НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»
Інститут прикладного системного аналізу
Кафедра системного проектування

До захисту допущено:

Завідувач кафедри

_____ А.І. Петренко

«___»_____2021 р

Дипломна робота
на здобуття ступеня бакалавра
за освітньо-професійною програмою «Інтелектуальні сервіс-орієнтовані розподілені обчислювання»
спеціальності 122 «Комп'ютерні науки»
на тему: «Сентимент аналіз як інструмент дослідження Telegram сповіщень»

Виконала:
студентка IV курсу, групи ДА-71
Опрісник Роксолана Романівна _____

Керівник:
Доцент, к.т.н.
Кисельов Геннадій Дмитрович _____

Консультант з економіки:
Доцент, к.е.н.
Рощина Надія Василівна _____

Рецензент:
Професор каф. ММСА, д.т.н.,
Петро Іванович Бідюк _____

Засвідчую, що у цій дипломній роботі немає запозичень з праць інших авторів без відповідних посилань.
Опрісник Роксолана Романівна _____

Київ – 2021

Завдання

на дипломну роботу студенту

Опрісник Роксолані Романівній

1. Тема роботи «Сентимент аналіз як інструмент дослідження Telegram сповіщень», керівник роботи Кисельов Геннадій Дмитрович, доцент, к.т.н., затверджені наказом по університету № _____

від "__" _____ 2021 р.

2.Термін подання студентом роботи _____

3. Вихідні дані до роботи

Методологія сентимент аналізу тексту за допомогою попередньо навчених моделей машинного навчання, мова програмування Python, телеграм бот створений за допомогою API TelegramBot

4. Зміст роботи

1. Дослідити області прикладних застосувань, для яких актуально дослідження емоціональної забарвленості тексту
2. Провести порівняльний аналіз методів сентимент аналізу текстів
3. Провести порівняння алгоритмів CNN, BERT, RNN з використанням готової моделі
4. Використання програми сентимент аналізу за допомогою чат - бота
5. Висновки та рекомендації

5. Перелік ілюстративного матеріалу (із зазначенням плакатів, презентацій тощо)

Презентація до захисту роботи.

6. Консультанти розділів роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
Економічний	Рощина Н. В		

7. Дата видачі завдання 02.03.2021

Календарний план

№ з/п	Назва етапів виконання дипломної роботи	Термін виконання етапів роботи	Примітка
1	Отримання завдання	02.03.2021	
2	Збір інформації	25.03.2021	
3	Ознайомлення з літературою і підготовка теоретичної частини роботи	12.04.2021	
4	Аналіз методів і засобів розв’язання поставленої задачі	25.04.2021	
5	Вибір моделі для використання в реалізації бота	04.05.2021	
6	Розробка телеграм - бота	07.05.2021	
7	Тестування розробленої системи	14.05.2021	
8	Розробка економічної частини дипломного проекту	28.05.2021	
9	Оформлення дипломної роботи	06.06.2021	

10	Отримання допуску до захисту та подача роботи в ДЕК	09.06.2021	

Студент Опрісник Роксолана Романівна

Керівник Кисельов Геннадій Дмитрович

АНОТАЦІЯ

Дипломна робота: 78 с., 23 рис., 13 табл., 1 додаток с., 16 джерел.

СЕНТИМЕНТ АНАЛІЗ ЯК ІНСТРУМЕНТ ДОСЛІДЖЕННЯ TELEGRAM СПОВІЩЕНЬ.

Предмет досліджень - методи вирішення задач аналізу тональності тексту.

Метою дипломної роботи було проаналізувати сучасні методи аналізу тональності тексту, та реалізувати телеграм - бот з використанням попередньо навченої моделі одного із розглянутих алгоритмів.

Після аналізу алгоритмів нейронних згорткових мереж, рекурентних нейронних мереж та двонаправленого трансформера BERT на точність та швидкість, було обрано модифіковану модель RoBERTa, алгоритму BERT, та реалізовано телеграм бот із використанням даної моделі для аналізу різного типу тексту на англійській мові.

ABSTRACT

Thesis: 78 pages, 23 figures, 13 tables, 1 appendix, 16 sources.

SENTIMENT ANALYSIS AS A TELEGRAM NOTIFICATION RESEARCH TOOL.

The subject of research - methods of solving problems of text tone analysis.

The aim of the thesis was to analyze modern methods of text tone analysis, and to implement a telegram - bot using a pre - trained model of one of the considered algorithms.

After analyzing the algorithms of neural convolutional networks, recurrent neural networks and bidirectional BERT transformer for accuracy and speed, a modified model of RoBERTa, BERT algorithm, was selected and a telegram bot was implemented using this model to analyze different types of text in English.

ВСТУП

Про обробку природної мови часто говорять не лише в наукових колах,де дані методи є особливо важливими,так як вважаються основою розвитку систем штучного інтелекту, але й також серед рядових програмістів, студентів та просто звичайних людей, які цікавляться IT-індустрією.

Одним із найцікавіших методів є саме сентимент аналіз. Який призначений для виявлення емоційного забарвлення в текстах різного типу. Завдання сентимент аналізу в всесвітній мережі Інтернет набирає шаленої актуальності,оскільки кількість користувачів мережі,зростає з експоненціальною швидкістю, також зростає кількість часу,який люди проводять на форумах, в соціальних мережах тощо.

Задача аналізу тональності тексту має велике значення в сучасному світі,адже значно легше і ефективніше отримувати аналіз даних про певні продукти чи послуги в автоматизованому режимі. Власне сентимент аналіз важливий не лише в бренд маркетингу, дослідженні ринку та галузевих тенденцій чи аналізі робочої сили, що є особливо важливими на сьогоднішній день, аналіз тональності тексту може допомогти також у вкладенні власного капіталу чи моніторингу кредитних ринків. Існує багато різних методів вирішення даної задачі, проте всі вони різні і по - своєму цінні, зважаючи на це є сенс в аналізі передових алгоритмів на швидкість, якість і точність. Також варто відмітити, що на даний момент нема доступних інструментів на просторі Telegram, які би дозволяли класифікувати текст в особистих цілях або бізнес - питань. Тому має сенс створення бота використовуючи навчену модель кращого з розглянутих алгоритмів.

1 Актуальність теми і мета дослідження

1.1 Актуальність теми

Аналітика та моніторинг соціальних мереж становить величезний інтерес для соціологів, лінгвістів, психологів, маркетологів і державних структур.[2]

Велика кількість компаній використовують аналіз тональності тексту для дослідження відгуків про свої товари чи послуги. Даний метод охоплює більшу кількість людей, аніж опитування. Тому що люди все більш відкрито висловлюють свою думку в мережі.

Сентимент аналіз має широке застосування у різноманітних галузях, починаючи від торгівлі до фінансів та бізнесу.

Однак також аналіз настрою може стати дуже цінним ресурсом для користувачів соціальних мереж, а саме в особистих чатах або бесідах, адже в першу чергу завжди варто розуміти якої думки притримується твій співрозмовник стосовно певної теми, а в час пандемії Covid-19 це стало ще актуальнішим,адже спілкування тет-а-тет є обмеженим заради ж безпеки самих людей. Та іноді важко зрозуміти яку емоцію співрозмовник хотів передати своїм повідомленням.

Також даний аналіз може бути корисним для політ- технологів, маркетологів та працівників мас медіа, для яких реакція, на певні новини або теми, людей є надзвичайно важливою.

Таким чином, метою дослідження є розробка системи оцінювання емоційного відклику на повідомлення користувачів в соціальній мережі Telegram, а також в бесідах, де люди можуть обговорювати новини та інше, опис її принципів роботи та її практична реалізація.

1.2 Мета дослідження

Метою дипломної роботи є порівняння алгоритмів задачі тональності тексту з використанням CNN,RNN,BERT. Та розробка системи оцінювання емоційного відклику на повідомлення користувачів в соціальній мережі Telegram, а саме телеграм - бот з функціоналом класифікації тональності тексту з використанням навченої моделі одного із алгоритмів.

Щоб досягти мети були поставлені такі завдання:

6. Дослідити області прикладних застосувань, для яких актуально дослідження емоціональної забарвленості тексту
7. Провести порівняльний аналіз методів сентимент аналізу текстів
8. Провести порівняння алгоритмів CNN, BERT, RNN з використанням готової моделі
9. Використання програми сентимент аналізу за допомогою чат - бота
10. Зробити висновки та вказати рекомендації

Нижче розглянемо найпопулярніші способи використання аналізу емоційного забарвлення текстів.

1.3 Області застосування сентимент аналізу

Ціллю аналізу настрою може думка автора або сам автор. Це одна із найцікавіших сфер застосування, оскільки тут використовується не тільки делегування повноважень вченого, але також за мету взято наближення мислення комп'ютера до людського. Як було сказано вище аналіз тональності є одним з найбільш важливих та серйозних кроків до розвитку штучного інтелекту.

Розглянемо області в яких використання даних методів є найбільш поширеним та актуальним.

1.3.1 Моніторинг соціальних мереж

Використання сентимент аналізу в соціальних мережах має вагоме значення,адже кількість користувачів мережі інтернет щодня зростає з експоненціальною швидкістю. Люди все більше роблять покупки онлайн, замовляють певні послуги та висловлюють свої думки щодо них.Саме тому визначення емоційної забарвленості тексту допомагає компаніям дізнатись про те, як клієнти насправді ставляться до певних послуг та товарів, і також дозволяє виявити проблеми до того як вони вийдуть із під контролю , а також зарадити можливим проблемам.

До прикладу одного доленосного вечора 9 квітня 2017 року United Airlines примусово вивела пасажира з перебронюваного рейсу. Жахливий інцидент був знятий іншими пасажирами на їх смартфони та негайно розміщений в мережі. Одне з відео, опубліковане у Facebook, було поширено понад 87 000 разів і переглянуто 6,8 мільйона разів до 18:00 у понеділок, лише через 24 години.[2]

Даний інцидент слугує хорошим прикладом того, що важливо розуміти як саме висловлюються люди про ваш бренд, тому що велика кількість поширень не є гарантією позитивної реакції.

Комунікуючи з клієнтами в мережі, можна завчасно зарадити конфлікту чи великим проблемам та зберегти репутацію бренду

1.3.2 Моніторинг бренду

Велика кількість інформації також є і на просторах новинних сайтів, форумах та блогах.

До прикладу можемо взяти іцедент згаданий вище, історія про те як United Airlines примусово вивели пасажира, поширилась не лише в соціальних мережах, а також і на новинних сайтах, поширився по США, Китаю та В'єтнаму, оскільки це був представник китайсько-в'єтнамського походження компанію засуджували за расизм. І даний конфлікт швидко став темою No1 в Китаї. І все це сталось за лічені години.

Моніторинг бренду пропонує аналіз всіх новинних сайтів, блогів та форумів,щоб визначити настрої клієнтів. Також є можливість орієнтування на певні регіони та інші показники.

Даний метод дозволяє отримати не лише сухі цифри статистики, але й дає змогу зрозуміти почуття та думки своїх клієнтів. Можна порівнювати імідж власного бренду з конкурентами.

Аналіз настроїв дає можливість зарадити кризам та проблемам ще до їх появи. Хорошим прикладом є Expedia Канада.

Близько Різдва Expedia в Канаді провела класичну маркетингову кампанію "рятуйся від зими". Все було добре, за винятком скрипучої скрипки, яку

вони обрали фоновою музикою. Зрозуміло, що люди відвідували соціальні мережі, блоги та форуми. Expedia відразу помітила і видалила рекламу.[3]

Сентимент аналіз дозволяє контролювати всі обговорення навколо вашої компанії, та знешкоджувати проблеми ще до їх появлення.

1.3.3 Голос клієнта (VoC)

Моніторинг соціальних мереж та брендів дає нам швидку, фільтровану та цінну інформацію, проте аналіз настроїв можна застосовувати також під час опитувань та взаємодії з клієнтами.

Опитування Net Promoter Score (NPS) - це один із найпопулярніших способів отримання підприємствами зворотного зв'язку за допомогою простого запитання: чи рекомендуєте Ви цю компанію, продукт та / або послугу другові чи члену сім'ї? Вони дають єдиний бал за числовою шкалою.[3]

Підприємства використовують ці оцінки, щоб визначити клієнтів як промоутерів, пасивних або недоброзичливців. Мета полягає в тому, щоб визначити загальний досвід споживачів та знайти способи підняти всіх клієнтів на рівень «промоутера», де вони, теоретично, будуть купувати більше, залишатися довше та направляти інших клієнтів.[3]

Аналіз настрою легко може використовуватись для будь якого виду опитування,щоб краще розуміти своїх клієнтів. Його можна використати для виявлення різко негативних відгуків і негайно вирішити проблему.

1.3.4 Дослідження ринку та розуміння галузевих тенденцій

Як було сказано вище, сайти , форуми та соціальні мережі є хорошими джерелами інформації. Люди обговорюють кожну тему, яка з'являється в мережі 24/7 і роблять вони це повністю добровільно.

Аналіз настрою добре справляється із проблемою обробки великої кількості даних. Використовуючи цей тип аналізу тексту, маркетологи відстежують та вивчають моделі поведінки споживачів у реальному часі,

щоб передбачити майбутні тенденції та допомогти керівництву приймати обґрунтовані рішення.[4]

Однію із вагомих переваг є невелика кількість ресурсів, що потрібні для аналізу.

1.3.5 Аналіз продукції

Кожен бізнесмен турбується про те,щоб продукт,який він вводить в продаж користувався попитом. Один із найбільш ефективних способів є - аналіз думок клієнтів. Комнанії збирають інформацію за допомогою способу зворотного зв'язку, щоб мати можливість вдосконалити продукт навіть після його випуску. Часто виникає проблема, як визначити, які групи споживачів запитувати та як правильно проаналізувати масу даних,щоб мати змогу потім використати результати.

Саме тут аналіз тональності дозволяє дізнатися про переваги та недоліки товару.

Дослідивши результати аналізу настроїв, компанія точно знатиме як випустити товар, який клієнти будуть з радістю купувати.

1.3.6 Аналіз робочої сили / моніторинг залучення працівників

Деякі компанії застосовують сентимент аналіз не лише для дослідження відгуків і думок людей щодо товару, але також і для внутрішніх процесів компанії, що пов'язані з персоналом. Ці компанії виявляють задоволеність працівників та фактори, що погано впливають на ентузіазм, щоб усунути дані проблеми та тим самим збільшити ефективність роботи.

Аналіз настрою допомагає відстежувати настрої співробітників в режимі реального часу. Дані опитування можуть проводитись з певною

періодичністю і компанія матиме змогу проводити моніторинг змін в настроях працівників.

1.3.7 Вкладення власного капіталу

Обробка природної мови виходить на новий рівень застосування, з'являється можливість автоматизувати завдання що пов'язані не лише з торгівлею, а й з інвестуванням.

