

Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук
(повна назва)

Кафедра Штучного інтелекту
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)

Емоційна нейтралізація текстів з збереженням семантики
(тема)

Виконав:
студент 2 курсу, групи СШМ-22-1
Білоконь Б.О.
(прізвище, ініціали)

Спеціальність 122 Комп'ютерні науки
(код і повна назва спеціальності)

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

Освітня програма Системи штучного інтелекту
(повна назва спеціалізації)

Керівник проф. Терзіян В.Я.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри _____
(підпис)

В.О. Філатов
(прізвище, ініціали)

2024 р.

Харківський національний університет радіоелектроніки

Факультет _____ Комп'ютерних наук _____
(повна назва)
Кафедра _____ Штучного інтелекту _____
(повна назва)
Рівень вищої освіти _____ другий (магістерський) _____
Спеціальність _____ 122 Комп'ютерні науки _____
(код і повна назва)
Тип програми _____ освітньо-наукова _____
(освітньо-професійна або освітньо-наукова)
Освітня програма _____ Системи штучного інтелекту _____
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
« _____ » _____ 20 ____ р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

студентові _____ Білоконю Богдану Олексійовичу _____
(прізвище, ім'я, по батькові)

1. Тема роботи _____ Емоційна нейтралізація текстів з збереженням семантики _____

затверджена наказом університету від 1 квітня 20 24 р. № 260Ст

2. Термін подання студентом роботи до екзаменаційної комісії 11 червня 20 24 р.

3. Вихідні дані до роботи _____ документація по мові програмування Python, згенеровані дані для експерименту _____

4. Перелік питань, що потрібно опрацювати в роботі _____

1) Аналіз предметної галузі та постановка задачі

2) Теоретичні дослідження

3) Програмна реалізація

4) Експериментальні дослідження

РЕФЕРАТ

Пояснювальна записка: 74 с., 33 рис., 7 табл., 1 дод., 26 джерел.

ВЕЛИКА МОВНА МОДЕЛЬ, ЕМОЦІЯ, МАШИННЕ НАВЧАННЯ,
НЕЙТРАЛІЗАЦІЯ, ПРОМПТ, СЕМАНТИКА, ТРАНСФОРМЕР.

Об'єкт дослідження – процес нейтралізації емоцій в тексті зі збереженням семантики.

Предмет дослідження – модель GPT.

Мета роботи – аналіз здатності моделі GPT нейтралізувати емоції в тексті зі збереженням семантики та вплив різних типів шаблонів підказок для моделі.

Методи дослідження – методи генерації, аналізу і візуалізації даних, методи prompt-engineering, методи порівняння текстів за семантикою.

У результаті проведених досліджень вирішено задачу аналізу здатності моделі GPT нейтралізувати емоції у тексті зі збереженням семантики, експериментально доведено ефективність використання методів prompt-engineering.

Отримані результати можуть використовуватись у сферах, де важлива розважливість, стриманість, безпристратність, таких як юриспруденція, журналістика, політика, військова справа тощо.

ABSTRACT

Master's thesis contains: 74 pp., 33 fig., 7 tabl., 1 ann., 26 references.

EMOTION, LARGE LANGUAGE MODEL, MACHINE LEARNING,
NEUTRALIZATION, PROMPT, SEMANTICS, TRANSFORMER.

The object of research is the process of neutralizing emotions in a text while preserving semantics.

The subject of the study is the GPT model.

Purpose – to analyze the ability of the GPT model to neutralize emotions in the text while preserving semantics and the impact of different types of prompt templates for the model.

Research methods – methods of data generation, analysis and visualization, prompt-engineering methods, methods of comparing texts by semantics.

As a result of the conducted research, the task of analyzing the ability of the GPT model to neutralize emotions in the text while preserving semantics was solved, and the effectiveness of using prompt-engineering methods was experimentally proved.

The obtained results can be used in areas where prudence, restraint, and impartiality are important, such as law, journalism, politics, military affairs, etc.

ЗМІСТ

Перелік умовних позначень, символів, одиниць, скорочень і термінів	8
Вступ	9
1 Аналіз предметної галузі	11
1.1 Аналіз проблеми	11
1.2 Огляд існуючих методів вирішення проблеми	14
1.2.1 Сумаризація тексту	14
1.2.2 Великі мовні моделі	16
1.3 Постановка задачі	18
2 Теоретичні дослідження	20
2.1 Теорія базових емоцій	20
2.2 Аналіз емоційного забарвлення в тексті	23
2.2.1 Лексичні методи	23
2.2.2 Статистичні методи	25
2.2.3 Методи машинного навчання	27
2.2.4 Трансформери	29
2.3 Prompt-engineering	34
2.3.1 Persona pattern	35
2.3.2 Audience persona pattern	37
2.3.3 Few-shot examples pattern	38
2.4 Проблема порівняння текстів	40
2.4.1 Косинус подібності	41
2.4.2 Word2Vec	43
2.4.3 Sentence-BERT	45
3 Експериментальні дослідження	47
3.1 План виконання експерименту	47
3.2 Підготовка даних	47
3.3 Рівень емоційності текстів для експерименту	48
3.4 Нейтралізація емоцій	51

3.5 Рівень емоційності текстів, отриманих у ході експерименту	56
3.5.1 Нейтралізація гніву.....	56
3.5.2 Нейтралізація радості	58
3.5.3 Нейтралізація смутку	60
3.5.4 Нейтралізація страху	62
3.5.5 Нейтралізація здивування	63
3.5.6 Нейтралізація огиди	65
3.6 Оцінка семантичної подібності даних до і після нейтралізації	66
Висновки.....	68
Перелік джерел посилання	70
Додаток А Відомість кваліфікаційної роботи	74

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

МН – машинне навчання;

ШІ – штучний інтелект;

BERT – Bidirectional Encoder Representations from Transformers –
двонаправлені кодувальні представлення з трансформерів;

GPT – Generative Pre-trained Transformer – генеративний попередньо
навчений трансформатор;

LLM – Large Language Model – велика мовна модель;

NLP – Natural Language Processing – обробка природної мови.

ВСТУП

Роль емоцій у виступах і повідомленнях є важливою. Емоції виступають як ключовий елемент у взаємодії між спікером і аудиторією. Їхня емоційна відкритість може створити атмосферу взаєморозуміння та співпраці. Відтак, важливою стає не лише сама змістовна частина виступу, але й тон та емоційна інтонація. Часто люди пам'ятають не лише те, що було сказано, а й те, як це було сказано. Таким чином, емоційний струмінь може стати мостом між ідеями спікера та їхнім сприйняттям аудиторією, роблячи виступ більш ефективним та запам'ятовуваним.

Проте коли в тексті або виступі присутні негативні емоції, це може мати широкий спектр негативних наслідків. По-перше, така емоційна насиченість може призвести до зміни тону і стилю висловлювання, що вплине на сприйняття аудиторією. Наприклад, якщо спікер або автор тексту перебуває в стані гніву або роздратування, його висловлювання може стати агресивним або образливим, що відверне аудиторію від сприйняття ідей. Негативні емоції можуть призвести до неясності та неоднозначності в повідомленні. Коли спікер або автор тексту переживає стрес або тривогу, його думки можуть бути заплутані або невиразні, що зробить їх менш зрозумілими для аудиторії. Це може призвести до втрати чіткості та логічної послідовності в експозиції ідей. Крім того, негативні емоції можуть викликати втрату довіри та авторитету спікера чи автора тексту. Аудиторія може менше довіряти інформації, яку отримує від особи, що проявляє негативність. Це може призвести до зменшення впливу та ефективності комунікації.

Отже, емоційний стан спікера чи автора тексту грає важливу роль у сприйнятті повідомлення. Контроль над емоційним фоном може допомогти забезпечити чіткість, довіру та ефективність комунікації з аудиторією.

Машинне навчання – великий підрозділ штучного інтелекту, що вивчає методи побудови алгоритмів, здатних навчатися. Машинне

навчання – це процес, під час перебігу якого система опрацьовує велику кількість прикладів, виявляє закономірності і використовує їх, щоб прогнозувати вихідні характеристики для нових вхідних даних [1]. Мета машинного навчання – частково або навіть повністю автоматизувати вирішення різноманітних складних аналітичних завдань.

Машинне навчання відкриває широкі можливості і для нейтралізації емоцій у тексті. Наприклад, популярними зараз є великі мовні моделі – LLM. LLM можна використовувати для аналізу та класифікації емоцій у тексті. За допомогою методів сентиментного аналізу, модель може визначати емоційний тон кожного речення або тексту та виконувати відповідні дії для нейтралізації, наприклад, виділяти емоційно заряджені фрази для подальшої обробки.

1 АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ

1.1 Аналіз проблеми

Емоції відіграють критичну роль у всіх аспектах комунікації, від щирості спілкування до ефективного передавання повідомлень. Вони впливають на сприйняття, реакції та рішення учасників комунікації.

Емоції можуть впливати на наші рішення та дії. Вони можуть підсилювати або приглушувати раціональне мислення та впливати на наші уподобання та вибори. Наприклад, люди можуть приймати рішення на основі своїх емоційних реакцій, навіть якщо ці реакції не відповідають раціональному аналізу ситуації.

Емоції можуть бути потужним інструментом впливу на інших людей. Говоріння з емоційним навантаженням може переконати, надихнути або мотивувати аудиторію до дії.

Емоційно-нейтральна комунікація є важливою в різних областях та ситуаціях, де об'єктивність, безпристрасність або відсутність емоційного упередження є критичними. Наприклад:

- обслуговування клієнтів: забезпечення того, що спілкування з клієнтами залишається нейтральним, допомагає підтримувати професійний та безпристрасний підхід, навіть в складних ситуаціях;

- юридичні та судові процеси: юридичні документи, судові засідання та офіційна комунікація часто потребують емоційної нейтральності для забезпечення справедливості та безпристрасності;

- охорона здоров'я та медичні записи: медичні записи, звіти та комунікація між медичними працівниками користуються емоційною нейтральністю для збереження професіоналізму та безпристрасності;

- журналістика та звітність: новини та журналістське письмо спрямовані на об'єктивність і безпристрасність, вимагаючи емоційно-нейтральної комунікації для представлення інформації без упередження;

– наука: наукові статті, статті та академічна комунікація повинні бути емоційно-нейтральними для збереження вірогідності та об’єктивності;

– корпоративна комунікація: звіти, офіційні накази та корпоративна комунікація часто прагнуть до емоційної нейтральності для об’єктивного та професійного викладу інформації;

– кадрова робота: комунікація, пов’язана з управлінням персоналом, така як оцінки результатів і дисциплінарні заходи, вигідна від емоційної нейтральності для забезпечення справедливості та послідовності;

– навчальні матеріали: освітній контент, такий як підручники та навчальні матеріали, часто спрямований на емоційну нейтральність для об’єктивного та безупинного представлення інформації;

– військова справа: у військових операціях важливо зберігати секретність і передавати інформацію об’єктивно, без емоційного упередження. Застосування емоційно-нейтральної комунікації у плануванні операцій, звітності та командуванні може сприяти точному розумінню ситуації та прийняттю об’єктивних рішень. У випадках взаємодії з громадськістю або дипломатичними інстанціями, важливо подавати інформацію нейтрально та без емоційного впливу. Це може допомагати зберігати стабільність та міжнародний спокій [2].

Наприклад, у 2017 році Дональд Трамп, 45-ий президент США, вимовив під час свого виступу: «When you see these thugs being thrown into the back of a paddy wagon, you just seen them thrown in, rough. I said, "Please don't be too nice"» (переклад з англійської мови – «Коли ви бачите, як цих хуліганів кидають назад у в’язницю, ви бачите, як їх кидають, грубо. Я сказав: "Будьте, будь ласка, не такими добрими".»). Ця фраза викликала громадське обурення.

Ще одним прикладом може бути виступ Гіларі Клінтон. Вона висловилася з приводу прихильників Дональда Трампа під час її передвиборчої кампанії у 2016 році, що стало одним із найбільш обговорюваних моментів виборчих перегонів. Цей коментар було зроблено

під час заходу зі збору коштів у Нью-Йорку 9 вересня 2016 року. Клінтон намагалася охарактеризувати дві різні групи прихильників Трампа, але її слова були сприйняті як зайве узагальнення та образа. Вона сказала: «To just be grossly generalistic, you can put half of Trump's supporters into what I call the "basket of deplorables". Unfortunately there are people like that, and he has lifted them up» (переклад з англійської мови – «Якщо узагальнити, то половину прихильників Трампа можна віднести до того, що я називаю "кошиком жалюгідних". На жаль, такі люди є, і він підняв їх нагору»). Ця заява викликала широкий резонанс у суспільстві та медіа. Критики та прихильники Трампа використали її слова, щоб зобразити Клінтон як відчужену від звичайних американців та надто готову до стигматизації великої частини електорату. Хоча Клінтон згодом висловила жаль з приводу своєї формулювання, стверджуючи, що не варто було говорити "половина", ця ситуація залишилась однією з ключових у її виборчій кампанії і, за деякими думками, могла негативно вплинути на її кандидатуру.

Такі приклади можна зустріти не тільки в сфері політики, а й, наприклад, журналістики. Яскравим прикладом можна назвати випадок, що стався у 2021 році в ефірі програми «Good Morning Britain» з одним із ведучих П'єром Морганом. П'єр викликав значне обурення своїми коментарями після інтерв'ю Меган Маркл та принца Гаррі з Опрою Вінфрі, що було випущено в березні 2021 року. В інтерв'ю Меган поділилася своїми переживаннями щодо психічного здоров'я та зазначила, що вона мала думки про самогубство під час перебування в королівській родині. На наступний день після трансляції інтерв'ю ведучий в ефірі сказав «I'm sorry, I don't believe a word she said, Meghan Markle. I wouldn't believe it if she read me a weather report.» (переклад з англійської мови – «Пробачте, але я не вірю в те, що вона сказала, Меган Маркл. Я б не повірив, навіть якщо б вона читала мені погоду»). Ці слова спричинили велику хвилю критики з боку глядачів, що підняло питання про серйозність відношення до психічного здоров'я та важливість його обговорення. Через обурення громадськості і більш

ніж 41 000 скарг, поданих на Офіс з комунікацій – регулюючий орган Великобританії, Морган покинув програму «Good Morning Britain». Цей інцидент став одним із найбільш резонансних у кар’єрі Моргана і викликав широке обговорення про важливість підтримки психічного здоров’я, а також про відповідальність медіа-персоналій за їх висловлювання.

У світі, де інформація поширюється з надзвичайною швидкістю, здатність чітко та зважено висловлювати свої думки стає вирішальним фактором успіху в будь-якій сфері. Аналіз негативного впливу емоцій на подачу думок підкреслює важливість емоційного інтелекту та самоконтролю як для індивідуального розвитку, так і для підтримки здорової соціальної взаємодії. Незважаючи на те, що емоції є невід’ємною частиною нашої природи, їх надлишковість може перешкоджати об’єктивному сприйняттю, аналізу і передачі інформації, створюючи бар’єри на шляху до взаєморозуміння.

1.2 Огляд існуючих методів вирішення проблеми

1.2.1 Сумаризація тексту

Сумаризація тексту – це процес створення короткого викладу основного змісту більш об’ємного текстового матеріалу. Ця процедура включає вибірку ключових пунктів, фактів та ідей оригінального тексту та їх подальше представлення у стислій формі.

Системи сумаризації тексту (ATS) можна класифікувати як однодокументні або багатодокументні системи узагальнення. Перші створюють реферат на основі одного документа, а другі – на основі групи документів [3]. Системи ATS розробляються із застосуванням одного з підходів до реферування тексту: екстрактивного, абстрактного або гібридного. Екстрактивний підхід вибирає найважливіші речення з вхідного тексту і використовує їх для створення резюме. Абстрактний підхід

представляє вхідний текст у проміжній формі, а потім генерує резюме зі словами та реченнями, які відрізняються від речень вихідного тексту. Гібридний підхід поєднує в собі як екстрактивний, так і абстрактний підходи.

Загальна архітектура таких систем включає три етапи:

1) попередня обробка: створення структурованого представлення оригінального тексту з використанням багатьох лінгвістичних методів, таких як сегментація речень, токенізація слів, видалення стоп-слів, позначення частин мови, стеммінг тощо [4];

2) обробка: використання одного з підходів до узагальнення тексту шляхом застосування одного або кількох методів для перетворення вхідного документа чи документів на резюме;

3) після-обробка: вирішення деяких проблем у згенерованих реченнях резюме, таких як розпізнавання анафори та переупорядкування вибраних речень перед створенням остаточного резюме.

Власне обробка може бути застосована з використанням різних підходів. До них відносяться:

– статистичні методи. Вони засновані на кількісних характеристиках тексту та можуть включати частотний аналіз, наприклад, вибірку речень на основі частоти виникнення певних слів, або ж позиційні методи, до яких відноситься оцінка важливості речення на основі його положення у тексті, виходячи з припущення, що вступ і висновок часто містять ключову інформацію;

– методи на основі графів. Ці методи представляють текст як граф, де вузли є реченнями, а ребра відображають відносини між ними (наприклад, семантичну схожість) та проводиться аналіз зв'язності для визначення важливості речення на основі його зв'язків з іншими реченнями;

– концептуальні методи. Ці методи фокусуються на ідентифікації та аналізі ключових концептів або тем у тексті, наприклад, групування речень

на основі їхньої близькості до певних тем або концептів для визначення найбільш репрезентативних висловлювань;

– методи, засновані на темах. Ці методи аналізують структуру та зміст тексту для виявлення основних тем. Вони використовують алгоритми, такі як Latent Dirichlet Allocation, для ідентифікації тем у тексті та вибірки речень, що найкраще їх представляють;

– машинне навчання та штучний інтелект;

– гібридні методи.

Сумаризація як метод нейтралізації емоцій у текстах зменшує емоційн заряд, зберігаючи при цьому ключовий зміст та інформаційну суть матеріалу. Сумаризація дозволяє не тільки ідентифікувати та видалити емоційно заряджені слова та фрази, але й концентрує увагу на фактичній інформації, відсіваючи суб'єктивні емоційні вираження.

