Хмельницький національний університет
Факультет інформаційних технологій
Кафедра комп'ютерних наук

КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА

на тему *Метод контент-аналізу коментарів засобами інтелектуального аналізу даних для IT блогу розробників платформи Unity*

Галузь знань *12 – Інформаційні технології*
Шифр і назва галузі знань

Спеціальність *122 – Комп'ютерні науки*
Шифр і назва спеціальності

Освітня програма *Комп'ютерні науки*
Назва освітньої програми

Виконав: *студент групи КН-20-1* — Група виконавця | Підпис | *Джорджо МІЗИН* — Ім'я, ПРІЗВИЩЕ

Керівник: *PhD, ст. викл. каф. КН* — Науковий ступінь, посада | Підпис | *Павло РАДЮК* — Ім'я, ПРІЗВИЩЕ

Нормоконтроль *к.т.н., доц. каф. КН* — Науковий ступінь, посада | Підпис | *Руслан БАГРІЙ* — Ім'я, ПРІЗВИЩЕ

До захисту допускаю:
Зав. кафедри КН, д.т.н., професор — Підпис | *Олександр БАРМАК* — Ім'я, ПРІЗВИЩЕ
24 червня 2024 р.

Хмельницький 2024

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
Факультет *інформаційних технологій*
Кафедра *комп'ютерних наук*
Освітній ступінь *бакалавр*
Галузь знань *12 – Інформаційні технології*
Спеціальність *122 – Комп'ютерні науки*

ЗАТВЕРДЖУЮ
Завідувач кафедри *комп'ютерних наук*

(підпис)
д.т.н., професор Олександр БАРМАК
«*16*» *02* *2024* року

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ БАКАЛАВРА

1. Тема кваліфікаційної роботи бакалавра: *«Метод контент-аналізу коментарів засобами інтелектуального аналізу даних для IT блогу розробників платформи Unity»*

2. Завдання видано студенту *Джорджо Мізину*
(Ім'я, прізвище)

3. Керівник роботи *старший викладач кафедри КН Павло Радюк*
(посада, ім'я, прізвище)

4. Затверджено наказом університету від «*15*» *02* *2024* р. № *8*

5. Дата видачі завдання студенту: «*16*» *02* *2024* р.

6. Зміст пояснювальної записки (перелік задач) та вихідні дані:

Мета цієї роботи – підвищення якості контент-аналізу коментарів IT блогу розробників платформи Unity засобами інтелектуального аналізу даних. Для цього потрібно: виконати дослідження предметної області контент-аналізу; створити метод контент-аналізу коментарів IT блогу розробників платформи Unity засобами інтелектуального аналізу даних; виконати розробку архітектури нейромережі для визначення сентименту коментаря; створити проектну архітектуру інформаційної системи IT блогу розробників платформи Unity; виконати проєктування БД; виконати вибір та підготовку робочих вхідних даних методу контент-аналізу коментарів IT блогу; виконати програмну реалізацію методу та провести його дослідження ефективності.

7. Календарний план виконання кваліфікаційної роботи бакалавра:

№	Назва етапів (розділів) кваліфікаційної роботи бакалавра	Термін виконання	Примітка
1	*Вибір напряму дослідження та узгодження тематики кваліфікаційної роботи бакалавра з керівником, складання календарного графіка виконання роботи*	*січень 2024*	Виконано
2	*Ознайомлення з предметною областю, формулювання мети та задач дослідження, визначення об'єкта та предмета дослідження*	*лютий 2024*	Виконано
3	*Проектування та розробка загальної архітектури програмного забезпечення, інтерфейсу користувача, вибір засобів реалізації програмного забезпечення*	*березень 2024*	Виконано
4	*Створення та тестування програмного забезпечення*	*квітень 2024*	Виконано
5	*Написання пояснювальної записки, урахування зауважень керівника, оформлення згідно вимог*	*травень 2024*	Виконано
6	*Розробка презентаційних матеріалів та попередній захист кваліфікаційної роботи*	*травень 2024*	Виконано
7	*Отримання відгуку керівника, рецензії, перевірка на плагіат, нормоконтроль*	*червень 2024*	Виконано
8	*Підготовка до захисту та захист кваліфікаційної роботи бакалавра*	*червень 2024*	Виконано

Виконавець: *студент групи КН-20-1* _____ *Джорджо МІЗИН*
Група виконавця | Підпис | Ім'я, ПРІЗВИЩЕ

Керівник: *ст. викл. каф. КН* _____ *Павло РАДЮК*
Науковий ступінь, посада | Підпис | Ім'я, ПРІЗВИЩЕ

Анотація

Тема кваліфікаційної роботи бакалавра: *«Метод контент-аналізу коментарів засобами інтелектуального аналізу даних для IT блогу розробників платформи Unity»*

Виконавець кваліфікаційної роботи бакалавра: *студент групи КН-20-1 Джорджо Мізин*

Керівник кваліфікаційної роботи бакалавра: *старший викладач кафедри КН Павло Радюк*

Кваліфікаційна робота бакалавра містить:

Пояснювальна записка				Кількість додатків
Сторінок	Рисунків	Таблиць	Джерел інформації	
68	30	11	35	5

Метою кваліфікаційної роботи бакалавра є підвищення якості контент-аналізу коментарів IT блогу розробників платформи Unity засобами інтелектуального аналізу даних. Результатом досягнутої мети є відповідний метод контент-аналізу коментарів засобами інтелектуального аналізу даних та його програмна реалізація у вигляді IT блогу розробників платформи Unity.

Розроблена система призначена для автоматизації процесу аналізу коментарів IT блогу розробників платформи Unity, забезпечуючи високу якість контент-аналізу завдяки застосуванню інтелектуальних методів обробки даних.

Напрямками практичного використання розробленої інформаційної системи визначено автоматизовану обробку коментарів та їх класифікацію, що дозволяє значно знизити навантаження на модераторів блогу та підвищити ефективність моніторингу дискусій.

Ключові слова: контент аналіз, інтелектуальний аналіз даних, IT блог, аналіз сентименту, GRU.

Виконавець: *студент групи КН-20-1* _____ *Джорджо МІЗИН*

Група виконавця | Підпис | Ім'я, ПРІЗВИЩЕ

Зміст

Перелік скорочень

Скорочення, термін, позначення	Пояснення
ІТ	Інформаційні технології
КН	Комп’ютерні науки
НМ	Нейронна мережа
GRU	Gated Recurrent Unit
ПЗ	Програмне забезпечення
КА	Контент-аналіз
SVM	Support Vector Machines
NB	Naive Bayes
LSTM	Long Short-Term Memory
WB	Word embeddings
КП	Косинусна подібність
NLU	Natural Language Understanding
API	Application Programming Interface
CNN	Convolutional Neural Network
HTML	HyperText Markup Language
БД	База даних
СКБД	Система керування базами даних
ER	Entity-relationship model
ID	Identifier Data
HTTP	HyperText Transfer Protocol
JSON	JavaScript Object Notation
URL	Uniform Resource Locator
SQL	Structured Query Language

Вступ

Кваліфікаційна робота бакалавра присвячена підвищенню якості контент-аналізу коментарів IT блогу розробників платформи Unity засобами інтелектуального аналізу даних. Для досягнення мети було створено метод контент-аналізу коментарів засобами інтелектуального аналізу даних, а також було створено відповідну інформаційну системи у вигляді IT блогу, що дозволяє за проведеним аналізом коментарів визначати їх релевантність.

Актуальність. У сучасному світі блогові платформи є важливим джерелом інформації для розробників, які обмінюються досвідом, обговорюють нові технології та діляться своїми знаннями. У зв’язку з цим виникає необхідність автоматизованого аналізу великої кількості коментарів, що дозволяє визначити їх релевантність та корисність. Ручний аналіз коментарів займає багато часу та ресурсів, що робить його неефективним для великих платформ.

Використання інтелектуального аналізу даних для обробки коментарів в IT блогах надає можливість підвищити якість обговорень та сприяти створенню цінного контенту. Це особливо важливо для платформи Unity, яка є однією з провідних у сфері розробки ігор та інших інтерактивних додатків. Завдяки автоматизації процесу аналізу коментарів можна швидко виявляти нерелевантні або шкідливі відгуки, тим самим підтримуючи високий рівень дискусій та зберігаючи професійність блогу.

Інтелектуальний аналіз коментарів також сприяє покращенню взаємодії між користувачами платформи. Завдяки аналізу тональності коментарів можна краще розуміти настрої та потреби користувачів, що дозволяє адміністраторам блогу оперативніше реагувати на запити та покращувати загальний досвід користувачів. Це, в свою чергу, підвищує лояльність користувачів до платформи та сприяє її подальшому розвитку.

В умовах постійного зростання кількості користувачів та обсягу інформації в інтернеті, створення автоматизованої системи аналізу коментарів

стає критично важливим завданням. Впровадження таких систем дозволяє ефективніше управляти інформаційним простором, підтримувати високу якість контенту та забезпечувати корисність і релевантність інформації для всіх користувачів платформи.

Об’єкт дослідження – процес інтелектуального аналізу коментарів IT блогу розробників платформи Unity засобами інтелектуального аналізу даних.

Предмет дослідження – методи контент-аналізу для коментарів IT блогу.

Мета кваліфікаційної роботи бакалавра – підвищення якості контент-аналізу коментарів IT блогу розробників платформи Unity засобами інтелектуального аналізу даних.

Завдання кваліфікаційної роботи бакалавра – Провести аналіз інформаційних моделей в області контент-аналізу. Розглянути засоби інтелектуального аналізу даних області контент-аналізу коротких текстових даних, та обрати підхід для реалізації. Виконати аналіз існуючих програмних засобів та наукових рішень. Створити метод контент-аналізу коментарів IT блогу розробників платформи Unity засобами інтелектуального аналізу даних. Виконати розробку архітектури нейромережі для визначення сентименту коментаря. Створити проектну архітектуру інформаційної системи IT блогу розробників платформи Unity. Виконати проєктування бази даних. Виконати вибір та підготовку робочих вхідних даних методу контент-аналізу коментарів IT блогу. Розглянути особливості використання спеціалізованих програмних компонентів для спрощення програмної розробки. Виконати вибір засобів програмної реалізації методу контент-аналізу коментарів IT блогу розробників платформи Unity засобами інтелектуального аналізу даних. Виконати програмну реалізацію створеного методу. Виконати тестування створеного IT блогу розробників платформи Unity та застосунку для тренування нейромереж. Виконати дослідження ефективності створеного методу з використанням розробленого ПЗ.

Розділ 1 Характеристика предметної області: аналіз моделей, методів та реалізацій для контент-аналізу коментарів блогів

1.1 Аналіз інформаційних моделей в області контент-аналізу

Контент-аналіз стає все більш актуальним у сучасному інформаційному середовищі з кількісним зростанням доступу до даних та швидким розвитком технологій обробки тексту. У сучасних умовах він знаходить широке застосування в різних сферах, наприклад, в маркетингу контент-аналіз дозволяє підприємствам розуміти реакцію споживачів на їхні рекламні кампанії та продуктові пропозиції. Аналізуючи коментарі та відгуки в соціальних мережах та інтернет-форумах, маркетологи можуть визначити ключові теми, що цікавлять аудиторію, а також ідентифікувати проблеми або недоліки продуктів, що потребують уваги [1].

У соціальних науках контент-аналіз є незамінним інструментом для дослідження публічної думки, політичних дискусій та інших аспектів соціокультурного життя. Він дозволяє аналізувати масштабні обсяги текстових даних, що включають статті, інтерв'ю, твіти, коментарі в новинах тощо, для виявлення тенденцій, настроїв та важливих подій в суспільстві, наприклад, в дослідженнях медійної лояльності контент-аналіз допомагає оцінити тональність і контекст медійних повідомлень та їх вплив на громадську думку [2].

Існує кілька методів контент-аналізу, які допомагають дослідникам отримувати корисні дані з різних типів тексту.

Одним з основних методів контент-аналізу є кількісний підхід. Він зосереджується на підрахунку частоти певних слів, фраз або тем у тексті. Кількісний підхід дозволяє дослідникам отримувати статистичні дані, які можуть бути використані для виявлення загальних тенденцій та патернів у великих обсягах тексту. Наприклад, кількісний контент-аналіз може бути застосований для вивчення зміни частоти вживання певних термінів у медіа-просторі протягом певного періоду часу, що допомагає зрозуміти зміни у суспільній думці або медійній політиці [3].

Іншим важливим методом є якісний контент-аналіз. Даний підхід зосереджується на детальному вивченні змісту та контексту тексту, щоб виявити глибинні значення, теми та мотиви. Якісний контент-аналіз включає в себе кілька технік, таких як тематичний аналіз, коли дослідники ідентифікують та аналізують головні теми тексту. Наративний аналіз, як частина якісного підходу, допомагає дослідникам вивчати структури та розповіді в тексті, що дозволяє зрозуміти, як інформація організована та представлена [4].

Лексичний аналіз є ще одним методом контент-аналізу, що зосереджується на вивченні слів та їхніх відносин у тексті. Метод включає аналіз лексичних полів, де дослідники вивчають групи слів, пов'язаних за значенням, та семантичні мережі, що представляють відносини між словами та поняттями у вигляді графів. Лексичний аналіз допомагає виявити семантичні структури тексту та зрозуміти, як різні слова та фрази взаємодіють між собою [5].

Сентимент-аналіз є ще одним важливим методом, який фокусується на визначенні емоційної тональності тексту. Аналіз сентименту дозволяє ідентифікувати, чи є текст позитивним, негативним або нейтральним за своєю емоційною складовою. Це дає змогу компаніям зрозуміти, як їхні продукти, послуги або контент сприймаються споживачами. Наприклад, у сфері маркетингу, аналіз сентименту допомагає визначити реакцію споживачів на нові рекламні кампанії або випуск продуктів. Позитивні коментарі свідчать про успіх ініціативи, тоді як негативні вказують на проблеми, які потребують вирішення [6].

Ще однією важливою задачею є визначення відповідності коментаря до теми обговорення або допису. Задача передбачає оцінку того, наскільки коментарі відповідають основній темі або контексту допису. Такий аналіз є важливим для підтримки релевантності обговорень у форумах, соціальних мережах та інших платформах, де користувачі взаємодіють через текстові повідомлення [7].

Визначення відповідності коментаря до теми може здійснюватися як вручну, так і автоматизовано. Ручний підхід, що включає читання та аналіз

коментарів дослідниками, є дуже точним, але надзвичайно трудомістким і неефективним для великих обсягів даних. Автоматизований підхід, навпаки, використовує алгоритми машинного навчання та обробки природної мови для класифікації коментарів на основі їх змісту та контексту [8].

Згідно проведеного аналізу предметної області можна виділити наступні сутності, що наведені у таблиці 1.1.