Наприклад, аналіз настроїв може використовуватись для збільшення масштабу дослідження, пов'язаного з інвестиціями в акціонерні капітали в банках.

1.3.8 Моніторинг відповідності

Як і у випадку з даними банківських звітів, підрозділи з дотримання вимог у банках історично збирали та зберігали записи про відповідність

нормативним актам, адже їм постійно потрібно оновлювати процеси , щоб дотримуватися цих повноважень. Це потрібно для уникнення штрафів та втрати ліцензій. Майже кожен банк має IT-команди для автоматизації певних процесів.Програмне забезпечення що використовується банками для перевірки цих нормативних вимог зазвичай базується на правилах.

Програми для аналізу настроїв, засноване на методах штучного інтелекту, може дати можливітсь банкам керувати процесами, досліджуючи великий обсяг даних та класифікувати кожне оновлення за певними факторами.

1.3.9 Моніторинг кредитних ринків

Банки також використовують сентимент аналіз для моніторингу кредитних настроїв. Різні статті про кредитні ринки містять в собі масу

даних про діяльність кредитних цінних паперів, що надаються компаніями. Досліджуючи дану сферу за допомогою аналізу настроїв, банки можуть отримати результати , пов'язані з кредитними ринками, що містять інформацію про певні облігації або комерційні папери.

У реальному випадку використання бізнесу Citigroup заявляє, що запустила CitiVelocity - інструмент, який використовує машиночитані дані медіа-новин від Thomson Reuters. Банк заявляє, що вони використовують цей інструмент для дослідження компаній щодо індексу свопів кредитних дефолтів у США та Європі.[5]

За допомогою даного способу можна отримати список компаній, що класифікується як компанії, про яких згадують частіше позитивно чи негативно. Такі дані можна використати для виявлення кореляції між новинними подіями та результатами діяльності цінних паперів на

кредитних ринках.

Ці інструменти забезпечують збереження часу та ресурсів для фінансових досліджень.

1.4 Висновки до розділу 1

Задача сентимент аналізу має велике значення в теперішньому світі,адже її можна використовувати в безлічі напрямків та сфер, починаючи спілкуванням в чатах,маркетингом, аналізом відгуків та закінчуючи передбаченнями щодо ринкових інвестицій,тощо.Існують

безліч методів для вирішення задач обробки природної мови. І кожен з них має одну ціль, але влаштований по - різному і показує різні результати точності та швидкості обчислень.

Хоча аналіз тональності має велике значення в сучасних умовах, проте досі не було реалізовано інструмент за допомогою якого кожен бажаючий міг би проаналізувати текст в соціальній мережі Telegram для своїх власних потреб або бізнесу.

Отже,метою дипломної роботи є аналіз методів машинного навчання для обробки текстів, а саме згорткові та нейронні мережі і двонаправлений трансформер BERT,спираючись на результати аналізу обрати найкращу попередньо навчену модель одного із розглянутих алгоритмів та реалізувати телеграм - бот з функціоналом аналізу тональності повідомлень.

2 Аналіз існуючих методів і моделей. Обрання моделей і методів для власної реалізації

2.1 Методи аналізу тональності тексту

На сьогоднішній день при використанні сентимент аналізу використовують одновимірний простір : позитив і негатив. Однак можливість використання

багатовимірного простору залишається. Основним завданням сентимент аналізу є визначення емоційного забарвлення тексту, тобто визначення чи є текст “позитивним” або “негативним” або ж “нейтральним”. Цілі і завдання сентимент аналізу можна показати на простих прикладах.

Є прості випадки,такі як[6]:

- Коронет має найкращі форми зі всіх круїзних суден.
- Бертрам має глибокий корпус V і легко проходить моря.
- У Флориді в 1980х роках робили потворні круїзні кораблі пастельних кольорів.

Та складні випадки, такі як[6]:

- Я би дійсно дуже хотів би піти прогулятись у таку погоду! (Можливий сарказм)
- Кріс Крафт виглядає краще, ніж Лаймстоун (Дві торгові марки, що роблять визначення цілі дуже важким)
- Кріс Крафт виглядає краще, ніж Лаймстоун, але Лаймстоун розробляє мореплавність та надійність. (Дві торгові марки, дві позиції)

2.2 Види класифікацій

2.2.1 Класифікація за бінарною шкалою

В даному випадку розглядають два види оцінок таких, як “позитив” і “негатив”. Основним недоліком даного методу є те,що не завжди можна точно сказати, що текст є саме позитивним чи негативним, бо він може містити ознаки двох класів оцінки.

Ранні роботи в цій області включають в себе праці Терні і Панга, які застосовують різні методи розпізнавання полярності оглядів товару і відгуків про фільмах відповідно.[6]

2.2.2 Класифікація за багатосмуговою шкалою

Полярність можна класифікувати за допомогою багатосмугових шкал, що було зроблено Пангом і Снайдером [6]. Вони зробили розширення класифікації кіновідгуків від оцінкок «позитивний або негативний» до рейтингу по 3-х або 4-бальною шкалою.

2.2.3 Системи шкалювання

Також відомий метод визначення тональності з використанням систем шкалювання, суть методу полягає в тому, що словам співвідносяться числа від -10 до 10 залежно від тональності (позитивна, негативна або нейтральна). Спочатку частини тексту аналізуються за допомогою алгоритмів обробки природної мови, а потім виділені з тексту об'єкти та терміни досліджуються з метою розуміння значення цих слів.

2.2.4 Суб'єктивність/об'єктивність

Дане завдання зазвичай визначається як відношення даного тексту в один з двох класів суб'єктивний й або об'єктивний.[6] Проблема суб'єктивності та об'єктивності часто може бути складнішою, ніж класифікація полярності, адже суб'єктивність слів і фраз залежить від контексту, а об'єктивний документ може містити суб'єктивні терміни.

2.3 Методи класифікації тональності

Комп'ютери можуть виконувати автоматичний аналіз цифрових текстів, використовуючи елементи машинного навчання, такі як

прихований семантичний аналіз, метод опорних векторів, «мішок слів». Більш складні методи намагаються визначити володаря настроїв (тобто людини) і мета (тобто сутність, щодо якої виражаються почуття). Щоб визначити думку з урахуванням контексту, використовують граматичні відносини між словами.[6]

Загалом аналіз тональності тексту розділяють на дві категорії,такі як:

- ручний
- автоматизований

Звичайно основну відмінність ми можемо бачити на точності результатів. Автоматизовані методи використовують методи машинного навчання,засоби і техніки обробки природної мови, що дозволяє обробляти великі пакети даних.

Аналіз тональності відносять до комп'ютерної лінгвістики, основною задачею є класифікація тексту за допомогою обробки природної мови з використанням різних засобів і технік. Аналіз тональності може виконуватись за допомогою різних методів:

- Методи, засновані на правилах і словниках
- Машинне навчання з вчителем
- Машинне навчання без вчителя
- Метод, заснований на теоретико-графових моделях

2.3.1 Методи, засновані на правилах і словниках

Даний алгоритм шукає певні думки і фрази в тексті по заздалегідь складеним словникам і правилам. Дальше текст оцінюється за шкалою відносно кількості знайденої позитивної та негативної лексики.

Цей метод використовує як списки правил, що підставляються в

регулярні вирази, так і спеціальні правила з'єднання тональної лексики всередині пропозиції. [6]

Для аналізу тексту використовують наступний алгоритм : привласнити кожному слову його значення тональності зі словника, і в кінці просумувати тональність всього тексту.

Основним мінусом методів, що засновані на словниках і правилах є складання словника,оскільки щоб отримати алгоритм з високою точністю необхідно врахувати всі аспекти вживання слів залежно від контексту. Бо велика кількість слів може вживатись як у негативному так і в позитивному напрямку, метод потребує багато ресурсів на складання правильного словника і великої кількості правил.

2.3.2 Машинне навчання з вчителем

На даний момент навчання з вчителем є доволі розповсюдженим у використанні, адже такий спосіб машинного навчання є в рази ефективнішим та швидшим на відміну від машинного навчання без вчителя. Під терміном вчителя часто мають на увазі втручання людини в процеси обробки інформації.[7]

Алгоритм можна описати так:

1. спочатку збирається колекція документів, на основі якої навчається машинний класифікатор;
2. кожен документ розкладається у вигляді вектора ознак (аспектів), за якими він буде досліджуватися;
3. вказується правильний тип тональності для кожного документа;
4. проводиться вибір алгоритму класифікації і метод для навчання класифікатора;
5. отриману модель використовуємо для визначення тональності документів нової колекції. [6]

До алгоритмів навчання з учителем належать такі типи задач, як регресія і класифікація. [7]

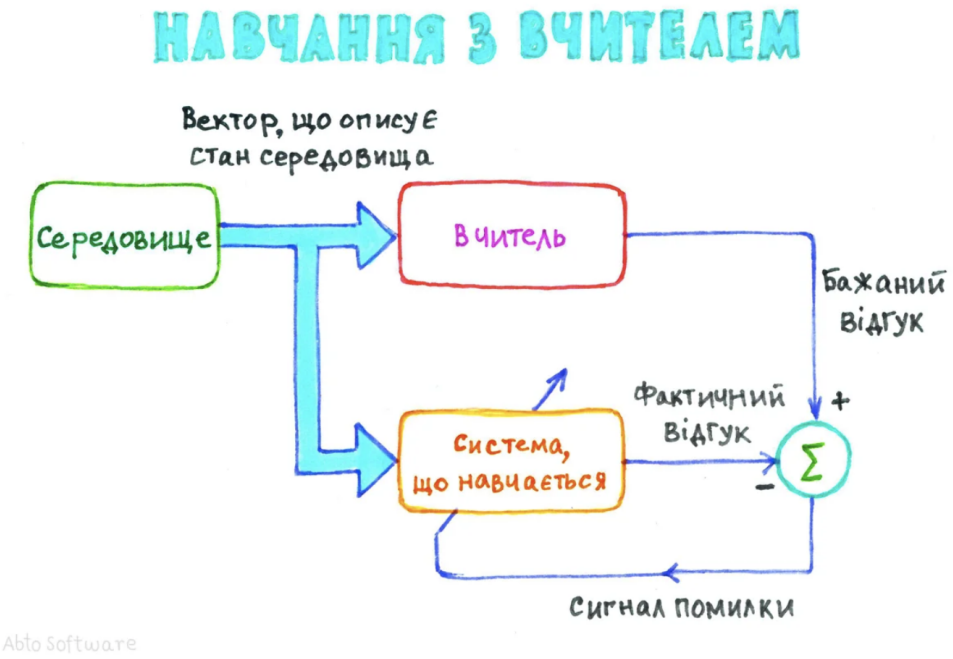

Рис. 2.1 Навчання з вчителем (7)

2.3.3 Машинне навчання без вчителя

Метод машинного навчання без вчителя полягає в тому, що машина намагається сама розібратись з класифікацією об'єктів чи тексту або інших схожих задач. Спочатку шукаються терміни,що найчастіше зустрічаються в певному тексті,і одночасно з'являються в незначній кількості текстів певної збірки текстів. В даній збірці ці слова мають зазвичай найбільшу вагу.Після визначення термінів та призначення їм тональності, можна зробити висновки про тональність всього тексту. На практиці такі алгоритми використовують зазвичай для попередньої підготовки даних.

Також великі компанії такі як Google для розмітки текстів залучають своїх клієнтів, до прикладу часто для підтвердження особи необхідно пройти авторизацію anticaptcha, виділяючи необхідні зображення ви не тільки проходити перевірку на бота, а також навчаєте

нейронну мережу розрізняти об'єкти правильно. До таких алгоритмів належать задачі кластеризації, зменшення розмірності і пошуку правил.[7]

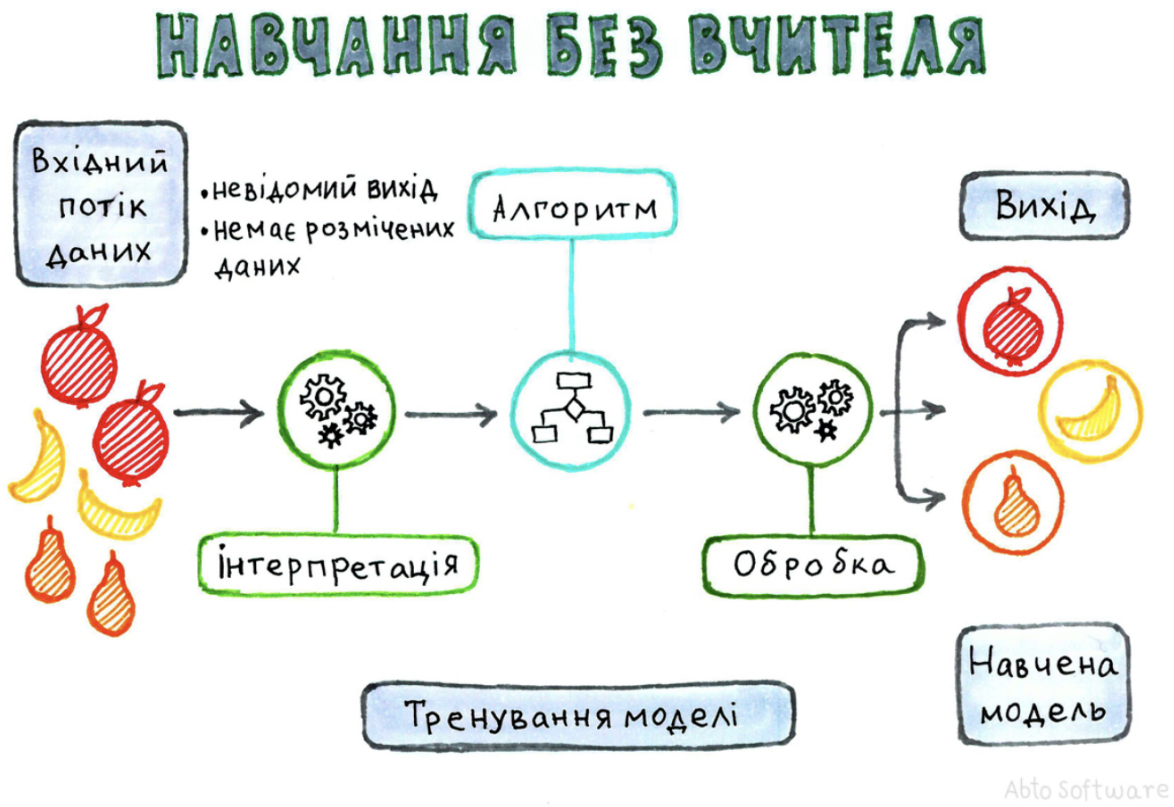

Рис.2.2 Навчання без вчителя (7)

2.3.4 Метод, заснований на теоретико-графових моделях

Основною відмінністю даного методу є припущення того,що не всі слова є рівнозначні за своєю тональністю. Різні слова можуть мати різну вагу, і тим самим більше впливати на тональність тексту в цілому.

