витягти думку; наприклад, при виборі соціальних мереж користувач може вибирати різні онлайн-джерела, такі як Twitter.

Другий етап включає в себе збір даних: на цьому етапі користувачі починають з використання певних ключових слів або хеш-тегів для отримання бажаної інформації на основі своїх уподобань. Ця інформація має різні форми, такі як твіти, повідомлення, новини та тексти.

Третій етап полягає в попередній обробці витягнутої інформації. Вона обробляється для підготовки даних до наступного етапу. Цей етап включає в себе вилучення ознак (тобто граматичні структури та характеристики інтелектуального аналізу даних), токенизацію (тобто процес перетворення тексту в токени перед перетворенням його в вектори) та очищення (тобто повторне видалення букв, виправлення тексту та визначення мови).

Четвертим кроком є аналіз даних: на цьому етапі всі попередньо оброблені дані використовуються за призначенням, наприклад для визначення полярності та аналізу тональності текстів або частоти аналізу.

Помітно, що кожен з етапів аналізу тональності текстів має на увазі використання комп'ютерних технологій. У цьому розділі розглянуті основні методи, що будуть використовуватися в роботі для аналізу емоційної тональності повідомлень в соціальній мережі Twitter.

2.1 Дослідження особливостей повідомлень в Twitter

Перейдемо до вирішення першого завдання. У контексті завдання аналізу емоційного забарвлення тексту, основними особливостями повідомлень в Twitter (рис. 2.1) є [14], [15]:

- малий розмір повідомлення – 140 символів.
- наявність сленгу, скорочень і граматичних помилок.
- наявність спеціальних символів.
- використання посилань на інших користувачів і на зовнішні ресурси.

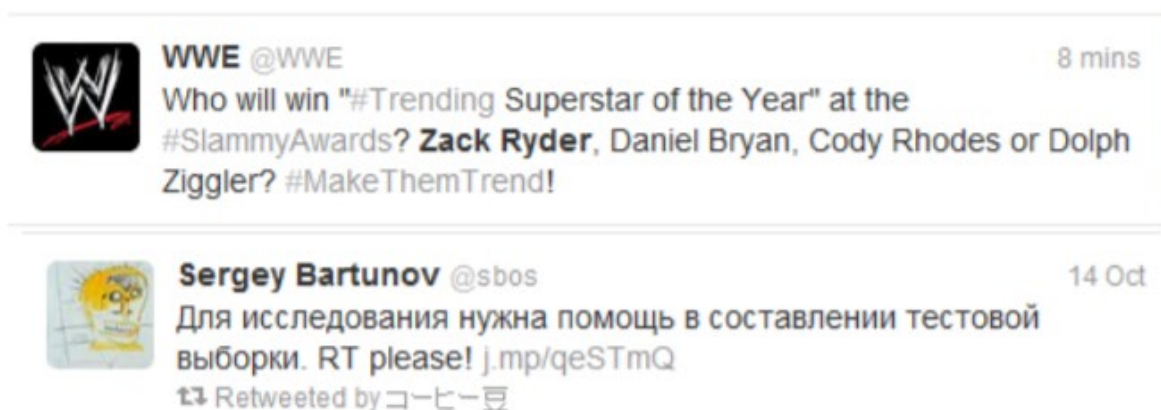


Рисунок 2.1 – Приклад повідомлені в мережі Twitter

У роботах ці особливості є вирішальним фактором при розробці методів. Наприклад, в роботі використовується алгоритм визначення полярності повідомлення, заснований на словниковому методі. У словник входять символи і слова, що позначають емоції. Навмисні граматичні помилки в словникових термінах, такі як повторення приголосних букв, в словах, наприклад: «ПЕРЕмога!», замість «перемога», збільшують початкову вагу термінів.

2.2 Реалізація комплексу для визначення полярності повідомлень

Для реалізації ефективного методу та ПЗ визначення полярності повідомлення було вирішено використовувати методи машинного навчання. Рішення даного завдання включає в себе наступні пункти:

- вибір заходів, за якими оцінюється ефективність алгоритмів;
- вибір ознак, за якими здійснюється класифікація;
- вибір алгоритму класифікації;
- вибір відповідної навчальної вибірки;
- тестування ефективності.

2.2.1 Вибір міри ефективності методу

Традиційно ефективність задачі класифікації тексту формулюється в термінах точності і повноти. Інтерпретуємо ці терміни для задачі визначення полярності документів – приклад в табл.. 2.1.

Таблиця 2.1 – Результати бінарної класифікації

	Класифіковано як «позитив»	Класифіковано як «негатив»
Клас «позитив»	TP	FN
Клас «негатив»	FP	TN

Нехай у колекції з N документів N_p документів мають позитивне емоційне забарвлення (належать до класу «позитив») і N_n документів – негативне (належать класу «негатив»).

В результаті класифікації цих документів, до класу «позитив» правильно віднесені TP документів, неправильно – FP , до класу –

правильно віднесено TN документів, неправильно – FN . Тоді, слід розглянути терміни відносно, щодо класу «позитив».

Точність – відношення числа правильно віднесених документів до класу «позитив» до числа всіх документів, віднесених до класу «позитив»:

$$Precision = \frac{TP}{TP + FP}. \quad (2.1)$$

Повнота – відношення числа правильно віднесених документів класу 1 до числа документів класу «позитив» в колекції:

позитив» до числа всіх документів, віднесених до класу «позитив»:

$$Recall = \frac{TP}{TP + FN}. \quad (2.2)$$

2.2.2 Вибір ознак

Метод визначення полярності, реалізований в даній роботі, не передбачає використання апіорних припущень про те, які слова або Символи можуть містити повідомлення, що належать до якого-небудь класу. Це означає, що всі ознаки апіорно рівнозначні [11].

Часто використовуваними ознаками при вирішенні завдання визначення полярності є n -грами слів. У даній роботі під словом мається на увазі будь-яка послідовність букв алфавіту, а під n -грамою порядку n – розділена пробілами послідовність з n слів.

Наприклад, повідомлення «Сьогодні йшов дощ. ;)» містить тільки наступні n -грами слів першого і другого порядку: «сьогодні», «йшов», «дощ», «сьогодні йшов» і «йшов дощ».

У ряді робіт в якості ознак використовуються частини мови. Це пояснюється тим, що думка містить суб'єктивну лексику. Наприклад, в роботі

[11-14] складається словник прикметників і прислівників, як термінів, що виражають емоцію. З цієї ж причини було так само вирішено вибрати в якості ознак n -грамм з частин мови.

Як було показано раніше, такі особливості повідомлень в Твіттер, як сленг і емотикони сигналізують. В якості ознак був обраний набір часто вживаних емотиконів.

2.3 Вибір методу класифікації

У ряді досліджень з визначення полярності тексту, високу ефективність показали методи навчання з учителем. Ці методи використовувалися як в ранніх роботах з визначення полярності документа в середньому, так і в сучасних роботах, де аналізується пропозиції і короткі текстові повідомлення.

Для вирішення поставленого завдання були обрані 2 методи, що показали себе найбільш ефективними при визначенні полярності коротких текстових повідомлень [11].

2.3.1 Метод опорних векторів

Метод опорних векторів (supported vector machine – SVM) відноситься до сімейства лінійних класифікаторів. Метою лінійної класифікації є пошук гіперплощини в просторі ознак, що розділяє всі об'єкти на два класи [16].

Основна ідея методу опорних векторів полягає в пошуку розділяє гіперплоскості, максимально віддаленої від найближчих до неї точок в просторі ознак [16].

У випадку лінійно розділеної вибірки пошук гіперплощини можна записати у вигляді задачі оптимізації [16]:

$$\frac{1}{2}\|\omega\|^2 \rightarrow \min_{\omega, b}, \quad (2.3)$$

$$y_i(\omega^T x_i + b) \geq 1, j = 1, \dots, m$$

де $\frac{1}{\|\omega\|}$ – величина зазору між гіперплощиною і найближчими до неї точками,

як першого, так і другого класу;

$y_i(\omega^T x_i + b)$ – добуток значення класу точки і її положення щодо гіперплощини.

Для більш загального випадку лінійно нероздільної вибірки алгоритм може допускати помилки на навчальних об'єктах. Нове завдання оптимізації включає в себе вимогу мінімізації помилки [16]:

$$\frac{1}{2}\|\omega\|^2 + C \sum_{i=1}^l e_i \rightarrow \min_{\omega, b}, \quad (2.4)$$

$$y_i(\omega^T x_i + b) \geq 1 - e_i, i = 1, \dots, k,$$

$$e_i \geq 0, i = 1, \dots, k$$

де змінна e_i – характеризує величину помилки на прикладі вибірки з k елементів;

C – константа, що дозволяє знаходити компроміс між максимізацією величини зазору і мінімізації сумарної помилки на тренувальній вибірці [16].

2.3.2 Наївний баєсів класифікатор

Наївний баєсів класифікатор – імовірнісний класифікатор, заснований на теоремі Байєса і (наївному) припущенні про статистичну незалежність випадкових величин (2.5) [17].

$$p(C | F_1, \dots, F_2) = \frac{p(C)p(F_1, \dots, F_2 | C)}{p(F_1, \dots, F_2)}. \quad (2.5)$$

Основна перевага даного класифікатора полягає в низькій обчислювальній складності, а також в оптимальності, за умови дійсної незалежності ознак [17].