1.2.2 Великі мовні моделі

Великі мовні моделі (LLM), такі як GPT, за останні роки досягли значного прогресу в обробці природної мови [5]. Ці моделі навчаються на величезних обсягах текстових даних і здатні генерувати людський текст, відповідати на запитання та виконувати інші завдання, пов'язані з мовою, з високою точністю. Одним із ключових досягнень у цій галузі є використання архітектур трансформерів та базового механізму уваги, які значно покращили здатність мовних моделей обробляти довготривалі залежності в природномовних текстах [6]. Зокрема, архітектура трансформера, представлена в роботі Васвані [7], використовує механізм самоуваги для визначення релевантності різних частин вхідних даних під час генерування прогнозів. Це дозволяє моделі краще розуміти зв'язки між словами в реченні, незалежно від їхньої позиції [8].

Існує два найпоширеніших типи LLM – GPT (Generative Pre-trained Transformer) і BERT (Bidirectional Encoder Representations from Transformers), які засновані на технології трансформерів.

Трансформери – це архітектура глибоких нейронних мереж, яка була вперше представлена в роботі «Attention is All You Need» в 2017 році. Ця архітектура революціонізувала обробку природної мови (NLP) і стала основою для багатьох потужних моделей, таких як GPT, BERT та інших.

Основна ідея трансформерів – це використання механізму уваги для ефективного аналізу та генерації послідовностей даних. Вони складаються з двох основних компонентів: кодувального (encoder) і декодувального (decoder).

Кодувальний шар (encoder) отримує на вхід послідовність токенів (слів або символів) і генерує складний векторний представлення цієї послідовності. Кожен токен подається через декілька блоків трансформера, які працюють паралельно і незалежно.

Декодувальний шар (decoder) використовується для генерації послідовності на виході. Він також складається з декількох блоків трансформера, але додатково має механізм уваги, який дозволяє моделі звертатися до векторних представлень вхідних токенів, що були згенеровані кодувальним шаром.

Механізм уваги (attention mechanism) – це ключовий компонент трансформерів. Він дозволяє моделі взаємодіяти з різними частинами вхідної послідовності при генерації вихідної послідовності. Механізм уваги дозволяє моделі визначати, на які елементи вхідної послідовності слід звернути більше уваги при генерації вихідного токена.

Трансформери виявилися дуже ефективними для багатьох завдань у обробці природної мови, завдяки їхній здатності розуміти довгострокові залежності в тексті та зберігати контекстуальну інформацію.

GPT – це серія моделей глибокого навчання з архітектурою трансформера, розроблених компанією OpenAI. GPT був випущений у кількох версіях, кожна з яких має певні вдосконалення та покращення.

Основна ідея за GPT полягає у використанні передбачення мови без учителя на великих корпусах текстів для попереднього навчання моделі. Після цього, ці попередньо навчені моделі можуть бути доопрацьовані для вирішення конкретних завдань, таких як машинний переклад, сентиментний аналіз, генерація тексту тощо.

BERT – це один з найбільш впливових та потужних моделей в області обробки природної мови. Розроблений у 2018 році дослідницькою групою в Google.

Основна особливість BERT полягає в тому, що він є бідирекційним (bidirectional), що означає, що він може враховувати контекст з обох сторін кожного слова при аналізі тексту. Це відрізняє BERT від більшості попередніх моделей, які були однонаправленими і аналізували текст тільки зліва направо або справа наліво.

BERT виявився дуже ефективним у багатьох завданнях NLP, завдяки його здатності до глибокого розуміння контексту тексту та збереження семантичних залежностей.

1.3 Постановка задачі

У сучасному світі, де інформаційний обмін відбувається надзвичайно швидко, важливим аспектом комунікації є здатність зберігати нейтральний тон у текстах. Особливо це актуально в бізнесі, журналістиці та соціальних мережах, де емоційний забарвлення тексту може вплинути на сприйняття та реакцію аудиторії. Відповідно, виникає потреба у методах, які дозволяють автоматично нейтралізувати емоційний компонент текстів.

Основною метою даного дослідження є оцінка ефективності нейтралізації тексту від емоцій за допомогою моделі GPT. Задача –

визначити, наскільки успішно сучасні мовні моделі здатні виявляти та усувати емоційні аспекти у текстах різного характеру.

У ході виконання роботи було зроблено:

– аналіз існуючих методів нейтралізації емоцій у тексті. Проведено огляд літератури щодо існуючих підходів та технологій, які використовуються для нейтралізації емоцій у текстах;

– вибір та підготовка даних. Згенеровано корпус текстів, які мають яскраво виражене емоційне забарвлення, та підготовлено їх для аналізу за допомогою моделі GPT;

– проведення експерименту. Проведено експеримент з нейтралізації емоцій у підготовлених даних різними способами;

– оцінка ефективності. Розроблено критерії для оцінки ефективності нейтралізації емоцій у текстах та проведено порівняльний аналіз результатів;

– висновки та рекомендації. На основі отриманих результатів зроблено висновки щодо можливостей та обмежень використання моделі GPT для нейтралізації емоцій у текстах, а також розробити рекомендації для подальших досліджень та вдосконалення методології.

2 ТЕОРЕТИЧНІ ДОСЛІДЖЕННЯ

2.1 Теорія базових емоцій

Емоції – це внутрішнє переживання або рух настрою, які можуть бути викликані ситуацією чи подією. Емоції також можуть виникати спонтанно. Емоції є суб'єктивними переживаннями і пов'язані з фізичними реакціями, виразом обличчя та поведінкою.

У біологічному розумінні емоція – це реакція мозку на афективний стимул. Як у людей, так і у тварин це майже автоматично виражається в певній моделі поведінки, наприклад у втечі або наближенні, а також у фізіологічних реакціях. Емоції це біологічно обґрунтовані психологічні стани, що виникають у результаті нейрофізіологічних змін, по-різному пов'язані з думками, почуттями, поведінковими реакціями і ступенем задоволення чи страждання [9]. Емоції часто переплітаються з настроєм, темпераментом, особистістю, диспозицією або креативністю [10].

Пол Екман – американський психолог, який здобув світову відомість завдяки своїм дослідженням у галузі емоцій та їх вираження на обличчі. Його праці значно вплинули на розвиток психології, криміналістики та навіть розважальної індустрії.

Екман почав свою кар'єру як дослідник в Університеті Каліфорнії в Сан-Франциско, де провів більшу частину своєї професійної діяльності. Протягом своєї кар'єри Екман досліджував взаємозв'язок між емоціями та їх невербальними проявами, зокрема мікроекспресіями – швидкими, мимовільними рухами м'язів обличчя, які відображають справжні емоції людини. Він також розробив методику виявлення брехні, засновану на аналізі мікроекспресій.

Пол Екман розробив теорію базових емоцій. Вона базується на ідеї, що існує обмежена кількість емоцій, які є універсальними для всіх людей, незалежно від їх культурного чи етнічного походження. Ці емоції мають

специфічні, автоматичні вирази обличчя, які можуть бути легко розпізнані. Основні положення теорії Екмана можна розбити на кілька ключових аспектів – універсальність емоцій, автоматичні вирази обличчя, мікрОВИРАЗИ та біологічне підґрунтя.

Під універсальністю емоцій Пол Екман має на увазі, що кожна людина незалежно від культури або походження має однаковий набір базових емоцій і виражає їх схожим чином. Цей висновок був підтверджений численними дослідженнями, які проводилися у різних культурах.

Кожна базова емоція має специфічний вираз обличчя, який виникає автоматично і майже не піддається контролю. Наприклад, страх супроводжується розширеними очима і відкритим ротом, а гнів – нахмуреними бровами і стиснутими губами.

МікрОВИРАЗИ – це дуже швидкі (тривають менше секунди) і мимовільні вирази обличчя, які відображають справжні емоції людини, навіть якщо вона намагається їх приховати. Екман розробив методи для розпізнавання цих мікрОВИРАЗІВ, що допомагає у виявленні брехні та розумінні справжніх почуттів людей.

Науковець припускає, що базові емоції мають біологічне підґрунтя і еволюційно обумовлені. Це означає, що емоції допомагали нашим предкам виживати, реагуючи на загрози, соціальні взаємодії та інші важливі аспекти життя [11].

Екман припустив, що існує сім емоційних проявів, універсальних для людей у всьому світі: щастя, сум, здивування, страх, гнів, огида, презирство. Згодом він вирішив, що огида і презирство дуже близькі за мімікою і у списку лишилося 6 базових емоцій [12]:

– радість. Згідно з дослідженнями, основними ознаками радості є підняття кутиків рота і напруження м'язів повік. Наш вираз обличчя може призвести до широкої посмішки, яка зачіпає наші очі. Ми також можемо сміятися, коли щасливі. Одним з важливих відкриттів Екмана є те, що люди з різних культур розпізнають радість однаково, незалежно від їхнього

культурного контексту. Це підтверджує теорію про універсальність базових емоцій;

– смуток. Коли ми відчуваємо щось, що емоційно ранить нас, ми можемо реагувати на це почуттям смутку. Сум часто виражається в тому, що куточки рота опускаються в насуپленість, опускаються повіки і піднімаються внутрішні куточки брів, а зовнішні залишаються нахиленими донизу. Сльозотеча або сльози також можуть супроводжувати смуток;

– гнів. Багато психологів вважають її вторинною, тобто вона може з'являтися після первинних емоцій, таких як смуток або страх, або приховувати їх. Реакція гніву може проявлятися у стиснутих губах, вирачених очах і опущених бровах, які концентруються в проміжку між бровами;

– здивування. Здивування як емоція може бути більше пов'язане з шоком, ніж з приємною або радісною несподіванкою в цьому контексті. Здивування – це емоція, яку ми зазвичай відчуваємо після новин, що змінюють життя. Коли людина здивована, її очі можуть розширюватися, брови підніматися вгору, а щелепа злегка відвисати, внаслідок чого рот набуває круглої або овальної форми;

– страх. Коли людина боїться, її очі можуть розширюватися, а брови підніматися. Страх у ділянці рота може проявлятися в тому, що губи витягнуті горизонтально по всій довжині рота. Ці дії м'язів створюють класичний вираз страху, який є впізнаваним у всіх культурах і допомагає людям комунікувати про наявність небезпеки або загрози;

– огида. Пол Екман описує огиду як базову емоцію з універсальними виразами обличчя, що включають зморщування носа, підняття верхньої губи, і опускання куточків губ. Цей вираз обличчя сигналізує про реакцію на щось неприємне або небезпечне, допомагаючи уникати потенційно шкідливих речовин або ситуацій.

2.2 Аналіз емоційного забарвлення в тексті

Міміка, жести, мова та текст загалом виражають настрій та емоції людини. На відміну від міміки та розпізнавання мови, текстове речення втрачає здатність визначати себе, оскільки воно позбавлене смаку. Через складність і неоднозначність тексту, з'ясувати емоції цього тексту є складним завданням. Розпізнати емоцію певного тексту стає складним завданням, оскільки кожне слово може мати різне значення та морфологічну форму [13].

2.2.1 Лексичні методи

Лексичні методи для визначення емоційної забарвленості тексту ґрунтуються на використанні словників, що містять слова і фрази, класифіковані за емоційними ознаками. Ці методи є одними з найпростіших і найпоширеніших підходів в аналізі тональності тексту.

Словники поділяються на дві групи: словники емоцій і словники тональності.

Словники емоцій містять слова або фрази, які пов'язані з певними емоціями, та можуть використовуватися для автоматичного визначення емоційного стану тексту. До найбільш популярних відносяться WordNet-Affect, NRC Emotion Lexicon та LIWC (Linguistic Inquiry and Word Count) [14].

WordNet-Affect є розширенням стандартного WordNet, яке додає інформацію про емоційні стани. У цьому словнику слова організовані у синонімічні ряди (синсети), де кожен ряд містить слова зі схожим значенням. Кожен синсет може бути пов'язаний з різними афективними мітками, які вказують на емоційні стани, такі як радість, сум, гнів тощо.

NRC Emotion Lexicon – словник, також відомий як EmoLex, містить понад 10 000 англійських слів, кожне з яких асоційоване з однією або

декількома з восьми основних емоцій: радість, довіра, страх, несподіванка, сум, відраза, гнів, очікування. Окрім цього, слова також класифікуються за позитивною або негативною тональністю. NRC Emotion Lexicon широко використовується завдяки своїй простоті і покриттю.

LIWC – це комерційний словник, який містить слова та категорії слів, пов'язані з емоціями, соціальними процесами, когнітивними процесами та іншими психологічними характеристиками. Він часто використовується в психологічних дослідженнях для аналізу текстів.

Словники тональності – це спеціалізовані словники, що містять слова, класифіковані за їхньою тональністю (позитивною, негативною або нейтральною). Вони використовуються для автоматичного визначення загальної тональності тексту. До них відносяться SentiWordNet, AFINN та VADER (Valence Aware Dictionary and sEntiment Reasoner).

SentiWordNet – це розширення WordNet, в якому кожен синсет містить оцінки тональності: позитивну, негативну та нейтральну.

AFINN – це словник тональності, який містить слова з оцінками тональності від -5 до +5, де негативні числа вказують на негативну тональність, а позитивні – на позитивну. Цей словник був розроблений для аналізу тону текстів у соціальних мережах.

VADER – це словник, який також розроблений спеціально для аналізу тональності текстів у соціальних мережах. Він враховує не тільки самі слова, але й контекст, зокрема капіталізацію, знаки пунктуації та інтенсивність.

Використання лексичних методів має очевидні переваги. Словники легко використовувати, пошук слів по словнику займає мало часу, а результати легко інтерпретувати. Але одночасно з цим ці методи мають недоліки, адже словники не можуть зберігати інформацію про усі можливі слова, особливо нові чи сленгові. До того ж такі методи не враховують контекст, з яким слова були використані, а також не враховують особливості

деяких слів мати більше одного значення, що може призвести до неправильного висновку.

Словники є популярними інструментами для базового аналізу текстів у різних сферах, таких як маркетинг, соціальні мережі, відгуки клієнтів. Вони можуть бути використані для швидкого аналізу великих обсягів текстів, щоб визначити загальну тональність і емоційний настрій. Проте, для більш точних і контекстуально обґрунтованих результатів, часто застосовують більш складні підходи.

2.2.2 Статистичні методи

Статистичні методи для визначення емоційної забарвленості тексту використовують математичні та статистичні підходи для аналізу частотності слів та фраз. Вони допомагають виявити патерни, характерні для різних емоційних станів. До основних відносяться N-грами, TF-IDF (Term Frequency-Inverse Document Frequency) та частотний аналіз.

Метод N-грам – це один з основних статистичних методів, що використовуються для аналізу тексту. Він базується на розділенні тексту на послідовності з N слів (або символів), які називаються N-грамами. Цей метод дозволяє виявляти патерни у тексті, які можуть бути характерними для певних емоцій або тональності.

Самі N-грами – це послідовності з N елементів (зазвичай слів), які використовуються для моделювання мови та аналізу тексту. Вони в більшості випадків поділяються на уніграми, що є окремими словами, тобто для тексту «I am very happy» уніграми будуть «I», «am», «very», «happy»; біграми, тобто послідовністю з двох слів, наприклад, «I am» і «very happy»; та триграми – «I am very» і «am very happy».

N-грами можуть бути використані для аналізу емоційної забарвленості тексту шляхом визначення частотності певних послідовностей слів, що часто зустрічаються у текстах з відомою емоційною

забарвленістю. Наприклад, біграми, такі як «very happy» або «feel sad», можуть бути характерними для позитивних або негативних емоцій відповідно.

Метод TF-IDF – це метод, що оцінює важливість слів у документі на основі їх частотності в окремому документі (TF) та загальної частотності у всіх документах колекції (IDF). Він дозволяє виділити слова, які є важливими для конкретного документа і менш поширеними у всій колекції.

TF вимірює, наскільки часто термін з'являється в документі. TF для терміну t у документі d обчислюється як відношення кількості появ цього терміну до загальної кількості слів у документі:

$$TF(t, d) = \frac{\text{Кількість появ терміну } t \text{ у документі } d}{\text{Загальна кількість термінів у документі } d}. \quad (2.1)$$

IDF вимірює важливість терміну. Вона обчислюється як логарифм відношення загальної кількості документів до кількості документів, які містять термін t :

$$IDF(t, D) = \log\left(\frac{\text{Загальна кількість документів}}{\text{Кількість документів, що містять термін } t}\right). \quad (2.2)$$

Об'єднуючи ці дві міри, отримуємо вагу TF-IDF терміну t у документі d :

$$TF-IDF(t, d, D) = TF(t, d) \times IDF(t, D). \quad (2.3)$$

TF-IDF залишається одним з найбільш популярних методів для попередньої обробки текстових даних перед застосуванням більш складних алгоритмів машинного навчання та аналізу даних.

Метод частотного аналізу є одним з найпростіших для визначення емоційної забарвленості тексту [15]. Він базується на підрахунку частоти

появи емоційно забарвлених слів і дозволяє оцінити загальний емоційний тон тексту. Цей метод використовує словники емоційних слів, наприклад, вищезгаданий LIWC. Кожне слово у тексті перевіряється на наявність у лексиконі емоційних слів. Якщо слово знайдено, його емоційна категорія враховується. Підраховується кількість слів, що належать до кожної емоційної категорії. На основі частоти появи емоційних слів визначається загальна емоційна забарвленість тексту. Наприклад, якщо у тексті багато слів, пов'язаних з радістю, можна зробити висновок, що текст має позитивне емоційне забарвлення.

Статистичні методи мають схожі переваги і недоліки на ті, що мають лексичні методи. Ці методи дуже легко реалізуються і інтерпретуються. Вони можуть бути використані для великих об'ємів даних. Проте ці методи не враховують контексту, в якому використовуються ті чи інші слова. Також ці методи потребують попередньої обробки тексту, наприклад, очищення від стоп-слів, таких як «і», «на», «або» тощо.

2.2.3 Методи машинного навчання

Методи машинного навчання (ML) для визначення емоцій у тексті значно підвищують точність та ефективність аналізу емоційної забарвленості.

Навчання з вчителем є однією з основних технік для визначення емоцій у тексті. Одним із найбільш простих і популярних методів у цій категорії є наївний Баєсів класифікатор. Цей метод базується на теоремі Баєса і дозволяє оцінювати ймовірність того, що певний текст належить до однієї з емоційних категорій. Він швидкий та ефективний для великих обсягів даних, але може бути менш точним для текстів із складними контекстами, де слова мають багатозначність. Дослідження показали ефективність наївних Баєсових класифікаторів у аналізі тональності текстів.