При використанні цього методу аналіз тональності розбивається на кілька етапів:[6]

1. побудова графа на основі досліджуваного тексту;

2. ранжування його вершин;

3. класифікація знайдених слів;

4. обчислення результату.

Для класифікації слів складають тональний словник, в якому кожному слову співвідноситься оцінка, наприклад «позитивна», «негативна» або «нейтральна». Кінцевий результат отримують за допомогою обчислення значень двох оцінок: позитивного забарвлення тексту і негативного. Відсоток позитивного забарвлення тексту знаходять за допомогою сумування всіх позитивних термінів тексту з урахуванням їх ваги. Частка негативного забарвлення тексту знаходиться аналогічним чином.

Щоб отримати кінцеву оцінку тональності всього тексту потрібно знайти відношення цих складових за формулою: $T = P/N$, де T — підсумкова оцінка тональності, P — оцінка позитивної складової тексту і N — негативна складова тексту. [6]

Відповідно до статті Меншикова, текст, в якому значення T близьке до одиниці, буде вважатися нейтральною, якщо трохи перевищує 1 — позитивним. Якщо сильно перевершує 1, то сильно позитивним. Зворотне вірно і для я текстів негативної тональності. [6]

2.4 Порівняльна характеристика та формули розрахунків точності

Таблиця 2.1 - Порівняльна характеристика методів аналізу тональності тексту

Метод	**Метод Методи, засновані на правилах і словниках**	**Машинне навчання з вчителем**	**Машинне навчання без вчителя**	**Метод, заснований на теоретико -графових моделях**
Переваги	Висока точність	Висока точність,ефективне та швидке навчання	Не потребує завчасно підготовлено го пакету даних	Висока точність
Недоліки	Велика трудомісткість	Потребує завчасно підготовлено	Низька точність	Складний у реалізації

	, висока затратність	го пакету даних		

Основними критеріями якості роботи алгоритму є точність та повнота результатів. *Повнота обчислюється за наступною формулою:*

R = correctly extracted opinions / total number of opinions

де *correctly extracted opinions* — правильно розпізнані думки, *total number of opinions* — загальна кількість думок (як знайдених системою, так і не знайдених). [6]

Точність результатів обчислюється за формулою:

P = correctly extracted opinions / total number of opinions found by system

де *correctly extracted opinions* — правильно розпізнані думки, *total number of opinions found by system* — загальна кількість думок знайдених системою. [6]

2.5 Порівняльний аналіз алгоритмів CNN,BERT,RNN

На сьогоднішній день існує безліч методів за допомогою яких можна вирішити задачу сентимент аналізу тексту.

Одними з основних методів є : *машинне навчання без вчителя,машинне навчання за допомогою вчителя,метод, заснований на теоретико-графових моделях і методи, засновані на правилах і словниках.* В даній роботі для вирішення задачі емоційної забарвленості тексту мною

було обрано метод машинного навчання за допомогою вчителя. А саме було проведено аналіз і тестування існуючих алгоритмів Двонаправлена нейронна мережа кодувальник
(BERT),Згорткова нейронна мережа(CNN),Рекурентна нейронна мережа(RNN).

Нижче розглянемо кожен з алгоритмів детальніше.

2.5.1 Двонаправлена нейронна мережа кодувальник - BERT

BERT — це методика машинного навчання, що ґрунтується на Трансформері, для попереднього тренування обробки природної мови, розроблена Google. [8] Методика BERT досягла найвищих показників в сфері розуміння природної мови.

Передумови

Дослідники неодноразово наголошували на важливості і цінності попередньої підготовки моделі нейронної мережі для певної задачі, а після певного налаштування використання навченої моделі як основну. Також відомим підходом є навчання на основі функцій, де нейронна мережа виробляє вбудування слів,які потім використовуються як функції в моделях.

Як працює BERT

Машинне навчання за допомогою алгоритму BERT використовує інструмент Transformer, який має завдання дослідження контексту співвідношень словами у тексті. Transformer застосовує два механізми - кодер, за допомгою якого відбувається зчитування тексту, і декодер, що виробляє прогноз для завдання. Метою алгоритму BERT є створення мовної моделі, тому потрібен лише механізм кодування.

Ключовою технічною інновацією BERT є застосування двостороннього навчання Transformer, популярної моделі уваги, до

моделювання мови. Це на відміну від попередніх зусиль, які розглядали текстову послідовність або зліва направо, або комбіновані тренування зліва направо та справа наліво. Результати статті показують, що мовна модель, яка двосторонньо навчена, може мати глибше відчуття мовного контексту та течії, ніж однонаправлені мовні моделі.[9]

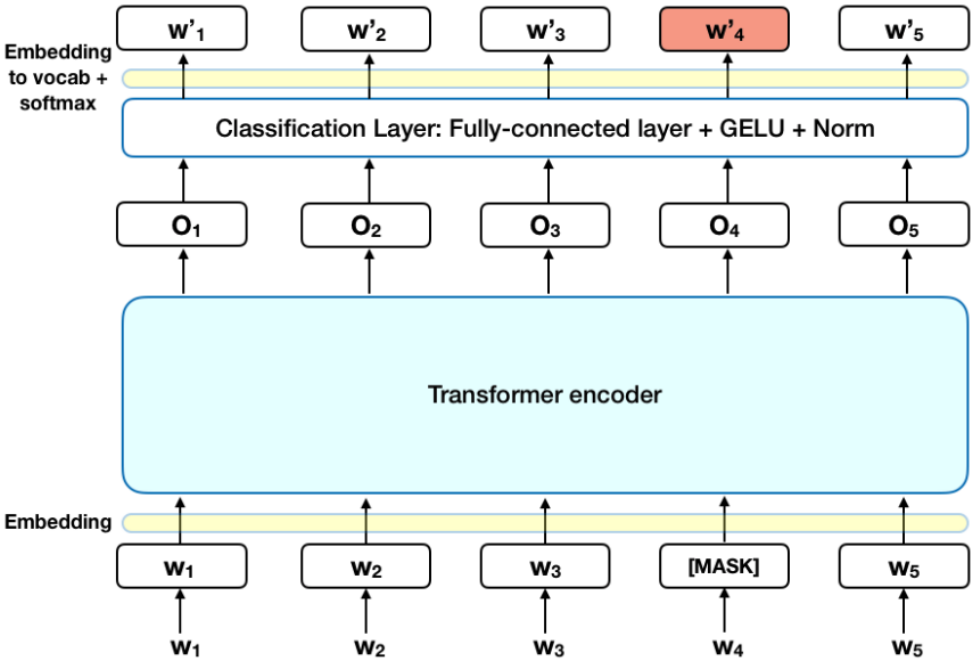

Рис.2.3 Високорівневий опис кодера Transformer(9)

Вхідні дані - це послідовність токенів, які спочатку вбудовуються у вектори, а потім обробляються в нейронній мережі. Вихід - це послідовність векторів розміром H, в якій кожен вектор відповідає вхідному маркеру з однаковим індексом.[9]

У машинному навчанні мовних моделей є проблема визначення мети прогнозування. Багато моделей використовують наступне слово в послідовності як спрямований підхід, що обмежує контекстне навчання.

Для вирішення цієї проблеми модель BERT використовує два підходи: LM у масках (MLM) та Прогнозування наступного речення (NSP).

LM у масках (MLM)

У механізмі BERT 15% слів у кожній послідовності замінюються маркером перед поданням даної послідовності. Наступним кроком моделі є процес передбачення вихідного значення замаскованих слів, виходячи з контексту, що надаються словами з послідовності, які не були попередньо замасковані.

У технічному плані передбачення вихідних слів вимагає[9]:

1. Додавання шару класифікації поверх виходу кодера.
2. Множення вихідних векторів на матрицю вбудовування, перетворюючи їх у вимір лексики.
3. Обчислення ймовірності кожного слова у словниковому запасі за допомогою softmax.

Функція втрати BERT враховує лише передбачення маскованих значень та ігнорує передбачення немаскованих слів.[9] В результаті, модель працює повільніше, ніж моделі спрямованості, однак це компенсується високою обізнаністю про контекст.

Прогнозування наступного речення (NSP)

Навчаючи мовну модель BERT отримує пари варіантів як вхідні дані і вчиться передбачати, чи є друге речення в отриманій парі наступним реченням у тексті. Під час навчання 50% вхідних даних - це пара, в якій друге речення є наступним реченням у вихідному документі, тоді як в інших 50% випадковим реченням з корпусу вибрано друге речення.[9]

Щоб модель могла правильно навчитись розрізняти два речення під час навчання, необхідно виконати наступні дії[9]:

1. На початку першого речення вставляється маркер [CLS], а в кінці кожного речення - маркер [SEP].
2. У кожну лексему додається вбудоване речення із зазначенням речення А або речення Б. Вкладання речень за своїм поняттям схожі на вкладання символів із словниковим запасом 2.
3. Позиційне вбудовування додається до кожного маркера, щоб вказати його положення в послідовності.

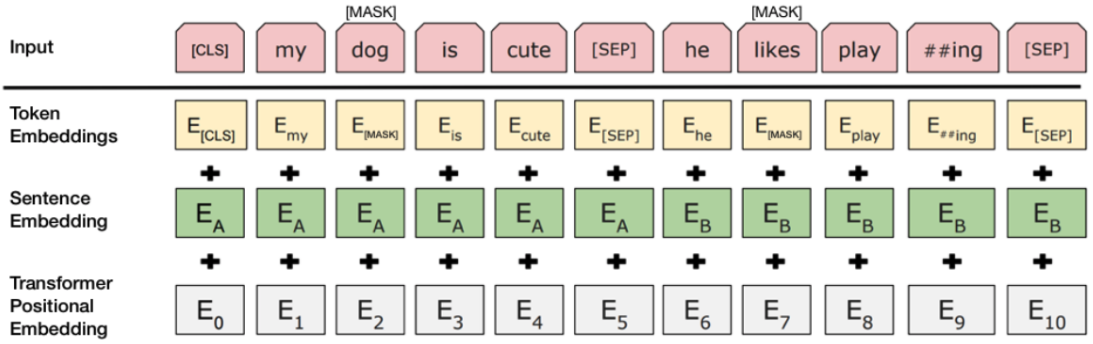

Рис.2.4 Візуалізація попередньої обробки даних (9)

Щоб передбачити, чи друге речення справді пов’язане з першим, виконуються такі дії[9]:

1. Вся послідовність введення проходить через модель трансформатора.

2. Вихід лексеми [CLS] перетворюється у вектор розміром 2 × 1, використовуючи простий шар класифікації (вивчені матриці ваг та упереджень).
3. Обчислення ймовірності IsNextSequence за допомогою softmax.

Під час навчання моделі BERT, Masked LM та Next Sentence Prediction проходять спільну обробку, щоб мінімізувати комбіновану функцію втрат двох стратегій.

Як використовувати BERT

Використання попередньо навчаної моделі є легким і доволі поширеним.

BERT використовують для:

1. Задач класифікації, таких як аналіз настрою, дані задачі виконуються за схожим методом класифікації наступних речень,а саме додають шар класифікації поверх вихідних даних Transformer для маркера [CLS].
2. Завдання відповіді на запитання,де програма отримує запитання відносно текстової послідовності та вимагає позначити відповідь у послідовності. Модель запитань та відповідей можна легко навчити за допомогою BERT, просто вивчивши два додаткові вектори, що будуть позначати початок і кінець відповіді.
3. Задачі розпізнавання іменованих об'єктів (NER) , де програма отримує текстову послідовність і вимагає позначити різні типи об'єктів, що є в тексті. Модель NER можна навчити за допомогою BERT, подаючи вихідний вектор кожного маркера в шар класифікації, який передбачає мітку NER.

Команда BERT застосувала техніку, в якій передбачається що гіперпараметри залишаються такими ж як при навчанні, за допомогою чого вдалося досягти найвищих показників у сфері обробки природної мови.

2.5.2 Згорткова нейронна мережа - CNN

Згорткова нейронна мережа (CNN) - це тип штучної нейронної мережі, що використовується для розпізнавання та обробки зображень, спеціально розробленої для обробки піксельних даних. Нейромережа - це система апаратних засобів та / або програмного забезпечення, створена за шаблоном після роботи нейронів в мозку людини. Однак згорткові нейронні мережі використовують також для сентимент аналізу тексту**.**

Як працює CNN

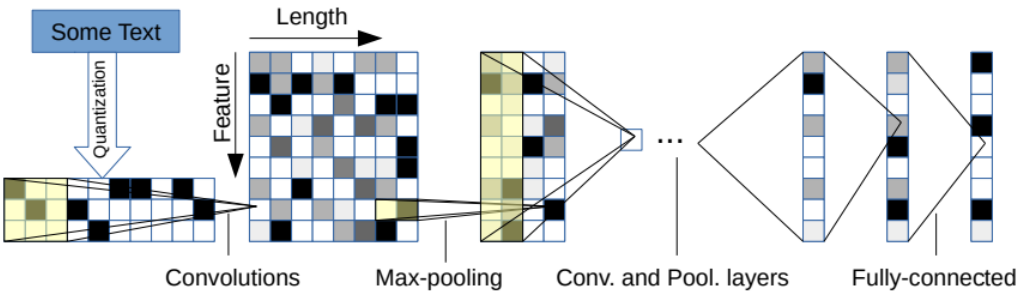

Рис.2.5 Візуалізація алгоритму CNN з текстом(11)

Для використання CNN нам необхідна матриця,оскільки згорткові нейронні мережі працюють саме з ними. Для цього використовують **embedding** - це зіставлення точки в багатовимірному просторі об'єкту, в нашому випадку - слову.

Word2Vec

Хорошим прикладом такого embedding є Word2Vec.

Word2vec - загальна назва для сукупності моделей на основі штучних нейронних мереж, призначених для отримання векторних уявлень слів на природній мові.[10]

Робота програми здійснюється наступним чином[10]:

word2vec приймає великий текстовий корпус в якості вхідних даних і зіставляє кожному слову вектор, видаючи координати слів на виході.

Спочатку він генерує словник корпусу, а потім обчислює векторне подання слів, «навчаючись» на вхідних текстах.

