Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет Комп'ютерної інженерії та управління
Кафедра Комп'ютерних інтелектуальних технологій та систем

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)

Інтелектуальна модель стилістичного аналізу гумористичних текстів

(тема)

Виконав:
здобувач 2 року навчання
групи КІТм-23-1
Левченко В.В.

Спеціальність 123 Комп'ютерна інженерія
Тип програми освітньо-професійна
Освітня програма Комп'ютерні інтелектуальні технології
Керівник доцент каф. КІТС, Ілюнін О.О.

Допускається до захисту

Зав. кафедри О.Г. Руденко

2024 р

Харківський національний університет радіоелектроніки

Факультет	Комп'ютерної інженерії та управління
Кафедра	Комп'ютерних інтелектуальних технологій та систем
Рівень вищої освіти	другий (магістерський)
Спеціальність	123 Комп'ютерна інженерія
Тип програми	освітньо-професійна
Освітня програма	Комп'ютерні інтелектуальні технології

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)

« 21 » 01 2025 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

здобувчеві Левченку Владиславу Васильовичу
(прізвище, ім'я, по батькові)

1. Тема роботи Інтелектуальна модель стилістичного аналізу гумористичних текстів

затверджена наказом університету від « 28 » жовтня 2024р. № 1156Ст

2. Термін подання студентом роботи до екзаменаційної комісії 20.01 2025р.

3. Вхідні дані до роботи (проекту)

1) Стилістичні характеристики текстів
2) Нейронні мережі, RNN, LTSM

4. Перелік питань, що потрібно опрацювати в роботі

1) Аналіз тексту для розуміння комп'ютером.
2) Підбір методу класифікації текста.
3) Розробка інтелектуальної моделі класифікації стилістики авторського тексту.
4) Дослідження моделі і тестування моделі. 6)Висновки

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій

10 слайдів перзентаційного матеріалу

6. Консультанти розділів роботи (п.6 включається до завдання за наявністю консультантів згідно до наказу, зазначеному у п.1)

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

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Термін виконання етапів роботи	Примітка
1	Видача завдання на проектування	28.10.24	
2	Ознайомлення з літературними джерелами, аналіз і вибір методу розв'язання поставленної задачі	26.11.24 - 01.12.24	
3	Розробка алгоритмів розв'язання, обгрунтквання системних засобів вирішення завдачі	01.12.24 – 10.12.24	
4	Розробка моделі	10.12.24 - 26.12.24	
5	Відладка і тестування моделі	26.10.24 - 30.11.24	
6	Оформлення пояснювальної записки	01.01.25 - 15.01.25	
7	Передзахист кваліфікаційної работи	23.01.25	
8	Захист кваліфікаційної роботи	24.01.25	

Дата видачі завдання «28» жовтня 2024р.

Здобувач _____
(підпис)

Керівник роботи _____ доцент каф. КІТС, Ілюнін О.О.
(підпис) (посада, ініціали, прізвище)

РЕФЕРАТ

Загальний обсяг роботи: 89 с., 19 рисунків, 21 джерел, 7 формул, 2 таблиці

НЕЙРОННІ МЕРЕЖІ, RNN, TF-IDF, LSTM, BERT, XLM-RoBERTa, РОЗПІЗНАВАННЯ ТЕКСТА, ГЛИБОКЕ НАВЧАННЯ, МОДЕЛЬ.

Метою роботи є створення моделі для виконання стилістичного аналізу гумористичного тексту.

Об’єктом дослідження є стилістичні характеристики тексту.

Предметом дослідження є алгоритми класифікації стилів тексту, що засновані на використанні формалізованних критеріїв аналізу стилістичних властивостей тексту, які використовуються при побудові глибокої нейронної мережі BERT.

В кваліфікаційній роботі запропоновано та досліджено варіанти класифікації тексту на основі методу «мішка слів» та TF-IDF, рекурентних нейронних мереж (RNN) й LSTM та трансформерів (XLM-RoBERTa).

ABSTRACT

Coursework: 89 pages, 19 figures, 21 sourses, 7 formulas, 2 tables

NEURAL NETWORKS, RNN, TF-IDF, LSTM, BERT, XLM-RoBERTa, TEXT RECOGNITION, DEEP LEARNING, MODEL.

The aim of the study is to create a model of a software for recognition the author's style text..

The object of study is stylistic characteristics of a text

The subject of research is text style classification algorithms based on the selection of property criteria for a deep neural network.

In the course work, variants of text classification based on the bag-of-words method and TF-IDF, recurrent neural networks (RNN) and LSTM, as well as transformers (XLM-RoBERTa) were proposed.

Міністерство освіти і науки України
Харківський національний університетрадіоелектроніки

Факультет Комп'ютерної інженерії та управління
Кафедра Комп'ютерних інтелектуальних технологій та систем

АНОТАЦІЯ

КВАЛІФІКАЦІЙНОЇ РОБОТИ

рівень вищої освіти другий (магістерський)

Інтелектуальна модель стилістичного аналізу гумористичних текстів

Виконав:
здобувач 2 року навчання
групи КІТм-23-1
Левченко В.В.

Спеціальність 123 Комп'ютерна інженерія
Тип програми освітньо-професійна
Освітня програма Комп'ютерні інтелектуальні технології
Керівник доцент каф. КІТС, Ілюнін О.О.

2024 р.

АНОТАЦІЯ

Левченко В.В. Тема кваліфікаційної роботи. – Інтелектуальна модель стилістичного аналізу гумористичних текстів

У магістерській кваліфікаційній роботі вирішено актуальну проблему створення моделі для стилістичного аналізу гумористичних текстів за допомогою сучасних технологій глибокого навчання. Розробка здійснена з використанням рекурентних нейронних мереж (RNN), LSTM та трансформерів (XLM-RoBERTa), що забезпечують високу точність класифікації та аналізу стилю тексту.

Метою кваліфікаційної роботи є створення інтелектуальної моделі, яка здатна автоматично визначати стилістичні характеристики текстів гумористичного характеру, сприяючи спрощенню аналізу текстів у таких сферах, як видавнича справа, освіта та засоби масової інформації.

Об'єктом дослідження є стилістичні характеристики гумористичних текстів. Предметом дослідження є інтелектуальні алгоритми класифікації текстів, що базуються на нейронних мережах та методах обробки природної мови.

Технологія класифікації авторських стилів тексту може додавати теги стилів до текстів на основі вмісту. Коли справа доходить до дослідження та впровадження таких аспектів, як ефективна організація пошуку та аналіз текстових ресурсів, це дуже важливо. Традиційні методи класифікації авторських стилів тексту використовують широкий спектр лінгвістичних характеристик. Розробка характеристик вимагає знань у галузі лінгвістики, а характеристики різних завдань класифікації не завжди узгоджені. Швидкий розвиток нейронних мереж і технологій обробки природної мови надав новий спосіб кращого розв'язання проблеми класифікації авторських стилів тексту.

Великі мовні моделі (LLM) є важливою віхою в розвитку технологій обробки природної мови. Їх розробка призвела до революційного покращення

розуміння, аналізу та використання мови в машинній комунікації. Цей процес розвитку призвів до розширених можливостей LLM для вирішення складних лінгвістичних проблем і суттєво змінив технологічний ландшафт, відкривши нові можливості та виклики в галузі обробки природної мови. Пропонується модель класифікації стилів тексту, заснований на виділенні властивостей для побудови глибокої нейронної мережі, який може ефективно покращити ефективність класифікації текстових стилів. За параметри класифікації стилю береться низка типів ознак: за родами (віршований, прозовий, драматичний), за знаковою системою (ноти, таблиці, формули), за місцем розташування на аркуші (епіграф, реферат, титульний текст, покажчик, тощо), за джерелами походження (препаровані, натуральні), за типами трансформацій вихідного натурального тексту (адаптовані, неадаптовані, повні, скорочені, змішані), за основними прагматичними функціями (інформативні, оцінні, інструктивні, змішані), за формою репрезентації (усні, письмові), за формою спілкування (діалогічні, монологічні), за сферою спілкування (стиль мовлення) (розмовні, наукові, офіційно-ділові, художні, публіцистичні.

Гумор відіграє важливу роль у світовій культурі комунікації. Вітчизняні фахівці з соціальних комунікацій радять починати публічний виступ з жарту, успішні політики повинні постійно демонструвати своє почуття гумору. Людина, яка не вміє дотепно жартувати, навряд чи досягне вершин ієрархії у будь-якій сфері діяльності, і щоб з'ясувати, чи є у кандидата на керівну посаду почуття гумору, його запрошують на неформальні прийоми, після чого вирішується його професійна доля. Людину, яка не жартує або не реагує на жарти, відкидають без вагань. Школи жартів особливо популярні в Японії, їх відвідують державні службовці, бізнесмени, поліцейські і навіть буддистські монахи. До комунікативних форм сміхової культури, що проявляються в міжособистісних стосунках, належать дотепні жарти, анекдоти, іронія, сарказм та ін. Гумор може бути добрим чи поганим, що визначає його прямий зв'язок з мораллю. Оскільки не всі варіанти культури сміху відповідають морально-етичним принципам поваги і такту, необхідно розрізняти культуру і

квазікультуру (лат. quasi - псевдо, ніби) гумору. Суть культури гумору полягає у створенні такого комічного смислу, який не ображає людину, не принижує її гідності і створює умови для взаємного емоційного задоволення при сприйнятті комічного. Тому найбільш прийнятною формою спілкування, що створює позитивний емоційний клімат у міжособистісних стосунках і сприяє взаємній насолоді від сміху, є добрий жарт - смішне висловлювання, комічна дія, що викликає сміх. Стилістичний аналіз тексту — це комплексний метод аналізу тексту, спрямований на визначення мовних засобів, використаних автором для досягнення комунікативної мети. Аналіз є важливим етапом у формуванні лінгвостилістичної компетентності, особливо для майбутніх учителів української мови і літератури. Мета стилістичного аналізу тексту - проаналізувати, як мовні засоби різних рівнів (фонетичного, лексичного, граматичного) впливають на художню образність, вираження думки та емоційний вплив тексту на читача. Розуміння специфіки функціонування мовних одиниць у тексті є основою у визначенні стилі твору а також дозволяє оцінити ефективність використання мовних засобів у контексті жанру та функціонального стилю. За повнотою виконання - частковий (аналіз окремих аспектів тексту, наприклад, лексичних засобів), повний (охоплює всі рівні мовної організації); за функціональними стилями - аналіз текстів художнього, наукового, публіцистичного, офіційно-ділового, розмовного стилів; за рівнями мовної організації - фонетико-стилістичний, лексико-стилістичний, морфолого-стилістичний, синтаксично-стилістичний.

У сучасному світі, що швидко змінюється, технологічний прогрес відіграє ключову роль у формуванні повсякденного життя. Однією з галузей, яка значно розвинулася за останні кілька років, є сфера штучного інтелекту (ШІ). Штучний інтелект, міждисциплінарна галузь, зосереджена на створенні комп'ютерних систем, здатних виконувати завдання, які зазвичай вимагають людського мислення. Мовна модель — це розподіл ймовірностей слів або послідовностей слів. На практиці вона визначає ймовірність того, що дана послідовність слів є «правильною». «Правильність» у цьому контексті означає

не граматичну правильність, а те, що послідовність слів схожа на тексти, які пишуть люди, тобто на те, чого навчає мовна модель. Це важливий момент. Мовна модель – це простий інструмент для включення великої кількості інформації в стислий спосіб, який можна повторно використовувати в контекстах поза вибіркою.

Процес створення моделі для оцінювання близкості авторського стилю тексту до еталонів, використовуючи передові методи NLP, зокрема, трансформерів, таких як модель BERT або XLM-RoBERTa.

Дані у вигляді текстів зберігаються в структурованому форматі, наприклад у таблиці (PSV або CSV), де кожен рядок містить текст і відповідного автора. Попередня обробка даних є обов'язковою, видаляються зайві символи, HTML-теги, а також приводиться текст до нижнього регістру для уніфікації. Текст розбивається на токени (слова або підслова), які будуть використовуватися моделлю BERT. Для цього можна використовувати вбудований токенізатор. Токени, введені користувачем, вважаються вхідними даними для моделей машинного навчання. Однак моделі розуміють тільки числа, а не текст, тому ці вхідні дані необхідно перетворити в числовий формат, званий «вбудованими вхідними даними». Вхідні вкладення представляють слова у вигляді чисел, які потім можуть оброблятися моделями машинного навчання.

Ці вкладення подібні до словника, який допомагає моделі зрозуміти значення слів, поміщаючи їх у математичний простір, де схожі слова розташовані поруч одне з одним. Починаючи з навчального набору даних, що містить вектор ознак і відповідну мітку класу, дерева рішень створюються за допомогою алгоритму RandomForest. Цей метод має низку переваг: висока точність – завдяки ансамблю дерев рішень, стійкість до перенавчання – алгоритм зменшує перенавчання шляхом усереднення декількох дерев і дає уявлення про важливість різних ознак для прогнозу. У контексті сучасних досліджень і застосувань машинного навчання значна увага приділяється проблемі дисбалансу класів у наборах даних. Методологія передбачає відбір

зразків з меншого класу, визначення *k*-найближчих сусідів та створення синтетичних вибірок. Спостереження з недопредставленого класу відбираються, і для кожного відібраного спостереження обчислюються *k*-найближчих сусідів у просторі ознак.

Основні результати дослідження включають в себе створення та налаштування моделі нейронної мережі для стилістичного аналізу текстів, аналіз ефективності традиційних методів класифікації тексту та порівняння з сучасними моделями на основі глибокого навчання, отримання високих показників точності класифікації текстів, що свідчить про ефективність запропонованого підходу.

Наукова новизна роботи полягає у поєднанні класичних та сучасних методів текстового аналізу для створення моделі, яка враховує специфічні стилістичні особливості текстів.

Практична цінність полягає у можливості використання моделі для автоматизованого аналізу текстів, що сприяє підвищенню ефективності роботи редакторів та письменників.

Однією з головних проблем є залежність точності класифікації від обсягу та якості вхідних даних. У випадках, коли обсяг даних є недостатнім, такі моделі не можуть належним чином навчитися розпізнавати патерни та особливості, що є критичними для точного класифікування. Також класичні класифікатори часто демонструють обмеження у своїй ефективності, особливо коли мова йде про обробку неповних або недостатніх даних. Вони базуються на попередньо визначених ознаках і передбачають, що всі необхідні характеристики даних доступні для аналізу, однак у реальних умовах ця передумова часто порушується. Неповні дані можуть виникати через різні причини: обмеженість ресурсів, технічні помилки, або ж особливості самих даних, які є важкодоступними або рідкісними, у таких випадках модель не здатна ефективно узагальнювати інформацію, що призводить до високої ймовірності помилкових класифікацій і зниження загальної продуктивності системи.

Таким чином, доречна розробка інтелектуальної моделі стилістичного аналізу гумористичного тексту при умові неповноти вхідних даних. Натомість, на заміну класичного класифікатора, запропоновано рішення з використанням методу навчання на неповних даних. Подібний метод дозволить класифікувати ті дані, з котрими раніше не можливо було б працювати, а саме приводячи більшість унікальних признаків до одного, за яким можно було б чітко ідентифікувати текст.

У процесі виконання кваліфікаційної роботи була розглянута концепція оцінювання текстів за допомогою класифікатора за набором ознак: за різноманітними лінгвістичними характеристиками, зокрема кількість слів, кількість речень, середня довжина речень, кількість знаків пунктуації та часткою великих літер. Створена модель для стилістичного розбору стилю тексту використовує трансформери, зокрема XLM-RoBERTa для навчання, адже це суттєво скоротить часові витрати у вирішенні завдань класифікації тексту за стилем. Проводиться аналіз тексту на предмет його жанру та лінгвістичних особливостей. Модель має середню точність оцінки 89% відсотків, що є досить великим показником, при відносно невеликому обсягу даних, на якому вона була навчена. При обсязі в 1000 рядків (даних), 3/4 жанрів текстів були визначені вірно.