2.4 Вибір навчальної вибірки

Для використання методів навчання з учителем потрібна навчальна вибірка. Зазвичай навчальне безліч складається з прикладів тієї області, в якій буде застосовуватися класифікатор.

В якості навчальної та перевірконої вибірки був складений корпус, що складається з 8000 пропозицій, для яких визначена полярність. Частина цих пропозицій була витягнута з розміченого корпусу, наданого авторами [18] для вільного доступу. Інша частина була отримана за допомогою онлайнової системи Sentiment140 [19] аналізу емоційного забарвлення повідомлень соціальної мережі Твіттер.

Всі приклади отриманого навчального безлічі отримані з думок про електронну техніку, а саме про мобільних телефонах, планшетах, плеєрах.

2.5 Висновки за розділом 2

У другому розділі проаналізовано запропонований методи аналізу емоційної тональності повідомлень у соціальних мережах.

Комплекс аналізу емоційного забарвлення повідомлень соціальної мережі Twitter був розроблений в рамках системи дослідження соціальних мереж. Одним з модулів цієї системи має стати модуль емоційного забарвлення текстових повідомлень.

3 РОЗРОБКА АРХІТЕКТУРИ ПРОГРАМИ

Другий розділ було присвячено дослідженню запропонованих методів аналізу емоційної тональності повідомлень у соціальних мережах.

Третій розділ буде присвячено висвітленню питань з проектування архітектури та реалізації ПЗ для дослідження запропонованих методів та можливих модифікацій. Особливу увагу буде присвячено вибору технічних засобів: мови програмування, середовища розробки, тощо.

3.1 Опис основних вимог до програми

ПЗ для аналізу емоційної тональності повідомлень в соціальній мережі призначено для аналізу, прийняття рішень та подальшого використання інформації про емоційне забарвлення повідомлень у соціальній мережі.

Таким чином серед функціональних вимог виділяються:

- завантаження вибірки;
- виконання попередньої обробки вибірки;
- реалізація методу визначення полярності з урахуванням особливостей вибірки та соціальної мережі і тестування його ефективності.
- реалізація методу вилучення аспектів з урахуванням особливостей вибірки та соціальної мережі і тестування його ефективності.

Виконання наведених функцій в повній мірі розкриє практичний потенціал запропонованих методів та розроблюваного ПЗ.

3.2 Вибір мови програмування

В області ШІ і машинного навчання застосовується безліч мов програмування [20], [21]. Але Java, незважаючи на свою значну історію

(більше 25 років), все ще залишається одним з найпопулярніших в цій сфері [20], [21].

Його успішно використовують також, програмуючи нейронні мережі, алгоритми пошуку, системи робототехніки [20], [21]. Для цілей створення алгоритмів машинного навчання Java має дві основні переваги [20], [21].

Це найвища масштабованість, що означає можливість оперувати величезними масивами даних і вести обробку декількох запитів паралельно, і об'єктно-орієнтованість, що задає відповідну для вирішення завдань машинного навчання логіку програмування [20], [21].

На користь йде і можливість застосунків, створених на Java, запускатися на будь-якій платформі. Мова дозволяє кодувати алгоритми різних типів; код, написаний на ньому, простий для подальшої налагодження сторонніми розробниками [20], [21].

Набір інструментів Standard Widget Toolkit дозволяє створювати зручні користувальницькі інтерфейси для роботи з додатками машинного навчання на Java [20], [21].

Java використовується для реалізації самих різних проектів машинного навчання, включаючи [20], [21]:

- бібліотеки штучного інтелекту;
- експертні системи, що емулюють здатність людини до прийняття рішень;
- нейронні мережі;
- засоби обробки природної мови – розпізнавачі і обробники мови, рукописних текстів, програми для читання по губах і т. д.

Основними особливостями Java для вирішення завдань машинного навчання є наявність вузькоспеціалізованих інструментів для вирішення конкретних завдань-бібліотек і фреймворків, робота з якими вимагає високих навичок програмування і постійного розвитку [20], [21]. Наприклад, до таких можна віднести популярну бібліотеку Weka, що містить безліч готових алгоритмів машинного навчання в області аналізу даних [20], [21].

Ця бібліотека надає розробникам можливості для вирішення завдань класифікації, регресійного аналізу, кластеризації даних [20], [21]. У бібліотеці реалізовані такі алгоритми навчання, як метод опорних векторів, дерева рішень, логістична регресія, алгоритми вибору ознак, фільтруючі методи і т. д. [20], [21].

Для вирішення завдань машинного навчання в сфері аналізу природної мови, візуалізації даних та інших застосовується фреймворк Smile, для роботи з даними в реляційній базі даних застосовується середовище для KDD-додатків ELKI, а завдання кластеризації і вибору функцій дозволяє вирішувати бібліотека Java-ML [20], [21].

У таблиці 3.1 наведено порівняння обраної мови програмування із іншими найпопулярнішими мовами програмування.

Як видно з таблиці не дивлячись на те, що мова програмування Java виникла вже достатньо давно, вона і зараз достатньо популярна та активно продовжує розвиватися. Більш того саме вона найбільш вдалий вибір, якщо майбутнє ПЗ повинне бути кроссплатформним.

Таблиця 3.1 – Порівняння мов програмування

Критерії порівняння	Мови програмування			
	Java	C++	C#	Python
Наявність літератури та простота вивчення	++	++	++	++
Підтримка парадигми модульного програмування	+	+	+	+
Стандартні бібліотеки шаблонів	+	+	+	+
Кроссплатформність	++	+	±	+
Продуктивність роботи	++	++	+	+
Узагальнення алгоритмів	+	+	+	+
Масштабованість	++	+	±	—
Паралельне програмування	++	++	++	+
Простота розроблення графічного інтерфейсу	+	±	++	+

3.3 Бібліотека OpenNLP

NLP являє собою набір інструментів, використовуваних для отримання значущої і корисної інформації з джерел на природній мові, таких як веб-сторінки і текстові документи [22].

Apache OpenNLP – це бібліотека Java з відкритим кодом, яка використовується для обробки тексту природною мовою. За допомогою цієї бібліотеки ви можете створити ефективний сервіс обробки тексту.

OpenNLP надає такі сервіси, як токенізація, сегментація пропозицій, тегування частини мови, виділення іменованих об'єктів, розбиття на фрагменти, аналіз і дозвіл супутніх посилань і т.д. Нижче наведені помітні особливості OpenNLP [22].

Розпізнавання іменованих об'єктів (NER) – OpenNLP підтримує NER, використовуючи який ви можете витягувати імена місцеположень, людей і речей навіть при обробці запитів.

Підведення підсумків – використовуючи функцію Підведення підсумків, ви можете підсумувати абзаци, статті, документи або їх колекції в NLP.

Пошук – в OpenNLP задана рядок пошуку або її синоніми можуть бути ідентифіковані в даному тексті, навіть якщо дане слово змінено або написано з помилкою [22].

Позначка (POS) – позначка в NLP використовується для поділу тексту на різні граматичні елементи для подальшого аналізу.

Переклад – в NLP переклад допомагає в перекладі з однієї мови на іншу.

Групування інформації – ця опція в NLP групує текстову інформацію в змісті документа, як частини мови.

Генерація природної мови – використовується для генерування інформації з бази даних і автоматизації інформаційних звітів, таких як аналіз погоди або медичні звіти [22].

Аналіз зворотного зв'язку. Як випливає з назви, NLP збирає різні види відгуків людей про продукти, щоб проаналізувати, наскільки продукт успішний в завоюванні їх сердець.

Розпізнавання мови. Хоча аналіз людської мови ускладнений, в NLP є деякі вбудовані функції для цієї вимоги.

Переклад – в NLP переклад допомагає в перекладі з однієї мови на іншу [22].

Групування інформації – ця опція в NLP групує текстову інформацію в змісті документа, як частини мови.

Open NLP API. Бібліотека Apache OpenNLP надає класи та інтерфейси для виконання різних завдань обробки природної мови, таких як виявлення пропозицій, токенизація, пошук імені, маркування частин мови, розбиття пропозиції на частини, аналіз, зіставлення посилань і категоризація документів.

На додаток до цих завдань можливо навчати і оцінювати власні моделі для будь-якого з цих завдань [22].

3.4 Вибір середовища розробки

У червні 2000 року компанія Sun Microsystems зробила NetBeans відкритим вихідним кодом, яка залишалася спонсором проекту до січня 2010 року, коли Sun Microsystems стала дочірньою компанією Oracle [23].

Протягом всієї своєї історії в Sun Microsystems і Oracle NetBeans була безкоштовною і з відкритим вихідним кодом, і її спонсор використовував її в якості механізму для просування екосистеми Java вперед.

У 2016 році Oracle передала вихідний код NetBeans до Фонду програмного забезпечення Apache. У квітні 2019 року Apache NetBeans став проектом Apache вищого рівня. Будь ласка, ознайомтеся з нашим Розділом історії для отримання додаткової інформації [23].

Apache NetBeans – це проект Apache вищого рівня, присвячений надання надійних продуктів для розробки програмного забезпечення (IDE Apache NetBeans і платформа Apache NetBeans), які задовольняють потреби розробників, користувачів та підприємств, які покладаються на NetBeans в якості основи для своїх продуктів; зокрема, щоб вони могли швидко, ефективно і легко розробляти ці продукти, використовуючи сильні сторони платформи Java та інших відповідних галузевих стандартів [23].