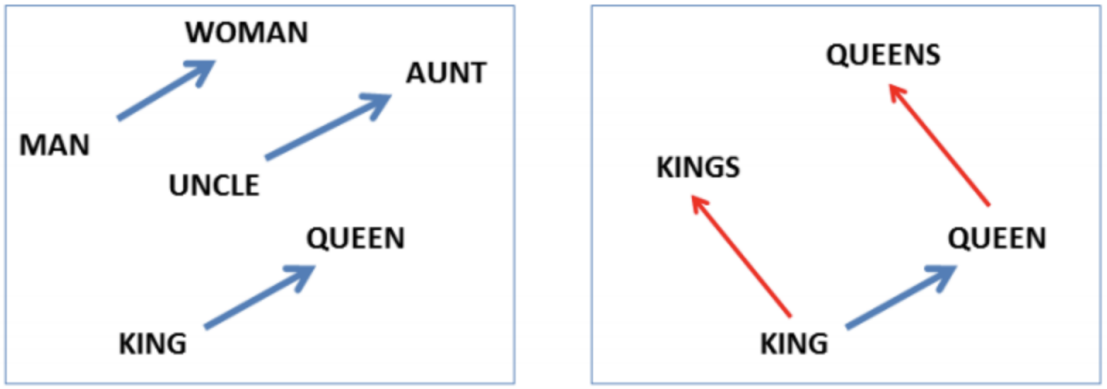

Рис.2.4 Word2Vec(11)

Векторне подання основане на контекстній близькості: слова мають схожий зміст, будуть мати близькі вектори. Отримані векторні уявлення слів використовуються для обробки природної мови та машинного навчання.

Алгоритм CNN

Використовуючи метод word2Vec ми отримуємо потрібну нам матрицю. В матриці зображення три виміри - ширина,висота та канали. У випадку з текстом у нас є тільки ширина і канали,оскільки embedding слова має значення тільки повністю, бо кожен окремий вимір втрачає свій сенс і не скаже нам нічого про слова.

Спочатку вхідна матриця обробляється шарами згортки. Як правило, фільтри мають певну фіксовану ширину, рівну розмірність призначеного простору, для підбору розмірів у фільтрів налаштовується лише один параметр - висота h. Отже, розмір вихідної матриці призначений для кожного фільтра змінюється в залежності від висоти фільтру та висоти вихідної матриці.

Наступним кроком є обробка карти ознак, що була отримана на виході кожного фільтра, шаром субдискретизації з визначеною функцією завантаження , тобто відбувається зменшення розміру карти признаків.

Таким чином ми отримуємо найважливішу інформацію для кожної згортки незалежно від її положення в тексті. Тобто ми виключаємо найбільш значущі n-грамми.

В результаті отримані карти ознак, що були розписані на виході кожного слова субдискретизації, об'єднуються в одну вектору ознаку. Вона подається на вхід в скритий повнозвязковий шар, а потім поступає на вихідний шар нейронної мережі, де і відбуваються розрахунки міток класів.

2.5.3 Рекурентна нейронна мережа (RNN)

Рекурентні нейронні мережі — це клас штучних нейронних мереж, у якому з'єднання між вузлами утворюють граф орієнтований у часі. [12] Дана нейронна мережа має блок функцій, що приймає два входи, активацію та вхідні дані та повертає вихід. Потім цей вихід передається в той самий блок з наступними вхідними даними доти, поки не будуть використані всі вхідні дані.

Рекурентні нейронні мережі працюють у три етапи. На першому етапі текст проходить вперед через прихований шар і робить прогноз. На другому етапі порівнюється передбачення із справжнім значенням за допомогою функції збитків. Функція збитків показує ефективність роботи моделі. На останньому кроці використовується значення помилок при зворотному розповсюдженні, яке додатково обчислює градієнт для кожної точки (вузла). Градієнт - це значення, яке використовується для регулювання ваги мережі в кожній точці.[13]

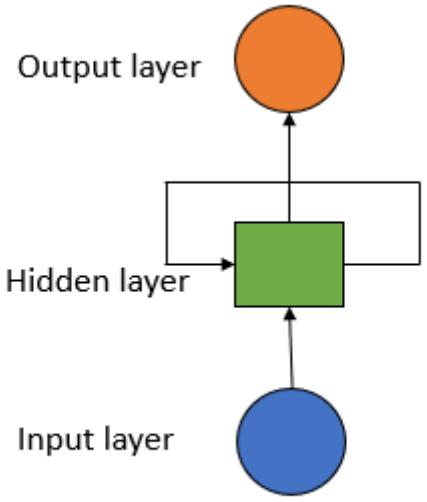

Рис.2.5 Візуалізація шарів (13)

Рекурентні нейронні мережі зазвичай використовуються до послідовних даних. Модель використовує шари, що надають їй короткочасну пам’ять. Використовуючи цю пам’ять, можна набагато точніше передбачити наступні дані. Час, протягом якого зберігатиметься інформація про минулі дані, не є фіксованим, але це залежить від відведених йому ваг. Таким чином, RNN використовується в аналізі настроїв, маркуванні послідовностей, міченні мовлення тощо.[13]

Шари рекурентної нейронної мережі

Першим шаром є embedding, так як і в CNN. Першим аргументом рівня вбудовування визначається кількість різних слів у наборі даних. Другим аргументом є кількість векторів вбудовування. Кожне слово в корпусі відповідає розміру тексту.

Другим шаром моделі є LSTM, що є найважливішою частиною рекурентних нейронних мереж. Цей шар вирішує проблему градієнта зникнення і надає можливість моделі передбачити наступне слово за допомогою пам'яті, що використовувалась останньою.

Зникаючі градієнти

Градієнт - значення , яке застосовують для регулювання ваги в кожній точці. Регулювання залежить від величини градієнту.

Якби значення градієнта попереднього шару було невеликим, значення градієнта на цьому вузлі було б меншим і навпаки.[13]

Коли градієнти мають малу величину,то відповідно вага в кожній точці погано регулюється, і в результаті чого навчання не може бути ефективним. При такому навчанні модель не розуміє контекстуальних даних.

Рішенням даної проблеми є використання LSTM. Довго-короткочасна пам’ять контролювала б потік даних при зворотному розповсюдженні. [13] Вбудований механізм LSTM, запобігає зменшенню ваги. Завдяки цьому модель стає глибшою, і ефективність моделі збільшується. На вхід структурі передають лише відповідну інформацію і відкидають нерелевантну інформацію, і, таким чином механізм передбачає послідовність правильно.

Функція активації

Модель RNN використовує функція активації "Гіперболічний тангенс (tanh (x))", оскільки вона зберігає значення від -1 до 1.[13] Ваги у вузлі множаться на градієнти, щоб отримати коригування. Чим більше значення градієнта, тим більше значення ваги буде для конкретного вузла. Через це ваги інших вузлів будуть мінімальними і не будуть враховуватись в процесі навчання, що призведе до високого зміщення моделі. Для вирішення даної проблеми використовується гіперболічна функція tanh (z). Вона надає такі ж значення як і (tanh (x))" і при цьому зберігає рівномірний розподіл між вагами мережі.

При цьому функція гіперболічного тангенса сходиться швидше, відповідно обчислення є менш затратними.

Вихідний шар

У вихідному рівні використовується функція активації “Sigmoid”. Перевагою даної функції є те, що вона дозволяє використовувати лише відповідне і важливе значення для прогнозування.

Задача класифікації тексту полягає в тому , що ми прогнозуємо позитивний або негативний окрас тексту. Отже, ми маємо справу з проблемою двійкової класифікації. Отже, ми використовуємо функцію втрати “binary_crossentropy”.

Модель навчається партіями. Замість підготовки окремого огляду за раз, ми ділимо його на групи. що дає можливість зменшити обчислювальну потужність.

Таблиця 2.2 - Порівняння моделей BERT,CNN,RNN за ознаками точності та втрат NSP

	BERT	CNN	RNN
Втрата NSP	-	35%	31%
Точність	93%	81%	86%

2.6 Обраний алгоритм для тестування за допомогою телеграм-бота

Для тестування за допомогою телеграм-бота мною було вибрано надійно оптимізовану модель RoBERTa методу BERT. А саме в роботі використовувалась попередньо навчена модель sentiment-roberta-large-english.

Ця модель є налаштована контрольно-пропускним пунктом великого RoBERTa, що надає можливість на надійний двійковий аналіз настроїв для різних типів англомовного тексту. Як результат ми отримуємо позитивні (1) або негативні (0) настрої. Дана модель була доопрацьована та оцінена

на основі 15 наборів даних з різних текстових джерел, щоб посилити узагальнення для різних типів текстів.[14]

Отже, дана модель значно перевершує, навчені лише одному типу тексту моделі.

Модель RoBERTa була запропонована в RoBERTa: Надійно оптимізований підхід до підготовки BERT від Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov. Вона заснований на моделі BERT від Google, випущеній у 2018 році.[14]

Модель базується на BERT та модифікує ключові гіперпараметри, що робить можливим не застосовувати попереднє навчання наступного речення та навчається за допомогою великих міні-партій , великого байт-рівня BPE та за допомогою динамічним маскуванням, повними реченнями без втрати NSP.

Підхід попереднього навчання моделі допоміг досягти зростання продуктивності.

Таблиця 2.3 - Гіперпараметри для попередньої підготовки RoBERTa - large[16]

Гіперпараметри	RoBERTa - large
Кількість шарів	24
Прихований розмір	1024
FFN внутрішній прихований розмір	4096
Головки уваги	16
Розмір головки уваги	64

Випадання	0.1
Увага, що випадає	0.1
Кроки підготовки	30k
Піковий рівень навчання	4e-4
Розмір партії	8k
Зниження ваги	0.01
Максимальний крок	500k
Зниження швидкості навчання	Лінійна
Adamє	1e-6
Adam β1	0.9
Adam β2	0.98
Відсікання градієнта	0.0

Таблиця 2.4 - Гіперпараметри для тонкого налаштування RoBERTa LARGE[16]

Гіперпараметри	RACE	SQuAD	GLUE
Швидкість навчання	1e-5	1.5e-5	{1e-5, 2e-5, 3e-5}
Розмір партії	16	48	{16,32}

Зниження ваги	0.1		
Зниження швидкості навчання	Лінійна	Лінійна	Лінійна
Коефіцієнт підготовки	0.06	0.06	0.06
Максимальна кількість епох	4	2	10

2.6.3 Порівняння моделі RoBERTa з іншими модифікаціями методу BERT

Оцінка моделі RoBERTa за трьома різними еталонами: GLUE, SQuaD та RACE.

Таблиця 2.5 - Результати отримані за еталоном RACE [16]

Модель	Точність
BERT_large	72%
XLNet_large	81.7%
RoBERTa	83.2%

Таблиця 2.6 - Результати отримані за еталоном GLUE [16]

Модель	EM	F1
BERT_large	79%	81.8%
XLNet_large	86.1%	88.8%
RoBERTa	86.5%	89.4%

Де EM - цей показник вимірює відсоток, який точно відповідає будь-якій із основних правдивих відповідей , F1 - ця метрика вимірює середнє перекриття між передбаченням та відповіддю на достовірність.

Таблиця 2.7 - Результати отримані за еталоном SQuaD[16]

Модель	MNLI	SST
BERT_large	86.6%/-	93.2%
XLNet_large	89.8%/-	95.6%
RoBERTa	90.2%/90.2%	96.4%

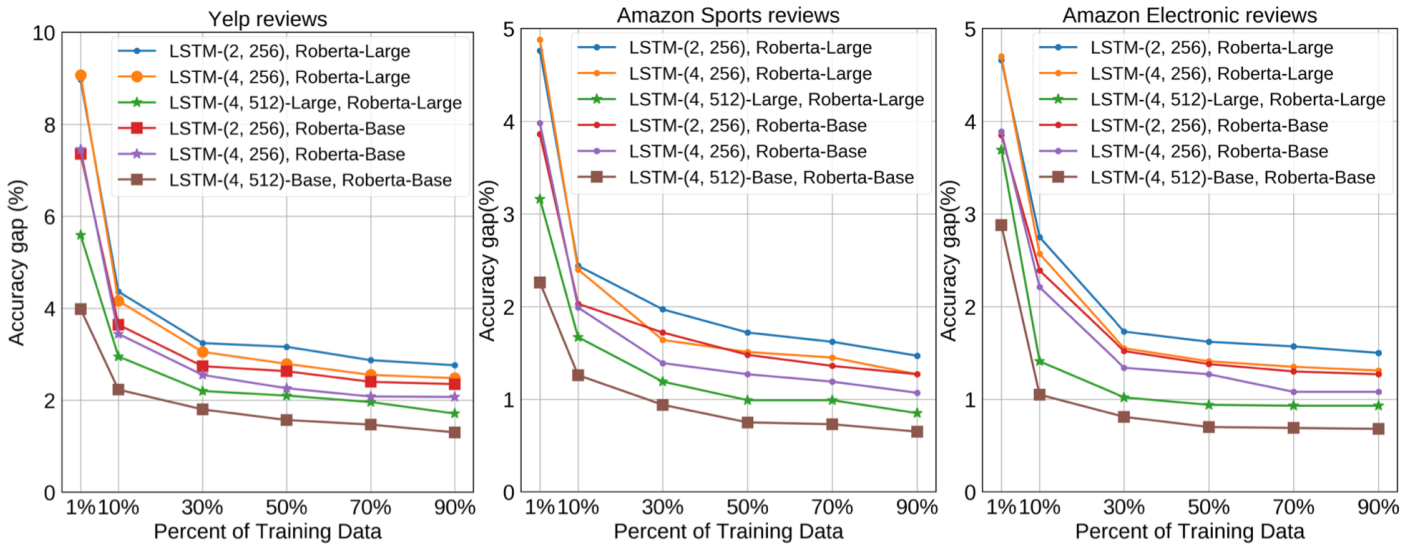

Рис.2.6 Розрив точності для RoBERTa, навченої на різних обсягах даних (15)

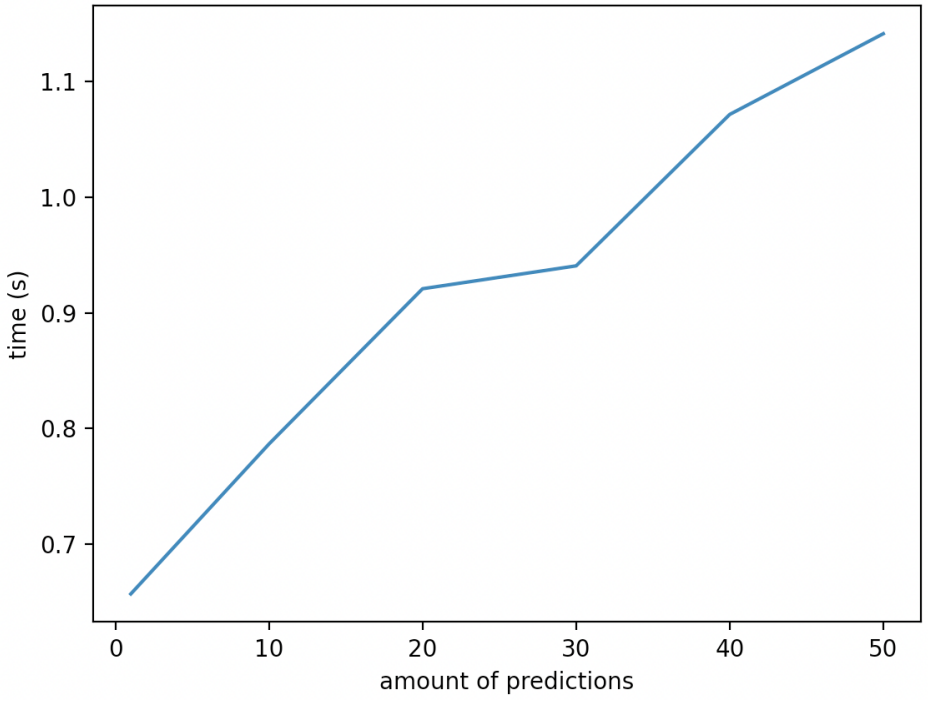

Рис.2.7 Графік швидкості залежно від кількості передбачень моделі Roberta(BERT)