Таблиця 1.1 – Інформаційна модель предметної області

№	**Сутність**	**Опис**	**Атрибути**
1	Вебресурс	Вебсистема, для якої проводиться контент-аналіз	Назва, адреса, тип, контактні дані
2	Користувач	Користувачі вебресурсу	Логін, пароль, тип користувача дата реєстрації, дата деактивації
3	Обговорення	Обговорення створене власниками вебресурсу	Тема, текст, автор, категорія, дата/час публікації
4	Коментар	Коментарі користувачів до обговорення	Текст коментаря, автор
5	Результат контент-аналізу	Результати проведеного контент-аналізу, що містять визначену тональність коментаря та відповідність коментаря темі обговорення.	Дата/час проведення, допис, коментар допису, тональність коментаря, відповідність коментаря темі

Отже, у сучасному цифровому середовищі, де щодня генерується величезна кількість текстової інформації, виявлення сентименту та відповідності

коментаря до теми обговорення стає надзвичайно важливим завданням. Отримана інформація дозволяє компаніям, дослідникам та аналітикам розуміти настрої користувачів та забезпечувати релевантність обговорень.

1.2 Засоби інтелектуального аналізу даних області контент-аналізу коротких текстів

Контент-аналіз коментарів засобами інтелектуального аналізу є інструментом для розуміння великих обсягів текстових даних, що генеруються користувачами в цифровому середовищі. Сентимент-аналіз є складною задачею в області обробки природної мови, яка вимагає використання різних підходів для досягнення високої точності і надійності в класифікації текстів за їх емоційним відтінком.

Робота з короткими текстами має свої особливості, які варто враховувати при їхньому аналізі і обробці. Одна з основних особливостей полягає в обмеженій кількості інформації, яка може бути міститься в одному текстовому повідомленні. У зв'язку з цим важливо використовувати методи інтелектуального аналізу даних, які ефективно працюють з обмеженими кількістю тексту.

Короткі тексти часто характеризуються відсутністю структури та відсутністю контексту, що може ускладнювати їхню обробку. Наприклад, в одному короткому тексті може бути відсутнім загальний контекст, який зазвичай присутній в довших тексти. Також короткі тексти можуть містити специфічні або мовні особливості, що потребують додаткової обробки перед аналізом [9].

Для ефективного аналізу коротких текстів часто застосовуються методи машинного навчання, зокрема класифікація текстів за сентиментом, виявлення тем, аналіз ключових слів тощо.

Класичні методи машинного навчання, такі як Support Vector Machines або Naive Bayes, використовуються для статистичного аналізу текстів і класифікації їх за сентиментом. Крім того, гібридні підходи, які поєднують у

собі різні методи для покращення точності результатів, є популярними в розробці продуктів для аналізу соціальних медіа та інших веб-платформ. Кожен з цих підходів має свої особливості і може бути використаний залежно від конкретної задачі і доступних ресурсів для тренування моделі. Один із підходів полягає в застосуванні нейромережевих архітектур, таких як Gated Recurrent Unit (GRU), що спроможні ефективно моделювати довгострокові залежності між словами в тексті [10].

GRU одним з типів рекурентних нейронних мереж, аналогічним до LSTM (Long Short-Term Memory). Введений у 2014 році, цей архітектурний підхід призначений для спрощення та прискорення процесу навчання порівняно з LSTM, зберігаючи при цьому більшість його ефективності, особливо в обробці послідовностей даних.

На відміну від LSTM, в GRU відсутня окрема довготривала станція комірки. Стан комірки в GRU є комбінацією минулого стану та нових вхідних даних, які модулюються через ворота оновлення та скидання. Цей стан оновлюється на кожному кроці і передає інформацію по всій мережі [11].

Визначення відповідності коментаря до теми обговорення можна виконати також за допомогою декількох підходів. Один з підходів полягає у використанні методів машинного навчання для автоматичного класифікації текстів. Наприклад, можна використовувати класифікаційні моделі на основі SVM або навчання з учителем для визначення тематичних зв'язків у текстах.

Інший підхід полягає в застосуванні алгоритмів обробки природної мови для аналізу семантики текстів. Такі алгоритми можуть враховувати контекстуальні особливості і семантичні зв'язки між словами, що дозволяє визначати, наскільки зв'язаним є коментар з основною темою.

Для точнішого визначення відповідності можна використовувати алгоритми аналізу текстів, які враховують структурні особливості речень і взаємодію між словами у контексті. Наприклад, алгоритми, що використовують векторні представлення слів (word embeddings). А для визначення відстані між векторами використовується косинусна подібність [12].

Косинусна подібність представляє собою метод для вимірювання ступеня семантичної близькості двох векторів у просторі високих вимірів. Вона обчислюється як косинус кута між двома векторами, що визначається як добуток їх значень, поділений на добуток їхніх норм. Значення косинусної подібності знаходиться в діапазоні від -1 до 1, де 1 вказує на максимальну схожість (однаковий напрямок), а -1 вказує на максимальну несхожість (протилежні напрямки). Даний метод широко використовується у природно-мовних обробках для порівняння семантичної схожості між словами, фразами або текстовими документами [13].

Отже, для визначення сентименту коментарів до обговорення доцільно використати GRU, а для визначення відповідності коментаря до теми обговорення косинусну подібність.

1.3 Аналіз існуючих програмних засобів та наукових рішень щодо контент-аналізу коротких текстів

В сучасному світі існує велика кількість різноманітних інструментів для аналізу контенту. Такі інструменти використовуються для обробки різних типів даних, таких як тексти, звуки, відео тощо, з метою отримання важливої інформації. Вони різняться за функціональністю, можливостями і методами обробки даних, але всі спрямовані на полегшення і автоматизацію процесів аналізу великих обсягів інформації. Далі наведено деякі з них.

NVivo – програмне забезпечення для аналізу даних, призначеним для досліджень у соціальних науках, гуманітарних науках, бізнес-аналітиці та інших областях, де важливим є глибоке розуміння текстових та мультимедійних даних. Основна перевага NVivo полягає в його здатності організовувати, управляти та аналізувати як структуровані, так і неструктуровані дані в одному зручному середовищі [14].

Програма дозволяє імпортувати та обробляти різноманітні типи даних, включаючи текстові документи, веб-сторінки, аудіо- та відеозаписи, електронні

таблиці та інші формати, що дозволяє дослідникам працювати з різноманітними джерелами інформації. Особливу увагу приділяється можливостям кодування даних, що дозволяє відзначати ключові концепції, теми та шаблони в тексті за допомогою кодів або тегів.

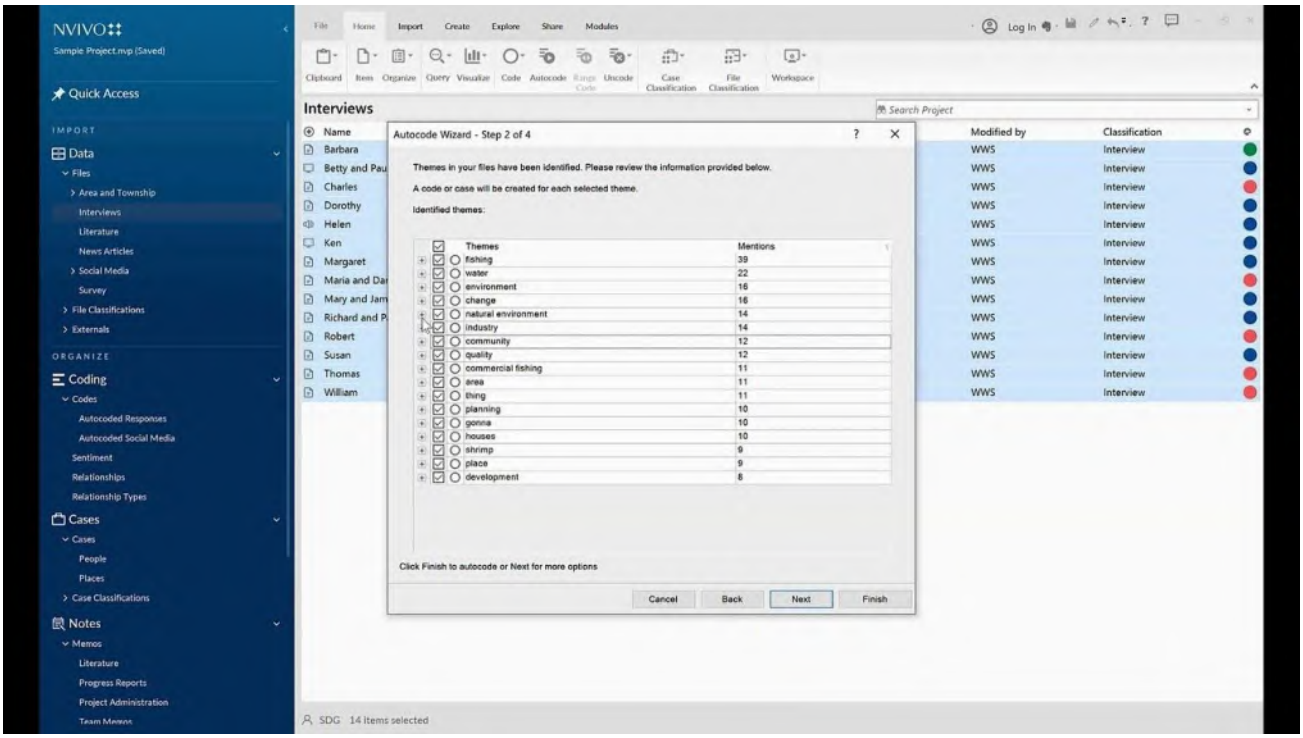


Рисунок 1.1 – Інтерфейс програми NVivo [15]

Однією з основних переваг NVivo є можливість проведення якісного і кількісного аналізу даних, використовуючи різноманітні аналітичні методики та інструменти. В програмі реалізовані засоби для візуалізації даних, порівняльного аналізу, а також можливості для спільної роботи та обміну даними між дослідниками. NVivo є необхідним інструментом для проведення складних досліджень, які вимагають розуміння контексту та змісту текстових матеріалів, забезпечуючи дослідникам інструменти для аналізу, інтерпретації та висновків.

IBM Watson Natural Language Understanding (NLU) – це розширений сервіс для аналізу текстів і розуміння природної мови, розроблений компанією IBM (рисунок 1.2). Він надає інструменти для обробки текстових даних, що дозволяють витягати інформацію з текстів у реальному часі [16].

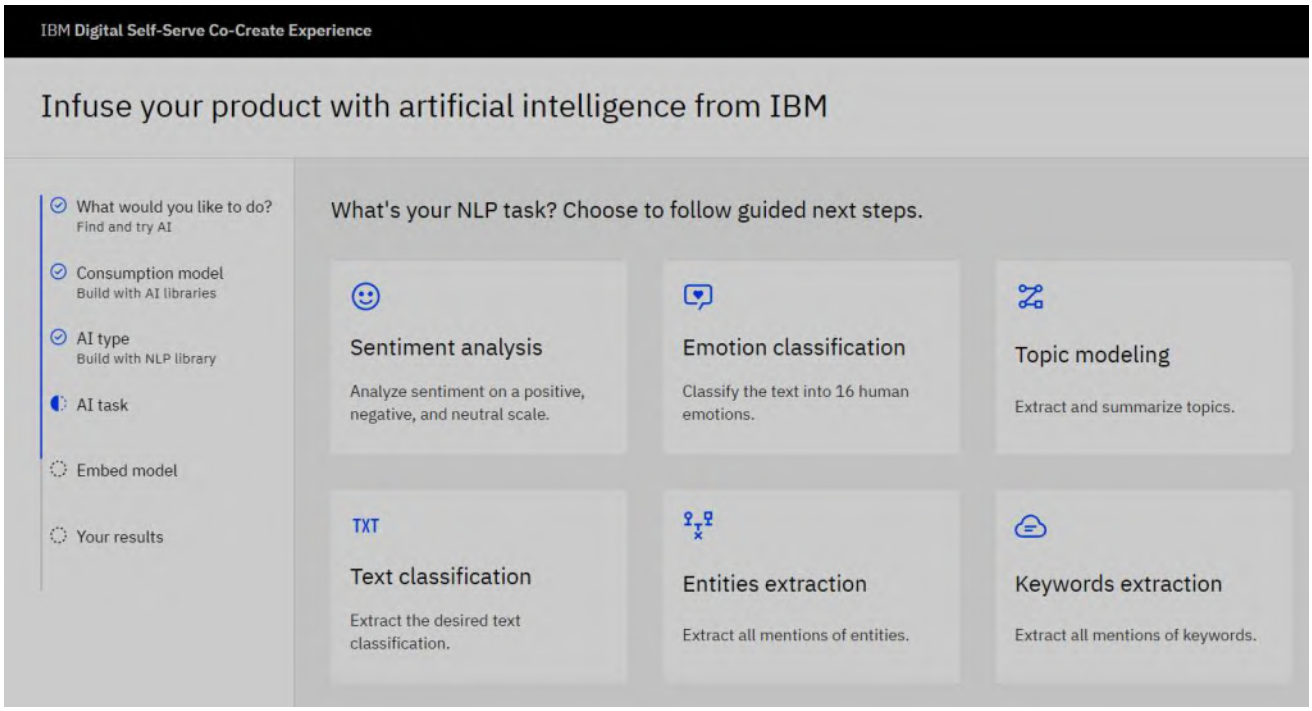


Рисунок 1.2 – Інтерфейс сервісу IBM Watson NLU [16]

IBM Watson NLU дозволяє визначати тон тексту (позитивний, негативний, нейтральний) і аналізувати семантику за допомогою алгоритмів машинного навчання. Він автоматично визначає емоційну насиченість тексту і дозволяє оцінити, як користувачі відгукувались до певного контенту. Сервіс може ідентифікувати ключові сутності (наприклад, іменовані особи, організації, місця) у тексті і визначати їхні зв'язки. Також здатний автоматично класифікувати тексти за темами або категоріями, що допомагає організувати великі обсяги текстової інформації і швидко отримувати основні висновки, виявляє синтаксичну структуру речень і визначає відносини між словами у тексті, що дозволяє розуміти контекст і смислові зв'язки.

Google Cloud Natural Language API – це інструмент для обробки природної мови, розроблений компанією Google (рисунок 1.3). API дозволяє автоматично аналізувати текстові дані для визначення сентименту, ідентифікації ключових сутностей, класифікації текстів за тематиками і розрізнення між різними частинами мови.

Основні переваги Google Cloud Natural Language API полягають у високій швидкості обробки і точності результатів завдяки використанню алгоритмів

машинного навчання. Сервіс інтегрується з іншими сервісами Google Cloud, що робить його ідеальним для використання в різних областях, включаючи аналітику даних, моніторинг соціальних медіа, аналіз настроїв користувачів та інші сценарії, де потрібно обробляти великі обсяги текстових даних в реальному часі [17].

