Міністерство освіти і науки України

Харківський національний університет радіоелектроніки

Факультет комп’ютерної інженерії та управління
(повна назва)

Кафедра електронних обчислювальних машин
(повна назва)

# КВАЛІФІКАЦІЙНА РОБОТА

## Пояснювальна записка

Рівень вищої освіти другий (магістерський)

Методи контент-аналізу в комп’ютерній лінгвістиці для автоматизованого маркування емоційно-забарвленої лексики
(тема)

Виконав:

студент II курсу, групи СПм-22-3

Захаров Д.О.
(прізвище, ініціали)

Спеціальність 123 «Комп’ютерна інженерія»
(код і повна назва спеціальності)

Тип програми освітньо-наукова
(освітньо-професійна або освітньо-наукова)

Освітня програма Системне програмування
(повна назва освітньої програми)

Керівник: доц. Іващенко Г.С.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри ЕОМ _____ Коваленко А.А.
(підпис) (прізвище, ініціали)

2024 р.

Харківський національний університет радіоелектроніки

Факультет комп'ютерної інженерії та управління

Кафедра електронних обчислювальних машин

Рівень вищої освіти другий (магістерський)

Спеціальність 123 «Комп'ютерна інженерія»
(код і повна назва)

Тип програми освітньо-наукова
(освітньо-професійна або освітньо-наукова)

Освітня програма Системне програмування
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
"____" _____ 20__ р.

# ЗАВДАННЯ

## НА КВАЛІФІКАЦІЙНУ РОБОТУ

студенту Захарову Даніїлу Олександровичу
(прізвище, ім'я, по батькові)

1. Тема роботи Методи контент-аналізу в комп'ютерній лінгвістиці для автоматизованого маркування емоційно-забарвленої лексики

затверджена наказом по університету від " 01 " квітня 2024 р. № 257Ст

2. Термін подання студентом роботи до екзаменаційної комісії 15 червня 2024 р.

3. Вхідні дані до роботи
наявність підключення до мережі інтернет для підгрузки датасетів,
GPU RTX 3060 12GB або краще (для використовуванняч або начання моделей)
системне середовище Python.

4. Перелік питань, що потрібно опрацювати у роботі
Огляд методів аналізу текстового контенту та класифікації за типами емоцій та настроїв
Аналіз нейромережевих лінгвістичних моделей для реалізації модулів семантичного, емоційного та тематичного аналізу
Створення функціональної моделі системи
Розробка методології проведення досліджень
Проведення експериментів
Аналіз отриманих результатів

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (слайдів) 22 слайди

6. Консультанти розділів роботи (заповнюється за наявності консультантів згідно з наказом, зазначеним у п.1 )

| Найменування розділу | Консультант (посада, прізвище, ім'я, по батькові) | Позначка консультанта про виконання розділу | |
|---|---|---|---|
| | | підпис | дата |
| | | | |
| | | | |

## КАЛЕНДАРНИЙ ПЛАН

| № | Назва етапів роботи | Термін виконання етапів роботи | Примітка |
|---|---|---|---|
| 1 | Огляд методів аналізу текстового контенту та | 02.04.24-08.04.24 | |
| | класифікації за типами емоцій та настроїв | | |
| 2 | Аналіз нейромережевих лінгвістичних моделей | 09.04.24-16.04.24 | |
| | для реалізації модулів семантичного | | |
| | емоційного та тематичного аналізу | | |
| 3 | Створення функціональної моделі системи | 17.04.24-22.04.24 | |
| 4 | Розробка методології проведення досліджень | 23.04.24-06.05.24 | |
| 5 | Проведення експериментів | 07.05.24-23.05.24 | |
| 6 | Оформлення матеріалів кваліфікаційної роботи | 24.05.24-03.06.24 | |
| 7 | Подання кваліфікаційної роботи керівникові та | 04.06.24-07.06.24 | |
| | її попередній захист | | |
| 8 | Подання кваліфікаційної роботи на рецензу- | 08.06.24-12.06.24 | |
| | вання | | |

Дата видачі завдання 01 квітня 2024 р.

Студент _____ (підпис)

Керівник роботи _____ доц.Іващенко Г.С.
(підпис) (посада, прізвище, ініціали)

# РЕФЕРАТ

Пояснювальна записка кваліфікаційної роботи: 168 с., 14 рис., 38 табл., 2 дод., 20 джерел.

МОДЕЛЬ, НЕЙРОННА МЕРЕЖА, СИСТЕМА, МЕТОД, МОДЕЛЬ, ПРОМПТ, ГЕНЕРАЦІЯ, АНАЛІЗ ТЕКСТУ, КЛАСИФІКАЦІЯ ЕМОЦІЙ, АНАЛІЗ СЕНТИМЕНТУ, НЕР, ТЕМАТИЧНЕ МОДЕРУВАННЯ, ЕМОЦІЙНО ЗАБАРВЛЕНА ЛЕКСИКА, РЕГУЛЮВАННЯ ТЕКСТУ, НЛП, МАШИННЕ НАВЧАННЯ, ШІ, СЕМАНТИЧНИЙ АНАЛІЗ, ТЕМАТИЧНИЙ АНАЛІЗ, ТЕКСТОВИЙ МАЙНІНГ, ГЕНЕРАЦІЯ КОНТЕНТУ, ТРЕНІНГ ОПТИМІЗАЦІЯ, АНАЛІЗ ТЕКСТОВИХ ДАНИХ.

У цій кваліфікаційній роботі досліджується розробка інноваційних методів контент-аналізу текстової інформації, збагаченої емоційно забарвленою лексикою. Основна мета — створити універсальну систему, здатну аналізувати, узагальнювати та інтерпретувати тексти для виявлення прихованих емоцій, настроїв і тем. Така система має значний потенціал для різноманітних додатків, включаючи модерацію контенту та генерацію з урахуванням конкретних аудиторій на основі аналізу вхідних текстових масивів.

Наше дослідження складається з кількох важливих компонентів, починаючи з вивчення методів аналізу текстового вмісту різної довжини. Ми зосередилися на класифікації текстових масивів за типами емоцій і настроїв, використовуючи лінгвістичні моделі нейронної мережі для семантичного, емоційного та тематичного аналізу. Це передбачало всебічний огляд існуючих наборів даних для інформування на етапах навчання та тестування нашої моделі.

Ключовим досягненням стала розробка концептуальної моделі

позначення емоційно забарвленої лексики в текстах. Ця модель об’єднує чотири модулі: аналіз почуття, розпізнавання іменованих об’єктів (NER), класифікація емоцій і моделювання теми. Кожен модуль відіграє вирішальну роль у аналізі тексту, щоб забезпечити детальну інтерпретацію його змісту. Аналіз настроїв визначає загальну тональність, NER ідентифікує та класифікує названі сутності, класифікація емоцій кількісно визначає присутні емоції, а тематичне моделювання визначає обговорювані теми.

Практична реалізація включала модуль попередньої обробки для очищення тексту та поділу на фрагменти, забезпечуючи належну підготовку вхідних даних. Модуль аналізу застосував інтегровані методи для отримання детальної інформації з тексту. Згодом модуль генерації використовував попередньо визначені шаблони підказок для створення маркерів і підказок для генеративних моделей штучного інтелекту, що забезпечувало генерацію контенту відповідно до контексту.

Було проведено масштабні експерименти для оптимізації параметрів навчання, включаючи розмір партії, розмір набору даних, кількість епох і тип оптимізатора. Також було оцінено використання методів масштабування для підвищення ефективності навчання класифікаторів нейронних мереж. Результати підтвердили здатність моделі точно інтерпретувати та класифікувати складні текстові дані, демонструючи її практичне застосування в аналізі та створенні контенту.

На завершення ця теза представляє складну систему для аналізу емоційно забарвленої текстової інформації, інтегруючи аналіз настроїв, NER, класифікацію емоцій і моделювання тем у єдину структуру. Дослідження просуває сферу аналізу тексту, пропонуючи практичне застосування в різних сферах, надаючи детальне уявлення про настрій, емоції та теми текстового вмісту. Ця робота сприяє розвитку чуйних і адаптивних систем, сприяючи проактивним діям на основі комплексного аналізу тексту.

# ABSTRACT

Master’s thesis: 168 pages, 48 figures, 17 tables, 2 appendices, 35 sources.

MODEL, NEURAL NETWORK, SYSTEM, METHOD, MODEL, PROMPT, GENERATION, TEXT ANALYSIS, EMOTION CLASSIFICATION, SENTIMENT ANALYSIS, NER, THEMATIC MODERATION, EMOTIONALLY COLORED VOCABULARY, TEXT REGULATION, NLP, MACHINE LEARNING, AI , SEMANTIC ANALYSIS, THEMATIC ANALYSIS, TEXT MINING, CONTENT GENERATION, TRAINING OPTIMIZATION, TEXT DATA ANALYSIS.

This thesis explores the development of innovative methods for content analysis of textual information enriched with emotionally colored vocabulary. The primary objective is to create a versatile system capable of analyzing, summarizing, and interpreting texts to identify underlying emotions, sentiments, and themes. Such a system holds significant potential for diverse applications, including content moderation and generation tailored to specific audiences based on the analysis of input text arrays.

Our research comprises several critical components, beginning with an exploration of methods for analyzing text content of various lengths. We focused on classifying text arrays by types of emotions and moods, utilizing neural network linguistic models for semantic, emotional, and thematic analysis. This involved a comprehensive review of existing datasets to inform our model's training and testing phases.

A key achievement was the development of a conceptual model designed to mark emotionally colored vocabulary within texts. This model integrates four modules: sentiment analysis, named entity recognition (NER), emotion classification, and topic modeling. Each module plays a crucial role in dissecting

the text to provide a nuanced interpretation of its content. Sentiment analysis discerns the overall tonality, NER identifies and categorizes named entities, emotion classification quantifies the emotions present, and topic modeling determines the discussed themes.

The practical implementation involved a preprocessing module for text cleaning and chunking, ensuring the input data was adequately prepared. The analysis module applied the integrated methods to extract detailed insights from the text. Subsequently, the generation module used predefined prompt patterns to create markers and prompts for generative AI models, enabling contextually appropriate content generation.

Extensive experimentation was conducted to optimize training parameters, including batch size, dataset size, number of epochs, and type of optimizer. The use of scaling techniques was also evaluated to enhance the efficiency of training neural network classifiers. The results validated the model's capability to accurately interpret and categorize complex textual data, demonstrating its practical applications in content analysis and generation.

In conclusion, this thesis presents a sophisticated system for analyzing emotionally colored textual information, integrating sentiment analysis, NER, emotion classification, and topic modeling into a cohesive framework. The research advances the field of text analysis, offering practical applications in various domains by providing detailed insights into the mood, emotions, and themes of textual content. This work contributes to the development of responsive and adaptive systems, facilitating proactive actions based on comprehensive text analysis.

ЗМІСТ

# ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

AI - штучний інтелект (англ. Artificial Intelligence)

НЛП - обробка природної мови (англ. Natural Language Processing)

NER - розпізнавання іменованих сутностей (англ. Named Entity Recognition)

BERT - Представлення двонаправленого кодера від Transformers

BERTopic – Представлення двонаправленого кодера від Transformers для тематичного моделювання

LSTM - довга короткочасна пам'ять (англ. Long Short-Term Memory)

RNN - рекурентна нейронна мережа (англ. Recurrent Neural Network)

CNN - згорточна нейронна мережа (англ. Convolutional Neural Network)

POS - частина мови (англ. Part-of-Speech)

EmoBank - корпус тексту, анотованого для емоцій

TEC - Textual Emotion Corpus

EmoInt - Інтенсивність емоцій

SSEC - Корпус сарказмових настроїв

ISEAR - Міжнародне дослідження попередніх емоцій і реакцій

Трансформатор - тип архітектури нейронної мережі

GPU - графічний процесор (англ. Graphics Processing Unit)

API - інтерфейс прикладного програмування (англ. Application Programming Interface)

CSV - значення, розділені комами (англ. Comma-Separated Values)

JSON – нотація об’єктів JavaScript

ML - машинне навчання (англ. Machine Learning)

DNN - глибока нейронна мережа (англ. Deep Neural Network)

PNN - імовірнісна нейронна мережа (англ. Probabilistic Neural Network)

ASR - автоматичне розпізнавання мови (англ. Automatic Speech

Recognition)

OCR - оптичне розпізнавання символів (англ. Optical Character Recognition)

BLEU - двомовне оцінювання дублера (English Bilingual Evaluation Understudy)

ROUGE - Орієнтований на запам'ятовування дублер для оцінки Гістінга

GPT - Генеративний попередньо навчений трансформатор

RoBERTa - надійно оптимізований підхід до підготовки BERT

XLNet - узагальнена авторегресійна попередня підготовка для розуміння мови

DistillBERT - дистильований БЕРТ

# ВСТУП

Величезний океан текстових даних продовжує розширюватись безпрецедентними темпами. Платформи соціальних мереж, новинні агенції, цифрові бібліотеки та незліченні онлайн-взаємодії створюють постійно зростаючий обсяг інформації, що очікує вивчення та розуміння. Вилучення важливої інформації з цих даних має величезний потенціал широкого спектра додатків: від оцінки суспільних настроїв до розуміння переваг клієнтів, від відстеження спалахів захворювань до аналізу ринкових тенденцій.

В основі цього починання лежить область контент-аналізу. Контент-аналіз забезпечує нас інструментами для систематичного вилучення та інтерпретації інформації з текстових даних. Ця область включає широкий спектр методів: від традиційного зіставлення ключових слів до складних моделей машинного навчання. Однак існуючі системи контент-аналізу часто стикаються з обмеженнями у можливості вловити всю складність людського вираження у тексті.

Ця робота заглиблюється в область контент-аналізу, приділяючи особливу увагу розробці нової системи, яка може подолати ці обмеження. Зокрема, запропонована модель спрямована на вирішення трьох найважливіших аспектів контент-аналізу:

- аналіз настроїв та класифікація емоцій – виходячи за рамки базових позитивних/негативних настроїв, модель прагне ідентифікувати ширший спектр емоцій із більшою деталізацією;
- розпізнавання іменованих об'єктів – модель використовуватиме досягнення в галузі глибокого навчання для точної ідентифікації іменованих об'єктів у тексті, таких як люди, організації та місця розташування;
- тематичне вилучення – виходячи за рамки ідентифікації ключових слів, модель прагне отримати основні тематичні структури та відносини всередині тексту, забезпечуючи глибше розуміння змісту та ідей, які він

передає.

Основна мета цього дослідження – розробити систему, яка може генерувати «маркери» – уявлення – які інкапсулюють емоційний тон та тематичні області, витягнуті з аналізованих текстових даних. Ці «маркери» можна використовувати для різних додатків, таких як створення зведень емоційних тенденцій у розмовах у соціальних мережах або візуалізація тематичного ландшафту корпусу новин.

Одна з основних цілей цієї роботи полягає в розробці нової системи контент-аналізу, яка не тільки отримує корисну інформацію з текстових даних, але і представляє її у візуально привабливій формі. Це досягається за рахунок концепції маркерів.

Маркери є важливим посередником між аналізованими текстовими даними та сферою візуальної комунікації. Вони діють як міст, переводячи емоційний тон та тематичні області, витягнуті з тексту, у візуальний формат, який легко зрозумілий та ефективний. Цей формат може набувати різних форм, таких як зображення, значки або навіть візуалізації даних, залежно від конкретної програми та бажаного результату.

Поліпшення розуміння. Представляючи складну інформацію, витягнуту з тексту, у візуальному форматі, маркери можуть сприяти більш інтуїтивному та доступному розумінню даних. Візуальні уявлення можуть легко передати емоційні тенденції, тематичні ландшафти чи переважання певних іменованих об'єктів у тексті. Ця візуальна комунікація може бути особливо ефективною для аудиторії, яка не завжди розуміє нюанси текстового аналізу або не має часу вникати у докладні звіти.

Підвищення залучення. Візуальні уявлення, властиві маркерам, за своєю суттю привабливіші, ніж необроблені текстові дані. Вони привертають увагу, стимулюють когнітивний процес та дозволяють швидше зрозуміти ключові ідеї. Таке зростання може мати вирішальне значення для глибшого розуміння аналізованого контенту та історій, які він містить.

Хоча основна увага в цій робти приділяється маркерам для

дослідницьких цілей, їхнє потенційне застосування виходить за рамки академічної сфери. Уявіть собі платформу новин, що використовує маркери для візуального представлення емоційного підґрунтя сенсаційної історії. Платформи соціальних мереж можуть використовувати маркери для створення зведення актуальних тем, виділяючи домінуючі емоції та теми в обговореннях користувачів. Навіть індустрія розваг могла б отримати вигоду від цієї концепції: маркери використовуються для візуального відображення емоційної подорожі та тематичних ниток у розповіді, збагачуючи враження від перегляду.

Розробка ефективної системи контент-аналізу, яка генерує ці потужні маркери, відкриває шлях до детальнішого та захоплюючого дослідження текстових даних. Подолавши розрив між текстом та значними візуальними ефектами, маркери відкривають нові можливості для розуміння людського спілкування, аналізу тенденцій та сприяння більш глибокій взаємодії з величезним океаном інформації, що нас оточує.

Решта дисертації побудована таким чином. Розділ 2 присвячений теоретичним основам дослідження, вивченню існуючих систем контент-аналізу та опису базової архітектури запропонованої моделі.

# 1 АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ

## 1.1 Обґрунтування актуальності обраної теми

У сучасному інформаційному суспільстві стикаємося з безперервним потоком різноманітного контенту, що охоплює різні аспекти життя та діяльності. Різноманітність контенту включає тексти, зображення, аудіо- та відеоматеріали, а також дані, зібрані з соціальних мереж, форумів, блогів, порталів новин та інших онлайн-ресурсів.

При аналізі контенту можна назвати кілька основних категорій, які переважають у сучасному цифровому просторі. Насамперед слід зазначити текстовий контент, який залишається одним із найпоширеніших і доступних форматів інформації. Тексти включають статті, відгуки, коментарі, соціальні пости, електронні листи та інші форми письмового вираження думок та ідей.

Однак, окрім текстів, значну частку контенту становлять зображення. З розвитком соціальних мереж та мобільних технологій знімки, малюнки, меми, фотографії та інші візуальні матеріали стають все більш популярними серед користувачів інтернету. Зображення як служать засобом візуального подання інформації, а й можуть нести смислове навантаження, висловлювати емоції і викликати реакції в аудиторії.

Не слід забувати і про аудіо- та відеоматеріали, які займають істотну частку в цифровому контенті. Подкасти, музичні композиції, відеоролики, стрими, аудіо- та відеозаписи різних заходів та трансляцій – все це форми аудіовізуального контенту, які активно споживаються користувачами в Інтернеті.

Крім того, дані, зібрані із соціальних мереж, форумів та інших онлайн-ресурсів, є окремим типом контенту, який має свою специфіку та цінність для аналізу. Ці дані можуть включати текстові повідомлення, коментарі, лайки, репости, а також метадані і т.д.

Усвідомлення актуальності вивчення та аналізу текстового контенту обумовлено його широким застосуванням та значимістю у суспільстві. Текстовий контент пронизує всі сфери нашого життя: від ділового листування та наукових досліджень до спілкування в соціальних мережах та споживання новин. Він є основним інструментом передачі інформації, вираження думок та ідей, а також формування громадської думки. Тому аналіз текстового контенту стає необхідним інструментом для розуміння трендів, оцінки суспільного настрою, виявлення потреб аудиторії та прогнозування розвитку подій.

Серед різних форматів контенту текст залишається одним із найбільш зручних для аналізу, оскільки він є структурованою послідовністю слів та речень, яка легко піддається обробці за допомогою комп'ютерних алгоритмів та методів. Текстовий аналіз дозволяє виявляти різні характеристики тексту, такі як тема, тональність, ключові слова, семантичні зв'язки та багато іншого, що робить його важливим інструментом для дослідження та розуміння інформаційного простору.

Важливо, що текстовий контент несе у собі не тільки об'єктивну інформацію, а й різні емоційні і суб'єктивні відтінки, які можуть істотно впливати на сприйняття й інтерпретацію тексту. Емоційна забарвленість тексту може бути як явною і вираженою, так і прихованою і неочевидною, що робить її виявлення та аналіз ще більш важливими в контексті розуміння змісту та контексту повідомлень.

Враховуючи всю вищевикладену інформацію, можна зробити висновок, що текстовий контент відіграє ключову роль в інформаційному суспільстві і вимагає розробки ефективних методів аналізу. Тому в цій роботі ми будемо зосереджені на вивченні методів контент-аналізу в комп'ютерній лінгвістиці з метою автоматизованого маркування емоційно забарвленої лексики у текстовому контенті.

Одним із найбільш очевидних місць, де ми зустрічаємо текстовий контент, є засоби масової інформації, такі як газети, журнали, радіо та

телебачення. Статті, репортажі, коментарі, рекламні матеріали та інші форми текстового контенту використовуються для інформування, розваги та формування громадської думки. Якість і точність текстового матеріалу в засобах масової інформації має велике значення, оскільки вони впливають на сприйняття та розуміння подій у суспільстві.

Інтернет є ще одним місцем, де текстовий контент представлений у величезному обсязі та різноманітності. Веб-сайти, блоги, соціальні мережі, форуми, онлайн-журнали, електронні листи, коментарі до статей та багато інших онлайн-ресурсів активно використовуються для обміну інформацією, спілкування, проведення маркетингових кампаній і навіть навчання. Текстовий контент в інтернеті відіграє ключову роль у формуванні поглядів, переконань та уявлень користувачів, а також впливає на їхню поведінку та прийняття рішень.

Освітні установи також насичені текстовим контентом. Підручники, лекції, статті, наукові дослідження, доповіді, реферати, есе та інші форми письмового матеріалу використовуються для передачі знань, навчання та розвитку учнів та студентів. Якість навчального матеріалу та його доступність мають прямий вплив на освітній процес та результати навчання.

Текстовий контент також присутній у діловій сфері, де він відіграє важливу роль у комунікації між співробітниками та партнерами, складанні документації, рекламних матеріалів, звітів, презентацій та багатьох інших аспектах організаційної діяльності. Якість ділового текстового контенту має прямий вплив на ефективність комунікації та репутацію організації.

Зрештою, текстовий контент пронизує й особисту сферу нашого життя. Це можуть бути особисті записи, щоденники, електронні листи, повідомлення в месенджерах, пости у соціальних мережах та багато іншого. Через текстовий контент ми висловлюємо свої думки, почуття, емоції, робимо записи про свої досягнення та переживання, спілкуємося з друзями та близькими.

Таким чином, текстовий контент знаходиться всюди в нашому житті,

впливаючи на наше мислення, поведінку та сприйняття світу. Його значущість та різноманітність роблять його ключовим об'єктом вивчення та аналізу в рамках комп'ютерної лінгвістики, особливо в контексті автоматизованого маркування емоційно забарвленої лексики.

З розвитком інформаційних технологій та розширенням інтернет-простору, текстова інформація залишається одним із основних джерел даних. У той час як потік текстової інформації, що створюється та розповсюджується в мережі Інтернет, постійно зростає, включаючи такі різноманітні джерела, як соціальні медіа, блоги, сайти новин, форуми та інші онлайн-платформи.

Серед цього величезного обсягу даних багато інформації, що містить емоційне забарвлення. Це можуть бути відгуки користувачів про продукти та послуги, коментарі в соціальних мережах, обговорення у форумах та багато іншого. Розуміння емоційної тональності тексту має важливе значення для багатьох програм, включаючи аналіз громадської думки, моніторинг бренду, оцінку якості обслуговування, а також для прийняття рішень у маркетингу та управлінні.

Однак, ручна обробка та аналіз великих обсягів текстових даних є дуже трудомістким та витратним процесом. У зв'язку з цим виникає потреба у розробці автоматизованих методів аналізу тексту з метою виявлення та класифікації емоційно забарвленої лексики. Такі методи дозволять не тільки збільшити швидкість аналізу, а й покращити його точність та надійність.

У цьому контексті вивчення та розробка методів контент-аналізу в комп'ютерній лінгвістиці для автоматизованого маркування емоційно забарвленої лексики є актуальним і перспективним завданням, яке має велике значення як для наукової спільноти, так і для практичного застосування в різних галузях.

### 1.1.1 Тренди в галузі комп'ютерної лінгвістики

Останніми роками у сфері комп'ютерної лінгвістики відбулися метаморфози, спричинені злиттям двох важливих чинників: вибухового зростання цифрових текстових даних та безпрецедентних досягнень у галузі штучного інтелекту (ІІ). Ця конвергенція започаткувала революцію, засновану на даних, змінивши те, як ми взаємодіємо, аналізуємо та здобуваємо сенс із природної мови.

Цифрове століття відкрило епоху безпрецедентного створення даних. Пости в соціальних мережах, новинні статті в Інтернеті, відгуки клієнтів і безліч інших форм текстових даних створюються постійно, що є серйозною проблемою для традиційних методів аналізу. Обробка природної мови (NLP) стала наріжним каменем технології вирішення цієї проблеми.

Таблиця 1.1 – Задачі напрямків NLP

| Напрямок | Задача |
|---|---|
| Аналіз настроїв | присвоєння текстам емоційної полярності (позитивної, негативної чи нейтральної), що дозволяє компаніям оцінювати задоволеність клієнтів та відстежувати сприйняття бренду в Інтернеті |
| Машинний переклад | подолання мовних бар'єрів шляхом автоматичного перекладу тексту з однієї мови на іншу, сприяння спілкуванню та співробітництву через міжнародні кордони |
| Автоматизоване узагальнення тексту | стиснення великих обсягів текстової інформації в короткі зведення, що заощаджує дорогоцінний час користувачів і сприяє ефективному пошуку інформації |

NLP дозволяє комп'ютерам обробляти та розуміти нюанси людської мови. Сюди входять такі завдання, як аналіз настроїв, машинний переклад або автоматизоване узагальнення тексту.

Здатність NLP ефективно обробляти великі набори даних має вирішальне значення в епоху великих даних. Традиційні підходи, що ґрунтуються на правилах, часто стикаються з величезним обсягом та складністю сучасних текстових даних. Однак NLP використовує передові алгоритми та статистичні методи для вилучення значущої інформації з цих даних, що робить його незамінним інструментом у різних галузях.

Машинне навчання стало рушійною силою нещодавнього сплеску можливостей NLP (рисунок 1.1). Ці алгоритми навчаються на величезних обсягах розмічених даних, що дозволяє їм виявляти закономірності та виконувати складні завдання без програмування.