Під час тесутвання роботи моделі було виявленно та проаналізовано низку недоліків. Серед них низька якість вхідних даних та їх очевидний недолік, необхідність в додаткових як матеріальних так і інтелектуальних ресурсів для створення більш точної, повної та якісної оцінки, а також нестача обчислювальних потужностей. Дані, на яких проводилось навчання моделі, мали дуже великий діапазон даних, це призводило до нездатності правильно визначити жанр тексту, через це довелось використовувати синтетичний балансувальник, що знижувало точність через шуми у вибірці. Важливу роль відіграє кількість даних для навчання. Доступним по темі роботи є датасет який складається 42000 текстів та 18 унікальних жанрів, проте таких даних як, стиль та інші було недостатньо для повноцінного функціоналу, наприклад, на

десять тисяч текстів з різноманітними жанрами, приходиться всього одна тисяча гумористичних, що у більшості випадків їх класифікації видавало результат «інше».

Підхід такого роду може бути корисним для різних застосувань, включно з літературним аналізом, захистом авторських прав і автоматичним розпізнаванням плагіату. Навчання на достатньо великому обьемі даних дозволить проводити повний стилістичний аналіз включаючи в себе увесь список параметрів аналізу. Інтегрування системи перевірки на плагіат та авторськи права, на основі парсингу мережі інтернет, та більш продвинутого варіанту, аналізу за допомогою штучного інтелекту.Надалі можливі поліпшення моделі завдяки використанню більших і різноманітніших даних, а також експериментування з різними архітектурами нейронних мереж.

Розроблена модель має місце в різних структурах та організаціях, таких як, видавництва, школи, університети, книгарні. Здібність швидко та точно працювати з текстом є користним інструментом в закладах, де безпосередньо відбуваєтсья робота з текстом. У школах та університетах це або додатковий способ для навчання, а скорочення часу викладачам на рутинних операціях. Видавництва також можуть використовувати дану модель для швидкого визначення та аналізу текста

СТИЛІСТИЧИЙ АНАЛІЗ, ГУМОРИСТИЧНІ ТЕКСТИ, НЕЙРОННІ МЕРЕЖІ, LSTM, BERT, XLM-RoBERTa, ОБРОБКА ПРИРОДНОЇ МОВИ

Публікації здобувача за темою роботи:

1. Левченко В.В., Ілюнін О.О. «МОДЕЛЬ ОЦІНКИ СТИЛІСТИЧНОГО АНАЛІЗУ ГУМОРИСТИЧНОГО ТЕКСТУ», Матеріали конференцій МНЛ, (14 червня 2024 р., м. Львів). С. 159–161.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ,СКОРОЧЕНЬ І ТЕРМІНІВ

ШІ — штучний інтелект.

LLM – (англ. Large language model) — велика мовна модель.

PTG – (англ. Pretrained Language Models) — предтренована мовна модель.

LSTM – (англ. Long short term memory) — мережа с довгою та короткотривалою пам’яттю.

Трансформер – архітектура глибокого навчання.

BERT – (Bidirectional Encoder Representations from Transformers) – Двонаправлена нейронна мережа від компанії Google.

ВСТУП

Технологія класифікації авторських стилів тексту може додавати теги стилів до текстів на основі вмісту. Коли справа доходить до дослідження та впровадження таких аспектів, як ефективна організація, пошук та аналіз текстових ресурсів, це дуже важливо.

Традиційні методи класифікації авторських стилів тексту використовують широкий спектр лінгвістичних характеристик. Розробка характеристик вимагає знань у галузі лінгвістики, а характеристики різних завдань класифікації не завжди узгоджені. Швидкий розвиток нейронних мереж і технологій обробки природної мови надав новий спосіб кращого розв'язання проблеми класифікації авторських стилів тексту.

У цій кваліфікаційній роботі пропонується метод на основі вилучення текстових характеристик та застосування глибоких нейронних мереж та для розв'язання проблеми низької точності традиційних методів стилістичного аналізу текстів.

Алгоритм класифікації авторських стилів тексту виділяє кілька типів ознак як класифікаційні характеристики: лексичні, синтаксичні та стилістичні ознаки. Традиційні методи, такі як мішок слів (Bag of Words) та TF-IDF, добре працюють для базового аналізу тексту, але мають обмеження в захопленні контексту та стилістичних особливостей.

Застосування сучасних методів, таких як рекурентні нейронні мережі (RNN), особливо LSTM, дозволяє враховувати послідовності слів і краще розпізнавати стилістичні патерни. Однак, найсучасніші результати досягаються з використанням трансформерів, таких як BERT, які здатні моделювати складні залежності в тексті.

В результаті запропоновано модель для класифікації авторських стилів тексту, яка поєднує переваги традиційних методів та сучасних підходів на

основі глибокого навчання. Для кращого представлення властивостей стилю автора до виводу застосовуються різні ваги та оптимізаційні техніки.

1 АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ

Великі мовні моделі (LLM) є важливою віхою в розвитку технологій обробки природної мови. Їх розробка призвела до революційного покращення розуміння, аналізу та використання мови в машинній комунікації. Цей процес розвитку призвів до розширених можливостей LLM для вирішення складних лінгвістичних проблем і суттєво змінив технологічний ландшафт, відкривши нові можливості та виклики в галузі обробки природної мови.

Аналіз Bowman [1] виявив розширені можливості LLM у виведенні авторських знань, розрізненні помилок і фактів, а також у демонстрації чутливості до навчального контенту. Він також підкреслив чутливість LLM до навчального контенту і появу швидкого інжинірингу для обходу їхніх недоліків. Розвиток мовних моделей має значний вплив на здатність міркувати з точки зору здорового глузду.

На даний момент мовні моделі є універсальним інструментом для вирішення проблем різного роду, наприклад, допомагають обробляти великі обсяги інформації і навіть спілкуватися з іншими людьми. Ці приклади дають нам зрозуміти, що мовні моделі впритул наблизилися до людської продуктивності. Однак для досягнення такої продуктивності дуже важливу роль відіграють правильно розроблені підказки. Yongchao Zhou [2] наводить докази прогресу, досягнутого великими мовними моделями, і підкреслює їхню подібну до людської продуктивність у різних завданнях обробки природної мови. Однак, як підкреслюється в статті, якість підказок, якими керуються ці моделі, все ще має вирішальне значення. Zhou зазначає, що найефективніші та найточніші підказки часто генеруються людиною вручну. Це підкреслює важливість людського вкладу у формування ефективного управління такими передовими системами. Але підказки - це не єдиний аспект. Перш ніж підказки можна буде використовувати для спілкування з мовними моделями, моделі потрібно протестувати та адаптувати до їхніх потреб. Саме ці та інші ключові

аспекти описані в [3]. Це всебічне і глибоке масштабне дослідження мовних моделей, що охоплює ключові аспекти їхнього функціонування. Zhou обговорює процеси початкового навчання, адаптації, використання та оцінки цих моделей.

Інженерія підказок є важливим напрямком досліджень у галузі великих мовних моделей (LLM), що робить значний внесок у їхнє ефективне використання та оптимізацію. Цей напрям зосереджений на формулюванні підказок, які є підказками або інструкціями, що подаються до LLM для отримання бажаних відповідей.

Розробка підказок передбачає стратегічний дизайн і вдосконалення вхідних інструкцій з метою підвищення ефективності моделі та адаптації її результатів до конкретних контекстів або завдань. Ретельно розроблені підказки можуть мати значний вплив на здатність LLM розуміти тонкі запити, генерувати контекстно-релевантний контент і демонструвати бажану поведінку в різних додатках.

З цією метою було введено поняття метапідказок [4].

Junyi Li [5] розглядає проблеми адаптації моделей навчання на основі підказок (Pretrained Language Models) до завдання генерації тексту. Інноваційний метод передачі підказок для генерації тексту (PTG) використовує навчання з перенесенням для впровадження динамічного підходу до вибору підказок. Результати показують потенціал кластеризації подібних завдань і наборів даних, проливаючи світло на ефективну розробку підказок для покращення генерації тексту. Розробка підказок займає центральне місце в роботі [4] Reynolds наголошує на необхідності ефективного контролю та взаємодії з LLM. Це вимагає обмеження підказок і включення наративів і культурних якорів з точки зору природної мови, щоб підвищити релевантність і якість створеного контенту. Нові дослідження ще більше розширюють цю перспективу.

Дослідження Webson [6] є важливим внеском у розуміння того, як моделі, засновані на підказках, інтерпретують і обробляють інструкції. Webson

піднімає важливі питання про повноту розуміння підказок моделями, припускаючи, що існує потенційна прогалина в їхній здатності повністю розуміти складні інструкції, що може призвести до обмежень у їхній здатності імітувати людське розуміння. Ця робота проливає світло на аспект, що має вирішальне значення для ефективного функціонування моделей, заснованих на підказках, а саме на здатність адекватно розуміти та інтерпретувати поставлені запитання або інструкції.

В дослідженні [7], з іншого боку, Kervadec розширює перспективу, зосереджуючись на тому, як мовні моделі реагують на підказки, що генеруються як машинами, так і людьми. Kervadec виявляє відмінності у моделях реагування, вказуючи на тонкі, але важливі відмінності у сприйнятті та обробці підказок, що генеруються різними джерелами. Цей висновок підкреслює необхідність глибшого розуміння того, як мовні моделі справляються з різними типами підказок, що має вирішальне значення для покращення їхньої здатності ефективно спілкуватися з різними користувачами, як машинами, так і людьми. Jules White та співавтори [8] знайомлять із захопливою галуззю інженерії підказок для великих мовних моделей. White та співавтори не лише аналізують роль підказок, але й представляють вичерпний каталог патернів підказок, чітко згрупованих у п'ять категорій: семантика введення, вирівнювання результатів, ідентифікація помилок, покращення підказок та взаємодія. Ця структурована класифікація патернів має вирішальне значення для розуміння різних аспектів інженерії підказок і дозволяє ефективно застосовувати їх на практиці. White та співавтори підкреслюють важливість документування цих патернів, з акцентом на створенні чітких і зрозумілих посилань, які є цінним ресурсом для користувачів.

Задокументовані шаблони підказок важливі для того, щоб допомогти розробникам моделей ефективно вирішувати типові проблеми, з якими вони можуть зіткнутися. Структуровані посилання на п'ять категорій патернів дозволяють більш систематично підходити до проектування підказок, що

призводить до більш точного контролю над великими мовними моделями. Qinyuan Ye [9] глибоко розглядає тему ключової ролі швидкого проектування в оптимізації продуктивності великих мовних моделей. Йє не лише представляє швидкий інжиніринг як надзвичайно складне завдання, але й вказує на те, що його ефективність залежить від просунутого виводу.

Аналіз моделі, висування гіпотез щодо можливих помилок або відсутньої інформації в поточній підказці та чітке формулювання завдання вимагають точних навичок. Ye представляє інноваційний метод під назвою PE2, який об'єднує ключові концепції оптимізації, такі як розмір партії, крок та імпульс, у мета-підказку. Метою PE2 є покращення швидкої інженерії, спрямовуючи мовні моделі на більш ефективну роботу. Зазначено, що PE2 добре справляється з різноманітними завданнями, перевершуючи встановлені еталони для автоматизованої швидкої інженерії. Така універсальність у виконанні різноманітних мовних завдань підкреслює практичне застосування PE2 і його потенціал для покращення продуктивності великих мовних моделей. Це відкриває нові перспективи в галузі швидкої інженерії, пропонуючи ефективний підхід на основі оптимізації.

Не менш актуальною є тема оцінювання якості текстів, згенерованих великими мовними моделями. Zishan Guo [10] досліджує оцінювання великих мовних моделей, зосереджуючись на їхній здатності генерувати текст і код. Guo не лише наголошує на необхідності ретельного оцінювання LLM перед їх впровадженням, але й підкреслює потенційні ризики недостатнього оцінювання, що може призвести до непередбачуваних наслідків. Zishan Guo представляє важливий перехід від традиційного тестування, орієнтованого на завдання, до оцінювання, орієнтованого на здібності, де межі між різними кінцевими завданнями стають нечіткими. Ця нова перспектива оцінювання розглядає цілісні здібності мовних моделей, що охоплюють генерацію тексту, коду та інструментів. Крім того, Guo дає зрозуміти, що можливості LLM сильно залежать від розміру їхніх параметрів і точного налаштування з використанням високоякісних даних. Варто зазначити, що Guo демонструє,

що менші моделі можуть працювати так само добре, як і більші, що є важливим висновком для практиків. Це визнання потенціалу менших моделей в контексті продуктивності відкриває нові можливості, а також підкреслює важливість кастомізації навчання як ключового елементу, що впливає на ефективність моделей.

Дослідження, що проводив Lochan Basyal [11] вводить нову сферу, пов'язану з LLM, а саме питання скорочення тексту. Basyal досліджує використання моделей великих мов для скорочення тексту, підкреслюючи цінність цього методу в епоху великих даних, коли велика кількість текстової інформації підкреслює необхідність ефективних методів скорочення тексту. Basyal представляє два основні підходи до скорочення тексту: абстрагування та екстрагування. Він також обговорює контрольовані та неконтрольовані методи скорочення тексту, підкреслюючи переваги кожного підходу та їх відповідні застосування.

У статті оцінюються можливості скорочення тексту різними LLM, такими як mpt-7b-instruct, falcon-7b-instruct і text-davinci-003, на двох наборах даних: CNN/Daily Mail 3.0.0 і XSum.

Результати показують виняткову продуктивність моделі OpenAI text-davinci-003 порівняно з іншими моделями. Basyal використовує алгоритми оцінки якості тексту, такі як BLEU, ROUGE, щоб визначити, яка мовна модель зі списку краще впоралася з поставленим завданням. BLEU і ROUGE є широко прийнятими метриками для оцінки якості машиногенерованих резюме. BLEU вимірює схожість між n-грамами в реченнях, перекладених машиною, і n-грамами в реченнях, перекладених людиною. ROUGE, з іншого боку, оцінює збіг n-грам між згенерованими анотаціями та еталонними анотаціями. Він враховує такі метрики, як ROUGE-N (уніграми, біграми тощо) і ROUGE-L (найдовша спільна підпослідовність), щоб оцінити збіг змісту. Ці метрики є цінними інструментами для оцінки ефективності різних LLM та вдосконалення їх моделей. Крім того, вони можуть бути використані для порівняння ефективності різних LLM на різних наборах даних.

Штучний інтелект широко використовується в журналістиці і змінює способи збору, обробки та поширення інформації. Однією з найважливіших функцій ШІ в журналістиці є автоматизація таких процесів, як створення новин і редагування текстів. Багато інформаційних агентств використовують алгоритми для створення коротких новин на основі фінансових звітів або спортивних результатів, що дозволяє редакторам відмовитися від монотонної роботи і зосередитися на аналізі та інтерв'ю. Такі алгоритми навчання аналізують вподобання користувачів, щоб запропонувати найбільш релевантні статті та теми, покращуючи користувацький досвід і підвищуючи залученість аудиторії. Наприклад платформа Google News використовує ШІ для створення персоналізованих стрічок новин, а також ШІ для перевірки фактів і виявлення дезінформації в системі. Штучний інтелект може аналізувати різні джерела, виявляти суперечливу інформацію та попереджати редакторів про потенційні помилки, щоб забезпечити точність і достовірність контенту, а також виключити можливість таблоїдизації.

Технології розпізнавання мови та зображень полегшують роботу журналістів, розшифровуючи аудіо- та відеоінтерв'ю та автоматично додаючи субтитри. Водночас використання штучного інтелекту в журналістиці порушує важливі етичні питання: Необхідно забезпечити прозорість алгоритмів, уникнути упередженості та захистити приватність користувачів, для чого потрібно розробити нові стандарти та кодекси.

Цей тип інструментів змінює журналістику, пропонуючи нові можливості для підвищення якості контенту та ефективності роботи редакційних команд. Однак важливо враховувати та вирішувати етичні та правові питання, які виникають, щоб забезпечити відповідальне та безпечне використання наданого контенту.