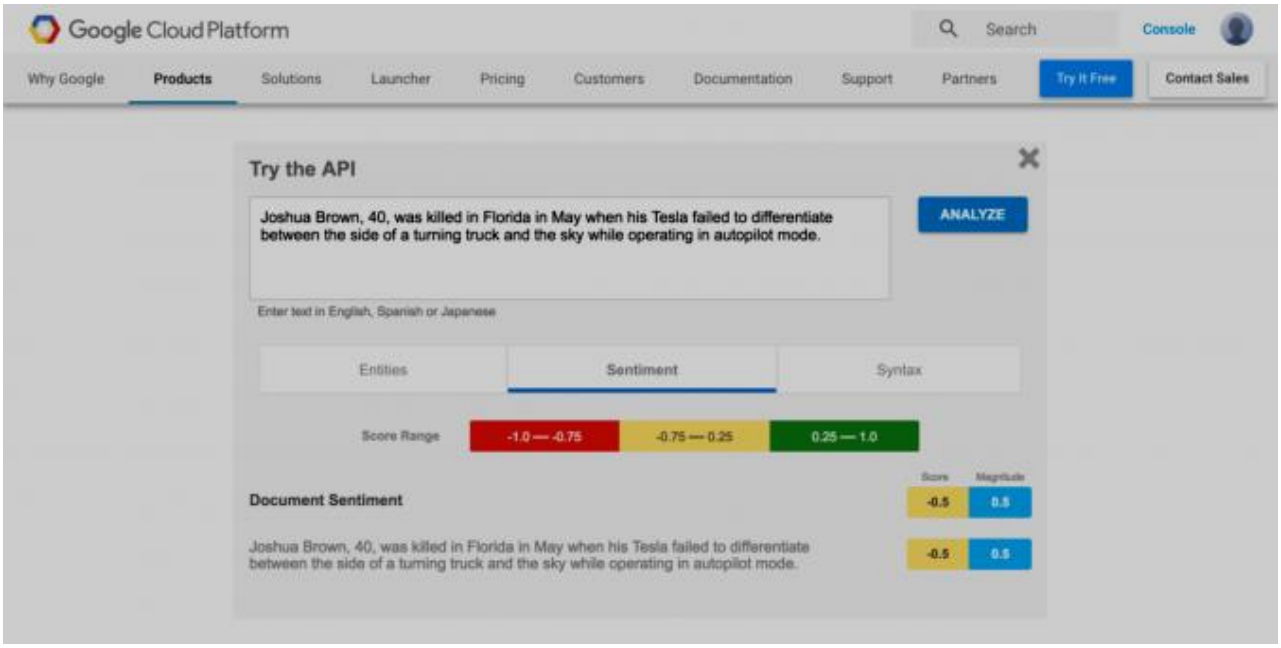


Рисунок 1.3 – Інтерфейс сервісу Google Cloud Natural Language API [17]

Крім того, Google Cloud Natural Language API підтримує різні мови, що робить його універсальним інструментом для глобальних застосувань, де необхідно аналізувати тексти різною мовою і враховувати культурні особливості та відтінки вираження емоцій.

Також проводяться наукові дослідження у сфері контент-аналізу, зокрема в задачах виявлення сентименту у коментарях та визначення їх відповідності темі обговорення. Сучасні наукові дослідження вивчають різні методи та підходи до аналізу текстових даних з метою визначення емоційного тону, ставлення користувачів до теми обговорення, а також оцінки рівня семантичної відповідності тексту до заданої теми.

У дослідженні [18] запропоновано навчену модель глибокого навчання на основі CNN і LSTM для проведення багатокласового аналізу сентименту у

коментарях бенгальських соціальних мереж. Метою дослідження є досягнення максимальної точності за допомогою запропонованої моделі та порівняльний аналіз з базовими моделями. Розглянуто шість моделей машинного навчання з двома різними методами вилучення ознак як базові моделі. Архітектура CLSTM значно покращує ефективність аналізу сентименту з точністю 85.8% та оцінкою F1 0.86 на позначеному наборі даних з 42,036 коментарів у Facebook. На основі запропонованої моделі та найефективнішої базової моделі було розроблено веб-додаток для визначення реального сентименту коментарів у соціальних мережах.

Автори статті [19] детально розглянули вплив розвитку технологій та введення Web 2.0 на соціальні медіа, що спричинило значний розвиток засобів комунікації між людьми з усього світу. Вони висвітлили в роботі, як коментарі стали популярним засобом вираження думок в інтернеті, а також зазначили, що разом із зростанням популярності коментарів з'явилося і їхнє негативне використання, включаючи вміст з ненависними або екстремістськими висловлюваннями, спрямованими проти певних осіб чи спільнот. Автори роботи запропонували використання техніки аналізу сентименту для автоматичного виявлення та фільтрації неприпустимих коментарів. Цей підхід був успішно впроваджений у клоні соціальної мережі, де дані для навчання моделі були зібрані з набору даних IMDb. Перед обробкою дані були попередньо очищені від усіх HTML-тегів, після чого вони були використані для тренування моделі машинного навчання на основі SVM. Після завершення навчання модель використовувалась для аналізу нових коментарів і визначення їхньої позитивності чи негативності. Позитивні коментарі дозволялися для публікації, тоді як негативні були автоматично видалені, а користувачі, що їх написали, позначалися як неприпустимі для уникнення подібних інцидентів у майбутньому.

Автори статті [20] досліджують вплив новинного агрегатора LINE Today на формування громадської думки в Індонезії щодо Covid-19. Вони досліджують коментарі користувачів до новин на платформі, яка дозволяє висловлювати відгуки і формувати громадську думку. Застосовуючи кількісний аналіз

контенту, автори вивчають найбільш коментовані новини про Covid-19 на LINE Today, виокремлюючи три типи чинників відгуків та три визначники громадської думки. Результати показують, що коментарі до новин про Covid-19 в LINE Today переважно спричинені відносинами між коментатором та темою новин. Громадські думки індонезійців щодо новин про Covid-19 в LINE Today в основному визначаються їхніми переконаннями, а не фактами.

Отже, як висновок варто зазначити, що контент-аналіз є актуальним в сучасному світі та дозволяє проводити різного роду дослідження. Особливо корисним є аналіз сентименту коментарів на дописи, адже це дозволяє визначити реакцію громадськості на різного роду інформацію.

1.4 Мета та задачі кваліфікаційної роботи бакалавра

Метою кваліфікаційної роботи бакалавра є підвищення якості контент-аналізу коментарів IT блогу розробників платформи Unity засобами інтелектуального аналізу даних.

Для досягнення мети, ставляться такі задачі:

– виконати аналіз інформаційних моделей в області контент-аналізу;

– розглянути засоби інтелектуального аналізу даних області контент-аналізу коротких текстових даних, та обрати підхід для реалізації;

– виконати аналіз існуючих програмних засобів та наукових рішень;

– створити метод контент-аналізу коментарів IT блогу розробників платформи Unity засобами інтелектуального аналізу даних;

– виконати розробку архітектури нейромережі для визначення сентименту коментаря;

– створити проектну архітектуру інформаційної системи IT блогу розробників платформи Unity;

– виконати проєктування бази даних;

– виконати вибір та підготовку робочих вхідних даних методу контент-аналізу коментарів IT блогу;

– розглянути особливості використання спеціалізованих програмних компонентів для спрощення програмної розробки;

– виконати вибір засобів програмної реалізації методу контент-аналізу коментарів IT блогу розробників платформи Unity засобами інтелектуального аналізу даних;

– виконати програмну реалізацію створеного методу;

– виконати тестування створеного IT блогу розробників платформи Unity та застосунку для тренування нейромереж;

– виконати дослідження ефективності створеного методу з використанням розробленої програмної реалізації.

Розділ 2 Метод контент-аналізу коментарів засобами інтелектуального аналізу даних для IT блогу розробників платформи Unity

2.1 Схема та кроки методу контент-аналізу коментарів засобами інтелектуального аналізу даних

Метод контент-аналізу коментарів засобами інтелектуального аналізу даних дозволяє за проведеним аналізом коментарів визначати їх релевантність. Схема та кроки методу наведені на рисунку 2.1.

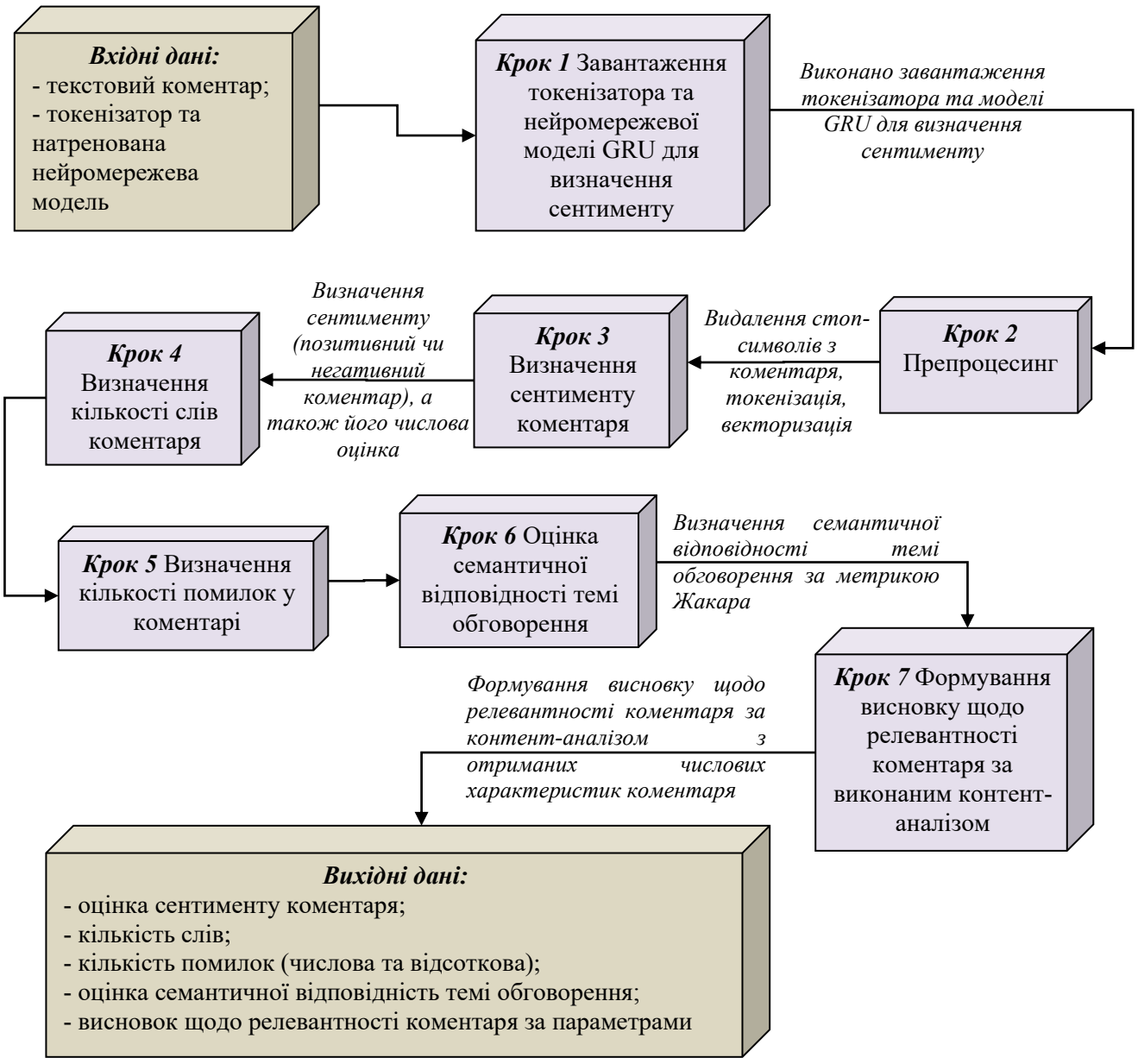


Рисунок 2.1 – Схема та кроки методу контент-аналізу коментарів засобами інтелектуального аналізу даних для IT блогу розробників платформи Unity

Метод контент-аналізу коментарів засобами інтелектуального аналізу даних працює шляхом перетворення вхідних даних у вигляді текстового коментаря до теми обговорення та токенізатора і натренованої нейромережевої моделі у вихідні дані у вигляді оцінка сентименту коментаря, кількості слів, кількість помилок (числова та відсоткова), оцінка семантичної відповідність темі обговорення та висновку щодо релевантності коментаря за параметрами.

Першим кроком є завантаження токенізатора та нейромережевої моделі GRU для визначення сентименту. На цьому ж кроці здійснюється перевірка працездатності компонентів.

Наступним кроком відбувається препроцесинг користувацького коментаря. Він включає в себе видалення стоп-символів з коментаря, токенізацію та векторизацію, адже нейромережа приймає на вхід числові вектора.

На кроці 3 відбувається визначення сентименту коментаря. Нейромережа видає результат від 0 до 1, якщо сентимент менше 0.5, він вважається негативним, в іншому випадку – позитивним.

На кроці 4 відбувається визначення кількості слів коментаря, для подальшої оцінки його інформативності. Якщо кількість слів менша від 5, коментар швидше за все має низьку інформативність, від 5 до 10 – середню, а більше 10 – достатню, однак, цей параметр є допоміжним.

На кроці 5 відбувається визначення кількості помилок у коментарі, що здійснюється як у кількісному вигляді , так і у відсотковому.

На кроці 6 обчислюється оцінка семантичної відповідності темі обговорення шляхом використання метрики Жаккара між описом теми та надісланим користувацьким коментарем.

Останнім кроком здійснюється формування висновку щодо релевантності коментаря за виконаним контент-аналізом. Він є підсумовуючим кроком для попередніх кроків.

На основі контент-аналізу довжини коментаря, кількості помилок та семантичної відповідності, можна зробити висновки щодо його релевантності, використовуючи набір правил.

Якщо коментар містить від 1 до 5 слів, він є доволі коротким і, швидше за все, нерелевантним. Коментарі середньої довжини (від 6 до 10 слів) можуть бути нерелевантними, тому варто звернути увагу на інші параметри. Довгі коментарі, що містять понад 10 слів, є більш інформативними та, ймовірно, релевантними.

Якщо відсоток помилок перевищує 20%, це свідчить про недостатній рівень володіння мовою автора. Якщо при цьому сентимент коментаря негативний, то коментар може бути нерелевантним. Особливо, якщо кількість помилок перевищує 40% і сентимент негативний, це з високою вірогідністю вказує на нерелевантність коментаря.

Семантична відповідність також важлива, коментар, семантична подібність якого менша за 15%, швидше за все не стосується теми обговорення. В той час, як коментар з показником від 15% до 25% має відношення до теми, але все ж варто звернути увагу на інші параметри. Коментар з семантичною подібністю понад 25%, то він є релевантним темі обговорення.