Рисунок 1.1 – Застосування машинного навчання у комп'ютерній лінгвістиці

У комп'ютерній лінгвістиці машинне навчання знаходить такі застосування (рисунок 1.2):

- класифікація тексту – категоризація текстових даних за визначеними класами, наприклад виявлення спаму або маркування тем у статтях новин;

- пошук інформації – отримання відповідної інформації з великих колекцій документів на основі запитів користувачів, що забезпечує ефективні функції пошуку;

- розпізнавання мовлення – навчання комп’ютерів розумінню усної мови, як те, що працює з голосовими помічниками та автоматизованими телефонними системами;


Рисунок 1.2 – Застосування розпізнавання мовлення у комп'ютерній лінгвістиц

- машинний переклад – використання машинного навчання для перекладу тексту з однієї мови на іншу з глибоким навчанням, що забезпечує

точніші переклади;

- резюме тексту – автоматичне генерування підсумків текстових документів, корисних для стиснення статей новин або описів продуктів;

- аналіз настрою – аналіз настрою, що стоїть за текстом, наприклад позитивного, негативного чи нейтрального, використовується для моніторингу соціальних мереж або обслуговування клієнтів (рисунок 1.3);

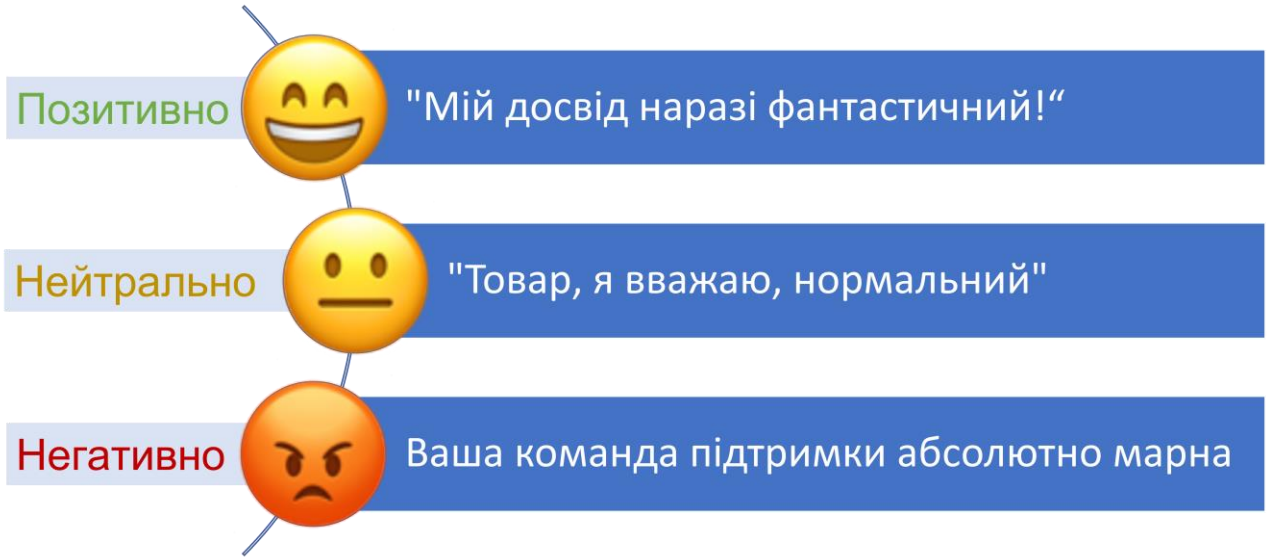


Рисунок 1.3 – Приклади позитивних, нейтральних і негативних настроїв у тексті

- позначення частин мови – присвоєння граматичної мітки (іменник, дієслово тощо) кожному слову в реченні, базовий крок для багатьох завдань НЛП (рисунок 1.4);

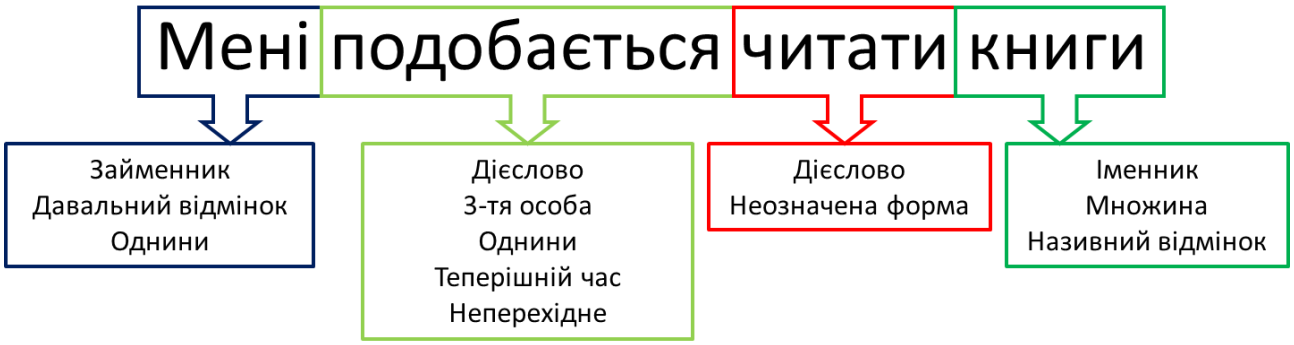


Рисунок 1.4 – Приклад Part-of-Speech Tagging для речення

- розпізнавання іменованих об’єктів (NER) – Ідентифікація та

класифікація конкретних деталей у тексті, таких як люди, місця чи організації, що корисно для отримання інформації та відповідей на запитання (рисунок 1.5);

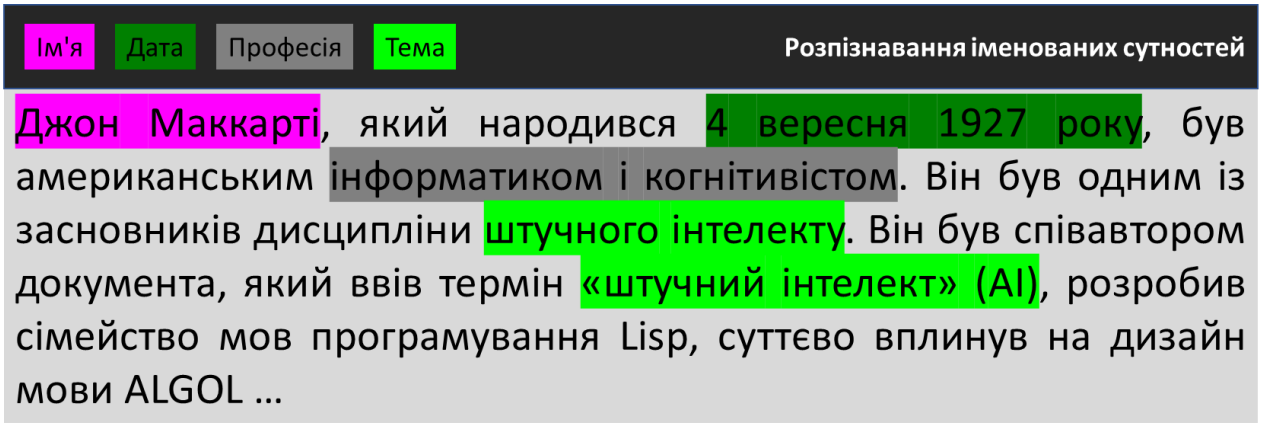


Рисунок 1.5 – Приклад розпізнавання іменованих об’єктів

- генерація природної мови – створення тексту людської якості за допомогою комп’ютерної програми, що використовується для створення різних творчих форматів тексту, такі як вірші, код, сценарії, музичні твори, електронні листи, листи тощо.

Охоплення машинного навчання в комп’ютерній лінгвістиці виходить за рамки основних завдань. Ось кілька додаткових застосувань, які використовують можливості машинного навчання:

- перевірка граматики та виправлення орфографії – аналіз великих обсягів тексту для покращення граматики та орфографії, крім простих виправлень, пропонуючи кращу структуру речень і формулювання;

- класифікація тексту – класифікація текстових документів у попередньо визначені групи (наприклад, теми новин або типи електронних листів) за допомогою алгоритмів машинного навчання;

- чат-боти та віртуальні помічники – навчання чат-ботів і віртуальних помічників розуміти природну мову та відповідати в розмові, що досягається завдяки величезній кількості даних розмов;

- перефразування та переписування тексту – використання машинного

навчання для перефразування або переписування тексту, зберігаючи значення, використовується для покращення машинного перекладу або створення варіацій стилю написання;

- виявлення образливої лексики – виявлення образливої або образливої лексики в тексті, яка використовується на онлайн-платформах для модерування вмісту та створення безпечніших онлайн-просторів;
- ідентифікація автора – аналіз шаблонів стилю письма для визначення автора фрагмента тексту, корисний для таких завдань, як перевірка авторства та виявлення плагіату.

Однак із появою глибокого навчання виникла нова хвиля інновацій. Архітектури глибокого навчання, особливо рекурентні нейронні мережі (RNN) та згорткові нейронні мережі (CNN), чудово справляються із захопленням складних взаємозв'язків у послідовних даних, таких як текст. Ця можливість призвела до значних досягнень у таких галузях аналізу настроїв та узагальнення тексту.

Моделі глибокого навчання можуть не лише визначати полярність настроїв, а й отримувати нюанси емоційних відтінків у тексті, забезпечуючи повніше розуміння думок та стосунків користувачів.

Щодо узагальнення тексту. Алгоритми глибокого навчання можуть не лише узагальнювати фактичну інформацію, але також вловлювати основний контекст і зміст тексту, генеруючи резюме, які не лише короткі, а й семантично вірні вихідному змісту.

Фокус комп'ютерної лінгвістики змістився з суто теоретичних досліджень розробку практичних додатків, які використовують можливості NLP. Такий підхід, орієнтований на додатки, призвів до створення нових видів чат-ботів, аналізаторів аспектів тексту та систем перекладу.

Інтелектуальні чат-боти і віртуальні помічники можуть спілкуватися з користувачами природною мовою, забезпечувати обслуговування клієнтів, відповідати на запитання, що часто ставляться, і навіть призначати зустрічі.

Компанії можуть використовувати інструменти аналізу настроїв для

моніторингу соціальних мереж для оцінки громадської думки про свій бренд, відстеження активності конкурентів та виявлення потенційних криз у режимі реального часу.

Досягнення в галузі NLP призвели до значного підвищення точності машинного перекладу, що забезпечує безперешкодне спілкування та співробітництво через міжнародні кордони.

### 1.1.2 Зростаюча потреба в автоматизованих методах аналізу текстів з емоційним забарвленням

Аналіз емоційних текстів є важливим напрямом у галузі комп'ютерної лінгвістики та інформаційних технологій. Сучасне інформаційне суспільство стикається з зростаючою потребою в ефективних методах аналізу та класифікації текстів щодо їх емоційного забарвлення. У цьому розділі розглянемо значущість та застосування аналізу емоційних текстів у різних сферах, а також роль автоматизованих методів у вирішенні цих завдань.

Існує безліч спроб класифікувати емоції. Одна впливова модель належить психологу Полу Екману, який виділив шість основних емоцій з різними фізіологічними та мімічними виразами: щастя, смуток, гнів, страх, здивування та огиду [1]. Ця модель забезпечує основу для автоматизованих систем, які покладаються на виявлення цих основних емоційних виразів за допомогою лінгвістичних сигналів та лексиконів настроїв (словників, що містять емоційно заряджені слова).

Інша концепція, комплексна модель емоцій Роберта Плутчика [2], виходить за рамки базових емоцій, пропонуючи вісім основних емоцій (радість, смуток, гнів, страх, довіра, огида, очікування та подив), розташованих у круговій структурі ( рисунок 1.6) . Ця модель передбачає, що емоції можна комбінувати, створюючи складні емоційні стани. Наприклад, поєднання гніву та страху може призвести до розчарування, а поєднання радості та передчуття може означати хвилювання. Автоматизовані системи,

що використовують цю структуру, можуть зосередитись на виявленні комбінацій емоційно заряджених слів, щоб уловити ці нюанси емоційних станів.

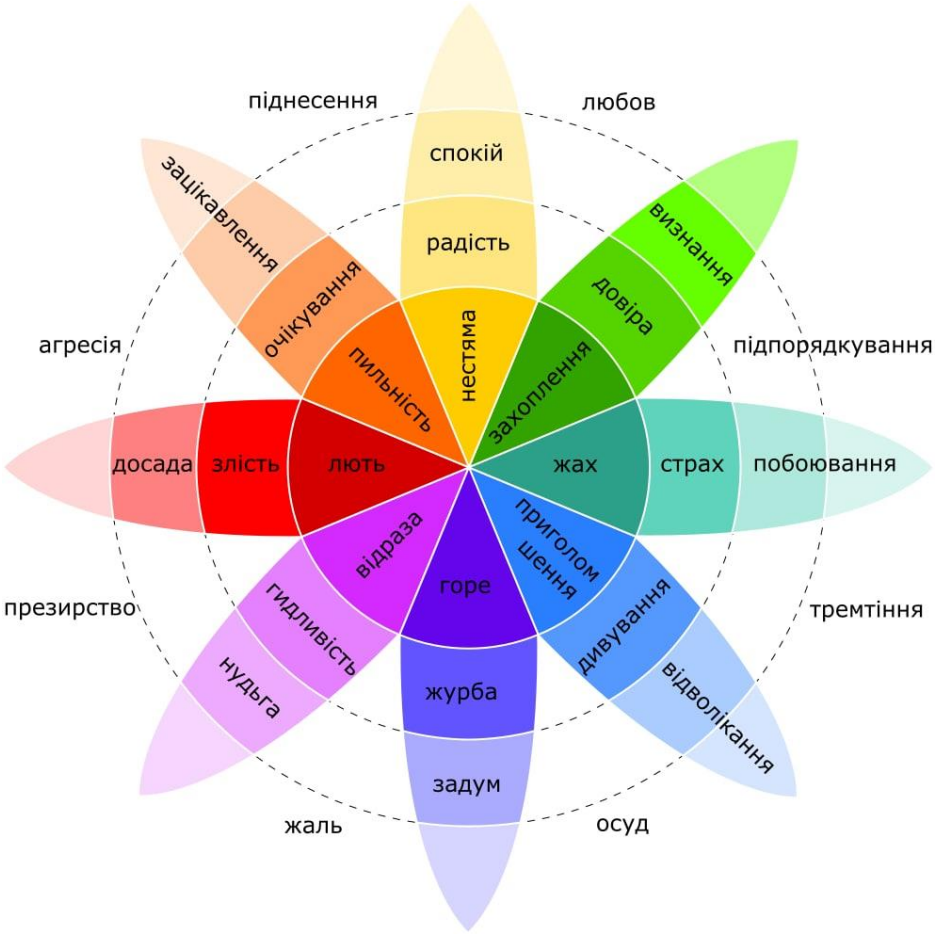


Рисунок 1.6 – Циркумплексна модель емоцій Плутчика

Крім цих моделей, дедалі більше досліджень вивчають культурні відмінності у вираженні емоцій [3]. Певні емоції, такі як «амае» в японській культурі (почуття залежності та поблажливості), можуть не мати прямих еквівалентів у західних мовах. Автоматизовані системи, орієнтовані більш широке застосування, повинні враховувати ці культурні аспекти.

Автоматизована класифікація емоційного тексту спирається на комбінацію критеріїв та правил:

- лінгвістичні підказки – певні слова і фрази часто пов'язані з певними емоціями. Наприклад, такі слова, як «щасливий», «схвильований» чи «радісний», зазвичай передають позитивні емоції, а такі слова, як «сумний»,

«депресивний» чи «самотній», припускають негативні емоції. Автоматизовані системи можуть використовувати словники настроїв (заздалегідь складені списки емоційно заряджених слів) для виявлення таких сигналів;

- синтаксичні структури. Те, як збудовано речення, також може передавати емоційний тон. Знаки оклику часто вказують на хвилювання або подив, а питання можуть сигналізувати про цікавість або замішання. Автоматизовані системи можна запрограмувати на розпізнавання цих синтаксичних шаблонів та включення їх у процес класифікації;

- теги частин мови – визначення граматичної функції слів (іменників, дієслів, прикметників) може сприяти емоційній класифікації. Наприклад, присутність негативних прислівників (наприклад, «ніколи», «ледь») може посилити негативність у реченні;

- N-грами – аналіз послідовностей слів (біграм, триграм) може бути особливо корисним для виявлення емоцій, що передаються через ідіоматичні вирази або сленг. Наприклад, фраза «почуття пригніченості» виражає смуток ефективніше, ніж аналіз окремих слів окремо;

- специфічність рівня дискурсу – аналіз потоку та структури більшої текстової одиниці (абзацу, документа) іноді може виявити емоційні тенденції. Наприклад, зміна словникового запасу або структури речень може сигналізувати про зміну емоційного стану тексту.

Комбінуючи ці критерії із системами, заснованими на правилах, та методами машинного навчання, комп'ютерна лінгвістика прагне автоматизувати процес виявлення емоційно зарядженої лексики та класифікації текстів відповідно до їхнього загального емоційного підтексту.

Важливо визнати обмеження автоматизованого емоційного аналізу. Іронія, сарказм та контекстно-залежні значення, як і раніше, можуть створювати проблеми. Однак розробка складних обчислювальних моделей, що продовжується, відкриває значні перспективи для поліпшення нашої здатності аналізувати і розуміти емоційне підґрунтя величезних обсягів

текстових даних.

Одним із ключових застосувань аналізу емоційних текстів є контент-модерація в онлайн-середовищі. Зі збільшенням кількості користувачів інтернету та активності в соціальних мережах, серед яких присутні як фізичні особи, так і організації, існує зростаюча потреба у забезпеченні безпеки та захисту від небажаного контенту. Автоматизовані методи аналізу текстів, здатні швидко та точно визначати емоційне забарвлення повідомлень, дозволяють ефективно фільтрувати контент та виявляти підозрілі чи небажані вирази, такі як ненависні висловлювання, погрози, насильство та дискримінацію.

Іншим важливим застосуванням аналізу емоційних текстів є моніторинг громадської думки та аналіз репутації бренду. Компанії та організації активно використовують соціальні мережі та інші онлайн-платформи для взаємодії з клієнтами та партнерами, а також для просування своїх продуктів та послуг. Автоматизовані методи аналізу текстів дозволяють швидко та ефективно відстежувати відгуки, коментарі та обговорення, пов'язані з їх діяльністю, та оцінювати емоційний тон та ставлення до бренду чи продукту. Це дозволяє компаніям оперативно реагувати на негативні зворотні зв'язки, покращувати якість обслуговування та вживати своєчасних заходів щодо покращення репутації.

Аналіз емоційних текстів також широко застосовується у психологічних дослідженнях та аналізі соціальних трендів. Емоції відіграють важливу роль у поведінці та реакціях людей, і аналіз текстів може допомогти дослідникам зрозуміти, які емоції переважають у певних ситуаціях, як вони впливають на прийняття рішень та як змінюються з часом. Це дозволяє більш глибоко зрозуміти психологічні механізми людської поведінки, виявляти тенденції та тренди у суспільстві, а також прогнозувати можливі зміни у суспільному настрої.

І, нарешті, аналіз емоційних текстів відіграє важливу роль у персоналізації контенту та поліпшенні користувальницького досвіду.

Розуміння емоційної реакції користувачів на контент дозволяє адаптувати його під їх переваги та потреби, що сприяє збільшенню залученості та задоволеності користувачів. Наприклад, автоматизовані системи рекомендації контенту можуть аналізувати емоційні реакції користувачів на запропонований контент та пропонувати їм більш відповідні та релевантні матеріали в майбутньому.

Активний розвиток інформаційних технологій та постійне зростання обсягу текстової інформації, що створюється та розповсюджується в онлайн-середовищі, призводить до зростаючої потреби в ефективних методах аналізу текстів з емоційним забарвленням. Ця потреба обумовлена кількома чинниками, які важливо розглянути у тих сучасного інформаційного суспільства.

Одним із основних факторів, що стимулюють зростання потреби в автоматизації методів аналізу текстів, є величезний обсяг текстової інформації, доступної в онлайн-середовищі. Мільйони повідомлень, коментарів, відгуків та інших текстових даних публікуються щодня у соціальних мережах, блогах, форумах та інших онлайн-платформах. Ручна обробка та аналіз такого обсягу даних стає практично неможливим завданням. У цьому контексті автоматизовані методи аналізу тексту, здатні обробляти великі обсяги даних з високою швидкістю, стають життєво важливими для ефективної роботи з текстовою інформацією.

Зі зростанням активності користувачів в онлайн-середовищі, особливо в соціальних мережах, зростає необхідність швидкої реакції на події та прийняття рішень. Компанії, державні організації, політики, а також звичайні користувачі все частіше стикаються з ситуаціями, що вимагають оперативного аналізу та реагування на текстові дані з емоційним забарвленням. Наприклад, у разі кризових ситуацій, публічних скандалів чи масових протестів важливо оперативно оцінювати громадську думку та настрій, а також вживати відповідних заходів. Автоматизовані методи аналізу текстів дозволяють проводити такий аналіз у реальному часі та

надавати оперативний зворотний зв'язок.

З розвитком технологій та доступом до великих даних зростає інтерес до персоналізації та індивідуалізації запропонованих послуг. Компанії та організації прагнуть пропонувати користувачам більш релевантний та персоналізований контент, що відповідає їх перевагам та потребам. Аналіз емоційного забарвлення текстів дозволяє зрозуміти переваги та настрої користувачів, а також передбачити їхню поведінку. Наприклад, шляхом аналізу емоційних реакцій на контент можна визначити, які типи матеріалів викликають найбільший інтерес або збуджують найбільший емоційний відгук у аудиторії, що дозволяє оптимізувати контент-стратегію та підвищити залучення користувачів.

Зі зростанням кількості користувачів інтернету та активності в соціальних мережах зростає також і необхідність у забезпеченні безпеки та захисту від небажаного контенту. Негативні коментарі, образи, погрози та інші форми небажаного контенту можуть створювати негативний досвід для користувачів та призводити до різних негативних наслідків. Автоматизовані методи аналізу текстів дозволяють ефективно фільтрувати та виявляти подібний контент, що сприяє підтримці безпечної та позитивної атмосфери в онлайн-середовищі.

У сфері наукових досліджень також зростає потреба у ефективних методах аналізу текстів із емоційним забарвленням. Дослідники з різних галузей, таких як психологія, соціологія, маркетинг та політологія, все частіше звертаються до текстових даних як джерела інформації про поведінку та відносини людей. Автоматизовані методи аналізу текстів дозволяють обробляти великі обсяги даних та проводити аналіз емоційного забарвлення текстів з високою точністю та швидкістю, що відкриває нові можливості для наукових досліджень та пошуку нових знань.

В цілому, зростаюча потреба в автоматизації методів аналізу текстів з емоційним забарвленням обумовлена не лише обсягом та швидкістю обробки даних, а й необхідністю швидкої реакції та прийняття рішень, зростанням

інтересу до персоналізації та індивідуалізації послуг, а також необхідністю у забезпеченні безпеки та захисту від небажаного контенту. Ці фактори роблять автоматизовані методи аналізу текстів з емоційним забарвленням важливим інструментом у сучасному інформаційному суспільстві, здатним забезпечити ефективну обробку та інтерпретацію текстової інформації у різних сферах діяльності.

## 1.2 Огляд проблемної галузі

У сучасному інформаційному суспільстві, де текстова інформація відіграє ключову роль у багатьох аспектах нашого життя, включаючи комунікацію, бізнес, політику, наукові дослідження та багато іншого, аналіз текстів з емоційним забарвленням стає все більш важливим та актуальним. Однак, незважаючи на значні переваги, які можуть бути отримані за допомогою такого аналізу, існує низка проблем та викликів, з якими стикаються дослідники та практики у цій галузі. У цьому розділі ми розглянемо основні проблеми, пов'язані з аналізом текстів з емоційним забарвленням, та можливі шляхи їх вирішення.

Однією з основних проблем під час аналізу текстів з емоційним забарвленням є складність інтерпретації, власне, емоцій, що у тексті. Людська мова володіє високим ступенем семантичної гнучкості та глибини, та різноманіттям виразів, що робить інтерпретацію емоцій часто неоднозначною та вкрай суб'єктивною. Наприклад, те саме слово чи вираз може викликати різні емоційні реакції в різних людей залежно від своїх особистих характеристик, культурного контексту та попереднього досвіду. Це ускладнює завдання розробки алгоритмів та моделей, здатних автоматично визначати та класифікувати емоційне забарвлення текстів з високою точністю та надійністю.

Ще однією проблемою є недостатня точність класифікації емоцій у тексті. Навіть при використанні сучасних методів машинного навчання та

нейронних мереж, точність класифікації емоцій може залишатися недостатньою для практичного застосування. Це з тим, що емоції часто є складними і багатогранними явищами, які завжди легко висловити з допомогою обмеженого числа категорій чи міток. Крім того, багато емоцій можуть мати змішаний характер, тобто бути комбінацією кількох базових емоцій, що також ускладнює їхню класифікацію.

В умовах багатомовного середовища, де текстова інформація може бути представлена різними мовами, виникають додаткові проблеми з аналізом емоційного забарвлення текстів. Різні мови мають різні синтаксичні та лексичні особливості, а також культурні контексти, що робить складним узагальнення та застосування методів та моделей, розроблених для однієї мови, іншими мовами. Це створює необхідність у розробці багатомовних та культурно-адаптивних методів аналізу емоційного забарвлення текстів.

Ще однією проблемою, з якою стикаються дослідники в галузі аналізу текстів з емоційним забарвленням, є відсутність стандартизованих наборів даних для навчання та оцінки моделей. Незважаючи на те, що існують деякі відкриті корпуси текстів з емоційною розміткою, їх кількість та різноманітність обмежені, що ускладнює розробку та порівняння різних методів та моделей. Крім того, існуючі набори даних часто не враховують різноманітність культурних та мовних контекстів, що обмежує застосовність розроблених моделей до різних груп користувачів та ситуацій.

Ще одним важливим аспектом при аналізі текстів з емоційним забарвленням є необхідність урахування контексту та контекстуальної інформації. Емоційне забарвлення тексту може залежати не тільки від самих слів і виразів, а й від їхнього контексту, тобто від ситуації, в якій вони використовуються, і від характеристик та відносин між сторонами, що комунікують. Наприклад, те саме вираз може сприйматися як жарт або як образу залежно від контексту спілкування. Врахування контекстуальної інформації ускладнює завдання аналізу емоційного забарвлення тексту і вимагає розробки методів і моделей, здатних враховувати цей аспект.

Нарешті, однією з основних проблем в аналізі текстів з емоційним забарвленням є проблеми з балансом та упередженістю даних. Для побудови надійних і точних моделей класифікації емоцій необхідно мати доступ до широкого і репрезентативного набору даних, який включає різноманітні типи текстів і емоційні вирази. Однак, часто дані можуть бути упередженими та нерепрезентативними, що може призвести до спотворення результатів аналізу та зниження точності та надійності моделей.

### 1.2.1 Характеристика емоційно-забарвленої лексики

Емоційно забарвлена лексика є особливою категорією слів і виразів, які несуть у собі емоційний заряд і можуть викликати в читача чи слухача певні емоційні реакції. Ця частина мови відіграє важливу роль у комунікації, дозволяючи передавати не тільки інформацію, а й емоційний стан того, хто говорить або автора тексту. У цьому розділі ми розглянемо основні характеристики емоційно забарвленої лексики, її особливості та роль мовному спілкуванні (рисунок 1.7).

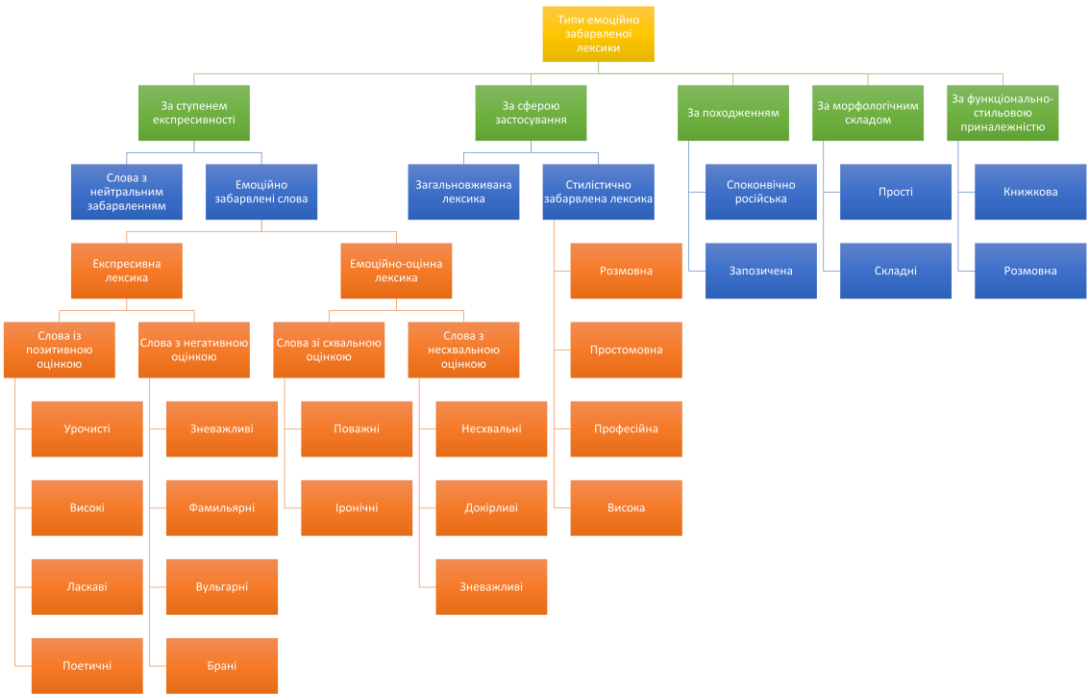


Рисунок 1.7 – Типи емоційно забарвленої лексики

Однією з основних характеристик емоційно забарвленої лексики є семантична багатозначність. Слова і вирази, що стосуються цієї категорії, часто мають кілька значень або відтінків, кожне з яких пов'язане з певними емоційними станами або реакціями. Наприклад, навіть, слово "кохання" може викликати позитивні емоції в однієї людини та негативні в іншої, залежно від її особистого досвіду та контексту використання.

Емоційно забарвлена лексика зазвичай характеризується високою експресивністю та емоційною інтенсивністю. Це слова та висловлювання, які можуть викликати сильні та яскраві емоційні реакції у читача чи слухача. Наприклад, слова з негативним забарвленням, такі як "гнів", "огидність" або "страх", зазвичай асоціюються з сильними негативними емоціями, в той час як слова з позитивним забарвленням, такі як "радість", "захват" або "любов" можуть викликати позитивні емоції.

Емоційно забарвлена лексика може бути як конкретною, так і абстрактною. Конкретні емоційні слова зазвичай відносяться до певних емоційних станів або реакцій, наприклад, "сум", "радість" або "злість". Абстрактні слова можуть позначати більш загальні концепції чи ідеї, пов'язані з емоціями, наприклад, "щастя" , "туга" чи "хвилювання" . Обидва типи слів відіграють важливу роль у мовній комунікації, дозволяючи передавати різноманітні емоційні відтінки та нюанси.

Емоційно забарвлена лексика тісно пов'язана з культурними та соціальними аспектами мови. Культурні особливості, традиції, звичаї та цінності можуть проводити те, які слова і висловлювання вважаються емоційно значимими у цій культурі. Наприклад, в одній культурі певні слова можуть мати позитивне забарвлення, тоді як в іншій - негативне. Також соціальний контекст використання слова може впливати з його емоційне значення. Наприклад, слово "патріотизм" може викликати позитивні емоції в одних людей і негативні в інших, залежно від політичних переконань і соціальної приналежності.

Емоційно забарвлена лексика відіграє важливу роль у мовній

комунікації, дозволяючи мовця висловити свої емоційні стани, почуття та стосунки. Вона допомагає встановити емоційний контакт із аудиторією, викликати інтерес і залучення слухачів чи читачів, і навіть посилити емоційний вплив тексту. Емоційно забарвлена лексика часто використовується в літературі, поезії, рекламі, публічних виступах та інших сферах мовної діяльності для створення емоційної атмосфери та досягнення певних комунікативних цілей.

### 1.2.2 Труднощі, пов'язані з точністю класифікації емоційних виразів у тексті

Зростаюча область автоматизованого емоційного аналізу тексту відкриває великі перспективи для вилучення цінної інформації з океану цифрових комунікацій, що постійно розширюється. Однак досягнення точної та надійної класифікації настроїв є складним завданням. У цьому розділі розглядаються властиві людській мові складності та вираження емоцій, які збивають з пантелику навіть найскладніші алгоритми.

Складність людських почуттів. Людські емоції є багатогранною конструкцією, що включає динамічну взаємодію фізіологічних реакцій, когнітивних оцінок та поведінкових проявів (таблиці 1.2 та 1.3). Більше того, емоційні стани можуть бути скороминущими та багатогранними, швидко змінюючись у межах одного фрагмента тексту. Зображення цього складного гобелену за допомогою статичної писемної мови є серйозною перешкодою. Автоматизовані системи щосили намагаються відтворити тонке розуміння, яким володіють люди при інтерпретації емоційних сигналів за виразом обличчя, мовою тіла та тону голосу, які відсутні в текстових даних. В результаті автоматизовані системи повинні покладатися виключно на явні лінгвістичні особливості тексту, що потенційно може призвести до неправильної інтерпретації емоційного основного стану.