Іншим потужним методом супервізованого навчання є підтримуючі векторні машини (SVM). SVM працює шляхом знаходження оптимальної гіперплощини, яка розділяє дані на класи. Цей метод забезпечує високу точність і ефективність у багатьох задачах класифікації тексту. Проте, він вимагає налаштування гіперпараметрів і може бути обчислювально затратним. Дослідження продемонстрували потужність SVM у категоризації текстів [16].

Метод логістичної регресії також широко використовується для прогнозування ймовірності належності тексту до певної категорії на основі незалежних змінних. Він є простішим у використанні та інтерпретації, але може бути менш точним у порівнянні з більш складними моделями.

Глибоке навчання надає більш складні методи для аналізу тексту. Рекурентні нейронні мережі (RNN), особливо їх варіації, такі як LSTM (Long Short-Term Memory) та GRU (Gated Recurrent Units), широко використовуються для обробки послідовних даних. Ці моделі враховують послідовність слів у тексті, що дозволяє краще розуміти контекст. RNN мають високу точність і здатність враховувати довготривалі залежності, але вимагають великого обсягу даних для навчання і значних обчислювальних ресурсів.

Конволюційні нейронні мережі (CNN), хоча зазвичай використовуються для обробки зображень, також ефективно застосовуються для тексту, особливо для виявлення ключових фраз і структур. CNN ефективні для виявлення локальних патернів у тексті, але менш ефективні для врахування довготривалих залежностей.

Методи машинного навчання для визначення емоцій у тексті є потужними інструментами, які забезпечують високу точність і можливість врахування контексту. Вони використовуються у різних галузях для аналізу емоційного фону текстів, що дозволяє краще розуміти настрої користувачів, їхні потреби та реакції. Використання цих методів

може значно покращити результати аналізу текстів і забезпечити глибше розуміння емоційної забарвленості текстового контенту.

Методи машинного навчання досягають значно більшої точності у порівнянні з лексичними чи статичними методами. Проте, для їх ефективного використання потрібні значні обчислювальні ресурси та великі обсяги навчальних даних. До того ж деякі моделі, особливо глибокі нейронні мережі, можуть бути складними для інтерпретації.

2.2.4 Трансформери

Трансформери є одними з найсучасніших та найпотужніших архітектур для обробки природної мови (NLP). Вони були вперше представлені в роботі у 2017, і з того часу трансформери стали основою для багатьох моделей NLP, включаючи BERT, GPT-3 та інші.

Трансформери складаються з двох основних компонентів: енкодера та декодера (рисунок 2.1). Енкодер складається з послідовності шарів, кожен з яких містить механізм самоперевірки та повнозв'язний шар. Декодер також складається з послідовності шарів, але додатково містить механізм крос-перевірки, який взаємодіє з виходами енкодера. Основним нововведенням трансформерів є механізм самоперевірки, який дозволяє моделі враховувати контекст кожного слова у вході безпосередньо, замість послідовного оброблення, як у рекурентних нейронних мережах.

Механізм самоперевірки дозволяє моделі визначати, які інші частини входу є важливими для поточного слова. Це досягається шляхом обчислення зважених сум всіх слів у вході з урахуванням їх релевантності до поточного слова. Це дозволяє трансформерам ефективно обробляти залежності на довгих відстанях, що є обмеженням для традиційних RNN.

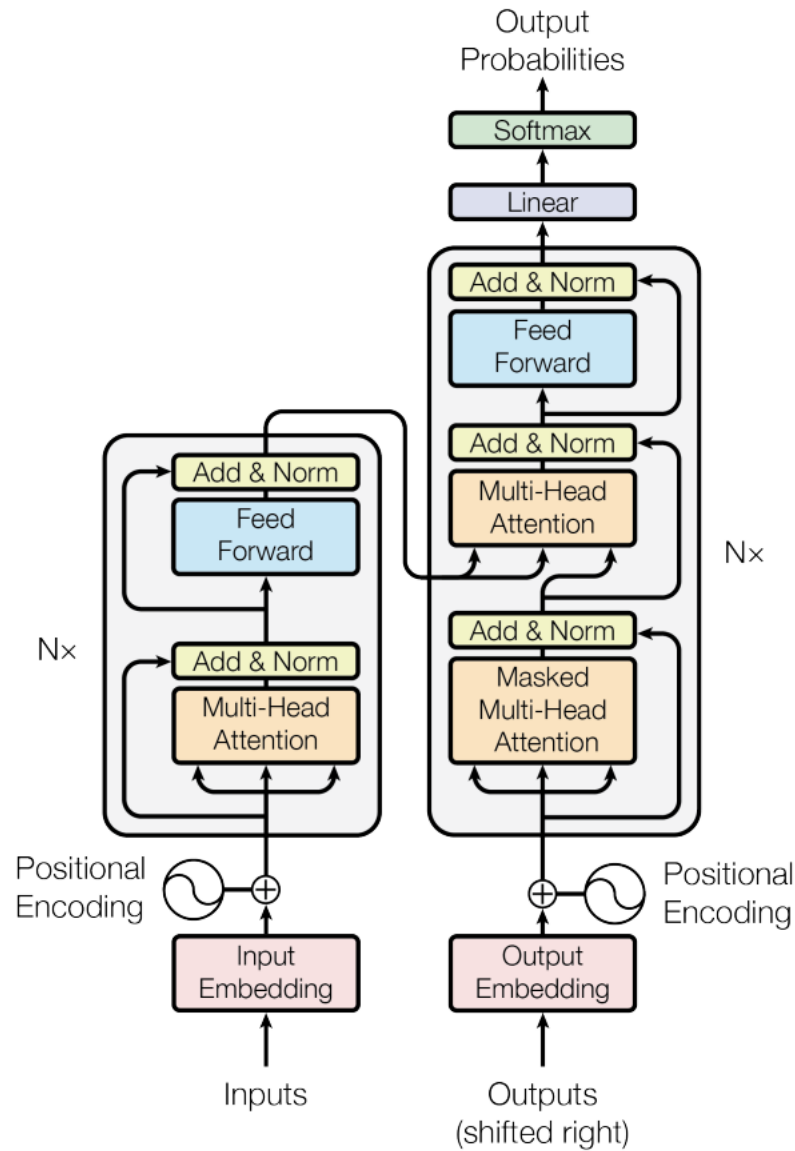


Рисунок 2.1 – Архітектура трансформера

2.2.4.1 BERT

BERT є однією з найвідоміших моделей на основі трансформерів, розробленою Google AI Language. BERT використовує двонаправлений механізм перевірки, що дозволяє моделі враховувати контекст слова як зліва, так і справа. Це робить BERT особливо ефективним для задач, де контекст має вирішальне значення, таких як класифікація емоцій у тексті. BERT був навчений на великому корпусі текстів з двома основними

завданнями: заповнення пропусків у тексті та прогнозування наступного речення.

Дослідження Девліна продемонструвало значні покращення в багатьох задачах NLP, включаючи класифікацію емоцій, завдяки використанню BERT. Інші дослідження також підтвердили ефективність BERT у різних задачах обробки природної мови, що підкреслює його важливість і вплив у цій галузі [6].

2.2.4.2 RoBERTa

У 2019 році дослідники з Facebook AI заявили у статті «RoBERTa: A Robustly Optimized BERT Pretraining Approach», що BERT був значно недонавченим, і запропонували покращений алгоритм навчання BERT-моделей, який називається RoBERTa, що може зрівнятися або навіть перевершити ефективність усіх майбутніх моделей на основі BERT [17]. Основною метою RoBERTa є покращення продуктивності BERT шляхом оптимізації процесу попереднього навчання та використання великих обсягів даних.

Одне з ключових вдосконалень RoBERTa полягає у використанні значно більших обсягів текстових даних для попереднього навчання. Модель була навчена на декількох великих корпусах, включаючи новинні статті, книги, веб-сторінки та інші джерела, що дозволяє отримати більш загальні та точні векторні представлення.

Ще одним покращенням є динамічне маскуванню. BERT покладається на випадкове маскуванню і передбачення токенів. Оригінальна реалізація BERT виконувала маскуванню один раз під час попередньої обробки даних, в результаті чого отримували одну статичну маску. Щоб уникнути використання однієї і тієї ж маски для кожного навчального прикладу в кожній епосі, навчальні дані для RoBERTa були продубльовані 10 разів, так що кожна послідовність була замаскована 10 різними способами

протягом 40 епох навчання. Таким чином, кожна навчальна послідовність була показана з тією самою маскою чотири рази під час навчання. Це стає критично важливим, коли потрібно навчити більше кроків або з більшими наборами даних.

Також дослідники зауважили, що попередні роботи в галузі нейронного машинного перекладу показали, що навчання з дуже великими міні-пакетами вхідних даних може підвищити швидкість оптимізації та продуктивність кінцевого завдання за умови відповідного збільшення швидкості навчання [18]. Одна з наступних робіт показала, що BERT також піддається навчанню великими партіями [19]. Тож вони збільшили розмір таких міні-пакетів, що показало кращі результати, ніж той розмір, що використовувався для навчання BERT.

RoBERTa зберігає базову архітектуру трансформера, що використовується в BERT, з багатьма шарами уваги та енкодерами. Вона також підтримує двонаправленість, що означає, що модель враховує контекст з обох боків слова. Це дозволяє моделі отримувати багатші та контекстуально точніші представлення слів і речень. Проте хоч в оригінальній моделі BERT використовується задача прогнозування наступного речення (NSP), RoBERTa відмовляється від цієї задачі, оскільки дослідження показали, що вона не приносить значної користі. Замість цього модель зосереджується на завданні маскуванню слів (MLM), що дозволяє моделі краще розуміти контекст.

2.2.4.3 DistilRoBERTa

DistilRoBERTa – це полегшена та оптимізована версія моделі RoBERTa, розроблена для забезпечення швидшої і менш ресурсоємної обробки тексту, зберігаючи при цьому більшу частину точності оригінальної моделі. DistilRoBERTa була створена командою Hugging Face, яка також розробила DistilBERT [20]. Основна мета цієї моделі – зменшити

обчислювальні витрати та вимоги до пам'яті, при цьому підтримуючи високу продуктивність у задачах обробки природної мови.

DistilRoBERTa має значно менше параметрів порівняно з оригінальною RoBERTa моделлю, що зменшує вимоги до пам'яті та робить її легшою для розгортання на обмежених ресурсах, таких як мобільні пристрої та вбудовані системи. Завдяки меншій кількості параметрів і оптимізованій архітектурі, DistilRoBERTa працює швидше, що дозволяє ефективніше використовувати її для реальних застосувань, де важлива швидкість обробки, наприклад, у чат-ботах і сервісах реального часу.

Незважаючи на зменшений розмір, DistilRoBERTa зберігає більшу частину точності оригінальної RoBERTa моделі, забезпечуючи високу продуктивність у різних задачах NLP, наприклад, класифікація текстів.

DistilRoBERTa використовує техніку дистиляції знань, яка включає в себе навчання меншої моделі-учня на основі знань, переданих від більшої, попередньо навченої моделі-вчителя. Основні кроки процесу дистиляції знань включають:

- навчання вчителя. Спочатку велика модель, така як RoBERTa, навчається на великому корпусі текстових даних;
- дистиляція знань. Менша модель DistilRoBERTa навчається на основі виходів вчителя. Це здійснюється шляхом мінімізації різниці між виходами вчителя та учня;
- оптимізація. Втрата дистиляції включає компонент традиційної крос-ентропії та компонент для мінімізації різниці між виходами вчителя та учня, що допомагає меншій моделі навчитися не тільки точним відповідям, але й розподілу ймовірностей, який генерує вчитель.

DistilRoBERTa є ефективною та компактною версією RoBERTa, яка зберігає високу продуктивність, але має менші розміри і швидшу обробку. Завдяки використанню техніки дистиляції знань, DistilRoBERTa забезпечує високу точність при значно менших вимогах до обчислювальних ресурсів,

що робить її ідеальним вибором для реальних застосувань у обробці природної мови, де важливі швидкість і ефективність.

2.3 Prompt-engineering

Prompt-engineering – це процес створення та налаштування запитів, тобто промтів, для взаємодії з мовними моделями, такими як GPT. Цей процес передбачає підбір слів, фраз та структур запитів, які допомагають отримати найбільш релевантні та точні відповіді від моделі. Prompt-engineering потрібен для оптимізації взаємодії з мовною моделлю, покращення якості генерованих відповідей та підвищення ефективності використання моделей у різних прикладних сценаріях, таких як автоматизація обробки текстів, створення контенту, дослідження ринку тощо.

Промпт – це набір інструкцій, наданих LLM, які програмують LLM, налаштовуючи його та розширюючи чи вдосконалюючи його можливості [21]. Підказка може впливати на подальшу взаємодію з LLM і на результати, що генеруються LLM, надаючи конкретні правила і вказівки для діалогу LLM з набором початкових правил. Зокрема, підказка задає контекст розмови і повідомляє LLM, яка інформація є важливою і якою має бути бажана форма і зміст вихідних даних. Наприклад, підказка може вказувати, що LLM повинен генерувати тільки код, який відповідає певному стилю кодування або парадигмі програмування. Так само можна вказати, що LLM повинен позначати певні ключові слова або фрази у створеному документі і надавати додаткову інформацію, пов'язану з цими ключовими словами. Впроваджуючи ці рекомендації, підказки сприяють створенню більш структурованих і деталізованих результатів, які допомагають у вирішенні широкого спектру завдань [22]. «What colour is the sun?» – це приклад одного з найпростіших промтів.

Дослідження у сфері prompt-engineering зосереджені на різних аспектах, включаючи вивчення різних підходів до створення ефективних запитів, таких як використання ключових слів, контекстних підказок та структурованих шаблонів; аналіз впливу різних запитів на якість відповідей, визначення найбільш оптимальних практик для конкретних завдань; розробка інструментів та алгоритмів для автоматичного генерування та оптимізації запитів, що дозволяє зменшити час та зусилля, необхідні для налаштування взаємодії з мовною моделлю.

Ця область активно розвивається, оскільки ефективне використання мовних моделей має великий потенціал для різних галузей, від освіти до бізнесу.

На сьогодні є різні шаблони промптів, які допомагають досягнути кращих результатів.

2.3.1 Persona pattern

Шаблон персони (Persona pattern) – це приклад одного з таких методів. Persona pattern полягає в тому, щоб задавати контекст, який визначає особистість, роль або стиль мовної моделі під час генерації відповіді. Це може включати інформацію про тон, манеру спілкування, знання та інші характеристики, що впливають на поведінку моделі. Це допоможе отримати більш індивідуалізовані відповіді, які краще відповідають потребам і очікуванням користувача; забезпечення узгодженості у відповідях, що важливо для брендів або проектів, які потребують єдиного стилю спілкування; наближення взаємодії з моделлю до реальної розмови, коли модель імітує певний тип особистості.

Такий шаблон є доволі простим – «You are X. Do Y.» або «Act as X. Do Y.» (рисунок 2.2).

You are a famous science fiction writer. Write the beginning of a story about time travel.



Title: **Echoes of Tomorrow**

In the year 2478, humanity had mastered many things—interstellar travel, artificial intelligence, and quantum entanglement. Yet, the concept of time travel remained elusive, wrapped in theoretical quandaries and existential risks. That is, until Dr. Elara Myles stumbled upon the Chrono-Key, a device buried deep within the ruins of an ancient civilization on the recently terraformed Mars.

Elara, a brilliant physicist with an unruly mop of curly hair and a penchant for vintage Earth rock music, had dedicated her life to unraveling the mysteries of time. Her discovery, a small, intricately carved obsidian disk, pulsed with a strange, iridescent light that seemed to dance just beneath its surface, resonating with the fabric of space-time itself.

The Chrono-Key, as she called it, had the power to open rifts in time, allowing one to step through the ages as easily as walking from one room to another. It operated on principles that bent the known laws of physics, principles that Elara was both thrilled and terrified to explore.

With the support of her team at the Temporal Research Institute, she conducted her first experiment. The laboratory was filled with equipment that buzzed and hummed with energy, the air crackling with anticipation. Elara held the Chrono-Key up, her hands trembling slightly as she activated its mechanism. A shimmering portal sprang to life, its edges swirling with a kaleidoscope of colors that defied description.

Elara stepped through the portal, her heart pounding in her chest. She emerged not in another place, but another time—Paris, 1925. The city buzzed with the energy of a post-war world, the air filled with the sounds of jazz and the smells of fresh pastries and espresso. As she walked the cobblestone streets, Elara realized the profound responsibility she held. The past was no longer a memory; it was alive, tangible, and incredibly fragile.

Рисунок 2.2 – Приклад використання persona pattern

Persona pattern – це потужний інструмент у prompt-engineering, який дозволяє створювати більш релевантні, персоналізовані та реалістичні відповіді від мовних моделей. Від чіткої формулювання ролі та контексту до встановлення відповідного тону і стилю, цей підхід може значно покращити якість взаємодії з моделлю.

2.3.2 Audience persona pattern

Ще одним шаблоном є шаблон аудиторії персон (audience persona pattern). Audience persona pattern в prompt-engineering є важливою технікою, що використовується для створення запитів, орієнтованих на певну аудиторію або групу користувачів. Цей підхід враховує особливості та потреби конкретної аудиторії, щоб мовна модель могла генерувати більш релевантні та адаптовані відповіді. Використання audience persona pattern передбачає налаштування промптів таким чином, щоб враховувати такі характеристики аудиторії, як вік, інтереси, рівень знань, а також специфічні потреби або запити.

Цей підхід є необхідним для забезпечення персоналізації контенту, що дозволяє створювати відповіді, які відповідають потребам і очікуванням конкретної групи користувачів. Це підвищує ефективність комунікації, роблячи відповіді більш релевантними та корисними, що в свою чергу сприяє кращій взаємодії з аудиторією. Залучення аудиторії до взаємодії з моделлю за допомогою орієнтованих відповідей підвищує їх задоволеність і інтерес до отриманого контенту.

Процес використання audience persona pattern починається з визначення цільової аудиторії, що включає демографічні характеристики, інтереси та потреби. Після цього підбирається відповідний тон і стиль комунікації, враховуючи ці характеристики. Наприклад, для студентів старших класів, які цікавляться наукою, можна використовувати просту мову та приклади з реального життя. Наступним кроком є включення релевантного контенту, який буде цікавим і корисним для аудиторії, та надання чітких інструкцій, як модель повинна відповідати, враховуючи потреби аудиторії.