Вихідними даними є оцінка сентименту коментаря, кількість слів, кількість помилок (числова та відсоткова), оцінка семантичної відповідність темі обговорення, висновок щодо релевантності коментаря за параметрами.

Отже, описано схему та кроки методу контент-аналізу коментарів засобами інтелектуального аналізу даних для IT блогу розробників платформи Unity, що дозволяє за проведеним аналізом коментарів визначати їх релевантність та досягає такого ефекту шляхом перетворення вхідних даних у вигляді текстового коментаря до теми обговорення та токенізатора і натренованої нейромережевої моделі у вихідні дані у вигляді оцінка сентименту коментаря, кількості слів, кількість помилок, оцінка семантичної відповідність темі обговорення та висновку щодо релевантності коментаря за параметрами.

2.2 Аналіз та автоматизація обробки потоків даних для IT блогу розробників платформи Unity

Схема навігації між сторінками інформаційної системи у вигляді IT блогу розробників платформи Unity наведена на рисунку 2.2.

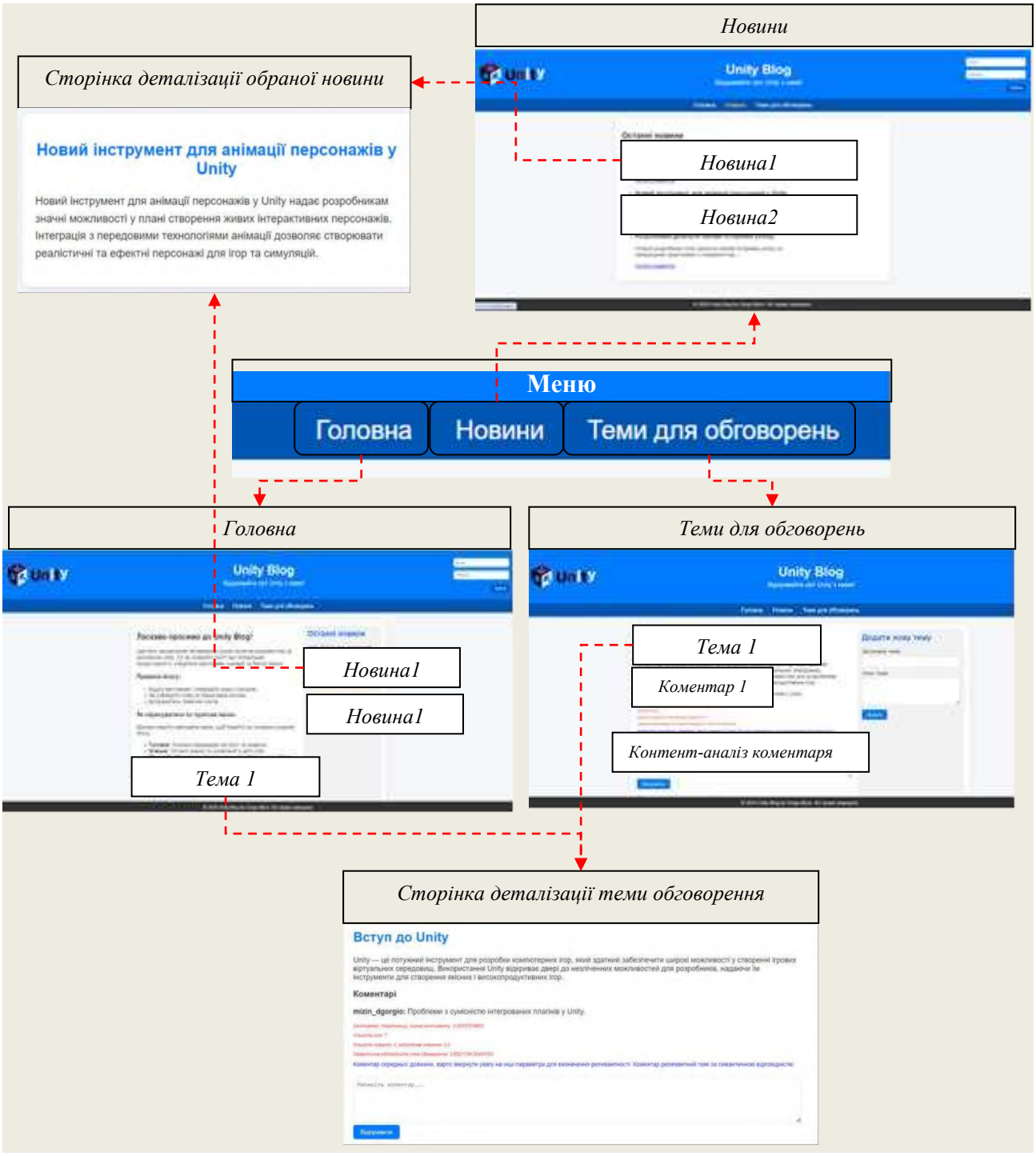


Рисунок 2.2 – Схема навігації між сторінками IT блогу розробників платформи Unity

Основана навігація між сторінками інформаційної системи організована за допомогою кнопок в меню, на які будуть застосовані тригерні переходи. З пункту меню «Головна» можна перейти до головної сторінки інформаційнох системи. Також з головної сторінки праворуч розташована інтерактивна зона для тригерного переходу на сторінку деталізації обраної новини, з якої був здійснений цей тригерний перехід. Внизу сторінки «Головна» також міститься область із організацією тригерних переходів до тем обговорень.

По кнопці головного меню «Новини» здійснюється тригерний перехід на сторінку інформаційної системи у вигляді IT блогу «Новини». Зі сторінки «Новини» організовано тригерні переходи для деталізації новин, а також перехід до інших пунктів горизонтального меню.

По кнопці меню «Теми для обговорень» здійснюється тригерний перехід на сторінку інформаційної системи у вигляді IT блогу «Теми для обговорень». Дана сторінка також містить інтерактивні компоненти тригерних переходів. З поточної сторінки також можна деталізувати обрану тему для обговорення, написати коментар та отримати його контент-аналіз.

Отже, наведено схему для автоматизації обробки потоків даних для IT блогу розробників платформи Unity.

2.3 Розробка архітектури нейронної мережі для оцінки сентименту коментарів IT блогу

Метод контент-аналізу коментарів засобами інтелектуального аналізу даних для IT блогу розробників платформи Unity вхідними даними має токенізатор та навчену нейромережеву модель.

Для оцінки сентименту коментарів IT блогу буде використано нейронну мережу класу рекурентних нейронних мереж, а саме – GRU. Спроєктована архітектура нейромережі наведена на рисунку 2.3.

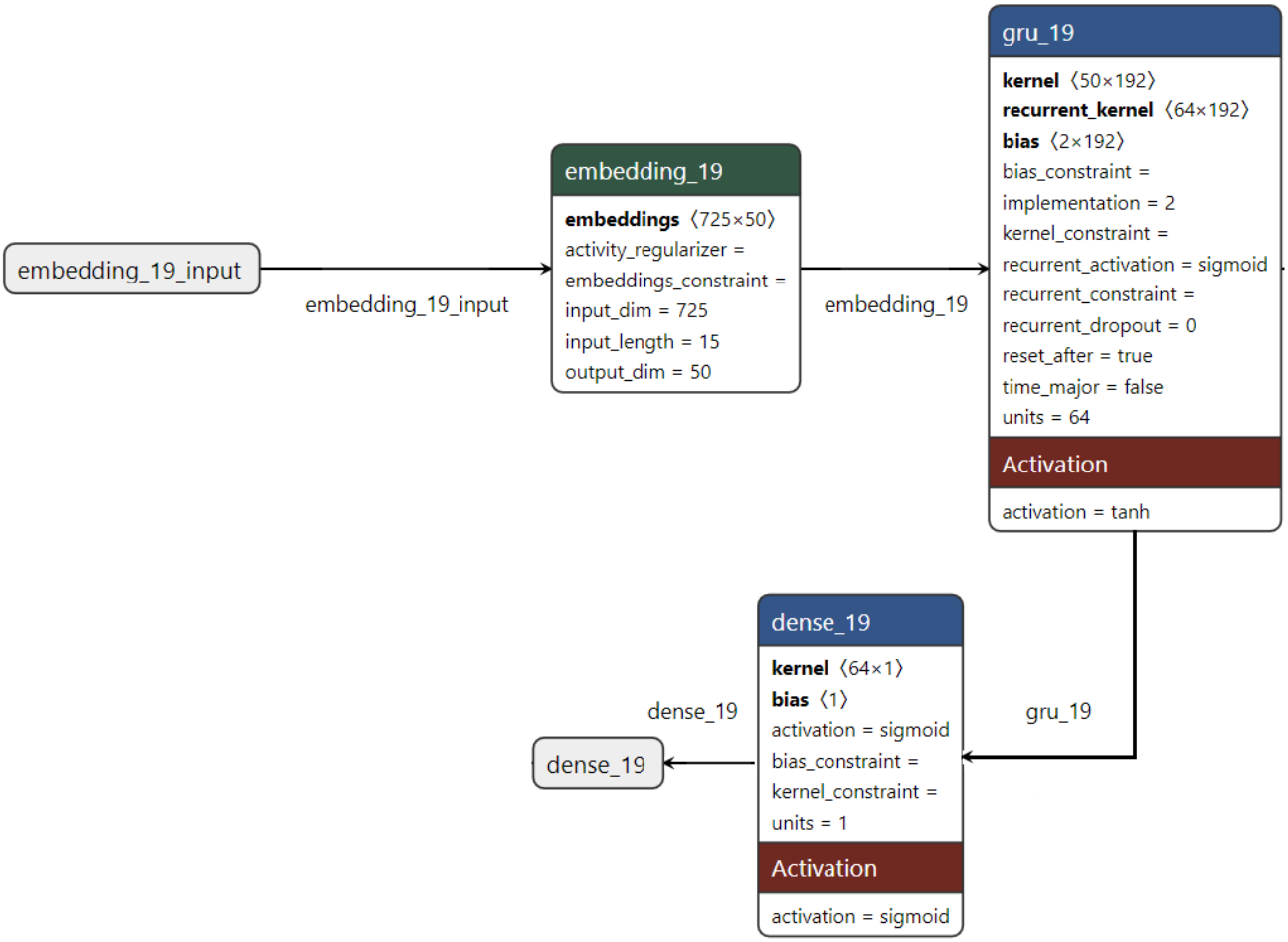


Рисунок 2.3 – Спроєктована архітектура нейромережі GRU

Вхідні дані для моделі є текстовими даними, які спочатку повинні бути перетворені на числовий формат. І тому використовується словник, де кожному слову відповідає унікальний індекс.

Перший шар моделі – це шар Embedding, який перетворює індекси слів у вектори фіксованої розмірності. Довжина цих векторів дорівнює 50 що означає, що кожне слово в послідовності буде представлено вектором з 50 чисел. Цей шар допомагає моделі зрозуміти контекст слів у реченні, перетворюючи дискретні індекси на більш інформативні вектори.

Другий шар – GRU, який має 64 нейрони. Цей шар призначений для обробки послідовних даних та отримання тимчасових залежностей у тексті. GRU шари ефективні у завданнях обробки послідовностей, оскільки вони можуть утримувати важливу інформацію про попередні слова, що особливо корисно у завданнях, пов'язаних з контент-аналізом.

Останній шар – Dense шар з одним нейроном та активацією сигмоїди. Цей шар призначений до виконання фінальної класифікації. Оскільки завдання – бінарна класифікація сентименту, вихідне значення буде в діапазоні від 0 до 1, де 0 належить класу «негативний сентимент», а 1 – «позитивний сентимент».

Отже, наведено архітектуру нейромережевої моделі, що буде використано для бінарної класифікації сентименту.

2.4 Проектна архітектура та взаємозв'язок компонентів інформаційної системи інтелектуального аналізу коментарів для IT блогу

Проектна архітектура інформаційної системи інтелектуального аналізу даних для IT блогу розробників платформи Unity представлена на рисунку 2.4.

Рисунок 2.4 – Архітектура інформаційної системи інтелектуального аналізу коментарів для IT блогу розробників платформи Unity

Інформаційна система інтелектуального аналізу коментарів для IT блогу розробників платформи Unity складається із 4-х підсистем: «Підсистема контент-аналізу коментарів», «Підсистема навчання нейромережевої моделі GRU», «Підсистема роботи з темами для обговорення», «Підсистема роботи з новинами» та бази даних.

Підсистема навчання нейромережевої моделі GRU призначена для тренування та валідації нейромережевих моделей для визначення сентименту. Вона виконує такі функції: навчання нейромережевої моделі для визначення сентименту, оцінка ефективності навченої нейромережевої моделі за метриками, збереження токенізатора, збереження навченої нейромережевої моделі для визначення сентименту. Є допоміжною підсистемою, яка надає вхідні дані для «Підсистеми контент-аналізу коментарів» у вигляді натренованої нейромережевої моделі.

Підсистема контент-аналізу коментарів призначена для визначати релевантність текстових коментарів IT блогу розробників платформи Unity, та виконує такі функції: оцінка сентименту коментаря, кількість слів, кількість помилок (числова та відсоткова), оцінка семантичної відповідність темі обговорення, висновок щодо релевантності коментаря за параметрами. Є допоміжною підсистемою для підсистеми роботи з темами для обговорення.

Підсистема роботи з темами для обговорення є головною підсистемою, яка дозволяє виконувати: додавання теми для обговорення, видалення існуючої теми для обговорення, перегляд наявних в БД тем для обговорення, перегляд наявних коментарів до теми обговорення, оцінка коментарів з використанням підсистеми контент-аналізу коментарів, виведення на екран оцінки коментарів.

Підсистема роботи з новинами призначена для взаємодії з новинами IT блогу розробників платформи Unity, та дозволяє: додавання новини, видаляти обрану новину, переглядати наявні в БД новини, деталізувати обрану новину.

Отже, сформовано проектну архітектуру інформаційної системи інтелектуального аналізу коментарів для IT блогу розробників платформи Unity, що складається із 4-х підсистем та БД.

2.5 Проєктування бази даних інформаційної системи інтелектуального аналізу коментарів для IT блогу

Згідно проведеного аналізу предметної області спроєктовано ER-діаграму (рисунок 2.5) для IT блогу розробників платформи Unity, яка відображає структуру даних та взаємозв'язки між усіма сутностями цієї структури. У наведеній ER-діаграмі зображено 5 сутностей із своїми атрибутами.

Користувачі належать до вебресурсу, на якому можуть створювати обговорення та коментувати їх. Також адміністраторам можна здійснювати та переглядати результати контент-аналізу для обраних коментарів та дописів.