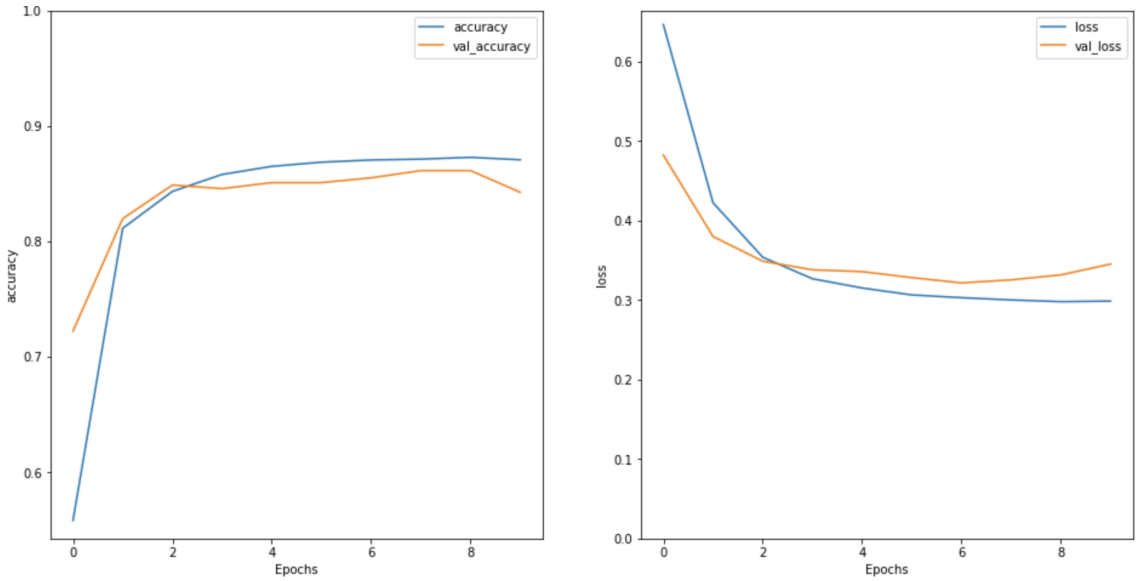

Рис.2.8 Графік точності та втрати NSP для алгоритму RNN в залежності від епох

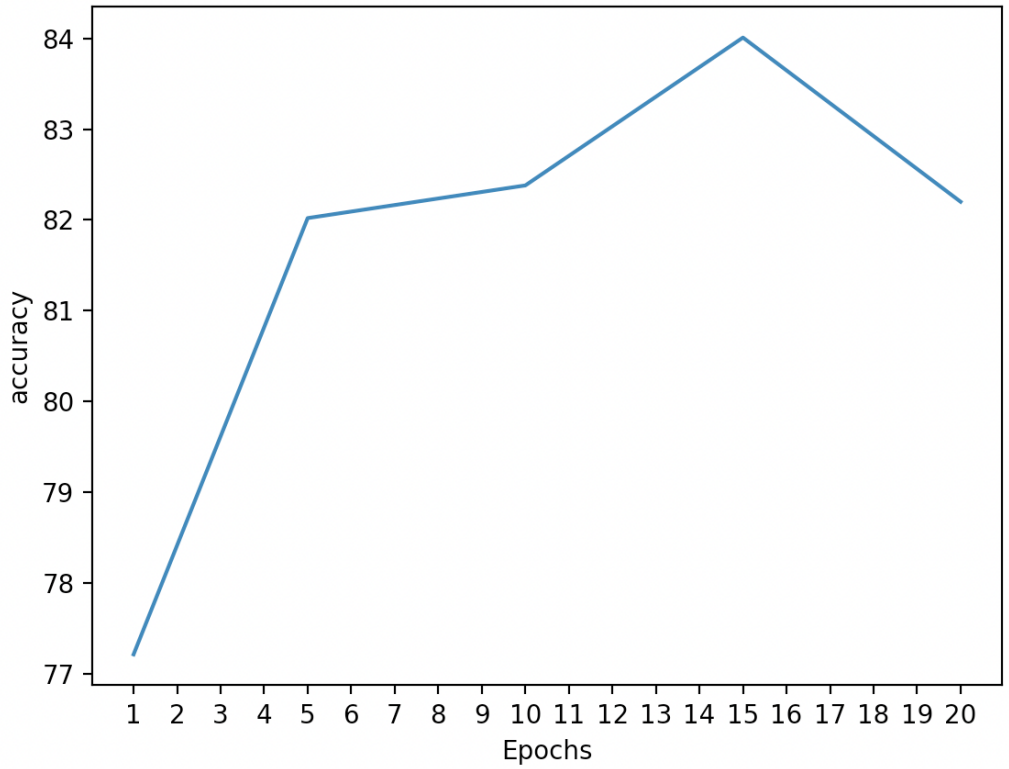

Рис.2.9 Графік точності для алгоритму CNN в залежності від епох

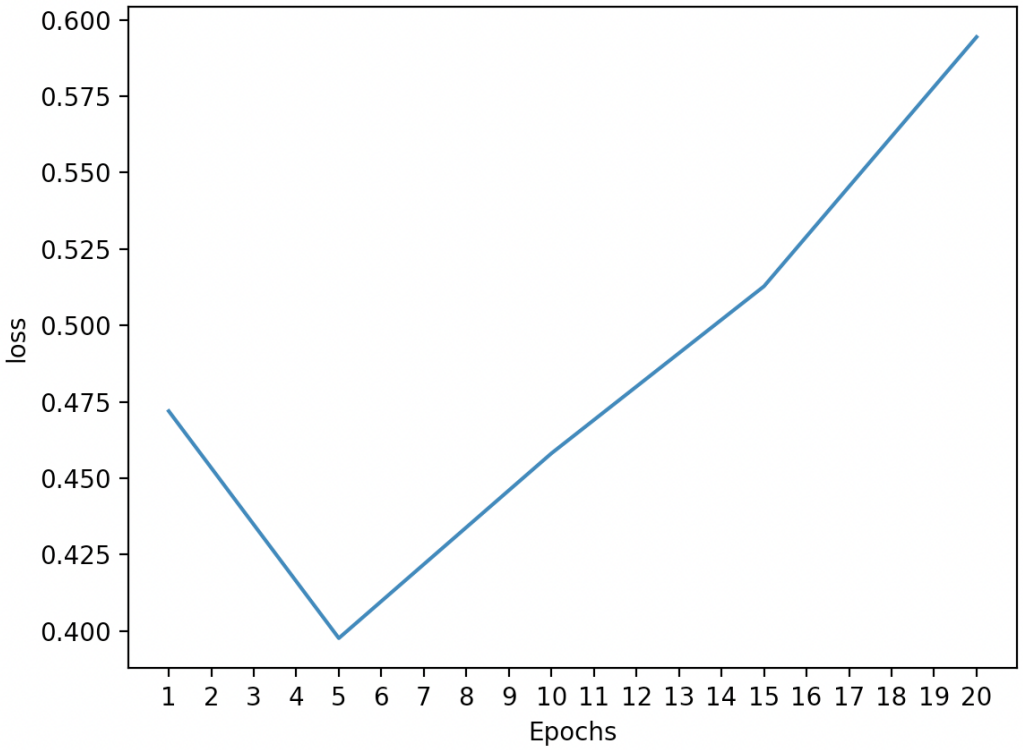

Рис.2.10 Графік втрати NSP для алгоритму CNN в залежності від епох

2.7 Висновки до розділу 2

Методи,що використовуються для вирішення задачі сентимент аналізу поділяють на ручні та автоматизовані. В сучасних умовах доцільно обирати автоматизовані методи,оскільки вони дають високу точність та швидкий результат. Алгоритми CNN,RNN та BERT є представниками машинного навчання з вчителем. Зазвичай згорткові та нейронні мережі використовують для класифікації зображень. Однак, вони також доволі добре показують себе у роботі з текстами.

Проте, алгоритм BERT є більш досконалим в цій сфері,адже спрямований на задачі обробки природної мови, хоча час навчання є

значно довшим, навідміну від методів нейронних мереж, але його моделі показують найкращі результати в тестах на швидкість та точність обробки текстів.

3 Реалізація чат-бота з вибором інструментарію реалізації

3.1 Версії використаних бібліотек та технологій

Для реалізації телеграм - бота на мові Python v3.7.9 було використано середовище розробки PyCharm Professional 2020.3. А також бібліотека telebot із TelegramBotAPI - що є інтерфейсом на основі HTTP, створений для розробки телеграм - ботів.

Для підключення моделі для використання сентимент аналізу було використано бібліотеки pipelines - для легкого використання інтерфейсу моделі та transformers, власне де і знаходиться попередньо навчена модель.

3.2 Опис використаних бібліотек та методів

За допомогою функції message.text бібліотеки telebot, отримуємо вхідне повідомлення користувача.

Для відправки повідомлення з проханням ввести текст для класифікації було використано такі функції, як send_message, для відправки також вказуємо id чату в який буде відправлено повідомлення за допомогою функції message.chat.id.

Для зміни структури отриманих результатів роботи моделі було використано метод update( ) , який додає пари ключів у словник.

Для безперервної роботи бота необхідно використати метод polling().

3.3 Застосування моделі до чат - бота у соціальній мережі Telegram

Застосування моделі sentiment-roberta-large-english для використання класифікації тексту за допомогою телеграм бота потребує часткової модифікації для отримання результатів, що будуть оброблятись навченою моделлю, а саме:

1. необхідно приймати текст, який надходить через телеграм - бот на вхід моделі,за допомогою функції message.text, яка є вбудованою в TelegramBotAP
2. також необхідно перетворити структуру яку повертає модель у вигляді списку словників у звичайний словник.

Словник буде містити два ключа "**label**", який надає нам інформацію про тональність тексту, а також "**score**",що зберігає значення від -1 до 1 ,наскільки текст є позитивно чи негативно забарвлений.

Після зміни структури уже є можливість отримати значення саме ключа “label", який будемо використовувати для надання результатів у чат-боті. Даний ключ може мати два значення - "**POSITIVE**", у разі якщо текст класифіковано, як позитивне емоційне забарвлення, або "**NEGATIVE**" - якщо текст класифіковано, як негативне емоційне забарвлення.

3.4 Опис програмного продукту

Для реалізації телеграм - бота було використано мову програмування високого рівня Python. Для забезпечення функціоналу бота було використано надійно оптимізований метод попередньої підготовки систем обробки природних мов (NLP), що вдосконалює подання двонаправлених кодерів від Transformers, або BERT, метод самоконтролю,що був випущений Google у 2018 році, який було вбудовано для функціоналу бота за допомогою мови програмування Python.

Обрана мова є однією з найшвидших та надійних, а також є кросплатформенною,що дає змогу запускати вихідний код на різних платформах без жодних змін і додаткових бібліотек.

Вибір моделі для реалізації телеграм - бота було зроблено в сторону **sentiment-roberta-large-english[15]** , оскільки дана модель вважається однією із найточніших в передбаченні результатів щодо задачі сентимент аналізу. Модель була доопрацьована та оцінена на основі 15 наборів даних з різних текстових джерел, щоб посилити узагальнення для різних типів текстів (огляди, твіти тощо).

Отже, вона перевершує моделі, навчені лише одному типу тексту.

3.5 Висновки до розділу 3

Для реалізації чат - бота в соціальній мережі Telegram, було обрано надійно оптимізований метод попередньої підготовки систем NLP, а саме модель RoBERTa алгоритму двонаправленого трансформера BERT. Для реалізації бота було обрано мову Python та бібліотеку TelegramBotAPI. Результати роботи моделі було частково модифіковано до зручнішого вигляду та використання в чаті.

4 Експериментальні дослідження реалізації. Пропозиції з оптимізації і до подальшої роботи

4.1 Доступ до телеграм - бота

Доступ до телеграм - бота можна отримати за наступним посиланням - https://t.me/sentimbot або за допомогою QR Code що наведено нижче.

Рис.4.1 QR Code для доступу до телеграм-бота Sentiment Analysis

4.2 Початковий екран

Після переходу по ссилці або за допомогою QR-Code ми потрапляємо на сторінку з якої можемо перейти власне в чат з телеграм - ботом.