Практичні приклади використання audience persona pattern можуть включати освітні матеріали, маркетинг та рекламу, медичні поради, а також туризм та подорожі.

Шаблон виглядає таким чином – «Assume that I am X. Explain Y to me.» (рисунок 2.3).

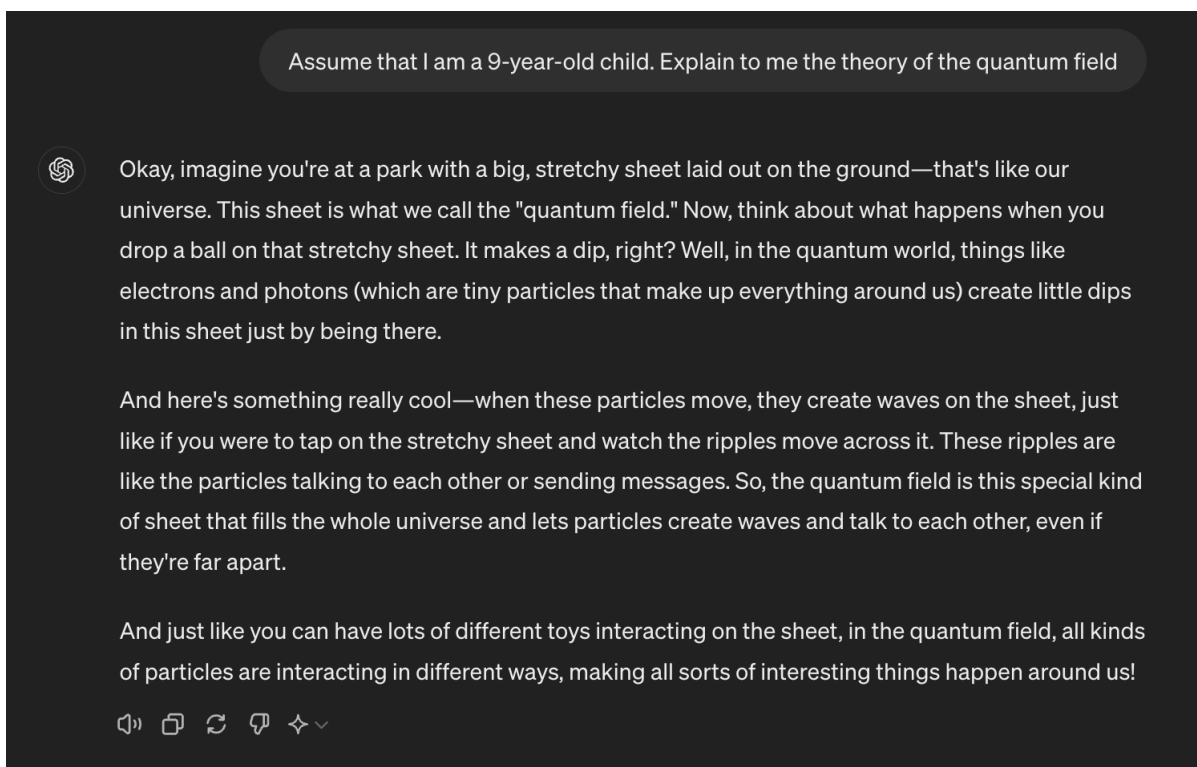


Рисунок 2.3 – Приклад використання audience persona pattern

Переваги використання audience persona pattern полягають у забезпеченні релевантності контенту, залученні користувачів та підвищенні їх задоволеності через адаптовані відповіді. Проте, є і виклики, пов'язані зі складністю налаштування, динамічністю аудиторії та необхідністю балансування між надто загальними та надто специфічними відповідями.

2.3.3 Few-shot examples pattern

Few-shot examples pattern – це техніка, що використовується для покращення якості відповідей мовної моделі шляхом надання їй декількох прикладів того, як повинні виглядати правильні відповіді. Цей підхід дозволяє моделі зрозуміти очікуваний формат, стиль і контекст відповідей,

навіть якщо модель спочатку не має достатньо інформації для правильного розуміння запиту.

Few-shot examples pattern полягає у включенні до запиту кількох прикладів (зазвичай від одного до п'яти) перед основним питанням або запитом. Ці приклади служать орієнтирами для моделі, демонструючи, яким чином слід відповідати. Такий підхід допомагає моделі краще зрозуміти завдання і генерувати більш релевантні та точні відповіді.

Цей підхід особливо корисний у випадках, коли модель має справу зі складними або спеціалізованими запитом. Few-shot examples pattern дозволяє моделі адаптуватися до нового завдання, навіть якщо вона не має значного досвіду в цій галузі. Це досягається шляхом надання прикладів, які моделюють поведінку, очікувану від моделі.

Процес використання few-shot examples pattern починається з підбору релевантних прикладів, що ілюструють правильний підхід до вирішення завдання. Ці приклади мають бути чіткими, зрозумілими та відповідати стилю і формату очікуваних відповідей. Після цього приклади включаються у запит, створюючи контекст для моделі. Наприклад, якщо потрібно створити опис продукту, можна включити кілька зразкових описів перед основним запитом.

Форматом такого патерну є визначення завдання з послідовними прикладами питань-відповідей, який закінчується певним питанням, відповідь на який має згенерувати LLM (рисунок 2.4). Тобто:

«Input: Question1.

Output: Answer1.

Input: Question2.

Output: Answer2.

Input: X.

Output:»

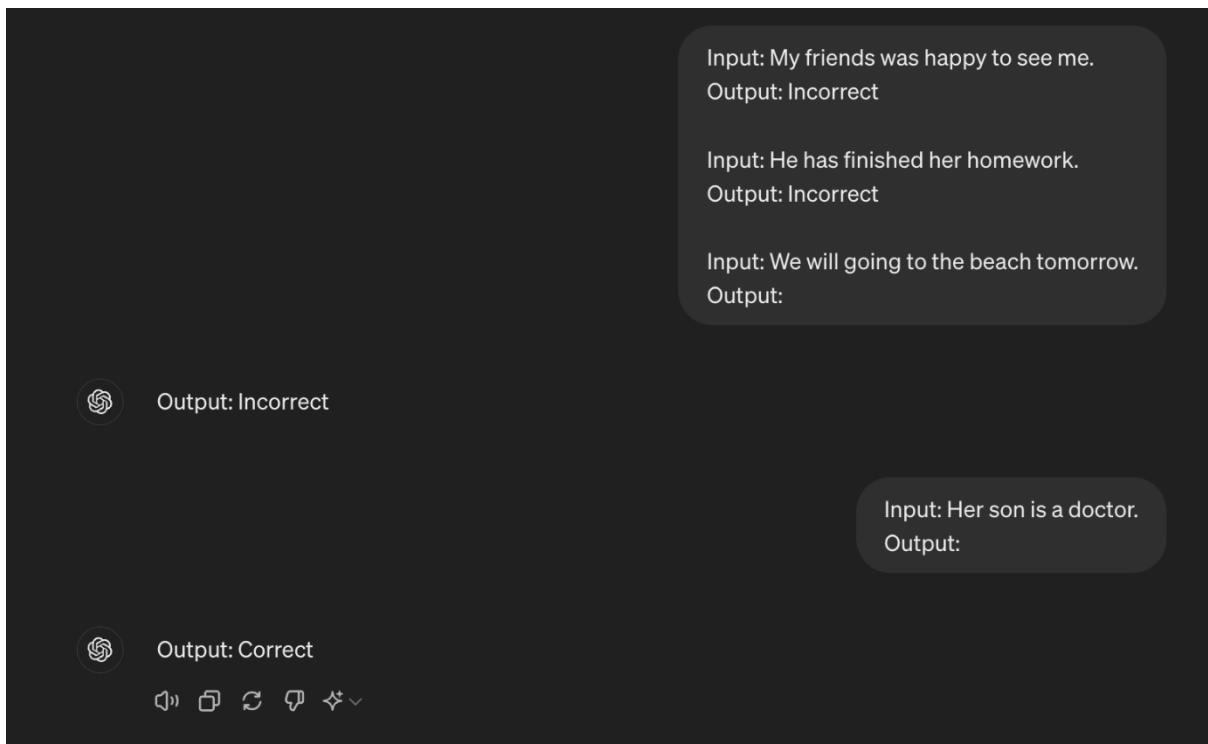


Рисунок 2.4 – Приклад використання few-shot examples pattern

Основні переваги використання few-shot examples pattern полягають у підвищенні точності та релевантності відповідей, покращенні адаптації моделі до нових завдань і зменшенні кількості помилок. Однак, цей підхід також має певні виклики, такі як необхідність ретельного підбору прикладів та потенційні обмеження у випадках, коли приклади не повністю відповідають новому контексту.

2.4 Проблема порівняння текстів

Порівняння текстів є фундаментальною проблемою у багатьох галузях, включаючи обробку природної мови (NLP), інформаційний пошук, машинне навчання та аналіз даних. У своїй суті це завдання спрямоване на визначення схожості або відмінностей між текстовими документами на основі їхнього змісту. Це дозволяє вирішувати такі завдання, як виявлення дублікатів, плагіату, класифікація документів за темами та багато інших.

Порівняння текстів стикається з низкою складних викликів. Наприклад, синонімія та полісемія створюють серйозні перешкоди. Різні слова можуть мати однакове значення, як-от «автомобіль» і «машина», що вимагає врахування синонімів при порівнянні текстів. Водночас, одне слово може мати кілька значень залежно від контексту, як у випадку зі словом «банк», яке може означати фінансову установу або берег річки. Ці аспекти ускладнюють точне визначення схожості текстів.

Контекстуальність є ще одним важливим фактором. Значення слів і речень сильно залежать від контексту, що робить порівняння текстів складним завданням. Наприклад, омоніми. Морфологічна та граматична різноманітність також впливають на точність порівняння. Складність мови, включаючи ідіоми, метафори та інші мовні конструкції, додає ще один рівень складності.

Існує кілька методів, що показали певні результати у вирішенні цієї проблеми. Глобально вони поділяються на лексичні і статистичні методи та семантичні методи.

Деякі лексичні та статистичні методи вже були описані вище для вирішення проблеми визначення емоційності тексту. До них відносяться метод N-грам, TF-IDF та інші. Також до таких методів відноситься метод косинусу подібності.

2.4.1 Косинус подібності

Косинус подібності – це широко застосовувана метрика в інформаційному пошуку та пов'язаних з ним дослідженнях. Ця метрика моделює текстовий документ як вектор термінів. За цією моделлю схожість між двома документами може бути отримана шляхом обчислення значення косинуса між векторами термінів двох документів [23]. Реалізація цієї метрики може бути застосована до будь-яких двох текстів (речення, абзацу або цілого документа). У випадку з пошуковими системами значення

схожості між запитом користувача та документами сортуються від найбільшого до найменшого. Чим вищий показник схожості між вектором термінів документа і вектором термінів запиту, тим більша релевантність між документом і запитом.

Косинус подібності рахується за формулою:

$$\cos \theta = \frac{A \cdot B}{\|A\| \|B\|}, \quad (2.4)$$

де $A \cdot B$ – скалярний добуток векторів A і B ,

$\|A\|$ і $\|B\|$ – довжини векторів A і B .

Значення косинусу подібності лежить в межах від -1 до 1. 1 означає, що вектори повністю однакові в напрямку. 0 означає, що вектори ортогональні (немає жодної подібності). -1 означає, що вектори повністю протилежні.

Косинус подібності для вимірювання схожості між документом і запитом користувача повинна враховувати значення слова. Однак вона все ще не може ідеально впоратися зі смисловим значенням тексту. Вимірювання косинусу подібності між двома векторами термінів синтаксично іноді дає ненадійний результат. Синтаксичний збіг може бути не в змозі вирішити проблему розбіжності семантичного значення. Для подальшого процесу це може дати хибний результат і призвести до погіршення продуктивності.

Проте лексичні методи порівняння текстів мають серйозні обмеження через свою нездатність враховувати контекст, синонімію, полісемію та морфологічні варіації. Ці недоліки роблять їх менш ефективними для багатьох завдань у порівнянні з семантичними методами, які можуть подолати ці обмеження, забезпечуючи більш точне і глибоке розуміння тексту.

2.4.2 Word2Vec

Семантичне порівняння текстів – це процес оцінювання схожості між текстами на основі їхнього змісту та значення, а не просто лексичних збігів. Основна мета семантичного порівняння полягає у виявленні того, наскільки схожими є тексти з точки зору їхньої інформаційної складової та сенсу, який вони передають. Це відрізняється від простого порівняння текстів за допомогою пошуку збігів слів або фраз, оскільки враховує контекст і взаємозв'язок між словами.

Одним із прикладів методів семантичного порівняння текстів є Word2Vec. Word2Vec є інноваційним методом, що трансформувал галузь обробки природної мови (NLP) завдяки здатності перетворювати слова у багатовимірні вектори, які відображають їх семантичні властивості. Розроблений дослідниками з Google, цей метод дозволяє моделювати слова так, що вони можуть бути розташовані у векторному просторі, де схожі слова мають близькі представлення. Це значно покращило ефективність алгоритмів NLP, зробивши їх більш точними і контекстуально чутливими.

Word2Vec використовує два основні підходи: CBOW (Continuous Bag of Words) і Skip-Gram. У моделі CBOW задача полягає у передбаченні цільового слова на основі його контексту, тобто навколишніх слів. Наприклад, для речення «кіт сидить на килимі» модель передбачатиме слово «сидить» на основі слів «кіт», «на» і «килимі». Цей підхід працює швидше і більш ефективно на великих обсягах даних. Натомість модель Skip-Gram працює навпаки, передбачаючи контекстуальні слова на основі цільового слова. Це дозволяє Skip-Gram краще обробляти рідкісні слова, оптимізуючи векторні представлення для передбачення контексту.

Навчання моделей Word2Vec є відносно простим процесом. Вхідні дані представляються у вигляді текстового корпусу, який попередньо обробляється для токенизації і очищення від шуму. Кожне слово у словнику представляється як one-hot вектор, де позиція відповідного слова

позначається як 1, а всі інші позиції – як 0. Нейронна мережа має вхідний шар, прихований шар і вихідний шар (рисунок 2.5), де оптимізація вагів за допомогою зворотного поширення помилки і градієнтного спуску дозволяє отримати векторні представлення слів у прихованому шарі [24].

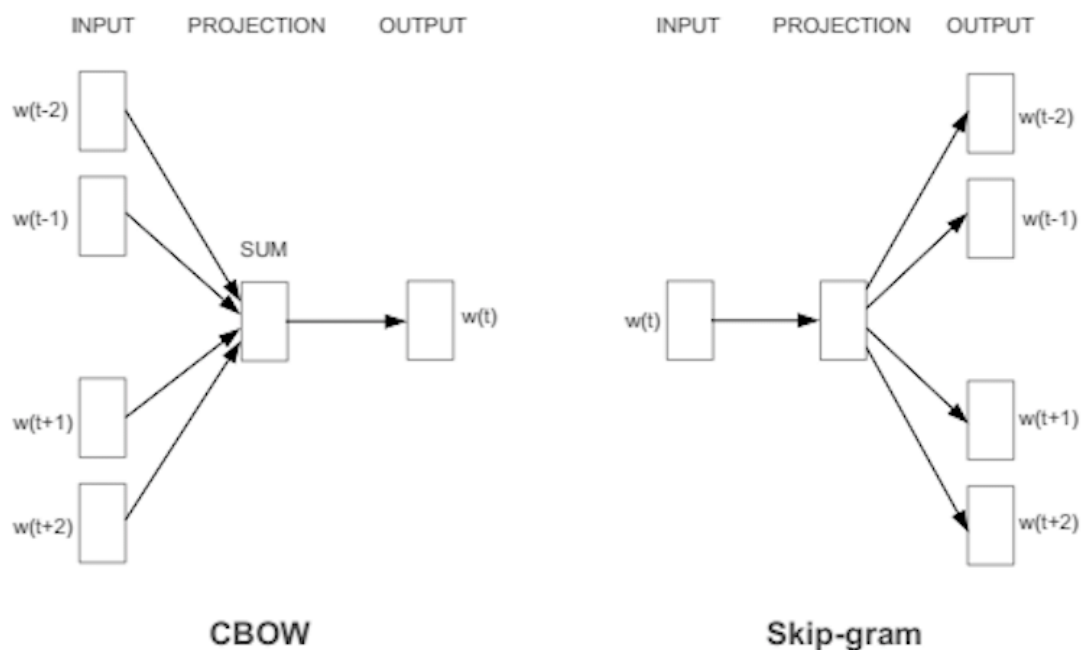


Рисунок 2.5 – Архітектура Word2Vec

Переваги Word2Vec очевидні. По-перше, цей метод забезпечує семантичну близькість між словами: слова, які використовуються у подібних контекстах, мають схожі вектори. Наприклад, слова «король» і «королева» будуть близькими у векторному просторі, відображаючи їхню семантичну спорідненість. Крім того, Word2Vec є обчислювально ефективним, що дозволяє застосовувати його до великих текстових корпусів. Відносна простота реалізації також робить цей метод популярним вибором серед дослідників і практиків у галузі NLP.

Однак, Word2Vec не позбавлений недоліків. Одним з головних обмежень є те, що він створює статичні вектори для слів, тобто значення слова не змінюється залежно від контексту. Це може бути проблемою для

слів з багатьма значеннями. Крім того, модель може мати упередження до частотних слів у тренувальному корпусі, що може вплинути на якість векторів. Для отримання якісних векторів потрібен великий обсяг текстових даних для навчання.

2.4.3 Sentence-BERT

Так як семантичне порівняння текстів – це задача NLP, то трансформери є ще одним ефективним методом вирішення такої задачі, наприклад, BERT.

BERT використовує перехресне кодування: Два речення передаються на мережу-трансформатор і прогнозується цільове значення. Однак така схема не підходить для різних завдань парної регресії через занадто велику кількість можливих комбінацій. Пошук у колекції з $n = 10\,000$ речень пари з найвищою схожістю вимагає від BERT $n - (n - 1) / 2 = 49\,995\,000$ обчислень виводу. На сучасному графічному процесорі V100 це займає близько 65 годин. Аналогічно, визначення того, яке з понад 40 мільйонів існуючих запитань на веб-сайті Quora є найбільш схожим для нового запитання, можна змоделювати як попарне порівняння за допомогою BERT, однак відповідь на один запит потребуватиме понад 50 годин.