Можливості автоматизованого аналізу есе майже не обмежені. Учні отримують зворотній зв'язок щодо їхніх стилістичних особливостей, включаючи оцінку тональності тексту, використання таких літературних прийомів, як метафора, алегорія та іронія, а також відповідність тексту з точки

зору жанру і стилю, що, в свою чергу, допоможе їм зрозуміти свої сильні і слабкі сторони і навчитися покращувати структуру і виразність своїх текстів. Також є вікно, яке вчителі можуть використовувати для аналізу текстів класичних і сучасних творів, наприклад, щоб показати учням розмаїття стилів і мовних прийомів, які використовують різні автори. Це допоможе вчителю не витрачати надто багато часу на підготовку матеріалів до урока або доповнювати їх на основі матеріалів, наданих штучним інтелектом, що, в свою чергу, допоможе ефективніше працювати з матеріалом і глибше зрозуміти особливості літератури, а також розвинути навички інтерпретації та аналізу.

Існує низка проблем та обмежень у сфері вивчення іноземних мов, які дане дослідження може допомогти вирішити. За допомогою такого інструменту аналіз письмових робіт учнів стає набагато простішим і швидшим, коли найпоширеніші та найпростіші помилки фіксуються попередньо навченою моделлю, яка оцінює, чи правильно використовуються мовні конструкції та стилістичні прийоми, характерні для мови, що вивчається. Це прискорює засвоєння мови та дозволяє уникнути типових помилок.

Модель стилістичного аналізу також може стати незамінним інструментом для вчителів: Вона полегшує оцінювання письмових робіт і дає більше часу для індивідуальної роботи зі студентами. У великих класах або на курсах з високим навантаженням, де викладачі не завжди можуть приділити достатньо часу кожному студенту, це полегшує роботу викладача та засвоєння матеріалу студентами. Такі практики сприяють розвитку здатності до саморефлексії, заохочують студентів до самостійної роботи над текстами, зміцнюють впевненість студентів у своїх здібностях і мотивують їх до постійного вдосконалення своїх навичок.

Інтеграція такої моделі в освітній процес також може підвищити якість навчання, оскільки воно стає значно інтерактивним та більш індивідуалізованим, що, в свою чергу, сприяє кращому засвоєнню матеріалу та розвитку ключових навичок і вмінь учнів.

В кваліфікаційній роботі пропонується модель класифікації стилів тексту, заснована на виділенні властивостей тексту для побудови глибокої нейронної мережі, яка може покращити ефективність класифікації текстових стилів:

– за параметрами класифікації стилю береться низка типів ознак:

– за родами — віршований, прозовий, драматичний,

– за знаковою системою — ноти, таблиці, формули,

– за місцем розташування на аркуші — епіграф, реферат, титульний текст, покажчик тощо,

– за джерелами походження — препаровані, натуральні, за типами трансформацій вихідного натурального тексту — адаптовані, неадаптовані, повні, скорочені, змішані,

– за основними прагматичними функціями — інформативні, оцінні, інструктивні, змішані,

– за формою репрезентації — усні, письмові,

– за формою спілкування — діалогічні, монологічні,

– за сферою спілкування — розмовні, наукові, офіційно-ділові, художні, публіцистичні, змішаного типу: побутово-ділові, науково-публіцистичні, науководілові, художньо-публіцистичні,

– за стилями та жанрами — наукові: наукові статті, тексти підручників, монографії, виступи, доповіді, реферати і под.; офіційно-ділові: характеристика, заява, пояснювальна записка, автобіографія, анкета, рекомендація тощо; інформаційнопубліцистичні: інформаційний огляд, передова стаття, коментар, фейлетон, нарис; художні: поеми, вірші, пісні, оповідання, казки, повісті тощо,

– за способом викладу — розповідні, описові, модально-полемічні, пояснювальні, тексти-роздуми, тексти-висновки, текстидоведення, тексти-визначення, змішані або комбіновані тексти,

– за експресивністю мовлення — художні, розмовні, публіцистичні, нейтральні, наукові, експресивно і стилістично забарвлені, за емоційністю мовлення неофіційні, офіційні, нейтральні,

– за прагматичними функціями — наукові, інформативні, газетно-інформаційні, інструктивні, оцінні, офіційно-ділові, інформативно-оцінні, офіційно-ділові, публіцистичні, оцінно-інструктивні,

– за функціонально-смисловими типами мовлення — розповідні, описові, аргументативні, пояснювальні, інструктивні, визначення.

Таким чином, використання сучаних методів побудови моделей - RNN, особливо LSTM та трансформерів буде оптимальним для досягнення якісного результату оцінювання тексту.

Використання великих мовних моделей (LLM) уможливлює автоматизоване створення та редагування текстів, поглиблений аналіз даних і покращення якості контенту, що сприяє прискоренню дослідницьких процесів і підвищенню якості послуг за рахунок зменшення рутинних завдань для редакторів і журналістів. Існуючі дослідження зосереджені на широкому спектрі застосувань ШІ для обробки текстів, але питання стилістичного аналізу текстів ще недостатньо вивчені. Відсутність ефективних інструментів для аналізу гумористичних елементів у текстах зумовлює необхідність розробки специфічних моделей, здатних визначати стиль написання, виявляти плагіат та аналізувати тексти на предмет запозичень.

Розробка інтелектуальної моделі для стилістичного аналізу гумористичних текстів є актуальною: гумористичні жанри стають дедалі популярнішими, і зростає потреба в автоматизації аналізу та оцінювання текстів. Гумор активно використовується для привернення уваги аудиторії в сучасній літературі, журналістиці та ЗМІ. Аналіз таких текстів допоможе краще зрозуміти їхню структуру та вплив, а інструменти штучного інтелекту можуть значно підвищити ефективність роботи редакторів і письменників.

Незважаючи на значний прогрес в обробці природної мови, деякі питання залишаються без відповідей:

– ідентифікація гумористичних елементів. Існуючі моделі не завжди правильно ідентифікують гумор, оскільки він може бути культурним або контекстуальним.

– оцінка стилістичних особливостей. Необхідно розробити моделі, здатні враховувати різні стилістичні тенденції, особливо в гумористичних текстах.

Основна мета – розробити модель для стилістичного аналізу гумористичних текстів із залученням ШІ для підтримки авторів і редакторів в аналізі текстів з точки зору стилю та автентичності. Дані оброблятимуться за допомогою попередньо навченої моделі на основі великого об’єму текстів, що дозволить точно визначити стиль та інші особливості тексту. Розробка інтелектуальної моделі для аналізу гумористичних текстів робить важливий внесок в обробку природної мови і підвищує якість і швидкість аналізу текстів. Це відкриває нові перспективи для дослідників і практиків, зацікавлених у використанні ШІ для аналізу складних текстових структур.

Класичні класифікатори часто демонструють обмеження у своїй ефективності, особливо коли мова йде про обробку неповних або недостатніх даних. Однією з головних проблем є їхня залежність від обсягу та якості вхідних даних. У випадках, коли обсяг даних є недостатнім, такі моделі не можуть належним чином навчитися розпізнавати патерни та особливості, що є критичним для точного класифікування. Ці моделі базуються на попередньо визначених ознаках і передбачають, що всі необхідні характеристики даних доступні для аналізу, однак у реальних умовах ця передумова часто порушується. Неповні дані можуть виникати через різні чинники: обмеженість ресурсів, технічні помилки, або ж особливості самих даних, які є важкодоступними або рідкісними. У таких випадках модель не здатна ефективно узагальнювати інформацію, що призводить до високої ймовірності помилкових класифікацій і зниження загальної продуктивності системи.

Класичні класифікатори зазвичай не мають механізмів для обробки неповних або нечітких даних, що ще більше ускладнює їхнє використання у

випадках неповноти інформації, вони схильні до переобучення на обмеженому наборі даних, що знижує їхню здатність до узагальнення на нових або незнайомих прикладах. У ситуаціях, коли доступні дані є неповними або недостатніми за обсягом, класичні класифікатори не можуть забезпечити надійну і точну класифікацію.

Для вирішення цієї проблеми використовують альтернативні підходи, такі як методи машинного навчання, здатні працювати з неповними даними або алгоритми, що враховують імовірнісні аспекти класифікації. Ці методи можуть бути більш адаптивними до умов нестачі даних і дозволяють покращити точність і надійність класифікації в умовах невизначеності. Використанная класичного класифікатору для вирішення такої задачі, ставить під сумнів точніть на коректність отриманих результатів, через нестачу вхідних даних. Мова навчання – українська, тому і дані для навчання повинні бути українською мовою. На даний момент в українському та зарубіжному інтернет просторі, не існує достатнього за обсягом датасету текстів з подібного роду стилістичними особливостями, з явною унікальністю особливо через те, що видання українською та уркаїнський гумор наразі не є широко розповсюдженими. Досліджень за даною та подібними напрямами у сфері інтелектуальних систем аналізу тексту та гумору в інтернет-ресурсах не виявлено.

2 ОГЛЯД ЛІТЕРАТУРИ ТА ТЕОРЕТИЧНІ ОСНОВИ СТИЛІСТИЧНОГО АНАЛІЗУ

2.1 Роль та стилі гумору у суспільстві

Гумор відіграє важливу роль у світовій культурі комунікації. Вітчизняні фахівці з соціальних комунікацій радять починати публічний виступ з жарту, успішні політики повинні постійно демонструвати своє почуття гумору тощо. Людина, яка не вміє дотепно жартувати, навряд чи досягне вершин ієрархії у будь-якій сфері діяльності, і щоб з'ясувати, чи є у кандидата на керівну посаду почуття гумору, його запрошують на неформальні прийоми, після чого вирішується його професійна доля. Людину, яка не жартує або не реагує на жарти, відкидають без вагань. Школи жартів особливо популярні в Японії, їх відвідують державні службовці, бізнесмени, поліцейські і навіть буддистські монахи. Японці вважають, що гумор - це один з інструментів, який можна успішно використовувати на практиці. [12]

Гумор, як відомо, є глобальним поняттям. Наприклад, висміювання такої людської слабкості, як дурість, є невід'ємною частиною гумористичної культури багатьох країн, що знайшло своє відображення в різноманітних народних приказках: «Дурня повчати - все одно, що небіжчика лікувати» (російська); “Дурневі голова і ноги - біда” (білоруська, українська); “Мудрому досить знака, а дурневі - палиці” (іврит); “Якщо прихистити дурня, він розмалює стіни” (англійська); “Якщо зустрінеш дурня, прикинься зайнятим” (іспанська) тощо. Національні сміхові культури мають відносно схожу структуру, в основі якої лежить принцип інтелектуального задоволення від гумору та почуття гумору. Почуття гумору — це здатність розпізнавати і розуміти сенс смішного.

Сприйняття комічного забезпечує людині естетичну насолоду і підносить її особисто над об'єктом висміювання. Здатність сприймати і

створювати гумор можна побачити в різних формах сміхової культури - в масових розвагах, народній творчості, народних піснях, гумористичних оповіданнях та анекдотах.

До комунікативних форм сміхової культури, що проявляються в міжособистісних стосунках, належать жарти, дотепні дотепи, анекдоти, іронія, сарказм та ін. Гумор може бути добрим чи поганим, що визначає його прямий зв'язок з мораллю. Гумор може бути добрим чи поганим, що визначає його прямий зв'язок з мораллю. Оскільки не всі варіанти культури сміху відповідають морально-етичним принципам поваги і такту, необхідно розрізняти культуру і квазікультуру (лат. quasi - псевдо, ніби) гумору. Суть культури гумору полягає у створенні такого комічного смислу, який не ображає людину, не принижує її гідності і створює умови для взаємного емоційного задоволення при сприйнятті комічного. Тому найбільш прийнятною формою спілкування, що створює позитивний емоційний клімат у міжособистісних стосунках і сприяє взаємній насолоді від сміху, є добрий жарт - смішне висловлювання, комічна дія, що викликає сміх. Культура гумору вимагає нестандартного мислення, інтелектуального напруження і не виставляє нікого на посміховисько. Поширеною формою культури гумору є дотеп.

Дотеп – влучне, стисле, афористичне висловлювання філософського змісту з відчутним жартівливим, сатиричним відтінком.

Неприйнятною формою гумору в міжособистісних стосунках є тенденційний жарт, який може привнести деструктивні моменти в спілкування. Навпаки, він є результатом вираження логічно правильної думки логічно неправильним способом. Усвідомлення цього протиріччя приносить емоційне задоволення. Однак задоволення від самої техніки жарту також пов'язане із задоволенням прихованого бажання принизити іншу людину і продемонструвати свою перевагу. Модний жарт - це інтелектуальний прийом, за допомогою якого створюється ілюзія «пристойності».

У повсякденному житті культура гумору проявляється в жартах, зміст яких не повинен бути осудливим і веселим і не повинен принижувати.

Анекдот (грец. anekdotos – нечуваний) – жанр міського фольклору, комічне оповідання-мініатюра, своєрідна гумористична притча. У 18 столітті анекдоти були історичними творами, літературними розповідями про неймовірні події. Вони не обов'язково були предметом для сміху. Характерними рисами сучасного анекдоту, окрім гумору, є лаконічність, легкість, дотримання головної думки чи ідеї та завершення смішної історії чимось дивовижним і несподіваним. У контексті культури спілкування важливо, щоб анекдот не втрачав своєї естетичної цінності. Культурі гумору протиставляється квазікультура гумору, до модальностей якої належать іронія та сарказм.

Іронія (грец. eironeia – видимість) – форма комізму, що розкриває суперечливу сутність явищ через уїдливе висміювання. В її основі лежить контраст між явним і прихованим, коли за словесно вираженою позитивною оцінкою приховується заперечення і насмішка. Як правило, іронія поєднується з позірно серйозним, поважним і співчутливим тоном, за яким ховається заперечення істини.

Сарказм (грец. sarkazo, буквально: «роздираю плоть») – уїдлива й образлива насмішка, що має на меті принизити об'єкт критики, показати його потворність; уїдливе заперечення у формі перебільшеного визнання. Засобом сарказму є гостре емоційне судження про явища, події та людей без підтексту. Цим сарказм відрізняється від іронії.

Вважається, що глузування, іронія і сарказм беруть свій початок з ритуального сардонічного сміху (сміху перед мертвими). Для багатьох первісних народів сміх символізував життя і слугував талісманом для захисту від помсти мертвих у разі вбивства. Сардонічним сміхом первісна людина намагалася відігнати смерть і продемонструвати повноту і цілісність власного життя. Це був сміх, сповнений страху. У сучасному трактуванні сардонічний

сміх – це злий, уїдливий, презирливий сміх, спрямований на моральне знищення ворога.

Більш зрілою формою сміху (інтелектуальний сміх через заперечення) була антична іронія. Вона виражала негативне ставлення до чогось, не кажучи про це прямо. В її основі лежить лицемірство розуму, ствердження того, що насправді заперечується. У Стародавній Греції слово «eirop» (іронічний), від якого етимологічно походить термін «іронія», перекладалося як обман, шахрайство, богохульство, лицемірство, дисимуляція тощо. Філософ Сократ (бл. 470-399 рр. до н.е.) використовував іронію для викриття невігластва своїх опонентів.

Французький інтуїтивіст Анрі Бергсон (1859-1941) називав іронічний сміх і його крайню форму – сарказм – «анестетиком серця», оскільки він байдужий і жорстокий до об'єкта глузування. У народній мові схоже значення виражає приказка: «Де сміх, там і гріх». З одного боку, він є інтелектуальним вираженням зла, оскільки принижує людську гідність через висміювання, а з іншого - часто виконує важливу соціальну функцію як засіб морального покарання людей розумово млявих і неуспішних.

2.2 Аналіз гумору як стилістичного феномену

Гумор – складне і багатошарове явище, яке відіграє важливу роль у людській комунікації. Як стилістичне явище гумор ґрунтується на різноманітних мовних і немовних засобах, які допомагають досягти комічного ефекту. Він не лише розважає, але й виконує важливі соціальні функції, такі як розрядка конфлікту, критика, висміювання суспільних проблем тощо. Для повного розуміння гумору необхідно розглянути його лінгвістичні, когнітивні та культурні аспекти.