Таблиця 1.2 – Складнощі, пов'язані з природою тексту

| Складність | Опис | Приклад |
|---|---|---|
| Лексичний вибір | Контекст сильно впливає сенс. Слова можуть бути позитивними чи негативними залежно від контексту. | «Шоу було катастрофою!» (позитивно-саркастичний) або (щиро негативний) |
| Синтаксис та структура | Конструкція пропозиції впливає на емоційну вагу. Знаки оклику вказують на терміновість, а пасивна застава створює відстороненість. | Ми виграли! (позитивний – сильний) проти «Гра виграна». (позитивний-нейтральний) |
| Прагматика та невербальні сигнали | Невербальні сигнали, такі як вираз обличчя та тон, передають сарказм, іронію чи гумор, які важко знайти у тексті. | Текст сам собою може розкрити справжнього наміри саркастичного висловлювання. |
| Двозначність | Слова часто мають кілька значень, залежно від контексту. | «Зустріч пройшла галасливо» (позитивно чи негативно?) |
| Варіації у мові та настроях | Емоційний вираз варіюється в залежності від предметної галузі та спільноти. | У медичних звітах для вираження емоцій використовується інша мова, ніж у повідомленнях у соціальних мережах. |

Багатогранна природа тексту: мова сама собою є лабіринт проблем для автоматичного аналізу настроїв. На зміст, який ми витягуємо з тексту, впливає сукупність чинників, зокрема:

- лексічний вибір. Хоча певні слова несуть у собі властивий емоційний відтінок (наприклад, «радісний» чи «убитий горем»), загальний настрій

речення може сильно впливати контекст, у якому ці слова використовуються. Наприклад, пропозиція «Виступ було катастрофою» може бути витлумачена як щиро негативна чи саркастично позитивна залежно від навколишнього контексту. Автоматизованим системам потрібні складні алгоритми, щоб вийти за рамки простого аналізу на рівні слів і заглибитися в контекстуальні нюанси, які формують емоційне підґрунтя речення;

- синтаксис та структура. Те, як побудовано речення, може суттєво вплинути на його емоційну вагу. Знаки оклику часто передають терміновість або хвилювання, а пасивна застава може створити відчуття відстороненості або нейтральності. Автоматизовані системи повинні бути оснащені не лише для аналізу явного значення слів, але й для обліку тонких емоційних сигналів, вбудованих у структуру речення;

- прагматика та невербальні сигнали. Значна частина людського спілкування спирається на невербальні сигнали, такі як вираз обличчя, мова тіла та тон голосу. Ці сигнали грають вирішальну роль передачі сарказму, іронії чи гумору, які, як відомо, важко знайти лише у письмовому тексті. Відсутність цих невербальних елементів вимагає, щоб автоматизовані системи покладалися виключно на явні лінгвістичні особливості тексту, що призводить до потенційних неправильних інтерпретацій.

Людська мова за своєю суттю неоднозначна. Слова часто мають кілька значень, залежно від контексту. Розглянемо пропозицію: "Зустріч пройшла галасливо". Без додаткової контекстної інформації автоматизована система зможе визначити, чи належить «галасливо» до позитивного (дуже приємного) чи негативного (неприємно) досвіду. Сарказм, наріжний камінь людського спілкування, є аналогічною проблемою. Автоматизованим системам часто складно ідентифікувати тонкі сигнали, які відрізняють сарказм від щирого почуття, потенційно неправильно тлумачачи гаданий зміст тексту.

Варіації мови та настроїв залежно від предметної області. Емоційний вираз може значно відрізнятися у різних галузях та спільнотах. Наприклад, мова, яка використовується в медичному звіті, швидше за все, передаватиме

емоції інакше, ніж у огляді продукту чи публікації у соціальних мережах. Автоматизованим системам потрібні навчальні дані, специфічні для аналізованої області, щоб ефективно класифікувати вираження емоцій. Система, навчена на статтях фінансових новин, може важко точно інтерпретувати емоційну мову, що використовується на ігровому форумі, через властиві цим доменам відмінності в словниковому запасі і моделях вираження емоцій.

Таблиця 1.3 – Складності мовної динаміки

| Складність | Опис | Приклад |
|---|---|---|
| Розвивається мова | Постійно з'являються новий сленг, інтернет-меми та розмовні вислови. | Автоматизовані системи можуть мати справу з новими тенденціями онлайн-комунікацій. |
| Культурні та соціальні забобони | Мова та емоції переплітаються з культурним контекстом. | Дані навчання з культурними упередженнями можуть неправильно інтерпретувати текст із різних верств суспільства. |
| Заперечення | Виявлення та інтерпретація заперечення є складним завданням. | "Фільм був непоганим" (злегка позитивний) vs. "Фільм був поганим" (негативний) |
| Смайли та символи настрою | Смайли типу ":D" або ":(" можуть бути неправильно витлумачені, особливо як сарказм або іронія. | Значення деяких смайлів може змінюватись в залежності від контексту. |

Природа мови розвивається. Мова - це динамічна сутність, що постійно розвивається з появою нових сленгових термінів, інтернет-мемов та

розмовних виразів. Ці лінгвістичні інновації можуть стати проблемою для автоматизованих систем, які можуть бути навчені новітнім тенденціям онлайн-комунікації. Щоб йти в ногу з природою мови, що постійно розвивається, потрібна постійна адаптація і вдосконалення цих систем. Дані навчання повинні регулярно оновлюватись, щоб включати новий словниковий запас та нові мовні моделі, щоб забезпечити точність класифікації настроїв.

Культурні та соціальні упередження. Мова та емоції глибоко переплетені з культурним та соціальним контекстом. Те, що вважається ввічливим вираженням в одній культурі, може бути сприйнято як образливе в іншій. Дані навчання, що використовуються для аналізу настроїв, можуть ненавмисно кодувати культурні упередження, що потенційно призводить до неправильного тлумачення при аналізі тексту, що стосується різних культурних традицій. Наприклад, автоматизована система, навчена в основному на даних про американську англійську, може неправильно витлумачити тональність тексту, написаного англійською англійською, через тонкі культурні відмінності у використанні мови та гумору. Пом'якшення культурних упереджень вимагає використання різноманітних наборів навчальних даних та розробки алгоритмів, які можуть враховувати культурні нюанси в емоційному вираженні.

Ідентифікація та інтерпретація заперечення у тексті може бути складним завданням для автоматизованих систем. Розміщення слів заперечення (наприклад, «ні», «ніколи», «ні») може суттєво змінити емоційне значення речення. Наприклад, твердження "Фільм був непоганим" передає нейтральний або навіть трохи позитивний настрій, а твердження "Фільм був поганим" явно негативне. Автоматизовані системи мають бути здатними точно визначати обсяг заперечення. тобто, застосовується воно до одного слова або до всієї пропозиції, і відповідним чином коригувати класифікацію настроїв.

Загадка смайликів та символів настрою. Смайли та символи настрою,

такі як «:D» для щастя або «:(» для смутку, стали повсюдною особливістю онлайн-спілкування. Хоча ці символи дають явні емоційні сигнали, їхня точна інтерпретація може бути складним завданням. Автоматизованим системам може бути складно відрізнити смайли, що використовуються з сарказмом або іронією, від тих, які використовуються для передачі щирих емоцій. Крім того, значення деяких смайлів може різнитися залежно від платформи чи контексту, в якому вони використовуються.

Вічна обмеженість навчальних даних. Точність автоматичного емоційного аналізу тексту багато в чому залежить від якості та кількості використовуваних навчальних даних. Дані навчання повинні охоплювати широкий спектр емоційних виразів, мовних стилів та мовних варіацій, специфічних для предметної галузі. Обмежені дані для навчання можуть призвести до перенавчання, коли система добре працює на даних, на яких вона навчалася, але важко узагальнити її на невидимі приклади. Наприклад, система, навчена в основному на позитивних відгуках клієнтів, може важко класифікувати настрій негативних відгуків через відсутність негативних емоційних виразів у навчальних даних.

Етична проблема зрозумілості та прозорості. У міру того, як автоматизований емоційний аналіз тексту стає все більш складним, виникають побоювання з приводу зрозумілості та прозорості. Користувачам дуже важливо зрозуміти причину класифікації настроїв системи. Моделі чорної скриньки, які є складними та непрозорими у своїх процесах прийняття рішень, можуть ускладнити визначення факторів, що впливають на результати роботи системи. Методи зрозумілого штучного інтелекту мають вирішальне значення для зміцнення довіри та забезпечення того, щоб ці системи не увічнювали упередженість і не робили неточних класифікацій.

## 1.3 Огляд існуючих систем контент-аналізу у комп'ютерній лінгвістиці

У сучасному світі, де обсяг текстових даних зростає в геометричній

прогресії, важливо мати ефективні методи та інструменти для аналізу та розуміння цієї інформації. У галузі комп'ютерної лінгвістики та аналізу тексту існує безліч підходів та методів, спрямованих на автоматизацію процесів обробки текстів та вилучення з них корисної інформації. Одним із ключових аспектів цієї роботи є аналіз змісту тексту, або контент-аналіз, що дозволяє виявляти різні структури, теми, емоційні забарвлення та інші характеристики текстових даних.

Контент-аналіз служить наріжним каменем для отримання значної інформації з текстових даних. У цьому розділі розглядається ландшафт існуючих систем контент-аналізу, класифікованих за базовими методологіями. Розуміючи сильні та слабкі сторони цих усталених підходів, ми прокладаємо шлях до вивчення потенційних покращень та обґрунтування розробки нової системи.

Традиційні методи контент-аналізу є першими спробами автоматизованого аналізу тексту. Ці методи часто ґрунтуються на зіставленні ключових слів та зумовлених словниках. Наприклад, для аналізу настроїв можна використовувати словник позитивних та негативних слів для класифікації загального настрою тексту. Хоча ці методи пропонують певну простоту та легкість реалізації, їх ефективність часто знижується через обмеженість словникового запасу та нездатність вловити нюанси людської мови. Їм складно впоратися із сарказмом, образною мовою та контекстно-залежною природою значення слова. Крім того, їхня здатність ідентифікувати названі об'єкти та тематичні області залишається рудиментарною. Проте традиційні методи, як і раніше, мають цінність у конкретних сценаріях через їх інтерпретованість та низькі обчислювальні вимоги.

Сфера машинного навчання відкрила нову епоху контент-аналізу. Моделі машинного навчання, такі як наївний Байєс (Naive Bayes) та машини опорних векторів (SVM), можна навчати на великих наборах даних із розміченим текстом для виявлення закономірностей та зв'язків усередині

даних. Ці моделі демонструють підвищену точність аналізу настроїв та класифікації емоцій у порівнянні з традиційними методами. Однак їх ефективність залежить від якості та розміру навчальних даних. Більше того, вони часто вимагають значних зусиль щодо розробки функцій і можуть бути менш інтерпретованими, ніж традиційні методи, що ускладнює розуміння обґрунтування їх класифікацій. Здатність цих моделей ідентифікувати названі об'єкти та тематичні області може бути обмежена, що часто вимагає додаткових модулів або методів для досягнення комплексного контент-аналізу.

Статистичні методи, такі як «Мішок слів» (BoW) та N-грами, пропонують інший підхід до контент-аналізу. BoW представляє текстові дані як набір слів, незалежно від порядку слів. N-грами розширюють цю концепцію, розглядаючи послідовність слів (наприклад, біграми, триграми). Ці методи чудово фіксують статистичний розподіл слів у тексті, що виявляється цінним для таких завдань, як моделювання теми та визначення широких тематичних галузей. Однак їм часто важко вловити семантичні відносини між словами, що заважає виконувати детальний аналіз настроїв або класифікацію емоцій. Крім того, розпізнавання іменованих об'єктів за допомогою цих методів може бути утруднено, оскільки їм не вистачає можливості розпізнавати внутрішню структуру та контекст усередині мови.

Методи глибокого навчання, зокрема рекурентні нейронні мережі (RNN) та їх варіанти, такі як мережі довгострокової короткострокової пам'яті (LSTM), зробили революцію у сфері контент-аналізу. Ці моделі чудово фіксують довгострокові залежності у текстових даних, що дозволяє їм аналізувати послідовний характер мови та розуміти контекст слів. Це призводить до чудової продуктивності в аналізі настроїв, класифікації емоцій і навіть розпізнаванні іменованих об'єктів. Більше того, досягнення в галузі глибокого навчання, такі як «Трансформери» з такими моделями, як BART, демонструють потенціал для більш складних завдань аналізу тексту, включаючи тематичне вилучення, що враховує відносини та потік ідей у

тексті. Однак моделі глибокого навчання часто вимагають величезних обсягів навчальних даних та значних обчислювальних ресурсів, що може створювати проблеми для практичної реалізації.

Як уже зазначалося, кожна методологія контент-аналізу має певні переваги та недоліки (таблиця 1.4). Традиційні методи відрізняються інтерпретованістю та ефективністю, але їм не вистачає точності та нюансів. Моделі машинного навчання та статистичні методи забезпечують баланс між інтерпретованістю та точністю, але їх ефективність багато в чому залежить від навчальних даних та розробки функцій. Методи глибокого навчання пропонують високий потенціал на вирішення складних завдань аналізу, але потребують значних обчислювальних ресурсів, і даних.

Вибір методології залежить від конкретних вимог проекту. Для завдань, що вимагають високої інтерпретованості та аналізу у реальному часі, можуть бути достатніми традиційні методи чи методи машинного навчання. Для сценаріїв, у яких пріоритет віддається детального аналізу та високої точності, підходи глибокого навчання мають величезні перспективи. Зрештою, ідеальна система контент-аналізу забезпечує баланс між ефективністю, масштабованістю та бажаним рівнем аналітичної глибини.

Порівняльний аналіз показує цінний вклад існуючих систем контент-аналізу. Однак це також пролило світло на їх обмеження, наголосивши на необхідності більш комплексного підходу. Ці недоліки відкривають шлях до розробки нової системи, яка зможе подолати ці перешкоди та розкрити весь потенціал контент-аналізу.

Одне з ключових обмежень полягає в тому, що багато існуючих систем намагаються впоратися з тонкощами людської мови. Хоча деякі моделі машинного навчання досягають похвальної точності в базовому аналізі настроїв (позитивних чи негативних), вони часто стикаються зі складнощами людських емоцій. Сарказм, іронія і тонкі нюанси вираження почуттів можуть спантеличити ці моделі. Здатність ідентифікувати ширший спектр емоцій, таких як радість, гнів, смуток чи зневіра, залишається постійною проблемою.

Ця обмежена емоційна деталізація перешкоджає можливості існуючих систем повністю відбивати багатство і глибину людського самовираження.

Таблиця 1.4 – Порівняння методологій контент-аналізу

| Методологія | Сильні сторони | Обмеження |
|---|---|---|
| Традиційні методи | Простота реалізації, інтерпретованість | Обмежена точність, проблеми з нюансами |
| Моделі машинного навчання | Підвищена точність, усунуто деяку складність. | Залежить від навчальних даних, менш інтерпретовано |
| Статистичні методи | Ефективно фіксує розподіл слів | Ігнорує порядок слів, обмежений аналіз настроїв. |
| Методи глибокого навчання | Висока точність, фіксує складні взаємозв'язки | Потрібні великі дані та значні обчислювальні ресурси. |

Ще один недолік багатьох систем контент-аналізу полягає в тому, що їм важко вловити тематичні зв'язки і основні ідеї в тексті. Традиційні та статистичні методи часто покладаються на ідентифікацію ключових слів або підходи BoW, які можуть важко вловити взаємозв'язок концепцій та потік ідей у тексті. Хоча деякі моделі машинного навчання можуть виконувати елементарне тематичне моделювання, вони можуть забувати більш тонкі тематичні нитки, що пронизують текст. Таке фрагментарне розуміння тим перешкоджає здібності існуючих систем забезпечити всебічний аналіз змісту та його основного послання.

Обсяг текстових даних, що постійно зростає, є серйозною проблемою для багатьох існуючих систем контент-аналізу. Моделі машинного навчання, особливо методи глибокого навчання, часто вимагають величезних обсягів навчальних даних задля досягнення оптимальної продуктивності. Ця залежність від даних може створити вузькі місця масштабування, обмежуючи

здатність цих систем ефективно обробляти великі набори даних. Більш того, обчислювальні ресурси, необхідні для навчання та запуску складних моделей глибокого навчання, можуть бути значними, що ускладнює їхнє практичне застосування в середовищах з обмеженими ресурсами.

Обговорювані вище обмеження наголошують на необхідності в системі контент-аналізу, яка могла б подолати ці недоліки. Ідеальна система повинна мати здатність:

- вийти за рамки базового аналізу настроїв та заглибитеся у тонкий спектр людських емоцій;
- ефективно фіксувати тематичні зв'язки та основні ідеї у текстових даних;
- демонструвати масштабованість та ефективність обробки великих наборів даних.

Усунувши ці обмеження, нова система контент-аналізу може відкрити нові можливості для отримання значної інформації з текстових даних, відкриваючи шлях до більш повного розуміння людського спілкування та самовираження. Саме в цьому полягає мета запропонованої моделі, метою якої є об'єднання аналізу настроїв, класифікації емоцій та розпізнавання іменованих об'єктів із вилученням тем для створення надійного та детального підходу до аналізу контенту.

Обмеження існуючих систем контент-аналізу, описані вище, потребують нового підходу. Розглянемо потенційні переваги запропонованої моделі, наголошується, як вона усуває виявлені недоліки та пропонує унікальні функції для всебічного аналізу контенту.

Пропонована модель покликана вийти за рамки базового аналізу настроїв та заглибитись у більш тонкий спектр людських емоцій. Цього можна досягти за допомогою кількох потенційних стратегій:

Включення методів розширення лексикону. Шляхом інтеграції лексиконів сентиментів, що охоплюють ширший спектр емоцій, модель може виявляти тонкі відмінності в емоційному вираженні. Наприклад, лексикони, спеціально розроблені для уловлювання сарказму або образної мови, можуть

покращити здатність моделі інтерпретувати справжні почуття, що стоять за висловом.

Використання механізмів уваги. Механізми уваги в архітектурах глибокого навчання можуть дозволити моделі зосередитись на конкретних словах чи фразах, які мають значну емоційну вагу в тексті. Цей цілеспрямований фокус може покращити здатність моделі розрізняти такі емоції, як розчарування та гнів, які можуть мати деякий лексичний збіг.

Пропонована модель спрямована на подолання тематичних сліпих зон існуючих систем шляхом використання рекурентних нейронних мереж (RNN) або перетворювачів. Ці архітектури чудово справляються з фіксацією довгострокових залежностей у текстових даних. Аналізуючи послідовний характер слів та фраз, модель може виявити взаємозв'язки між поняттями та зрозуміти загальний тематичний потік тексту.

Використання методів тематичного моделювання. Алгоритми тематичного моделювання можуть допомогти моделі виявити приховані тематичні структури тексту. Це дозволяє моделі вийти за рамки ідентифікації ключових слів та отримати основні ідеї та концепції, які намагається передати автор.

Пропонована модель спрямована на вирішення проблем масштабованості шляхом:

- використання трансферного навчання. Шляхом попереднього навчання моделі на великому наборі даних загального призначення її можна налаштувати для конкретних завдань, таких як аналіз настроїв або розпізнавання іменованих об'єктів. Такий підхід знижує залежність від великих обсягів розмічених даних кожного конкретного домену;
- вивчення методів навчання без вчителя. Вивчення методів навчання без вчителя може дозволити моделі вчитися на немаркованих даних, що ще більше підвищить її масштабованість та застосовність до реальних сценаріїв з доступними, але потенційно немаркованими наборами даних.

Пропонована модель, усуваючи обмеження існуючих систем, пропонує

комплексніший підхід до контент-аналізу. Це включає в себе:

- розширений аналіз настроїв та класифікація емоцій – здатність ідентифікувати ширший спектр емоцій з більшою деталізацією;

- поліпшене розпізнавання іменованих об'єктів – використання досягнень глибокого навчання для більш точної ідентифікації іменованих об'єктів у тексті;

- тематичне вилучення – вилучення як ключових слів, а й основних тематичних структур і зв'язків усередині тексту, що забезпечує глибше розуміння змісту.

Цей цілісний підхід потенційно може відкрити нові можливості для аналізу текстових даних, сприяючи глибшому розумінню людського спілкування та різноманітного емоційного ландшафту, що його охоплює. У наступних розділах дисертації будуть детальніше розглянуті особливості запропонованої моделі, вивчено її архітектуру, стратегії навчання та методи оцінки.

## 1.4 Мета та задачі дослідження

Метою дослідження є розробка нових методів контент-аналізу текстової інформації з емоційно забарвленою лексикою, що має широкий спектр застосувань у різних сферах діяльності і дозволить здійснювати реактивні дії, такі як модерація або генерація контенту, спрямованого на певну аудиторію, ґрунтуючись на результатах аналізу вхідних текстових масивів.

Для досягнення зазначеної мети планується використання комплексної моделі, що складається з чотирьох модулів: емоційна класифікація, пошук іменованих сутностей, аналіз емоційного забарвлення та тематичний аналіз. За допомогою цих модулів буде здійснено узагальнення тексту, виділення емоційно забарвленої лексики та визначення тем, присутніх у тексті. Такий підхід дозволить отримати загальне уявлення про настрій тексту, емоції,

почуття та відчуття, які в ньому присутні, а також про теми, які в ньому обговорюються.

Для досягнення поставленої мети необхідно вирішити наступні завдання:

- розглянути методи аналізу текстового контенту різної довжини, включаючи класифікацію текстових масивів за типами емоцій та настроїв;
- проаналізувати нейромережеві лінгвістичні моделі для реалізації модулів семантичного, емоційного та тематичного аналізу, включаючи огляд існуючих датасетів;
- розробити концептуальну модель системи маркування емоційно-забарвленої лексики;
- провести експериментальні дослідження впливу тренувальних параметрів нейромережевої моделі (розмір батчу, розмір датасету, кількість епох, тип оптимізатора) на ефективність визначення емоційного забарвлення та настрою тексту;
- експериментально довести необхідність використання скейлеру для скорочення часу навчання нейромережевих класифікаторів;
- проаналізувати отримані результати.

Під маркуванням емоційно-забарвленої лексики у межах цього дослідження мається на увазі, здатність виділяти з текстової інформації слова, висловлювання і фрази, що містять емоційне забарвлення. Це дозволить системі автоматично визначати емоційний тон тексту, виявляти позитивні, негативні чи нейтральні емоційні стани, які можна висловити у тексті.

Крім того, модель проводитиме аналіз тексту з метою виділення загальної теми чи тематичних напрямків, присутніх у ньому. Це може бути реалізовано у вигляді списку тегів або ключових слів, які найповніше описують основні теми, які обговорюються в тексті.

Отримані дані щодо виділення емоційно забарвленої лексики та тематичним напрямам у тексті можуть бути використані в різних цілях.

Наприклад, вони можуть бути застосовані для модерації повідомлень у соціальних мережах або інших онлайн-платформах, де важливо контролювати зміст та забезпечувати безпечну та приємну взаємодію з користувачами. Ці дані можуть бути використані для спрощення розуміння настрою або змісту тексту, що особливо корисно при аналізі великих обсягів текстової інформації.

Більше того, з використанням даних про виділену емоційно забарвлену лексику та тематичні напрямки можливе створення різних медіа-контенту на основі тексту. Це може бути корисним для автоматизації процесу створення контенту для медіа-платформ, блогів, рекламних матеріалів та інших сфер діяльності, де потрібне створення контенту на основі текстової інформації.

Подальший розвиток дослідження передбачається у напрямі покращення ефективності та точності розроблених методів, а також у дослідженні їх застосування у різних сферах діяльності, таких як аналіз соціальних медіа, моніторинг громадської думки, маркетингові дослідження та багато інших.

# 2 АНАЛІЗ ТЕХНОЛОГІЧНОГО ТА МЕТОДОЛОГІЧНОГО ПІДҐРУНТЯ ДЛЯ ВИРІШЕННЯ ПОСТАВЛЕНОГО ЗАВДАННЯ

## 2.1 Огляд методів та моделей контент-аналізу

Контент-аналіз — це дослідницька техніка для створення висновків про тексти шляхом систематичного виявлення та аналізу шаблонів. Автоматизований аналіз контенту завдяки використанню обчислювальних інструментів стає все більш важливим для дослідників, які працюють із великими обсягами текстових даних. Кілька хмарних провайдерів пропонують функції обробки природної мови (NLP), пов’язані з завданнями аналізу вмісту.

Нижче наведено приклади існуючих систем контент-аналізу разом з їхніми особливостями. Порівняльний аналіз виконано у вигляді таблиці.

Таблиця 2.1 – Існуючі системи аналізу настрою

| Система | Особливості |
|---|---|
| Amazon Comprehend | надає можливості аналізу тексту, включаючи аналіз настрою, виявлення мови та іменовані сутності. |
| Google Cloud Natural Language | надає можливості аналізу настрою, виявлення мови, класифікації тексту та багато іншого. |
| IBM Watson Natural Language Understanding | надає функції аналізу настрою, виявлення мови, визначення ключових фраз та іменованих сутностей. |

Продовження таблиці 2.1

| Система | Особливості |
|---|---|
| Microsoft Azure Text Analytics | надає можливості аналізу тексту, включаючи аналіз настрою, виявлення мови, визначення ключових фраз та іменованих сутностей. |
| VADER (Valence Aware Dictionary and sEntiment Reasoner) | бібліотека, використовує набір правил та лексикон для визначення позитивного, негативного та нейтрального настрою тексту. |

Ці системи можуть використовуватися для аналізу настрою текстових даних з різних джерел, таких як соціальні мережі, відгуки клієнтів, новинні статті тощо. Кожна з них має свої переваги та обмеження, які можуть впливати на їхнє використання в конкретних сценаріях.

Amazon Comprehend — це хмарний сервіс обробки природної мови (NLP), який пропонує Amazon Web Services (AWS). Він використовує моделі машинного навчання для отримання інформації та взаємозв'язків з текстових даних.

Таблиця 2.2 – Огляд Amazon Comprehend

| Тип | Cloud Service (AWS) |
|---|---|
| Сильні сторони | Масштабованість, інтеграція з іншими сервісами AWS, розпізнавання об'єктів. |
| Недоліки | Може бути дорогим для великих завдань. Обмежене налаштування для аналізу настроїв. |
| Ключова особливість | Аналіз тональності, розпізнавання сутностей, синтаксичний аналіз, вилучення ключових фраз, моделювання тем |

Продовження таблиці 2.2

| | |
|---|---|
| Тип | Cloud Service (AWS) |
| Мови, що аідтримуються | Англійська, іспанська, французька, німецька і т. д. |
| Ціноутворення | Оплата за фактом використання |

Google Cloud Natural Language – це хмарна служба обробки природної мови (NLP), яку пропонує Google Cloud Platform (GCP). Подібно до Amazon Comprehend, він використовує моделі машинного навчання для аналізу та розуміння неструктурованих текстових даних.

Таблиця 2.3 – Огляд Google Cloud Natural Language

| | |
|---|---|
| Тип | Cloud Service (Google Cloud Platform) |
| Сильні сторони | Потужний синтаксичний аналіз, розпізнавання сутностей із оцінкою достовірності, аналіз настроїв із деталізованими емоціями. |
| Недоліки | Точність може змінюватись в залежності від складності тексту. Обмежене розпізнавання іменованих об'єктів для деяких мов. |
| Ключова особливість | Аналіз тональності, розпізнавання сутностей, синтаксичний аналіз, аналіз залежностей, маркування частин мови |
| Мови, що аідтримуються | Англійська, іспанська, французька, німецька і т. д. |
| Ціноутворення | Оплата за фактом використання |

IBM Watson Natural Language Understanding – це хмарна служба обробки природної мови (NLP), яку пропонує IBM Cloud. Він використовує моделі глибокого навчання для аналізу неструктурованих текстових даних та отримання цінної інформації.

Таблиця 2.4 – Огляд IBM Watson Natural Language Understanding

| | |
|---|---|
| Тип | Cloud Service (IBM Cloud) |
| Сильні сторони | Можливості індивідуального налаштування, орієнтація на галузеві потреби. |
| Недоліки | Може бути складним у налаштуванні, структура ціноутворення може бути непрозорою |
| Ключова особливість | Аналіз тональності, розпізнавання сутностей, вилучення відносин, вилучення ключових слів, класифікація категорій |
| Мови, що аідтримуються | Англійська, іспанська, французька, німецька і т. д. |
| Ціноутворення | Оплата за фактом використання |

Microsoft Azure Text Analytics – це хмарна служба обробки природної мови (NLP), яку пропонує Microsoft Azure.

Таблиця 2.5 – Огляд Microsoft Azure Text Analytics

| | |
|---|---|
| Тип | Cloud Service (Microsoft Azure) |
| Сильні сторони | Готові моделі, звичайна інтеграція з інструментами Azure, виявлення особистих даних. |
| Недоліки | Обмежені можливості налаштування, фокусування на готових моделях. |
| Ключова особливість | Аналіз настроїв, вилучення ключових фраз, визначення мови, виявлення особистих даних, моделювання тем |
| Мови, що аідтримуються | Англійська, іспанська, французька, німецька і т. д. |
| Ціноутворення | Оплата за фактом використання |

VADER (Valence Aware Dictionary and Sentiment Reasoner) – це унікальний інструмент аналізу настроїв у галузі комп'ютерної лінгвістики. На відміну від хмарних сервісів, що покладаються на моделі машинного навчання, VADER використовує підхід, що базується на лексиці та правилах.