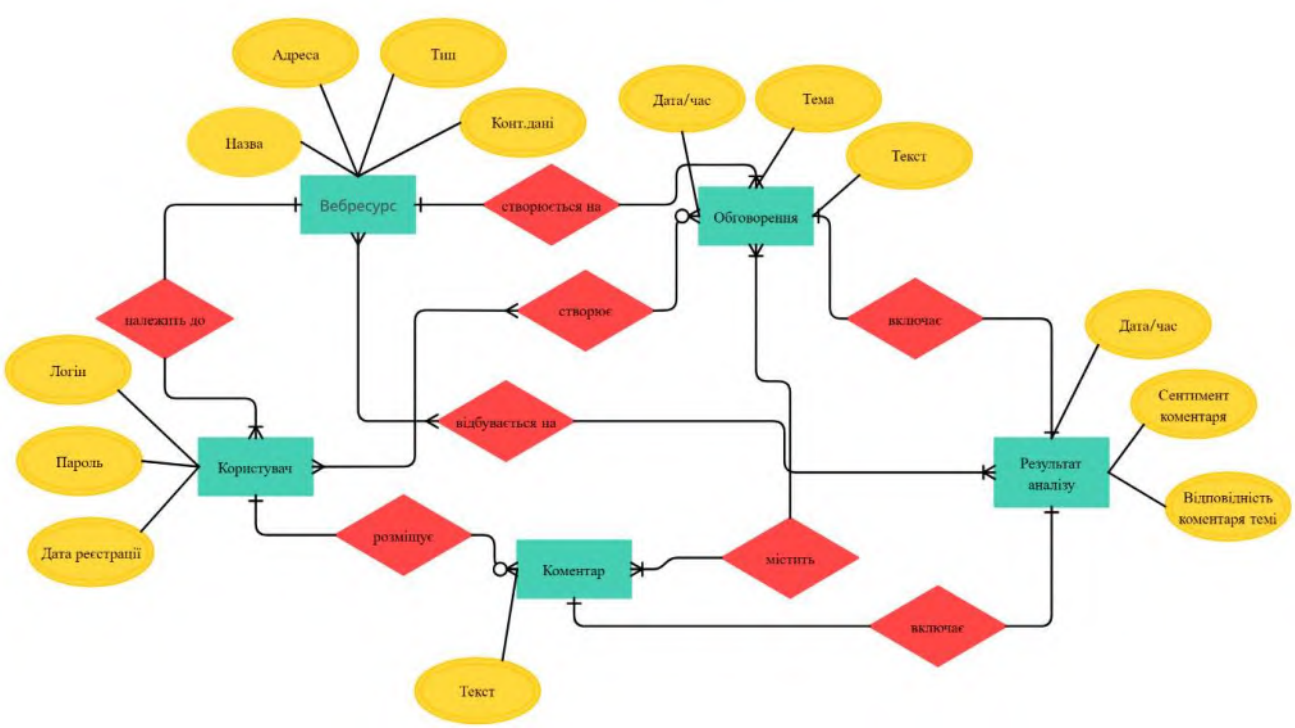


Рисунок 2.5 – ER-діаграма БД для інформаційної системи інтелектуального аналізу коментарів для IT блогу розробників платформи Unity

Згідно зі поданою ER-діаграмою IT блогу розробників платформи Unity спроєктовано даталогічну модель бази даних, яка забезпечує зберігання усіх необхідних даних. Даталогічна модель бази даних зображена на рисунку 2.6.

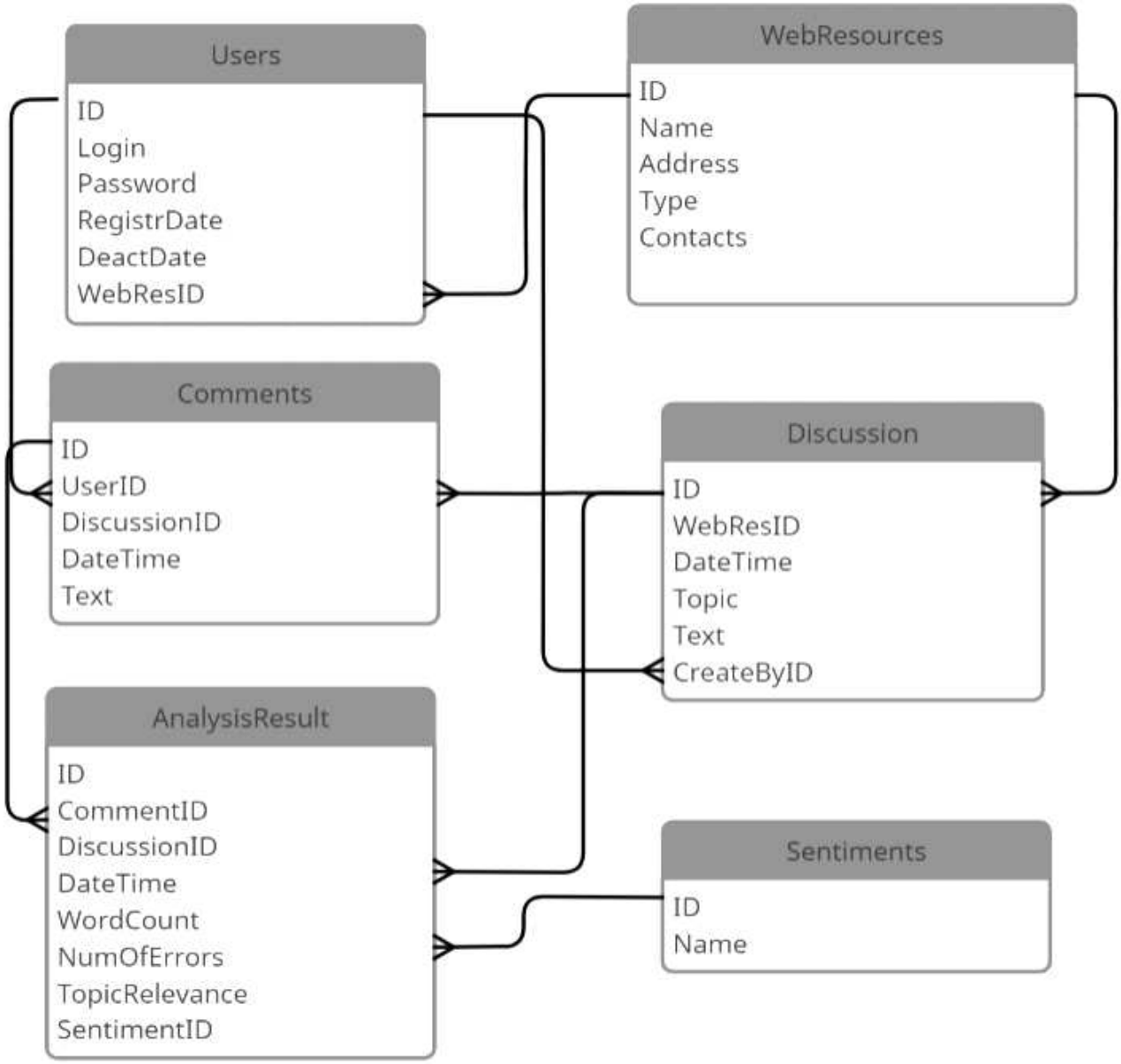


Рисунок 2.6 – Даталогічна модель бази даних інформаційної системи інтелектуального аналізу коментарів для IT блогу розробників платформи Unity

У базі даних буде 6 таблиць: WebResources, Users, Discussions, Comments, AnalysisResults, Sentiments.

Таблиця «Users» призначена для збереження даних про користувачів вебресурсу. Атрибутами таблиці є: ID, Login, Password, RegistrDate, DeactDate, WebResID (таблиця 2.1).

Таблиця 2.1 – Атрибути таблиці «Users»

№ п/п	Назва	Тип даних	Опис
1.	ID	int	Первинний ключ. Унікальний ідентифікатор користувача
2.	Login	varchar(255)	Логін користувача
3.	Password	varchar(255)	Пароль користувача
4.	RegistrDate	date/time	Дата/час реєстрації
5.	DeactDate	date/time	Дата/час деактивації профіля
6.	WebResID	int	Вторинний ключ. Посилання на запис із таблиці «WebResources». Вказує на приналежність користувача до певного вебресурсу

Таблиця «Sentiments» призначена для збереження даних про сентимент. Атрибутами таблиці є: ID, Name, Mark (таблиця 2.2).

Таблиця 2.2 – Атрибути таблиці «Sentiments»

№ п/п	Назва	Тип даних	Опис
1.	ID	int	Первинний ключ. Унікальний ідентифікатор сентименту
2.	Name	varchar(255)	Назва сентименту
3	Mark	float	Оцінка сентименту

Таблиця «Discussions» призначена для збереження даних про обговорення на вебресурсі. Атрибутами таблиці є: ID, WebResID, DateTime, Topic, Text, CreateByID (таблиця 2.3).

Таблиця 2.3 – Атрибути таблиці «Discussions»

№ п/п	**Назва**	**Тип даних**	**Опис**
1.	ID	int	Первинний ключ. Унікальний ідентифікатор обговорення
2.	WebResID	int	Вторинний ключ. Посилання на запис із таблиці «WebResources». Вказує на приналежність обговорення до певного вебресурсу
3.	DateTime	date/time	Дата/час створення обговорення
4.	Topic	varchar(255)	Тема обговорення
5.	Text	text	Текст обговорення
6.	CreateByID	int	Вторинний ключ. Посилання на запис із таблиці «Users». Вказує на користувача, що створив обговорення

Таблиця «Comments» призначена для збереження даних про коментарі на вебресурсі. Атрибутами таблиці є: ID, WebResID, DateTime, Topic, Text, CreateByID (таблиця 2.4).

Таблиця 2.4 – Атрибути таблиці «Comments»

№ п/п	Назва	Тип даних	Опис
1.	ID	int	Первинний ключ. Унікальний ідентифікатор коментаря
2.	DiscussionID	int	Вторинний ключ. Посилання на запис із таблиці «Discussions». Вказує на обговорення для якого створено коментар
3.	UserID	int	Вторинний ключ. Посилання на запис із таблиці «Users». Вказує на користувача, що створив коментар
4.	DateTime	date/time	Дата/час створення коментаря
5.	Text	varchar(255)	Текст коментаря

Таблиця 2.5 – Атрибути таблиці «WebResources»

№ п/п	Назва	Тип даних	Опис
1.	ID	int	Первинний ключ. Унікальний ідентифікатор вебресурсу
2.	Name	varchar(255)	Назва вебресурсу
3.	Address	varchar(255)	Адреса в інтернеті вебресурсу
4.	Type	varchar(255)	Тип вебресурсу
5.	Contacts	varchar(255)	Контактні дані розробників та власників

Таблиця «WebResources» призначена для збереження даних про вебресурс. Атрибутами таблиці є: ID, Name, Address, Type, Contacts (таблиця 2.5).

Таблиця 2.6 – Атрибути таблиці «AnalysisResult»

№ п/п	Назва	Тип даних	Опис
1.	ID	int	Первинний ключ. Унікальний ідентифікатор результаті контент-аналізу
2.	CommentID	int	Вторинний ключ. Посилання на запис із таблиці «Comments». Вказує на коментар для якого зберігаються результати контент-аналізу
3.	DiscussionID	int	Вторинний ключ. Посилання на запис із таблиці «Discussions». Вказує на обговорення до якого належить коментар з результату дослідження
4.	DateTime	date/time	Дата/час проведення контент-аналізу
5.	WordCount	int	Кількість слів у тексті коментаря
6.	NumOfErrors	int	Кількість помилок у тексті коментаря
7.	TopicRelevance	varchar(255)	Текст коментаря
8.	SentimentID	int	Вторинний ключ. Посилання на запис із таблиці «Sentiments». Вказує на сентимент тексту коментаря, що визначено

Таблиця «AnalysisResult» призначена для збереження даних про проведені контент-аналізи коментарів до обговорень на вебресурсі. Атрибутами

таблиці є: ID, CommentID, DiscussionID, DateTime, WordCount, NumOfErrors, TopicRelevance, SentimentID (таблиця 2.6).

Отже, була спроєктована даталогічна модель бази даних, яка містить 6 таблиць, а саме WebResources, Users, Discussions, Comments, AnalysisResults, Sentiments. Спроєктованя таблиці дозволять зберігіати усі необхідні дані для роботи методу контент-аналізу коментарів засобами інтелектуального аналізу даних для IT блогу розробників платформи Unity.

2.6 Підготовка робочих вхідних даних для інформаційної системи інтелектуального аналізу коментарів для IT блогу

Для реалізації методу контент-аналізу коментарів засобами інтелектуального аналізу даних для IT блогу розробників платформи Unity необхідно підготувати робочих вхідних даних для системи. Для цього було обрано «IMDB Dataset of 50K Movie Reviews» [21].

Набір даних IMDB з 50 тис. рецензій на фільми доступний на Kaggle [22] і містить велику колекцію рецензій на фільми із супровідними мітками настроїв (позитивних чи негативних). Цей набір даних використовується для задач обробки природної мови, таких як аналіз настроїв і класифікація текстів. Він містить близько 50 000 рецензій, розділених на навчальну та тестову вибірки. Кожен відгук позначений як позитивний або негативний, що робить його зручним для побудови та оцінки моделей машинного навчання.

На рисунку 2.7 відображено вигляд записів в датасеті.

Також було проведено статистичне дослідження вмісту датасету:

- кількість унікальних відгуків: 49,582;
- кількість унікальних відгуків: 2 (позитивний і негативний);
- загальна кількість відгуків: 50,000;
- позитивні: 25 000 відгуків;
- негативні: 25 000 відгуків.

	review	sentiment
0	One of the other reviewers has mentioned that ...	positive
1	A wonderful little production. The...	positive
2	I thought this was a wonderful way to spend ti...	positive
3	Basically there's a family where a little boy ...	negative
4	Petter Mattei's "Love in the Time of Money" is...	positive

Рисунок 2.7 – Попередній перегляд записів в датасеті

На рисунку 2.8 наведено діаграму розподілу позитивних та негативних відгуків у датасеті.

Рисунок 2.8 – Діаграма розподілу позитивних та негативних відгуків у датасеті

Цей набір даних є повним, без пропущених значень в колонці відгуків та колонці настроїв. Існує збалансований розподіл позитивних і негативних

відгуків, кожна з яких становить 50% набору даних. Більшість відгуків є унікальними, з невеликою кількістю дублікатів.

Також було здійснено машинний переклад датасету на українську мову за допомогою бібліотеки googletrans. В результаті виконання перекладу було отримано датасет з відгуками українською мовою. На рисунках 2.9 та 2.10 наведено хмару слів для позитивних та негативних відгуків в отриманому датасеті.