З лінгвістичної точки зору, гумор часто є результатом гри слів, подвійних значень, алітерацій, каламбурів та інших стилістичних прийомів. Каламбури, наприклад, ґрунтуються на двозначності слів або їхній фонетичній

схожості, що може створити комічний ефект через неочікуване значення речення. Інші мовні засоби, такі як перебільшення, іронія та сарказм, також сприяють гумористичному ефекту, заохочуючи сміх через перебільшення або протилежність того, що мається на увазі [13].

Когнітивний аспект гумору - це обробка інформації, яка відбувається в мозку під час сприйняття жартів або кумедних ситуацій. Механізм, який викликає гумор, часто базується на різниці між очікуваним і фактичним результатом, що призводить до когнітивного дисонансу і, зрештою, до сміху. Цей ефект можна пояснити теоріями невідповідності, які стверджують, що гумор виникає тоді, коли існує різниця між тим, чого очікує аудиторія, і тим, що відбувається насправді.

Культурний контекст також має великий вплив на сприйняття гумору. Те, що є смішним в одній культурі, може бути незрозумілим або навіть образливим в іншій. Це пов'язано з різними культурними цінностями, традиціями та соціальними нормами. Наприклад, гумор про політику чи релігію може по-різному сприйматися в різних країнах.

Гумор також має важливу соціальну функцію. Він може слугувати засобом соціальної адаптації, об'єднуючи людей у групи або, навпаки, розмежовуючи різні соціальні класи. Однак гумор також може бути засобом впливу або маніпуляції, якщо він свідомо використовується автором для висміювання певних соціальних явищ або людей.

У літературі та мистецтві гумор часто використовується у формі сатири або пародії, щоб висміяти соціальні негаразди або особисті недоліки. Художники, письменники та сценаристи використовують гумор, щоб привернути увагу до соціальних проблем і зробити їх доступними для широкої аудиторії.

Підсумовуючи, можна сказати, що гумор як стилістичне явище є невід'ємною частиною мовної комунікації і виконує важливі соціальні, психологічні та культурні функції. Він базується на складній взаємодії мовних засобів, когнітивних процесів і культурного контексту, що робить його

багатогранним і цікавим об'єктом дослідження. Сучасні технології, такі як штучний інтелект, дозволяють більш детально аналізувати гумористичні тексти і відкривають нові можливості для розуміння цього складного явища.

2.3 Методологія сучасних лінгвістичних досліджень

Методологія сучасних лінгвістичних досліджень є важливим розділом мовознавства, який вивчає методи, підходи та принципи дослідження мови, вона охоплює не лише способи аналізу мовних явищ, але й теоретичні засади, що лежать в основі наукового пізнання.

Методологія науки в широкому розумінні включає прийоми, засоби, а також метанаукові переконання і цінності, яких дотримуються дослідники, що у вузькому розумінні це вчення про методи й методики дослідження. У межах лінгвістики методологія орієнтована на розкриття природи мови у взаємозв'язку з її носіями, соціумом, культурою, а також на розробку інструментарію для аналізу мовних продуктів.

Методологія має кілька рівнів:

– Філософський – загальні принципи пізнання.

– Загальнонауковий – підходи, спільні для багатьох наук.

– Конкретно-науковий – методи, специфічні для мовознавства.

– Процедурний – методики та техніка дослідження.

Ключовими поняттями методології є метод, методика, процедура, прийом та аспект. Метод визначається як сукупність прийомів, спрямованих на вирішення наукових завдань, а методика – це конкретні дії, що реалізують метод у практиці.

Класифікація методів дослідження за різними критеріями:

– За сферою застосування: міждисциплінарні (загальнонаукові) і внутрішньодисциплінарні (специфічні для лінгвістики).

– За характером відношень між фактами: детерміністичні (жорсткі причинно-наслідкові зв’язки) та імовірнісні (статистичні).

– За підходом до явищ: дедуктивні (від теорії до фактів) та індуктивні (від фактів до теорії).

– Методологія сучасних лінгвістичних досліджень включає як традиційні методи (описовий, історичний, зіставний), так і сучасні (психолінгвістичний, статистичний, метод моделювання). Кожен із них має свої переваги і використовується залежно від об’єкта і завдань дослідження.

–

– 2.4 Парадигми та емпіричні методи лінгвістичних досліджень

–

– Розвиток лінгвістики пов'язаний зі зміною наукових парадигм, до яких належать генетична (еволюційна), таксономічна (структурна) та антропоцентрична парадигми. Кожна з цих парадигм визначає пріоритети у вивченні мови, її структури та функцій. Важливими аспектами сучасної методології лінгвістичних досліджень, що визначають підходи до вивчення мови, є наукові парадигми та емпіричні методи. Парадигми в лінгвістиці відображають панівні концепції та підходи, які змінюються з розвитком науки. Найважливішими парадигмами є:

– Генетична (еволюційна) – зосереджується на історичному розвитку мов, використовуючи порівняльно-історичний метод. Вона вивчає походження мов, їхню еволюцію та зв'язки між спорідненими мовами.

– Таксономічна (структурна) – досліджує внутрішню організацію мови, класифікує її одиниці та вивчає системні зв’язки між ними. Вона стала основою для структуралізму, який трактує мову як цілісну знакову систему.

– Антропоцентрична (комунікативна) – розглядає мову як засіб комунікації, акцентуючи увагу на людському факторі, функціональності мовних одиниць та їхній ролі у взаємодії.

– У сучасній лінгвістиці ці парадигми співіснують, доповнюючи одна одну, що сприяє більш глибокому вивченню мови. Емпіричні методи дослідження є практичним інструментом для збору, аналізу та інтерпретації мовного матеріалу.

– До них належать:

– Спостереження — вивчення мовних явищ у природних умовах без втручання дослідника.

– Експеримент — створення спеціальних умов для перевірки гіпотез, наприклад, у психолінгвістичних дослідженнях.

– Аналіз текстів — дослідження писемних або усних текстів для виявлення закономірностей.

– Статистичні методи — кількісний аналіз мовних явищ для встановлення частотності та зв’язків між ними.

Емпіричні методи дозволяють отримувати об'єктивні дані, необхідні для підтвердження теоретичних концепцій, що є важливим на різних етапах дослідження: від збору матеріалу до побудови висновкі

2.5 Традиційні методи аналізу текстів у лінгвістиці

Традиційні методи аналізу тексту в лінгвістиці [14] формують основу для аналізу структури, змісту та функцій мови. Вони включають низку підходів, які дозволяють дослідникам глибоко аналізувати різні аспекти текстів, виявляти закономірності у використанні мови та розуміти її соціальну і культурну роль.

Одним із найважливіших традиційних методів є граматичний і синтаксичний аналіз, який використовується для визначення структури речення, його компонентів і зв'язків між ними. Цей метод спрямований на визначення граматичних категорій, таких як частини мови, відмінки, часи та інші граматичні особливості. Це допомагає зрозуміти правила синтаксису та розпізнати специфічні стилістичні особливості тексту.

Лексико-семантичний аналіз – ще один важливий метод аналізу значення слів і речень та їхньої ролі у створенні сенсу тексту. За допомогою цього підходу можна проаналізувати семантичні поля, синоніми та багатозначність слів, що допомагає виявити такі стилістичні прийоми, як

метафора, метонімія та алегорія. Лексико-семантичний аналіз також використовується для аналізу особливостей авторського стилю та визначення жанрових характеристик текстів.

Одним з найважливіших традиційних методів є аналіз тексту, який розглядає текст як смислову одиницю. Це фокусується на таких аспектах, як зв'язність і когерентність тексту, зв'язки між реченнями, абзацами та частинами тексту аналізуються для того, щоб визначити логіку тексту, його тематичні лінії та структуру.

Стилістичний аналіз — ще один важливий метод аналізу засобів вираження, які використовуються для досягнення певного художнього чи емоційного ефекту. Він включає такі аспекти, як використання тропів, ідіом, інтонації та інших стилістичних прийомів, він допомагає зрозуміти, як автор досягає певного впливу на читача і які засоби використовує для створення художнього образу.

Інтертекстуальний аналіз – це вивчення зв'язків між текстом та іншими текстами з метою виявлення впливів, алюзій та ремінісценцій. Цей підхід дає можливість зрозуміти, як текст пов'язаний з культурним контекстом і як цитати або відсилання до інших творів використовуються для створення додаткових смислів.

Функціональний аналіз текстів зосереджується на їхній комунікативній меті та аналізі мовленнєвих актів. Він досліджує, як текст виконує свою основну функцію – інформативну, експресивну, фатичну тощо, і також – яким чином структура тексту є значущою щодо цієї функції.

Традиційні методи аналізу тексту в лінгвістиці пропонують багатовимірний підхід до вивчення мови, що дозволяє глибоко аналізувати як окремі тексти, так і цілі мовні системи, вони є основою для подальших досліджень і розробки новітніх методів аналізу, які враховують як класичні підходи, так і сучасні технологічні інструменти.

Види аналізу:

– Герменевтичний аналіз — цей метод зосереджується на інтерпретації текстів, особливо художніх і філософських. Він враховує історичний та культурний контексти, що допомагає зрозуміти глибинні смисли, які можуть бути прихованими або багатозначними.

– Прагматичний аналіз — вивчає, як текст функціонує у комунікативній ситуації, враховуючи контекст, наміри автора і очікування аудиторії. Це дозволяє аналізувати, як мовні одиниці використовуються для досягнення практичних цілей.

– Квантитативний аналіз — включає використання статистичних методів для вивчення тексту. Це можуть бути підрахунки частоти використання певних слів або граматичних конструкцій, що дає змогу виявити домінантні теми або стилістичні тенденції.

– Когнітивний аналіз — зосереджений на тому, як текст відображає мисленнєві процеси, метафоричні структури та концептуальні моделі. Це допомагає зрозуміти, як людина обробляє інформацію і як це впливає на структурування тексту.

– Діахронічний аналіз — досліджує зміни у використанні мови та текстових структур у різні історичні періоди. Це дозволяє простежити еволюцію мовних норм і стилістичних прийомів.

2.6 Стилістичний аналіз тексту

Стилістичний аналіз тексту — це комплексний метод аналізу тексту, спрямований на визначення мовних засобів, використаних автором для досягнення комунікативної мети. Аналіз є важливим етапом у формуванні лінгвостилістичної компетентності, особливо для майбутніх учителів української мови і літератури.

Мета стилістичного аналізу тексту — проаналізувати, як мовні засоби різних рівнів (фонетичного, лексичного, граматичного) впливають на художню образність, вираження думки та емоційний вплив тексту на читача.

Розуміння специфіки функціонування мовних одиниць у тексті є основою у визначенні стилі твору а також дозволяє оцінити ефективність використання мовних засобів у контексті жанру та функціонального стилю. Види стилістичного аналізу:

– За повнотою виконання:

1) частковий (аналіз окремих аспектів тексту, наприклад, лексичних засобів);
2) повний (охоплює всі рівні мовної організації).

– За функціональними стилями:

1) аналіз текстів художнього, наукового, публіцистичного, офіційно-ділового, розмовного стилів.

– За рівнями мовної організації:

1) фонетико-стилістичний;
2) лексико-стилістичний;
3) морфолого-стилістичний;
4) синтаксично-стилістичний.

Етапи стилістичного аналізу:

– Підготовчий етап:

1) ознайомлення з теоретичними основами стилістики;
2) вивчення стилістичних характеристик мовних одиниць.

– Початковий етап:

1) читання тексту, його сприйняття та розуміння;
2) визначення жанру та стилю тексту.

– Основний етап:

1) виявлення стилістично маркованих мовних одиниць;
2) аналіз тропів, стилістичних фігур, експресивних засобів;
3) визначення їхньої ролі у створенні художнього образу та вираженні авторського задуму.

– Завершальний етап:

1) підготовка письмового звіту про аналіз;
2) обговорення результатів.

Принципи стилістичного аналізу:

– Системність — текст розглядається як цілісна система.

– Міждисциплінарність — враховуються зв’язки з літературознавством та культурологією.

– Поетапність — аналіз проводиться поступово, з акцентом на окремих аспектах тексту.

Стилістичний аналіз тексту — це важливий інструмен, що дозволяє глибше зрозуміти природу мови та її стилістичні ресурсию. За допомогою нього не лише підвищує рівень професійної компетентності, але й допомагає ефективно використовувати отримані знання у власній педагогічній, науковій, дослідницькій діяльності та роботі.

Таким чином доречна розробка інтелектуальної моделі стилістичного аналізу гумористичного тексту при умові неповноти вхідних даних. Натомість, на заміну класичного класифікатора, запропоновано рішення з використанням методу навчання на неповних даних. Подібний метод дозволить класифікувати ті дані, з котрими раніше не можливо було працювати, а саме приводячи множину унікальних ознак до одного, за яким можно було б чітко ідентифікувати текст. Проводити аналіз тексту на основі отриманого результату, використовуючи такі параметри як: кількість речень, кількість слів, середня довжина речень, коєфіцієнт заголовних букв, кількість знаків пунктуації, кількість емоційних слів.

Подібний аналіз надає можливість швидко та ефективно працювати з текстами, що надає можливість звільнити час від рутиного опрацювання тексту, на більш важливі речі у роботі редакторів, керівників видавництв, учителів, викладачів в університетах.

3 РОЗРОБКА СТРУКТУРИ МОДЕЛІ

3.1 Штучний інтелект

Сучасний штучний інтелект (ШІ) — одна з найцікавіших і найдинамічніших галузей досліджень і технологічних розробок. Він працює на перетині інформатики, математики, психології та інших дисциплін, трансформуючи наше уявлення про технології та їхню роль у сучасному суспільстві.

AI is ...

"The use of algorithms. The term 'algorithm' refers to a specific instruction for solving a problem or performing a calculation."

"The imitation of all human intellectual abilities by computers."

"The imitation of various complex human skills by machines."

"Technology that can function appropriately and with foresight in its environment."

"Systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals."

Рисунок 3.1 – Різні визначення ШІ

Існує кілька визначень штучного інтелекту, як показано на рисунку 3.1, і важко виокремити лише одне, оскільки всі вони мають своє місце, але якщо узагальнити всі ці терміни, то штучний інтелект - це галузь комп'ютерних наук, яка фокусується на створенні систем і технологій, здатних виконувати завдання, які зазвичай вимагають людського інтелекту. Хоча саме поняття ШІ

може здатися абстрактним, це область, орієнтована на конкретні цілі, такі як розпізнавання мови, аналіз даних, вирішення проблем, планування, машинне навчання і багато іншого.

Одне з найпоширеніших визначень ШІ - це здатність комп'ютера виконувати завдання, які зазвичай вимагають людського мислення. Ці завдання включають розуміння природної мови, аналіз зображень, розпізнавання образів і багато інших. ШІ прагне імітувати можливості людського мозку таким чином, щоб комп'ютер міг приймати рішення, вчитися на досвіді та адаптуватися до мінливого середовища [15].

3.1.1 Ключові поняття у сфері штучного інтелекту

Щоб краще зрозуміти штучний інтелект, корисно розглянути деякі ключові поняття, пов'язані з цією галуззю:

– Машинне навчання — це галузь ШІ, яка фокусується на розробці алгоритмів, що дозволяють комп'ютерним системам навчатися на основі даних і досвіду. Машинне навчання дозволяє комп'ютерам вирішувати завдання, які раніше здавалися неможливими для автоматизації.

– Нейронні мережі — тип математичної моделі, натхненний структурою людського мозку. Нейронні мережі широко використовуються в машинному навчанні, особливо в задачах, пов'язаних з розпізнаванням зображень, аналізом тексту і обробкою мови.

– Обробка природної мови (NLP) — це область ШІ, яка займається аналізом, розумінням і створенням тексту на природній мові. NLP дозволяє комп'ютерам спілкуватися і співпрацювати з людьми в більш природний спосіб.