Поширеним методом кластеризації та семантичного пошуку є зіставлення кожного речення з векторним простором так, щоб семантично схожі речення були близькими. Дослідники почали вводити окремі речення в BERT і отримувати вставки речень фіксованого розміру. Найпоширенішим підходом є усереднення вихідного шару BERT (відомого як вставки BERT) або використання виходу першої лексеми (лексеми [CLS]). Проте ця поширена практика дає досить погані вставки речень [25].

Щоб вирішити цю проблему, була розроблена Sentence-BERT або SBERT. Архітектура цієї мережі (рисунок 2.6) дозволяє отримувати вектори фіксованого розміру для вхідних речень.

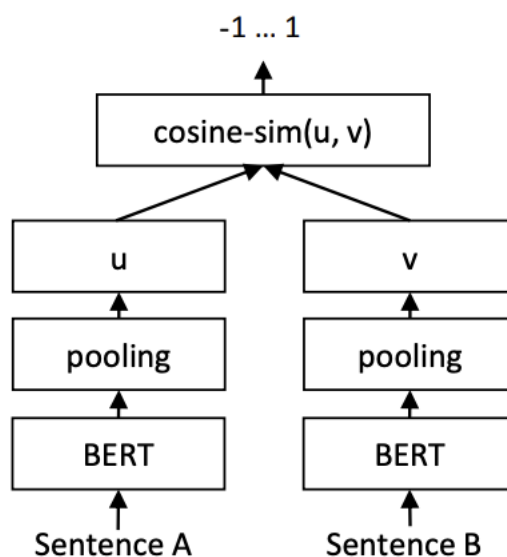


Рисунок 2.6 – Архітектура Sentence-BERT для задачі семантичного порівняння

Використовуючи міру подібності, таку як косинус подібності або Евклідову відстань, можна знайти семантично схожі речення. Ці міри подібності можуть виконуватися надзвичайно ефективно на сучасному обладнанні, що дозволяє використовувати SBERT як для пошуку семантичної подібності, так і для кластеризації. Складність пошуку найбільш схожої пари речень у колекції з 10 000 речень зменшується з 65 годин з BERT до обчислення 10 000 вкладених речень (~5 секунд з SBERT) і обчислення косинусу подібності (~0,01 секунди). Використовуючи оптимізовані структури індексів, пошук найбільш схожих запитань Quora можна скоротити з 50 годин до кількох мілісекунд [26].

3 ЕКСПЕРИМЕНТАЛЬНІ ДОСЛІДЖЕННЯ

3.1 План виконання експерименту

Основною задачею цього експеримент є визначення, наскільки успішно GPT модель здатна виявляти та усувати емоційні аспекти у текстах різного характеру. Для цього треба провести наступні етапи експерименту:

- підготовка даних, які мають яскраво виражене емоційне забарвлення з різними базовими емоціями;
- оцінка рівня емоційного забарвлення даних, на яких проводитиметься експеримент;
- основна частина експерименту, тобто генерація нейтралізованого від емоцій тексту;
- оцінка рівня емоційного забарвлення даних, що отримані в результаті основної частини експерименту;
- оцінка рівня семантичної схожості початкового і переробленого текстів;
- формування висновків на основі отриманих результатів.

3.2 Підготовка даних

Для ефективного оцінювання нейтралізації емоцій у тексті за допомогою моделі GPT-4-turbo було підготовлено спеціалізований корпус даних, що складається з 240 текстів англійською мовою. Приклад декількох з них можна побачити на рисунку 3.1. Кожен з цих текстів містить 3–5 речень і виражає одну з шести базових емоцій: радість, смуток, страх, гнів, здивування та огиду. Для цього модель була запущена для генерації прикладів під кожен емоцію окремо.

- 1) It was completely unexpected when a judge ruled that an oral agreement made over a decade ago, and only witnessed by a single individual, constituted a binding contract in a multi-million dollar real estate dispute, significantly altering the expected outcome of the case.
- 2) There are rumors of biological weapons being developed by rogue states. The possibility of facing an unseen, airborne agent that could devastate populations is a nightmarish thought that keeps me up at night.
- 3) Every piece of equipment must be accounted for and in prime condition. I don't care if it's a bullet or a bulldozer; if it's not in perfect working order, then you're not only failing yourself, you're failing your brothers and sisters in arms!

Рисунок 3.1 – Приклад згенерованих даних

Для забезпечення різноманітності контекстів та тематики тексти були розподілені по чотирьох основних темах: політика, журналістика, військова справа та юриспруденція. Ці теми були обрані, адже питання нейтралізації емоцій в кожній з них є актуальним. Кожна з цих тем містить по 60 текстів, рівномірно розподілених за шістьма базовими емоціями. Такий підхід дозволяє оцінити ефективність нейтралізації емоцій у різних контекстах та забезпечує об'єктивність результатів дослідження.

Структура текстів була обрана таким чином, щоб забезпечити достатню кількість контексту для моделі GPT-4-turbo для виявлення та нейтралізації емоцій. Кожен текст містить 3–5 речень, що робить їх достатньо компактними для аналізу, але водночас забезпечує достатній обсяг інформації для вираження емоційного забарвлення.

Підготовлений корпус даних дозволяє всебічно оцінити здатність моделі GPT-4 нейтралізувати емоційний компонент у текстах різної тематики. Використання збалансованих даних забезпечує об'єктивність результатів та дозволяє зробити вагомі висновки щодо ефективності та універсальності підходу.

3.3 Рівень емоційності текстів для експерименту

Для оцінки рівня емоційності текстів, що були згенеровані на попередньому етапі, було вирішено скористатись моделлю j-

hartmann/emotion-english-distilroberta-base. Ця модель зроблена на основі DistilRoBERTa для класифікації емоцій в англomовних текстах. Вона передбачає шість базових емоцій за Екманом – гнів, відразу, страх, радість, смуток і здивування, а також нейтральний клас. Ця модель була навчена на шести різних наборах даних, що включають тексти з Twitter, Reddit і діалоги з телебачення, що забезпечує широкий спектр типів тексту для надійного визначення емоцій. Для використання моделі було використано бібліотеку transformers на Python від Hugging Face.

Ця модель на виході віддає 7 числових значень від 0 до 1, кожен з яких означає рівень певної емоції в конкретному тексті. Приклад використання можна побачити на рисунку 3.2.

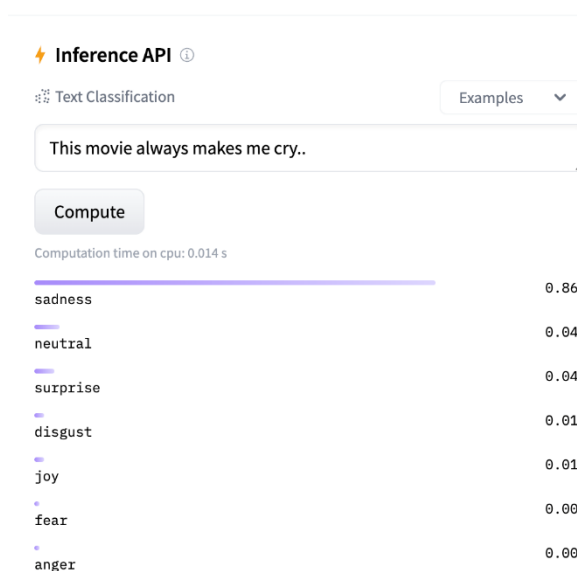


Рисунок 3.2 – Приклад використання моделі j-hartmann/emotion-english-distilroberta-base

Експериментальні дані були оцінені цією моделлю. Кожен текст, що був згенерований під конкретну емоцію, має найвищим показником саме цю емоцію. Наприклад, текст «The state of our national infrastructure is a disgrace – an utter mockery of what we promise our citizens! Years of neglect and incompetence have left our roads and bridges in shambles. We demand immediate

action, not more bureaucratic tape and empty promises!», який має бути з емоцією злості як найбільш вираженішою, був оцінений моделлю так, як зображено на рисунку 3.3.

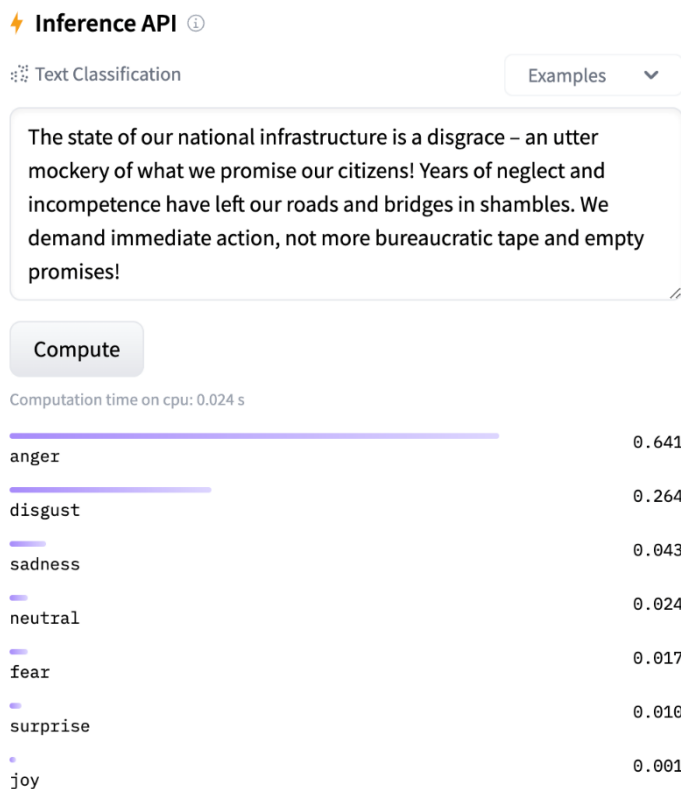


Рисунок 3.3 – Приклад оцінки рівня емоційності експериментальних даних

У результаті до початку основної частини експерименту маємо наступний розподіл рівня емоційності згенерованих текстів, як зображено на рисунку 3.4. На ньому зображено шість графіків – по одному на кожну базову емоцію: гнів, радість, смуток, страх, здивування, огида. На кожному графіку присутні дві криві. Синя – це рівень емоції для усіх 40 прикладів тексту, де переважає ця певна емоція. В свою чергу помаранчева лінія – це рівень нейтральності тих самих текстів. Як можна бачити, більшість прикладів мають високий рівень її основної емоції і відповідно дуже низький рівень нейтральності.

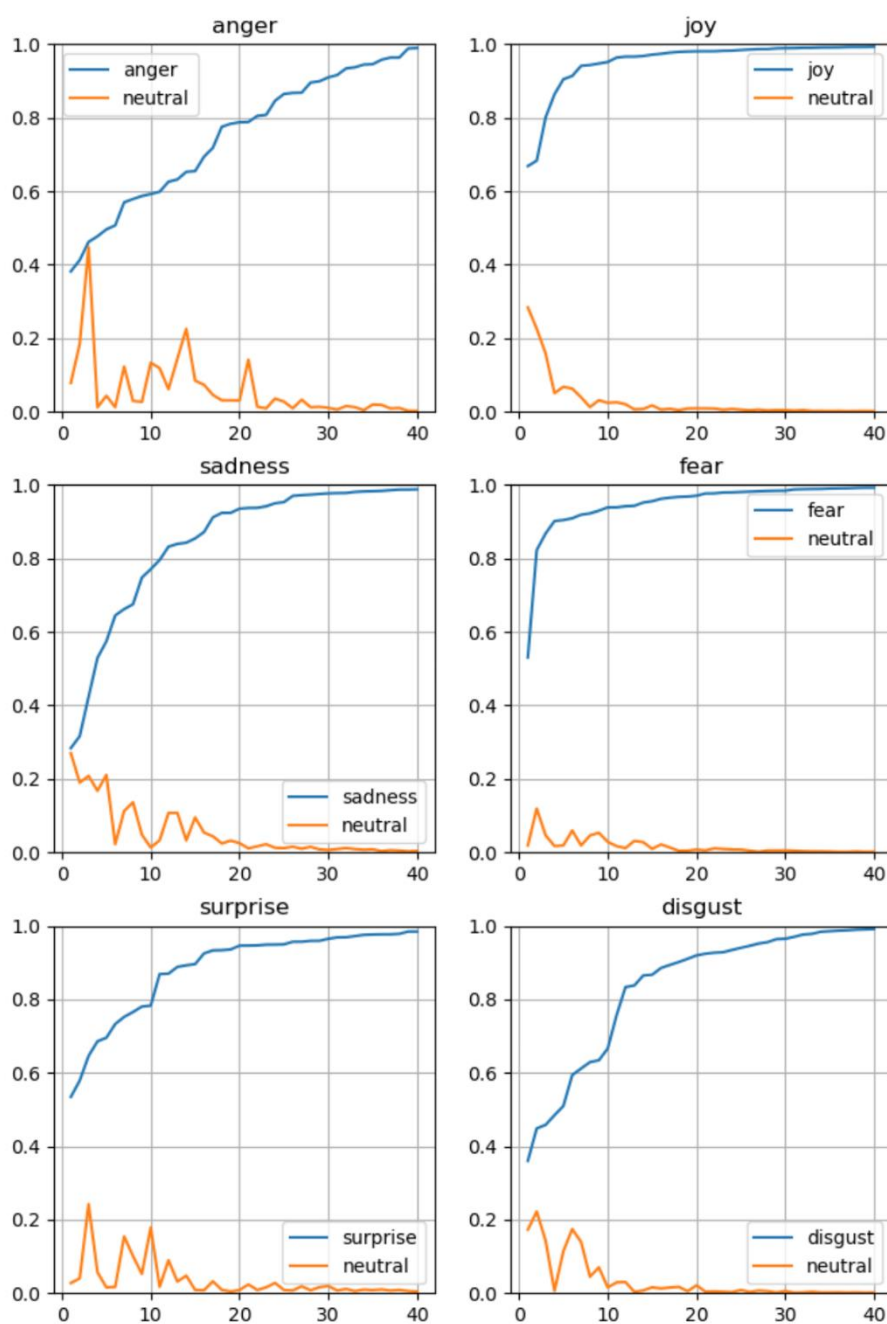


Рисунок 3.4 – Рівень емоційного забарвлення основної емоції і рівень нейтральності початкових даних

3.4 Нейтралізація емоцій

Експеримент проводився на моделі GPT-4-turbo за допомогою OpenAI API.

У процесі проведення експерименту з нейтралізації тексту від емоцій за допомогою моделі GPT використовувалися різні шаблони вхідних даних. Цей підхід був спрямований на визначення найбільш ефективних методів для досягнення нейтрального тону в текстах, насичених емоційним забарвленням. Основний експеримент базувався на використанні чотирьох різних типів промптів, серед яких три є шаблонами prompt engineering.

Перший промпт передбачає пряме формулювання завдання для моделі GPT. Для цієї задачі було використано запит сформований наступним чином: «You need to rewrite the given text so that it is emotionless, but with the same semantic load. <Text>». Цей підхід дозволяє оцінити базову здатність моделі справлятися з завданням нейтралізації емоцій без будь-якої додаткової інформації чи контексту. Приклад такого використання можна побачити на рисунку 3.5.

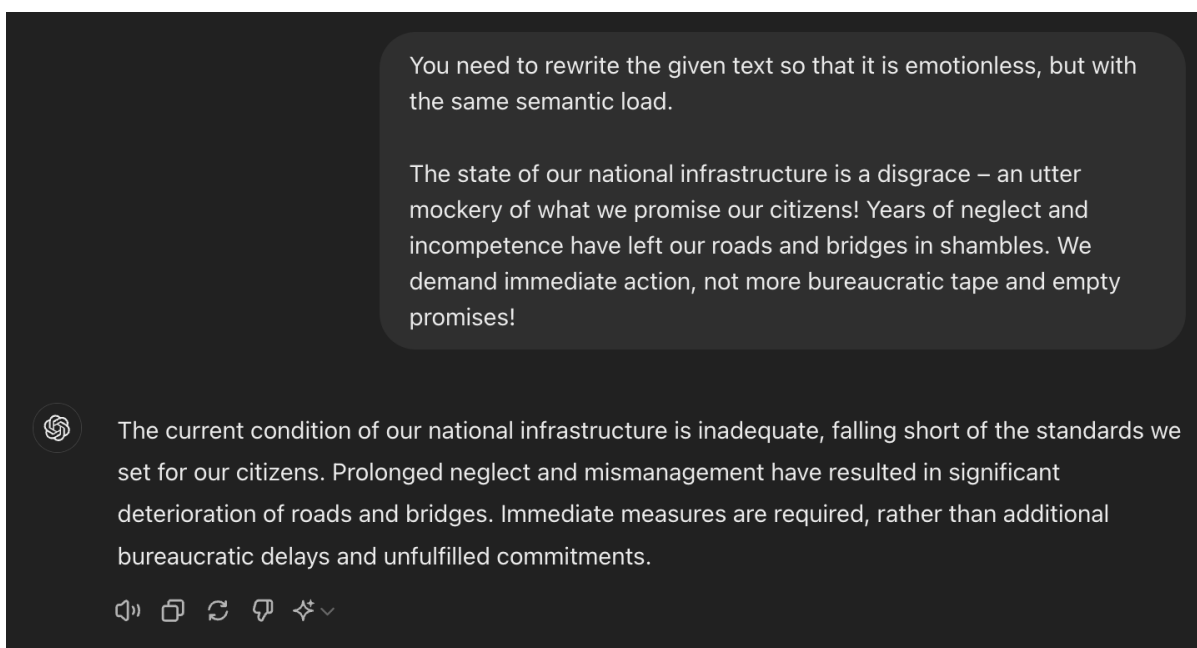


Рисунок 3.5 – Приклад використання промпту з прямим формулюванням завдання

Другий шаблон, persona pattern, включає надання моделі інформації про уявного автора тексту або його персональні характеристики. Цей

шаблон допомагає моделі краще зрозуміти контекст та особливості мови, якою користується автор, що сприяє більш точному виконанню завдання з нейтралізації емоцій. Для цієї задачі було використано наступне формулювання шаблону: «Act as a person who is objective and unbiased who does not show emotion. Rewrite the given text. <Text>». Приклад використання такого шаблону можна побачити на рисунку 3.6.