Рисунок 2.9 – Хмара слів для позитивних відгуків в україномовному датасеті

Рисунок 2.10 – Хмара слів для негативних відгуків в україномовному датасеті

Таким чином, було сформовано датасет для подальшої роботи над реалізацією методу контент-аналізу коментарів засобами інтелектуального аналізу даних для IT блогу розробників платформи Unity, що буде використано для навчання нейромережі для визначення сентименту.

2.7 Особливості використання спеціалізованих програмних компонентів

Для створення методу контент-аналізу коментарів засобами інтелектуального аналізу даних для IT блогу розробників платформи Unity необхідно обрати спеціалізовані програмні компоненти.

Python має багату екосистему бібліотек, які значно спрощують розробку програм та виконання складних завдань. Однією з таких бібліотек є pickle, яка дозволяє серіалізувати і десеріалізувати об’єкти, забезпечуючи зручне збереження стану програм і модулів.

Бібліотека pickle [23] є стандартною бібліотекою в Python, призначеною для серіалізації та десеріалізації об'єктів. Серіалізація – це процес перетворення об'єкта в байтовий потік, що дозволяє зберігати об'єкти в файлах або передавати їх по мережі. Відповідно, десеріалізація – це процес відновлення об'єктів з байтового потоку до їх оригінальної форми. Pickle підтримує серіалізацію більшості вбудованих типів Python, таких як числа, рядки, списки, кортежі, словники та користувацькі об'єкти. Це робить його корисним для збереження стану програм, збереження моделі машинного навчання та інших об'єктів для подальшого використання

Також необхідно використати Flask [24], що є мікрофреймворком для веб-розробки на Python, який забезпечує мінімальну базу для створення веб-додатків, надаючи розробникам простоту та гнучкість. Flask базується на концепції Web Server Gateway Interface та використовує Jinja2 як шаблонізатор. Це дозволяє легко створювати динамічні веб-сторінки, які можуть відображати дані з серверної частини додатку. Flask також підтримує розширення, які додають

функціональність до основного фреймворку, такі як Flask-SQLAlchemy для роботи з базами даних, Flask-WTF для інтеграції форм та Flask-Login для управління аутентифікацією користувачів.

Однією з ключових особливостей Flask є його модульна структура, яка дозволяє розробникам вибирати тільки ті компоненти, які їм потрібні, що робить додаток легшим та швидшим. Flask також добре підходить для розробки RESTful API, оскільки він забезпечує зручні засоби для обробки HTTP-запитів і відповіді в форматі JSON.

Бібліотека Keras [25] є високорівневим API для побудови та навчання нейронних мереж, яка працює поверх TensorFlow. Keras спрощує процес створення складних моделей глибокого навчання завдяки своїй інтуїтивно зрозумілій і зручній для користувача архітектурі. Вона розроблена для швидкої розробки прототипів, що робить її ідеальною для досліджень і експериментів.

Keras підтримує як послідовні, так і функціональні API, що дозволяє створювати як прості, так і складні моделі з декількома вхідними і вихідними каналами. Бібліотека містить різноманітні інструменти для підготовки даних, включаючи утиліти для обробки послідовностей, що полегшує роботу з текстовими та іншими даними. Це включає функції для токенізації тексту, перетворення його в послідовності чисел, а також паддінг для приведення послідовностей до однакової довжини.

Однією з ключових особливостей Keras є її модульність і розширюваність. Вона підтримує широкий спектр шарів нейронних мереж, включаючи щільні (Dense), згорткові (Convolutional), рекурентні (Recurrent) шари та багато інших.

Keras також підтримує різні оптимізатори, функції втрат і метрики, що дозволяє тонко налаштовувати процес навчання моделей. Крім того, бібліотека включає в себе інструменти для візуалізації процесу навчання та оцінки продуктивності моделей, що полегшує діагностику та усунення помилок.

Завдяки інтеграції з TensorFlow [26], Keras може використовувати потужні можливості цієї бібліотеки для розподіленого навчання та роботи на графічних процесорах, що значно прискорює процес навчання моделей.

Бібліотека NLTK [27] є потужним інструментом для обробки природної мови, яка була розроблена для підтримки досліджень і розробок в галузі комп'ютерної лінгвістики та обробки тексту. Вона включає в себе велику кількість текстових корпусів, таких як Браунівський корпус, корпус Гутенберга, корпус веб-текстів і багато інших. Крім корпусів, NLTK надає лексичні ресурси, зокрема, WordNet, який є великою лексичною базою даних англійської мови.

Серед інструментів NLTK варто виділити токенізацію, яка дозволяє розбивати текст на окремі слова або речення, що є першим кроком в більшості задач обробки тексту. NLTK також надає інструменти для тегування частин мови (POS tagging), що дозволяє автоматично визначати граматичні категорії слів у тексті, такі як іменники, дієслова, прикметники тощо. Іншою важливою функцією є синтаксичний аналіз, який дозволяє визначати граматичну структуру речень.

Бібліотека Stanza [28], розроблена Стенфордським університетом, є ще одним потужним інструментом для обробки природної мови. Вона забезпечує багатомовну підтримку та надає інструменти для різних аспектів аналізу тексту. Stanza включає інструменти для морфологічного аналізу, який дозволяє визначати граматичні властивості слів, такі як відмінок, число, рід і час.

Синтаксичний аналіз у Stanza дозволяє будувати дерева залежностей для речень, що допомагає зрозуміти структуру та взаємозв'язки між словами в реченні. Це особливо корисно для більш глибокого аналізу тексту та розробки додатків, які потребують розуміння контексту, таких як автоматичне узагальнення текстів або запитання-відповіді системи.

Stanza також забезпечує високоточне розпізнавання іменованих сутностей, що дозволяє автоматично виділяти та класифікувати імена осіб, організацій, місць та інших сутностей у тексті. Це робить Stanza цінним

інструментом для завдань, пов'язаних з вилученням інформації з текстових даних.

Бібліотека requests [29] є однією з найпопулярніших бібліотек для роботи з HTTP-запитами в Python. Вона забезпечує простий і зручний інтерфейс для надсилання HTTP-запитів та обробки відповідей, що робить її ідеальною для взаємодії з веб-сервісами та API. Використовуючи requests, розробники можуть легко здійснювати GET, POST, PUT, DELETE та інші HTTP-запити. Бібліотека автоматично обробляє такі аспекти, як кодування URL, формування заголовків запитів, управління сесіями та обробка файлів кукі. Крім того, requests підтримує аутентифікацію, SSL-сертифікати та обробку помилок, що робить її надзвичайно зручною для розробки додатків, які потребують інтеграції з іншими веб-службами або API.

Бібліотека datetime [30] надає класи для роботи з датами та часом у Python. Вона включає в себе різні функції для маніпуляції з датами та часом, обчислення тривалості подій, встановлення дедлайнів і виконання арифметичних операцій з датами. Основні класи, такі як datetime, date, time та timedelta, дозволяють легко створювати, порівнювати та обчислювати дати і час. Наприклад, за допомогою datetime можна визначити поточний час, додати або відняти дні, години або хвилини, обчислити різницю між двома датами та форматувати дати у різні рядкові представлення. Це робить datetime незамінною для задач, пов'язаних з управлінням подіями, розкладом, таймерами та будь-якими іншими операціями, які вимагають точного контролю над датами і часом.

Таким чином, ці бібліотеки разом забезпечують потужний інструментарій для розробки методу контент-аналізу коментарів засобами інтелектуального аналізу даних для IT блогу розробників платформи Unity. Requests спрощує взаємодію з веб-сервісами та API, Flask надає платформу для створення веб-додатків, а Keras і Stanza забезпечують інструменти для машинного навчання та обробки природної мови, Keras буде використано для завантаження навченої нейромережевої моделі визначення сентименту. NLTK додає можливості для

аналізу тексту, а datetime дозволяє ефективно працювати з датами та часом, що буде доцільним для контролю над життєвим циклом сесії.

2.8 Висновки до розділу 2

У ході виконання другого розділу було створено метод контент-аналізу коментарів засобами інтелектуального аналізу даних для IT блогу розробників платформи Unity, що дозволяє за проведеним аналізом коментарів визначати їх релевантність, а також сприяє підвищенню якості контент-аналізу коментарів IT блогу розробників платформи Unity.

Метод працює шляхом перетворення вхідних даних у вигляді текстового коментаря до теми обговорення та токенізатора і натренованої нейромережевої моделі у вихідні дані у вигляді оцінка сентименту коментаря, кількості слів, кількість помилок, оцінка семантичної відповідність темі обговорення та висновку щодо релевантності коментаря за параметрами.

Наведено схему автоматизації обробки потоків даних для інформаційної системи інтелектуального аналізу коментарів для IT блогу розробників платформи Unity, описано основні моменти взаємодії між сторінками інформаційної системи.

Наведено архітектуру нейромережевої моделі, що буде використано для бінарної класифікації сентименту.

Сформовано проектну архітектуру інформаційної системи інтелектуального аналізу коментарів для IT блогу розробників платформи Unity, що складається із 4-х підсистем та БД, та виконує такі основні функції:

- оцінка сентименту коментаря;
- обчислення кількості слів;
- обчислення кількості помилок (числова та відсоткова);
- оцінка семантичної відповідність темі обговорення;
- висновок щодо релевантності коментаря за параметрами;
- додавання теми для обговорення;

– видалення існуючої теми для обговорення;

– перегляд наявних в БД тем для обговорення;

– перегляд наявних коментарів до теми обговорення;

– оцінка коментарів з використанням підсистеми контент-аналізу коментарів;

– виведення на екран оцінки коментарів;

– навчання нейромережевої моделі для визначення сентименту;

– оцінка ефективності навченої нейромережевої моделі за метриками;

– збереження токенізатора;

– збереження навченої нейромережевої моделі для визначення сентименту;

– додавання новини;

– видалення обраної новини;

– перегляд наявних в БД новин;

– деталізація обраної новини.

Інформаційна система інтелектуального аналізу коментарів для IT блогу розробників платформи Unity складається із таких підсистем: «Підсистема контент-аналізу коментарів», «Підсистема навчання нейромережевої моделі GRU», «Підсистема роботи з темами для обговорення», «Підсистема роботи з новинами» та бази даних.

Спроєктовано даталогічну модель бази даних, яка містить 6 таблиць, а саме WebResources, Users, Discussions, Comments, AnalysisResults, Sentiments. Спроєктованя таблиці дозволять збергіати усі необхідні дані для роботи методу контент-аналізу коментарів засобами інтелектуального аналізу даних для IT блогу розробників платформи Unity.

Сформовано датасет для подальшої роботи над реалізацією методу контент-аналізу коментарів засобами інтелектуального аналізу даних для IT блогу розробників платформи Unity, що буде використано для навчання нейромережі для визначення сентименту.

Визначено набір спеціалізованих програмних компонентів, що будуть використані для подальшої розробки інформаційної системи у вигляді IT блогу розробників платформи Unity.

Розділ 3 Експериментальне дослідження методу контент-аналізу коментарів для IT блогу розробників платформи Unity

3.1 Визначення шляхів дослідження та засобів створення інформаційної системи інтелектуального аналізу коментарів

За створеним методом контент-аналізу коментарів засобами інтелектуального аналізу даних необхідно створити програмну реалізацію у вигляді застосунку для навчання типових моделей GRU для визначення сентименту, а також необхідно створити програмну реалізацію у вигляді інформаційної системи інтелектуального аналізу коментарів для IT блогу розробників платформи Unity. Дана програмна реалізація буде слугувати для дослідження ефективності розробленого методу. Програмну реалізацію необхідно протестувати з використанням тест-кейсів.

Щодо проведення дослідження ефективності – необхідно дослідити ефективність нейромережі визначати сентимент для коментарів IT блогу розробників платформи Unity з використанням метрик. Буде використано метрику Accuracy та Loss.

Метрика Accuracy (точність) вимірює частку правильних прогнозів моделі щодо загальної кількості прогнозів. У контексті завдання визначення сентименту, точність показує, наскільки добре модель може правильно класифікувати коментарі як позитивні чи негативні.

Метрика Loss (втрати) вказує на ступінь розбіжності передбачень моделі з реальними значеннями. У цій задачі буде використано бінарну крос-ентропію, яка вимірює різницю між передбаченим значенням ймовірності та фактичною міткою (позитивною чи негативною). Найменші значення функції втрат вказують на кращу відповідність передбачень моделі реальних даних, що свідчить про успішніше навчання моделі.

Для дослідження методу контент-аналізу коментарів засобами інтелектуального аналізу даних буде застосовано підхід з залученням експерта,

який дасть відповідь щодо релевантності набору коментарів, які будуть порівняні із відповідями програмної реалізації.

3.2 Вибір засобів розробки інформаційної системи інтелектуального аналізу коментарів для IT блогу розробників платформи Unity

Для розробки інформаційної системи інтелектуального аналізу коментарів для IT блогу розробників платформи Unity важливо обрати оптимальні засоби, які відповідають специфіці завдання і забезпечать ефективне функціонування програмного продукту. Серед найважливіших засобів є мова програмування, яка визначає можливості системи та швидкість розробки, середовище розробки, яке надає необхідні інструменти для написання, тестування та налагодження коду. Так як інформаційна система інтелектуального аналізу коментарів для IT блогу розробників платформи Unity буде у вигляді сайту, то необхідно обрати фрейморк для реалізації веб-інтерфейсу. Також важливим є вибір сервісу для тренування нейромережі. Оскільки інформаційна система базується на обробці даних, необхідно визначити систему керування базами даних (СКБД), яка забезпечить ефективне зберігання та обробку інформації, а також мову запитів, що визначає спосіб взаємодії з даними.

Мова програмування Python є обрана для розробки інформаційної системи завдяки своїм високорівневим інструментам для швидкого прототипування і аналізу даних. Його простота синтаксису сприяє швидкій інтеграції нових ідей та концепцій у програмному коді. Зокрема, у сфері штучного інтелекту Python є популярним вибором, оскільки має розгалужені бібліотеки для машинного навчання, такі як TensorFlow, PyTorch і Scikit-learn, що спрощують тренування моделей та аналіз результатів [31].

Для створення веб-інтерфейсу для інформаційної системи інтелектуального аналізу коментарів для IT блогу розробників платформи Unity було обрано фреймворк Flask з урахуванням його простоти у використанні і

здатності швидко розгортати веб-додатки. Flask надає мінімальну конфігурацію і гнучкість у побудові RESTful API, що дозволяє швидко і ефективно інтегрувати функціональність інформаційної системи з веб-інтерфейсом. Велика кількість розширень для Flask також робить його хорошим вибором для розширення можливостей системи за потреби дослідження та розвитку [32].