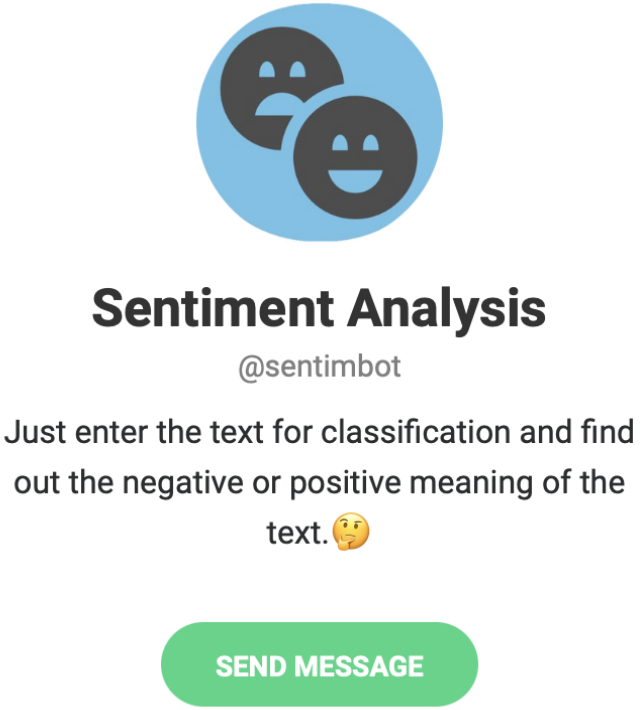

Рис.4.2 Перехід за посиланням

4.3 Запуск бота

Після повного переходу в чат, нам необіхдно запустити бота за допомогою команди **START**

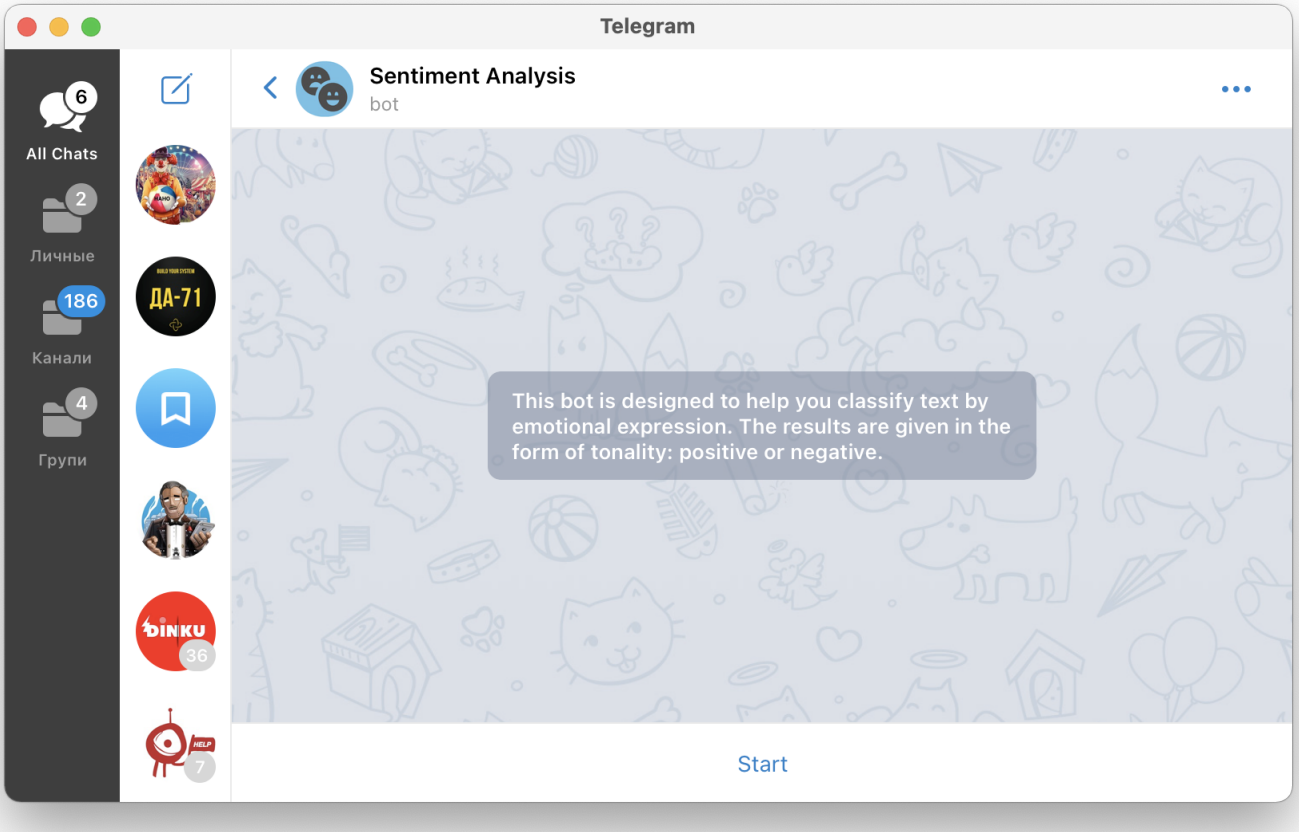

Рис.4.3 Запуск роботи бота

4.4 Введення тексту

Після чого ми отримуємо повідомлення з проханням ввести текст на англійській мові ,який необхідно класифікувати за емоційним забарвленням.

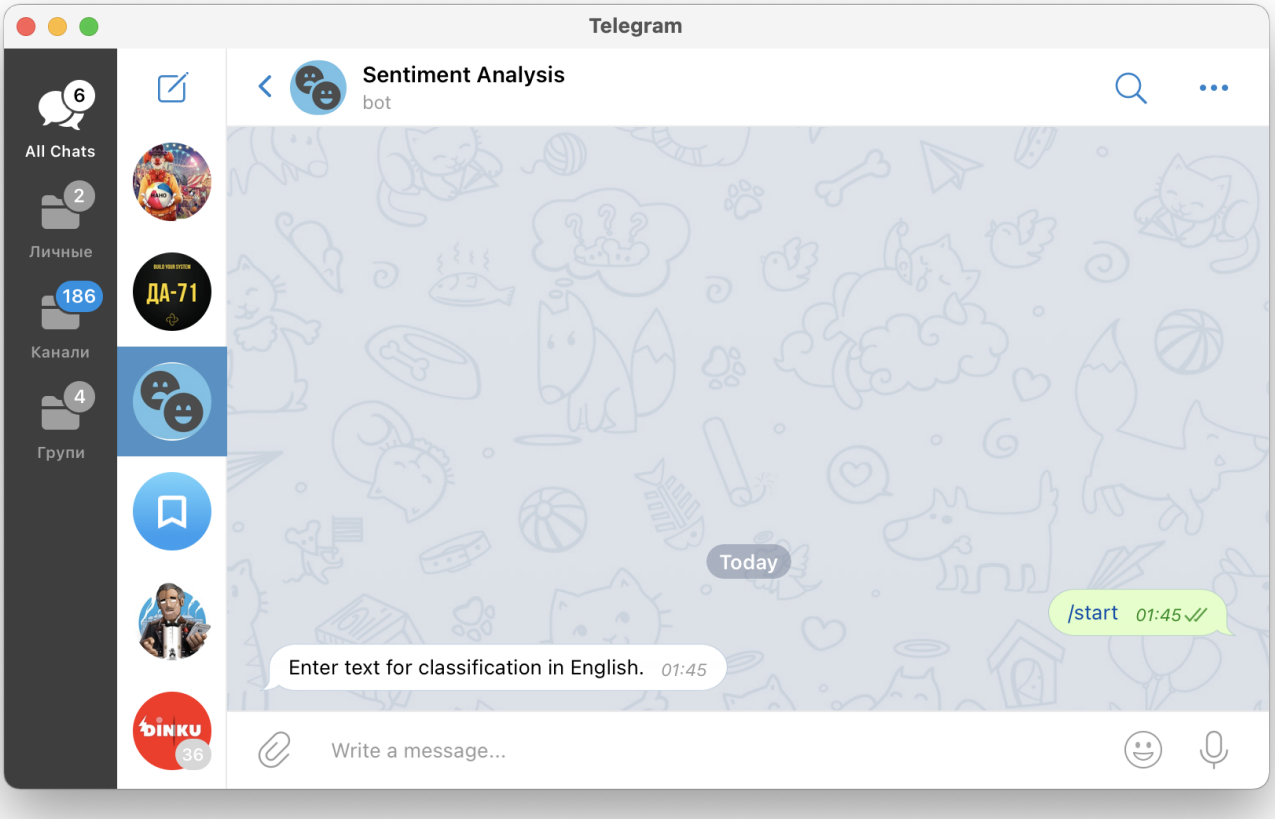

Рис.4.4 Інформація про введення тексту для класифікації

Після того як ми ввели текст,що бажаємо класифікувати необхідно зачекати доки бот опрацює введені дані і видасть кінцевий результат.

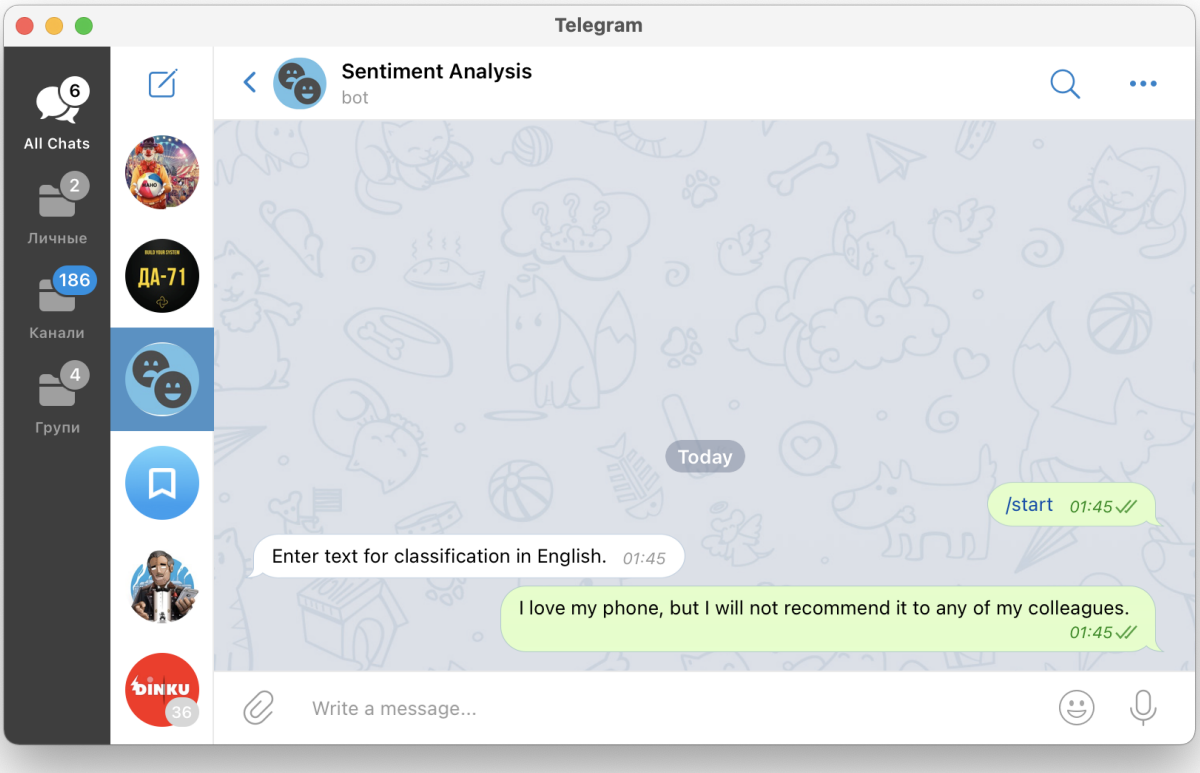

Рис.4.5 Приклад введення тексту

4.5 Отримання результатів

Через деякий час телеграм бот надсилає нам повідомлення в якому йдеться про результат класифікації.

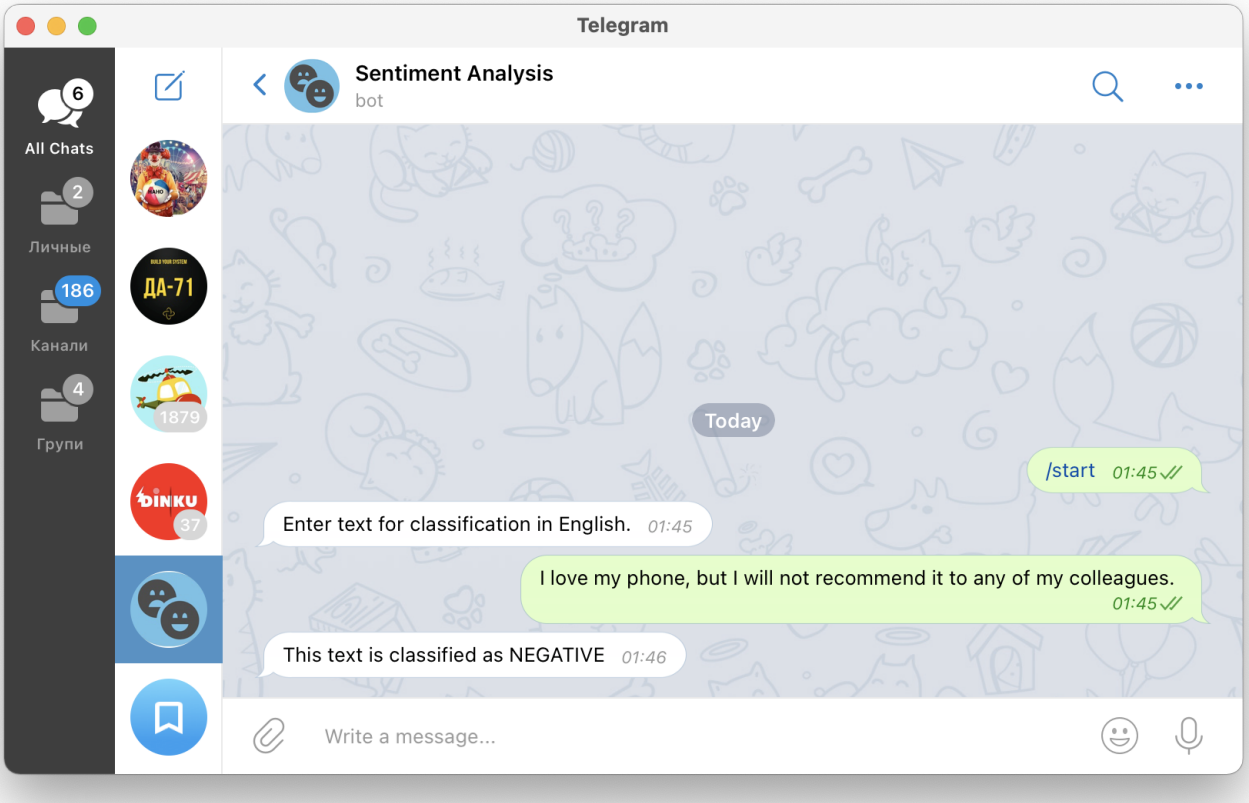

Рис.4.6 Результати класифікації

Телеграм - бота також можна використовувати в групових чатах, послідовність дій аналогічна під час часу безпосередньо із ботом в приватному чаті.

4.6 Рекомендації

Важливо відмітити потребу захисту та приватності інформації, що надходить на вхід моделі з чату з телеграм - ботом. Рекомендовано для повного використання даного бота забезпечити приватність даних, що надходять на вхід моделі, адже бот може використовуватись не тільки для бізнес - питань, але й також у особистих цілях.

Бот може бути корисним для людей з обмеженими можливостями, наприклад аутизм. Людям із аутизмом часто в офлайн бесідах буває важко розуміти почуття співбесідника, а отже бот може стати їм в нагоді при онлайн спілкуванні.

На даному етапі безпека інформації не була опрацьована, чат - бот можна використовувати зі згоди обох сторін на обробку інформації, або власне особисто для себе чи певних питань бізнесу.

Також для повного функціонування телеграм - бота рекомендовано використати модель, що підтримує українську та російську мови.

Необхідно налаштувати повне використання бота в бесідах,що складаються з 3 осіб та більше.

4.7 Висновки до розділу 4

Телеграм - бот , що було реалізовано з використанням попередньо навченої моделі є зручним застосунком у використанні для сентимент аналізу текстів англійською мовою. Його функціонал повністю виконує поставлене завдання. Використана модель надає результати з високою точністю.

5 Економічна частина

5.1 Постановка задачі проектування

Спроектований програмний продукт для автоматизованого аналізу тональнсоті тексту за допомогою телеграм - бота з використанням попередньо навченої моделі. Реалізовано за допомогою мови Python та бібліотеки pyTelegramBotAPI та TensorFlow.