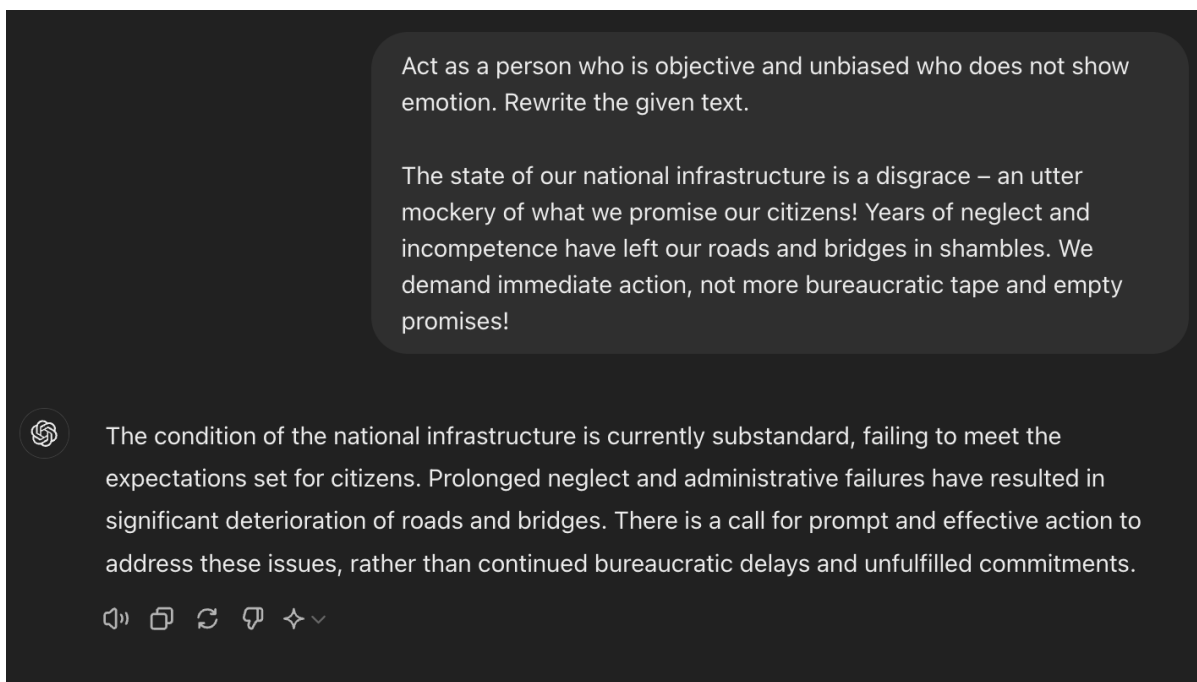


Рисунок 3.6 – Приклад використання persona pattern

Третій шаблон, audience persona pattern, фокусується на характеристиках цільової аудиторії тексту. У цьому випадку модель отримує інформацію про очікування та емоційні потреби аудиторії, для якої призначений текст. Такий підхід дозволяє моделі коригувати емоційний тон тексту відповідно до очікувань та реакцій цієї аудиторії, сприяючи досягненню більш нейтрального тону. Для цієї задачі було використано наступне формулювання шаблону: «Rewrite the given text. Assume I am a person who will use this text in my work and my work requires neutrality and

impartiality. <Text>». Приклад використання такого шаблону можна побачити на рисунку 3.7.

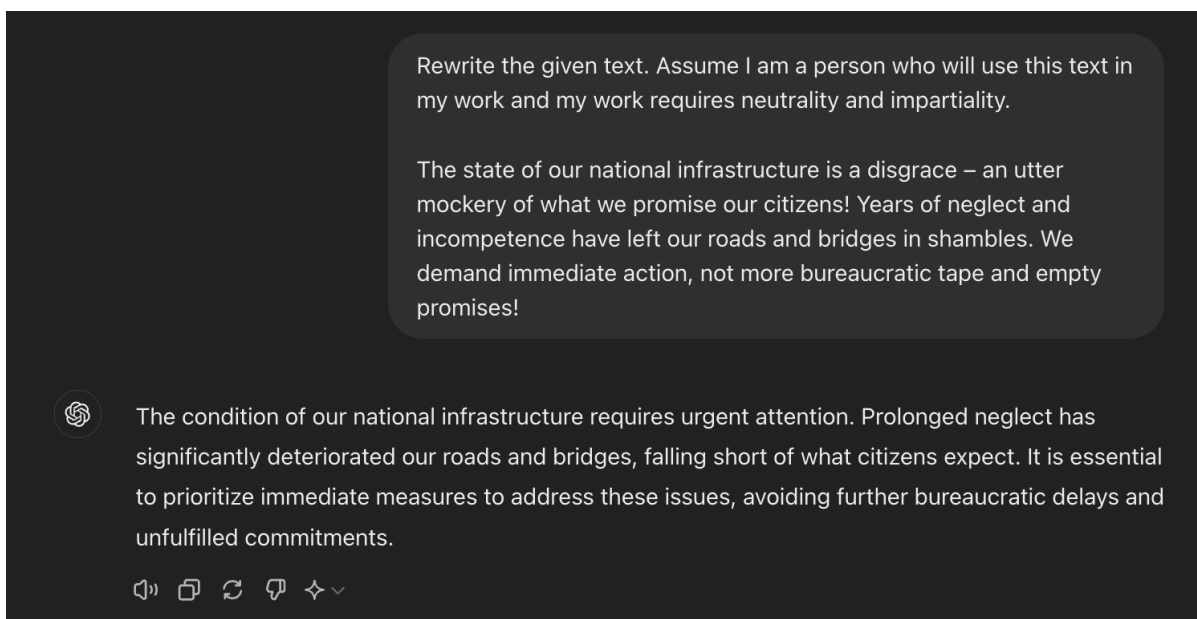


Рисунок 3.7 – Приклад використання audience persona pattern

Четвертий шаблон, few-shot examples pattern, передбачає надання моделі декількох прикладів текстів із вже нейтралізованими емоціями. Ці приклади служать орієнтиром для моделі, допомагаючи їй краще зрозуміти завдання та очікуваний результат. Використання таких прикладів підвищує точність та ефективність моделі у процесі нейтралізації емоційного забарвлення текстів.

Для цієї задачі було використано наступне формулювання few-shot examples pattern:

«Make the text neutral.

Input: How dare they delay the project again? This is utterly unacceptable! Every single time, it's the same ridiculous excuses that push us further from our goals!

Output: The project has experienced another delay. This has been a common occurrence, with similar reasons cited each time, which impacts our timeline for reaching the goals.

Input: I'm absolutely thrilled to announce that we've secured the funding we needed! This is a monumental achievement for us all. Our future projects are now fully supported, paving the way for incredible innovations. Thanks to everyone's dedication, we're set to make groundbreaking advancements. Let's keep up this fantastic momentum!

Output: The required funding has been secured, ensuring support for our upcoming projects. This development allows us to continue with planned innovations. The team's consistent effort has contributed to achieving this milestone, positioning us for future advancements.

Input: <Text>

Output:».

Приклад використання такого шаблону можна побачити на рисунку 3.8.

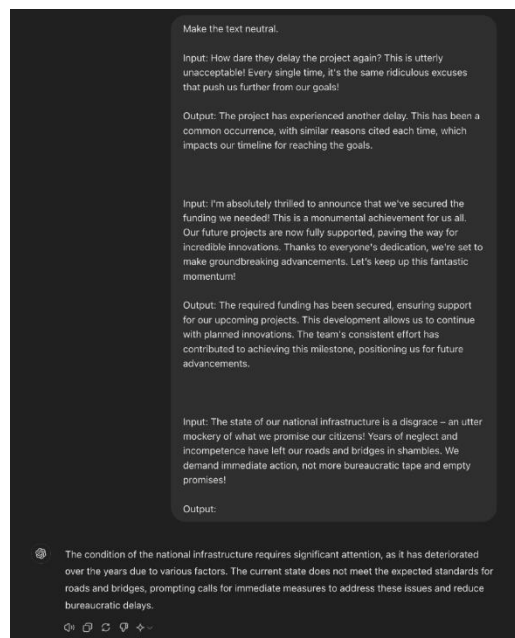


Рисунок 3.8 – Приклад використання шаблону few-shot examples

Застосування різних шаблонів prompt-engineering у рамках експерименту дозволило всебічно оцінити здатність моделі GPT

нейтралізувати емоції в текстах, а також визначити найбільш ефективні підходи для досягнення бажаного результату. Отримані результати сприяють покращенню методології обробки природної мови та можуть бути використані у подальших дослідженнях та практичних застосуваннях.

3.5 Рівень емоційності текстів, отриманих у ході експерименту

Загалом в результаті було згенеровано по чотири результати для кожної з 240 рядків експериментальних даних – по одному на кожен шаблон промптів, тобто всього 960 нових версій текстів. Кожен з них був оцінений все тією ж моделлю j-hartmann/emotion-english-distilroberta-base.

3.5.1 Нейтралізація гніву

На рисунку 3.9 можна бачити рівень гніву для тих 40 експериментальних даних, у яких гнів є найбільш вираженою емоцією. В свою чергу на рисунку 3.10 можна побачити рівень нейтральності тих самих прикладів відповідно.

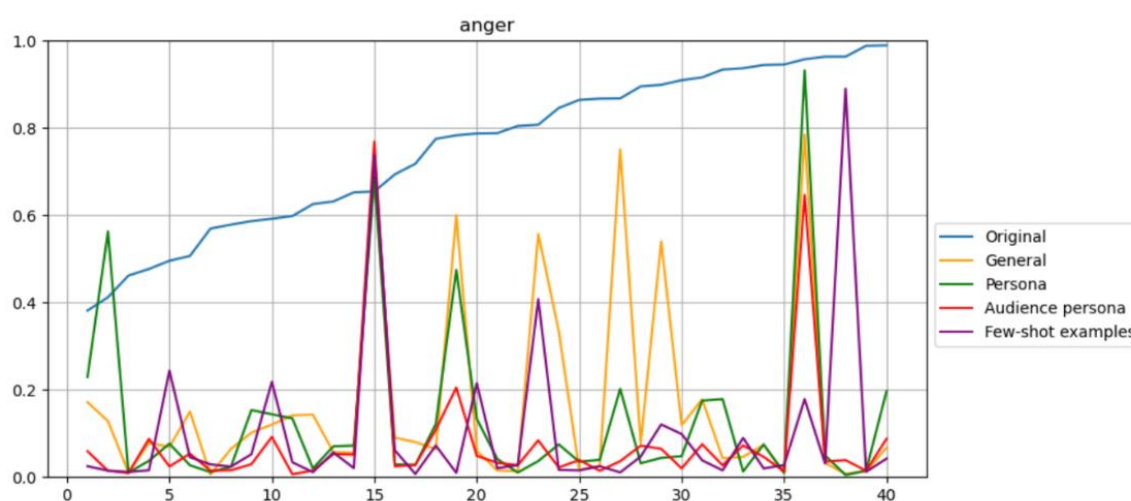


Рисунок 3.9 – Рівень гніву до та після нейтралізації

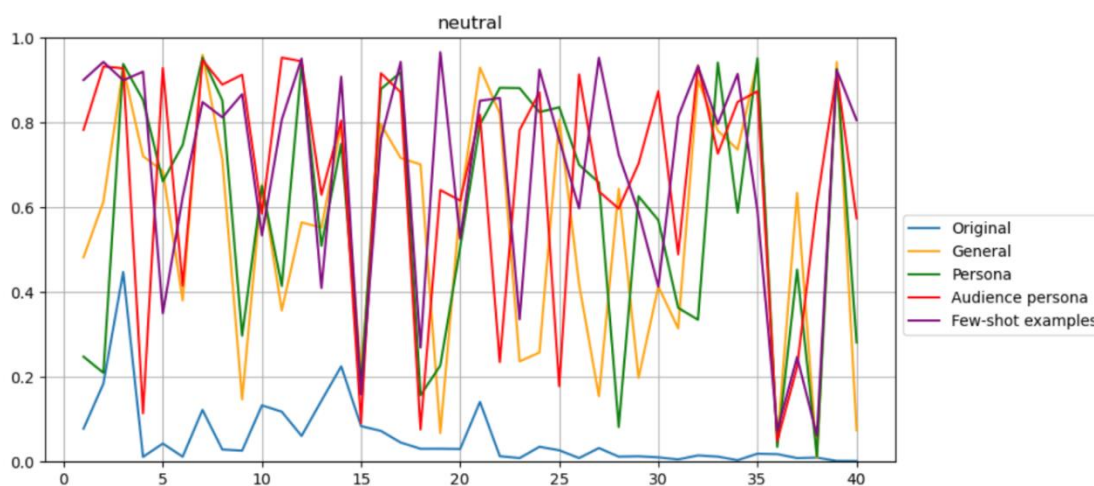


Рисунок 3.10 – Рівень нейтральності до та після нейтралізації гніву

Отже GPT-модель впоралась з нейтралізацією емоції гніву у 39 з 40 випадків. У випадку Persona pattern – у 38 випадках.

Наявність піків на рисунку 3.9 пояснюється тим, що у деяких випадках емоція настільки тісно пов'язана з семантикою тексту, що його складно нейтралізувати, незалежно від вибраного шаблону промптів.

Отже, відповідаючи на запитання, чи може GPT-модель нейтралізувати емоції гніву, то відповідь ствердна, до того ж з використанням будь-якого із запропонованих шаблонів. Проте щоб дізнатись, який промпт дає кращі показники, треба відсортувати дані, щоб побачити загальну тенденцію промпту вздовж всієї вибірки. Це можна побачити на рисунку 3.11.

У таблиці 3.1 чітко видно, що audience persona pattern найкраще справляється з нейтралізацією гніву, зменшуючи рівень гніву з 75% до 8%, при цьому зберігаючи найвищі показники нейтральності. Few-shot examples pattern також досяг майже таких самих значень.

При цьому промпт з прямим запитанням впорався найгірше у порівнянні з іншими шаблонами, знизивши рівень гніву до 17%.

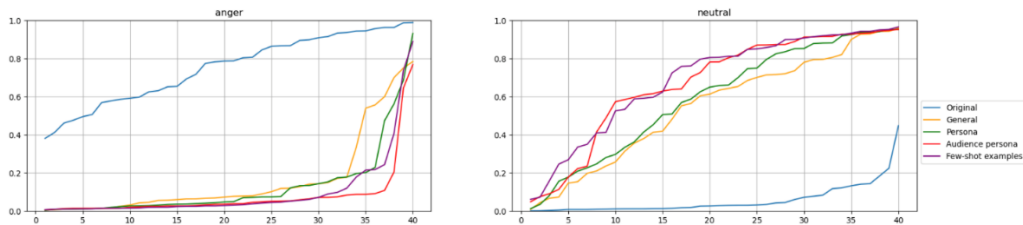


Рисунок 3.11 – Загальна тенденція нейтралізації емоції гніву

Таблиця 3.1 – Статистика результатів нейтралізації гніву

	Середній показник емоції	Мін.	Макс.	Кількість зразків, де емоція <0.1	Кількість зразків, де емоція <0.33	Кількість зразків, де емоція <0.5	Нейтр. >0.5	Нейтр. >0.67	Нейтр. >0.9
Original	0.75177	0.381583	0.989470	0	0	5	0	0	0
General	0.1651	0.006182	0.784828	24	33	34	24	17	6
Persona	0.1315	0.004671	0.931732	26	36	37	26	18	7
Audience persona	0.079515	0.006693	0.768949	36	38	38	31	23	11
Few-shot examples	0.101038	0.007252	0.890045	32	37	38	31	25	12

Шаблон з прямим запитанням має найгірші показники порівняно з шаблонами з prompt-engineering, що доводить ефективність використання спеціалізованих шаблонів для роботи з GPT-моделлю.

3.5.2 Нейтралізація радості

На рисунку 3.12 можна бачити рівень радості для тих 40 експериментальних зразків, у яких радість є найбільш вираженою емоцією.

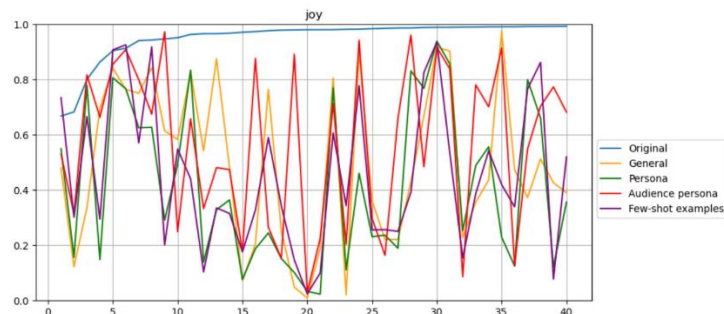


Рисунок 3.12 – Рівень радості до та після нейтралізації

В свою чергу на рисунку 3.13 можна побачити рівень нейтральності тих самих прикладів відповідно.

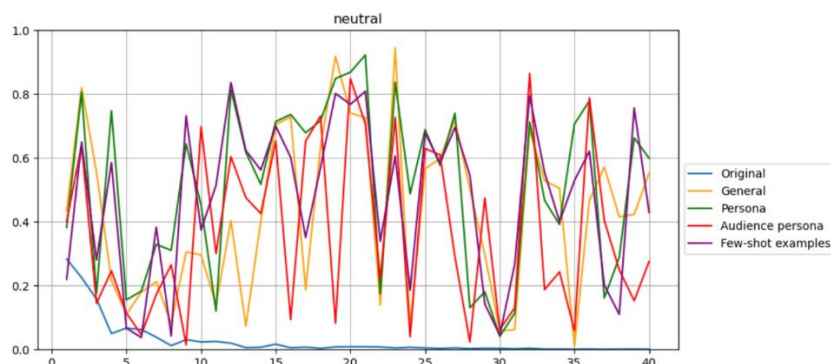


Рисунок 3.13 – Рівень нейтральності до та після нейтралізації радості

Хоча модель і змогла зменшити рівень радості і підвищити рівень нейтральності майже на усіх вхідних зразках, проте цей показник гірше, ніж при нейтралізації гніву.