У якості середовища розробки обрано PyCharm з огляду на його потужність і зручність для роботи з мовою програмування Python. PyCharm забезпечує інтегровану підтримку відладки, автодоповнення коду, контроль версій і керування проектами, що робить його ідеальним інструментом для продуктивної розробки. Його підтримка великої кількості плагінів і інтеграція з популярними системами управління версіями, такими як Git, дозволяють ефективно керувати проектом і співпрацювати у команді. Крім того, PyCharm має потужні інструменти для аналізу коду і рефакторингу [33]

Для тренування нейромережі в рамках розробки інформаційної системи було обрано сервіс Google Colab з урахуванням його безкоштовної доступності та потужності обчислювальних ресурсів. Google Colab надає віртуальне середовище з передвстановленими бібліотеками для машинного навчання, такими як TensorFlow, PyTorch та інші, що дозволяє швидко налаштувати і почати роботу з проектом без необхідності установки додаткових програмних засобів. Крім того, сервіс автоматично масштабує обчислювальні ресурси в залежності від потреб користувача, що забезпечує можливість ефективно працювати з великими обсягами даних та складними моделями нейронних мереж

У якості мови запитів для бази даних в інформаційній системі було обрано SQL з огляду на його широке поширення, стандартизованість і потужність у взаємодії з реляційними базами даних. SQL дозволяє ефективно створювати, змінювати, управляти та оптимізувати бази даних, надаючи зручний і чіткий синтаксис для виконання різноманітних операцій, від простих запитів до складних аналітичних операцій [34].

Для зберігання та керування даними в інформаційній системі було обрано систему керування базами даних MySQL з огляду на її надійність, широке поширення. MySQL є однією з найпопулярніших відкритих реляційних СКБД, що відома своєю швидкодією, ефективним управлінням транзакціями та підтримкою великих обсягів даних. Вона підтримує широкий спектр операцій та має велику кількість розширень і інтеграцій, що робить її ідеальним вибором для складних інформаційних систем, де важлива якість та надійність обробки даних. Крім того, MySQL є безкоштовним програмним забезпеченням з активною спільнотою користувачів і розробників, що забезпечує постійну підтримку та оновлення системи [35].

Отже, для розробки інформаційної системи інтелектуального аналізу коментарів для IT блогу розробників платформи Unity було обрано мову програмування Python, фреймворк Flask для створення веб-інтерфейсу, середовище розробки PyCharm, сервіс Google Colab для тренування нейромережі, мова запитів SQL для взаємодії з базою даних, СКБД MySQL, яка відповідає за зберігання та керування даними.

3.3 Структура та функціональне призначення програмних складових інформаційної системи інтелектуального аналізу коментарів для IT блогу

Структура складових інформаційної системи інтелектуального аналізу коментарів для IT блогу розробників платформи Unity наведена на рисунку 3.1. Складається наведена структура із 6-и класів: «Post», «NewsItem», «Comment», «CommentProcessor», «User», «AppRoutes».

Клас «Post» відповідає за зберігання інформації про пости в базі даних. Він також містить методи для додавання коментарів до посту та отримання всіх постів або посту за його ідентифікатором.

Клас «NewsItem» відповідає за зберігання новин в базі даних і має методи для отримання всіх новин або новини за її ідентифікатором.

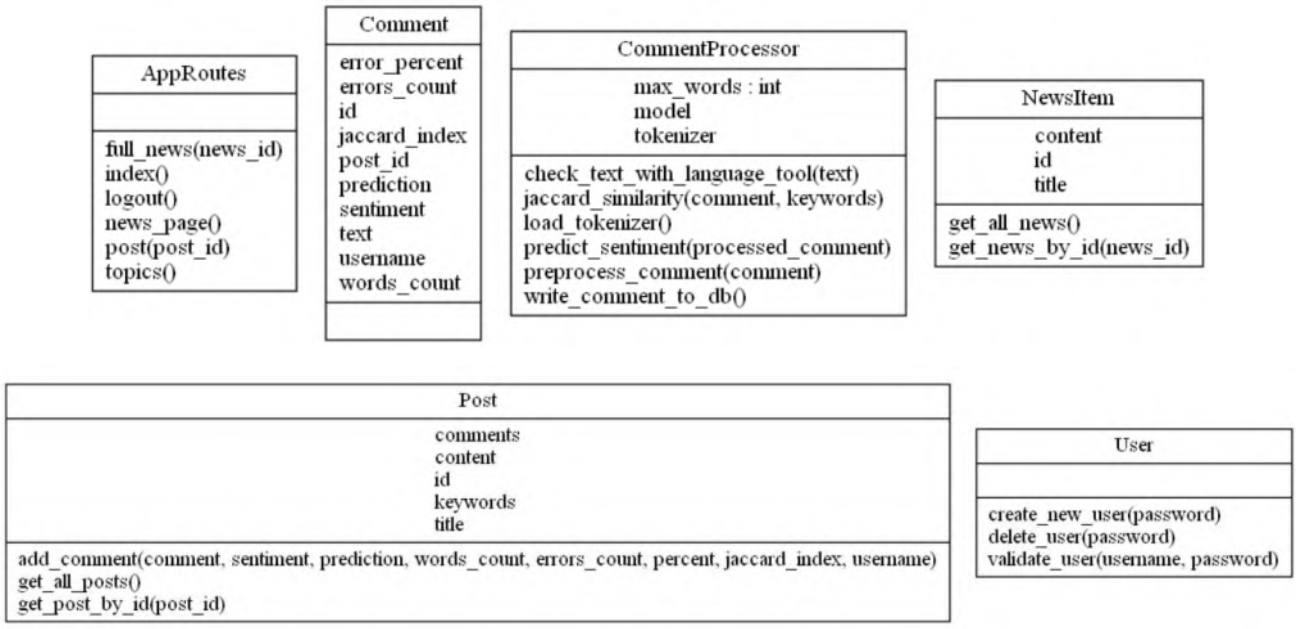


Рисунок 3.1 – Діаграма класів інформаційної системи інтелектуального аналізу коментарів для IT блогу розробників платформи Unity

Клас «Comment» відповідає за зберігання коментарів до постів у базі даних. Він містить інформацію про текст коментаря, його емоційний тон, кількість слів, кількість помилок та інші метрики.

Клас «CommentProcessor» відповідає за обробку коментарів, включаючи передобробку тексту, передбачення емоційного тону коментаря, підрахунок кількості помилок та обчислення коефіцієнта Жаккара між коментарем та ключовими словами посту.

Клас «User» відповідає за валідацію користувачів, зберігання та перевірку даних про користувачів, які зберігаються у БД. Він також містить методи для створення та видалення користувачів.

Клас «AppRoutes», який визначає маршрути для обробки запитів у Flask-додатку, але він не є типовим класом для зберігання даних чи логіки обробки.

Отже, наведено структуру та описано функціональне призначення програмних складових системи IT блогу розробників платформи Unity. Структура майбутнього вебзастосунку представлена у вигляді діаграми класів, кожен з яких описано згідно його функціонального призначення.

3.4 Особливості реалізації програмних складових інформаційної системи інтелектуального аналізу коментарів для IT блогу

Першою підсистемою інформаційної системи інтелектуального аналізу коментарів для IT блогу розробників платформи Unity, що буде розглянуто з точки зору програмної реалізації буде «Підсистема навчання нейромережевої моделі GRU».

Для реалізації підсистеми виконується кілька ключових етапів для підготовки даних і навчання моделі GRU для задачі знаходження сентименту коментаря IT блогу розробників платформи Unity. Спочатку проводиться токенізація текстових коментарів за допомогою класу Tokenizer, який перетворює текст на послідовності чисел, де кожне число відповідає певному слову з словника. Потім ці послідовності приводяться до однакової довжини (max_words), використовуючи функцію pad_sequences (), щоб усі коментарі мали однакову кількість слів.

Після цього токенізатор зберігається у файл для подальшого використання, що дозволяє відтворити процес токенізації на нових даних. Дані розділяються на тренувальний і тестовий набори за допомогою функції train_test_split(), що забезпечує випадковий поділ даних на 80% для навчання і 20% для тестування.

Далі будується модель нейронної мережі з використанням архітектури GRU. Модель складається з трьох основних шарів: шар вбудовування (Embedding), шар GRU і вихідний щільний шар (Dense) з активацією sigmoid для виконання класифікації. Модель компілюється з оптимізатором Adam і функцією втрат binary_crossentropy(), а також метрикою accuracy.

Нарешті, модель навчається на тренувальних даних протягом визначеної кількості епох, причому кожна епоха оцінюється на тестових даних, щоб відстежувати продуктивність моделі. Приклад логів процесу навчання моделі на 20 епох наведено на рисунку 3.2.

```
Epoch 1/20
9/9 [==============================] - 3s 76ms/step - loss: 0.6864 - accuracy: 0.6385 - val_loss: 0.6854 - val_accuracy: 0.6212
Epoch 2/20
9/9 [==============================] - 0s 15ms/step - loss: 0.6704 - accuracy: 0.8115 - val_loss: 0.6764 - val_accuracy: 0.6364
Epoch 3/20
9/9 [==============================] - 0s 14ms/step - loss: 0.6483 - accuracy: 0.8346 - val_loss: 0.6623 - val_accuracy: 0.6667
Epoch 4/20
9/9 [==============================] - 0s 14ms/step - loss: 0.6099 - accuracy: 0.8731 - val_loss: 0.6377 - val_accuracy: 0.7727
Epoch 5/20
9/9 [==============================] - 0s 14ms/step - loss: 0.5455 - accuracy: 0.8846 - val_loss: 0.5931 - val_accuracy: 0.7727
Epoch 6/20
9/9 [==============================] - 0s 15ms/step - loss: 0.4337 - accuracy: 0.9269 - val_loss: 0.5171 - val_accuracy: 0.7576
Epoch 7/20
9/9 [==============================] - 0s 14ms/step - loss: 0.3150 - accuracy: 0.9269 - val_loss: 0.4480 - val_accuracy: 0.7576
Epoch 8/20
9/9 [==============================] - 0s 15ms/step - loss: 0.2001 - accuracy: 0.9538 - val_loss: 0.3989 - val_accuracy: 0.8182
Epoch 9/20
9/9 [==============================] - 0s 14ms/step - loss: 0.1224 - accuracy: 0.9692 - val_loss: 0.3781 - val_accuracy: 0.8182
Epoch 10/20
9/9 [==============================] - 0s 15ms/step - loss: 0.0710 - accuracy: 0.9846 - val_loss: 0.3450 - val_accuracy: 0.8485
Epoch 11/20
9/9 [==============================] - 0s 17ms/step - loss: 0.0566 - accuracy: 0.9923 - val_loss: 0.3214 - val_accuracy: 0.8788
Epoch 12/20
9/9 [==============================] - 0s 15ms/step - loss: 0.0318 - accuracy: 0.9962 - val_loss: 0.3067 - val_accuracy: 0.8636
Epoch 13/20
9/9 [==============================] - 0s 14ms/step - loss: 0.0260 - accuracy: 0.9962 - val_loss: 0.3035 - val_accuracy: 0.8636
Epoch 14/20
9/9 [==============================] - 0s 15ms/step - loss: 0.0197 - accuracy: 0.9962 - val_loss: 0.2827 - val_accuracy: 0.8636
Epoch 15/20
9/9 [==============================] - 0s 13ms/step - loss: 0.0133 - accuracy: 0.9962 - val_loss: 0.2799 - val_accuracy: 0.8788
Epoch 16/20
9/9 [==============================] - 0s 15ms/step - loss: 0.0091 - accuracy: 1.0000 - val_loss: 0.2718 - val_accuracy: 0.9091
Epoch 17/20
9/9 [==============================] - 0s 15ms/step - loss: 0.0068 - accuracy: 1.0000 - val_loss: 0.2876 - val_accuracy: 0.8939
Epoch 18/20
9/9 [==============================] - 0s 17ms/step - loss: 0.0056 - accuracy: 1.0000 - val_loss: 0.2794 - val_accuracy: 0.9091
Epoch 19/20
9/9 [==============================] - 0s 15ms/step - loss: 0.0039 - accuracy: 1.0000 - val_loss: 0.2878 - val_accuracy: 0.9091
Epoch 20/20
9/9 [==============================] - 0s 15ms/step - loss: 0.0033 - accuracy: 1.0000 - val_loss: 0.2962 - val_accuracy: 0.9091
Модель збережено успішно.
```

Рисунок 3.2 – Процес навчання нейромережевої моделі GRU на 20-и епохах

Як видно з рисунку 3.2, після 16-ї епохи нейромережа перестала покращувати свій результат, тому для використання у підсистемі контент-аналізу коментарів буде використано нейромережу що навчалась на 15-и епохах, однак параметри і їх вплив на результат навчання ще буде досліджено окремо.

Далі будуть наведені фрагменти опису реалізації підсистеми роботи з темами для обговорення, та підсистема контент-аналізу коментарів. Підсистема контент-аналізу коментарів використовується підсистемою роботи з темами для обговорення. Підсистема контент-аналізу коментарів реалізована сторінкою «topics», основна функція обробляє GET і POST запити до URL «/topics». Якщо це POST-запит, тобто користувач надіслав коментар, код виконує кілька кроків для аналізу та зберігання коментаря.

Перший крок включає отримання коментаря користувача та текст теми обговорення. Потім коментар попередньо обробляється для аналізу сентименту, і модель передбачає сентимент коментаря, класифікуючи його як позитивний або

негативний. Коментар очищається від зайвих пробілів та підраховується кількість слів. Далі, за допомогою сервісу LanguageTool, визначається кількість граматичних помилок у коментарі, після чого обчислюється відсоток помилок від загальної кількості слів. Також обчислюється індекс Жаккара для оцінки семантичної відповідності коментаря до ключових слів теми.

Отримані результати, включаючи текст коментаря, сентимент, оцінку, кількість слів, кількість помилок, відсоток помилок, індекс Жаккара і ім'я користувача, додаються до списку коментарів посту. Після цього відбувається перенаправлення на сторінку обговорення, де новий коментар відображається разом з іншими. Якщо виконуваний запит GET, сторінка обговорення просто відображається з поточними постами та коментарями.

Приклад залишення позитивного коментаря та його контент-аналізу наведено на рисунку 3.3.