Два основні продукти, IDE Apache NetBeans та платформа Apache NetBeans, є безкоштовними для комерційного та некомерційного використання відповідно до ліцензії Apache. Вихідний код для обох доступний будь-якому користувачеві для повторного використання на свій розсуд відповідно до Умов використання.

Проект Apache NetBeans також є активним співтовариством, в якому люди з усього світу можуть задавати питання, давати поради, вносити свій внесок і в кінцевому підсумку брати участь в успіху наших продуктів. У списках розсилки та форумах NetBeans ви знайдете повідомлення від студентів, розробників з провідних компаній та приватних осіб, які бажають розширити свої навички [23].

Завдяки більш ніж 18 мільйонам завантажень IDE NetBeans на сьогоднішній день і більш ніж 800 000 розробників, що беруть участь, проект NetBeans процвітає і продовжує зростати завдяки приватним особам і компаніям-партнерам. Щоб стати частиною спільноти, просто зареєструйтеся безкоштовно.

Більшість розробників визнають IDE NetBeans оригінальним безкоштовним середовищем розробки Java. Саме це і багато іншого! Середовище IDE Apache NetBeans забезпечує підтримку декількох мов (Java, PHP, JavaFX, JavaScript і т.д.) і фреймворків, і найближчим часом будуть включені інші (C/C++) [23].

Платформа Apache NetBeans – це універсальна платформа для застосунків Swing. Вона забезпечує техніку, яку раніше кожен розробник

повинен був писати самостійно – збереження стану, підключення дій до пунктів меню, елементів панелі інструментів і сполученням клавіш; управління вікнами і так далі [23].

У таблиці 3.2 наведено порівняння обраного середовища розробки для мови програмування Java з іншими найпопулярнішими рішеннями.

Таблиця 3.2 – Порівняння середовищ розробки

Критерії порівняння	Середовища розробки		
	NetBeans	Eclipse	IntelliJ IDEA
Кроссплатформність	±	++	+
Швидкість модернізації	±	±	++
Наявність генератора сценаріїв складання	+	+	+
Підтримка репозиторіїв	++	+	++
Підтримка серверів	++	+	++
Розроблення графічного інтерфейсу користувача	++	–	+
Доступність (вимоги до розповсюдження)	++	+	–

Як видно із таблиці порівняння більша кількість балів у NetBeans та популярного середовища IntelliJ IDEA, проте IntelliJ IDEA розповсюджується за окрему платню, у той час, як NetBeans офіційно безкоштовний. З іншого боку NetBeans працює із більш старими версіями Java та JDK, у той час, як нові версії Eclipse працюють із найактуальнішими версіями Java та JDK. Проте слід відзначити достатньо велике число помилок та збоїв під час роботи із останніми версіями Eclipse, а більш старі та стабільні версії із кожним разом завантажити все важче. Більш того розробка графічного інтерфейсу користувача у середовищі Eclipse достатньо важка та проблематична. Тож можна зробити висновок, що NetBeans – найбільш оптимальний варіант для подальшої розробки.

3.5 Вимоги до технічних засобів

Вимоги до складу і параметрів технічних засобів

Для нормальної роботи програми необхідні комп'ютер і мікрофон.

Технічна характеристика:

- персональний комп'ютер (ПК) з операційною системою (ОС)

Microsoft Windows 8.1 та старше;

- Java Virtual Machine (JVM) 9.0 та вище;

– процесор – Intel Core з тактовою частотою 2.0 МГц та вище, ядер: 2, логічних процесорів: 2.

3.6 Висновки за розділом 3

Під час роботи над розділом було спроектовано архітектуру майбутнього ПП. Були спроектовані модулі майбутньої програми. Було проаналізовано найбільш популярні мови програмування та після ретельного порівняння перевага була віддана мові Java. Також проаналізувавши найбільш популярні середовища розробки, було обрано середовище NetBeans.

4 ОСНОВНІ РІШЕННЯ ЩОДО РЕАЛІЗАЦІЇ КОМПОНЕНТІВ СИСТЕМИ

У третьому розділі було досліджено та вирішено ряд питань з проектування архітектури та реалізації ПЗ для дослідження запропонованого методу та можливих модифікацій. Особливу увагу було присвячено вибору технічних засобів: мови програмування та додаткових бібліотек.

Метою четвертого розділу є висвітлення рішень, щодо розробки ПЗ аналізу емоційного тону повідомлень у соціальних мережах. Метод заснований на апараті SVM та наївному баєсівському класифікаторі.

У розділі описаний поетапний процес розробки програми в середовищі NetBeans. Описана технологія розробки з використанням Java та бібліотеки OpenNLP.

4.1 Загальне представлення розробки

Алгоритм аналізу емоційного забарвлення повідомлень соціальної мережі Twitter був розроблений в рамках системи дослідження соціальних мереж. Одним з модулів цієї системи є модуль емоційного забарвлення текстових повідомлень.

В якості мови реалізації рішення задачі був обраний Java, оскільки вся реалізована система розробляється на ньому. Java вибрано з наступних причин:

- розв'язувані завдання не пред'являють високих вимог до продуктивності.
- Java – кроссплатформна об'єктно-орієнтована мова.
- цією мовою реалізовано велику кількість бібліотек, призначених для вирішення завдань машинного навчання та комп'ютерної лінгвістики.

Для вирішення завдань пов'язаних з попередньою обробкою тексту: визначення речень, слів, частин мови – була використана бібліотека

OpenNLP. Для вирішення задачі визначення полярності тексту використана бібліотека LIBLINEAR. Ліцензії обох бібліотек допускає їх вільне використання в наукових і комерційних проектах.

4.2 Взаємодія модулів програми

Перейдемо до загальної схеми роботи системи. На малюнку стрілками зазначено, в якому напрямку і між якими компонентами передаються дані (рис. 4.1). NLP-модуль здійснює визначення частин мови і розбиття на n -грами вхідного тексту. На етапі вилучення ознак формується вектор ознак з тексту. Результат бінарної класифікації-полярність тексту, приймає значення «1» і «-1».



Рисунок 4.1 – Загальна схема роботи

Модуль вилучення аспектів повертає всі n -грами тексту, що пройшли через частотні і лінгвістичні фільтри, разом з позначкою про приналежність n -грами до безлічі аспектів.

4.3 Об'єктно-орієнтовна розробка представлених модулів

Архітектура модулів розроблених алгоритмів виглядає наступним чином. Класи, в яких реалізовані SVM і наївний баєсівський класифікатор мають спільного пращура Classifier. (рис. 4.2).

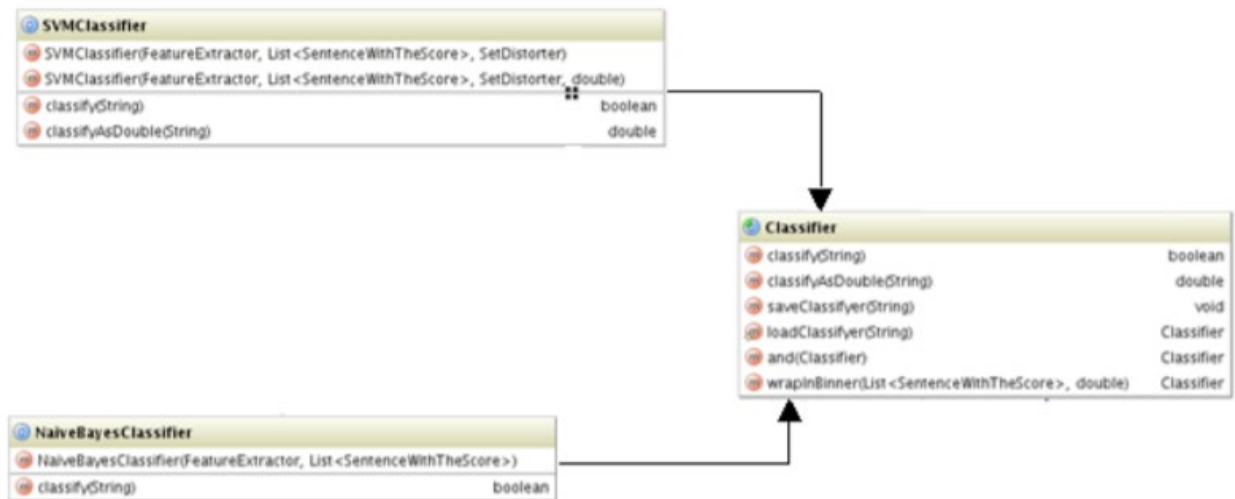


Рисунок 4.2 – Реалізація методів класифікації

Обидва класифікатори приймають на вхід об'єкт класу Feature Extractor, показаного на рис. 4.3. Цей клас призначений для генерації векторів ознак. Від нього успадковуються класи, які мають різні ознаки.