Audience persona pattern для нейтралізації радості призвів до найгірших результатів у порівнянні з іншими видами промптів, навіть порівняно з шаблоном з прямим запитанням (рисунок 3.14). Шаблон Persona для нейтралізації радості підходить найкраще, адже в середньому знижує рівень емоційності з 95% до 46% (таблиця 3.2). Трошки гірший результат показав шаблон few-shot examples.

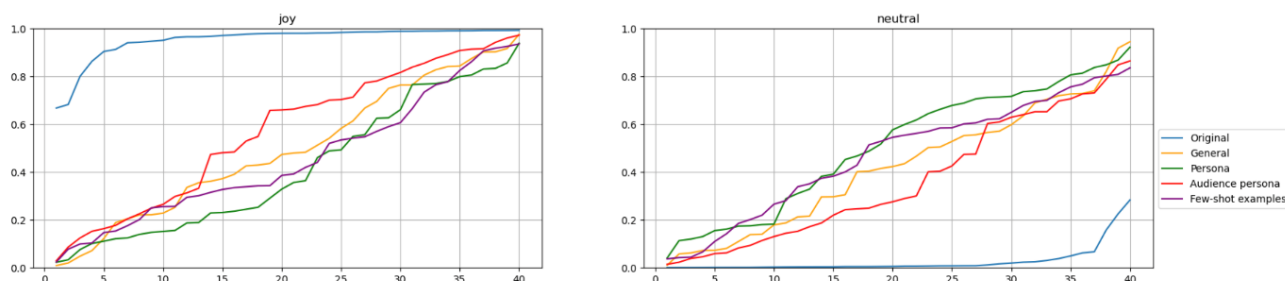


Рисунок 3.14 – Загальна тенденція нейтралізації емоції радості

Таблиця 3.2 – Статистика результатів нейтралізації радості

	Середній показник емоції	Мін.	Макс.	Кількість зразків, де емоція <0.1	Кількість зразків, де емоція <0.33	Кількість зразків, де емоція <0.5	Нейтр. >0.5	Нейтр. >0.67	Нейтр. >0.9
Original	0.952558	0.668136	0.993060	0	0	0	0	0	0
General	0.498243	0.009498	0.976154	4	11	22	18	9	2
Persona	0.417960	0.023207	0.938996	3	20	25	22	16	1
Audience persona	0.571298	0.029481	0.972928	2	12	16	13	7	0
Few-shot examples	0.455542	0.024324	0.936207	3	15	23	23	10	0

3.5.3 Нейтралізація смутку

На рисунках 3.15 і 3.16 можна побачити результати нейтралізації емоції смутку. Можна помітити, що деякі приклади, особливо ті, що створені за допомогою промпта few-shot examples, дали в результаті показник смутку вище, ніж в оригінальному тексті.

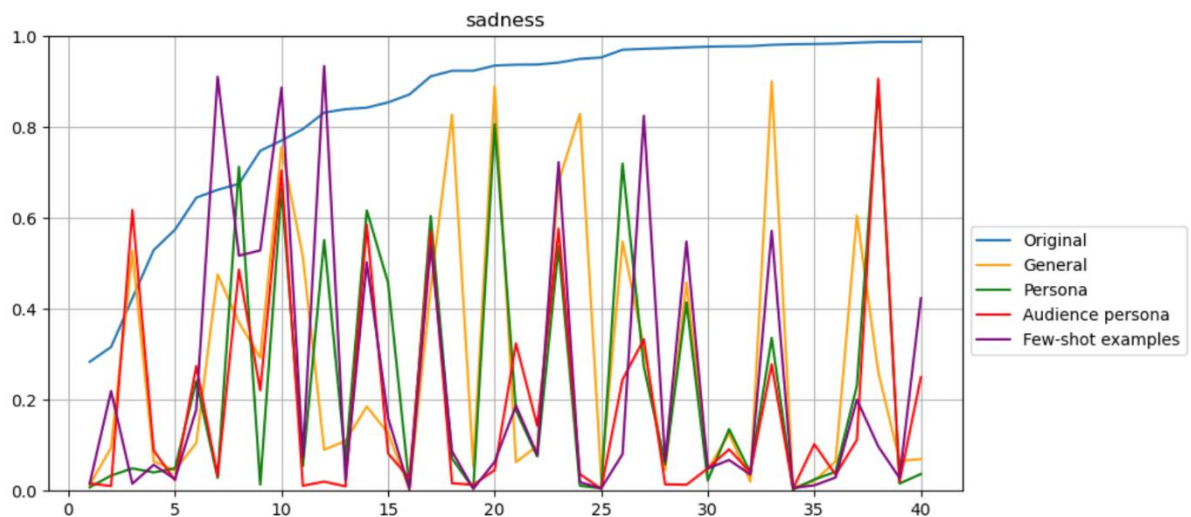


Рисунок 3.15 – Рівень смутку до та після нейтралізації

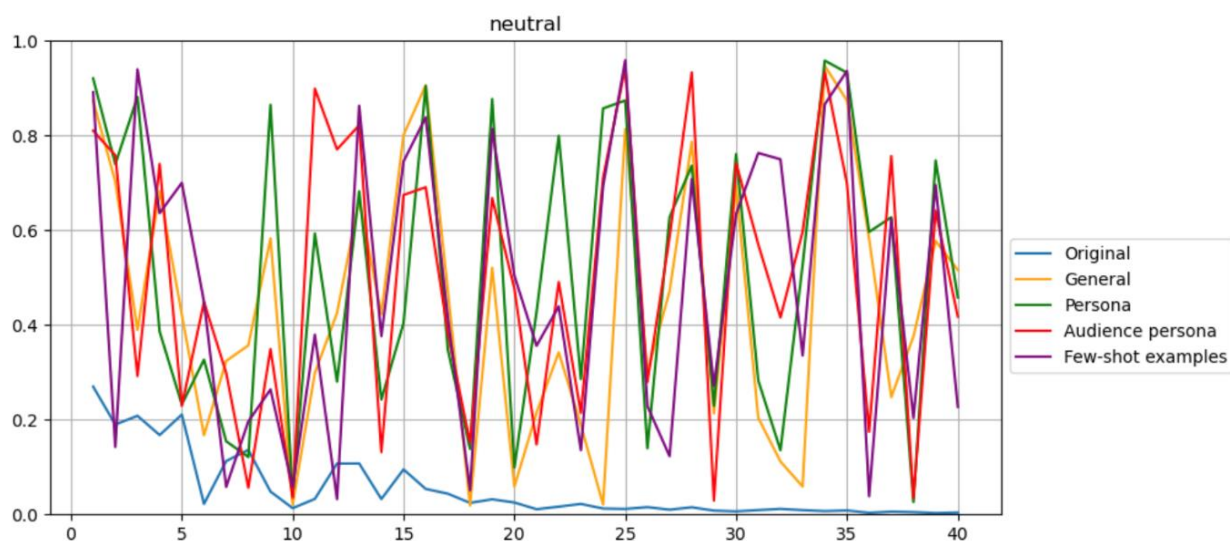


Рисунок 3.16 – Рівень нейтральності до та після нейтралізації смутку

Якщо дивитись на загальну тенденцію на рисунку 3.17, то можна зробити висновок, що модель здатна нейтралізувати текст від радості. Найкраще для нейтралізації радості підходить шаблон audience persona, яка знизила середній показник емоційності з 84% до 19%, найгірше – шаблон з прямою постановкою завдання, який знизив лише до 28% (таблиця 3.3).

Майже половина зразків «смутку» були позбавлені цієї емоції зовсім, чого не було у випадку з радістю.

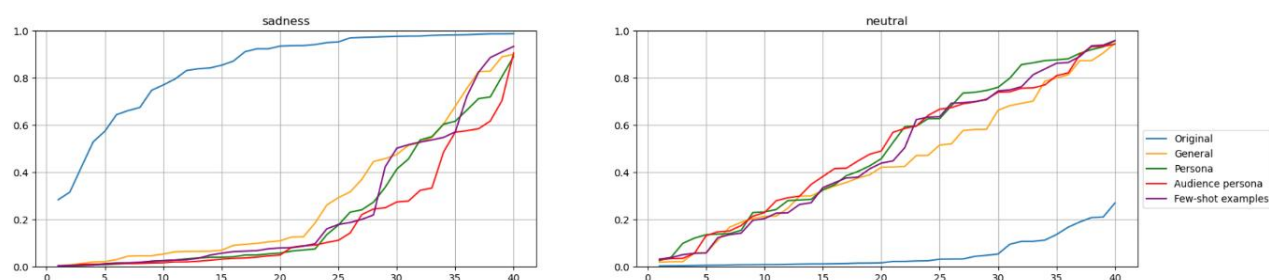


Рисунок 3.17 – Загальна тенденція нейтралізації емоції смутку

Таблиця 3.3 – Статистика результатів нейтралізації смутку

	Середній показник емоції	Мін.	Макс.	Кількість зразків, де емоція <0.1	Кількість зразків, де емоція <0.33	Кількість зразків, де емоція <0.5	Нейтр. >0.5	Нейтр. >0.67	Нейтр. >0.9
Original	0.844704	0.283845	0.988210	0	2	3	0	0	0
General	0.281322	0.002874	0.900679	18	26	30	16	10	2
Persona	0.228341	0.002936	0.891260	23	28	31	20	15	4
Audience persona	0.185935	0.004254	0.906573	23	32	34	20	15	3
Few-shot examples	0.245153	0.003782	0.934380	23	28	29	19	15	3

3.5.4 Нейтралізація страху

На рисунках 3.18 і 3.19 можна побачити таку ж ситуацію, які у випадку з нейтралізацією емоції радості.

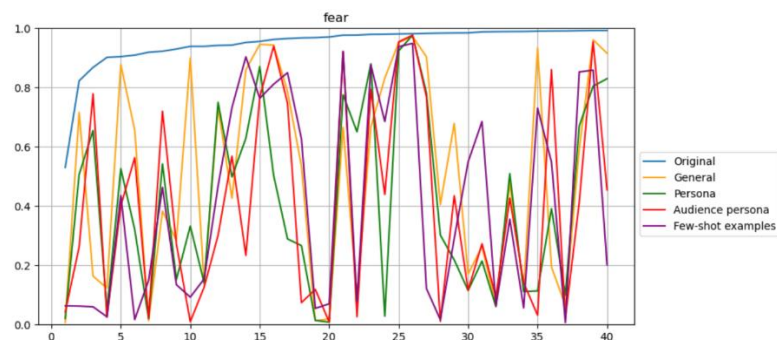


Рисунок 3.18 – Рівень страху до та після нейтралізації

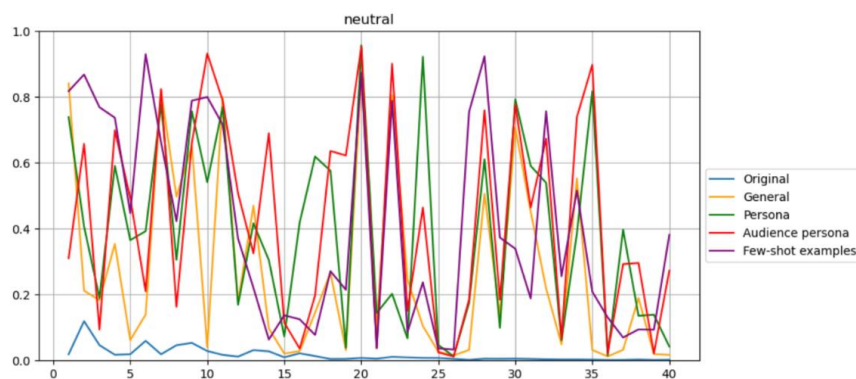


Рисунок 3.19 – Рівень нейтральності до та після нейтралізації страху

Шаблон прямої постановки завдання впорався найгірше для цієї задачі, аніж інші типи промптів (рисунок 3.20). В свою чергу audience persona pattern в більшості прикладів показує кращий результат, знижуючи середній показник емоційності страху з 95% до 40% (таблиця 3.4).

Таблиця 3.4 – Статистика результатів нейтралізації страху

	Середній показник емоції	Мін.	Макс.	Кількість зразків, де емоція <0.1	Кількість зразків, де емоція <0.33	Кількість зразків, де емоція <0.5	Нейтр. >0.5	Нейтр. >0.67	Нейтр. >0.9
Original	0.948307	0.530451	0.992639	0	0	0	0	0	0
General	0.511521	0.007527	0.973009	7	15	19	9	6	0
Persona	0.413828	0.008262	0.976886	8	19	22	15	8	2
Audience persona	0.402446	0.008445	0.976878	11	26	26	17	12	3
Few-shot examples	0.418172	0.006835	0.949329	13	19	23	15	13	2

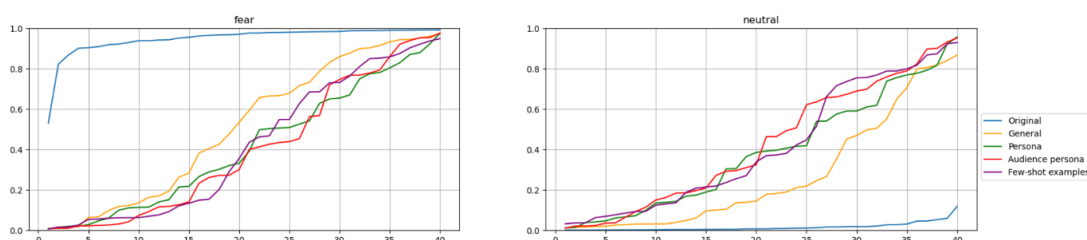


Рисунок 3.20 – Загальна тенденція нейтралізації емоції страху

3.5.5 Нейтралізація здивування

Згідно рисунків 3.21, 3.22 і 3.23 можна зробити висновок, що здивування найлегше піддається нейтралізації у порівнянні з іншими емоціями, на рівні з нейтралізацією гнів.

Усі тестовані шаблони показали майже однаковий рівень ефективності, проте persona pattern виявився найефективнішим серед усіх шаблонів, знизивши рівень емоційності у середньому з 88% до 9% (таблиця 3.5).

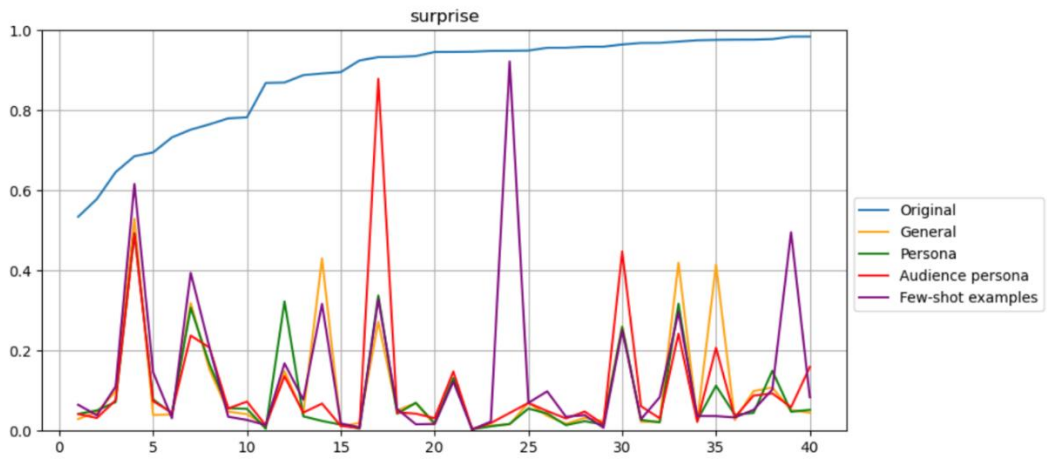


Рисунок 3.21 – Рівень здивування до та після нейтралізації

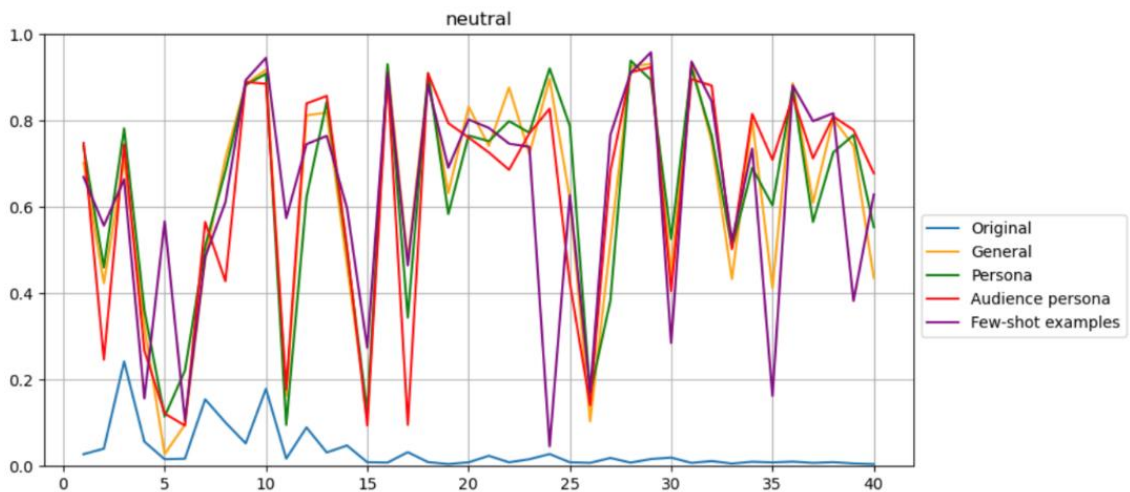


Рисунок 3.22 – Рівень нейтральності до та після нейтралізації здивування

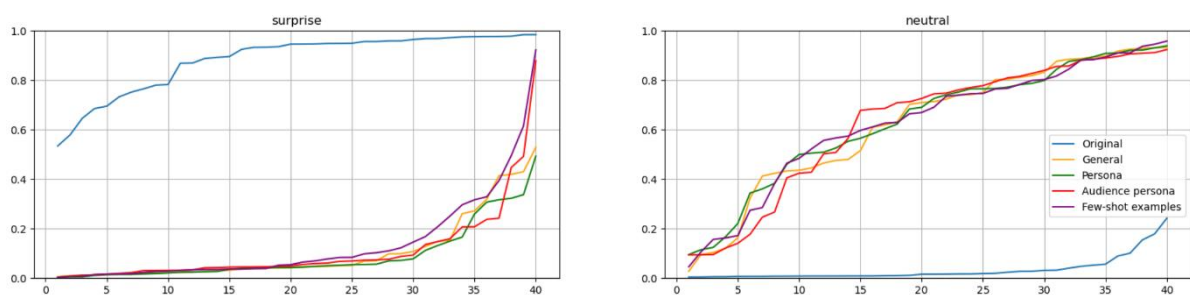


Рисунок 3.23 – Загальна тенденція нейтралізації емоції здивування

Таблиця 3.5 – Статистика результатів нейтралізації здивування

	Середній показник емоції	Мін.	Макс.	Кількість зразків, де емоція <0.1	Кількість зразків, де емоція <0.33	Кількість зразків, де емоція <0.5	Нейтр. >0.5	Нейтр. >0.67	Нейтр. >0.9
Original	0.883503	0.534378	0.984499	0	0	0	0	0	0
General	0.106411	0.006074	0.528858	29	36	39	26	22	5
Persona	0.091383	0.003873	0.493255	30	38	40	31	22	6
Audience persona	0.112263	0.003709	0.878810	30	37	39	29	26	4
Few-shot examples	0.137881	0.002966	0.922369	26	36	38	30	20	5

3.5.6 Нейтралізація огиди

Рисунки 3.24, 3.25 і 3.26 демонструють рівень нейтралізації тексту моделлю GPT від емоції огиди. У порівнянні з іншими емоціями, модель з емоцією огиди впоралась найгірше. Для цієї емоції найкраще спрацьовує audience persona pattern, який знизив середній показник емоційності з 83% до 26% (таблиця 3.6).