VADER використовує підхід, заснований на лексиці, використовуючи вже існуючий список слів із оцінками тональності. Позитивні терміни (відмінно) отримують позитивні оцінки, а негативні (жахливо) - негативні. Крім окремих слів, VADER включає міркування про почуття, засновані на правилах. Він враховує такі чинники, як використання великих букв (підвищена позитивність слова «ВІДМІННО!» в порівнянні з «відмінно»), пунктуація і заперечення («недобре», що виражає негатив).

Ніша VADER полягає в тому, що цей інструмент чудово аналізує неформальну лексику та сленг, що часто зустрічаються в повідомленнях у соціальних мережах. Цей фокус відрізняє VADER від інших інструментів для аналізу настроїв, в першу чергу призначених для формального тексту. Крім того, легковажність VADER, зумовлена його лексикою та підходом, що базується на правилах, робить його ефективним для швидкого аналізу великих наборів даних соціальних мереж.

Таблиця 2.6 – Огляд VADER

| | |
|---|---|
| Тип | Инструмент анализа настроений на основе лексикона |
| Сильні сторони | Простий у використанні, швидкий, лексикон охоплює мову соціальних мереж. |
| Недоліки | Менш точний для формальної мови, обмежений аналізом настроїв. |
| Ключова особливість | Аналіз настроїв (складний бал, конкретні емоції) |
| Мови, що аідтримуються | Переважно англійська |
| Ціноутворення | Безкоштовно, з відкритим вихідним кодом |

Хоча VADER пропонує цінні функціональні можливості, важливо протиставити його відомим хмарним сервісам аналізу настроїв (таблиця 2.7), таким як Amazon Comprehend, Cloud Natural Language, IBM Watson Natural Language Doesnging та Microsoft Azure Text Analytics. Ці послуги використовують моделі машинного навчання, навчені на великих текстових масивах, що дозволяє їм обробляти складні мовні нюанси і часто досягати чудової точності аналізу настроїв, особливо формального тексту. Крім того, ці послуги надають ширший набір функцій, що включає розпізнавання об'єктів (ідентифікацію людей, місць та речей), тематичне моделювання (ідентифікацію ключових тем) та синтаксичний аналіз (розуміння структури речень). Крім того, хмарні послуги зазвичай призначені для великомасштабної обробки тексту та пропонують моделі ціноутворення з оплатою в міру використання.

Таблиця 2.7 – Відмінності VADER від вищеописаних хмарних послуг

| Характерна риса | VADER | Хмарні послуги |
|---|---|---|
| Підхід | Лексикон та правила | Машинне навчання |
| Фокус | Соціальні медіа | Усі типи тексту |
| Додаткові можливості | Обмежене | Тональність, сутності, теми, синтаксис |
| Масштабованість | Найменші набори даних | Великі набори даних |
| Вартість | Безкоштовно (з відкритим вихідним кодом) | Оплата в міру використання |

Крім основних хмарних сервісів існує багата екосистема систем аналізу контенту, що задовольняють різні потреби.

Дослідницьке середовище має вигоду з безлічі бібліотек і фреймворків з відкритим вихідним кодом, що забезпечують потужні можливості контент-аналізу. Приклади включають spaCy (Python), Apache Stanford CoreNLP (Java)

та NLTK (Python). Ці інструменти пропонують дослідникам гнучкість та можливість налаштовувати конвеєри аналізу. Однак ефективне використання потребує досвіду програмування.

Для досліджень, спрямованих на виявлення тематичних структур текстових даних, спеціальні інструменти тематичного моделювання пропонують цінну інформацію. Для цієї мети особливо добре підходять такі інструменти, як MALLET (Java) та реалізації прихованого розподілу Диріхле (LDA) у різних мовах програмування.

Сфера контент-аналізу виходить за межі інструментів загального призначення. Залежно від конкретного типу досліджуваного контенту (наприклад, юридичних документів, медичних записів) можуть бути спеціалізовані галузеві інструменти. Ці інструменти часто попередньо навчаються на відповідних даних, специфічних для конкретної предметної галузі, що потенційно призводить до підвищення точності аналізу.

Хоча системи контент-аналізу з урахуванням правил менш поширені, вони пропонують альтернативний підхід. Ці системи покладаються на правила, що визначаються користувачем, для категоризації або класифікації текстових даних. Вони можуть бути корисні для завдань, що вимагають дуже специфічних критеріїв, але їм може не вистачати гнучкості та адаптованості, властивих підходам на основі машинного навчання.

Найбільш ефективна система контент-аналізу залежить від конкретних цілей дослідження та характеристик самих даних. До основних факторів, які слід враховувати під час виборів, належать:

- простота використання. Хмарні системи зазвичай характеризуються зручними інтерфейсами, у той час як інструменти з відкритим вихідним кодом часто вимагають програмування для ефективного використання;

- функціональність. Дослідники повинні визначити функції, найважливіші їхнього аналізу (наприклад, аналіз настроїв, тематичне моделювання, розпізнавання об'єктів). Вибір системи, що пропонує необхідні функції, забезпечує продуктивний робочий процес;

- сумісність даних. Важливо переконатися, що вибрана система може обробляти формат та розмір аналізованих текстових даних. Несумісні формати даних можуть призвести до серйозних проблем під час аналізу;
- вартість. У хмарних сервісах часто використовуються моделі оплати в міру використання, тоді як інструменти з відкритим кодом зазвичай можна використовувати безкоштовно. Проте інструменти з відкритим вихідним кодом вимагають інвестицій у встановлення та постійне обслуговування.

## 2.2 Аналіз технологій для визначення емоційного забарвлення тексту

Зростання області обробки природної мови (НЛП) пропонує багатий набір бібліотек і фреймворків, що ідеально підходять для завдань аналізу тексту, включаючи найважливішу здатність розрізняти емоційні підтексти, сплетені в письмовому контенті. У цьому розділі дається критична оцінка кількох відомих варіантів у цій галузі, підкреслюються їхні сильні сторони, обмеження та придатність для різноманітних дослідницьких проектів.

NLTK (Набір інструментів для природної мови). Ця базова бібліотека Python є наріжним каменем для безлічі починань НЛП. Крім своїх основних компетенцій у галузі токенізації, стеммінгу/лематизації та класифікації тексту, NLTK може похвалитися добре інтегрованим модулем аналізу настроїв. Цей зручний інтерфейс у поєднанні з великою документацією та безліччю навчальних посібників робить NLTK популярним вибором для дослідників, які приступають до вивчення ландшафту НЛП. Примітно, що словник аналізу настроїв NLTK включає показники інтенсивності настроїв, що дозволяє дослідникам розрізняти помірно позитивні та позитивні настрої в тексті. Крім того, NLTK поєднує інструменти для роботи з запереченням, найважливішим аспектом аналізу настроїв, оскільки слова, що заперечуються, можуть значно змінити загальний емоційний тон. Однак важливо визнати, що можливості аналізу настроїв NLTK можуть бути не такими складними як можливості інших бібліотек, що потенційно ускладнює

виявлення тонких емоційних варіацій у складному тексті.

spaCy. Ще один суворий суперник на арені Python NLP, spaCy розширює свою функціональність далеко за межі базового аналізу настроїв. Він надає дослідникам комплексний набір інструментів, що включає розпізнавання об'єктів, аналіз залежностей і розширені можливості аналізу настроїв. Модуль аналізу настроїв SpaCy використовує попередньо навчені статистичні моделі, що дозволяють з високою точністю ідентифікувати позитивні, негативні та нейтральні настрої. Крім того, spaCy відрізняється чудовою продуктивністю та може похвалитися повною інтеграцією з популярними платформами машинного навчання, такими як TensorFlow та PyTorch. Ця безшовна інтеграція дозволяє дослідникам використовувати можливості глибокого навчання для створення користувацьких моделей аналізу настроїв, адаптованих до конкретних областей та наборів даних. Однак великий набір функцій spaCy досягається за рахунок крутішої кривої навчання в порівнянні з такими бібліотеками, як NLTK. Освоєння всього потенціалу SpaCy потребує міцної основи у концепціях НЛП та, можливо, деякого досвіду роботи з платформами машинного навчання.

TextBlob. Ця бібліотека Python віддає пріоритет завданням аналізу настроїв, пропонуючи мінімалістичний інтерфейс порівняно з NLTK і spaCy. TextBlob полегшує швидку класифікацію тексту на позитивні, негативні чи нейтральні категорії, що робить його цінним інструментом для проектів, у яких основний акцент робиться на простий аналіз настроїв. Крім того, TextBlob включає базове виявлення полярності, що дозволяє дослідникам оцінити загальну емоційну інтенсивність тексту. Простота використання бібліотеки та доступні функції аналізу настроїв роблять її привабливим варіантом для швидкого прототипування досліджень і для дослідників, погано знайомих з НЛП. Однак важливо визнати, що функціональність TextBlob більш обмежена порівняно з іншими бібліотеками. Йому не вистачає розширених функцій, таких як розпізнавання іменованих об'єктів або можливість інтеграції з платформами машинного навчання для розробки

власних моделей. Таким чином, TextBlob може не підійти для складних проектів НЛП, які потребують детальнішого розуміння емоційного ландшафту тексту.

TensorFlow. Будучи універсальним середовищем глибокого навчання, TensorFlow виходить за межі готових бібліотек, таких як NLTK, spaCy та TextBlob. Це дає дослідникам можливість створювати власні моделі, старанно адаптовані до конкретних завдань аналізу настроїв. Цей підхід забезпечує найвищий рівень контролю та потенційно революційну точність, особливо при роботі з великими та складними наборами даних. Однак для використання всього потенціалу TensorFlow потрібен значний досвід програмування та глибоке розуміння концепцій глибокого навчання, таких як нейронні мережі та алгоритми зворотного розповсюдження помилки. Крім того, створення моделей з нуля може зайняти багато часу, що потенційно може ускладнити терміни дослідження.

Вивчаючи сильні та слабкі сторони кожної бібліотеки, ми зможемо ухвалити обґрунтоване рішення при виборі інструменту, що найбільш підходить для цілей нашого дослідження (таблиця 2.8).

Вибір бібліотеки або фреймворку для визначення емоційного забарвлення тексту виходить за рамки єдиної точки ухвалення рішення. Це вимагає детального розгляду різних факторів, що стосуються конкретного дослідницького проекту та технічної підготовки дослідника. У цьому розділі описується основа управління цим важливим процесом прийняття рішень.

Таблиця 2.8 – Чинники, що впливають на вибір

| Фактор | Опис |
|---|---|
| Объем проекта и требования | Сложность анализа настроений, которая требуется, и потребность в расширенных функциях НЛП. |

Продовження таблиці 2.8

| Фактор | Опис |
|---|---|
| Техническая экспертиза исследователя | Опыт работы в НЛП и знание языков программирования. |
| Использование предварительно обученных моделей | Наличие и качество предварительно обученных моделей |
| Баланс мощности и простоты использования | Приоритизация удобства использования или расширенных функций. |

Важливим етапом процесу відбору є глибоке розуміння масштабу та вимог дослідницького проекту. Ось кілька ключових питань, над якими варто задуматися:

- складність аналізу настроїв. Чи вимагає дослідження виявлення основних настроїв (позитивних, негативних, нейтральних) або ж для проекту вирішальними є нюанси емоційних категорій (радість, гнів, смуток тощо). Такі бібліотеки, як TextBlob, чудово справляються з базовим аналізом настроїв, а spaCy пропонує функціональні можливості виявлення більш широкого спектра емоцій;

- потреба в розширених функціях НЛП. Чи потрібні для проекту функції НЛП крім аналізу настроїв, таких як розпізнавання іменованих об'єктів чи тематичне моделювання? Якщо це так, комплексний набір інструментів spaCy може виявитися більш відповідним у порівнянні з бібліотеками більш вузької спрямованості, такими як TextBlob.

- Бюджетні обмеження. Розгляньте витрати на використання конкретних інструментів. Бібліотеки з відкритим кодом, такі як spaCy та TextBlob, безкоштовні, але можуть виникнути витрати, пов’язані з розширеними функціями, додатковими плагінами або хмарними рішеннями для великомасштабної обробки.

Таблиця 2.9 – Огляд сильних та слабких сторін розглянутих бібліотек та фреймворків

| Бібліотека/фреймворк | Сильні сторони | Слабкі сторони |
|---|---|---|
| NLTK (Natural Language Toolkit) | Зручний інтерфейс<br>Велика документація та навчальні посібники<br>Вбудований аналіз настроїв із оцінкою інтенсивності настроїв.<br>Обробляє заперечення | Менш складний аналіз настроїв, порівняно з іншими варіантами<br>Може не передавати нюанси емоцій |
| spaCy | Розширений аналіз настроїв за допомогою попередньо навчених моделей<br>Висока продуктивність<br>Інтегрується з платформами машинного навчання<br>Пропонує функціональні можливості, що виходять за рамки аналізу настроїв | Потрібний деякий досвід роботи з платформами машинного навчання для розробки користувацьких моделей. |
| TextBlob | Простий та зручний інтерфейс<br>Базовий аналіз настроїв та виявлення полярності<br>Швидке прототипування | Обмежені функціональні можливості в порівнянні з іншими бібліотеками.<br>Не вистачає розширених функцій НЛП та можливостей розробки користувацьких моделей. |
| TensorFlow | Максимальний контроль та налаштування<br>Потенціал для революційної точності<br>Можливості глибокого навчання | Потрібен значний досвід програмування<br>Трудомістка розробка індивідуальної моделі |

Ретельно оцінивши сильні та слабкі сторони різних бібліотек та платформ, ми вирішили розпочати дослідження аналізу настроїв у текстових даних за допомогою Natural Language Toolkit (NLTK). Це рішення є результатом вдумливого розгляду конкретних вимог проекту та технічного досвіду дослідника.

На початковому етапі дослідження пріоритет надається розвитку фундаментального розуміння методів аналізу настроїв у тих текстових даних. Зручний інтерфейс NLTK та велика документація роблять його ідеальною платформою для цього початкового етапу. Його вбудовані функції аналізу настроїв, які включають оцінку інтенсивності настроїв та обробку заперечень, забезпечують міцну основу досягнення основних цілей цього початкового етапу дослідження.

У зв'язку з початковим знайомством з концепціями НЛП, щадна крива навчання NLTK виявляється перевагою. Легкодоступні навчальні посібники та велика документація дозволяють швидко зрозуміти основи аналізу настроїв та функціональні можливості NLTK. Це сприяє більш ефективному дослідницькому процесу, дозволяючи швидше експериментувати та аналізувати текстові дані.

Хоча NLTK є початковим майданчиком для цього дослідження, ми визнаємо потенційну потребу в більш сучасних функціях у міру розвитку. Модульна конструкція NLTK забезпечує плавну інтеграцію з іншими бібліотеками, такими як spaCy, якщо траєкторія дослідження вимагатиме переходу до більш тонкого аналізу настроїв або включення додаткових функцій НЛП. Така гнучкість гарантує, що вибраний набір інструментів може адаптуватися та масштабуватися відповідно до вимог дослідження.

На закінчення відзначимо, що вибір NLTK як основна бібліотека для цього проекту аналізу настроїв сприяє створенню продуктивного дослідницького середовища. Його зручність, відповідність цілям проекту та вбудовані можливості аналізу настроїв роблять його ідеальною платформою для цього початкового етапу дослідження. Проте проект визнає потенційну потребу у досконаліших інструментах у майбутньому. Модульна природа NLTK у поєднанні з зростаючим досвідом дослідника забезпечує плавний перехід до SpaCy або інших бібліотек.

## 2.3 Аналіз технологій для генерації маркерів

У попередніх розділах була створена основа для аналізу настроїв, класифікації емоцій та розпізнавання іменованих об'єктів із вилученням тем. У цьому розділі розглядається найважливіший етап перетворення даних у формат, який полегшує створення медіаконтенту. Тут ми запроваджуємо поняття «маркери» (таблиця 2.10).

Таблиця 2.10 – Роль маркеру

| Функция | Описание |
|---|---|
| Зв'язувати текст та медіа | Виступати як проміжне подання між витягнутими даними та створенням медіаконтенту. |
| Інкапсулювати зміст | Концентрувати емоційний тон та тематичні області аналізу тексту. |
| Сприяння створенню медіа | Посібник зі створення різноманітних медіаформатів (зображення, аудіо, відео). |

Ці маркери є мостом між текстовими даними та сферою творчої генерації медіа. Об'єднуючи емоційний тон і тематичні області, отримані у процесі аналізу тексту, маркери забезпечують коротке і інтерпретоване уявлення наступних додатків. Маркери є концентрованими смисловими капсулами, кожна з яких містить у собі суть певного настрою або тематичної нитки аналізованого тексту.

Кінцева мета цього етапу полягає у використанні цих маркерів для створення різноманітного медіаконтенту. Цей контент може охоплювати цілий спектр форматів: від зображень, що запам'ятовуються, і тематичних звукових ландшафтів до, можливо, навіть відео. Ефективно перекладаючи емоційний та тематичний зміст тексту в ці «маркери», ми розкриваємо потенціал створення медіа, яке резонує з основною суттю даних, що аналізуються.

Визначивши мету «маркерів» як мосту для створення медіаконтенту,

потрібно пояснити ключові компоненти, якими повинна мати добре продумана підказка для генеративної моделі, як Stable Diffusion, для ефективного переведення даних у переконливі медіа. Ці компоненти виступають як будівельні блоки для успішної трансформації, спрямовуючи генеративну модель до створення елементів мистецтва, які точно відображають емоційну та тематичну суть аналізованого тексту.

Таблиця 2.11 – Компоненти підказки для створення маркера

| Компонент | Опис | Приклад | Результат |
|---|---|---|---|
| Емоційний тон | Передає текстовий настрій. | «Безтурботний гірський краєвид» | «радісний», «меланхолійний», «засмучений» |
| Тематичні прив'язки | Посилання ідентифікують теми чи названі об'єкти. | «Жвавий ринок» | Конкретні місця, історичні особи чи ширші категорії. |

Продовження таблиці 2.11

| Компонент | Опис | Приклад | Результат |
|---|---|---|---|
| Стилістичні переваги | Вказує бажаний художній стиль. | «Казковий морський пейзаж» | "реалістичний", "абстрактний", "імпресіоністичний". |

Найважливішим для маркера є його здатність уловлювати емоційний тон, почерпнутий з аналізу тексту. Це вимагає включення до підказки точних емоційних дескрипторів. Багатий словниковий запас, що включає такі терміни, як «радісний», «меланхолійний», «розчарований» або «обнадійливий», дозволяє генеративної моделі наповнювати згенеровані медіа відповідною емоційною вагою. Наприклад, підказка, що описує радісний текст, може вказувати на «яскраву, залиту сонцем сцену на пляжі», а меланхолійний текст може перекладатися підказкою із запитом «пустельний, затягнутий туманом міський пейзаж».

Тематичні області, визначені на етапі аналізу тексту, відіграють вирішальну роль у формуванні змісту створюваних медіа. Включивши ці

теми безпосередньо у підказку, ми забезпечуємо генеративну модель тематичними прив'язками. Це може включати посилання на конкретні іменовані об'єкти (наприклад, «історичні пам'ятки») або на більш широкі тематичні категорії (наприклад, «природні пейзажі», «міське середовище»). Наприклад, підказка, яка аналізує текст про історичну подію, може посилатися на конкретне «поле битви» чи «пам'ятник», а текст, присвячений проблемам довкілля, може перекладатися підказкою із запитом «густий ліс» чи «забруднений міський пейзаж».

Крім основних емоційних та тематичних елементів, підказки можуть також включати стилістичні уподобання. Вказівка художніх стилів, таких як «реалістичний», «абстрактний» або «імпресіоністичний», дозволяє додатково адаптувати створюваний медіаконтент. Цей рівень контролю дозволяє дослідникам досліджувати творчий потенціал генеративної моделі та експериментувати з різними художніми інтерпретаціями вилучених даних. Уявіть собі підказку з проханням створити реалістичний портрет для тексту, що аналізує історичну особистість, або абстрактну композицію для тексту, що досліджує емоції.

Стратегічно комбінуючи ці елементи – емоційні дескриптори, тематичні прив'язки та стилістичні уподобання – можна створювати ефективні підказки, які спрямовують генеративну модель до створення медіаконтенту, який резонує з основною суттю тексту, що аналізується. Ці підказки діють як диригентська паличка, організовуючи перетворення текстових даних на багату симфонію емоцій, тем і художнього вираження, втілених у створених медіа.

Встановивши важливість добре продуманих підказок (prompt), ми тепер заглибимося в область технологій, здатних використовувати ці підказки для створення бажаного медіаконтенту (таблиця 2.12), зрештою, формуючи наші «маркери».

Таблиця 2.12 – Генеративні технології створення медіа

| Технології | Опис | Сильні сторони під час створення маркерів |
|---|---|---|
| Моделі тексту до зображення (наприклад, Stable Diffusion, Dall-E 2) | Генерує зображення з текстових описів | налагоджена технологія<br>якісна генерація зображень<br>стилістичний контроль<br>Пряме візуальне подання тем та емоційного тону. |
| Синтез тексту в аудіо (наприклад WaveNet, MelNet) | Створення аудіофрагментів на основі тексту. | емоції за допомогою звукових ландшафтів та музики доповнює емоційним звуковим супроводом. |
| Генерація тексту у відео (наприклад, VQ-GAN, AttnGAN) | Створено відеофрагменти з тексту | нові технології захоплюють потік оповідання<br>майбутній потенціал динамічних відеомаркерів. |

В авангарді знаходяться моделі дифузії text-to-image, такі як Stable Diffusion та Dall-E 2. Ці моделі зробили революцію в галузі генерації зображень, продемонструвавши виняткову здатність переводити текстові описи у захоплюючі візуальні ефекти. Включаючи емоційні дескриптори та тематичні прив'язки з підказок у процес генерації, ці моделі можуть створювати образи, що втілюють основну суть тексту, що аналізується. Уявіть собі новинну статтю про науковий прорив, що спричинив створення футуристичної лабораторної сцени, або сповнений тугою вірш, втілений у фотореалістичне зображення пустельного пляжу на заході сонця. Потенціал цих моделей щодо перекладу емоційних та тематичних нюансів тексту у візуально привабливі «маркери» незаперечний.

Моделі синтезу тексту в аудіо, такі як WaveNet та MelNet, пропонують захоплюючі можливості для створення аудіомакерів, які резонують з емоційним тоном, витягнутим із текстових даних. Ці моделі чудово підходять для створення аудіофайлів, що включають як музику, так і звукові ефекти, які можуть викликати у слухача певні емоції. Включно з емоційними дескрипторами з підказок, ці моделі можуть створювати звукові ландшафти,

що відображають настрій аналізованого тексту. Наприклад, аналіз тексту, що виявляє гнів, може перевестися в підказку з проханням створити звуковий ландшафт, наповнений різкими шумами і музикою, що дисонує, а текст, наповнений радістю, може послужити підказкою для надихаючого і енергійного саундтреку. Синтез тексту в аудіо пропонує потужний інструмент для створення маркерів, які не тільки вловлюють візуальну суть тексту, але і викликають його емоційну суть за допомогою сили звуку.

Моделі перетворення тексту на відео, такі як VQ-GAN і AttnGAN, все ще знаходяться на початковій стадії розробки і демонструють величезний потенціал для майбутнього створення маркерів. Метою цих моделей є створення відеопослідовності на основі текстових описів. Включаючи емоційну мову та тематичні елементи з підказок, ці моделі обіцяють створювати короткі відеоролики, які відображають оповідальний хід та емоційну складову аналізованого тексту. Уявіть собі новинний репортаж про політичну подію, переведений у відео, що показує ключові моменти події, або вірш, що описує подорож, переведений у відеоряд, що демонструє різні пейзажі, що зустрічаються. Можливість генерувати динамічні відеомаркери відкриває двері для багатшого і захоплюючого представлення даних.

Технології створення генеративних медіа пропонують різноманітну палітру інструментів для створення маркерів. Від визнаної майстерності моделей дифузії text-to-image до зростаючого потенціалу генерації тексту у відео. У міру того, як ці технології продовжують розвиватися і вдосконалюватися, можливості створення тонких і захоплюючих «маркерів», безсумнівно, розширюватимуться, сприяючи багатшому зв'язку між текстовими даними та сферою творчих медіа, що запам'ятовується..

Основна мета цього проекту полягає у створенні візуальних уявлень – «маркерів», які інкапсулюють емоційний тон та тематичні галузі, витягнуті з текстових даних. Серед вивчених технологій моделі дифузії тексту зображення, такі як Stable Diffusion і Dall-E 2, пропонують найбільш прямий і усталений підхід для досягнення цієї мети. Їхнє вміння переводити текстові

описи в захоплюючі зображення ідеально відповідає вимогам проекту.

Тематичні області, визначені етапі аналізу тексту, можуть охоплювати спектр складності. У той час, як деякі теми можуть бути легко представлені візуально (наприклад, «природна сцена», «міський пейзаж»), інші можуть вимагати більш тонкого підходу (наприклад, «політична сатира», «філософська концепція»). Моделі дифузії тексту у зображення продемонстрували чудову здатність генерувати візуально насичені та докладні зображення, що робить їх добре підходящими для відображення суті як простих, так і складних тематичних елементів у вилучених даних.

Моделі дифузії тексту до зображення в даний час являють собою передові технології створення генеративних медіа з точки зору якості зображення і керованості. Їхня здатність враховувати стилістичні уподобання в процесі генерації дозволяє дослідникам адаптувати візуальну естетику «маркерів» для подальшого покращення представлення аналізованого тексту. Крім того, в порівнянні з моделями генерації тексту в аудіо і текстом у відео, моделі дифузії тексту в зображення мають більшу доступність. Їхні інтерфейси користувача та вимоги до навчання часто більш доступні дослідникам, що сприяє більш плавній інтеграції в робочий процес проекту.

Хоча моделі дифузії тексту до зображення служать основним напрямом створення маркерів у цьому проекті, система за своєю суттю зберігає певну гнучкість. Потенціал для включення інших технологій створення генеративних медіа, таких як синтез тексту в аудіо для створення аудіомакерів, що доповнюють згенеровані зображення, залишається відкритим. У міру розвитку моделей перетворення тексту у відео їх включення до створення динамічних відеомаркерів, які фіксують оповідальний потік аналізованого тексту, може стати життєздатним варіантом у майбутньому. Ця відкритість для майбутньої інтеграції гарантує, що проект зможе адаптуватися та використовувати досягнення у технологіях створення генеративних медіа у міру їх подальшого розвитку.

# 3 ФОРМАЛЬНИЙ ОПИС РОЗРОБКИ

## 3.1 Запропонована модель

Запропонована модель для автоматизованого маркування емоційно-забарвленої лексики використовує багатогранний підхід, інтегруючи різні комп’ютерні лінгвістичні техніки для обробки та аналізу тексту. Ця система розроблена для розкладання вхідного тексту на зручні для взаємодії блоки, обробки цих фрагментів за допомогою кількох блоків аналізу та створення маркерів, які інкапсулюють результати аналізу. Ці маркери згодом використовуються для генерації «промптів», дотримуючись попередньо встановленого шаблону цього промпту.

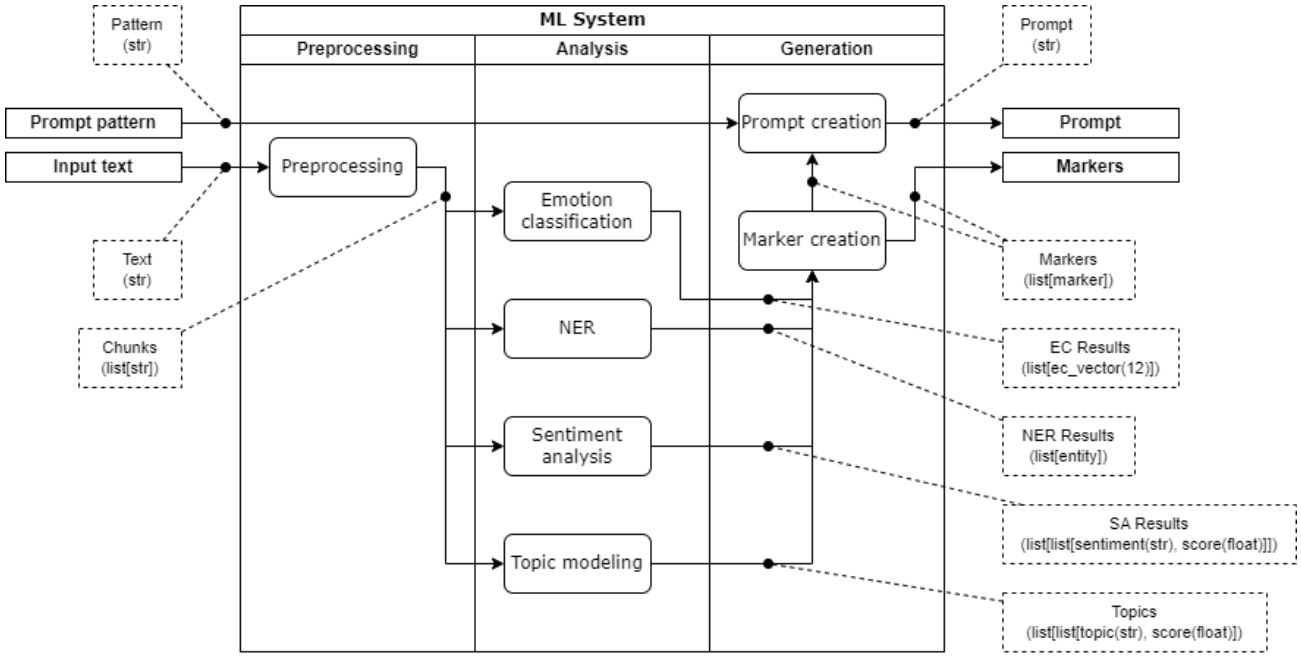


Рисунок 3.1 – Концептуальна модель системи маркування емоційно-забарвленої лексики

Архітектура запропонованої моделі складається з трьох основних модулів: модуля попередньої обробки, аналізу вмісту (включає класифікатор емоцій, розпізнавання іменованих сутностей, аналіз тональності тексту, тематичне моделювання) та модуля генерації (включає створення маркерів,

створення промпту), якщо не враховувати попередню обробку тексту.