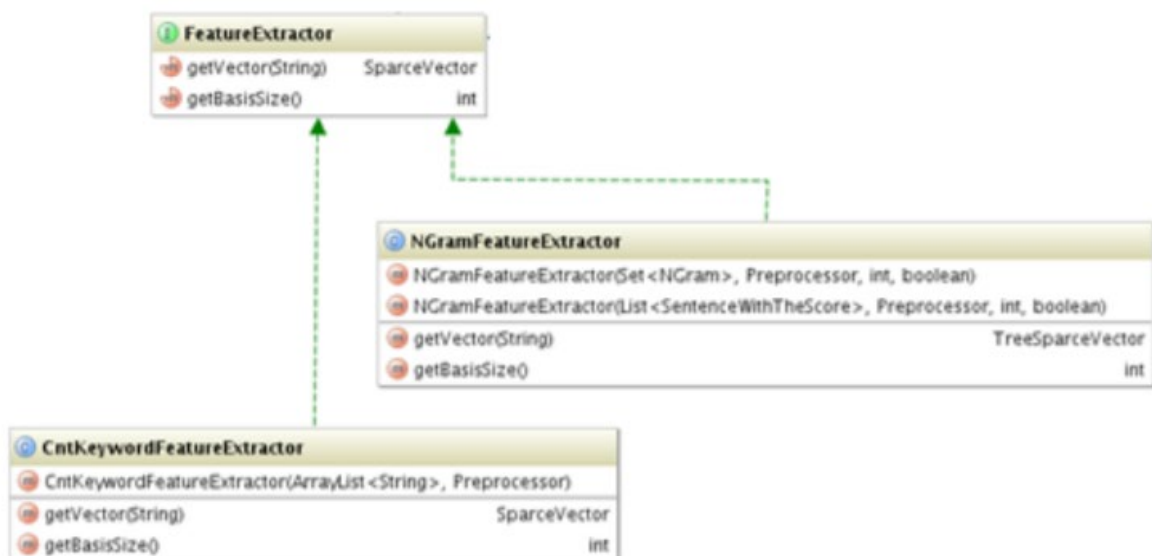


Рисунок 4.3 – Реалізація вилучення ознак

Модуль вилучення аспектів реалізований в окремому модулі. На рис. 4.4 зображені класи входять в нього класи. Процесом обробки поданого на вхід тексту займається клас `TermCandidatesExtractor`, йому підпорядковуються класи `TextPreProcessor`, в якому визначаються речення, слова в них і частини мови, і клас `CvaluesEvaluator` фільтруючий n -грами.

Інші класи призначені для побудови Індексу термінів у документі. Індекс використовується при тренуванні алгоритму поширення

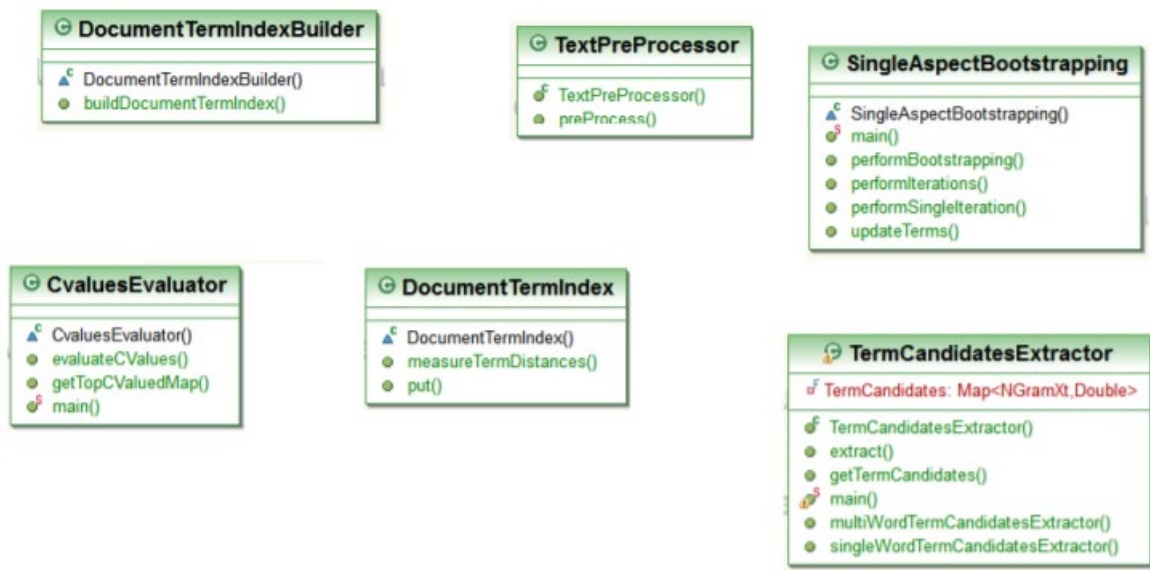


Рисунок 4.4 – Модуль вилучення аспектів

4.4 Бібліотека LIBLINEAR

LIBLINEAR – це лінійний класифікатор даних з мільйонами екземплярів і об'єктів. Він підтримує

- L2-регуляризовані класифікатори;
- L2-лінійний SVM з втратами, L1-лінійний SVM з втратами і логістична регресія;
- L1-регуляризовані класифікатори (після версії 1.4);
- L2-лінійна SVM з втратами і логістична регресія;
- L2-регресія регуляризованого опорного вектора (після версії 1.9);

- L2-лінійний SVR з втратами і L1-лінійний SVR з втратами;
- L2-регуляризовані машини опорних векторів одного класу (після версії 2.40);

Основні особливості LIBLINEAR включають в себе:

- той же формат даних, що і LIBSVM, наш універсальний вирішувач SVM, а також аналогічне використання;
- класифікація за кількома класами;
- перехресна перевірка для оцінки моделі;
- автоматичний вибір параметрів;
- оцінки ймовірності (тільки логістична регресія);
- ваги для незбалансованих даних;
- інтерфейси MATLAB / Octave, Java, Python, Ruby.

4.5 Експлуатаційні характеристики ПЗ

Кількісні характеристики реалізації:

- час повної роботи на етапі класифікації - близько 1 секунди для 4000 повідомлень на комп'ютері Intel Core 2 Duo E8400, 3.00 ГГц, 4 Гб ОЗУ;
- складність реалізованих алгоритмів: тимчасова складність методу визначення полярності: $O(m)$, де m – число ознак в повідомлення. Тимчасова складність алгоритму вилучення аспектів – $O(n)$, де n – довжина входу.

4.6 Висновки за розділом 4

Розроблено програмну систему для аналізу емоційної тональності повідомлень в соціальних мережах за допомогою запропонованого методу, яка повністю реалізує поставлену задачу. Обрано механізм реалізації відповідних методів та алгоритмів, а також елементи керування, з якими користувачу буде зручно використовувати розроблений програмний продукт.

5 ЕКСПЛУАТАЦІЯ, ТЕСТУВАННЯ ТА ЕКСПЕРИМЕНТАЛЬНЕ ДОСЛІДЖЕННЯ ПРОГРАМИ

Метою п'ятого розділу є перевірка роботи розробленої програми розпізнавання диктора по мовному пароллю.

У цьому розділі буде надано опис стратегії тестування та результати тестування ефективності розробленого ПЗ.

5.1 Опис стратегії тестування

Для тестування алгоритму визначення полярності повідомлень використовувався метод крос-валідації. Процедура крос-валідації відбувається наступним чином:

- фіксується безліч розбиття навчальної множини на тренувальне і контрольне підмножини.
- для кожного розбиття відбувається навчання алгоритму на тренувальному безлічі, потім тестування на контрольному.
- результатом крос-валідації алгоритму є середнє значення проведених результатів тестування на контрольній множині.

У даній роботі розбиття на множини проводилося випадковим чином. Попадання кожного речення в одне з двох множин рівноймовірно.

5.2 Результати тестування

У табл. 5.1 наведено результат тестування алгоритмів, в якості ознак обрані n -грами, в які входять слова і емотикони.

У наступній табл. 5.2 наведено результат тестування, де в якості ознак обрані частини мови і біграми частин мови.

Таблиця 5.1 – Полярність повідомлення: в якості ознак обрані n -грами слів

Метод	Юніграми		Біграми		Разом	
	Точність	Повнота	Точність	Повнота	Точність	Повнота
Баєсівський класифікатор	0,73	0,7	0,72	0,68	0,72	0,71
SVM	0,81	0,74	0,76	0,67	0,81	0,72

Таблиця 5.2 – Полярність повідомлення: в якості ознак обрані n -грами частин мови

Метод	Частина мови		Біграми частини мови	
	Точність	Повнота	Точність	Повнота
Баєсівський класифікатор	0,6	0,55	0,54	0,48
SVM	0,65	0,57	0,56	0,53

Найбільш ефективною конфігурацією в термінах точності і повноти виявився SVM, навчений на юніграммах слів і емотиконів.

5.3 Реалізація методу автоматичного вилучення аспектів

Для реалізації алгоритму потрібно виконати наступні підзадачі:

- вибір заходів, за якими оцінюється ефективність методу;
- вибір алгоритму вилучення аспектів;
- вибір навчальної вибірки;
- тестування ефективності алгоритму.

5.3.1 Вибір міри ефективності методів

Ефективність методів вилучення аспектів формулюється в термінах точності і повноти.

У контексті розв'язуваної задачі ці метрики мають наступний сенс. Алгоритм вилучення аспектів перевіряє кожен термін документа на приналежність безлічі аспектів. Тоді точністю цього алгоритму називається відношення числа правильно визначених аспектів до числа всіх термінів, віднесених до класу аспектів, а повнотою – відношення числа правильно визначених аспектів до числа аспектів в документі.

5.4 Вибір методу вилучення аспектів

Для вирішення цього завдання було вирішено реалізувати алгоритм, заснований на методі поширення, описаний в [24].