5.2 Обґрунтування функцій та параметрів програмного продукту

F1 – вибір мови програмування;

a. Python
b. C#

F2 – вибір алгоритму;

a. CNN/RNN
b. BERT

F3 – вибір моделі сентимент аналізу

a. RoBERTa
b. BERT

Варіанти реалізації основних функцій наведені у морфологічній карті системи

Морфологічна карта відображає множину всіх можливих варіанти основних функцій.

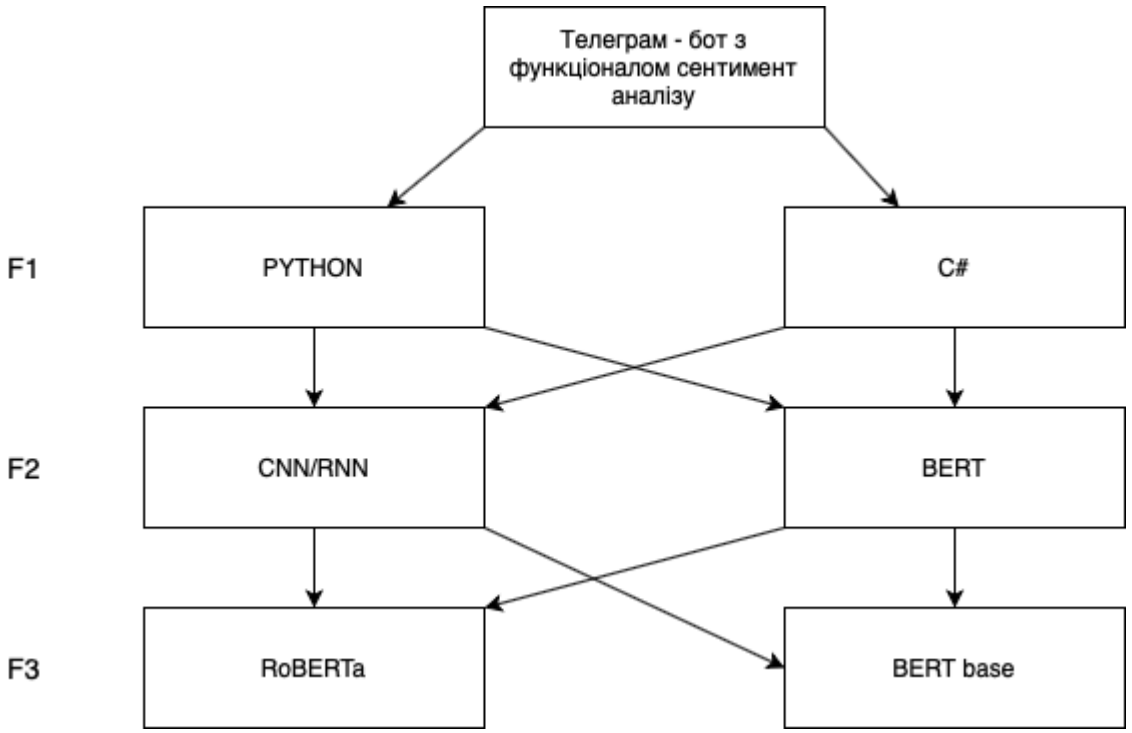

Рис 5.1 Морфологічна карта

Таблиця 5.1 Позитивно - негативна матриця

Функції	Варіанти реалізації	Переваги	Недоліки
F1	a	Легка розробка бота, кросплатформеність, доступність бібліотек	Низька швидкість обробки при обробці великих наборів даних
	b	Кросплатформеність,автоматичний збір сміття	менш гнучкий,низький рівень

			паралельності
F2	a	Хороша точність результатів	Велика кількість параметрів, для яких необхідна конфігурація,до вгий час навчання
	b	Висока точність результатів, простіша конфігурація	Деякі типи моделей BERT можуть псувати деякі слова в послідовності
F3	a	Найвища точність результатів	Навчання займає багато часу, в 4-5 разів більше ніж BERT
	b	Досить швидкий час навчання	втрата NSP

На основі карти побудовано позитивно-негативну матрицю варіантів основних функцій. Можемо дійти до висновку, що при розробці програмного продукту деякі варіанти реалізації функцій варто відкинути, тому що вони не відповідають поставленим перед програмним продуктом задачам.

Розглянемо більш деталально кожну з функцій.

1. Функція F1: Перевагу віддамо простоті реалізації, доступності бібліотек та кросплатформеності . Отже, варіант b відкидаємо.
2. Функція F2: Обидва варіанти буде доцільно використати в розробці.
3. Функція F3: Перевагу надаємо найвищій точності результатів. Варіант b відкидаємо.

Таким чином, будемо розглядати такі варіанти релазції програмного продукту:

F1a - F2a - F3a

F1a - F2b - F3a

Для оцінювання якості розглянутих функцій обрана система параметрів, описана нижче.

5.3 Обґрунтування системи параметрів програмного продукту

На основі даних, розглянутих вище, визначаються основні параметри вибору, які будуть використані для розрахунку коефіцієнта технічного рівня.

Для того, щоб охарактеризувати програмний продукт, будемо використовувати наступні параметри:

X1 – швидкодія мови програмування;

X2 – зручність введення даних;

X3 – швидкодія програмного продукту;

X4 – точність результатів.

Оцінка значень параметрів наведена в таблиці нижче.

За данними наведеної вище таблиці побудуємо графічні характеристики параметрів X1 − X4.

Таблиця 5.2 - Оцінка значень параметрів

Назва параметра	Позначення	Одиниці виміру	Значення параметра		
			Гірші	Середні	Кращі
швидкодія мови програмування	X1	оп/мс	10000	15000	20000
зручність введення даних	X2	кількість кроків	5	3	2
швидкодія програмного продукту	X3	секунди	30	15	5
точність результатів	X4	відсотки	80	85	92

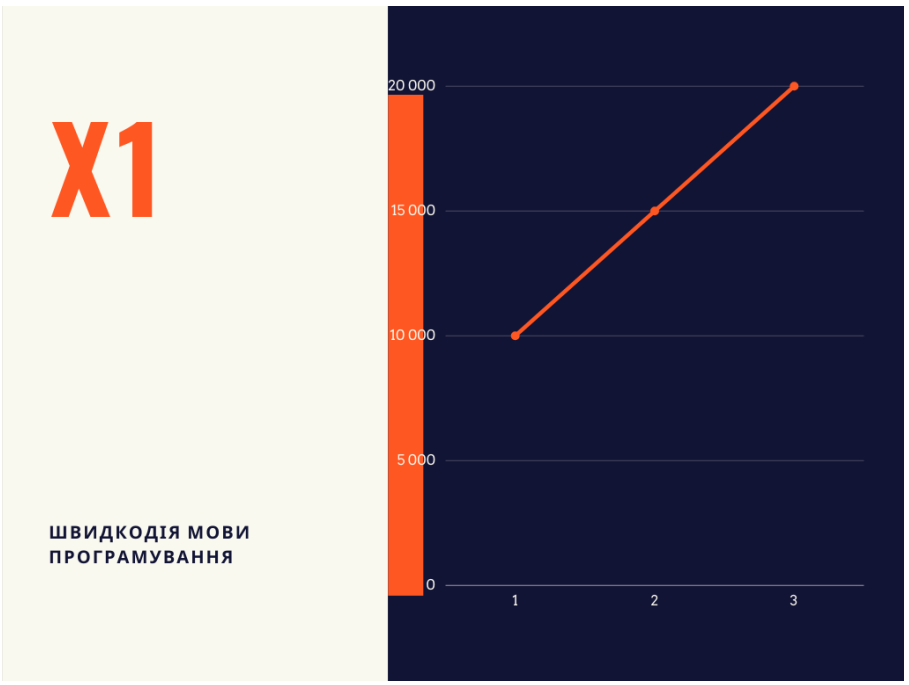

Рисунок 5.2 – X1, швидкодія мови програмування

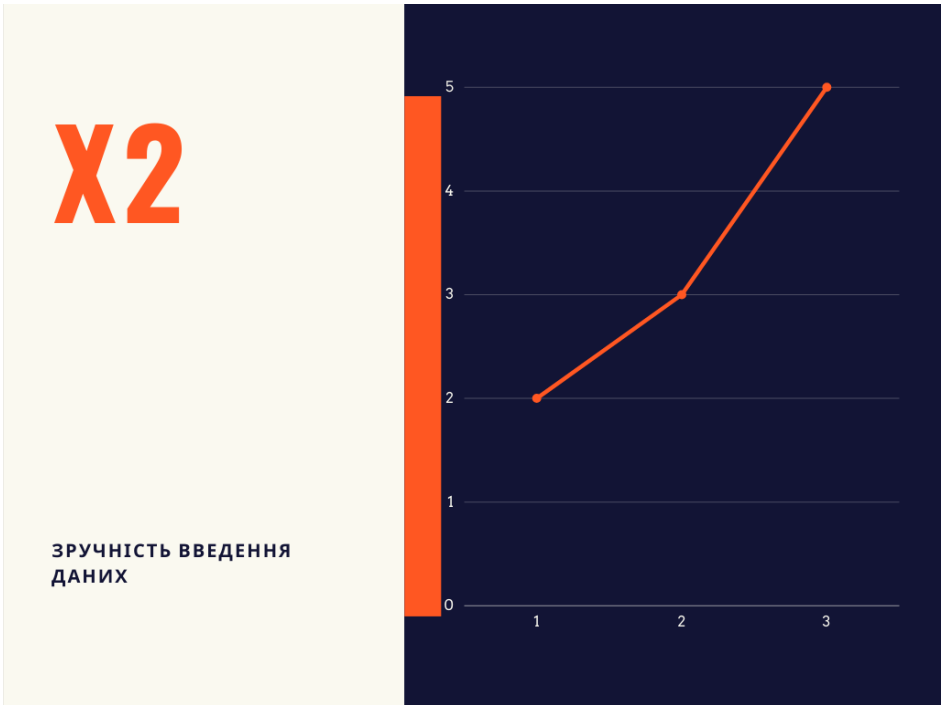

Рисунок 5.3 – X2,зручність введення даних

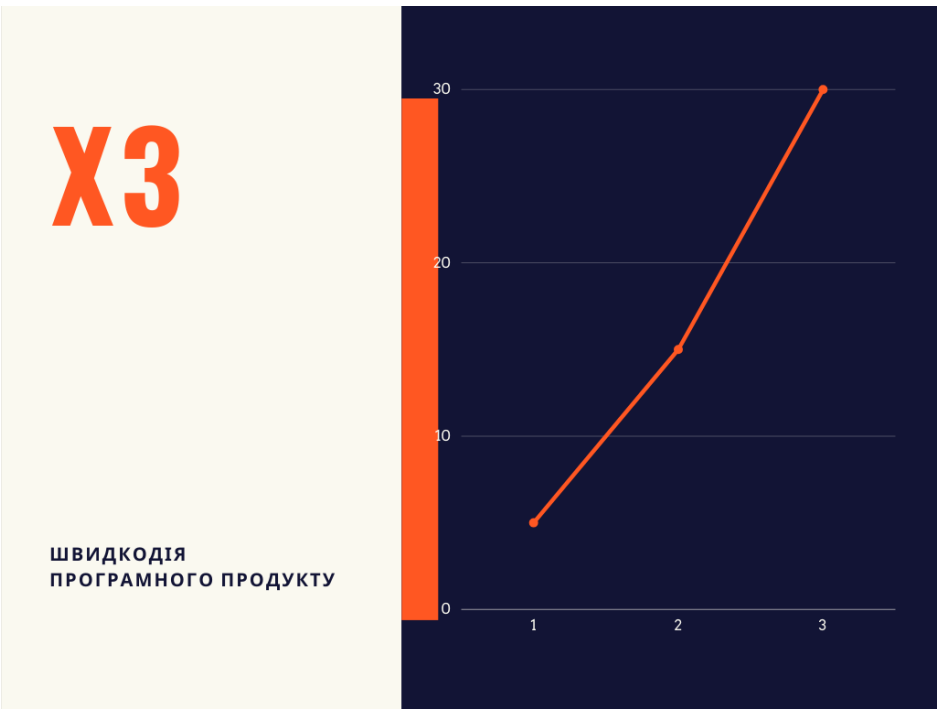

Рисунок 5.4 – X3, швидкодія програмного продукту

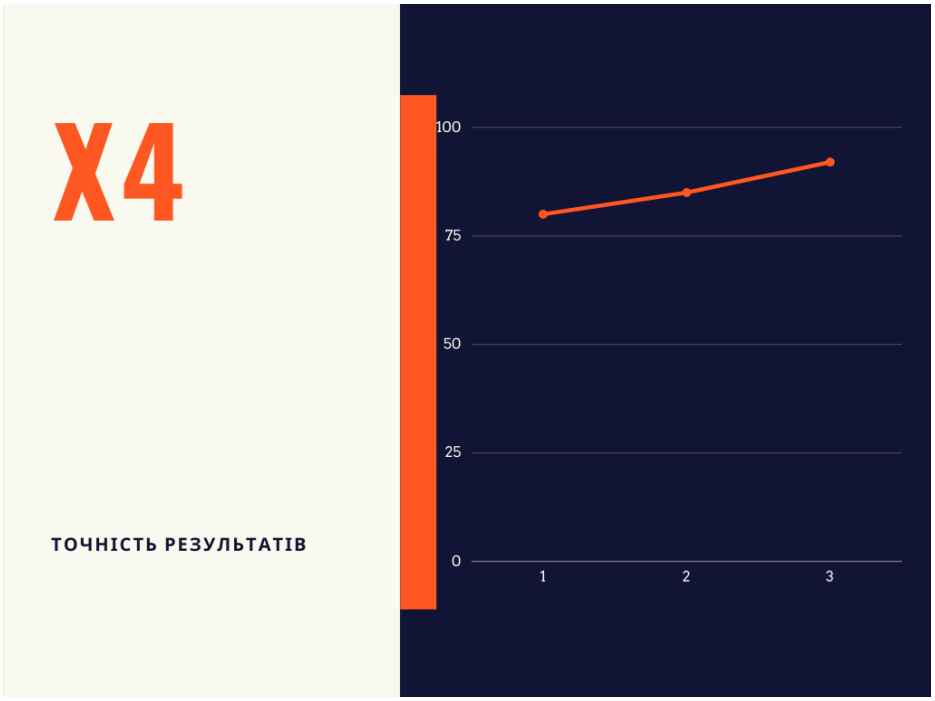

Рисунок 5.5 – X4, точність результатів

5.4 Аналіз експертного оцінювання параметрів

Після детального обговорення й аналізу кожний експерт оцінює ступінь важливості кожного параметру для конкретно поставленої цілі – розробка програмного продукту, який дає найбільш точні результати при знаходженні параметрів моделей адаптивного прогнозування і обчислення прогнозних значень.

Значимість кожного параметра визначається методом попарного порівняння. Оцінку проводить експертна комісія із 7 людей. Ранги варіюються від 1 до 4, де 1 — найменший, а 4 — найбільший.