– Розпізнавання образів (Pattern Recognition) — це здатність систем ШІ виявляти закономірності в даних, що корисно в задачах аналізу зображень, аудіо або даних.

– Обчислювальний інтелект (ШІ) включає в себе різні техніки і методи, що використовуються в ШІ, такі як еволюційні алгоритми, системи на основі нечіткої логіки, системи на основі агентів та інші.

Алгоритми машинного навчання в основному поділяються на чотири категорії: контрольоване навчання, неконтрольоване навчання, напівконтрольоване навчання та навчання з підкріпленням, як показано на рисунку 3.2 Кожен тип методів навчання розглядається нижче, а також ступінь, до якого вони можуть бути застосовані до реальних проблем [16].

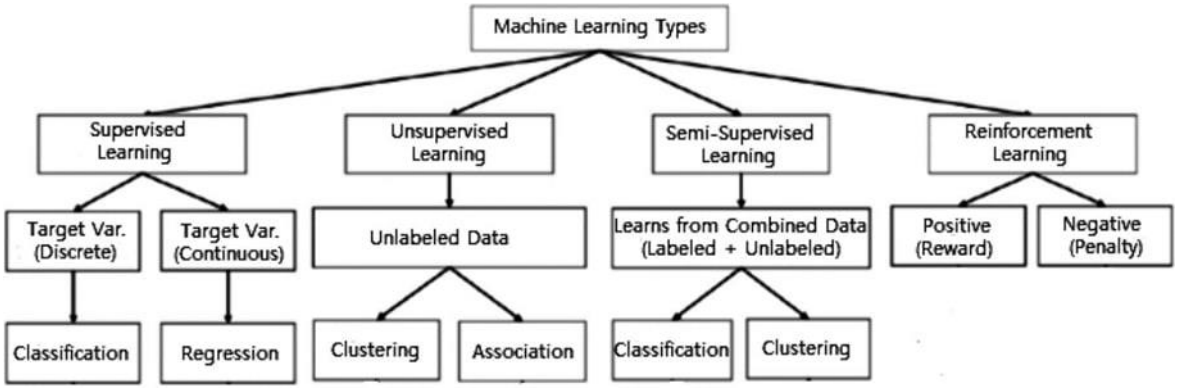

Рисунок 3.2 – Методи машинного навчання

Контрольоване навчання: контрольоване навчання – це, як правило, завдання машинного навчання, яке вивчає функцію, що відображає вхідні дані у вихідні дані на основі прикладів пар вхідних і вихідних даних. Для виведення функції використовуються марковані навчальні дані та набір навчальних прикладів. Навчання під контролем здійснюється, коли потрібно досягти певних цілей на основі визначеного набору вхідних даних, тобто підхід на основі завдань. Найпоширенішими завданнями під контролем є «класифікація», яка розділяє дані на різні класи, і «регресія», яка узгоджує дані. Наприклад, передбачення класу або настрою фрагмента тексту, такого як твіт або відгук про товар, тобто класифікація тексту, є прикладом керованого навчання.

Неконтрольоване навчання: Некероване навчання аналізує немарковані набори даних без втручання людини, тобто це процес, керований даними. Він широко використовується для вилучення генеративних ознак, виявлення значущих тенденцій і структур, кластеризації результатів і дослідницьких цілей. Найпоширенішими завданнями неконтрольованого навчання є кластеризація, оцінка щільності, вивчення особливостей, зменшення розмірності, пошук асоціативних правил, виявлення аномалій тощо.

Напівконтрольоване навчання: напівконтрольоване навчання можна визначити як гібридизацію вищезгаданих контрольованих і неконтрольованих методів, оскільки воно оперує як з міченими, так і з немічеными даними. Таким чином, воно знаходиться між «неконтрольованим» навчанням і «контрольованим» навчанням. У реальному світі мічені дані можуть бути рідкісними в деяких контекстах, а немічені дані є численними, тому напівконтрольоване навчання є корисним. Кінцевою метою напівкерованого навчання є забезпечення кращого результату прогнозування, ніж той, що отриманий за допомогою лише мічених даних моделі. Деякі з областей застосування, в яких напівкероване навчання включають машинний переклад, виявлення шахрайства, маркування даних і класифікацію текстів.

Навчання з підкріпленням: Навчання з підкріпленням - це тип алгоритму машинного навчання, який дозволяє програмним агентам і машинам автоматично оцінювати оптимальну поведінку в конкретному контексті або середовищі з метою підвищення продуктивності, тобто підхід, заснований на навколишньому середовищі. Цей тип навчання ґрунтується на винагороді або покаранні, а кінцевою метою є використання знань, отриманих від суб'єктів навколишнього середовища, для вжиття заходів, спрямованих на збільшення винагороди або мінімізацію ризику. Це потужний інструмент для навчання моделей штучного інтелекту, який може допомогти підвищити рівень автоматизації або оптимізувати роботу сучасних систем, таких як робототехніка, завдання автономного водіння, виробнича логістика та

логістика ланцюгів поставок, але не рекомендується використовувати його для вирішення базових або простих завдань.

3.1.2 Застосування NLP

Наступним важливим питанням, є обробка природної мови (NLP). Обробка природної мови поєднує обчислювальну лінгвістику зі статистичними моделями та машинним навчанням, що дозволяє комп'ютерам і цифровим пристроям розпізнавати, розуміти і генерувати текст і мову [17].

Як галузь штучного інтелекту, NLP лежить в основі додатків і пристроїв, які можуть:

– перекладати текст з однієї мови на іншу;

– реагувати на письмові або усні команди;

– розпізнавати або аутентифікувати користувачів на основі голосу;

– узагальнювати великі обсяги тексту;

– оцінювати наміри або настрій тексту або мовлення;

– генерувати текст, графіку або інший контент на вимогу, часто в режимі реального часу.

Виходячи з вищесказаного, можна виділити місця, де використовується NLP [17]. Машинний переклад Google Translate є прикладом широко доступної технології NLP. Корисний машинний переклад вимагає більше, ніж простої заміни слів однієї мови словами іншої. Ефективний переклад повинен точно передавати зміст і тональність вхідної мови і перекладати його в текст з тим же значенням вихідною мовою. Інструменти машинного перекладу роблять величезні кроки в плані точності. Чудовий спосіб протестувати будь-який інструмент машинного перекладу - це перекласти текст однією мовою, а потім перекласти його назад мовою оригіналу.

Віртуальні агенти та чат-боти, віртуальні агенти, такі як Siri від Apple і Alexa від Amazon, використовують розпізнавання мови для розпізнавання шаблонів у голосових командах і генерування природної мови, щоб

відповідати відповідними діями або корисними коментарями. Чат-боти виконують ту ж саму магію у відповідь на введені текстові повідомлення. Найкращі з них також вчаться розпізнавати контекстні підказки про людські запити і використовувати їх, щоб з часом надавати ще кращі відповіді або варіанти. Ще одне вдосконалення для цих додатків — це відповіді на запитання, або здатність відповідати на наші запитання доречними та корисними відповідями своєю мовою.

Аналіз настроїв у соціальних мережах NLP стало важливим бізнес-інструментом для виявлення прихованих даних з каналів соціальних мереж. Аналіз настроїв може аналізувати мову, яка використовується в постах, відповідях, відгуках тощо, щоб виявити ставлення та емоції у відповідь на продукти, акції та події - інформацію, яку бізнес може використовувати в дизайні продуктів, рекламних кампаніях тощо.

Реферування тексту використовує методи NLP для «перетравлення» величезних обсягів цифрового тексту і створення резюме та обговорень для індексів, дослідницьких баз даних або для зайнятих читачів, які не мають часу читати повні тексти. Найкращі програми для узагальнення тексту використовують семантичний висновок і генерацію природної мови (NLG), щоб додати корисний контекст і висновки до анотацій.

3.2 Мовні моделі

У сучасному світі, що швидко змінюється, технологічний прогрес відіграє ключову роль у формуванні повсякденного життя. Однією з галузей, яка значно розвинулася за останні кілька років, є сфера штучного інтелекту (ШІ). Штучний інтелект – міждисциплінарна галузь, зосереджена на створенні комп'ютерних систем, здатних виконувати завдання, які зазвичай вимагають людського мислення. У цій галузі мовні моделі стали одним із найважливіших інструментів, що революціонізували лінгвістичну комунікацію, аналіз текстів та обробку мовлення. У цьому розділі розглянуто принципи побудови мовних

моделей, їх визначення, завдання і функції, а також на їхній вплив на сферу штучного інтелекту.

3.2.1 Визначення мовних моделей

Мовна модель використовує машинне навчання для створення розподілу ймовірностей слів, які використовуються для прогнозування найбільш ймовірного наступного слова в реченні на основі попереднього. Мовні моделі навчаються на тексті і можуть використовуватися для [18]: отримання оригінального тексту, прогнозування наступного слова в тексті, розпізнавання мови, оптичного розпізнавання символів і розпізнавання рукописного тексту.

Мовна модель – це розподіл ймовірностей слів або послідовностей слів. На практиці вона дає ймовірність того, що дана послідовність слів є «правильною». Правильність у цьому контексті не означає граматичну правильність. Натомість це означає, що вона схожа на те, як люди пишуть, тобто на те, чого навчає мовна модель, це важливо, адже мовна модель — це просто інструмент для включення великої кількості інформації в стислий спосіб, який можна повторно використовувати в контекстах поза вибіркою [18].

3.2.2 Типи мовних моделей

Існує два типи мовних моделей:

– нейронно-імовірнісна мовна модель.

– сучасні мовні моделі на основі нейронних мереж.

Проста ймовірнісна модель мови (рис. 3.3) будується шляхом обчислення ймовірності $n-$грам. $N-$грама — це послідовність з n слів, де n - ціле число, більше нуля. Ймовірність $n-$грами — це умовна ймовірність того, що останнє слово $n-$грами слідує за конкретною $n-1$ грамою (за винятком останнього слова).

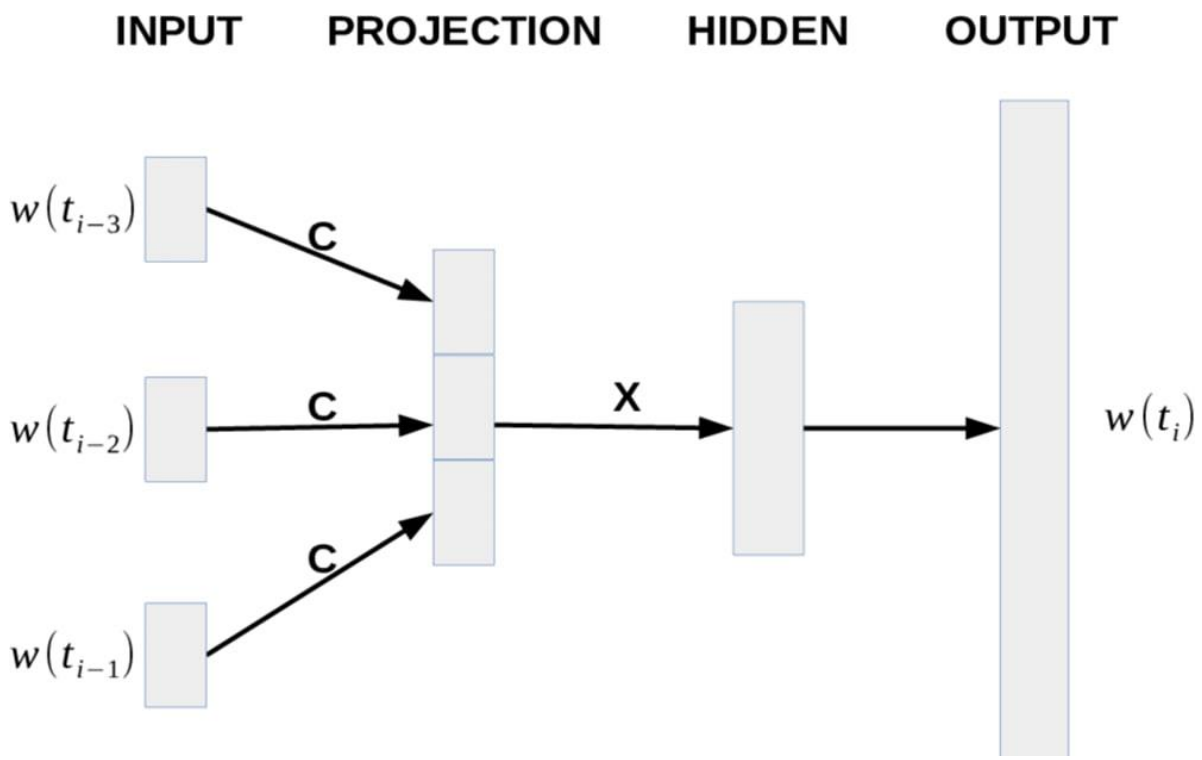

Рисунок 3.3 – Нейронно-імовірнісна мовна модель.

Це частка входжень останнього слова після $n-1$ грами, за винятком останнього слова. Це поняття є припущенням Маркова. За наявності $n-1$ грама (теперішнього), ймовірність появи $n-$ грамів (майбутнього) не залежить від $n-2, n-3$ і т.д. грамів (минулого). Такий підхід має очевидні недоліки. Найголовніший з них полягає в тому, що лише попередні n слів впливають на розподіл ймовірності наступного слова. Складні тексти мають глибокий контекст, який може мати вирішальний вплив на вибір наступного слова. Тому наступне слово може бути неочевидним з попередніх n слів, навіть якщо n дорівнює 20 або 50. Термін впливає на вибір попереднього слова: слово United є набагато більш вірогідним, якщо за ним слідує States of America. Це можна назвати контекстуальною проблемою. Крім того, очевидно, що такий підхід погано масштабується. Зі збільшенням розміру (n) кількість можливих перестановок стрімко зростає, навіть якщо більшість з них ніколи не зустрічаються в тексті. І всі ймовірності, що зустрічаються (або всі номери $n-$

грам), повинні бути обчислені і збережені. Крім того, $n-$грами, що не зустрічаються, створюють проблему розрідженості, оскільки деталізація розподілу ймовірностей може бути досить низькою. Ймовірності слів мають мало різних значень, тому більшість слів мають однакову ймовірність [18].

Для подолання цих недоліків були розроблені різні методи. Одним з них є використання згладжування Лапласа або інших технік згладжування, щоб зменшити вплив рідкісних або невідомих n-грамів. Також використовуються нейронні мережі, такі як рекурентні нейронні мережі (RNN) та трансформери, які можуть захоплювати довгострокові залежності та контекст у тексті. Ці підходи дозволяють моделі враховувати більш широкий контекст, ніж прості n-грамові моделі, що покращує точність передбачень.