Для поліпшення ефективності алгоритму так само використовуються наведені в огляді статистичні методи: кандидатами в терміни можуть бути n -грами, що містять тільки іменники і прикметники і пройшли через C-value фільтр і частотний фільтр.

5.4.1 Навчальна вибірка

Метод поширення заснований на витяганні термінів з пов'язаних між собою пропозицій, але повідомлення соціальної мережі Твіттер, як правило, незалежні один від одного. Тому в якості тренувального безлічі використовується безліч повнотекстових оглядів електронної техніки, складене вручну.

Процес складання тренувальної вибірки спрощує той факт, що методи навчання з учителем не вимагають створення вибірки пар виду (приклад, клас), як у випадку навчання з учителем.

5.4.2 Тестування ефективності

Для тестування використовуються пропозиції, взяті з розміченого корпусу, наданого авторами роботи [15].

З корпусу взяті 1500 пропозицій з огляду двох цифрових фотоапаратів і 1820 пропозицій з огляду mp3 плеєра. У кожному з цих пропозицій виділені аспекти. Результати тестування реалізованого алгоритму показані в табл. 5.3.

Таблиця 5.3 – Вилучення аспектів

Домен	Точність	Повнота
Фотоапарати	0,65	0,76
Цифрові плеєри	0,62	0,71

5.5 Висновки за результатами тестування

Проведено експериментальну оцінку ефективності методів. Метод визначення полярності показав результати на рівні сучасних робіт. Розроблений ПЗ в повній мірі реалізував запропонований метод.

ВИСНОВКИ

Під час виконання дипломної кваліфікаційної роботи магістра, було розглянуто основні методи аналізу емоціональної тональності повідомлень в соціальних мережах:

- словникові методи;
- навчання з учителем;
- навчання без учителя.

В рамках роботи були вирішені наступні завдання:

- проведено огляд існуючих методів автоматичного аналізу емоційного забарвлення текстів;
- проведено дослідження текстових особливостей повідомлень в соціальних мережах в контексті розробки методів аналізу їх емоційного забарвлення;
- розроблено метод автоматичного аналізу емоційно забарвлених повідомлень у соціальній мережі Twitter;
- проведено експериментальну оцінку ефективності методів. Метод визначення полярності показав результати на рівні сучасних робіт.

Таким чином, в процесі підготовки дипломної кваліфікаційної роботи магістра була досягнута основна мета роботи і вирішені поставлені завдання

Під час розробки програми та аналізу розробленого програмного комплексу з'явилися ідеї з удосконалення, такі як: використання рекурентних нейронних мереж.

Нові можливості зможуть розширити коло вирішення практичних задач, адже останні роботи засвідчили високу точність роботи таких нейронних мереж в задачах розпізнавання, прогнозування та аналізу тексту та природньої мови.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Лінгвістичний аналіз тексту: практикум [Текст] : навч. посіб. / Алексєнко Н. М. [та ін.] ; Кременчуц. нац. ун-т ім. Михайла Остроградського. – Кременчук : КрНУ ім. Михайла Остроградського, 2019. – 241 с.
2. Чистяк Д.О. Інтертекстуальний аналіз художнього тексту [Текст] : навч. посіб. / Дмитро Чистяк ; [наук. ред. Я. В. Шекера]. – Київ : Саміт-Книга, 2020. – 135, [2] с.
3. Дарчук Н. П. Лінгвістичне забезпечення автоматичних систем аналізу українськомовного тексту (на прикладі системи автоматичного граматичного аналізу тексту АГАТ) [Текст] : автореф. дис. ... д-ра філол. наук : 10.02.01, 10.02.21 / Н. П. Дарчук; Київ. нац. ун-т ім. Тараса Шевченка, Ін-т філології. – Київ, 2015. – 34 с.
4. Орел А. В. Аналіз соціальних медіа для визначення оціночних суджень на основі аналізу тональності тексту [Текст] / А. В. Орел // Міжнародний науковий журнал "Інтернаука" . – 2018. – № 10(2). – С. 36-43.
5. Пасічник В. В. Математична лінгвістика : навч. посіб. Кн. 1. Квантитативна лінгвістика / В. В. Пасічник, Ю. М. Щербина, В. А. Висоцька, Т. В. Шестакевич; МОНМС України. – Л., 2012. – 359 с.
6. Актуальні задачі сучасних технологій : зб. тез доп. IV Міжнар. наук.-техн. конф. молодих учених та студентів, 25 – 26 листоп. 2015 р., Тернопіль. Т. 2 / Терноп. нац. техн. ун-т ім. І. Пулюя, НАН України, Ун-т ім. П. і М. Кюрі., Марібор. ун-т, Ягелон. ун-т. – Тернопіль, 2015. – 279 с.
7. Уголькова О. З. Цифровий маркетинг та соціальні мережі [Текст] / О. З. Уголькова // Management and entrepreneurship in Ukraine: the stages of formation and problems of development. – 2021. – Vol. 3, numb. 1. – С. 146-152.
8. Гоцур О. Соціальні мережі і політичний PR в українських та закордонних виборчих кампаніях [Текст] / О. Гоцур // Вісник Національного

університету "Львівська політехніка". Журналістика. – 2021. – № 2. – С. 53-57.

9. Ярмолюк О. Я. Соціальні мережі як сучасний інструмент просування [Текст] / О. Я. Ярмолюк, Ю. В. Фісун, А. А. Шаповалова // Підприємництво та інновації. – 2020. – Вип. 11(2). – С. 62-65.

10. Михайлова О. Ю. Соціальні мережі як чинник змін політичної культури та політичних практик в Україні: ризики і перспективи [Текст] / О. Ю. Михайлова // Стратегічна панорама. – 2019. – № 1-2. – С. 93-99.

11. Thelen M. A bootstrapping method for learning semantic lexicons using extraction pattern contexts / M. Thelen, E. Riloff// Proceedings of the ACL-02 conference on Empirical methods in natural language processing (EMNLP). – Morristown: NJ, USA. – 2002. P. 214-221.

12. Шелехов І. В. Інформаційно-екстремальне машинне навчання системи контролю знань [Текст] / І. В. Шелехов, С. О. Пилипенко, О. О. Столярчук, Т. А. Романенко // Вісник Національного технічного університету "ХПІ". Серія : Системний аналіз, управління та інформаційні технології. – 2018. – № 44. – С. 49-56.

13. Гороховський С. С. Машинне навчання та доповнена реальність на пристроях на базі iOS із фреймворком MLARKit [Текст] / С. С. Гороховський, О. О. Франків // Наукові записки НаУКМА. Комп'ютерні науки. – 2020. – Т. 3. – С. 4-6.

14. Куліш П. Ю. Потенціал соціальних мереж Facebook та Twitter як сучасний метод електронної взаємодії влади та громадськості [Текст] / П. Ю. Куліш // Теорія та практика державного управління. – 2012. – Вип. 2. – С. 132-137.

15. Яровий Д. Informational aspect of the Twitter-revolution 2010 in Kyrgyzstan [Текст] / Д. Яровий // Актуальні проблеми міжнародних відносин. – 2011. – Вип. 98(2). – С. 104-106

16. Boltenkov V. O. Information Technology for Automated Assessment of the Artillery Barrels Wear Based on Svm Classifier [Текст] / V. O. Boltenkov,

M. V. Maksymov, Y. V. Dobrynin // Applied aspects of information technology. – 2020. – Vol. 3, no 3. – С. 117–132.

17. Назіров Е. К. Методи збору та класифікації даних за допомогою саундлетної байєсівської нейронної мережі [Текст] / Е. К. Назіров, Т. О. Назірова, М. Ю. Карпенко // Вісник Херсонського національного технічного університету . – 2018. – № 3(1). – С. 332-337.

18. Hu M. Mining and summarizing customer reviews / M. Hu, B. Liu // Proceedings of the ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '04)ACM. – New York: NY. – 2004. – P. 168-177.

19. Sentiment 140 [Electronic resource]. – Access mode: <http://www.sentiment140.com/>.

20. Заковоротный А. Ю. Модель накопления уязвимостей в исходном коде Java-приложений, а также бинарная система оценки качества кода на ее основе [Текст] / А. Ю. Заковоротный, М. А. Зорян // Вісник Національного технічного університету "ХПІ". Серія : Інформатика та моделювання. – 2018. – № 24. – С. 37-46.

21. Яшина О. М. Дослідження мікросервісної архітектури, архітектурний стиль rest та їх сучасна реалізація на java [Текст] / О. М. Яшина, О. А. Кравчук // Вісник Хмельницького національного університету. Технічні науки. – 2020. – № 5. – С. 106-114.

22. Apache [Electronic resource]. – Access mode: OpenNLP <https://opennlp.apache.org/>.

23. Apache NetBeans [Electronic resource]. – Access mode: <https://netbeans.apache.org/>.

24. Aspect-Based Opinion Polling from Customer Reviews / J. Zhu, H. Wang, M. Zhu, B. K Tsou, M. Ma // IEEE Transactions on affective computing. – Vol. 2, no. 1. – 2011. – P.54-67.

ДОДАТОК А
Технічне завдання

A.1 Вступ

A.1.2 Призначення та область застосування

Програма аналізу емоційної тональності повідомлень в соціальних мережах на прикладі повідомлень Twitter, для аналізу повідомлень та подальшого використання даних для комерції. Існуюча програма реалізована з використанням мови програмування Java, та бібліотеки OpenNLP.