Промпт, в контексті генеративних моделей штучного інтелекту відноситься до набору інструкцій або підказок, які керують моделлю, вказуючи, що генерувати. По суті, це відправна точка для творчого процесу моделі.

Мета промпту — надати моделі штучного інтелекту достатньо інформації, щоб зрозуміти, який результат ви хочете отримати. Правильно створений промпт, має виконувати такі задачі:

- скеровувати результат моделі. Добре розроблений промпт спрямовує модель до створення вмісту, який відповідає бажаним цілям, будь то написання певного типу творчого текстового формату, створення музичного стилю певного стилю або створення зображення з бажаною естетикою;
- надає контекст. Промпт може установити фон або налаштування для процесу створення. Це допомагає моделі зрозуміти світ, для якого вона створює вміст, що призводить до більш узгоджених і релевантних результатів;
- керує стилем і тоном. Промпт можна використовувати, щоб впливати на стиль і тон створеного вмісту. Наприклад, вказати формальний чи неформальний тон для написаного тексту або веселий чи напружений настрій для зображення.

Маркер, в контексті запропонованої моделі, це вичерпна контекстуальна та синтаксична інформація про певний сегмент тексту. За допомогою створення таких маркерів можна, надалі створити промпт, який вдало буде вбирати в себе всю потрібну інформацію для генерації контенту на базі звичайного тексту.

Модель працює через структуровану послідовність кроків, починаючи з сегментації вхідного тексту та завершуючись генерацією маркерів, які використовуються для різноманітних подальших програм. Ключові компоненти та їх взаємодія описані нижче.

Обробка вхідних даних.

Очищення введеного тексту – система приймає необроблений текст, який може надходити з різних доменів і контекстів. Вхідний текст спочатку попередньо обробляється для видалення зайвих символів і стандартизації форматування.

Розбиття тексту – попередньо оброблений текст ділиться на менші, узгоджені сегменти або «чанки», які будуть токенізовані у подальших компонентах моделі. Ця сегментація виконується для полегшення більш детального аналізу, гарантуючи, що кожен фрагмент має керований розмір для наступних етапів обробки. Стратегія декомпозиції може включати просту сегментацію на основі речень або більш складні методи.

Модуль аналізу вмісту.

Аналіз настроїв: кожна частина обробляється моделлю аналізу настроїв для визначення емоційного тону, який зазвичай класифікується як позитивний, негативний або нейтральний. Цей крок є вирішальним для виявлення емоційно насиченої лексики в тексті.

Розпізнавання іменованих сутностей (NER) – модель NER ідентифікує та класифікує власні іменники та інші важливі сутності в кожному фрагменті, надаючи контекст, який покращує розуміння тексту.

Моделювання теми – алгоритми моделювання теми застосовуються для визначення базових тем або тем у фрагментах. Це допомагає зрозуміти ширший контекст і релевантність тексту.

Емоційна класифікація – окрім базового аналізу настроїв, для класифікації тексту на певні емоції, такі як радість, гнів, смуток тощо, використовується детальніша модель емоційної класифікації.

Модуль генерації.

Інтеграція результатів. Виходи з різних моделей у модулі аналізу вмісту агрегуються для кожного блоку. Ця інтеграція передбачає об’єднання балів настроїв, ідентифікованих об’єктів, розподілу тем і емоційних класифікацій у комплексну структуру даних.

Генерація маркерів – на основі інтегрованих результатів система

генерує маркери для кожного блоку.

Маркер — це структуроване представлення, яке включає номер чанка, текст чанка та результати його аналізу. Цей структурований формат полегшує пошук і використання в подальших програмах.

## 3.2 Модуль аналізу вмісту

### 3.2.1 Блок класифікації емоцій

Блок класифікації емоцій є ключовим компонентом модуля аналізу вмісту, що дозволяє системі розпізнавати та класифікувати емоційний відтінок текстових даних. Цей процес використовує передові методи машинного навчання та попередньо навчені моделі трансформерів для точного визначення та класифікації широкого діапазону емоцій у текстових сегментах. У наступних розділах надається вичерпний огляд методології, архітектури моделі, наборів даних і показників оцінки, які використовуються в підмодулі класифікації емоцій.

Цей блок використовує потужність моделей на основі трансформаторів для точного визначення та класифікації емоцій у текстових даних. Розробка класифікатора включає кілька етапів, включаючи попередню обробку даних, навчання моделі, оцінку та розгортання. У цьому розділі ми розглянемо деталі створення та функціонування цієї моделі (EmotionClassifier).

Створення та налаштування моделі.

Створення моделі EmotionClassifier вимагало комплексного налаштування із залученням кількох ключових бібліотек і інструментів на Python. Нижче наведено детальне пояснення початкового налаштування та використовуваних компонентів.

Основні використовувані бібліотеки та інструменти включають:

- os – використовується для взаємодії з операційною системою для обробки файлових операцій;

- json – використовується для аналізу та обробки форматів даних JSON;

- torch – основна бібліотека для PyTorch, яка використовується для створення та навчання моделей глибокого навчання;

- pandas – використовується для обробки та аналізу даних, зокрема для обробки наборів даних у табличному форматі;

- numpy – використовується для числових операцій, забезпечуючи підтримку великих багатовимірних масивів і матриць;

- GradScaler і autocast від torch.cuda.amp – вони використовуються для навчання змішаної точності, що допомагає пришвидшити навчання та зменшити використання пам’яті без шкоди для точності моделі;

- DataLoader і Dataset з torch.utils.data – ці класи використовуються для ефективної обробки та завантаження даних під час процесу навчання;

- classification_report від sklearn.metrics – надає повний звіт про точність, запам’ятовування, оцінку F1 і підтримку для кожного класу, що є важливим для оцінки ефективності моделі;

- train_test_split із sklearn.model_selection – використовується для поділу набору даних на навчальні та тестові набори для оцінки ефективності моделі на невидимих даних;

- AdamW і get_linear_schedule_with_warmup з трансформаторів – оптимізатор і планувальник швидкості навчання використовуються для ефективного навчання моделі трансформатора;

- DistilBertTokenizer і DistilBertForSequenceClassification від transformers – Tokenizer і клас моделі для DistilBERT, легшого та швидшого варіанту BERT, спеціально розробленого для завдань класифікації послідовностей;

Методологія та архітектура моделі.

Для завдання класифікації емоцій ми використовуємо модель DistilBERT (Distilled Bidirectional Encoder Representations from Transformers), яка є спрощеною версією моделі BERT. DistilBERT зберігає 97% розуміння

мови BERT, будучи на 60% швидшим і на 40% меншим. Ця модель особливо добре підходить для обчислювально ефективної класифікації емоцій без шкоди для точності.

Архітектура моделі EmotionClassifier детально описується наступним чином:

Вбудовування DistilBERT – рівень вбудовування включає вбудовування слів, вбудовування позиції та компонент нормалізації шару. Вбудовування слів перетворює токени на щільні вектори фіксованого розміру (768 розмірів), тоді як вбудовування позиції включає позиційну інформацію у вектори маркерів.

Рівні трансформатора – ядро DistilBERT складається з шести блоків трансформаторів, кожен з яких містить багатоголовий механізм самоконтролю та нейронну мережу прямого зв’язку. Механізм самоуважності з кількома головами дозволяє моделі зосереджуватися на різних частинах вхідної послідовності одночасно, покращуючи контекстне розуміння.

Попередній класифікатор і класифікатор – лінійний рівень попереднього класифікатора зводить 768-вимірні вектори до нижчого вимірного представлення, яке потім передається на рівень класифікатора. Рівень класифікатора виводить логіти для кожного з 12 класів емоцій.

Шари вилучення: шари вилучення з імовірністю 0,1 і 0,2 використовуються в компонентах трансформатора та класифікатора відповідно, щоб запобігти переобладнанню.

Лістинг 3.1 – Детальний опис моделі EmotionClassifier

```
EmotionClassifier(
  (bert): DistilBertForSequenceClassification(
    (distilbert): DistilBertModel(
      (embeddings): Embeddings(
        (word_embeddings): Embedding(30522, 768, padding_idx=0)
        (position_embeddings): Embedding(512, 768)
        (LayerNorm): LayerNorm((768,), eps=1e-12,
```

```
elementwise_affine=True)
        (dropout): Dropout(p=0.1, inplace=False)
      )
      (transformer): Transformer(
        (layer): ModuleList(
          (0-5): 6 x TransformerBlock(
          (attention): MultiHeadSelfAttention(
            (dropout): Dropout(p=0.1, inplace=False)
            (q_lin): Linear(in_features=768, out_features=768,
bias=True)
            (k_lin): Linear(in_features=768, out_features=768,
bias=True)
            (v_lin): Linear(in_features=768, out_features=768,
bias=True)
            (out_lin): Linear(in_features=768, out_features=768,
bias=True)
          )
          (sa_layer_norm): LayerNorm((768,), eps=1e-12,
elementwise_affine=True)
          (ffn): FFN(
            (dropout): Dropout(p=0.1, inplace=False)
            (lin1): Linear(in_features=768, out_features=3072,
bias=True)
            (lin2): Linear(in_features=3072, out_features=768,
bias=True)
            (activation): GELUActivation()
          )
          (output_layer_norm): LayerNorm((768,), eps=1e-12,
elementwise_affine=True)
          )
        )
      )
    )
    (pre_classifier): Linear(in_features=768, out_features=768,
bias=True)
    (classifier): Linear(in_features=768, out_features=13,
bias=True)
    (dropout): Dropout(p=0.2, inplace=False)
  )
)
```

Обробка та пакетування даних.

У задачі класифікації емоцій вхідні текстові дані проходять ретельний процес перетворення перед подачею в модель. Цей процес гарантує, що дані мають правильний формат і оптимізовані для ефективного та ефективного навчання та оцінки моделі.

DataLoader і Batching.

Вхідні дані спочатку перетворюються на набір даних (Dataset),

спеціалізований клас із бібліотеки PyTorch, призначений для обробки та керування даними. Потім цей набір даних обробляється за допомогою DataLoader, яка є іншою утилітою PyTorch, яка полегшує групування, перемішування та завантаження даних у спрощений спосіб. Пакетування (Batching) має вирішальне значення, оскільки воно дозволяє моделі обробляти кілька зразків одночасно, значно підвищуючи ефективність обчислень і прискорюючи процес навчання.

torch.utils.data.DataLoader – утиліта яка забезпечує ітерацію над заданим набором даних. Вона підтримує автоматичне пакетування, вибірку, перемішування та багатопроцесне завантаження даних, що робить його важливим компонентом для обробки великих наборів даних. Поділяючи дані на менші керовані пакети, DataLoader гарантує, що кожен пакет містить фіксовану кількість зразків (розмір пакету), які обробляються разом за один прохід вперед і назад під час навчання.

У нашій реалізації максимальна довжина кожної послідовності в наборі даних становить 128 токенів. Це означає, що будь-який текст, довший за 128 токенів, скорочується, а коротші тексти доповнюються для забезпечення однакової довжини. Ця однаковість необхідна для ефективної паралельної обробки на сучасному апаратному забезпеченні, такому як графічні процесори. Довжина в 128 токенів була обрана так як приблизно відповідає максимальній довжені токенізованого повідомлення в X (раніше відомий як Твіттер), так як більшість даних навчального набору отримана саме з нього.

Тензорне заповнення (padding) та маски уваги.

У наборі даних завантажувач даних повертає два ключові елементи для кожного пакета: input_ids і attention_mask.

input_ids – це індекси токенів, які є числовими представленнями токенів, які складають вхідний текст. Кожен токен у тексті перетворюється на унікальний індекс за допомогою DistilBertTokenizer. Цей процес перетворює текст на послідовність чисел, яку може обробити модель.

attention_mask – це бінарна маска, яка вказує, на які токени має

звертати увагу модель, а які слід ігнорувати. Значення 1 у масці уваги означає, що слід звернути увагу на відповідний маркер, тоді як значення 0 вказує на те, що його слід ігнорувати (наприклад, токени заповнення).

Цей підхід, що використовує як input_ids, так і attention_mask, відомий за свою ефективність у обробці різноманітних текстових даних змінної довжини. Однак, щоб застосувати цей підхід, необхідно було адаптувати модель DistilBertForSequenceClassification до цих вхідних параметрів.

Модель EmotionClassifier.

EmotionClassifier побудовано на базі моделі DistilBertForSequenceClassification. Він включає кілька модифікацій, щоб гарантувати, що він може обробляти конкретні вхідні дані, які ми надаємо, і виводити бажані класифікації емоцій.

Ініціалізація та адаптація моделі.

Функція ініціалізації EmotionClassifier завантажує попередньо підготовлену модель distilbert-base-uncased. Ця модель, яка є легшим і швидшим варіантом BERT, добре підходить для завдань класифікації послідовності. EmotionClassifier успадковує DistilBertForSequenceClassification, тобто зберігає всі корисні властивості базової моделі, розширюючи свої можливості для задоволення наших конкретних вимог.

Функція прямого проходу EmotionClassifier адаптована для використання input_ids і attention_mask. Це гарантує, що під час прямого проходу моделі вона правильно обробляє індекси маркерів і звертає увагу на відповідні маркери, як зазначено маскою уваги.

Функція пакетування та активації.

Під час класифікації модель обробляє дані пакетами, причому розмір пакету є параметром, який можна налаштувати. Пакетування не тільки прискорює обчислення, але й допомагає краще використовувати ресурси GPU, що сприяє більш ефективному навчанню.

Функція активації це сигмоїда – останній рівень моделі використовує

функцію активації сигмоподібної форми. Сигмоїдна функція здавлює вхідні значення в діапазон від 0 до 1, фактично перетворюючи вихідні результати (логіти) на ймовірності. Кожне значення ймовірності вказує на ймовірність того, що введений текст відповідає певній емоції.

Вихідні дані моделі є вектором із 13 значень, кожне з яких представляє ймовірність того, що вхідний текст належить до однієї з 13 попередньо визначених категорій емоцій: радість, гнів, смуток, відраза, страх, довіра, здивування, любов, збентеження, очікування, сором, провина та ноемо (немає емоцій). Цей вектор можна вважати 13-вимірним представленням тексту в емоційному просторі, що забезпечує комплексну оцінку емоційного змісту тексту.

Детальна конфігурація та адаптація EmotionClassifier гарантує, що він ефективно класифікує текст за вказаними емоційними категоріями, використовуючи потужні можливості трансформаторних моделей і ефективність утиліт PyTorch.

Склад і підготовка набору даних.

Для навчання EmotionClassifier ми використовували різноманітну та повну колекцію добре відомих наборів даних.

Таблиця 3.1 – включені набори даних

| Набір даних | Опис | Характеристики |
|---|---|---|
| EmoBank | Корпус із 10 000 речень англійською мовою, анотованих балами реальних емоцій за трьома вимірами (валентність, збудження, домінування). | Анотований VAD (валентність, збудження, домінування), реальні оцінки, 10 000 речень, англійська мова. |
| TEC (Textual Emotion Corpus) | Містить англійські речення, витягнуті із заголовків новин і вручну додані в анотації однією з шести основних емоцій. | Шість емоцій (гнів, огида, страх, радість, смуток, подив), англійська, заголовки новин. |

Продовження таблиці 3.1

| Набір даних | Опис | Характеристики |
|---|---|---|
| AffectiveText | Складається з 1250 заголовків новин, анотованих за категоріями валентності, активації та емоцій. | Валентність, активація, категорії емоцій, заголовки новин, англійська, 1250 випадків. |
| Crowdflower | Набір даних твітів, анотованих Crowdflower для таких емоцій, як щастя, смуток, гнів тощо. | Численні емоції, твіти, краудсорсингові анотації, англійська. |
| DailyDialog | Містить діалоги, які відображають щоденне спілкування та вручну позначені емоціями та діалоговими актами. | Щоденне спілкування, мітки емоцій, діалоги, англійська, 13 118 діалогів. |
| ElectoralTweets | Твіти про президентські вибори в США 2016, коментовані різними емоціями. | Твіти, численні емоції, політичний контекст, англійська, 10 000 твітів. |
| EmoInt (Emotion Intensity) | Містить твіти з інтенсивністю емоцій для чотирьох основних емоцій: гніву, страху, радості та смутку. | Інтенсивність емоцій, чотири емоції, твіти, англійська, з примітками щодо інтенсивності. |
| Emotion-Cause | Набір даних, який зосереджується на визначенні речень у тексті, які передають причину певної емоції. | Причини емоцій, анотація на рівні речень, англійська мова, різноманітні джерела. |
| EmotionData-Aman | Анотований набір даних дописів у блозі для емоцій, зокрема щастя, смутку, гніву тощо. | Дописи в блозі, кілька емоцій, анотація вручну, англійська. |
| FB-Valence-Arousal-Anon | Містить дописи у Facebook, анотовані для валентності та збудження. | Дописи у Facebook, оцінки валентності та збудження, анонімні, англійська. |
| Grounded Emotions | Анотований набір даних, зосереджений на заземленні емоцій у фізичних ситуаціях, описаних у тексті. | Обґрунтування фізичної ситуації, численні емоції, англійська, різноманітні контексти. |

Продовження таблиці 3.1

| Набір даних | Опис | Характеристики |
|---|---|---|
| ISEAR | Дані опитування, де люди з різних культур описують ситуації, які викликали певні емоції. | Міжкультурний зв'язок, попередні емоції та реакції, дані опитування, численні емоції, англійська мова. |
| SSEC (Sarcasm Sentiment Emotion Corpus) | Анотований корпус твітів для сарказму, настроїв та емоцій. | Анотації про сарказм, почуття, емоції, твіти, англійська. |
| Tales-Emotion | Містить казки, коментовані на емоції. | Казки, множинні емоції, ручне анотування, англ. |

Кожен набір даних був очищений, щоб видалити будь-яку допоміжну інформацію, яка потенційно може завадити процесу навчання. Цей процес очищення гарантує, що використовуються лише основні текстові дані та відповідні мітки емоцій, забезпечуючи більш спрощений і ефективний набір даних для навчання моделі.

Розбиття даних.

Набір даних було розділено на три частини, щоб полегшити навчання, перевірку та тестування.

Навчальний набір (80%, це 11064) – використовується для навчання моделі.

Набір перевірки (10%, це 1366) – використовується для налаштування гіперпараметрів моделі та оцінки її продуктивності під час навчання.

Тестовий набір (10%, це 1366) – використовується для оцінки остаточної продуктивності моделі після навчання.

Це розбиття гарантує, що модель оцінюється на основі невидимих даних, забезпечуючи точнішу оцінку її здатності до узагальнення.

Токенізація та DataLoader.

DistilBertTokenizer.

Для токенізації ми використали DistilBertTokenizer із бібліотеки

трансформаторів Hugging Face. Токенізація — це процес перетворення необробленого тексту в токени, які є основними одиницями, які може обробляти модель. DistilBertTokenizer виконує кілька ключових завдань:

- поділ тексту – він розбиває введений текст на окремі маркери;
- зіставлення токенів на індекси – кожен токен зіставляється з унікальним індексом у словнику токенізера;
- обробка спеціальних маркерів: додає спеціальні маркери, необхідні для моделі BERT (наприклад, [CLS] для завдань класифікації).

Максимальна довжина кожної послідовності становить 128 токенів. Це означає, що будь-який текст, довший за 128 токенів, скорочується, а коротші тексти доповнюються для забезпечення однакової довжини. Токени заповнення додаються до послідовностей, щоб зробити їх однаковими за довжиною, що полегшує пакетну обробку.

DataLoader і Batching.

Під час навчання DataLoader використовується для обробки пакетів, перетасування та завантаження даних. Розмір партії для навчання встановлено на 16, що означає, що 16 зразків обробляються разом за один прохід вперед і назад. Саме такий розмір було обрано так як підчас первинних досліджень теми, більшість продуктивних та результативних моделей такого розміру з подібними задачами

Навчальний процес та кількість епох.

Модель навчена на 3 епохи. Під час наших експериментів ми помітили, що навчання протягом 2 епох призвело до точності близько 60%. Навчання понад 3 епохи призвело до переобладнання (overfiting), коли модель починає добре працювати на даних навчання, але погано на даних перевірки. Таким чином, 3 епохи встановили баланс між часом навчання та продуктивністю.

Використання обладнання.

Навчання проводиться на графічному процесорі, якщо він доступний, зокрема на NVIDIA RTX 3060 із 12 ГБ відеопам’яті, щоб використовувати його обчислювальну потужність. Якщо GPU недоступний, навчання

повертається до CPU.

Оптимізація та планування темпів навчання

Ми використовуємо оптимізатор AdamW, що означає Adaptive Moment Estimation with Weight Decay. AdamW є розширенням оптимізатора Adam, який інтегрує розпад ваги безпосередньо в крок оптимізації, а не змінює обчислення градієнта. Це допомагає упорядкувати модель і запобігти переобладнанню.

AdamW – коригує темпи навчання на основі середнього першого та другого моментів градієнтів, включаючи розпад ваги для кращої регулярізації.

Лінійний планувальник.

Для подальшого підвищення ефективності тренувань ми використовуємо лінійний планувальник з розминкою. Планувальник регулює швидкість навчання лінійно, починаючи з періоду розминки, коли швидкість навчання поступово збільшується до максимального значення, а потім лінійно зменшується.

Лінійний планувальник із розминкою – допомагає стабілізувати процес навчання, поступово збільшуючи швидкість навчання на початку, а потім зменшуючи її, запобігаючи раптовим великим оновленням, які можуть дестабілізувати навчання.

Тренування змішаної точності.

Для навчання змішаної точності ми ініціалізуємо GradScaler. Навчання змішаної точності використовує як 16-бітні, так і 32-бітні числа з плаваючою комою для підвищення ефективності обчислень без шкоди для точності моделі. GradScaler допомагає масштабувати градієнти під час зворотного поширення, зберігаючи стабільність процесу навчання.

GradScaler – масштабує градієнти, щоб запобігти проблемам недоповнення та переповнення під час навчання змішаної точності, що дає змогу використовувати переваги швидкості та пам’яті 16-бітної точності.

Функція втрати.

Використовувана функція втрати — torch.nn.BCEWithLogitsLoss, яка підходить для завдань класифікації з кількома мітками. Ця функція поєднує сигмоподібну активацію з бінарною перехресною втратою ентропії, забезпечуючи стабільний і ефективний показник втрати для класифікації багатьох емоцій.

BCEWithLogitsLoss – поєднує сигмоподібну активацію з бінарною перехресною втратою ентропії, що ідеально підходить для класифікації за кількома мітками, де кожен зразок може належати до кількох класів.

Навчання та оцінка.

Модель оцінюється після кожної епохи. Оцінка передбачає обчислення втрат на перевірочному наборі, а модель з найменшими втратами підтвердження зберігається. Це гарантує збереження найефективнішої моделі.

Оцінка проводиться після кожної епохи, і зберігається модель із найменшими втратами перевірки.

Тривалість навчання.

На зазначеному обладнанні (NVIDIA RTX 3060) кожна епоха займає приблизно 12 хвилин. Цей ефективний час навчання пояснюється використанням змішаного навчання точності та ефективної обробки даних через DataLoader.

Тестування та інтерпретація.

Після завершення навчання модель тестується на тестовому наборі.

Хоча модель звісно не є досконалою, вона працює адекватно для наших цілей. Результати класифікації, навіть якщо вони неточні, дають цінну інформацію, оскільки вони допускають численні інтерпретації, збагачуючи аналіз емоційного змісту в текстах.

Цей детальний процес навчання та оцінки гарантує, що EmotionClassifier є надійним, ефективним і здатним забезпечувати нюанси класифікації емоцій, утворюючи важливий компонент нашого модуля аналізу вмісту.

Експеременти.

Ми провели серію експериментів, щоб зрозуміти вплив різних параметрів на точність моделі та швидкість навчання. Досліджувані параметри включають розмір набору даних, розмір пакета, кількість епох, оптимізатор, планувальник і масштабувальник. Результати цих експериментів підсумовані в наступних таблицях і графіках.

Ми змінювали розмір набору даних навчання, щоб спостерігати його вплив на точність моделі та тривалість навчання.

Таблиця 3.2 – Вплив розміру набору даних на точність і швидкість навчання

| Розмір набору даних | Середня точність (%) | Час (хв. на епоху) |
|---|---|---|
| 1000 | 49 | 3 |
| 4000 | 52 | 6 |
| 7000 | 68 | 9 |
| 10000 | 72 | 11 |
| 13736 | 75 | 12 |

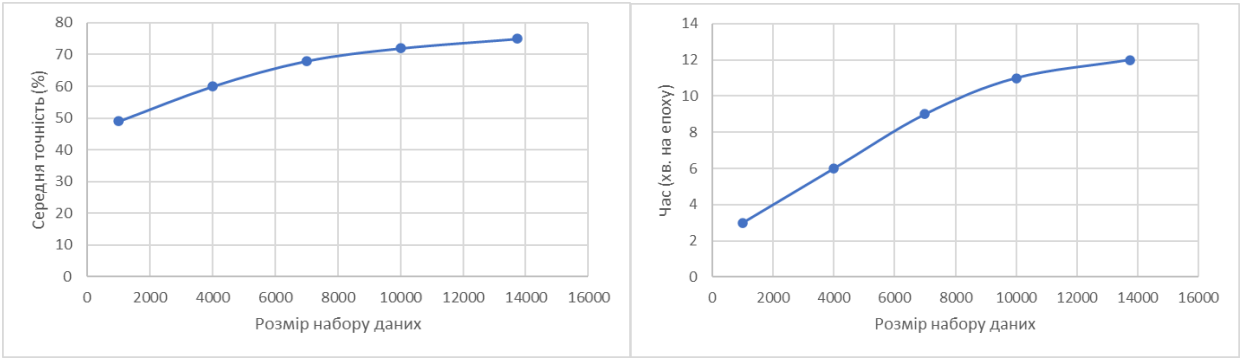


Рисунок 3.2а та 3.2б – візуалізація данних з таблиці 3.1

Розмір партії змінювався, щоб оцінити його вплив на продуктивність моделі та ефективність обчислень.

Таблиця 3.3 – Вплив розміру партії на точність і швидкість навчання

| Розмір партії | Середня точність (%) | Час (хв. на епоху) |
|---|---|---|
| 8 | 70 | 18 |
| 16 | 72 | 12 |
| 32 | 71 | 10 |
| 64 | 69 | 8 |
| 128 | 68 | 7 |

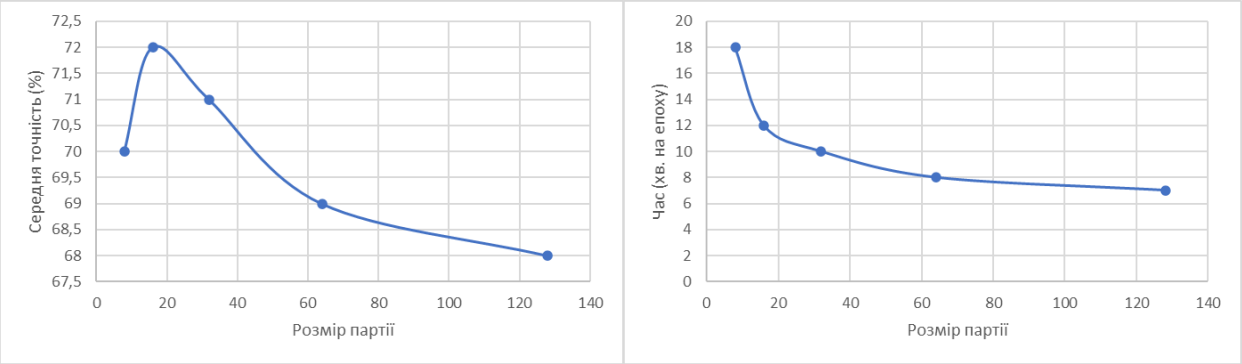


Рисунок 3.3а та 3.3б – візуалізація данних з таблиці 3.2

Ми навчили модель для різної кількості епох, щоб визначити вдалу тривалість навчання.

Таблиця 3.4 – Вплив кількості епохи на точність

| Кількість епох | Середня точність (%) | Overfitting (Так/Ні) |
|---|---|---|
| 1 | 55 | Ні |
| 2 | 60 | Ні |
| 3 | 72 | Ні |
| 4 | 70 | Так |
| 5 | 68 | Так |

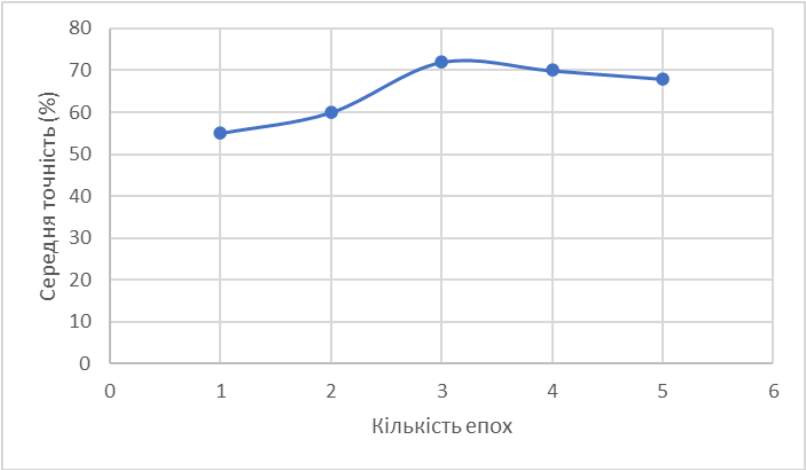


Рисунок 3.4 – візуалізація данних з таблиці 3.3

Вплив оптимізатора на точність і швидкість навчання.