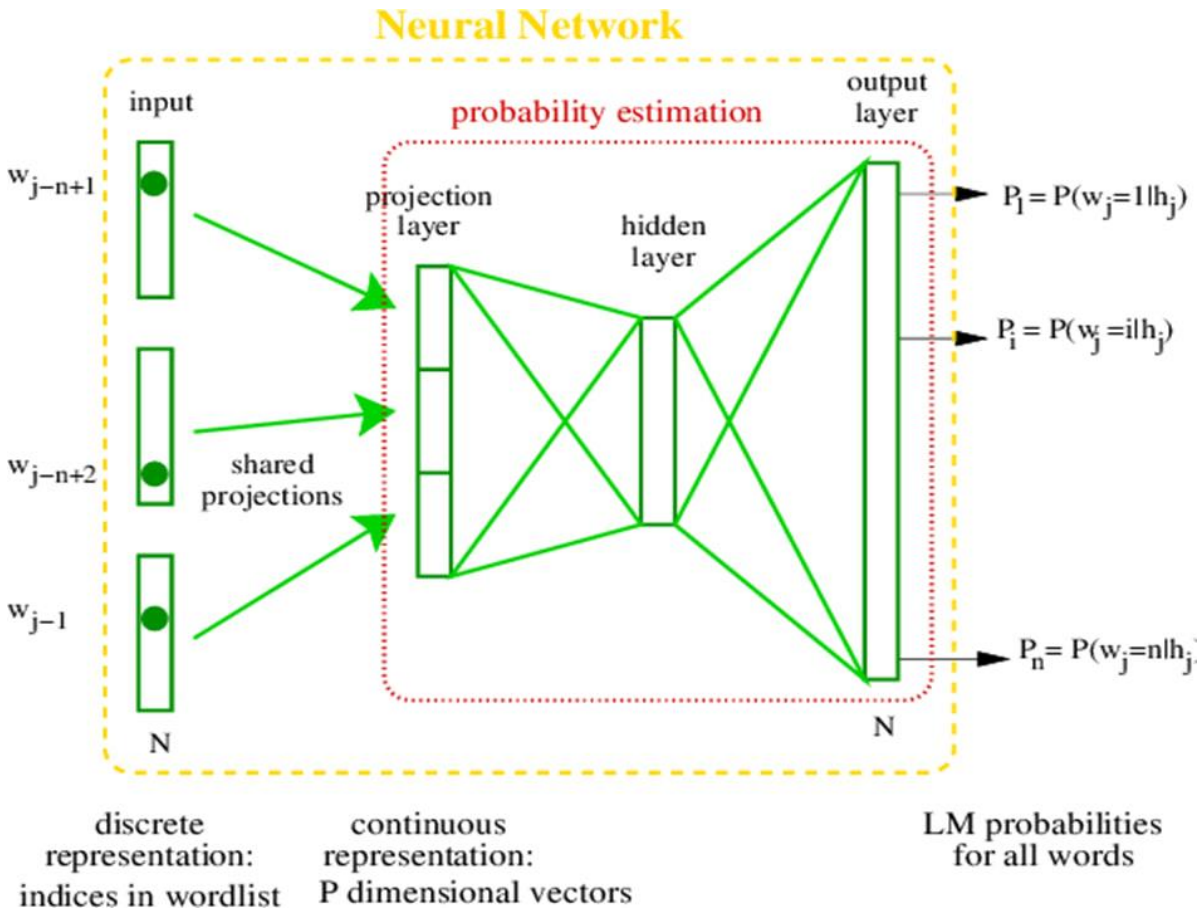

Рисунок 3.4 – Мовні моделі на основі нейронних мереж

Мовні моделі на основі нейронних мереж (рисунок 3.4) полегшують проблему розрідженості за рахунок способу кодування вхідних даних. Шари

вбудовування слів створюють вектор довільного розміру для кожного слова, який також містить семантичні зв'язки. Ці безперервні вектори створюють необхідну деталізацію розподілу ймовірності наступного слова. Крім того, мовна модель є функцією, як і всі нейронні мережі з множинними матричними обчисленнями, тому немає необхідності зберігати всі n-грами для отримання розподілу ймовірності наступного слова [18].

3.2.3 Завдання та функції мовних моделей

Мовні моделі мають ряд ключових функцій і використовуються для вирішення різноманітних завдань в області обробки природної мови. Ось деякі з основних завдань, які можуть виконувати мовні моделі:

Узагальнення тексту, мовна модель здатна узагальнювати довгі тексти, скорочуючи їхню довжину, зберігаючи при цьому суть і ключову інформацію. Це корисно для аналізу документів, створення резюме наукових статей і новин, що дозволяє швидше зрозуміти зміст.

Генерація тексту мовні моделі можуть генерувати текст різних стилів і тематик. Їх можна використовувати для створення описів продуктів, статей і творчих текстів, де модель може генерувати тексти на основі заданого контексту.

Розпізнавання та синтез мовлення на додаток до обробки тексту, моделі використовуються для розпізнавання та генерування мови. Завдяки цій функції можна спілкуватися з мовною моделлю так само, як і з людиною (наприклад, OpenAI нещодавно представила GPT-4o, яка має цю функцію).

Анотацію зображень мовної моделі можна використовувати для опису та категоризації вмісту зображень. Цю технологію використовують такі компанії, як Google і Bing, для навчання нейронних мереж розрізняти різні об'єкти на зображеннях. Іншим відомим прикладом є автопілот Tesla, який також працює, розпізнаючи об'єкти в камері, такі як дорожні знаки, інші автомобілі та пішоходів.

Машинний переклад — це інструмент для автоматичного перетворення тексту з однієї мови на іншу. Ці системи перекладу використовують передові мовні моделі, які навчаються на великих наборах текстових даних різними мовами. У бізнесі, науці, подорожах або на онлайн-платформах машинний переклад дає змогу спілкуватися між людьми, які розмовляють різними мовами.

Рекомендаційні системи використовуються у сфері аналізу вподобань користувачів і надання рекомендацій щодо різних продуктів, відео, контенту або послуг. Такі системи працюють, аналізуючи великі обсяги даних про поведінку користувачів, їхні вподобання, історію покупок або взаємодію з контентом. В електронній комерції рекомендаційні системи використовують моделі для аналізу купівельної поведінки користувачів, а потім рекомендують їм продукти та послуги на основі отриманих результатів. У випадку з потоковими сервісами лінгвістичні моделі використовуються для аналізу поведінки користувачів, наприклад, переглянутих відео, рейтингів або жанрових уподобань. На основі цих даних рекомендаційні системи можуть пропонувати користувачам контент, який відповідає їхнім користувачеві контент, який відповідає його вподобанням, рекомендувати нові фільми на основі його історії переглядів або подібних вподобань інших користувачів.

Виявлення шахрайства мовної моделі можна використовувати для виявлення шахрайства шляхом аналізу тексту, виявлення підозрілих шаблонів у повідомленнях, транзакціях та інших типах даних. Їх можна використовувати для створення автоматизованих систем виявлення шахрайства в різних сферах, таких як фінанси, електронна комерція, або для аналізу поведінки користувачів з метою виявлення шахрайських дій чи виявлення ботів шляхом аналізу навігації користувача в додатку, а також кількості та інтервалу між запитами, що надсилаються користувачем.

Генерація коду, ще одна сфера, де можна використовувати мовні моделі. Їх можна навчити або налаштувати так, щоб вони генерували фрагменти вихідного коду різними мовами програмування. Ці моделі можуть

підтримувати програмістів, пропонуючи код, пропонуючи відповідні фрагменти коду або автоматично генеруючи фрагменти коду на основі опису проблеми або завдання. Однак згенерований код може знадобитися перевірити і виправити вручну, оскільки мовні моделі іноді можуть генерувати неправильні або нефункціонуючі фрагменти коду, особливо для більш складних питань програмування. За словами автора, генерація коду добре працює для невеликих завдань, таких як завершення розпочатого методу, написання коментарів до методу або створення базового шаблону сторінки. Наочним прикладом моделі генерації коду є Github's Copilot. Ця модель призначена виключно для генерації коду і прискорює процес розробки.

3.3 Опис структури нейронної мережі

Процес створення моделі для оцінювання близкісті авторського стилю

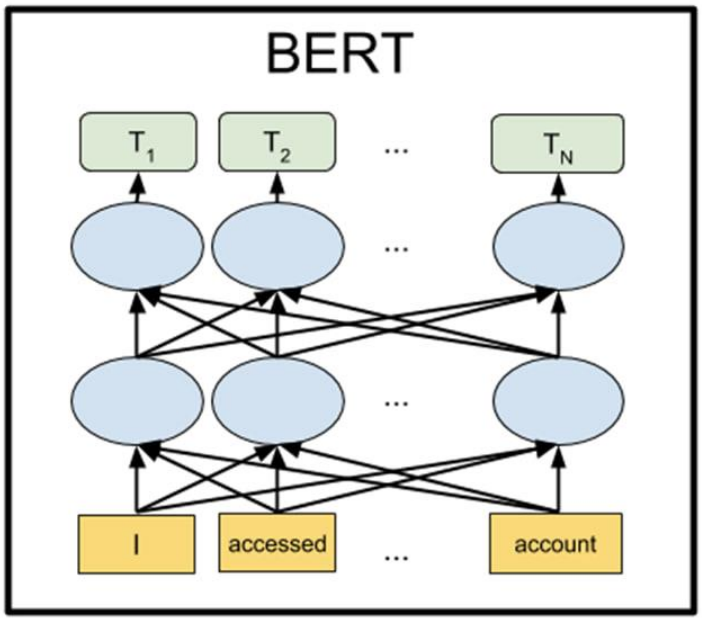

Рисунок 3.5 – Структура моделі BERT

тексту до еталонів, використовуючи передові методи NLP, зокрема, трансформери, такі як модель BERT (рисунок 3.5) або XLM-RoBERTa .

Формат даних у вигляді текстів зберігаються в структурованому форматі, наприклад, у таблиці, PSV або CSV , де кожен рядок містить текст і відповідного автора. Попередня обробка даних є обов'язковою, видаляються зайві символи, HTML-теги, а також приводиться текст до нижнього регістру для уніфікації

Текст розбивається на токени (слова або підслова), які будуть використовуватися моделлю BERT. Для цього можна використовувати вбудований токенізатор.

Токени, введені користувачем, вважаються вхідними даними для моделей машинного навчання. Однак моделі розуміють тільки числа, а не текст, тому ці вхідні дані необхідно перетворити в числовий формат, званий «вбудованими вхідними даними». Вхідні вкладення представляють слова у вигляді чисел, які потім можуть оброблятися моделями машинного навчання.

Ці вкладення подібні до словника, який допомагає моделі зрозуміти значення слів, поміщаючи їх у математичний простір, де схожі слова розташовані поруч одне з одним. Під час навчання модель вчиться створювати ці вкладення, щоб схожі вектори представляли слова зі схожим значенням.

Вхідний текст подається у вигляді:

$$input = [CLS]token1token2 \dots tokenN[SEP]input = \\ = [CLS]token1token2 \dots tokenN[SEP] \quad (3.1)$$

Дані розділяються на тренувальну (80%) і тестову (20%) вибірки для оцінки продуктивності моделі. Використовучи модель BERT бібліотеки Hugging Face Transformers. Попередньо навчена версія моделі буде донавчатися на зібраних даних для завдання класифікації. Параметри моделі, як довжина послідовності (наприклад 128 токенів), розмір мініпакета (наприклад 16) і кількість епох (наприклад 3).

Тексти перетворюються на числові послідовності з використанням токенізатора , де кожному тексту додаються спеціальні токени ([CLS] і [SEP]).

Під час опрацювання природної мови порядок слів у реченні має вирішальне значення для визначення сенсу речення. Однак традиційні моделі машинного навчання, такі як нейронні мережі, за своєю суттю не розуміють порядок вхідних даних. Щоб вирішити цю проблему, можна використовувати позиційне кодування для кодування положення кожного слова у вхідній послідовності у вигляді набору чисел. Ці числа можна ввести в модель Transformer разом із вхідними поданнями. Включивши позиційне кодування в архітектуру Transformer, GPT може більш ефективно розуміти порядок слів у реченні та генерувати граматично правильний і семантично значущий результат.

Кодер складається зі стеку з $N = 6$ однакових шарів. Кожен шар має два підшарів. Перший – це механізм самоуваги з декількома головками, а другий – проста, позиційно повністю зв'язана мережа прямого зв'язку. Ми використовуємо залишковий зв'язок навколо кожного з двох підшарів, з подальшою нормалізацією шарів . Тобто, вихід кожного підшару має вигляд

$LayerNorm(x + Sublayer(x))$, де Sublayer(x) – функція, реалізована підшаромсамим підшаром.

Щоб полегшити ці залишкові зв'язки, всі підшари в моделі, а також шари, що вбудовуються шари, створюють вихідні дані розмірністю $dmodel = 512$.

Замість того, щоб виконувати одну функцію уваги з ключами, значеннями та запитами розмірності $dmodel$, ми вирішили лінійно спроектувати запити, ключі та значення h разів, що корисніше лінійно спроектувати запити, ключі та значення h разів з різними, вивченими лінійними проекціями на dk, dk, dv розмірності відповідно. На кожну з цих спроектованих версій запитів, ключів і значень ми потім паралельно виконуємо функцію уваги, отримуючи $dv-$вимірну. Точкові добутки стають великими, припустимо, що компоненти q і k є незалежними випадковими випадкові величини із середнім значенням 0 та дисперсією 1. Тоді їхній точковий добуток:

$$q \cdot k = \sum_{i=1}^{dk} q_i k_i \tag{3.2}$$

має середнє значення 0 і дисперсію dk вихідні значення. Вони об'єднуються і ще раз проектуються, в результаті чого отримуємо остаточні значення, як показано на рисунку 3.6.

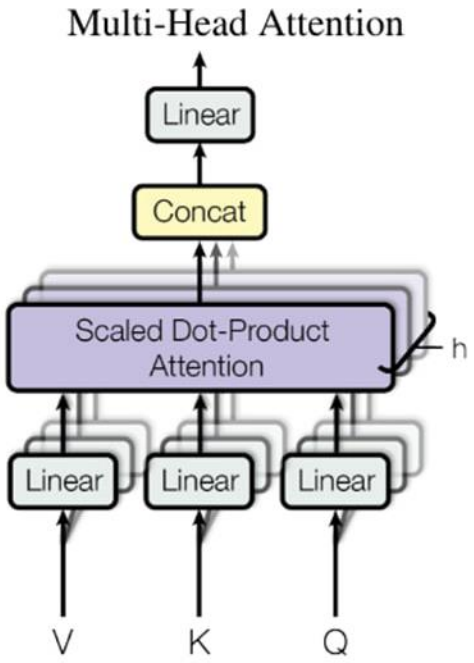

Рисунок 3.6 – Багатоголова увага складається з декількох рівнів уваги, що працюють паралельно.

Багатоголова увага дозволяє моделі спільно обробляти інформацію від різних респондентів.

Для навчання моделі використовую XLM-RoBERTa, придатну для завдання багатокласової класифікації, також оцінювання моделі виконується за рахунок метрика accuracy, яка показує частку правильно класифікованих текстів, а також незалежних алгоритмів оцінювання текстів. Кросс-ентропійна функція втрат для задачі класифікації виглядає так:

$$\mathcal{L} = -\sum_{i=1}^{N} \sum_{j=1}^{C} y_{ij} \log \hat{y}_{ij} \ , \tag{3.3}$$

де N – кількість зразків в навчальному наборі;

C – кількість класів;

y_{ij} – істинна мітка класу;

$\hat{y}_{ij}$ – передбачена ймовірність класу.

Процес навчання та валідації включає кілька епох, протягом яких модель багаторазово проходить через тренувальні дані, оптимізуючи свої параметри.

BERT, завдяки своїй архітектурі трансформерів і здатності захоплювати контекст в обох напрямках, є потужним інструментом для оцінки стилю авторського тексту. Використовуючи методи, описані вище, можна навчити модель ефективно розрізняти стилі авторськиъ текстів, що є важливим кроком в аналізі та розумінні тексту природною мовою.

3.4 Алгоритм навчання задач класифікації RandomForestClassifier

RandomForestClassifier – це універсальний і широко використовуваний алгоритм для машинного навчання задач класифікації - метод ансамблевого навчання, який створює кілька дерев рішень під час навчання і виводить клас, який відповідає типу класу (класифікації) кожного дерева.

RandomForestClassifier був представлений у 2001 році Лео Брейманом як покращення алгоритму дерева рішень. Алгоритм працює за принципом бутстреп-агрегування, згідно з яким кожне дерево будується з випадкової підмножини набору навчальних даних, а для розділення вузлів використовується випадкова підмножина ознак:

Ключові кроки алгоритму RandomForest:

– Bootstrap вибірка — випадкові вибірки беруться з заміною з початкових навчальних даних для створення декількох підмножин.

– Вибір ознак — у кожному вузлі дерева вибирається випадкова підмножина ознак для визначення найкращого розбиття.

– Побудова дерева — кожне дерево вирощується до максимально можливого розміру без обрізки.

– Агрегація — для задач класифікації, в якості остаточного прогнозу використовується режим передбачення, для выбору найкращого рішення.