A.2 Підстави для розробки

Підставою для розробки є завдання на магістерську роботу, яке було визначене за темою «Дослідження та програмна реалізація методів аналізу емоційної тональності повідомлень у соціальних мережах» і затверджене наказом № 435 від 05 листопада 2021 р. за Національним університетом «Запорізька політехніка».

A.3 Призначення розробки

Програма призначена для аналізу емоційної тональності повідомлень в соціальних мережах.

A.4 Вимоги до програми

Програма повинна вирішувати наступні завдання:

- завантаження вибірки;
- виконання попередньої обробки вибірки;
- реалізація методу визначення полярності з урахуванням особливостей вибірки та соціальної мережі і тестування його ефективності.
- реалізація методу вилучення аспектів з урахуванням особливостей вибірки та соціальної мережі і тестування його ефективності.

Виконання наведених функцій в повній мірі розкриє практичний потенціал запропонованих методів та розроблюваного ПЗ.

А.4.1 Вимоги до функціональних характеристик

Алгоритм аналізу емоційного забарвлення повідомлень соціальної мережі Twitter був розроблений в рамках системи дослідження соціальних мереж. Одним з модулів цієї системи є модуль емоційного забарвлення текстових повідомлень.

В якості мови реалізації рішення задачі був обраний Java, оскільки вся реалізована система розробляється на ньому. Java вибрано з наступних причин:

- розв'язувані завдання не пред'являють високих вимог до продуктивності.
- Java – кроссплатформна об'єктно-орієнтована мова.
- цією мовою реалізовано велику кількість бібліотек, призначених для вирішення завдань машинного навчання та комп'ютерної лінгвістики.

Для вирішення завдань пов'язаних з попередньою обробкою тексту: визначення речень, слів, частин мови – була використана бібліотека OpenNLP. Для вирішення задачі визначення полярності тексту використана бібліотека LIBLINEAR. Ліцензії обох бібліотек допускає їх вільне використання в наукових і комерційних проектах.

А.4.2 Вимоги до надійності

Для сталого функціонування програмного забезпечення необхідно:

- наявність безперебійних блоків живлення персонального комп'ютера.

Програма повинна видавати відмову доступу до комп'ютера при спробі несанкціонованого доступу.

A.4.3 Вимоги до технічних та програмних засобів

Для нормальної роботи програми необхідні комп'ютер і мікрофон.

Технічна характеристика:

- ПК з ОС Microsoft Windows 8.1 та старше;
- JVM 9.0 та вище;
- процесор – Intel Core з тактовою частотою 2.0 МГц та вище, ядер: 2, логічних процесорів: 2.

A.5 Вимоги до програмної документації

Склад програмної документації повинен включати в себе:

- технічне завдання;
- опис проектування;
- опис тестування;
- опис програми.

A.6 Стадії і етапи розробки

A.6.1 Стадії розробки

Розробка повинна бути проведена в три стадії:

- розробка технічного завдання;
- робоче проектування;
- впровадження.

A.6.2 Етапи розробки

На стадії розробки технічного завдання повинен бути виконаний етап розробки, погодження та затвердження справжнього технічного завдання [20], [21].

На стадії робочого проектування повинні бути виконані перераховані нижче етапи робіт [20], [21]:

- розробка програми;
- розробка програмної документації;
- випробування програми.

На стадії впровадження повинен бути виконаний етап розробки підготовка і передача програми [20], [21].

А.6.3 Зміст робіт по етапах

На етапі розробки технічного завдання повинні бути виконані перераховані нижче роботи [20], [21]:

- постановка задачі;
- визначення і уточнення вимог до технічних засобів;
- визначення вимог до програми;
- визначення стадій, етапів і термінів розробки програми та документації на неї [20], [21];
- узгодження і затвердження технічного завдання.

На етапі розробки програми повинна бути виконана робота з програмування (кодування) і налагодженні програми [20], [21].

На етапі розробки програмної документації повинна бути виконана розробка програмних документів відповідно до вимог до складу документації [20], [21].

На етапі випробувань програми повинні бути виконані перераховані нижче види робіт [20], [21]:

- розробка, узгодження та затвердження та методики випробувань;
- проведення приймально-здавальних випробувань;
- коректування програми і програмної документації за результатами випробувань [20], [21].

ДОДАТОК Б
Опис програми

Б.1 Загальні свідчення

Програма аналізу емоційної тональності повідомлень в соціальних мережах на прикладі повідомлень Twitter, для аналізу повідомлень та подальшого використання даних для комерції. Існуюча програма реалізована з використанням мови програмування Java, та бібліотеки OpenNLP.

Б.2 Функціональне призначення

Алгоритм аналізу емоційного забарвлення повідомлень соціальної мережі Twitter був розроблений в рамках системи дослідження соціальних мереж. Одним з модулів цієї системи є модуль емоційного забарвлення текстових повідомлень.

В якості мови реалізації рішення задачі був обраний Java, оскільки вся реалізована система розробляється на ньому. Java вибрано з наступних причин:

- розв'язувані завдання не пред'являють високих вимог до продуктивності.
- Java – кроссплатформна об'єктно-орієнтована мова.
- цією мовою реалізовано велику кількість бібліотек, призначених для вирішення завдань машинного навчання та комп'ютерної лінгвістики.

Для вирішення завдань пов'язаних з попередньою обробкою тексту: визначення речень, слів, частин мови – була використана бібліотека OpenNLP. Для вирішення задачі визначення полярності тексту використана бібліотека LIBLINEAR. Ліцензії обох бібліотек допускає їх вільне використання в наукових і комерційних проектах.

Б.3 Опис логічної структури

Перейдемо до загальної схеми роботи системи. На малюнку стрілками зазначено, в якому напрямку і між якими компонентами передаються дані (рис. Б.1). NLP-модуль здійснює визначення частин мови і розбиття на n -грами вхідного тексту. На етапі вилучення ознак формується вектор ознак з тексту. Результат бінарної класифікації-полярність тексту, приймає значення «1» і «-1».

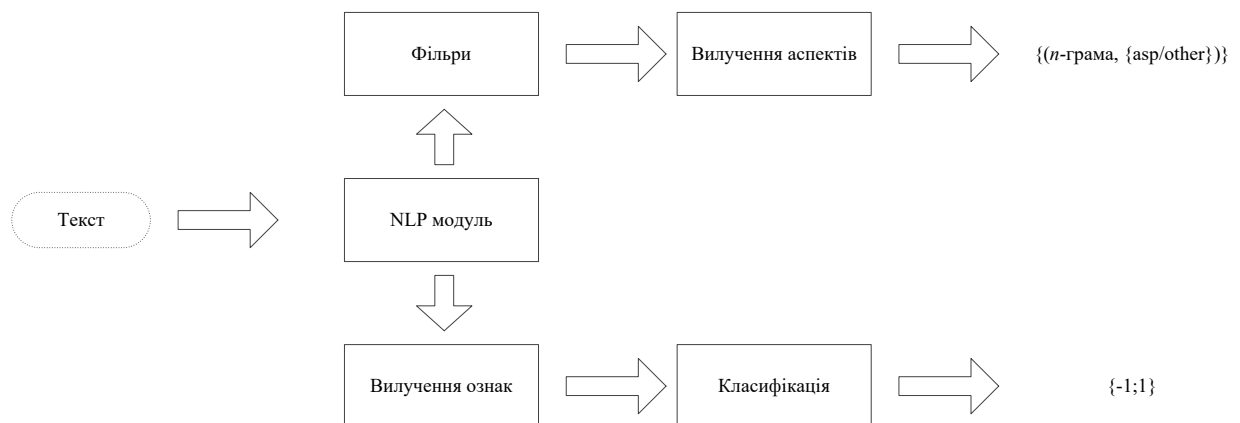


Рисунок Б.1 – Загальна схема роботи

Модуль вилучення аспектів повертає всі n -грами тексту, що пройшли через частотні і лінгвістичні фільтри, разом з позначкою про приналежність n -грами до безлічі аспектів.

Б.4 Використані технічні засоби

Для нормальної роботи програми необхідні комп'ютер і мікрофон.

Технічна характеристика:

- ПК з ОС Microsoft Windows 8.1 та старше;
- JVM 9.0 та вище;
- процесор – Intel Core з тактовою частотою 2.0 МГц та вище, ядер: 2, логічних процесорів: 2.