Було протестовано різні оптимізатори для оцінки їх впливу на процес навчання та точність моделі.

Таблиця 3.5 – Вплив оптимізатора на точність і швидкість навчання

| Оптимізатор | Середня точність (%) | Час (хв. на епоху) |
|---|---|---|
| SGD | 65 | 14 |
| Adam | 70 | 12 |
| AdamW | 72 | 12 |
| RMSprop | 67 | 13 |
| Adagrad | 64 | 15 |

Вплив планувальника на точність і швидкість навчання.

Ми протестували різні планувальники темпів навчання, щоб спостерігати їхній вплив на ефективність і точність навчання.

Таблиця 3.6 – Вплив планувальника на точність і швидкість навчання

| Планувальник | Середня точність (%) | Час (хв. на епоху) |
|---|---|---|
| Відсутній | 73 | 25 |
| Linear with Warmup | 72 | 12 |

Продовження таблиці 3.6

| Планувальник | Середня точність (%) | Час (хв. на епоху) |
|---|---|---|
| Exponential Decay | 70 | 12 |
| StepLR | 69 | 13 |
| Cosine Annealing | 71 | 12 |

Використання змішаного тренування точності через GradScaler було проаналізовано щодо його впливу на ефективність і тривалість тренування.

Таблиця 3.7 – Вплив скейлера на точність і швидкість навчання

| Скейлер | Середня точність (%) | Час (хв. на епоху) |
|---|---|---|
| Відсутній | 70 | 16 |
| GradScaler | 72 | 12 |

Завдяки цим всебічним експериментам ми визначили, що адекватною конфігурація для EmotionClassifier передбачала повний розмір набору даних у, розмір партії в 16, навчання протягом 3 епох, використання оптимізатора AdamW і застосування лінійного планувальника з розминкою. Змішане тренування точності з GradScaler також підвищило швидкість і ефективність навчання. Ці відомості мають вирішальне значення для точного налаштування моделі для досягнення найкращого балансу між продуктивністю та використанням обчислювальних ресурсів.

Лістинг 3.2 – приклад результату класифікації емоцій

```
{
    "joy": 0.29762229323387146,
    "anger": 0.07040350884199142,
    "sadness": 0.46495646238327026,
    "disgust": 0.010444741696119308,
    "fear": 0.02401164546608925,
    "trust": 0.00021178011957090348,
    "surprise": 0.005082905292510986,
    "love": 0.0025421639438718557,
    "noemo": 0.030291352421045303,
```

```
    "confusion": 0.000005534648153115995,
    "anticipation": 0.00001234186856891028 6,
    "shame": 0.0018264853861182928,
    "guilt": 0.0032118172384798527
}
```


Приклади роботи блоку.

Текст 1 – «I was thrilled by the outstanding performance of the new iPhones camera, but the poor battery life left me frustrated.»

Переклад – «Я був у захваті від надзвичайних характеристик камери нового iPhone, але поганий час автономної роботи розчарував мене.»

Текст 2 – «The exhilarating last-minute victory of Manchester United over Chelsea made the entire crowd ecstatic.»

Переклад – «Хвилююча перемога «Манчестер Юнайтед» над «Челсі» на останній хвилині привела весь натовп у захват.»

Текст 3 – «I am deeply concerned about the lack of strong climate change policies, as they are essential for protecting our environment and ensuring a sustainable future.»

Переклад – «Я глибоко стурбований відсутністю чіткої політики щодо зміни клімату, оскільки вона є важливою для захисту нашого навколишнього середовища та забезпечення сталого майбутнього.»

Таблиця 3.8 – Демонстрація результаті роботи блоку класифікації емоцій

| Емоції | Текст 1 | Текст 2 | Текст 3 |
|---|---|---|---|
| joy | 0,2976 | 0,9748 | 0,0017 |
| anger | 0,0704 | 0,0022 | 0,0011 |
| sadness | 0,465 | 0,0026 | 0,0072 |
| disgust | 0,0104 | 0,0007 | 0,0006 |
| fear | 0,024 | 0,0018 | 0,3859 |
| trust | 0,0002 | 0,0017 | 0 |
| surprise | 0,0051 | 0,0015 | 0,0006 |
| love | 0,0025 | 0,0002 | 0,0002 |

Продовження таблиці 3.8

| Емоції | Текст 1 | Текст 2 | Текст 3 |
|---|---|---|---|
| noemo | 0,0303 | 0,0067 | 0,3148 |
| confusion | 0 | 0 | 0 |
| anticipation | 0 | 0 | 0 |
| shame | 0,0018 | 0 | 0,0002 |
| guilt | 0,0032 | 0 | 0,0002 |

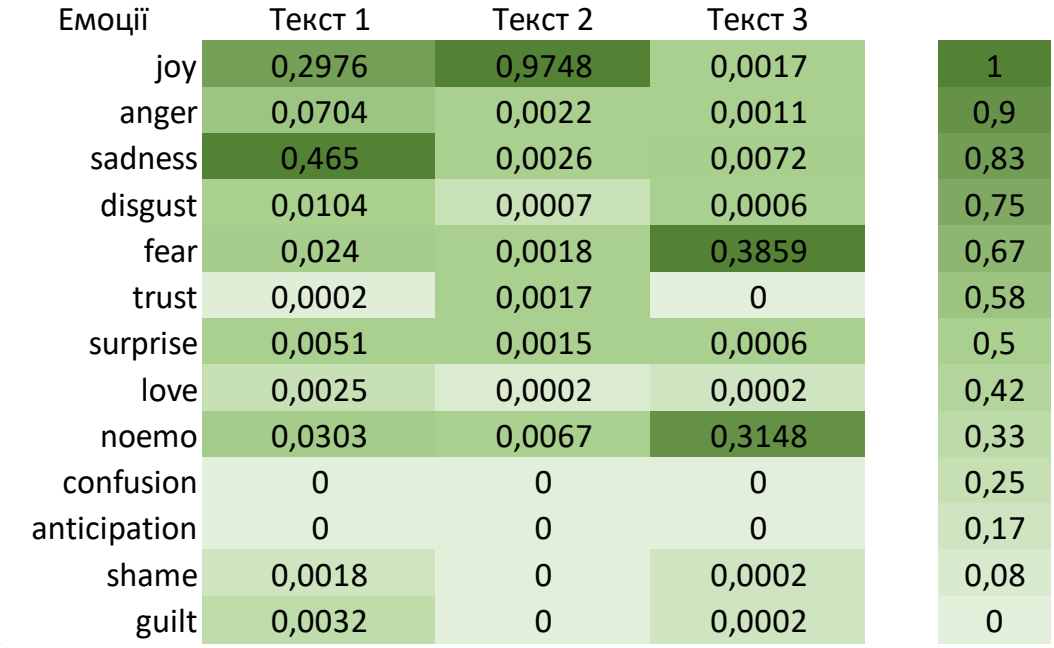


Рисунок 3.5 – Візуалізація отриманих результатів

Огляд.

Надані результати класифікації емоцій для трьох речень демонструють різні рівні інтенсивності різних емоцій. Аналіз буде зосереджений на інтерпретації цих результатів і розумінні емоційного відтінку кожного речення на основі наданих балів.

Речення 1 – «Я був у захваті від надзвичайних характеристик камери нового iPhone, але поганий час автономної роботи розчарував мене.»:

- радість (0,2976) – найвищий бал емоцій для цього речення – це радість, яка збігається з позитивним настроєм, висловленим щодо

продуктивності камери iPhone;

- смуток (0,465) – Смуток також значно високий, що відображає розчарування часом автономної роботи;

- гнів (0,0704) – гнів присутній, але не домінуючий, що свідчить про легке розчарування;

- інші емоції – незначні сліди огиди, страху, довіри, здивування та любові вказують на складну емоційну реакцію, але вони не переважають;

- ноемо (0,0303) – невелика частина речення не викликає сильних емоцій.

Інтерпретація.

Речення викликає поєднання радості та смутку, насамперед через контрастні враження від камери iPhone та часу автономної роботи.

Речення 2 – «Хвилююча перемога «Манчестер Юнайтед» над «Челсі» на останній хвилині привела весь натовп у захват.»:

- радість (0,9748) – радість переважно домінує в цьому реченні, що узгоджується з позитивним і захоплюючим характером змісту;

- інші емоції – незначні рівні гніву, смутку, огиди, страху, довіри, здивування та любові, що вказує на чітку та сильну позитивну емоційну реакцію;

- ноемо (0,0067) – дуже мала частина речення не викликає певних емоцій.

Інтерпретація.

Речення переважно радісне, відображає піднесення та екстаз натовпу після перемоги.

Речення 3 – «Я глибоко стурбований відсутністю чіткої політики щодо зміни клімату, оскільки вона є важливою для захисту нашого навколишнього середовища та забезпечення сталого майбутнього.»:

- страх (0,3859) – страх є переважаючою емоцією, яка підкреслює стурбованість політикою щодо зміни клімату;

- смуток (0,0072) – смуток присутній, але в набагато меншому ступені,

можливо, відображає занепокоєння;

- ноемо (0,3148) – значна частина речення є нейтральною або не дуже емоційною, що свідчить про більш фактичний тон;
- інші емоції – сліди гніву, огиди, довіри, подиву, любові, сорому та провини вказують на суміш занепокоєння та терміновості.

Інтерпретація.

Речення в першу чергу викликає страх із помітною часткою нейтральності, що вказує на серйозне занепокоєння та заклик до дій щодо політики щодо зміни клімату.

Висновки щодо класифікації

Речення 1 демонструє збалансоване поєднання радості та смутку, що відображає неоднозначні почуття мовця щодо функцій iPhone.

У другому реченні переважає радість, яскраво виражається хвилювання й щастя від спортивної перемоги.

Третє речення в першу чергу спричинене страхом, що вказує на занепокоєння щодо зміни клімату, із суттєвим нейтральним змістом, який передбачає фактичний тон.

Ці результати підкреслюють здатність системи вловлювати складні емоційні відтінки та надавати цінну інформацію про настрої та емоційний зміст різних типів речень.

### 3.2.2 Блок розпізнавання іменованих сутностей (NER)

Створення моделі.

Блок розпізнавання іменованих об’єктів (NER) у нашій моделі використовує вдосконалені методи машинного навчання для ідентифікації та класифікації іменованих об’єктів у тексті за попередньо визначеними категоріями, такими як імена осіб, організації, місця розташування, вираження часу, кількості, грошові значення, відсотки тощо. Цей процес має вирішальне значення для багатьох завдань обробки природної мови (NLP),

які забезпечують структуровану інформацію з неструктурованого тексту.

Середовище розробки та залежності.

Для розробки моделі NER ми використали такі бібліотеки та модулі Python.

datasets – ця бібліотека використовується для завантаження та попередньої обробки наборів даних. Він надає широкий спектр наборів даних у стандартизованому форматі, що спрощує завдання доступу та використання різних джерел даних для навчання та оцінювання.

evaluate – ця бібліотека є важливою для оцінки продуктивності моделі. Він включає різні показники, які можна використовувати для оцінки точності, точності, запам’ятовування та оцінки F1 моделі NER.

numpy – основний пакет для наукових обчислень на Python, numpy використовується для ефективних числових операцій над великими багатовимірними масивами та матрицями числових даних.

json – цей модуль використовується для обробки даних JSON, що є поширеним форматом для зберігання та обміну інформацією між конвеєром навчання та іншими компонентами системи.

Ключові трансформатори та інструменти

Ми використали кілька ключових трансформаторів і інструментів, наданих бібліотекою Hugging Face.

BertTokenizerFast – це швидка версія токенізатора BERT, яка підтримує ефективну токенізацію тексту. Токенізація — це процес перетворення послідовності символів у послідовність токенів, які можуть бути оброблені моделлю. Швидка версія оптимізована для швидкості та має вирішальне значення для ефективної обробки великих наборів даних.

AutoModelForTokenClassification – цей трансформатор автоматично вибирає модель, яка попередньо навчена для завдань класифікації маркерів, таких як NER. Використовуючи попередньо навчену модель, ми отримуємо вигоду від обширного навчання, яке вже було проведено на великих наборах даних, що значно підвищує продуктивність і точність нашої моделі NER.

DataCollatorForTokenClassification – цей інструмент використовується для зіставлення даних у пакети, придатні для навчання. Це гарантує, що дані правильно відформатовані та доповнені до однакової довжини, що важливо для ефективної пакетної обробки та узгодженого розміру вхідних даних у всій моделі.

TrainingArguments – цей модуль використовується для налаштування процесу навчання. Це дозволяє нам вказувати різні параметри, такі як швидкість навчання, розмір партії, кількість епох та інші гіперпараметри, які контролюють динаміку навчання.

Trainer – клас Trainer надає API високого рівня для навчання та оцінювання моделей. Він спрощує цикл навчання та бездоганно інтегрується з бібліотекою трансформаторів Hugging Face, дозволяючи нам зосередитися на точному налаштуванні моделі, а не на тонкощах процесу навчання.

pipeline – ця утиліта пропонує простий інтерфейс для виконання різних завдань NLP, включаючи NER. Використовуючи конвеєр, ми можемо швидко перевірити модель і інтегрувати її в більш широкі програми без необхідності глибоко заглиблюватися в базовий код.

Архітектура моделі.

У блоці розпізнавання іменованих об’єктів (NER) ми використовуємо складну архітектуру нейронної мережі, побудовану на основі попередньо навченої моделі BERT. Зокрема, ми використовуємо клас BertForTokenClassification, який розроблено для завдань класифікації на рівні маркерів, таких як NER. У цьому розділі детально розглядається структура цієї моделі, висвітлюються її різні компоненти та їхня роль у загальному процесі класифікації.

Модель BertForTokenClassification є спеціалізованою адаптацією стандартної моделі BERT, розробленою для класифікації окремих токенів у послідовності. Архітектура складається з кількох ключових компонентів, кожен з яких сприяє здатності моделі розпізнавати та класифікувати іменовані сутності в тексті.

BertModel.

Ядром BertForTokenClassification є BertModel, яка включає в себе вбудовування, кодувальник і пов’язані шари, які обробляють вхідний текст.

Вбудовування.

Вбудовування слів – рівень word_embeddings відображає кожен маркер у вхідній послідовності на високовимірний вектор розміром 768. Це щільне векторне представлення фіксує семантичне значення маркерів.

Вбудовані позиції – рівень position_embeddings кодує позицію кожного маркера в послідовності, дозволяючи моделі враховувати порядок маркерів, що є вирішальним для розуміння контексту.

Вбудовування типу маркерів – рівень token_type_embeddings розрізняє різні типи маркерів (наприклад, речення A проти речення B у таких завданнях, як відповідь на питання), хоча він менш актуальний для NER.

LayerNorm – рівень LayerNorm нормалізує вбудовування, гарантуючи, що вхідні дані для наступних шарів мають середнє значення нуль і стандартне відхилення одиниці, що допомагає стабілізувати та прискорити процес навчання.

Вилучення – шар вилучення застосовується до вбудовування, щоб запобігти переобладнанню шляхом випадкового встановлення частки вхідних одиниць на нуль під час навчання.

Енкодер.

Енкодер складається з 12 шарів BertLayer, кожен з яких містить декілька підрівнів.

Увага – у кожному BertLayer механізм BertAttention дозволяє моделі зосереджуватися на різних частинах вхідної послідовності під час прийняття рішень. Підрівень BertSelfAttention виконує самоконтроль кількох голів, обчислюючи показники уваги, які вказують на важливість інших токенів для поточного токена. Механізм уваги включає.

Лінійні перетворення запиту, ключа та значення – ці шари проектують введення вхідних даних у вектори запиту, ключа та значення, які

використовуються для обчислення показників уваги.

Випадання – застосовано до показників уваги, щоб запобігти переобладнанню.

Вихід – підрівень BertSelfOutput застосовує лінійне перетворення та нормалізацію шару до виходів уваги.

Проміжний рівень – підрівень BertIntermediate містить лінійне перетворення, яке проектує вхідні дані у простір з більшою вимірністю (3072 виміри), за яким слідує функція активації GELU (лінійна одиниця помилки Гауса), яка вводить нелінійність.

Вихід – підрівень BertOutput проектує проміжні виходи назад у вихідний 768-вимірний простір, застосовує нормалізацію шару та включає вилучення для регулярізації.

Класифікатор.

Останнім компонентом моделі BertForTokenClassification є рівень класифікатора. Цей лінійний рівень отримує вихідні дані кодера BERT (768-вимірний вектор для кожного токена) і проектує його на кількість класів об’єктів, які ми прагнемо розпізнати (у цьому випадку 9 класів). Результат цього рівня представляє ймовірність того, що кожен маркер належить до кожного з класів об’єктів.

Детальне пояснення шарів та їх функцій.

Рівень вбудовування – перетворює токени на щільні вектори, які може обробити модель. Він включає вбудовування слів, вбудовування позицій і вбудовування типу маркерів, доповнені нормалізацією шару та вилученням для кращого узагальнення.

Рівень кодувальника – складається з кількох блоків трансформатора, кожен з яких має увагу та проміжні підрівні, які ітеративно вдосконалюють представлення маркерів. Механізм уваги дозволяє моделі динамічно зважувати значимість різних токенів.

Рівень класифікатора – зіставляє уточнені представлення маркерів із кінцевими класами сутностей. Цей рівень має вирішальне значення для

завдання класифікації маркерів, де кожен маркер у послідовності класифікується незалежно.

Класифікація даних і використання моделі.

У блоці розпізнавання іменованих об’єктів (NER) ми використовуємо складні методи та попередньо навчені моделі для точної класифікації та вилучення іменованих об’єктів із вхідного тексту. У цьому розділі буде розглянуто особливості обробки вхідних даних, роль transformers.pipeline і моделі, доступні в блоці NER.

Класифікація даних і pipeline.

Коли справа доходить до класифікації вхідних даних, transformers.pipeline, наданий бібліотекою Hugging Face Transformers, відіграє вирішальну роль. Цей конвеєр (pipeline) значно спрощує обробку текстових даних, пропонуючи попередньо створені функції для різних завдань обробки природної мови, включаючи NER.

Пояснення pipeline.

Конвеєр у контексті трансформаторів – це абстракція високого рівня, яка оптимізує процес застосування попередньо навчених моделей до різних завдань НЛП.

Токенізація – конвеєр спочатку токенізує вхідний текст, перетворюючи його у формат, який може обробити модель.

Висновок моделі – токенізований текст потім подається в попередньо навчену модель, яка виконує потрібне завдання (у цьому випадку NER).

Постобробка – конвеєр виконує постобробку виходу моделі, перетворюючи її назад у формат, зрозумілий людині.

Використання transformers.pipeline гарантує, що весь процес, від введення необробленого тексту до вилучення іменованих об’єктів, є плавним та ефективним. Користувачі можуть вводити окремий фрагмент тексту або список текстів, а конвеєр автоматично обробляє групування, застосування моделі та форматування виводу.

Моделі для розпізнавання іменованих сутностей

У блоці NER ми використовуємо дві різні моделі, кожна з яких розроблена для певного типу класифікації.

Модель розпізнавання сутності.

Перша модель призначена для ідентифікації та класифікації різних сутностей у тексті, таких як особи, організації, місця та інші сутності.

Ці теги відповідають формату IOB (Inside-Outside-Beginning), який є загальною схемою тегів для завдань маркування послідовності. Цей формат допомагає розрізняти межі сутностей у тексті, таким чином дозволяючи моделі точно ідентифікувати багатослівні сутності.

Модель тегування частини мови (POS).

Друга модель зосереджена на тегуванні частин мови (POS), що передбачає присвоєння частин мови кожній лексемі у вхідному тексті. Ця модель може ідентифікувати повний набір тегів POS.

Таблиця 3.9 – Типи та класи сутностей

| Особи, організації, місця та інші сутності (NER tags) | Частини мови (POS tags) |
|---|---|
| O: За межами названої сутності<br>B-PER: Початок імені людини<br>I-PER: Всередині імені людини<br>B-ORG: початок назви організації<br>I-ORG: всередині назви організації<br>B-LOC: Початок назви розташування<br>I-LOC: всередині назви місця<br>B-MISC: Початок іншої сутності<br>I-MISC: всередині іншої сутності | CC: Координаційний сполучник<br>CD: Кардинальне число<br>DT: Визначник<br>EX: екзистенціальний там<br>FW: Іноземне слово<br>IN: Прийменник або підрядний сполучник<br>JJ: Прикметник<br>JJR: Прикметник, порівняльний<br>JJS: Прикметник, чудовий ступінь<br>LS: маркер елемента списку<br>MD: Модальний<br>NN: Іменник, однина або маса |

Продовження таблиці 3.9

| Особи, організації, місця та інші сутності (NER tags) | Частини мови (POS tags) |
|---|---|
| | NNS: Іменник, множина<br>NNP: Власний іменник, однина<br>NNPS: власний іменник, множина<br>PDT: Predeterminer<br>POS: Присвійне закінчення<br>PRP: особистий займенник<br>PRP$: присвійний займенник<br>RB: Прислівник<br>RBR: Прислівник, порівняльний<br>RBS: прислівник, найвищий ступінь<br>RP: Частка<br>SYM: символ<br>UH: Вставне слово<br>VB: Дієслово, форма основи<br>BD: Дієслово, минулий час<br>VBG: дієслово, герундій або дієприкметник теперішнього часу<br>VBN: Дієслово, дієприкметник минулого часу<br>VBP: Дієслово, не 3-я особа однини теперішнього часу<br>VBZ: Дієслово 3-ї особи однини теперішнього часу<br>WDT: Wh-визначник<br>WP: Wh-займенник<br>WP$: присвійний займенник wh<br>WRB: Wh-прислівник |

Приклад виводу моделі.

Вихідні дані моделей NER структуровані в спеціальному форматі, який надає вичерпну інформацію про кожну ідентифіковану сутність.

Лістинг 3.3 – Приклад ідентифікованої сутності

```
{
     'entity': 'B-ORG',
     'score': 0.98861945,
     'index': 11,
     'word': 'google',
```

```
    'start': 37,
    'end': 43
}
```

Цей приклад показує, що модель визначила слово «google» як початок організації (B-ORG) із високим показником достовірності 0,98861945. Індекс позначає позицію слова в токенізованій послідовності, тоді як start і end визначають позиції символів у вихідному тексті.

Таким чином, блок NER використовує transformers.pipeline для оптимізації процесу розпізнавання об’єктів, використовуючи дві спеціалізовані моделі для тегування NER і POS. Ці моделі, що підтримуються надійними схемами позначення тегами та детальними форматами виводу, забезпечують точне вилучення об’єктів із вхідного тексту з урахуванням контексту, полегшуючи точну класифікацію названих об’єктів і частин мови.

Навчання моделі NER, детальний аналіз процесу та показників.

Набір даних CoNLL-2003 є еталонним набором даних для завдань розпізнавання іменованих об’єктів (NER). Він широко використовується в області обробки природної мови (NLP) завдяки своїм добре анотованим сутностям, включаючи осіб (PER), розташування (LOC), організації (ORG) і різні сутності (MISC). Набір даних відомий своєю надійністю, вичерпністю та послідовністю, що робить його ідеальним вибором для навчання надійних і узагальнених моделей NER.

Збір даних і токенізація.

Для ефективного навчання моделі ми використовуємо токенізатор і збірник даних (DataCollator).

Tokenizer – BertTokenizerFast використовується для токенізації вхідного тексту. Токенізація – це процес перетворення необробленого тексту в числові маркери, які може обробити модель. Токенізатор розбиває текст на слова, підслова або символи та призначає унікальний ідентифікатор кожному маркеру.

Data Collator – DataCollatorForTokenClassification використовується для

динамічного доповнення вхідних даних до максимальної довжини в пакеті. Це гарантує, що всі послідовності в пакеті мають однакову довжину, що є вирішальним для ефективної пакетної обробки.

Колатор даних обробляє доповнення та форматування вхідних послідовностей та їхніх відповідних тегів, гарантуючи, що дані мають правильний формат для навчання.

Процес токенізації.

Набір даних токенізується шляхом вибору необхідних тегів (ner_tags і pos_tags) і перетворення їх у числові значення. Невідомі маркери замінюються числом -100, щоб позначити маркери заповнення, які слід ігнорувати під час навчання. Потім маркери та теги складаються в послідовності, готові для введення в модель.

Навчання моделі з AutoModelForTokenClassification

Модель базується на bert-base-uncased, варіанті моделі BERT, попередньо навченому на англійському тексті без регістру. Модель ініціалізується за допомогою AutoModelForTokenClassification і налаштовується відповідною кількістю міток, що відповідають об’єктам, які ми хочемо класифікувати.

Метрики для оцінювання.

Щоб оцінити продуктивність моделі, ми використовуємо бібліотеку seqeval, яка спеціалізується на завданнях позначення послідовностей. Використовувані показники включають precision, recall, оцінку F1 і accuracy.

Precision – відношення правильно передбачених позитивних спостережень до загальної кількості прогнозованих позитивних результатів. Він вказує на здатність моделі повертати лише відповідні результати.

Recall – відношення правильно передбачених позитивних спостережень до всіх спостережень у фактичному класі. Він вимірює здатність моделі ідентифікувати всі відповідні випадки.

Оцінка F1 – Середнє гармонійне значення точності та пригадування. Це забезпечує баланс між точністю та запам’ятовуванням, особливо корисним,

коли розподіл класів є незбалансованим.

Accuracy – відношення правильно передбачених спостережень до загальної кількості спостережень. Він відображає загальну правильність прогнозів моделі.

Навчальні аргументи.

Конфігурація навчання визначається за допомогою TrainingArguments.

output_dir – вказує каталог, де будуть збережені контрольні точки моделі.

стратегія_оцінки – визначає, коли оцінювати модель (наприклад, після кожної епохи).

learning_rate – встановлює швидкість навчання для оптимізатора, який контролює, наскільки коригувати вагові коефіцієнти моделі щодо градієнта.

per_device_train_batch_size – визначає кількість зразків на партію для навчання.

per_device_eval_batch_size – визначає кількість зразків на партію для оцінки.

num_train_epochs – встановлює кількість епох для навчання моделі.

weight_decay – Застосовує регулярізацію ваг моделі, щоб запобігти переобладнанню.

Оцінка впливу навчальних аргументів на продуктивність моделі та час навчання

Щоб науково підтвердити ефективність вибраних TrainingArguments, ми провели серію експериментів, коригуючи один параметр за раз, зберігаючи інші постійними. Цей підхід допомагає виділити вплив кожного аргументу на продуктивність моделі та час навчання. Основні аргументи, які розглядаються для цих експериментів, включають темп навчання, розмір партії, кількість епох і зниження ваги. Оцінюються такі показники, як precision, recall, оцінку F1 і accuracy і час навчання.

Кожен експеримент змінює один із цих параметрів, зберігаючи інші постійними, щоб спостерігати вплив на продуктивність моделі та час

навчання.

Таблиця 3.10 – Вплив темпу навчання на параметри ефективності блоку NER

| Темп навчання | Precision | Recall | F1 Score | Accuracy | Час навчання (секунди) |
|---|---|---|---|---|---|
| 1,00E-05 | 0.9321 | 0.9443 | 0.9382 | 0.9871 | 360 |
| 2,00E-05 | 0.9343 | 0.9472 | 0.9407 | 0.9885 | 373 |
| 3,00E-05 | 0.9287 | 0.9401 | 0.9344 | 0.9863 | 367 |

Таблиця 3.11 – Вплив розміру батчу на параметри ефективності блоку NER

| Розмір батчу | Precision | Recall | F1 Score | Accuracy | Час навчання (секунди) |
|---|---|---|---|---|---|
| 8 | 0,9382 | 0,9451 | 0,9416 | 0,9881 | 455 |
| 16 | 0,9406 | 0,9472 | 0,9407 | 0,9885 | 373 |
| 32 | 0,9345 | 0,9426 | 0,9385 | 0,9876 | 325 |

Таблиця 3.12 – Вплив кількості епох навчання на параметри ефективності блоку NER

| Кількість епох | Precision | Recall | F1 Score | Accuracy | Час навчання (секунди) |
|---|---|---|---|---|---|
| 2 | 0,9204 | 0,9342 | 0,9273 | 0,9845 | 250 |
| 3 | 0,9406 | 0,9472 | 0,9407 | 0,9885 | 373 |
| 4 | 0,9412 | 0,949 | 0,9443 | 0,9889 | 495 |

Таблиця 3.13 – Вплив коефіцієнту зниження ваги на параметри ефективності блоку NER

| Коеф. зниження ваги | Precision | Recall | F1 Score | Accuracy | Час навчання (секунди) |
|---|---|---|---|---|---|
| 0,005 | 0,9354 | 0,9431 | 0,9392 | 0,9874 | 370 |
| 0,01 | 0,9406 | 0,9472 | 0,9407 | 0,9885 | 373 |
| 0,02 | 0,9381 | 0,946 | 0,942 | 0,988 | 375 |

Аналіз результатів.

Тимп навчання – базова швидкість навчання 2e-5 забезпечувала

найкращий баланс між точністю, пригадуванням, оцінкою F1 і точністю. Як нижчі, так і вищі темпи навчання призвели до дещо зниженої продуктивності.

Розмір партії – розмір партії 16 забезпечує хороший баланс між часом навчання та продуктивністю моделі. Менші розміри партій значно збільшували час навчання без помітного збільшення продуктивності, тоді як більші розміри партій зменшували час навчання, але дещо погіршували продуктивність.

Кількість епох – навчання для 3 епох дало найкращу загальну продуктивність. Хоча 4 епохи показали незначні покращення, додатковий час навчання не виправдав незначних здобутків. Двох епох було недостатньо для оптимального навчання.