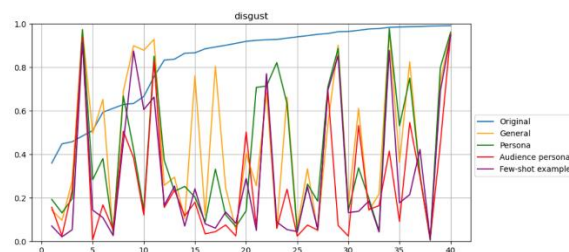


Рисунок 3.24 – Рівень огиди до та після нейтралізації

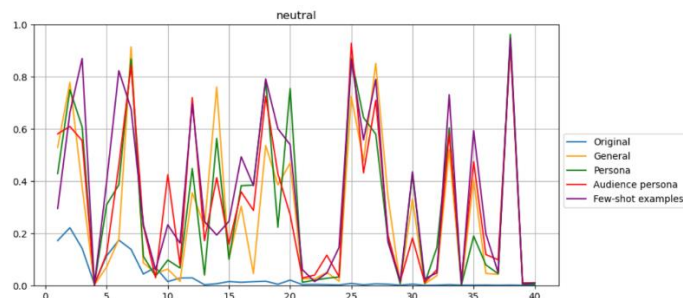


Рисунок 3.25 – Рівень нейтральності до та після нейтралізації огиди

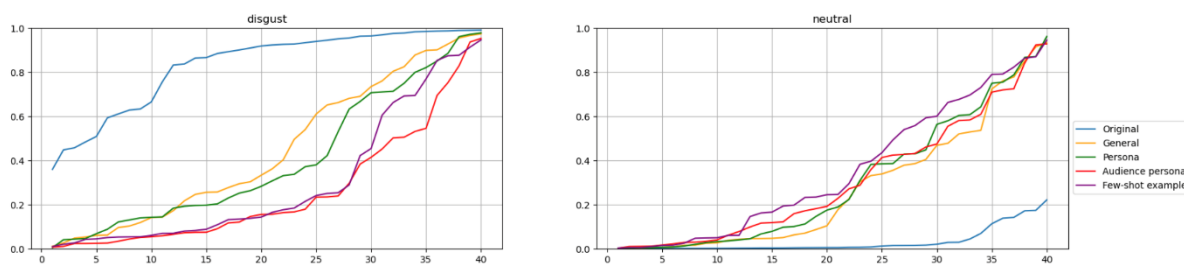


Рисунок 3.26 – Загальна тенденція нейтралізації емоції огиди

Таблиця 3.6 – Статистика результатів нейтралізації огиди

	Середній показник емоції	Мін.	Макс.	Кількість зразків, де емоція <0.1	Кількість зразків, де емоція <0.33	Кількість зразків, де емоція <0.5	Нейтр. >0.5	Нейтр. >0.67	Нейтр. >0.9
Original	0.833482	0.360153	0.991277	0	0	4	0	0	0
General	0.449290	0.011583	0.974477	7	19	23	9	6	2
Persona	0.403263	0.005612	0.978638	6	21	26	11	6	1
Audience persona	0.261933	0.007923	0.953583	16	28	31	10	6	2
Few-shot examples	0.300814	0.010537	0.947195	15	28	30	14	9	1

3.6 Оцінка семантичної подібності даних до і після нейтралізації

Для порівняння семантичної подібності між текстами до і після нейтралізації було використано модель sentence-transformers/all-MiniLM-L6-v2. Ця модель – це один із варіантів Sentence-Transformers, що базується на архітектурі MiniLM. Вона призначена для створення векторних уявлень речень, які відображають їхній семантичний зміст. Ця модель є ефективнішою і швидшою альтернативою більш великим моделям, таким як BERT або RoBERTa, завдяки своїй компактності та меншій кількості параметрів.

Sentence-transformers/all-MiniLM-L6-v2 використовує ті ж основні принципи, що й SBERT, але застосовує архітектуру MiniLM замість повної версії BERT. Це робить модель менш вимогливою до ресурсів та швидшою у виконанні, що є важливим для реальних застосунків і обмежених обчислювальних ресурсів.

Основна ідея SBERT та Sentence-Transformers полягає в тому, що вони адаптують трансформери для створення векторних уявлень, які можна легко порівнювати за допомогою стандартних метрик відстані, таких як косинусна схожість. Це дозволяє використовувати моделі для широкого спектра задач, де необхідно оцінювати семантичну подібність між текстами.

Згідно результатів, що зображені на рисунку 3.27, а також таблиці 3.7, шаблон із загальним питанням зберігає більшу частину семантики, аніж шаблони з prompt-engineering.

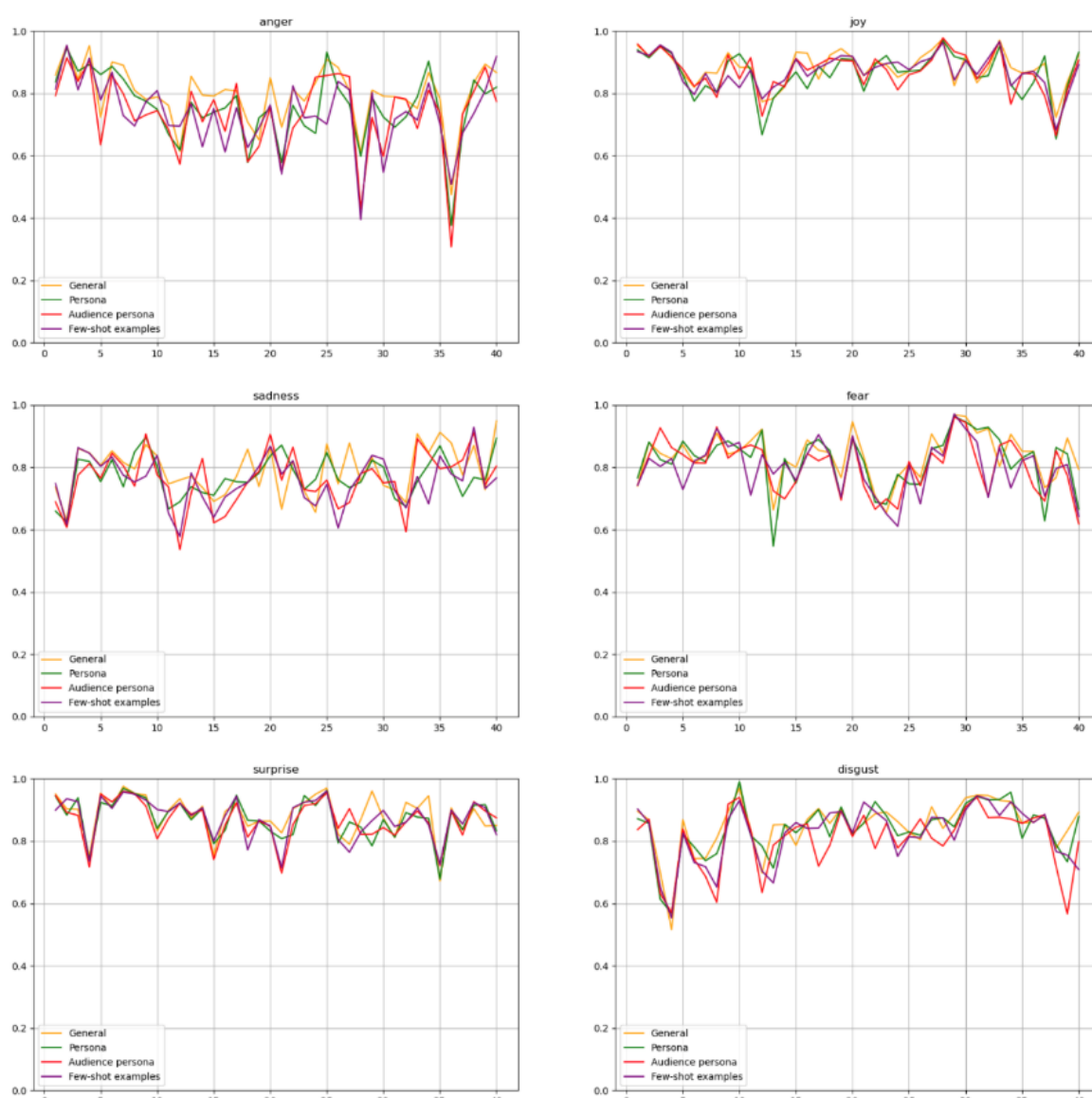


Рисунок 3.27 – Рівень семантичної схожості текстів до та після нейтралізації емоцій

Таблиця 3.7 – Середній рівень семантичної подібності

	Гнів	Радість	Смуток	Страх	Здивування	Огида
General	0.79643 2	0.884094	0.791774	0.836963	0.883686	0.849329
Persona	0.75668 1	0.865382	0.770124	0.820771	0.872852	0.838628
Audience persona	0.73586	0.871035	0.760486	0.803604	0.867952	0.805733
Few-shot examples	0.73509 5	0.871911	0.755177	0.799102	0.875309	0.823556

Модель GPT має найбільші труднощі зі збереженням семантики при нейтралізації емоцій гніву та смутку, де зберігається від 73% і 75% відповідно. В свою чергу емоції страху, огиди, радості та здивування зберігають як мінімум 79%, 80%, 86% і 86% відповідно.

ВИСНОВКИ

У дослідженні було використано сучасні методи штучного інтелекту для нейтралізації емоцій у текстах з метою збереження їхньої семантики. Експерименти показали, що модель GPT-4-Turbo ефективно знижує рівень емоційності в текстах, при цьому зберігаючи високий відсоток семантичної подібності (73–88%). Зокрема, шаблон промπτу persona демонструє високу ефективність у досягненні нейтралізації емоцій радості та здивування, а шаблон audience persona – для нейтралізації від емоцій гніву, страху, смутку і огиди.

Одні з найбільш вражаючі результати були отримані у зниженні рівня гніву (з 75% до 8%), що свідчить про високу ефективність використаних підходів для обробки текстів з негативними емоціями.

Найлегше моделі було нейтралізувати емоцію здивування, знизивши середній рівень емоції з 88% до 9%.

Аналогічні результати були досягнуті для смутку, страху та огиди, що додатково підтверджує універсальність та адаптивність методики.

Найгірший результат показала модель у задачі нейтралізації емоції радості, знизивши середній її рівень з 95% лише до 42%.

Тема цієї роботи має багато можливостей для розширення дослідження. Роботу можна продовжити у напрямку розширення спектру емоцій, не зупиняючись лише на базових. Також можна дослідити, як впливатиме на результат механізм подвійної нейтралізації, коли вже нейтралізований текст передається на вхід моделі для повторної нейтралізації. В свою чергу можна розширити спектр використаних шаблонів зі сфери prompt-engineering, щоб дослідити, чи вдасться знайти більш ефективний спосіб вказати великій мовній моделі, що саме треба зробити.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

- 1) Могильний С. Б. Машинне навчання з використанням мікрокомп'ютерів. Київ : Нац. центр «МАН України», 2019. 226 с.
- 2) Білоконь Б. LLM як інструмент для емоційної нейтралізації тексту. *28-й Міжнародний молодіжний форум «Радіоелектроніка та молодь у XXI столітті»*, м. Харків, 16–18 квіт. 2024 р. 2024. С. 958.
- 3) Automatic text summarization: a comprehensive survey / W. S. El-Kassas et al. Expert systems with applications. 2021. Vol. 165. P. 113679. URL: <https://doi.org/10.1016/j.eswa.2020.113679> (дата звернення: 29.03.2024).
- 4) Gupta V., Lehal G. S. A survey of text summarization extractive techniques. Journal of emerging technologies in web intelligence. 2010. Vol. 2, no. 3. URL: <https://doi.org/10.4304/jetwi.2.3.258-268> (дата звернення: 29.03.2024).
- 5) Floridi L., Chiriatti M. GPT-3: its nature, scope, limits, and consequences. Minds and machines. 2020. Vol. 30, no. 4. P. 681–694. URL: <https://doi.org/10.1007/s11023-020-09548-1> (дата звернення: 29.03.2024).
- 6) BERT: pre-training of deep bidirectional transformers for language understanding / J. Devlin et al. URL: <https://arxiv.org/pdf/1810.04805.pdf> (дата звернення: 29.03.2024).
- 7) Attention is all you need / A. Vaswani et al. URL: <https://arxiv.org/abs/1706.03762> (дата звернення: 29.03.2024).
- 8) ChatGPT for good? On opportunities and challenges of large language models for education / E. Kasneci et al. Learning and individual differences. 2023. Vol. 103. P. 102274. URL: <https://doi.org/10.1016/j.lindif.2023.102274> (дата звернення: 29.03.2024).
- 9) Damasio A. R. Emotion in the perspective of an integrated nervous system. Brain research reviews. 1998. T. 26, № 2-3. С. 83–86. URL: [https://doi.org/10.1016/s0165-0173\(97\)00064-7](https://doi.org/10.1016/s0165-0173(97)00064-7) (дата звернення: 27.05.2024).

10) Averill J. R. Individual differences in emotional creativity: structure and correlates. *Journal of personality*. 1999. Т. 67, № 2. С. 331–371. URL: <https://doi.org/10.1111/1467-6494.00058> (дата звернення: 27.05.2024).

11) Ekman P. An argument for basic emotions. *Cognition and emotion*. 1992. Т. 6, № 3-4. С. 169–200. URL: <https://doi.org/10.1080/02699939208411068> (дата звернення: 27.05.2024).

12) Ekman P. Basic emotions. *Handbook of cognition and emotion*. Chichester, UK, 2005. С. 45–60. URL: <https://doi.org/10.1002/0470013494.ch3> (дата звернення: 27.05.2024).

13) Chopade C. R. Text based emotion recognition: a survey. *International Journal of Science and Research (IJSR): Qualis CAPES Approved, Call for Paper*. URL: <https://www.ijsr.net/archive/v4i6/SUB155271.pdf> (дата звернення: 27.05.2024).

14) Seyeditabari A., Tabari N., Zadrozny W. Emotion detection in text: a review. *arXiv.org*. URL: <https://doi.org/10.48550/arXiv.1806.00674> (дата звернення: 27.05.2024).

15) Detection of emotion by text analysis using machine learning / K. Machová та ін. *Frontiers in psychology*. 2023. Т. 14. URL: <https://doi.org/10.3389/fpsyg.2023.1190326> (дата звернення: 27.05.2024).

16) Thorsten J. Text categorization with support vector machines. *Eldorado*. URL: <https://doi.org/10.17877/DE290R-5097> (дата звернення: 27.05.2024).

17) RoBERTa: A robustly optimized BERT pretraining approach / Y. Liu та ін. *arXiv.org*. URL: <https://doi.org/10.48550/arXiv.1907.11692> (дата звернення: 27.05.2024).

18) Scaling neural machine translation / M. Ott та ін. Proceedings of the third conference on machine translation: research papers, м. Belgium, Brussels.

Stroudsburg, PA, USA, 2018. URL: <https://doi.org/10.18653/v1/w18-6301> (дата звернення: 27.05.2024).

19) Large batch optimization for deep learning: training BERT in 76 minutes / Y. You та ін. *arXiv.org*. URL: <https://doi.org/10.48550/arXiv.1904.00962> (дата звернення: 27.05.2024).

20) DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter / V. Sanh та ін. *arXiv.org*. URL: <https://doi.org/10.48550/arXiv.1910.01108> (дата звернення: 27.05.2024).

21) Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing / P. Liu та ін. *arXiv.org*. URL: <https://doi.org/10.48550/arXiv.2107.13586> (дата звернення: 27.05.2024).

22) A prompt pattern catalog to enhance prompt engineering with chatgpt / J. White та ін. *arXiv.org*. URL: <https://doi.org/10.48550/arXiv.2302.11382> (дата звернення: 27.05.2024).

23) Salton G., Buckley C. Term-weighting approaches in automatic text retrieval. *Information processing & management*. 1988. Т. 24, № 5. С. 513–523. URL: [https://doi.org/10.1016/0306-4573\(88\)90021-0](https://doi.org/10.1016/0306-4573(88)90021-0) (дата звернення: 27.05.2024).

24) Training neural language models with SPARQL queries for semi-automatic semantic mapping / G. Futia та ін. *Procedia computer science*. 2018. Т. 137. С. 187–198. URL: <https://doi.org/10.1016/j.procs.2018.09.018> (дата звернення: 27.05.2024).

25) Reimers N., Gurevych I. Sentence-BERT: sentence embeddings using siamese bert-networks. *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, м. Hong Kong, China. Stroudsburg, PA, USA, 2019. URL: <https://doi.org/10.18653/v1/d19-1410> (дата звернення: 27.05.2024).

26) Johnson J., Douze M., Jegou H. Billion-scale similarity search with GPUs. *IEEE transactions on big data*. 2019. С. 1.

URL: <https://doi.org/10.1109/tbdata.2019.2921572>

(дата

звернення:

27.05.2024).