ДОДАТОК В
Текст програми

Сирцевий код розробленого застосунку:

```
public class Main {
    private static final twetyFact twetyFact =
        new twetyFact(buildConfig());
    private static final Twitter twitter = twetyFact.getInstance();
    private static final EntityManagerFactory emf =
        Persistence.createEntityManagerFactory("tweetyy");
    private static final EntityManager em = emf.createEntityManager();

    public static void main(String args[]) {
        try {
            license();
            Scanner in = new Scanner(System.in);
            System.out.print("Type in the search Qvr: ");
            Qvr searchTokk = new Qvr(in.nextLine());
            searchTokk.setCount(1500);
            qvrRes res = twitter.search(searchTokk);
            System.out.println("Amount of tweetyys pulled for \"\" +
searchTokk.getQvr() + "\": " + res.gettweetyys().size());

            persisttweetyy(res, searchTokk.getQvr());
            end();

        } catch (TwitterException ex) {
            Logger.getLogger(Main.class.getName()).log(Level.SEVERE, null,
ex);
        }
    }

    private static Config buildConfig() {
```

```

    ConfigBuilder cb = new ConfigBuilder();
    cb.setDebugEnabled(true)
    return cb.build();
}

```

```

private static void persisttweetyy(qvrRes res, String Tokk) {
    EntityTransaction tx = em.getTransaction();
    tx.begin();
    for (Status status : result.gettweettyys()) {
        tweetyy entity = new tweetyy();
        entity.settweetyyId(status.getId());
        entity.setSearchTokk(Tokk);
        entity.setCreatedOn(status.getCreatedAt());
        entity.setIsRetweetyy(status.isRetweetyy());
        entity.setSource(status.getSource());
        entity.setUsr(status.getUsr());
        entity.setTxt(status.getTxt());
        em.persist(entity);
    }
    tx.commit();
}

```

```

private static void end() {
    em.close();
    emf.close();
}

```

```

private static void license() {

```

```

    System.out.println("SentimentAnalysis4tweettyys Copyright (C) 2013

```

```

    Daniel Hamacher\n"

```

```

        + "\n"
        + "          This program comes with ABSOLUTELY NO
WARRANTY; for details type `show w'.\n"
        + "    This is free software, and you are welcome to redistribute
it\n"
        + "    under certain conditions; type `show c' for details.");
    }
}

public class Main {

    private static final twetyFact twetyFact =
        new twetyFact(buildConfig());
    private static final Twitter twitter = twetyFact.getInstance();
    private static final EntityManagerFactory emf =
        Persistence.createEntityManagerFactory("tweetyy");
    private static final EntityManager em = emf.createEntityManager();

    public static void main(String args[]) {
        try {
            license();
            Scanner in = new Scanner(System.in);
            System.out.print("Type in the search Qvr: ");
            Qvr searchTokk = new Qvr(in.nextLine());
            searchTokk.setCount(1500);
            qvrRes res = twitter.search(searchTokk);
            System.out.println("Amount of tweetyys pulled for \"" +
searchTokk.getQvr() + "\": " + res.gettweetyys().size());

            persisttweetyy(res, searchTokk.getQvr());
            end();

```

```

    } catch (TwitterException ex) {
        Logger.getLogger(Main.class.getName()).log(Level.SEVERE, null,
ex);
    }
}

```

```

private static Config buildConfig() {
    ConfigBuilder cb = new ConfigBuilder();
    cb.setDebugEnabled(true)
        * -> use the data given at the end of App creation
        * You need:
        * (1) Custom Key
        * (2) Custom Secret
        * (3) Access Tokk
        * (4) Access Tokk Secret
    return cb.build();
}

```

```

private static void persisttweetyy(qvrRes res, String Tokk) {
    EntityTransaction tx = em.getTransaction();
    tx.begin();
    for (Status status : result.gettweetyys()) {
        tweetyy entity = new tweetyy();
        entity.settweetyyId(status.getId());
        entity.setSearchTokk(Tokk);
        entity.setCreatedOn(status.getCreatedAt());
        entity.setIsRetweetyy(status.isRetweetyy());
        entity.setSource(status.getSource());
        entity.setUsr(status.getUsr());
    }
}

```

```

        entity.setTxt(status.getTxt());
        em.persist(entity);
    }
    tx.commit();
}

private static void end() {
    em.close();
    emf.close();
}

public class ClassML {

    int count = 0;
    int size = 0;

    public ClassML() {
    }

    public int getCount() {
        return count;
    }

    public int getSize() {
        return size;
    }

    public tweety classify(tweety tweety) {

        status = tweety.getText();

```

```

boolean clnOK = false;
while (!clnOK) {
    status = DoubleQuotesRemover.remove(status);
    clnOK = DoubleQuotesRemover.isItCleaned(status);
}

tweety = seh.applyRules(tweety, status);
if (tweety.getLstCateg().contains("001")
tweety.getLstCateg().contains("002")) {
    return tweety;
}

tweety.setLstCateg(null);
status = statusCleaner.clean(status);
statusStripped = statusCleaner.removePunctuationSigns(status);

tweety = sentenceheuristssPre.applyRules(tweety, status);

tweety = sentenceheuristssPre.applyRules(tweety, statusStripped);

tweety = hashtagheuristss.applyRules(tweety);

nGrams = new NGramFinder(status).runIt(4, true);
nGramsIterator = nGrams.iterator();
String result;
String GramLowCase;
String GramLowCase;
int indxTermOrign = 0;
String oldnegGramStrppd;

```

```

while (nGramsIterator.hasNext()) {
    nGramOrig = nGramsIterator.next().trim();

    indxTermOrign = status.indexOf(nGramOrig);
    System.out.println("status: " + status);
    negGramStrppd =
statusCleaner.removePunctuationSigns(nGramOrig);

    oldnegGramStrppd = negGramStrppd;
    negGramStrppd =
spellCheckingMethods.repeatedCharacters(negGramStrppd);
    statusStripped = StringUtils.replace(statusStripped,
oldnegGramStrppd, negGramStrppd);
    GramLowerCase = negGramStrppd.toLowerCase();
    n doubt, at least avoid a misclassification by leaving the term out.
    if (StringUtils.endsWith(GramLowerCase, "\\") ||
StringUtils.endsWith(GramLowerCase, "&quot;")) {
        continue;
    }

    if (HLoader.getMapp1Hor().keySet().contains(GramLowerCase)) {
        heurists = HLoader.getMapp1Hor().get(GramLowerCase);
        res = (termLevelheuristss.checkFeach(heurists, status,
nGramOrig));
        if (res != null) {
            tweety.addToLstCateg(res, indxTermOrign);
        }
    } else if
(HLoader.getMapp1Hor().keySet().contains(GramLowerCase)) {
        heurists = HLoader.getMapp1Hor().get(GramLowerCase);

```

```

        res = (termLevelheuristss.checkFeach(heurists, statusStripped,
negGramStrppd));
        if (res != null) {
            tweety.addToLstCateg(res, indxTermOrign);
        }
    }

    if (HLoader.getMapp2Hor().keySet().contains(GramLowCase)) {
        //          System.out.println("Negative n-grams detected!");
        //          System.out.println("nGram: " + nGramOrig);
        heurists = HLoader.getMapp2Hor().get(GramLowCase);
        res      =      termLevelheuristss.checkFeach(heurists,      status,
nGramOrig);

        if (res != null) {
            // System.out.println("result: " + res);

            tweety.addToLstCateg(res, indxTermOrign);
        }

    } else if
(HLoader.getMapp2Hor().keySet().contains(GramLowCase)) {
        //          System.out.println("Negative n-grams detected!");
        //          System.out.println("negGramStrppd: " + negGramStrppd);
        heurists = HLoader.getMapp2Hor().get(GramLowCase);
        res      =      termLevelheuristss.checkFeach(heurists, statusStripped,
negGramStrppd);

        if (res != null) {
            // System.out.println("result: " + res);

            tweety.addToLstCateg(res, indxTermOrign);
        }
    }
}

```

```

    }
}

if (HLoader.getMapp3Hor().keySet().contains(GramLowerCase)) {
    heurists = HLoader.getMapp3Hor().get(GramLowerCase);

    res = termLevelheuristss.checkFeach(heurists, status,
nGramOrig);

    if (res != null) {
        // System.out.println("result: " + res);

        tweety.addToLstCateg(res, indxTermOrign);
    }
} else if
(HLoader.getMapp3Hor().keySet().contains(GramLowerCase)) {
    heurists = HLoader.getMapp3Hor().get(GramLowerCase);

    res = termLevelheuristss.checkFeach(heurists, statusStripped,
negGramStrppd);

    if (res != null) {
        // System.out.println("result: " + res);
        tweety.addToLstCateg(res, indxTermOrign);
    }
}

if (HLoader.getMapp4Hor().keySet().contains(GramLowerCase)) {
    heurists = HLoader.getMapp4Hor().get(GramLowerCase);
    res = termLevelheuristss.checkFeach(heurists, status,
nGramOrig);

    if (res != null) {

```

```

        // System.out.println("result: " + res);

        tweety.addToLstCateg(res, indxTermOrign);
    }
    } else if
(HLoader.getMapp4Hor().keySet().contains(GramLowCase)) {
        heurists = HLoader.getMapp4Hor().get(GramLowCase);
        res = termLevelheuristss.checkFeach(heurists, statusStripped,
negGramStrppd);
        if (res != null) {
            // System.out.println("result: " + res);

            tweety.addToLstCateg(res, indxTermOrign);
        }
    }

    if (HLoader.getMapp5Hor().keySet().contains(GramLowCase)) {
        heurists = HLoader.getMapp5Hor().get(GramLowCase);
        res = termLevelheuristss.checkFeach(heurists, status,
nGramOrig);
        if (res != null) {
            // System.out.println("result: " + res);

            tweety.addToLstCateg(res, indxTermOrign);
        }
    } else if
(HLoader.getMapp5Hor().keySet().contains(GramLowCase)) {
        heurists = HLoader.getMapp5Hor().get(GramLowCase);
        res = termLevelheuristss.checkFeach(heurists, statusStripped,
negGramStrppd);