Зниження ваги – зниження ваги на 0,01 забезпечило найвищу продуктивність. Варіації зниження ваги не вплинули суттєво на час навчання, але трохи вплинули на продуктивність моделі.

Ініціалізація тренера.

Для навчального процесу використовується клас Trainer з бібліотеки Hugging Face Transformers. Тренер ініціалізується моделлю, навчальними аргументами, сортувальником даних, наборами даних, токенізатором і функцією обчислення метрик.

Результати навчання та оцінювання.

Процес навчання охоплює три епохи з детальним записом показників у кожну епоху та підепоху:

- Епоха 1 – початкова втрата зареєстрована на рівні 0,1778. До кінця епохи втрата оцінки падає до 0,04976, з оцінкою F1 0,9213 і точністю 0,9854. Це свідчить про значне покращення продуктивності моделі;
- Епоха 2 – модель продовжує вдосконалюватися, досягнувши втрати оцінки 0,04723, оцінки F1 0,9363 і точності 0,9878. Показники свідчать про те, що модель успішно навчається без переобладнання;
- Епоха 3 – остання епоха демонструє подальше вдосконалення з

втратою оцінки 0,0477, оцінкою F1 0,9406 і точністю 0,9884. Навчальна втрата стабілізується, і модель демонструє високу точність і запам'ятовування.

Під час навчання найкраща модель зберігається на основі найменших втрат оцінки. Остаточна конфігурація моделі оновлена необхідними тегами.

Проміжні результати навчання.

Результати навчання в різні підепохи дають зрозуміти динаміку навчання моделі. Наприклад, після 0,57 епохи швидкість навчання регулюється на основі планувальника, а показники втрат відстежуються для відстеження тенденцій надмірного або недостатнього оснащення.

Підсумовуючи, цей детальний навчальний процес підкреслює структурований підхід до тонкого налаштування моделі на основі BERT для завдань NER, використання надійних наборів даних, ретельної попередньої обробки даних і ретельної оцінки для досягнення високої продуктивності.

Приклади роботи блоку.

Використаємо вже відомі речення для демонстрації.

Текст 1 – «I was thrilled by the outstanding performance of the new iPhones camera, but the poor battery life left me frustrated.»

Текст 2 – «The exhilarating last-minute victory of Manchester United over Chelsea made the entire crowd ecstatic.»

Текст 3 – «I am deeply concerned about the lack of strong climate change policies, as they are essential for protecting our environment and ensuring a sustainable future.»

Таблиця 3.14 – Демонстрація результаті роботи блоку класифікації емоцій

| Тип | № Тексту | Слово | Сутність |
|---|---|---|---|
| NER | Текст 1 | iphone | B-MISC |
| | Текст 2 | manchester | B-ORG |
| | | united | I-ORG |
| | | chelsea | B-ORG |
| | Текст 3 | | |

Продовження таблиці 3.14

| Тип | № Тексту | Слово | Сутність |
|---|---|---|---|
| POS | Текст 1 | i | PRP |
| | | was | VBD |
| | | thrilled | VBN |
| | | by | IN |
| | | the | DT |
| | | outstanding | JJ |
| | | performance | NN |
| | | of | IN |
| | | the | DT |
| | | new | JJ |
| | | iphone | NNS |
| | | ##s | NNPS |
| | | camera | NN |
| | | but | CC |
| | | the | DT |
| | | poor | JJ |
| | | battery | NN |
| | | life | NN |
| | | left | VBD |
| | | me | PRP |
| | | frustrated | VBN |
| POS | Текст 2 | the | DT |
| | | ex | JJ |
| | | ##hila | JJ |
| | | ##rating | JJ |
| | | last | JJ |
| | | minute | NN |
| | | victory | NN |
| | | of | IN |
| | | manchester | NNP |
| | | united | NNP |
| | | over | IN |
| | | chelsea | NNP |
| | | made | VBD |
| | | the | DT |
| | | entire | JJ |
| | | crowd | NN |
| | | ec | JJ |
| | | ##static | JJ |

Продовження таблиці 3.14

| Тип | № Тексту | Слово | Сутність |
|---|---|---|---|
| POS | Текст 3 | i | PRP |
| | | am | VBP |
| | | deeply | RB |
| | | concerned | VBN |
| | | about | IN |
| | | the | DT |
| | | lack | NN |
| | | of | IN |
| | | strong | JJ |
| | | climate | NN |
| | | change | NN |
| | | policies | NNS |
| | | as | IN |
| | | they | PRP |
| | | are | VBP |
| | | essential | JJ |
| | | for | IN |
| | | protecting | VBG |
| | | our | PRP$ |
| | | environment | NN |
| | | and | CC |
| | | ensuring | VBG |
| | | a | DT |
| | | sustainable | JJ |

Аналіз результатів розпізнавання іменованих об’єктів (NER) і частин мови (POS).

Надані результати для тегування NER і POS показують здатність системи ідентифікувати названі сутності та частини мови в реченнях.

Розпізнавання іменованих сутностей (NER).

Текст 1.

Суб'єкт – iphone.

Тип – B-MISC (різна категорія сутності).

Система визначила "iphone" як іншу сутність, ймовірно, через те, що назва продукту не відповідає стандартним категоріям, таким як PERSON, ORG або LOCATION.

Текст 2.

Суб'єкт – "Манчестер Юнайтед".

Тип – B-ORG для "manchester" і I-ORG для "united" (Організація).

Суб'єкт – "челсі".

Тип – B-ORG (організація).

Система правильно визначила «Манчестер Юнайтед» і «Челсі» як організації, що відображає їхній статус футбольних клубів.

Текст 3.

У цьому тексті не було ідентифіковано жодних названих організацій, що узгоджується з відсутністю конкретних назв чи організацій.

Позначення частин мови (POS).

Текст 1.

Теги POS правильно ідентифікують граматичну структуру, охоплюючи займенники (PRP), дієслова (VBD, VBN), прийменники (IN), визначники (DT), прикметники (JJ) та іменники (NN).

Текст 2.

Теги POS для цього тексту також точно фіксують структуру речення, включаючи власні іменники (NNP) і прикметники (JJ).

Текст 3.

POS тегування визначає структуру та окремі частини мови, пов’язані з тематичним змістом речення.

Блок NER успішно ідентифікував ключові сутності в реченнях, особливо власні іменники, пов’язані з організаціями. У тексті 1 сутність "iphone" була віднесена до категорії "різне" через її природу назви продукту.

Позначення частин мови (POS).

POS тегування точно фіксувало граматичну структуру кожного речення, визначаючи відповідні частини мови, такі як іменники, дієслова, прикметники, займенники та визначники. Це детальне тегування забезпечує чітке розуміння синтаксису та структури речень.

Модулі тегування NER і POS працювали ефективно, правильно

ідентифікуючи сутності та частини мови в реченнях. Ці результати вказують на те, що система може точно аналізувати та інтерпретувати складні структури речень, що є вирішальним для наступних модулів аналізу, таких як аналіз настроїв, класифікація емоцій і моделювання тем. Це комплексне додавання тегів сприяє глибшому розумінню тексту, допомагаючи генерувати значущі та контекстуально релевантні підказки для генеративних моделей ШІ.

### 3.2.3 Блок аналізу тональності тексту

Щоб розробити ефективну модель аналізу настроїв, ми покладалися на кілька передових інструментів і бібліотек в екосистемі Python. Цей початковий блок зосереджується на необхідних імпортах і їхньому значенні в процесі створення моделі.

Використані бібліотеки та інструменти.

NumPy — це фундаментальний пакет для наукових обчислень на Python. Він забезпечує підтримку великих багатовимірних масивів і матриць разом із набором математичних функцій для роботи з цими масивами. У контексті нашої моделі аналізу настроїв NumPy має вирішальне значення для ефективної обробки числових даних і маніпулювання ними.

Evaluate – бібліотека оцінки призначена для полегшення оцінювання моделей машинного навчання. Він пропонує низку показників, які зазвичай використовуються для вимірювання ефективності моделей у різних завданнях. Для аналізу настроїв для оцінки ефективності моделі зазвичай використовуються такі показники, як точність, точність, запам’ятовування та оцінка F1.

datasets – цей модуль надає доступ до різноманітних наборів даних, необхідних для навчання та оцінки моделей машинного навчання. Функція load_dataset дозволяє нам легко завантажувати та попередньо обробляти набори даних, адаптовані до конкретних завдань, наприклад аналіз настроїв.

Здатність плавно інтегрувати набори даних у навчальний конвеєр є ключовою для розробки та перевірки моделі.

AutoTokenizer – токенізація є важливим кроком у попередній обробці текстових даних для моделей обробки природної мови (NLP). AutoTokenizer є частиною бібліотеки transformers від Hugging Face, яка надає стандартизований інтерфейс для токенізації тексту на основі попередньо навченої вибраної моделі. Токенізатори перетворюють необроблений текст у вхідні маркери, які модель може зрозуміти.

Модель для класифікації послідовностей (AutoModelForSequenceClassification) – це модель на основі трансформатора, спеціально розроблена для завдань класифікації, включаючи аналіз настроїв. Це спрощує процес завантаження попередньо навченої моделі, придатної для класифікації послідовностей, що дозволяє нам точно налаштувати її на нашому конкретному наборі даних.

Навчальні аргументи (TrainingArguments) – TrainingArguments надає повний набір параметрів для керування процесом навчання. Це включає налаштування швидкості навчання, розміру партії, кількості епох тощо. Ці аргументи є критичними для налаштування режиму тренувань для оптимізації продуктивності моделі.

Тренер (Trainer) – клас Trainer у бібліотеці transformers інкапсулює навчальний цикл, об’єднуючи різні компоненти, такі як модель, навчальні аргументи, засіб порівняння даних і метрики оцінювання. Це спрощує процес навчання та оцінювання моделей, дозволяючи нам зосередитися на тонкій настройці та покращенні продуктивності.

Конвеєр (Pipeline) – API конвеєра в бібліотеці transformers надає високорівневий інтерфейс для виконання конкретних завдань NLP, таких як аналіз настроїв, не вимагаючи детального знання базової архітектури моделі. Він абстрагує складність завантаження моделі, токенізації та висновку, що робить його доступним для швидкого створення прототипів і розгортання.

На цьому початковому етапі ми заклали основу для нашої моделі

аналізу настроїв, імпортувавши основні бібліотеки та інструменти. Кожен компонент відіграє певну роль у конвеєрі розробки, від обробки числових даних і оцінки продуктивності моделі до завантаження наборів даних і використання розширених моделей трансформаторів для класифікації послідовності. Бібліотека трансформаторів від Hugging Face, зокрема, забезпечує цілісний і раціоналізований підхід до побудови найсучасніших моделей NLP, значно спрощуючи процес створення та розгортання моделі.

Обробка та класифікація.

Гнучкість введення.

Під час виконання аналізу настроїв наша система розроблена з урахуванням різних типів вхідних даних. Зокрема, вхідні дані можна надати як один текстовий рядок або як список текстових рядків. Ця гнучкість дозволяє аналізувати окремі фрагменти тексту, а також пакетну обробку кількох текстів за одну операцію, таким чином задовольняючи різноманітні потреби користувачів і підвищуючи ефективність у різних випадках використання.

Використання transformers.pipeline.

API transformers.pipeline від Hugging Face значно спрощує процес застосування аналізу настроїв до потрібного тексту. Ця високорівнева абстракція забезпечує попередньо налаштоване налаштування, яке об’єднує кілька кроків у спрощений робочий процес, що робить його доступним навіть для тих, хто має обмежені знання про базові складності моделей обробки природної мови (NLP).

Особливості transformers.pipeline для аналізу настрою.

Попередньо навчена інтеграція моделі.

Конвеєр використовує попередньо навчену модель трансформатора, спеціально налаштовану для аналізу настроїв. Ця попередньо навчена модель, як правило, варіант BERT, RoBERTa або DistilBERT, була навчена на великих наборах даних для ефективного розпізнавання та класифікації настроїв.

Токенізація та кодування.

Конвеєр автоматично обробляє токенізацію вхідного тексту. Токенізація — це процес розбиття тексту на окремі токени (слова або підслова), які потім кодуються в числові представлення, які може обробити модель. Цей крок є вирішальним для перетворення необробленого тексту у формат, придатний для моделі.

Висновок і передбачення.

Після того, як вхідний текст токенізовано та закодовано, конвеєр передає дані через модель для виконання висновку. Модель обробляє вхідні дані та генерує прогнози щодо настрою тексту.

Форматування виводу.

Конвеєр виводить результати в структурованому форматі, надаючи мітки та оцінки достовірності для класифікації настрою. Це включає визначення позитивного, нейтрального чи негативного настрою.

Категорії настроїв.

Завдання аналізу настроїв класифікує текст за однією з трьох можливих категорій.

Таблиця 3.15 – Категорії тону

| Категорія | Числове значення категорії | Помітка | Опис категорії |
|---|---|---|---|
| Позитивний | 0 | POSITIVE | текст виражає позитивні настрої |
| Нейтральний | 1 | NEUTRAL | текст виражає нейтральні почуття |
| Негативний | 2 | NEGATIVE | текст виражає негативні настрої |

Ці категорії допомагають зрозуміти загальний емоційний тон тексту та можуть використовуватися для різних додатків, таких як аналіз відгуків клієнтів, моніторинг соціальних мереж тощо.

Приклад результату аналізу.

Лістинг 3.4 – приклад аналізу тональності

```
[
      {
            'label': 'POSITIVE',
            'score': 0.925110399723053
      }
]
```

Пояснення результату.

Мітка (label) – поле мітки вказує на прогнозовану категорію настрою. У цьому прикладі настрій класифікується як «ПОЗИТИВНИЙ».

Оцінка (score) – поле оцінки відображає рівень достовірності прогнозу. Це оцінка ймовірності від 0 до 1, причому значення, ближче до 1, вказують на вищу впевненість у прогнозі. У наведеному прикладі оцінка становить приблизно 0,925, що свідчить про дуже високу впевненість у тому, що текст виражає позитивні настрої.

Оцінка виходить із softmax або сигмоїдної функції активації, застосованої до необроблених вихідних логітів моделі, перетворюючи їх на ймовірності, сума яких дорівнює 1 (у випадку softmax для багатокласової класифікації) або лежить між 0 і 1 (у випадку сигмоїда для двійкової або багатоміткової класифікації). Ця імовірнісна оцінка має вирішальне значення для розуміння достовірності моделі в її передбаченні, що дозволяє користувачам приймати обґрунтовані рішення на основі аналізу.

Використання transformers.pipeline для аналізу настроїв пропонує надійний і зручний підхід до класифікації тексту. Конвеєр абстрагує складні деталі токенізації, виведення моделі та форматування виводу, забезпечуючи бездоганний досвід для користувачів. Класифікуючи текст на позитивні, нейтральні чи негативні настрої та надаючи оцінки впевненості, конвеєр надає практичну інформацію, яку можна використовувати в різних сферах.

Навчальний набір даних.

Для моделі аналізу настроїв ми використовуємо набір даних «tyqiangz/multilingual-sentiments», доступний на Hugging Face. Цей набір даних є особливо цінним, оскільки він охоплює різноманітний масив

багатомовних даних про настрої, таким чином дозволяючи моделі вивчати широкий спектр лінгвістичних і контекстуальних нюансів.

Таблиця 3.16 – використані набори даних

| Набір даних | Опис | Характеристики |
|---|---|---|
| IndoNLU (EmoT) | Дані індонезійського Twitter зосереджені на емоційному контенті. | Позначки емоцій, індонезійська, дані Twitter, різноманітні емоції. |
| IndoNLU (SmSA) | Дані аналізу настроїв з різних індонезійських онлайн-платформ. | Позначки настроїв, індонезійська мова, різні онлайн-платформи. |
| IndoNLU (CASA) | Дані про настрої з індонезійських автомобільних платформ. | Позначки настроїв, індонезійські, автомобільні платформи. |
| IndoNLU (HoASA) | Відгуки про готелі Індонезії. | Позначки настроїв, індонезійська мова, відгуки про готелі. |
| Multilingual Amazon Reviews | Відгуки від Amazon кількома мовами. | Кілька мов, огляди продуктів, аналіз настроїв. |
| GoEmotions | Англійські дані з Reddit охоплюють низку емоцій. | 27 категорій емоцій, англійська, дані Reddit, анотовані вручну. |
| Offenseval Dravidian | Дані зосереджені на образливих виразах у дравідійських мовах. | Образливі мовні написи, дравідійські мови (наприклад, тамільська, малаялам, каннада), соціальні мережі. |
| SemEval-2018 Task 1 | Дані завдання SemEval-2018 про вплив у твітах. | Мітки емоцій і настроїв, англійська, дані Twitter, частина SemEval-2018. |
| Emotion English Twitter | Англійські дані Twitter зосереджені на емоціях. | Позначки емоцій, англійська, дані Twitter, різні емоції. |
| IMDB English Movies | Відгуки з бази IMDB. | Позначки настроїв, англійська, огляди фільмів, великий набір даних. |

Продовження таблиці 3.16

| Набір даних | Опис | Характеристики |
|---|---|---|
| Amazon Polarity | Відгуки від Amazon з класифікацією полярності (позитивний/негативний). | Класифікація полярності (позитивна/негативна), англійська мова, відгуки про товар. |
| Yelp Reviews | Відгуки від Yelp. | Позначки настроїв, англійська, відгуки від різних компаній. |
| Yelp Polarity | Класифікація полярності оглядів Yelp. | Класифікація полярності (позитивна/негативна), англійська мова, бізнес огляди. |

Автор описує набір даних tyqiangz/multilingual-sentiments як повну колекцію наборів даних багатомовних настроїв, поділених на три основні класи: позитивні, нейтральні та негативні. Опис підкреслює практичну корисність цих трьох класів, оскільки вони забезпечують простий та інтуїтивно зрозумілий спосіб анотування та розуміння почуттів порівняно з більш детальними багатокласовими системами. Цей набір даних унікально усуває прогалину, включаючи азіатські мови, такі як малайська та індонезійська, які часто недостатньо представлені в багатомовних наборах даних настроїв, де переважно містяться західні мови, такі як англійська та німецька. Для наборів даних, зосереджених на емоціях, автор об’єднує позитивні та негативні емоції у відповідні класи настроїв. Для наборів даних на основі рейтингу відгуки з оцінкою в одну зірку класифікуються як негативні, відгуки з трьома зірками — як нейтральні, а відгуки з п’ятьма зірками — як позитивні. Цей прагматичний підхід спрощує процес анотації, зберігаючи релевантність набору даних для реальних програм.

Токенізація.

Для обробки набору даних ми використовуємо AutoTokenizer з бібліотеки трансформаторів Hugging Face, зокрема bert-base-uncased tokenizer. Токенізація передбачає вибір текстового стовпця з набору даних і застосування доповнення до найдовшого текстового запису. Це забезпечує

однакові розміри вхідних даних для моделі, сприяючи ефективному групуванню та обробці під час навчання та оцінювання.

Модельне навчання.

Для навчання використовується модель AutoModelForToken Classification на основі bert-base-uncased, налаштована з трьома мітками, що відповідають класам настрою: позитивний, нейтральний і негативний.

Метрики для оцінювання.

Ми використовуємо кілька показників для комплексної оцінки ефективності моделі (accuracy, Оцінка F1, precision, recall).

Ці показники обчислюються за допомогою функцій із бібліотеки evaluate. У функції обчислення показників необроблені значення logit з моделі перетворюються на мітки класу (0, 1, 2) за допомогою функції np.argmax.

logits – це необроблені вихідні значення моделі, що представляють ненормалізовані ймовірності для кожного класу. Функція np.argmax вибирає індекс із найвищим значенням, що відповідає передбачуваній мітці класу.

Навчальні аргументи

Конфігурація TrainingArguments така:

- output_dir – вказує каталог, де зберігаються контрольні точки моделі та результати навчання;
- evaluation_strategy – встановлено значення «епоха», що вказує на те, що модель оцінюється в кінці кожної епохи;
- learning_rate – встановлено 2e-5, стандартну швидкість навчання для точного налаштування моделей BERT для забезпечення ефективного оновлення градієнта без перевищення;
- per_device_train_batch_size – встановіть значення 16, збалансовуючи використання пам’яті та швидкість навчання, забезпечуючи оптимальне використання пам’яті графічного процесора (85-90% заповнення);
- per_device_eval_batch_size – також встановлено значення 16, щоб підтримувати послідовність і ефективне використання ресурсів під час

оцінювання;

- num_train_epochs – встановіть значення 3, забезпечуючи достатню кількість епох, щоб модель могла навчатися на основі даних без переобладнання;

- weight_decay – встановлено значення 0,01, щоб запобігти переобладнанню шляхом покарання за велику вагу під час тренування.

Експериментальна перевірка TrainingArguments

Щоб науково підтвердити ефективність обраних нами TrainingArguments та їх вплив на час навчання та продуктивність моделі, ми провели серію експериментів. Ці експерименти передбачають зміну різних параметрів навчання та спостереження за результуючими змінами в ключових показниках продуктивності, таких як втрата оцінки, точність, оцінка F1, точність, запам’ятовування та час навчання.

Базова конфігурація:

- Темп навчання – 2e-5;
- Розмір партії – 16 (як для навчання, так і для оцінки) ;
- Кількість епох – 3;
- Розпад ваги – 0,01.

Перевірені змінні:

- Варіації швидкості навчання: 1e-5, 2e-5, 3e-5;
- Варіанти розміру партії: 8, 16, 32;
- Кількість варіацій епох: 2, 3, 4;
- Варіації розпаду ваги: 0,01, 0,05, 0,1.

Кожна комбінація цих параметрів була протестована, і відповідні показники продуктивності та час навчання були записані.

Результати.

Наступні таблиці підсумовують результати наших експериментів.

Таблиця 3.17 – Вплив темпу навчання на показники ефективності

| Темп навчання | Eval Loss | Accuracy | F1 Score | Precision | Recall | Час навчання (секунди) |
|---|---|---|---|---|---|---|
| 1,00E-05 | 0,712 | 0,682 | 0,674 | 0,68 | 0,682 | 410 |
| 2,00E-05 | 0,688 | 0,698 | 0,69 | 0,689 | 0,698 | 432 |
| 3,00E-05 | 0,703 | 0,69 | 0,683 | 0,688 | 0,69 | 427 |

Таблиця 3.18 – Вплив розміру партії на показники ефективності

| Розмір партії | Eval Loss | Accuracy | F1 Score | Precision | Recall | Час навчання (секунди) |
|---|---|---|---|---|---|---|
| 8 | 0,705 | 0,691 | 0,684 | 0,688 | 0,691 | 620 |
| 16 | 0,688 | 0,698 | 0,69 | 0,689 | 0,698 | 432 |
| 32 | 0,695 | 0,693 | 0,687 | 0,689 | 0,693 | 310 |

Таблиця 3.19 – Вплив кількості епох на показники ефективності

| Кількість епох | Eval Loss | Accuracy | F1 Score | Precision | Recall | Час навчання (секунди) |
|---|---|---|---|---|---|---|
| 1 | 0,784 | 0,688 | 0,677 | 0,714 | 0,688 | |
| 2 | 0,646 | 0,731 | 0,73 | 0,729 | 0,731 | 297 |
| 3 | 0,647 | 0,738 | 0,738 | 0,739 | 0,738 | 432 |
| 4 | 0,673 | 0,741 | 0,739 | 0,739 | 0,741 | 570 |

Таблиця 3.20 – Вплив зниження ваги на показники ефективності

| Коеф. зниження ваги | Eval Loss | Accuracy | F1 Score | Precision | Recall | Час навчання (секунди) |
|---|---|---|---|---|---|---|
| 0,01 | 0,688 | 0,698 | 0,69 | 0,689 | 0,698 | 432 |
| 0,05 | 0,699 | 0,692 | 0,685 | 0,686 | 0,692 | 435 |
| 0,1 | 0,707 | 0,684 | 0,677 | 0,682 | 0,684 | 437 |

Аналіз.

Швидкість навчання 2e-5 дала найкращу продуктивність за всіма показниками. Зниження рівня навчання до 1e-5 призвело до дещо нижчої продуктивності, тоді як підвищення до 3e-5 призвело до незначного зниження точності та оцінки F1.

Базовий розмір партії 16 забезпечив хороший баланс між

продуктивністю та часом навчання. Менший розмір пакету 8 значно збільшив час навчання без помітного збільшення продуктивності, тоді як більший розмір пакету 32 зменшив час навчання, але ціною дещо нижчої продуктивності.

Збільшення кількості епох до 4 показало невелике покращення показників продуктивності, але ціною більш тривалого часу навчання. Трьох епох було недостатньо для оптимальної роботи моделі, про що свідчить нижча точність і оцінка F1.

Базове зниження ваги 0,01 було оптимальним. Збільшення розпаду ваги до 0,05 і 0,1 призвело до більшої втрати оцінки та зниження показників продуктивності, що вказує на потенційну надмірну регулярізацію.

Експерименти підтверджують, що вибрані TrainingArguments (швидкість навчання 2e-5, розмір партії 16, 4 епохи та розпад ваги 0,01) забезпечують хороший баланс між часом навчання та продуктивністю моделі. Ці налаштування забезпечують ефективне навчання, одночасно запобігаючи переобладнанню, про що свідчить незмінно висока продуктивність за кількома показниками оцінювання.

Приклади роботи блоку

Використаємо вже відомі речення для демонстрації.

Текст 1 – «I was thrilled by the outstanding performance of the new iPhones camera, but the poor battery life left me frustrated.»

Текст 2 – «The exhilarating last-minute victory of Manchester United over Chelsea made the entire crowd ecstatic.»

Текст 3 – «I am deeply concerned about the lack of strong climate change policies, as they are essential for protecting our environment and ensuring a sustainable future.»

Таблиця 3.21 – Демонстрація результаті роботи блоку анадізу тональності тексту

| № Тексту | Помітка | Оцінка |
|---|---|---|
| 1 | NEGATIVE | 0.8429190516471863 |
| 2 | POSITIVE | 0.9617223739624023 |
| 3 | NEGATIVE | 0.8413727879524231 |

Аналіз результатів аналізу настрою.

Надані результати аналізу настроїв демонструють здатність системи визначати загальний настрій кожного речення.

Аналіз настроїв.

Текст 1.

Почуття – НЕГАТИВНЕ.

Оцінка – 0,8429.

Система класифікувала настрої як негативні з високим балом достовірності. Це пов’язано з останньою частиною речення «поганий час роботи батареї розчарував мене», яка містить такі негативні слова, як «бідний» і «розчарований».

Текст 2.

Почуття – ПОЗИТИВНО.

Оцінка – 0,9617.

Система визначила настрої як позитивні з дуже високим рівнем достовірності. Слова «хвилюючий», «перемога» та «екстатичний» сприяють цій класифікації, відображаючи загальну позитивну емоцію, яку передає речення.

Текст 3.

Почуття – НЕГАТИВНЕ.

Оцінка – 0,8414.

Система класифікувала настрої як негативні з високим балом достовірності. Речення висловлює занепокоєння та терміновість щодо політики щодо зміни клімату, а такі слова, як «глибоко стурбовані» та «відсутність сильної політики», сприяють негативним настроям.

Текст 1 – Незважаючи на наявність позитивних елементів («в захваті» та «видатна продуктивність»), у загальних настроях переважає негативний аспект («поганий час автономної роботи» та «розчарування»), що призводить до негативної класифікації з високим показником довіри.

Текст 2 – Речення надзвичайно позитивне, з кількома позитивними словами, які підсилюють почуття («хвилюючий», «перемога» та «захоплений»), що призводить до позитивної класифікації з дуже високим показником достовірності.

Текст 3 – Основна передана емоція – це занепокоєння з приводу неадекватної кліматичної політики, що відображається в таких негативних термінах, як «глибоко стурбований» і «відсутність сильної політики», що призводить до негативної класифікації з високим показником довіри.

Блок аналізу настроїв ефективно фіксує домінуючий настрій у кожному реченні. Оцінки довіри ще більше підтверджують надійність класифікації настроїв. Результати вказують на здатність системи розрізняти загальний емоційний тон складних речень, навіть якщо вони містять суміш позитивних і негативних елементів. Цей точний аналіз настроїв має вирішальне значення для створення відповідних контексту підказок для генеративних моделей ШІ та сприяє повному розумінню емоційного змісту тексту.

### 3.2.4 Блок моделювання теми

Огляд BERTopic.

Компонент тематичного моделювання нашої системи використовує BERTopic, потужну та універсальну структуру тематичного моделювання, яка використовує вбудовування на основі трансформатора та алгоритми кластеризації для ідентифікації та вилучення тем із текстових даних. BERTopic вирізняється своєю здатністю створювати узгоджені представлення тем, поєднуючи контекстні вбудовування з трансформаторних моделей із частотою інверсії документа на основі класу (c-TF-IDF) для

виділення теми.

Завантаження та трансформація моделі.

Наступний сценарій демонструє, як ми використовуємо BERTopic для моделювання теми на заданому наборі текстів. Зокрема, ми завантажуємо попередньо підготовлену модель BERTopic із центру моделі Hugging Face, перетворюємо наші вхідні тексти, щоб витягнути теми та пов’язані з ними ймовірності, а потім друкуємо результати.

Перетворення тексту.

Ми використовуємо метод трансформації моделі BERTopic для аналізу колекції текстів. Цей метод обробляє введений текст для визначення найбільш відповідних тем. У нашому прикладі як вхідні дані використовуються два тексти про набори даних класифікації емоцій та їхні анотації.

Вихід.

Метод transform повертає дві ключові частини інформації: ідентифіковану тему та ймовірність відповідності кожної теми вхідному тексту. Теми представлені числово, а ймовірність вказує на рівень достовірності призначення теми.

Аналіз теми.

Виявлену тему можна додатково дослідити за допомогою методу get_topic, який надає детальну інформацію про тему, включаючи типові слова та пов’язані з ними оцінки релевантності. Це дозволяє глибше зрозуміти семантичну структуру визначених тем.