Визначення коефіцієнтів значущості передбачає:

1. визначення рівня значимості параметра шляхом присвоєння різних рангів;
2. перевірку придатності експертних оцінок для подальшого використання;
3. визначення оцінки попарного пріоритету параметрів;
4. обробку результатів та визначення коефіцієнту значущості.

Розрахуємо коефіцієнт узгодженості:

$$W = \frac{12S}{N^2(n^3-n)} = \frac{12*201}{7^2(4^3-4)} = 0.82$$

Коефіцієнт узгодженості більше нормативного, тому результати можна вважати достовірними.

Скориставшись результатами ранжирування, проведемо попарне порівняння всіх параметрів і результати заносимо у таблицю

Таблиця 5.3 - Результати ранжування параметрів

Назва параметра	Позначення	Одиниці виміру	Ранг параметра за оцінкою експерта							Сума рангів	Відхилення	
			1	2	3	4	5	6	7			
швидкодія мови програмування	X1	оп/мс	4	3	4	3	4	3	4	25	7.5	56.25
зручність введення даних	X2	кількість кроків	1	1	1	2	1	1	1	8	-9.5	90.25
швидкодія програмного продукту	X3	секунди	2	2	3	2	1	2	2	14	-3.5	12.25
точність результатів	X4	відсотки	3	4	2	4	3	4	4	24	6.5	42.25
Разом			10	9	10	11	9	10	11	70	0	201

Таблиця 5.4 – Попарне порівняння параметрів

Параметри	Експерти							Кінцева оцінка	Числове значення
	1	2	3	4	5	6	7		
X1 і X2	>	>	>	>	>	>	>	>	1,5
X1 і X3	>	>	>	>	>	>	>	>	1,5
X1 і X4	>	<	>	<	>	<	=	=	1
X2 і X3	<	<	<	=	=	<	<	<	0,5
X2 і X4	<	<	<	<	<	<	<	<	0,5
X3 і X4	<	<	>	<	<	<	<	<	1,5

Проведемо розрахунок вагомості параметрів. Результати наведено в Таблиці нижче.

Таблиця 5.5 – Розрахунок вагомості параметрів

Параметри	Параметри				Перша ітер.		Друга ітер.		Третя ітер	
	X1	X2	X3	X4	b_i	$K_{вi}$	b_i^1	$K_{вi}^1$	b_i^2	$K_{вi}^2$
X1	1,0	1,5	1,5	1,0	5,0	0,29	21	0,28	91	0,29
X2	1,5	1,0	0,5	0,5	3,5	0,21	15.25	0,21	63,38	0,20
X3	1,5	0,5	1,0	1,5	4,5	0,26	19.75	0,27	85,13	0,27
X4	1,0	0,5	1,5	1,0	4,0	0,24	17.5	0,24	75,75	0,24
Всього:					17,0	1	73.5	1	315,26	1

5.5 Аналіз рівня якості варіантів реалізації функцій

Визначимо рівень якості кожного варіанта виконання основних функцій

окремо. Розрахунки наведено в таблиці нижче.

Таблиця 5.6 – Розрахунок показників рівня якості варіантів реалізації основних функцій програмного продукта

Основна функція	Варіант реалізації	Параметри	Абсолютне значення параметра	Бальна оцінка параметра	Коефіцієнт вагомості параметра	Коефіцієнт якості
F1	a	X1	15000	6	0,29	1,74
F2	a	X4	85	2	0,24	0,48
	b	X4	92	5	0,24	1,2
F3	a	X3	15	2	0,27	0,54

За даними з таблиці 4 визначимо якість кожного з варіантів.

$K_{k1} = 1,74+0,48+0,54 = 2.76$

K_{k2} =1,74 + 1,2 + 0,54 = 3.48

Як видно з розрахунків, кращим є 2й варіант, для якого коефіцієнт технічного рівня має найбільше значення.

5.6 Економічний аналіз варіантів розробки програмного продукту

Для визначення вартості розробки програмного продукту спочатку проведемо розрахунок трудомісткості. Всі варіанти включають в себе два окремих завдання:

- Розробка алгоритму програмного продукту
- Розробка графічного інтерфейса користувача програмного продукту

Завдання 1 за ступенем новизни відноситься до групи А, завдання 2 – до групи Б. За складністю алгоритми, які використовуються в завданні 1 належать до групи 1; а в завданні 2 – до групи 3.

Для реалізації завдання 1 використовується довідкова інформація, а завдання 2 використовує інформацію у вигляді даних. Проведемо розрахунок норм часу на розробку та програмування для кожного з завдань.

$$T_1 = 90 * 1,7 * 0,8 = 122.4 \text{ людино - днів}$$

$$T_2 = 27 * 0,9 * 0,8 = 19,44 \text{ людино - днів}$$

Розрахуємо трудомісткість для кожного з варіантів розробки програми.

$$T_I = (19,44 + 122,4 + 122,4 + 122,4) * 8 = 3093,12$$

$$T_{II} = (19,44 \ + 122,4 + 122,4 + 19,44) * 8 = 2269.44$$

Найбільшу трудомісткість має варіант 1.

Нехай в розробці беруть участь 3 програмісти з окладом в 12000 грн. Визначаємо середню заробітну плату за годину.

$$СЧ \ = \frac{3*12000}{3*21*8} = 71,42$$

Заробітна плата для кожного з варіантів реалізації:

$$Сзп_1 = 71,42 * 3093.12 = 220910,63грн$$

$$Сзп_2 = 71,42 * 2269.44 = 162083,41 \text{ грн}$$

Відрахування на соціальне страхування(22%):

$$Свід_1 = 220910,63 * 0,22 = 48\ 600,3 \text{ грн}$$

$$Свід_2 = 162083,41 * 0,22 = 35\ 658,35 \text{ грн}$$

Визначимо витрати на оплату однієї машиногодини. З ЕОМ обслуговують 3 програмістів з окладом 12’000 грн та коефіцієнтом зайнятості 0.35, для однієї ЕОМ отримаємо.

$$С_Г = 12 * 12\ 000 * 0,35 = 50400 \text{ грн}$$

З урахуванням додаткової заробітної плати:

$$С_{зп} = 50\ 400 * (1 + 0.25) = 63\ 000 \text{ грн}$$

Відрахування на соціальний внесок:

$$С_{від} = 63\ 000 * 0,22 = 13\ 860 \text{ грн}$$

Розрахуємо амортизаційні підрахунки (амортизація 25%, вартість ЕОМ 25000 грн)

$$С_А = К_{ТМ} \cdot К_А \cdot Ц_{ПР}$$

$$С_А = 1,15 * 0,25 * 25\,000 = 7187,5 \text{ грн}$$

Розрахуємо витрати на ремонт та профілактику

$$С_Р = К_{ТМ} \cdot Ц_{ПР} \cdot К_Р$$

$$С_Р = 1,15 * 0,05 * 25\,000 = 1437,5 \text{ грн}$$

Розрахуємо ефективний годинний фонд часу ПК за рік

$$Т_{ЕФ} = (Д_К – Д_В – Д_С – Д_Р) \cdot t_З \cdot К_В = (365 – 138 – 16) \cdot 8 \cdot 0.8 =$$

$$= 1350.4 \text{ годин},$$

де $Д_К$ – календарна кількість днів у році;

$Д_В$, $Д_С$ – відповідно кількість вихідних та святкових днів;

$Д_Р$ – кількість днів планових ремонтів устаткування;

t –кількість робочих годин в день;

$К_В$– коефіцієнт використання приладу у часі протягом зміни.

Розрахуємо витрати на електроенергію

$$С_{ЕЛ} = Т_{ЕФ} \cdot N_С \cdot К_З \cdot Ц_{ЕН} = 1350.4 \cdot 0,3 \cdot 0,7 \cdot 3,52 = 998,21 \text{ грн.},$$

Накладні витрати дорівнюють:

$$С_Н = Ц_{ПР} \cdot 0.67 = 25000 \cdot \ 0{,}67 = 16750 \text{ грн.}$$

Отже експлуатаційні витрати:

$$С_{екс} = 63\,000 + 13\,860 + 7187{,}5 + 1437{,}5 + 998{,}21 + 16750 = 103233.21 \text{ грн}$$

Тоді собівартість однієї машиногодини ЕОМ дорівнюватиме:

$$С_{м-г} = \frac{103233.21}{1350{,}4} = 76{,}45$$

Враховуючи, що всі роботи ведуться на ЕОМ, витрати на оплату машинного часу:

$$С_{м1} = 76{,}45 * 3093.12 = 236469.02 \text{ грн}$$

$$С_{м2} = 76{,}45 * 2269.44 = 173498.69 \text{ грн}$$

Накладні витрати складають 67% від заробітної плати:

$$С_{н1} = 220910{,}63 * 0{,}67 = 148010.12 \text{ грн}$$

$$С_{н2} = 162083{,}41 * 0{,}67 = 108595.88 \text{ грн}$$

Отже, вартість розробки ПП за варіантами становить:

$$С_{пп1} = 220910{,}63 + 48\,600{,}3 + 236469.02 + 148010.12 = 653990.07 \text{ грн}$$

$$С_{пп2} = 162083{,}41 + 35\,658{,}35 + 173498.69 + 108595.88 = 479836.33 \text{ грн}$$

5.7 Вибір кращого варіанта програмного продукту техніко - економічного рівня

$$К_{к1} = 2.76$$

$$К_{к2} = 3.48$$

$$К_{тер1} = \frac{2,76}{653990.07} = 4,22 * 10^{-6}$$

$$К_{тер2} = \frac{3,48}{479836.33} = 7,25 * 10^{-6}$$

5.8 Висновки до розділу 5

Проаналізувавши розрахунки, можемо зробити висновок, що найбільш ефективним буде варіант розробки 2 з коефіцієнтом техніко - економічного рівня $7,25 * 10{-}6$. Таким чином, після проведеного функціонально – вартісного аналізу було прийняте рішення реалізовувати варіант за допомогою мови програмування Python, алгоритм BERT, а саме модель RoBERTa.

ВИСНОВКИ

В даній дипломній роботі було досліджено області прикладних застосувань, для яких актуально дослідження емоціональної забарвленості тексту, яких існує безліч від особистого використання, маркетингу, бізнес - питань і до моніторингу кредитних ринків.

Задача сентимент аналізу є поширеною,тому існують безліч методів машинного навчання з вчителем/без вчителя,методи засновані на правилах та словниках тощо. В даній роботі було розглянуто автоматизовані методи рекурентних нейронних мереж, згорткових нейронних мереж та двонаправленого трансформера BERT для задачі аналізу тональності тексту. Спираючись на результати аналізу методів,для використання попередньо навченої моделі в телеграм - боті , було обрано RoBERTa - LARGE.

В ході роботи було реалізовано телеграм - бот з функціоналом сентимент аналізу з підтримкою англійської мови. Бот повертає результат про класифікацію емоційного забарвлення повідомлення,результат надається у вигляді повідомлення з текстом "NEGATIVE" , якщо дані були класифіковані як негативний окрас, та "POSITIVE", якщо дані були класифіковані як позитивний окрас.

ЛІТЕРАТУРА

1. Анализ тональности текста: концепция, методы, области применения [Електронний ресурс] - Режим доступу до ресурсу: http://datareview.info/article/analiz-tonalnosti-teksta-kontseptsiya-metody i-oblasti-primeneniya/
2. АНАЛИЗ ЭМОЦИОНАЛЬНОЙ ОКРАСКИ СООБЩЕНИЙ В СОЦИАЛЬНЫХ СЕТЯХ**,** И. Е. Воронина, В. А. Гончаров [Електронний ресурс] - Режим доступу до ресурсу: http://www.vestnik.vsu.ru/pdf/analiz/2015/04/2015-04-21.pdf
3. SentimentAnalysis:ADefinitiveGuide [Електронний ресурс] - Режим доступу до ресурсу: https://monkeylearn.com/sentiment-analysis/
4. SentimentAnalysis:Types,Tools,andUseCases [Електронний ресурс] - Режим доступу до ресурсу: https://www.altexsoft.com/blog/business/sentiment-analysis-types-tools-and-use-cases/
5. SentimentAnalysisinBanking–4CurrentUse-Cases [Електронний ресурс]- Режим доступу до ресурсу: https://emerj.com/ai-sector-overviews/sentiment-analysis-banking/
6. Аналіз тональності тексту [Електронний ресурс]- Режим доступу до ресурсу: https://uk.wikipedia.org/wiki/Аналіз_тональності_тексту
7. Вступ до машинного навчання [Електронний ресурс]- Режим доступу до ресурсу: http://specials.kunsht.com.ua/machinelearning2
8. BERT (модель мови) [Електронний ресурс]- Режим доступу до ресурсу: https://uk.wikipedia.org/wiki/BERT_(модель_мови)
9. BERT Explained: State of the art language model for NLP [Електронний ресурс]- Режим доступу до ресурсу: https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

10. Word2vec [Електронний ресурс]- Режим доступу до ресурсу: https://ru.wikipedia.org/wiki/Word2vec
11. Применение сверточных нейронных сетей для задач NLP [Електронний ресурс]- Режим доступу до ресурсу: https://habr.com/ru/company/ods/blog/353060/
12. Рекурентна нейронна мережа [Електронний ресурс]- Режим доступу до ресурсу: https://uk.wikipedia.org/wiki/Рекурентна_нейронна_мережа
13. Text Classification with RNN [Електронний ресурс]- Режим доступу до ресурсу: https://towardsai.net/p/deep-learning/text-classification-with-rnn
14. Модель RoBERTa[Електронний ресурс]- Режим доступу до ресурсу: https://huggingface.co/siebert/sentiment-roberta-large-english
15. To Pretrain or Not to Pretrain: Examining the Benefits of Pretraining on Resource Rich Tasks [Електронний ресурс]- Режим доступу до ресурсу: https://www.aclweb.org/anthology/2020.acl-main.200.pdf
16. RoBERTa: A Robustly Optimized BERT Pretraining Approach [Електронний ресурс]- Режим доступу до ресурсу: https://arxiv.org/pdf/1907.11692.pdf

ДОДАТОК

ДОДАТОК А
ЛІСТИНГ КОДУ

```
import telebot
import config
from transformers import pipeline
bot = telebot.TeleBot(config.TOKEN)
@bot.message_handler(commands=['start'])
def handle_text (message):
    bot.send_message(message.chat.id, "Enter text for classification in English.")
    @bot.message_handler(content_types=['text'])
    def handle_text(message):
        txt = message.text
sentiment_analysis = pipeline("sentiment-analysis",model="siebert/sentiment-roberta-large-english")
        msg = sentiment_analysis(txt)
        dict = { }
        for i in msg:
            dict.update(i)
            res = dict["label"]
        bot.send_message(message.chat.id,"This text is classified as " + x)
if __name__ == '__main__':
    bot.polling(none_stop=True)
```