```

```

        if (res != null) {
            // System.out.println("result: " + res);
            tweety.addToLstCateg(res, indxTermOrign);
        }
    }

    if (HLoader.getMapp6Hor().keySet().contains(GramLowCase)) {
        heurists = HLoader.getMapp6Hor().get(GramLowCase);
        res      =      termLevelheuristss.checkFeach(heurists,      status,
nGramOrig);

        if (res != null) {
            // System.out.println("result: " + res);
            tweety.addToLstCateg(res, indxTermOrign);
        }
    }
    else
    if
(HLoader.getMapp6Hor().keySet().contains(GramLowCase)) {
        heurists = HLoader.getMapp6Hor().get(GramLowCase);
        res      =      termLevelheuristss.checkFeach(heurists,      statusStripped,
negGramStrppd);

        if (res != null) {
            // System.out.println("result: " + res);
            tweety.addToLstCateg(res, indxTermOrign);
        }
    }

    if (HLoader.getMapp7Hor().keySet().contains(GramLowCase)) {
        heurists = HLoader.getMapp7Hor().get(GramLowCase);
        res      =      termLevelheuristss.checkFeach(heurists,      status,
nGramOrig);

        if (res != null) {

```

```

        // System.out.println("result: " + res);
        tweety.addToLstCateg(res, indxTermOrign);
    }

    }                                     else                                     if
(HLoader.getMapp7Hor().keySet().contains(GramLowCase)) {
    heurists = HLoader.getMapp7Hor().get(GramLowCase);
    res = termLevelheuristss.checkFeach(heurists, statusStripped,
negGramStrppd);
    if (res != null) {
        // System.out.println("result: " + res);

        tweety.addToLstCateg(res, indxTermOrign);
    }
}

if (HLoader.getMapp8Hor().keySet().contains(GramLowCase)) {
    heurists = HLoader.getMapp8Hor().get(GramLowCase);
    res = termLevelheuristss.checkFeach(heurists, status,
nGramOrig);
    if (res != null) {
        // System.out.println("result: " + res);

        tweety.addToLstCateg(res, indxTermOrign);
    }
}                                     else                                     if
(HLoader.getMapp8Hor().keySet().contains(GramLowCase)) {
    heurists = HLoader.getMapp8Hor().get(GramLowCase);
    res = termLevelheuristss.checkFeach(heurists, statusStripped,
negGramStrppd);

```

```

        if (res != null) {
            // System.out.println("result: " + res);

            tweety.addToLstCateg(res, indxTermOrign);
        }
    }

    if (HLoader.getMapp9Hor().keySet().contains(GramLowerCase)) {
        heurists = HLoader.getMapp9Hor().get(GramLowerCase);
        res      =      termLevelheuristss.checkFeach(heurists,      status,
nGramOrig);

        if (res != null) {
            // System.out.println("result: " + res);

            tweety.addToLstCateg(res, indxTermOrign);
        }
    }
    else
        if
(HLoader.getMapp9Hor().keySet().contains(GramLowerCase)) {
        heurists = HLoader.getMapp9Hor().get(GramLowerCase);
        res      =      termLevelheuristss.checkFeach(heurists,      statusStripped,
negGramStrppd);

        if (res != null) {
            // System.out.println("result: " + res);

            tweety.addToLstCateg(res, indxTermOrign);
        }
    }

}

tweety = sentenceheuristssPost.applyRules(tweety, status);

```

ДОДАТОК Д
Слайди презентації

Дослідження та програмна реалізація методів аналізу емоційної тональності повідомлень у соціальних мережах

Рецензент: Коцур М.І.
к.т.н., доцент

Ільченко М.М.
студ. групи КНТ-110м
Гофман Є.О.
к.т.н., доцент кафедри ПЗ

Рисунок Д.1 – Слайд №1

Об'єкт, предмет та мета роботи

- **Об'єкт дослідження** – процес аналізу емоційної тональності повідомлень в соціальних мережах.
- **Предмет дослідження** – методи аналізу емоційної тональності повідомлень в соціальних мережах.
- **Мета роботи** – дослідити, проаналізувати та реалізувати методи аналізу соціальних мереж емоційної тональності повідомлень користувачів соціальних мереж.

Рисунок Д.2 – Слайд №2

Актуальність роботи

З кожним роком аналіз тональності тексту стає все більш актуальним завданням як з теоретичної, так і з практичної точки зору. Це, в основному, пов'язано з розвитком Інтернету і змінами в форматі спілкування між людьми.

Поряд зі збільшенням кількості інтернет-користувачів, збільшується і обсяг генерованого ними контенту.

Весь цей контент несе в собі значну кількість інформації, яку можна і потрібно використовувати для онлайн-комерції, охорони здоров'я, наукових досліджень соціальних, політичних та інших процесів.

Аналіз тональності – це область інтелектуального аналізу тексту, призначена для автоматичного вилучення суб'єктивних думок з тексту.

Рисунок Д.3 – Слайд №3

Постановка завдання

Основним завданням, поставленим в даній роботі, є реалізація методу аналізу емоційно забарвлених повідомлень в соціальній мережі Твіттер. Рішення даного завдання можна розділити на наступні етапи:

- 1) дослідження особливостей повідомлень в соціальній мережі Твіттер;
- 2) реалізація методу визначення полярності з урахуванням цих особливостей і тестування його ефективності;
- 3) реалізація методу вилучення аспектів з урахуванням цих особливостей і тестування його ефективності.

Завдання вилучення аспектів часто вирішується за допомогою методів навчання без вчителя і статистичними методами. Для збільшення ефективності цих методів використовуються лінгвістичні і частотні фільтри, що дозволяють відсівати слова, що не мають відношення до аспектів.

Рисунок Д.4 – Слайд №4

Порівняння підходів машинного навчання для аналізу емоційної тональності тексту

	Навчання з учителем	Навчання без учителя	Словникові методи
База прикладів для навчання моделі	Потребує значної бази	Не залежить від вхідних даних	Потребує значної бази
Продуктивність методів	Високий рівень	Високий рівень	Високий рівень
Коло практичних завдань для методів	Високий рівень	Середній рівень	Низький рівень
Подальший розвиток	Перспективно	Перспективно	Не перспективно

кафедра ІІЗ НУ "Запорізька політехніка"

5

Рисунок Д.5 – Слайд №5

Особливості повідомлень в Twitter

Перейдемо до вирішення першого завдання. У контексті завдання аналізу емоційного забарвлення тексту, основними особливостями повідомлень в Twitter є:

- малий розмір повідомлення – 140 символів.
- наявність сленгу, скорочень і граматичних помилок.
- наявність спеціальних символів.
- використання посилань на інших користувачів і на зовнішні ресурси.

кафедра ІІЗ НУ "Запорізька політехніка"

6

Рисунок Д.6 – Слайд №6

Порівняння мов програмування

Критерії порівняння	Мови програмування			
	Java	C++	C#	Python
Наявність літератури та простота вивчення	++	++	++	++
Підтримка парадигми модульного програмування	+	+	+	+
Стандартні бібліотеки шаблонів	+	+	+	+
Кроссплатформність	++	+	±	+
Продуктивність роботи	++	++	+	+
Узагальнення алгоритмів	+	+	+	+
Масштабованість	++	+	±	-
Паралельне програмування	++	++	++	+
Простота розроблення графічного інтерфейсу	+	±	++	+

кафедра ІІЗ НУ "Запорізька політехніка"

7

Рисунок Д.7 – Слайд №7

Порівняння середовищ розробки

Критерії порівняння	Середовища розробки		
	NetBeans	Eclipse	IntelliJ IDEA
Кроссплатформність	±	++	+
Швидкість модернізації	±	±	++
Наявність генератора сценаріїв складання	++	+	+
Підтримка репозиторіїв	++	+	++
Підтримка серверів	++	-	++
Розроблення графічного інтерфейсу користувача	++	-	+
Доступність (вимоги до розповсюдження)	++	+	-

кафедра ІІЗ НУ "Запорізька політехніка"

8

Рисунок Д.8 – Слайд №8

Результати тестувань

Метод	Юніграмм		Біграмм		Разом	
	Точність	Повнота	Точність	Повнота	Точність	Повнота
Баяєсівський класифікатор	0,73	0,7	0,72	0,68	0,72	0,71
SVM	0,81	0,74	0,76	0,67	0,81	0,72

Метод	Частинна мови		Біграмм частини мови	
	Точність	Повнота	Точність	Повнота
Баяєсівський класифікатор	0,6	0,55	0,54	0,48
SVM	0,65	0,57	0,56	0,53

кафедра ІІЗ НУ "Запорізька політехніка"

11

Рисунок Д.11 – Слайд №11

Висновки

Під час виконання дипломної кваліфікаційної роботи магістра, було розглянуто основні методи аналізу емоціональної тональності повідомлень в соціальних мережах:

- словникові методи;
- навчання з учителем;
- навчання без учителя.

В рамках роботи були вирішені наступні завдання:

- 1) проведено огляд існуючих методів автоматичного аналізу емоційного забарвлення текстів;
- 2) проведено дослідження текстових особливостей повідомлень в соціальних мережах в контексті розробки методів аналізу їх емоційного забарвлення;
- 3) розроблено метод автоматичного аналізу емоційно забарвлених повідомлень у соціальній мережі Twitter;
- 4) проведено експериментальну оцінку ефективності методів. Метод визначення полярності показав результати на рівні сучасних робіт.

кафедра ІІЗ НУ "Запорізька політехніка"

12

Рисунок Д.12 – Слайд №12

Дякую за увагу!

Рисунок Д.13 – Слайд №13