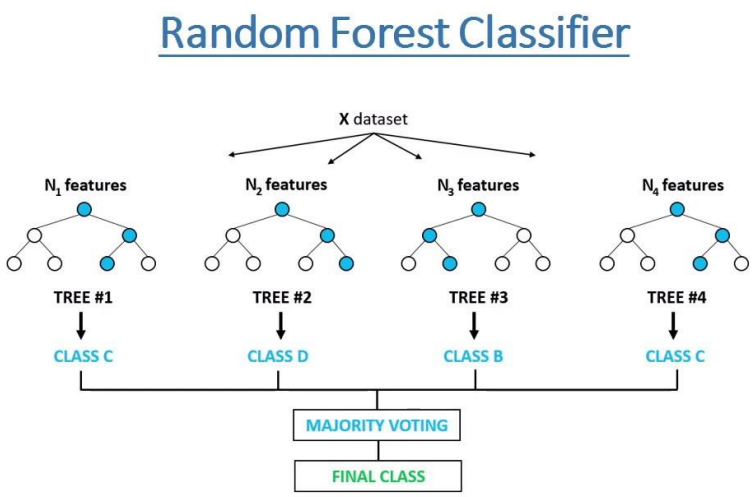

Рисунок 3.7 – Алгоритм роботи RandomForest

Починаючи з навчального набору даних, що містить вектор ознак і відповідну мітку класу, дерева рішень створюються за допомогою алгоритму RandomForest. Цей метод має низку переваг: висока точність - завдяки ансамблю дерев рішень, стійкість до перенавчання - алгоритм зменшує перенавчання шляхом усереднення декількох дерев і дає уявлення про важливість різних ознак для прогнозу.

З одного боку, використання дуже ефективного алгоритму призводить до труднощів його застосування до великих наборів даних, оскільки навчання великої кількості дерев може вимагати значних обчислювальних затрат, а також виникають труднощі в розумінні моделі, коли є велика кількість дерев рішень.

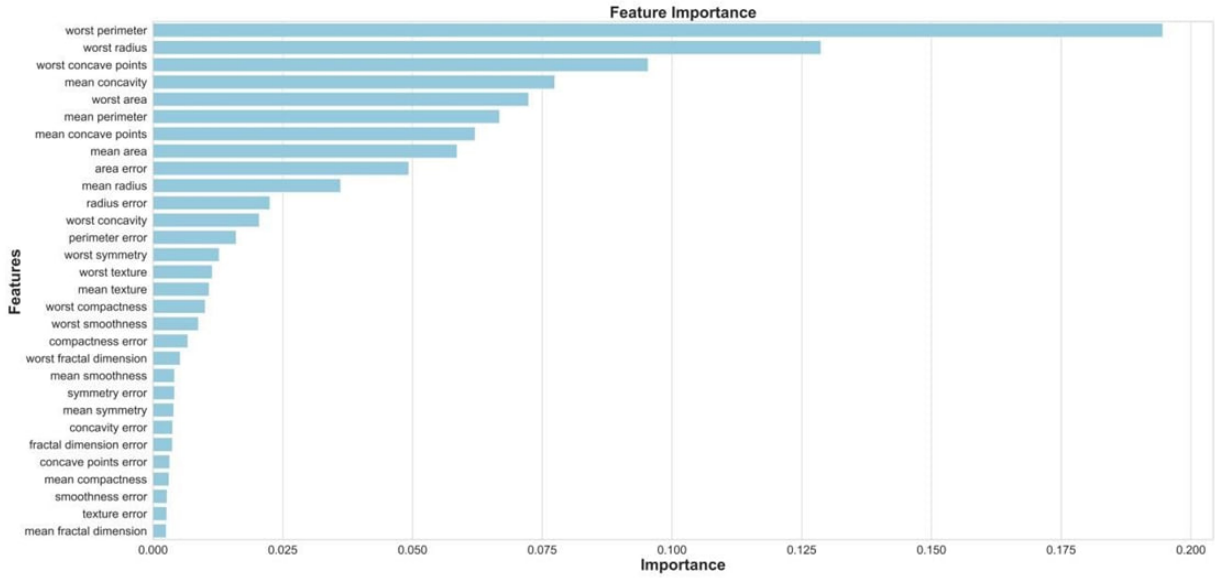

Рисунок 3.8 – Візуалізація «важливості» ознак

Ефективність RandomForestClassifier оцінюється з точки зору точності, достовірності, повернення і $F1-$рахунків, а також інших методів, які зазвичай використовуються для оцінки узагальненості моделі, що робить його потужним інструментом в інструментарії машинного навчання, особливо придатним для задач класифікації зі складними структурами даних.

3.5 Балансування даних за допомогою методу SMOTE

У контексті сучасних досліджень і застосувань машинного навчання значна увага приділяється проблемі дисбалансу класів у наборах даних. Цей дисбаланс виникає через невідповідність кількості спостережень для різних класів, що призводить до того, що моделі машинного навчання демонструють упередженість до домінуючого класу і нехтують менш представленими класами. Перспективним рішенням цієї проблеми є реалізація методу надмірної вибірки синтетичної меншості (SMOTE).

Алгоритм SMOTE призначений для усунення цього дисбалансу шляхом генерації нових вибірок для недостатньо представленого класу, інтерполюючи між існуючими спостереженнями цього класу. Ключова перевага SMOTE

полягає в тому, що він пом'якшує проблему перенавчання моделі на обмеженій кількості даних з недостатньо представленого класу, що часто зустрічається в традиційних методах доповнення даних.

Методологія передбачає відбір зразків з меншого класу, визначення $k-$ найближчих сусідів та створення синтетичних вибірок. Спостереження з недопредставленого класу відбираються, і для кожного відібраного спостереження обчислюються $k-$ найближчих сусідів у просторі ознак. Згодом створюються нові синтетичні вибірки шляхом обчислення лінійної комбінації обраного спостереження та одного з його сусідів за наступною формулою:

$$\text{Новий}_{\text{зразок}} = \text{Вибраний}_{\text{зразок}} + Y * \text{Сусід} - \text{Вибраний}_{\text{зразок}}, \qquad (3.4)$$

де Y – випадкове число в діапазоні [0,1].

SMOTE підвищує ефективність моделей машинного навчання на незбалансованих наборах даних завдяки ефективнішому використанню меншого класу та вирішенню проблеми перенавчання (рисунок 3.9).

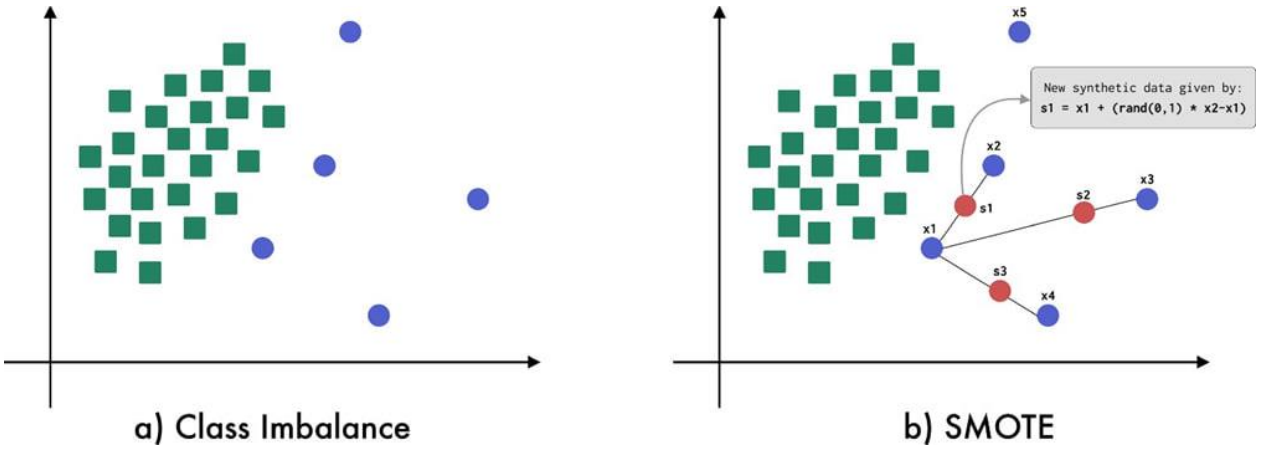

Рисунок 3.9 – Приклад роботи SMOTE: a) класичний дізбаланс;b) вирішення дізбалансу за допомгою SMOTE

Таке підвищення продуктивності пояснюється синтетичною природою згенерованих зразків, які демонструють більшу різноманітність, ніж просте

дублювання. Однак необхідно визнати обмеження, притаманні цьому методу. Генерація синтетичних вибірок, які не відповідають реальному розподілу даних, може призвести до зниження точності класифікації через чутливість параметра k, що, в свою чергу, впливає на якість вибірок.

Метод застосовується в контекстах, де класова нерівність є поширеною проблемою, оскільки він слугує потужним інструментом для подолання класової нерівності в процесі машинного навчання. При правильному використанні він може суттєво підвищити якість моделі та її узагальнюваність. Однак необхідно враховувати потенційні обмеження методу і використовувати його в поєднанні з альтернативними підходами.

3.6 Підказки та швидке проектування

Великі мовні моделі зробили революцію у взаємодії зі штучним інтелектом, усунувши потребу в кодуванні завдяки впровадженню інтерфейсу на основі природної мови. Ця трансформація відкрила широкий спектр можливостей для різних галузей, уможлививши спілкування з розширеними мовними моделями в більш інтуїтивно зрозумілий і природний спосіб, однак ключовим аспектом використання LLM є вміле використання підказок - інструкцій, які керують роботою цих моделей. У бізнес-контексті правильне узгодження підказок стає вирішальним для отримання цінних результатів.

3.7 Оцінка якості тексту

Метрики оцінювання використовуються для кількісної оцінки продуктивності конкретних моделей і алгоритмів NLP, що дозволяє приймати обґрунтовані рішення про їхню придатність для різноманітних застосувань. У практиці НЛП оціночні метрики визначають точність, надійність і загальну якість результатів, отриманих за допомогою моделей NLP. Вони забезпечують стандартизований спосіб порівняння продуктивності різних систем та

алгоритмів НЛП [19]. У сфері НЛП існує багато різних метрик оцінювання. Нижче наведено список найпоширеніших оціночних метрик у сфері НЛП, а також їхні визначення та типові застосування [20]:

Точність: Відноситься до частки випадків, коли модель робить правильний прогноз порівняно із загальною кількістю зроблених прогнозів. Найкраще використовувати, коли вихідна змінна є категоричною або дискретною. Наприклад, як часто алгоритм класифікації настроїв є правильним.

Точність (Precision): Оцінює відсоток істинних позитивних результатів від усіх позитивних випадків. Особливо корисний, коли ідентифікація позитивних результатів важливіша за загальну точність.

Наприклад, якщо ми визначаємо рак, який зустрічається в 1% випадків, модель, яка завжди дає результат «негативний», буде мати точність 99%.

«негативний», буде на 99% достовірною, але на 0% точною.

Відсоток істинно-позитивних результатів у порівнянні з комбінованими істинними і хибно-позитивними результатами. У прикладі з рідкісним видом раку, який зустрічається в 1% випадків, якщо модель робить абсолютно випадковий прогноз (50/50), він матиме 50% точності (50/100), 50% достовірності (0,5/1) і 1% вірогідності (0,5/50).

Оцінка $F1$ поєднує точність і пригадування для отримання єдиного показника - як повноти, так і точності. Використовується разом з точністю і є корисним для завдань маркування послідовності, таких як вилучення об'єктів і пошук відповідей на запитання.

$$F_1 = \left(\frac{recall^{-1} + precision^{-1^{-1}}}{2}\right) = 2 * \frac{precision * recall}{precision + recall} \tag{3.5}$$

4 ОПИС МОДЕЛІ ТА АНАЛІЗ РЕЗУЛЬТАТІВ

4.1 Опис моделі

У дослідженні використано комбінацію трансформерної моделі XLM-RoBERTa (рисунок 4.1) для формування множини мультимовних ознак текстів і класифікатора Random Forest для класифікації жанрів. Основні етапи реалізації моделі: Ініціалізація, формування множини, обробка даних, оцінка.

Ініціалізація моделі та токенізатора: Спочатку завантажується попередньо навчена багатомовна модель XLM-RoBERTa та її токенізатор з бібліотеки transformers.

```
model_name = "xlm-roberta-base"
tokenizer = XLMRobertaTokenizer.from_pretrained(model_name)
model = XLMRobertaModel.from_pretrained(model_name)
```

Рисунок 4.1 – Ініціалізація моделі XLM RoBERTa

Формування множини лінгвістичних ознак здійснюється таким чином, що з кожного тексту виділяються різноманітні лінгвістичні характеристики, зокрема кількість слів, кількість речень, середня довжина речень, кількість знаків пунктуації та частка великих літер, як показано на рисунку 4.2.

```
# Extracting linguistic features
features = {
    "word_count": len(tokens), "sentence_count": len(sentences),
    "avg_sentence_length": sum(len(s.split()) for s in sentences) / len(sentences) if sentences else 0,
    "uppercase_ratio": sum(1 for c in text if c.isupper()) / len(text) if text else 0,
    "punctuation_count": sum(1 for c in text if c in "!?.,"),
    "emotion_words_count": emotion_words_count,
}
```

Рисунок 4.2 – Формування множини лінгвістичних ознак об’єктів

Формування множини мультимовних векторних представлень (рисунок 4.3), де для кожного тексту використовуються токенізатор і модель XLM-RoBERTa для отримання контекстуальних векторів, які представляють текст у високорозмірному просторі ознак.

```
inputs = tokenizer(text, return_tensors="pt", truncation=True, padding=True, max_length=128)
outputs = model(**inputs)
return outputs.last_hidden_state.mean(dim=1).detach().numpy().flatten()
```

Рисунок 4.3 – Формування множини мультимовних векторних представлень

Обробка даних — це зібрані ознаки (рисунок 4.4), включаючи лінгвістичні та мультимовні, стандартизуються, а для категоріальних ознак застосовується one-hot кодування, це необхідно для уникнення розбіжності у форматах даних та для коректної роботи методу «Random Forest».

```
# Обробка числових і категоріальних ознак
numerical_columns = features_df.select_dtypes(include=['float64', 'int64']).columns
categorical_columns = features_df.select_dtypes(include=['object']).columns

if scaler is None:
    scaler = StandardScaler()
    features_df[numerical_columns] = scaler.fit_transform(features_df[numerical_columns])
else:
    features_df[numerical_columns] = scaler.transform(features_df[numerical_columns])

encoder = OneHotEncoder(sparse_output=False, drop='first')
if categorical_columns.any():
    encoded_cat_features = encoder.fit_transform(features_df[categorical_columns])
    encoded_cat_df = pd.DataFrame(encoded_cat_features, columns=encoder.get_feature_names_out(categorical_columns))
    features_df = features_df.drop(columns=categorical_columns)
    features_df = pd.concat( objs [features_df, encoded_cat_df], axis=1)
```

Рисунок 4.4 – Приклад обробки даних

SMOTE (рисунок 4.5) для балансування класів: З метою усунення дисбалансу між класами в датасеті застосовувався метод SMOTE для синтетичного збільшення кількості зразків у малих класах.

```
# Apply SMOTE, but with minimal k_neighbors for small classes
smote = SMOTE(sampling_strategy='auto', k_neighbors=3, random_state=42)
```

Рисунок 4.5 – Налаштування балансувальника

Для класифікації жанрів використовується алгоритм Random Forest, при побудові якого були задані наступні параметри, як на рисунку 4.6.

```
# Application of SMOTE with a check for the number of examples in the class
X_resampled, y_resampled = smote.fit_resample(X_combined, y_combined)

# Separate the data into training and test data
X_train, X_test, y_train, y_test = train_test_split( *arrays: X_resampled, y_resampled, test_size=test_size, random_state=42)

# Use Random Forest or XGBoost with cross validation
model = RandomForestClassifier(random_state=42)
model.fit(X_train, y_train)
```

Рисунок 4.6 – Параметри алгоритму Random Forest

Оцінка моделі: Модель оцінюється на тестових даних за допомогою крос-валідації як на рисунку 4.7 , а також через матрицю плутанини як на рисунку 4.8.

```
# Cross-validation
cv = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)
scores = cross_val_score(model, X_resampled, y_resampled, cv=cv, scoring='accuracy')
print(f"Cross-validation scores: {scores}")
print(f"Mean accuracy: {scores.mean()}")
```

Рисунок 4.7 – Приклад кросс валідації

```python
# Visualization of the confusion matrix
def plot_confusion_matrix(y_true, y_pred, classes):  1 usage  new *
    disp = ConfusionMatrixDisplay.from_predictions(
        y_true, y_pred, display_labels=classes, cmap=plt.cm.Blues, normalize='true'
    )
    disp.ax_.set_title("Матриця плутанини")
    plt.show()
```

Рисунок 4.8 – Приклад матриці плутанини

4.2. Методика проведення дослідження

Основна мета дослідження – розробка моделі, здатної класифікувати тексти за жанром. Через обмежену кількість даних у датасеті [21] (42000 текстів, 18 унікальних жанрів) виникла необхідність використання методів балансування, таких як SMOTE, а також обрати моделі, які ефективно працюють на малих обсягах даних. Загальна оцінка точності отриманих результатів здійснюється за допомогою крос-валідації.

$$yn - \hat{y}n = (1 - Hi_i) * (yn - \hat{y} - n) \quad (4.1)$$

Використані дані представлені у вигляді об’єктів тексту, один такий об’єкт може бути розміром від ста до ста тисяч символів. Оцінка кожного обє’кту відбувається однаково за допомогою методу формування множини лінгвістичних ознак та попередньою токенізацією тексту.