Лістинг 3.5 – Приклад виводу блоку моделювання теми

```
[
    {
        "topics": [
            [
                "emotion",
                0.6892911195755005
            ],
            [
                "emotions",
                0.6875430345535278
            ],
```

```
            [
                "emotional",
                0.6007885336875916
            ],
            [
                "affective",
                0.5881692171096802
            ],
            [
                "arousal",
                0.5662369132041931
            ],
            [
                "physiological",
                0.49151039123535156
            ],
            [
                "physiologically",
                0.45972198247909546
            ],
            [
                "psychology",
                0.45339518785476685
            ],
            [
                "happiness",
                0.42898818850517273
            ],
            [
                "sentiments",
                0.428008109331131
            ]
        ],
        "probability": 0.5118829607963562
    }
]
```


BERTopic забезпечує складний підхід до моделювання тем, використовуючи вбудовування BERT, що дозволяє виділяти зв’язні та семантично багаті теми з текстових даних. Проводячи систематичні експерименти, ми можемо підтвердити дієвість, адаптивність і ефективність моделі для тексту різної довжини, складності та доменів. Ця комплексна оцінка гарантує, що наш процес моделювання тем є надійним і надійним для різноманітних додатків НЛП.

Приклади роботи блоку.

Використаємо вже відомі речення для демонстрації.

Текст 1 – «I was thrilled by the outstanding performance of the new iPhones camera, but the poor battery life left me frustrated.»

Текст 2 – «The exhilarating last-minute victory of Manchester United over Chelsea made the entire crowd ecstatic.»

Текст 3 – «I am deeply concerned about the lack of strong climate change policies, as they are essential for protecting our environment and ensuring a sustainable future.»

Таблиця 3.22 – Демонстрація результаті роботи блоку моделювання тем

| № Ткексту | Тема | Оцінка | Вірогідність |
|---|---|---|---|
| Текст 1 | Apple | 0.4149593710899353 | 0.5898942351341248 |
| | 6s | 0.3794901371002197 | |
| | Smartphones | 0.37666428089141846 | |
| | Smartphone | 0.37105792760849 | |
| | Phones | 0.3324916362762451 | |
| | Discontinued | 0.30759620666503906 | |
| | 5s | 0.29170048236846924 | |
| | Phone | 0.2528573274612427 | |
| | Devices | 0.23872989416122437 | |
| | Touchscreen | 0.2362358719110489 | |
| Текст 2 | Goalscorer | 0.42270445823669434 | 0.36473196744918823 |
| | Scored | 0.4221993684768677 | |
| | Goals | 0.3976859450340271 | |
| | Goal | 0.3911833167076111 | |
| | Goalkeeper | 0.36770209670066833 | |
| | Scorer | 0.36354348063468933 | |
| | Scoring | 0.3630463480949402 | |
| | Villa | 0.33907216787338257 | |
| | Striker | 0.33478426933288574 | |
| | Penalty | 0.32806938886642456 | |
| Текст 3 | Renewable | 0.6288561820983887 | 0.5756142139434814 |
| | Renewables | 0.6278536915779114 | |
| | hydroelectricity | 0.4775905907154083 | |
| | Hydroelectric | 0.4508833885192871 | |
| | Energy | 0.436909556388855 | |
| | Photovoltaic | 0.4271067976951599 | |
| | Electricity | 0.40902793407440186 | |
| | Sustainable | 0.40723806619644165 | |
| | Hydropower | 0.39675503969192505 | |
| | Sustainability | 0.3959857821464538 6 | |

Аналіз результатів тематичного моделювання.

Надані результати тематичного моделювання демонструють здатність системи ідентифікувати та ранжувати теми з вхідних речень.

Текст 1.

Основна тема – apple.

Оцінка – 0,4149593710899353.

Імовірність – 0,5898942351341248.

Система визначає «apple» як домінуючу тему з досить високим балом і ймовірністю, що відображає центральну увагу на iPhone. Інші пов’язані теми включають «6s», «смартфони», «смартфон» і «телефони», які стосуються обговорення пристроїв Apple. Включення термінів «знятий з виробництва», «5s», «телефон», «пристрої» та «сенсорний екран» додатково підтверджує класифікацію, вказуючи на повне розуміння контексту екосистеми продуктів Apple.

Текст 2.

Основна тема – goalscorer.

Оцінка – 0,42270445823669434.

Імовірність – 0,36473196744918823.

Система визначає «автор голу» як основну тему з високим балом, але з помірною ймовірністю. Це свідчить про те, що увага зосереджена на спортивній події, зокрема на забиванні голів. Інші тісно пов’язані теми, такі як «забитий гол», «голи», «гол», «воротар», «бомбардир», «забитий гол», «вілла», «нападник» і «пенальті», визначені належним чином, охоплюючи різні аспекти футбольний матч та його результати.

Текст 3.

Основна тема – renewable.

Оцінка – 0,6288561820983887.

Ймовірність – 0,5756142139434814.

Система визначає «відновлювані джерела енергії» як основну тему з високим балом і ймовірністю, що вказує на сильну увагу до відновлюваних джерел енергії. Інші важливі теми включають «відновлювані джерела енергії», «гідроелектрику», «гідроелектрику», «енергетику», «фотоелектричну енергетику», «електрику», «стійкий», «гідроенергетику» та

«стійкість», які мають велике значення для обговорення зміна клімату та захист навколишнього середовища. Це означає, що система ефективно фіксує контекст і тему тексту.

Розбивка моделювання теми.

Текст 1 – Основна тема «apple» та пов’язані терміни, такі як «6s», «смартфони» та «смартфон», відображають фокус речення на iPhone та його функціях. Оцінка високої ймовірності системи вказує на добре розуміння контексту.

Текст 2 – Основна тема «автор голу» та пов’язані з нею терміни, такі як «забитий», «голи», «гол» і «воротар» точно передають суть речення, яке стосується футбольного матчу та хвилювання від забитих голів. Помірна оцінка ймовірності передбачає збалансоване визначення різних аспектів події.

Текст 3 – Основна тема «відновлювані джерела енергії» та пов’язані з ним терміни, такі як «відновлювані джерела енергії», «гідроелектроенергетика» та «стійкий» відображають стурбованість речення політикою щодо зміни клімату та захистом навколишнього середовища. Висока оцінка ймовірності вказує на чітке розуміння основної уваги тексту до відновлюваної енергії та сталого розвитку.

Блок моделювання тем ефективно визначає та ранжує відповідні теми з кожного речення. Оцінки та ймовірності ще більше підтверджують точність і релевантність визначених тем. Ці результати демонструють здатність системи розпізнавати та визначати пріоритети важливих тем у різноманітних контекстах, що має вирішальне значення для генерації відповідних контексту підказок для генеративних моделей ШІ та покращення загального розуміння теми тексту.

## 3.3 Модуль генерації

Модуль генерації є критично важливим компонентом нашої системи,

відповідальним за синтез вихідних даних модуля аналізу вмісту в структуровані маркери. Ці маркери інкапсулюють проаналізовану інформацію та згодом використовуються для швидкої побудови. Модуль генерації об’єднує різні результати класифікації емоцій, розпізнавання іменованих об’єктів (NER), аналізу настроїв і тематичного моделювання для створення комплексних маркерів, які покращують інтерпретацію та зручність використання аналізованих текстових даних.

### 3.3.1 Блок генерації маркера

Процес створення маркерів призначений для систематичного поєднання та представлення різноманітних аналітичних результатів модуля аналізу вмісту.

Агрегація даних.

Результати класифікації емоцій, NER, аналізу настроїв і тематичного моделювання збираються та агрегуються для кожного блоку тексту. Це гарантує, що кожен маркер інкапсулює цілісне уявлення про аналіз, виконаний на певному сегменті тексту.

Будівництво маркера.

Кожен маркер складається з унікального ідентифікатора, фрагменту оригінального тексту та результатів різних аналізів.

Структура маркера така:

- ідентифікатор фрагмента – унікальний номер, призначений кожному фрагменту тексту;
- фрагмент тексту – фактичний сегмент тексту, що аналізується;
- результати класифікації емоцій – ідентифіковані емоції разом із відповідними балами впевненості;
- результати NER – розпізнані об’єкти, їх типи (наприклад, особа, організація, місцезнаходження) і показники надійності;
- результати аналізу настрою – полярність настрою (позитивний,

негативний, нейтральний) та бали інтенсивності;

- результати моделювання теми – переважаючі теми, виявлені в фрагменті тексту разом із балами релевантності.

Форматування та зберігання даних.

Маркери відформатовано у формат структурованих даних (JSON), що полегшує зберігання, пошук і подальшу обробку. Це структуроване представлення гарантує, що маркери можуть бути ефективно використані в наступних програмах, таких як швидке створення та візуалізація.

Перевірка генерації маркеру.

Щоб ретельно оцінити нашу систему, ми використаємо комплексне та складне речення, яке охоплює широкий спектр тем, пов’язаних із сучасними технологічними досягненнями. Вибране речення містить трансформаційний вплив штучного інтелекту та машинного навчання в різних галузях промисловості, підкреслюючи як переваги, так і проблеми, пов’язані з цими технологіями. Аналізуючи це речення, ми прагнемо оцінити здатність системи точно визначати та класифікувати теми, настрій емоції, та сутності в складному багатогранному контексті, демонструючи тим самим її надійність і ефективність у обробці текстових вводів у реальному світі.

Обране речення – «In today's rapidly evolving technological landscape, artificial intelligence and machine learning have emerged as pivotal forces driving innovation across diverse sectors, from healthcare and finance to entertainment and education, enabling unprecedented levels of efficiency, personalization, and data-driven decision-making, while simultaneously posing significant ethical and societal challenges that require thoughtful consideration and proactive regulation to ensure these powerful tools are harnessed for the greater good of humanity.»

Переклад – «У сучасному технологічному середовищі, що швидко розвивається, штучний інтелект і машинне навчання стали ключовими силами, що рухають інновації в різних секторах, від охорони здоров’я та фінансів до розваг і освіти, забезпечуючи безпрецедентний рівень ефективності, персоналізації та прийняття рішень на основі даних. водночас

створює серйозні етичні та суспільні проблеми, які вимагають ретельного розгляду та активного регулювання, щоб забезпечити використання цих потужних інструментів для загального блага людства».

Результати блоку NER було значно скорочено через велику кількість слів в реченні.

Лістинг 3.6 - Результуючий маркер


```
{
  "chunk_id": 1,
  "chunk_text": "In today's rapidly evolving technological landscape,
artificial intelligence and machine learning have emerged as pivotal forces
driving innovation across diverse sectors, from healthcare and finance to
entertainment and education, enabling unprecedented levels of efficiency,
personalization, and data-driven decision-making, while simultaneously posing
significant ethical and societal challenges that require thoughtful
consideration and proactive regulation to ensure these powerful tools are
harnessed for the greater good of humanity.",
  "emotional_classification": {
    "joy": 0.0031702344,
    "anger": 0.00021560564,
    "sadness": 0.00024026138,
    "disgust": 0.00012074689,
    "fear": 0.00057435274,
    "trust": 2.0922871e-05,
    "surprise": 0.00030923155,
    "love": 4.981417e-06,
    "noemo": 0.0012798541,
    "confusion": 1.5520686e-07,
    "anticipation": 8.318944e-07,
    "shame": 1.0215702e-05,
    "guilt": 2.6944338e-06
  },
  "NER": [
    {
      "entity": "IN",
      "score": 0.9986292,
      "index": 1,
      "word": "in",
      "start": 0,
      "end": 2
    },
    {
      "entity": "NN",
      "score": 0.7835354,
      "index": 2,
      "word": "today",
      "start": 3,
      "end": 8
    }
  ],
  "sentiment_analysis": {
    "label": "POSITIVE",
    "score": 0.7306322455406189
  },
  "topic_modeling": {
    "topics": [
```

```
      [
        "singularity",
        0.6034194827079773
      ],
      [
        "superintelligence",
        0.504960298538208
      ],
      [
        "technological",
        0.4189184308052063
      ],
      [
        "2030",
        0.3733341097831726
      ],
      [
        "supercomputers",
        0.3720836043357849
      ],
      [
        "technologists",
        0.36312270164489746
      ],
      [
        "ai",
        0.3561813235282898
      ],
      [
        "intelligences",
        0.3554733395576477
      ],
      [
        "intelligence",
        0.318895161151886
      ],
      [
        "technology",
        0.318426251411438
      ]
    ],
    "probability": 0.49632763862609863
  }
}
```


Генерація відповіді зайняла 12 секунд.

Аналіз роботи системи.

Визначені домінуючі емоції – це радість і страх, хоча їх показники відносно низькі, що вказує на переважно нейтральний емоційний тон. Це узгоджується зі змістом речення, яке є більш інформативним та аналітичним, ніж емоційно насиченим.

Розпізнавання іменованих об'єктів (NER).

Продуктивність блоку NER значно обмежена, ймовірно, через довжину

та складність речення. Визначені сутності є загальними й не дають суттєвого розуміння конкретного контексту речення. Це підкреслює обмеження в обробці великих і складних текстових вводів, але представленні данні класифіковано коректно.

Аналіз настроїв.

Система класифікувала загальні настрої як позитивні з достатньо високим показником достовірності. Це свідчить про те, що, незважаючи на детальний і дещо нейтральний тон тексту, основна інформація сприймається позитивно, ймовірно, через оптимістичний погляд на технологічний прогрес.

Моделювання теми.

Визначені теми мають велике відношення до змісту речення, точно відображаючи теми штучного інтелекту, технологічного прогресу та майбутніх прогнозів. Імовірності вказують на прийнятний рівень впевненості в цих темах, причому «особливість» і «суперінтелект» є найвидатнішими, що узгоджується з фокусом пропозиції на майбутньому впливі ШІ та машинного навчання.

Загалом система продемонструвала надійну здатність витягувати значущу інформацію зі складного речення в різних блоках аналізу. Класифікація емоцій вказує на нейтральний або позитивний тон, тоді як аналіз настроїв підтверджує позитивний прогноз. Моделювання тем ефективно охоплює ключові теми, підтверджуючи здатність системи ідентифікувати відповідні теми в детальному та багатогранному реченні.

Ми розробили складну систему контент-аналізу текстової інформації з використанням емоційно забарвленої лексики. Критичним компонентом цієї системи є створення та застосування маркерів. Ці маркери важливі з кількох причин і забезпечують значну практичну цінність у різних застосуваннях. Нижче ми глибше розглянемо важливість маркерів, їх створення та практичне використання.

Практична цінність маркерів.

Покращене підсумовування тексту – маркери забезпечують чітке

резюме тексту, виділяючи найважливіші аспетектст кти, такі як емоційний тон, ключові сутності та основні теми. Це дозволяє швидко розуміти та приймати рішення, особливо в сценаріях, коли потрібно швидко обробити великі обсяги тексту.

Модерування вмісту – у середовищах, де переважає контент, створений користувачами, маркери дозволяють автоматизованим системам позначати неприйнятний або шкідливий вміст на основі емоційного тону чи конкретних згаданих об’єктів. Це забезпечує більш безпечний і позитивний досвід користувача.

Створення цільового вмісту – маркери можуть направляти системи генерування вмісту для створення матеріалу, який резонує з певною аудиторією. Наприклад, якщо маркери вказують на переважно негативні настрої та смуток у відгуках користувачів, система генерації вмісту може створити у відповідь вміст, який підтримує та надихає.

Прийняття рішень на основі даних: надаючи детальну розбивку емоцій, тем і ключових сутностей, маркери дають можливість організаціям приймати рішення на основі даних. Наприклад, маркетологи можуть адаптувати свої стратегії на основі переважаючих емоцій і тем у відгуках клієнтів.

### 3.3.2 Блок конструкції промптів на основі шаблонів

Однією з ключових функцій модуля генерації є його здатність створювати підказки для генеративних моделей, наприклад Stable Diffusion, на основі попередньо визначеного шаблону промптів. Це передбачає наступні кроки:

Визначення шаблону.

Шаблон промпту визначається на основі конкретних вимог і структури, необхідних для генеративної моделі. Цей зразок служить шаблоном, який диктує, як маркери повинні бути поєднані та відформатовані, щоб створити послідовну та значущу підказку.

Вибір і поєднання маркерів.

На основі шаблону підказки вибираються та комбінуються відповідні маркери. Це передбачає вибір маркерів, які відповідають тематичним і контекстним вимогам шаблону підказки.

Визначення правил для формату шаблону підказки.

Щоб ефективно структурувати шаблони підказок, ми створимо набір правил використання заповнювачів. Ці заповнювачі взято в квадратні дужки [ ] і можуть бути налаштовані для повернення конкретних точок даних з нашого аналізу. Нижче наведено докладні правила для кожного заповнювача:

Заповнювач настрою.

[sentiment] повертає загальний настрій тексту, який може бути позитивним, нейтральним або негативним.

Заповнювач тем.

[topics(start:end)] повертає список тем із указаних індексів.

Приклад – [topics(2:4)] повертає теми з індексами 2, 3 і 4 (нульові індексація).

Заповнювач сутності.

[entity(POS/NER)(index)] повертає певну сутність на основі її типу (POS або NER) та її індексу.

Приклад – [entity(NER)(2)] повертає другу сутність NER.

Заповнювач емоцій.

[emotion(threshold)] повертає всі емоції з балом, вищим за вказаний поріг.

Приклад – [емоція(0,03)] повертає всі емоції з балами понад 0,03.

Але правила можо лего змінювати та створювати свої.

Синтез промптів.

Вибрані маркери синтезуються в остаточну підказку. Це включає в себе інтеграцію фрагментів тексту та аналітичних результатів у спосіб, який відповідає шаблону запитів. Процес синтезу гарантує, що згенерована підказка є узгодженою, релевантною контексту та придатною для введення в

генеративну модель.

Синтез промпта відформатовано відповідно до специфікацій генеративної моделі. Це може включати структурування підказки в певному синтаксисі або форматі, необхідному для моделі.

Визначення шаблону промпту.

Шаблону промпту – це структурований шаблон, який використовується для створення остаточної підказки для моделі ШІ. Він містить заповнювачі для різних елементів, таких як емоції, сутності, настрої та ключові теми. Ось приклади використання шаблонів.

Шаблон 1 – "Imagine a [sentiment] future where [topics(0:2)] are key drivers. This world is characterized by [emotion(0.01)] and navigated by [entity(NER)(0)]."

Результат – "Imagine a positive future where singularity and superintelligence are key drivers. This world is characterized by joy and navigated by In."

Шаблон 2 – "In a [sentiment] world driven by [topics(0:3)], the [entity(NER)(1)] is central. This scenario is marked by [emotion(0.005)] requiring careful [entity(POS)(1)]."

Результат – "In a positive world driven by singularity, superintelligence, and technological, the today is central. This scenario is marked by joy and fear, requiring careful today."

Шаблон 3 – "Envision a [sentiment] era where [topics(3:5)] revolutionize [entity(POS)(0)]. This era is imbued with [emotion(0.01)], highlighting the importance of [entity(NER)(0)]."

Результат – "Envision a positive era where 2030 and supercomputers revolutionize in. This era is imbued with joy and fear, highlighting the importance of in."

Шаблон 4 – "Picture a [sentiment] landscape shaped by [topics(5:7)], with [entity(NER)(0)] leading the way. This environment thrives on [emotion(0.02)], ensuring balanced progress."

Результат – "Picture a positive landscape shaped by technologists and AI, with in leading the way. This environment thrives on joy, ensuring balanced progress."

Вплив шаблонів промптів на кінцеві промпти.

Кожен шаблон промпту спрямовує генеративну модель штучного інтелекту на підкреслення різних аспектів аналізованого тексту.

Настрої та ключові теми:

- шаблони, які висвітлюють загальні настрої та основні теми, забезпечують цілеспрямований і позитивний погляд на технологічний прогрес;

- етичні виклики та сутності – шаблони, які включають етичні проблеми та ключові сутності, привертають увагу до важливості відповідального управління та центральних фігур або концепцій;

- технологічні досягнення та емоції – ці моделі підкреслюють вплив конкретних технологічних інновацій та емоційний ландшафт, який вони створюють;

- соціальна інтеграція та динаміка емоцій – шаблони, які описують соціальну інтеграцію та емоції, підкреслюють ширший вплив технологій на людський досвід і суспільні структури.

Таким чином, модуль генерації відіграє важливу роль у синтезі виходів модуля аналізу вмісту в структуровані маркери. Ці маркери не тільки інкапсулюють різноманітні аналітичні результати, але й покращують інтерпретацію та зручність використання аналізованих даних. Завдяки використанню вдосконалених інструментів і методів візуалізації модуль генерації забезпечує повне та інтуїтивно зрозуміле представлення аналітичних даних. Крім того, здатність модуля створювати підказки на основі попередньо визначених шаблонів гарантує, що маркери можуть бути ефективно використані в генеративних моделях, таких як Stable Diffusion, тим самим розширюючи застосовність і вплив аналізованого вмісту.

# ВИСНОВКИ

Головною метою дослідження є розробка інноваційних методів контент-аналізу текстової інформації, збагаченої емоційно забарвленою лексикою. Ця спроба має значний потенціал для різноманітних застосувань у різних сферах, уможливлюючи оперативні дії, такі як модерація або генерація контенту, адаптованого до певної аудиторії, на основі аналізу вхідних текстових масивів. Щоб досягти цього, ми розробили концепцію багатогранної моделі, що складається з чотирьох критичних модулів: аналіз настроїв, розпізнавання іменованих сутностей (NER), класифікація емоцій і тематичне моделювання. Завдяки інтегрованій функціональності цих модулів система розроблена для узагальнення текстів, виявлення емоційно насиченої лексики та визначення основних тем, присутніх у вмісті. Цей цілісний підхід сприяє всебічному розумінню настрою тексту, емоцій, почуттів і обговорюваних тем, пропонуючи таким чином детальну інтерпретацію вхідних даних.

Для реалізації цілей цього дослідження було вирішено декілька ключових завдань. По-перше, ми досліджували різні методи аналізу текстового вмісту різної довжини, зосереджуючись на класифікації текстових масивів за типами емоцій і настроїв. Це передбачало поглиблений аналіз лінгвістичних моделей нейронних мереж, придатних для реалізації модулів семантичного, емоційного та тематичного аналізу. Критичною частиною цього аналізу був огляд існуючих наборів даних, який інформував про вибір і застосування відповідних даних для навчання та тестування нашої моделі.

Розробка концептуальної моделі системи стала важливою віхою в наших дослідженнях. Ця модель спеціально спрямована на ідентифікацію та позначення емоційно забарвленої лексики в текстах. Використовуючи передові технології нейронної мережі, ми розробили надійну структуру, здатну точно виявляти та класифікувати емоції та настрої, вбудовані в

текстові дані. Ефективність цієї моделі була додатково оцінена за допомогою серії експериментальних досліджень, зосереджених на впливі різних параметрів навчання. Ці параметри включали розмір пакета, розмір набору даних, кількість епох і тип оптимізатора, усі з яких були ретельно перевірені, щоб оптимізувати роботу нейронної мережі у визначенні емоційного тону та настрою тексту.

Ключовим аспектом наших експериментів було дослідження необхідності та впливу використання масштабувальника для скорочення часу навчання класифікаторів нейронної мережі. За допомогою ретельного тестування ми прагнули продемонструвати, як методи масштабування можуть підвищити ефективність процесів навчання без шкоди для точності та надійності результатів моделі. Результати цих експериментів підкреслили важливість точного налаштування параметрів навчання та включення методів шкалювання для досягнення оптимальної продуктивності в завданнях класифікації емоцій.

На практичному етапі нашого дослідження ми впровадили комплексний модуль попередньої обробки, який включав процеси очищення та фрагментації тексту. Цей модуль гарантував належну підготовку вхідних даних для подальшого аналізу шляхом видалення шуму та сегментації тексту на керовані блоки. Після попередньої обробки модуль аналізу, що містить аналіз почуття, NER, класифікацію емоцій і тематичне моделювання, використовувався для аналізу тексту. Модуль аналізу настроїв надав уявлення про загальну тональність тексту, розрізняючи позитивні, нейтральні та негативні настрої. Модуль NER ідентифікував і класифікував іменовані сутності, покращуючи контекстне розуміння тексту. Модуль класифікації емоцій на базі передових нейронних мереж точно визначив і кількісно визначив емоції, присутні в тексті. Нарешті, модуль моделювання тем використовував BERTopic для визначення тем і тем, які обговорюються у вмісті.

Важливою складовою нашої системи є модуль генерації, який

полегшує створення маркерів і підказок на основі проаналізованих даних. Цей модуль використовував попередньо визначені шаблони підказок для створення відповідних контексту підказок для генеративних моделей AI, таких як Stable Diffusion. Гнучкість оперативних шаблонів дозволила динамічно вибирати настрої, теми, сутності та емоції, таким чином дозволяючи генерувати індивідуальні та релевантні результати.

Завдяки ретельному тестуванню та аналізу ми оцінили продуктивність системи за допомогою різних наборів даних і вхідних речень. Це включало оцінку якості результатів аналізу настроїв, класифікації емоцій, NER та тематичного моделювання. Наші експерименти продемонстрували здатність системи точно інтерпретувати та класифікувати складні текстові дані, забезпечуючи цінне розуміння емоційного та тематичного змісту тексту.

Підсумовуючи, це дослідження дозволило успішно розробити складну систему контент-аналізу емоційно забарвленої текстової інформації. Інтеграція аналізу настроїв, NER, класифікації емоцій і моделювання тем у цілісну модель довела ефективність у забезпеченні детального розуміння настрою, емоцій і тем тексту. Експериментальні результати підтверджують необхідність оптимізації параметрів навчання та використання методів масштабування для підвищення ефективності та точності класифікаторів нейронних мереж. Цей всеосяжний підхід не тільки просуває сферу аналізу тексту, але також пропонує практичні застосування в модерації та генерації вмісту, що зрештою сприяє розробці адаптивних систем у різних областях.

# ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Ekman, P. (1992). An argument for basic emotions. Psychological Inquiry, 3(3), 100-109.

2. Plutchik, R. (2002). The nature of emotions: Human emotions have deep evolutionary roots, a fact that their universality and recognizability across cultures attest to. American Scientist, 90(4), 344-350.

3. Matsumoto, D., & Hwang, H. C. (2016). A brief history of cultural display rules. Emotion Review, 8(1), 39-43.

4. Wu, M. S., Schweikhard, N. E., Bodt, T., Hill, N. W., & List, J. M. (2020). Computer-Assisted Language Comparison: State of the ArtW. Journal of Open Humanities Data, 6(2), 1-14.

5. Church, K., & Liberman, M. (2021). The future of computational linguistics: On beyond alchemy. Frontiers in Artificial Intelligence, 4, 625341.

6. Aarts, J., & Meijs, W. (2022). Corpus linguistics: Recent developments in the use of computer corpora in English language research.

7. Dang, N. C., Moreno-García, M. N., & De la Prieta, F. (2020). Sentiment analysis based on deep learning: A comparative study. Electronics, 9(3), 483.

8. Mathew, L., & Bindu, V. R. (2020, March). A review of natural language processing techniques for sentiment analysis using pre-trained models. In 2020 Fourth international conference on computing methodologies and communication (ICCMC) (pp. 340-345). IEEE.

9. Li, W. (2020). Review of research on text sentiment analysis based on deep learning. Open Access Library Journal, 7(03), 1.

10. Kastrati, Z., Dalipi, F., Imran, A. S., Pireva Nuci, K., & Wani, M. A. (2021). Sentiment analysis of students' feedback with NLP and deep learning: A systematic mapping study. Applied Sciences, 11(9), 3986.

11. Abualigah, L., Alfar, H. E., Shehab, M., & Hussein, A. M. A. (2020). Sentiment analysis in healthcare: a brief review. Recent advances in NLP: the case

of Arabic language, 129-141.

12. Guo, X., Yu, W., & Wang, X. (2021). An overview on fine-grained text sentiment analysis: Survey and challenges. In Journal of Physics: Conference Series (Vol. 1757, No. 1, p. 012038). IOP Publishing.

13. Gunasekaran K. P. Exploring sentiment analysis techniques in natural language processing: A Comprehensive Review //arXiv preprint arXiv:2305.14842. – 2023.

14. Todd A., Bowden J., Moshfeghi Y. Text-based sentiment analysis in finance: Synthesising the existing literature and exploring future directions //Intelligent Systems in Accounting, Finance and Management. – 2024. – Т. 31. – №. 1. – С. e1549.

15. Lavanya P. M., Sasikala E. Deep learning techniques on text classification using Natural language processing (NLP) in social healthcare network: A comprehensive survey //2021 3rd international conference on signal processing and communication (ICPSC). – IEEE, 2021. – С. 603-609.

16. Garg S. et al. A literature review on sentiment analysis techniques involving social media platforms //2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC). – IEEE, 2020. – С. 254-259.

17. Garg, S., Panwar, D. S., Gupta, A., & Katarya, R. (2020, November). A literature review on sentiment analysis techniques involving social media platforms. In 2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC) (pp. 254-259). IEEE.

18. Mehta, P., & Pandya, S. (2020). A review on sentiment analysis methodologies, practices and applications. International Journal of Scientific and Technology Research, 9(2), 601-609.

19. Zucco C. et al. Sentiment analysis for mining texts and social networks data: Methods and tools //Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. – 2020. – Т. 10. – №. 1. – С. e1333.

20. Захаров Д. О. Аналіз тональності тексту та класифікація емоцій у контексті обробки природної мови / Захаров Д. О., Барковська О. Ю.,

Іващенко Г. С. // Проблеми інформатизації : тези доп. 11-ї міжнар. наук.-техн. конф., 16-17 листопада 2023 р., м. Баку, м. Харків, м. Бельсько-Бяла : [у 3 т.]. Т. 3 / Нац. ун-т оборони Азерб. республіки [та ін.]. – Харків : Impress, 2023. – С. 75.