4.2.1 XLM-RoBERTa

У роботі використано багатомовну модель XLM-RoBERTa – це модель трансформерного типу, яка була розроблена компанією Facebook AI. Вона має 125 мільйонів параметрів, що дозволяє їй ефективно розпізнавати особливості

текстів різними мовами. Модель була попередньо навчена на великій кількості текстів (1.5 ТБ даних) із використанням методу Masked Language Modeling (MLM). У дослідженні XLM-RoBERTa використовувалася для формування множини контекстуальних ознак текстів, які пізніше подавалися на вхід класифікатору Random Forest.

4.2.2 Використання SMOTE

Метод SMOTE (Synthetic Minority Oversampling Technique) застосовувався для синтетичного збільшення кількості зразків у малих класах. Це було необхідно через дисбаланс жанрів у датасеті, що могло призвести до переважання частіших класів і зниження точності моделі.

4.2.3 Використання Random Forest

Random Forest був обраний як основний класифікатор завдяки його стійкості до перенавчання та здатності працювати з даними високої розмірності. Він добре підходить для задач, де доступні ручні ознаки та обмежений обсяг даних. Основними параметрами моделі були: кількість дерев – 100, критерій розбиття – Gini. Через обмежені обчислювальні ресурси процес навчання був значно уповільнений, що не дозволило розширити загальну кількість експериментів і використовувати більш складні алгоритми. Крім того, спроби навчання інших моделей (наприклад, SVM та логістичної регресії) не дали суттєвого покращення результатів.

4.3. Аналіз результатів моделювання

4.3.1 Аналіз якісних показників моделі

Аналіз роботи моделі здійснювався на 1000 тестових елементах датасету. Результати класифікації (рисунок 4.9) показали, що використання XLM-RoBERTa разом із Random Forest і SMOTE дало змогу досягти високої середньої точності (майже 90%) на крос-валідації. Розкид між оцінками невеликий, а це означає, що модель стабільна при різних розбиттях даних.

```
Cross-validation scores: [0.8699187  0.90406504 0.90731707 0.88780488 0.92357724]
Mean accuracy: 0.8985365853658536
```

Рисунок 4.9 – Результати класифікації

Оцінки перехресної валідації: [0.8699, 0.9041, 0.9073, 0.8878, 0.9236].

Середня точність: 0.8985.

Під час тестування, було виявлено, що збільшення кількості епох навчання погіршувало точність роботи моделі, через те, що використана XLM-RoBERTa для навчання має достатню кількість параметрів щоб показувати вискоий результат у точності з однієї епохи навчання, в той час збільшення епох навчання навпаки погіршувало результат, що можно помабичити у таблиці 4.1, а для отримання більшого показника точності необхідно збільшити об’єм даних який використовується для навчання.

Таблиця 4.1 – Відношення точності моделі до епох навчання

Епохи	Кількість об’єктів	Кількість витраченого часу, хв	Точність роботи, %
1	1000	63	89
2	1000	64	78
3	1000	62	73
4	1000	60	69
5	1000	58	65

Відповідно до результатів на одній тисячі об’єктів, навчання на більшій кількості буде збільшувати витрачаємий час та необхідні вираховувальні потужності, що можно побачити у таблиці 4.2

Таблиця 4.2 – Відношення кількості об’єктів до часу навчання

Кількість об’єктів навчання	Необхідний час навчання, хв
200	8.5
1000	43
10000	432
42000	2148

Розрахувати необхідний час для навчання можно за формулою

$$T = \frac{N*E*P}{F}, \quad (4.2)$$

де T – час навчання,

N – кількість даних (об’єктів),

E – кількість епох,

P – кількість параметрів моделі,

F – обчислювальна потужність (FLOPS).

Навчання моделі проводилось за допомогою відеокарти Nvidia Gforce GTX 1080, та центрального процесора Intel Core i5-10600KF.

4.3.2 Розподіл класів у даних

Клас OTR (Other) домінує в обох наборах. Це вказує на значний дисбаланс класів. Інші класи, такі як MON(Монолог), DIA(Щоденник), і DRA(Драма), подані мінімально, від чого може знижуватись точність прогнозів для них.

Тестова вибірка (1000 рядків):

{'OTR': 601, 'LET': 342, 'DRA': 41, 'DIA': 11, 'MON': 5}

4.3.3 Матриця плутанини на тестовій вибірці

На рисунку 4.10 представлено матрицю плутанини (confusion matrix), яка демонструє розподіл прогнозів моделі за жанрами.

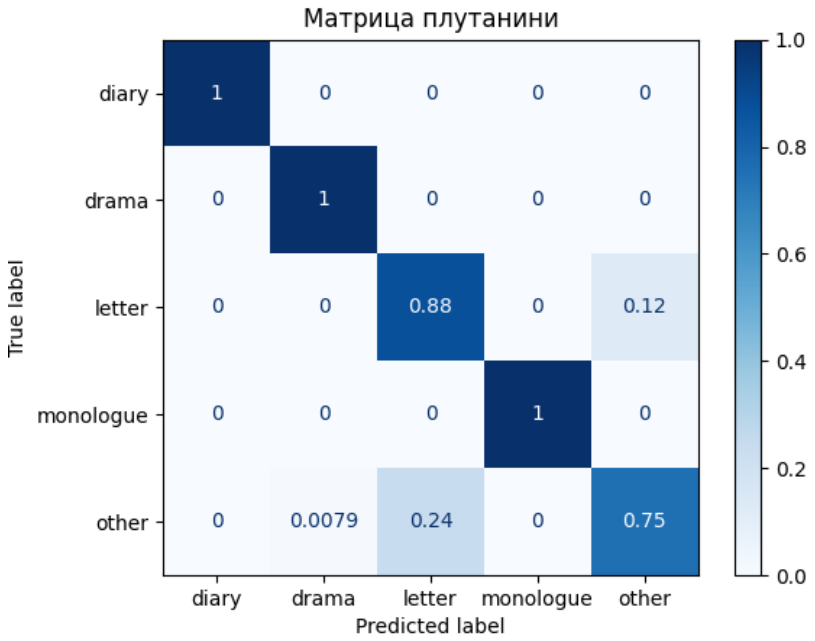

Рисунок 4.10 – Матриця плутанини

Як можна побачити на рисунку 4.10 модель демонструє високу точність для класів *diary, drama, monologue* і *letter*. Для класу *other* спостерігається плутанина з *letter* (24% випадків), що пов'язано з дисбалансом у даних.

4.4 Приклади розбору текстів

У магістерській роботі було проведено аналіз тестових текстів, на основі яких здійснено класифікацію за допомогою ручних ознак та жанрів. У цьому розділі розглянуто приклади кількох таких текстів, на яких здійснено розбір:

- Приклад 1 (монолог):
  1) Кількість слів: 1018
  2) Кількість речень: 78
  3) Середня довжина речень: 9.9 слів
  4) Коєфіцієнт заголовних букв: 3.38%
  5) Кількість знаків пунктуації: 226
  6) Кількість емоційних слів: 0
- Приклад 2 (лист):
  1) Кількість слів: 745
  2) Кількість речень: 51
  3) Середня довжина речень: 11.53 слів
  4) Коєфіцієнт заголовних букв: 2.73%
  5) Кількість знаків пунктуації: 140
  6) Кількість емоційних слів: 0
- Приклад 3 (лист):
  1) Кількість слів: 826
  2) Кількість речень: 31
  3) Середня довжина речень: 19.84 слів
  4) Коєфіцієнт заголовних букв: 2.64%
  5) Кількість знаків пунктуації: 166
  6) Кількість емоційних слів: 1

З аналізу видно, що модель опирається на структурні характеристики тексту, такі як кількість слів, кількість речень, середня довжина речень та кількість знаків пунктуації, проте в кожному з наведених прикладів кількість емоційних слів мала мінімальний вплив на класифікацію, що свідчить про те, що для конкретних жанрів емоційна складова не є основним фактором для визначення жанру тексту.

ВИСНОВКИ

У процесі виконання кваліфікаційної роботи, розглянута концепція оцінювання текстів за допомогою класифікатора за набором ознак: за різноманітними лінгвістичними характеристиками, зокрема кількість слів, кількість речень, середня довжина речень, кількість знаків пунктуації та часткою великих літер. Запропонований метод вирішення задачі виявився найбільш ефективний у роботі з умовами неповноти вхідних даних.

Створена модель для стилістичного розбору стилю тексту використовує трансформери, з використанням нейромережі типу XLM-RoBERTa для навчання, адже це суттєво скоротить часові витрати у вирішенні завдань класифікації тексту за стилем. Проводиться аналіз тексту на предмет його жанру та лінгвістичних особливостей. Модель має середню точність оцінки 89% відсотків, що є досить великим показником, при відносно невеликому обсягу даних, на якому вона була навчена. При обсязі в 1000 рядків (даних), 3/4 жанри текстів були визначені вірно.

Під час тесутвання роботи моделі було виявленно та проаналізовано низку недоліків. Серед них низька якість вхідних даних та їх очевидний недолік, необхідність в додаткових як матеріальних так і інтелектуальних ресурсів для створення більш точної, повної та якісної оцінки, а також нестача вираховувальних потужностей. Дані на яких проводилось навчання моделі мали дуже великий діапазон, це призводило до нездатності правильно визначити жанр тексту, через це довелось використовувати синтетичний балансувальник, що знижувало точність через шуми у вибірці. Важливу роль відіграє кількість даних для навчання. Доступними по темі роботи є датасет який складається 42000 текстів та 18 унікальних жанрів, проте таких даних як, стиль та інші було недостатньо для повноцінної роботи всього функціоналу моделі. Наприклад, на десять тисяч текстів з жанрами, приходиться всього одна тисяча стилей, що у більшості випадків видавало результат «інше».

Наразі подібна модель має декілька вітчизняних аналогів та низку світових, але лише за деякими лінгвістичними параметрами. Для роботи з простими текстами існують більш прості алгоритми, проте наразі у такій комплектації функціоналу, а саме роботи з авторськими текстами на українській мові модель є унікальною. Слід зауважити, що у відношені вже до готових рішень, вона потребує низки покращень та налашутвань.

Використаний для навчання датасет був створений об’єднанням українських студентів, викладачів, та інших учасників. Зокрема студенти Українського католицького університету, Національного університету “Львівська політехніка”, Київського національного університету ім. Тараса Шевченка, Національного університету "Києво-Могилянська академія" та інші.

Підхід такого роду може бути корисним для різних застосувань, включно з літературним аналізом, захистом авторських прав і автоматичним розпізнаванням плагіату. Надалі можливі поліпшення моделі завдяки використанню більших і різноманітніших даних, а також експериментування з різними архітектурами нейронних мереж.

Розроблена модель може бути використана в різних структурах та організаціях, таких як, видавництва, школи, університети, книгарні. Здібність швидко та точно працювати з текстом є користним інструментом в закладах, де безпосередньо відбуваєтсья робота з текстом. У школах та університетах це або додатковий способ для навчання, а скорочення часу викладачам на рутинних операціях. Видавництва також можуть використовувати дану модель для швидкого визначення та аналізу текста.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Bowman, S. R. Eight Things to Know about Large Language Models. Матеріали конференцій ICLR, (10 березня 2023 р.) С. 1-3.

2. Yongchao Zhou. A. I. Large Language Models Are Human-Level Prompt Engineers, Матеріали конференцій ICLR, (10 березня 2023 р.) С. 3-5.

3. Wayne Xin Zhao, K. Z. A Survey of Large Language Models. – URL: https://arxiv.org/abs/2303.18223 (дата звернення: 24.12.2024).

4. Laria Reynolds, K. M. Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. – URL: https://arxiv.org/abs/2102.07350 (дата звернення: 15.02.2021).

5. Junyi Li, T. T. Learning to Transfer Prompts for Text Generation. Сиетл, 2022 р. С. 3506-3518.

6. Albert Webson, E. P. Do Prompt-Based Models Really Understand the Meaning of their Prompts? Сиетл, 2022 р. С. 2340-2344.

7. Corentin Kervadec, F. F. Unnatural language processing: How do language models handle machine-generated prompts? Сінгапур, 2023. С. 14377–14379.

8. Jules White, Q. F. A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. – URL: https://arxiv.org/abs/2302.11382 (дата звернення: 21.02.2023).

9. Qinyuan Ye, M. A. Prompt Engineering a Prompt Engineer. – URL: https://arxiv.org/abs/2311.05661 (дата звернення: 09.11.2023).

10. Zishan Guo, R. J. Evaluating Large Language Models: A Comprehensive Survey. Матеріали конференцій ICLR, (30 жовтня 2023 р.) С. 8-10

11. Lochan Basyal, M. S. Text Summarization Using Large Language Models: A Comparative Study of MPT-7b-instruct, Falcon-7b-instruct, and OpenAI Chat-GPT Models. – URL: https://arxiv.org/abs/2310.10449 (дата звернення: 16.11.2023).

12. Міжнародні культурні традиції: мова та етика ділової комунікації. – URL:http://megalib.com.ua/content/8303_35_Gymor_yak_skladova_kyltyri_ spilkyvannya.html

13. Наталя Висоцька. Проблема класифікації текстів у мовознавстві. Донецьк, 2014 р. С. 1-5.

14. Анжеліка Попович. Стилістичний аналіз тексту як метод навчання майбутніх учителів української мови і літератури. Одеса, 2017 р. С. 1-6.

15. Sheikh H., P. C. Artificial Intelligence: Definition and Background. Париж, 2023. С. 2662-2668

16. Sarker, I. Machine Learning: Algorithms, Real-World Applications and Research Directions. Париж, 2021. С. 1 -5

17. What is natural language processing (NLP)? – URL: https://www.ibm.com/topics/natural-language-processing

18. Kapronczay, M. A Beginner's Guide to Language Models. – URL: https://builtin.com/data-science/beginners-guide-language- models (дата звернення: 13.12.2022).

19. Evaluation Metrics. – URL: https://www.alooba.com/skills/concepts/ natural-language-processing/evaluation- metrics/

20. Smirnov, M. (2022, February 18). Common metrics for evaluating natural language processing (NLP) models. – URL: https://medium.com/@mikeusru/common-metrics-for-evaluating-natural-language-processing-nlp-models-e84190063b5f

21. Maria Shvedova, Arsenii Lukashevskyi (2024): PluG: Corpus of Old Ukrainian Texts. Electronic resource: Kharkiv, Jena. Available at https://github.com/Dandelliony/pluperfect_grac