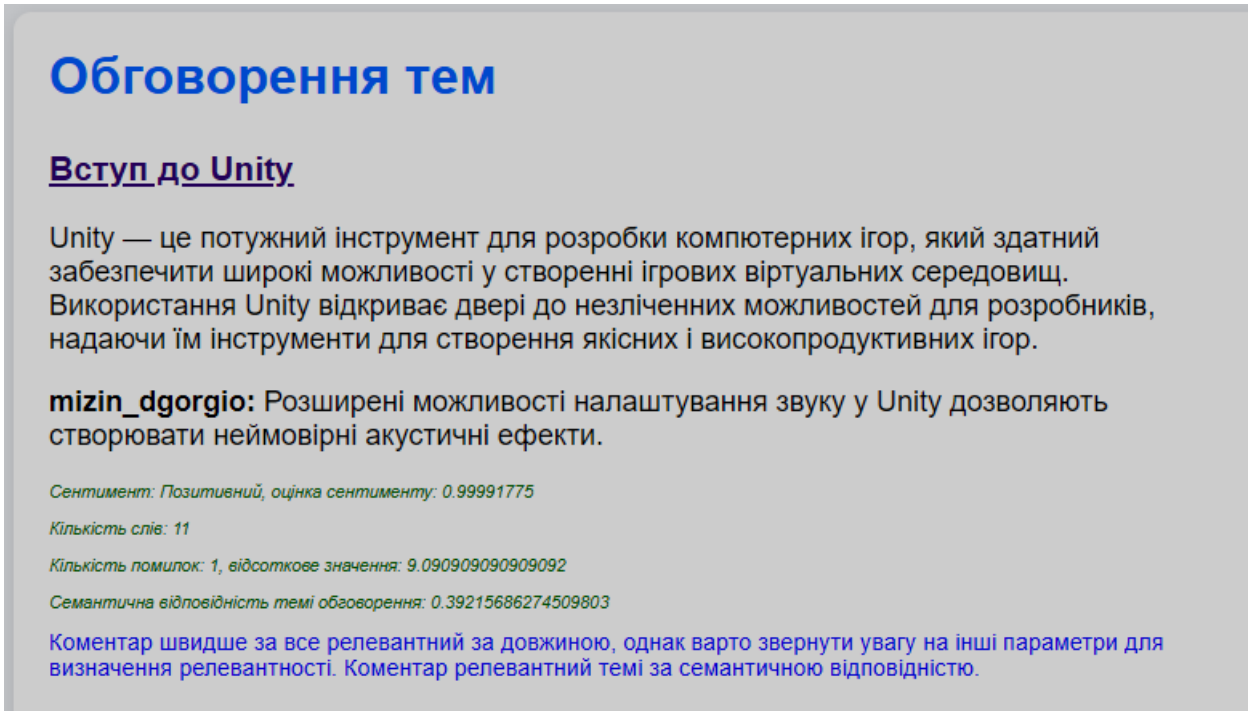


Рисунок 3.3 – Приклад роботи підсистем контент-аналізу коментарів та роботи з темами для обговорення

Як видно з рисунку 3.3, виводиться не тільки статистика показників по коментарю, а і відповідь щодо його релевантності.

Отже, наведено особливості реалізації програмних складових інформаційної системи інтелектуального аналізу коментарів для IT блогу розробників платформи Unity.

3.5 Тестування інформаційної системи інтелектуального аналізу коментарів для IT блогу розробників платформи Unity

Тестування інформаційної системи інтелектуального аналізу коментарів для IT блогу розробників платформи Unity необхідне для забезпечення якості та надійності аналітичних функцій, які допомагають користувачам ефективно взаємодіяти з контентом блогу. Тестування буде проведено з використанням засобів тест-кейсів.

Першим тестовим випадком буде перевірка коректності авторизації користувача. Кроки тестового випадку наведені у таблиці 3.1.

Таблиця 3.1 – Тест-кейс 00001

Тест-кейс ID: 00001	Пріоритет: 1	Створено:20.05.2024
Назва: Тест-кейс для перевірки коректності авторизації користувача		
Кроки		Очікуваний результат
1. Відкрити головну сторінку сайту. 2. В полі для уведення логіна та пароля ввести значення «mizin_dgorgio» та «1236» відповідно. 3. Натиснути кнопку «Увійти». 4. Перевірити наявність напису у верхній правій частині екрану «Ви авторизовані як: mizin_dgorgio» та наявність кнопки для можливості виходу з системи.		Відкрито головну сторінку Уведено дані Напис і кнопка наявні
Результат виконання тест-кейсу: пройдено успішно		

Скрін з прикладом виконання тест-кейсу 00001 наведено на рисунку 3.4.

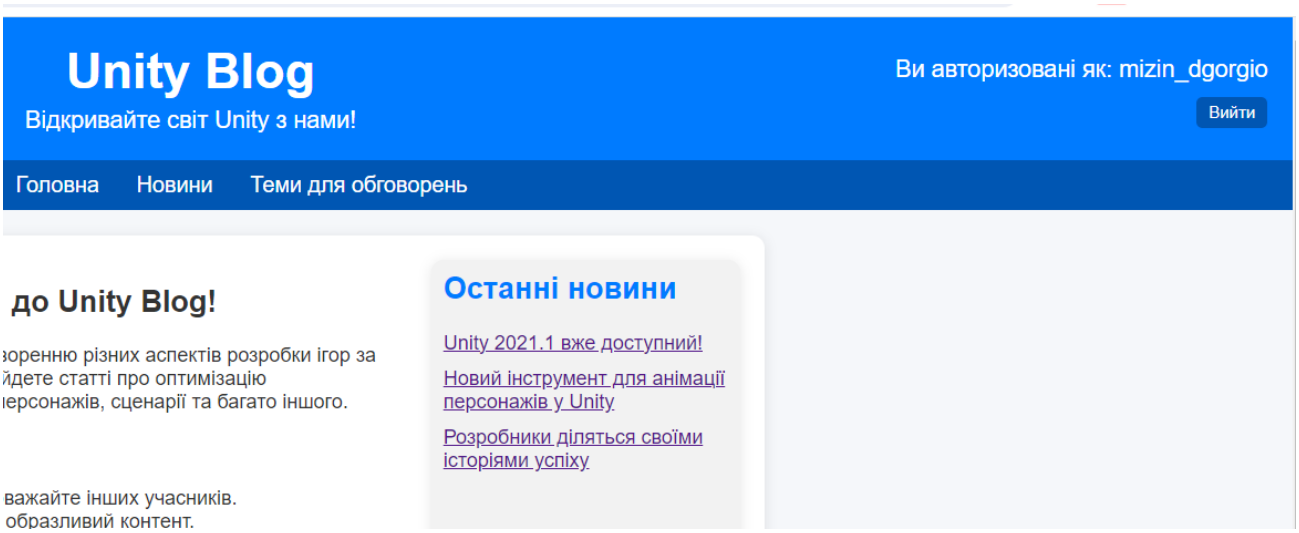


Рисунок 3.4 – Успішне виконання тест-кейсу 00001

Таблиця 3.2 – Тест-кейс 00002

Тест-кейс ID: 00002	Пріоритет: 1	Створено: 21.05.2024
Назва: Тест-кейс для перевірки додавання нового коментаря та його контент-аналіз		
Кроки		Очікуваний результат
1. Відкрити головну сторінку сайту. 2. Авторизуватись. 3. Перейти на сторінку «Теми для обговорень». 4. Ввести негативний коментар до теми «Вступ до Unity». 5. Натиснути на кнопку «Відправити». 6. Перевірити наявність коментаря та його контент-аналізу		Відкрито головну сторінку Авторизовано користувача mizin_dgorgio Відкрито сторінку «Теми для обговорень» Коментар відображено на сторінці «Теми для обговорень» Відображено виконаний контент-аналіз коментаря
Результат виконання тест-кейсу: пройдено успішно		

Наступним тестовим випадком буде перевірка додавання нового коментаря та його контент-аналіз.

Результат успішного виконання Тест-кейс 00002 наведено на рисунку 3.5.

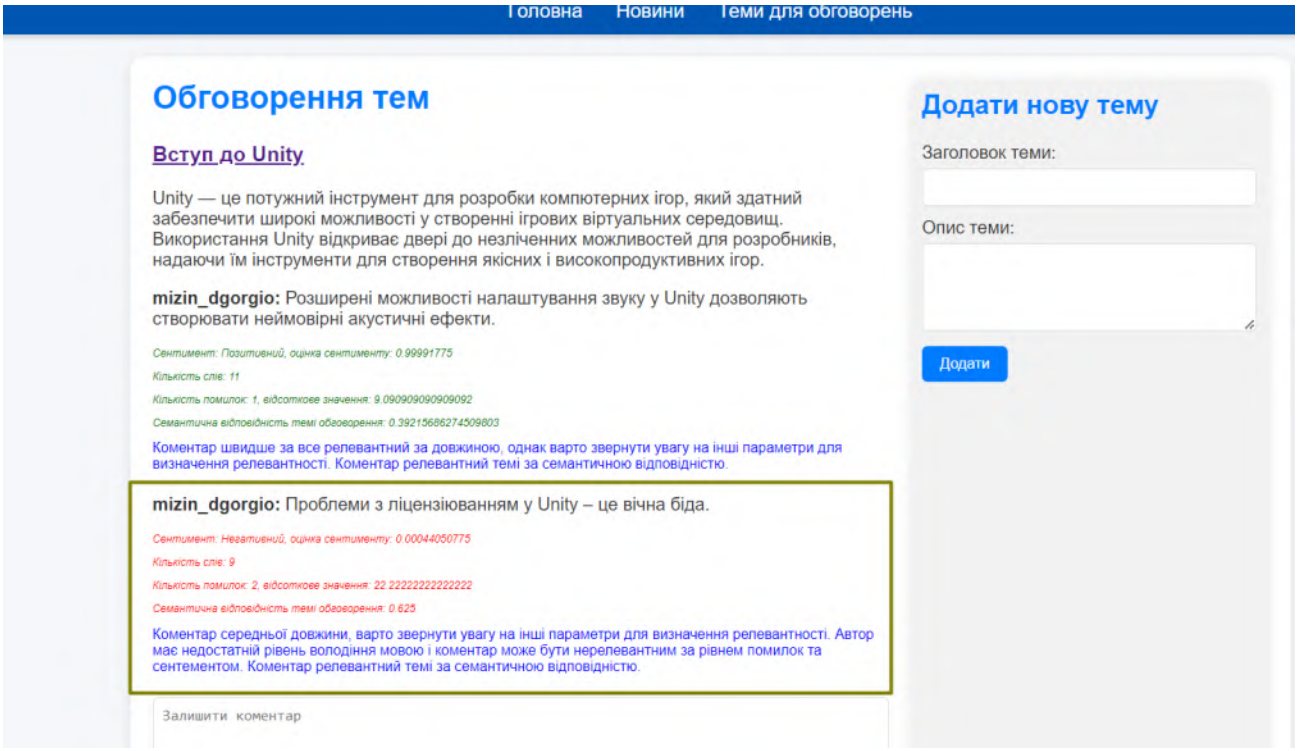


Рисунок 3.5 – Приклад додавання негативного коментаря та його контент-аналіз

Отже, проведене тестування некоректних функцій не виявило. Весь функціонал інформаційної системи інтелектуального аналізу коментарів для IT блогу розробників платформи Unity працює згідно заявлених функцій, та сприяє покращенню користувацького досвіду, підвищує релевантність обговорень та забезпечує надання більш якісного контенту, що, у свою чергу, підтримує активну та конструктивну взаємодію всередині спільноти розробників Unity.

3.6 Аналіз функціональності інформаційної системи інтелектуального аналізу коментарів для IT блогу

Для досконалого використання розробленої інформаційної системи інтелектуального аналізу коментарів для IT блогу розробників платформи Unity є потреба провести аналіз функціональності системи інтелектуального аналізу

даних. При запуску користувач побачить головну сторінку, з якої нікуди не зможе перейти, поки не виконає авторизацію (рисунок 3.6).

Рисунок 3.6 – Перший запуск IT блогу розробників платформи Unity

Для авторизації необхідно увести логін та пароль у відповідні текстові поля, після чого натиснути кнопку «Увійти» (рисунок 3.7).

Рисунок 3.7 – Приклад авторизації

Після виконаної авторизації на головній додасться права боковина із блоком останніх новин, з якої можна перейти на детанізацію новини, яка є цікавою для користувача (рисунок 3.8).

Рисунок 3.8 – Фрагмент деталізації обраної новини

При переході на сторінку «Новини» користувач буде бачити перелік новин (рисунок 3.9), та при бажанні натиснувши кнопку «Читати повністю» також зможе перейти на деталізацію новини, як на рисунку 3.8.

Рисунок 3.9 – Сторінка «Новини»

При натисненні на пункт меню «Теми для обговорень» користувач перейде на сторінку з темами для обговорень, до яких можна залишати коментарі (рисунок 3.10).

Для додавання нової теми для обговорень необхідно в правій боковині увести заголовок теми та її опис, після чого натиснути кнопку «Додати», фрагмент для додавання нової теми наведено на рисунку 3.11.

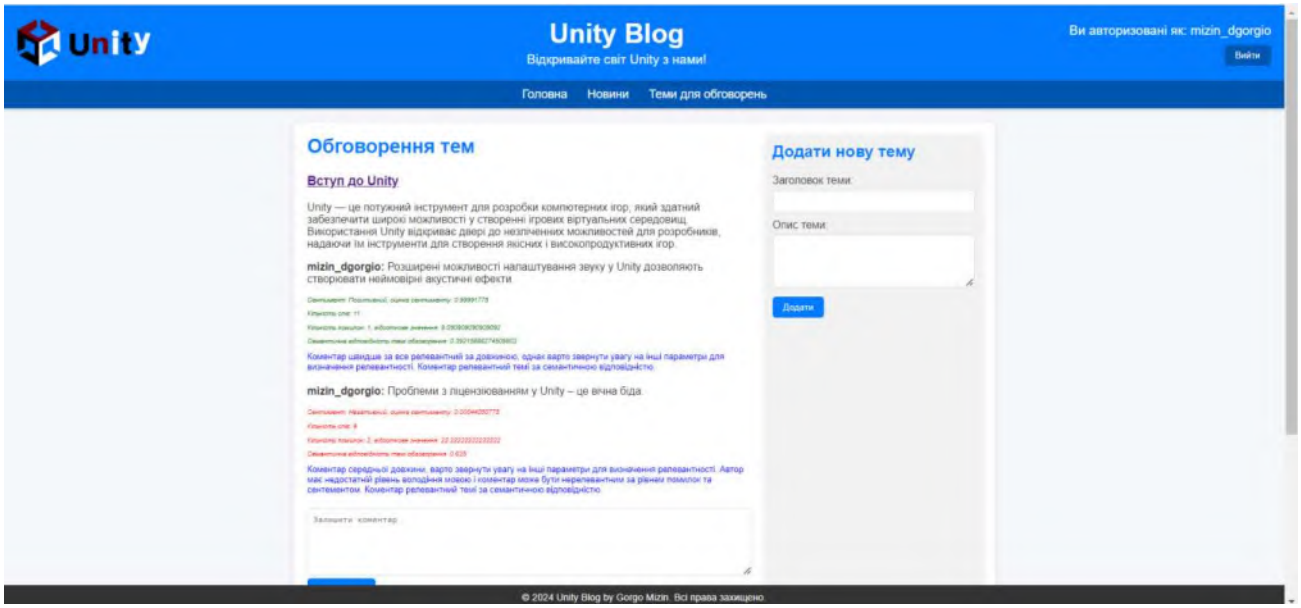


Рисунок 3.10 – Сторінка «Теми для обговорень»

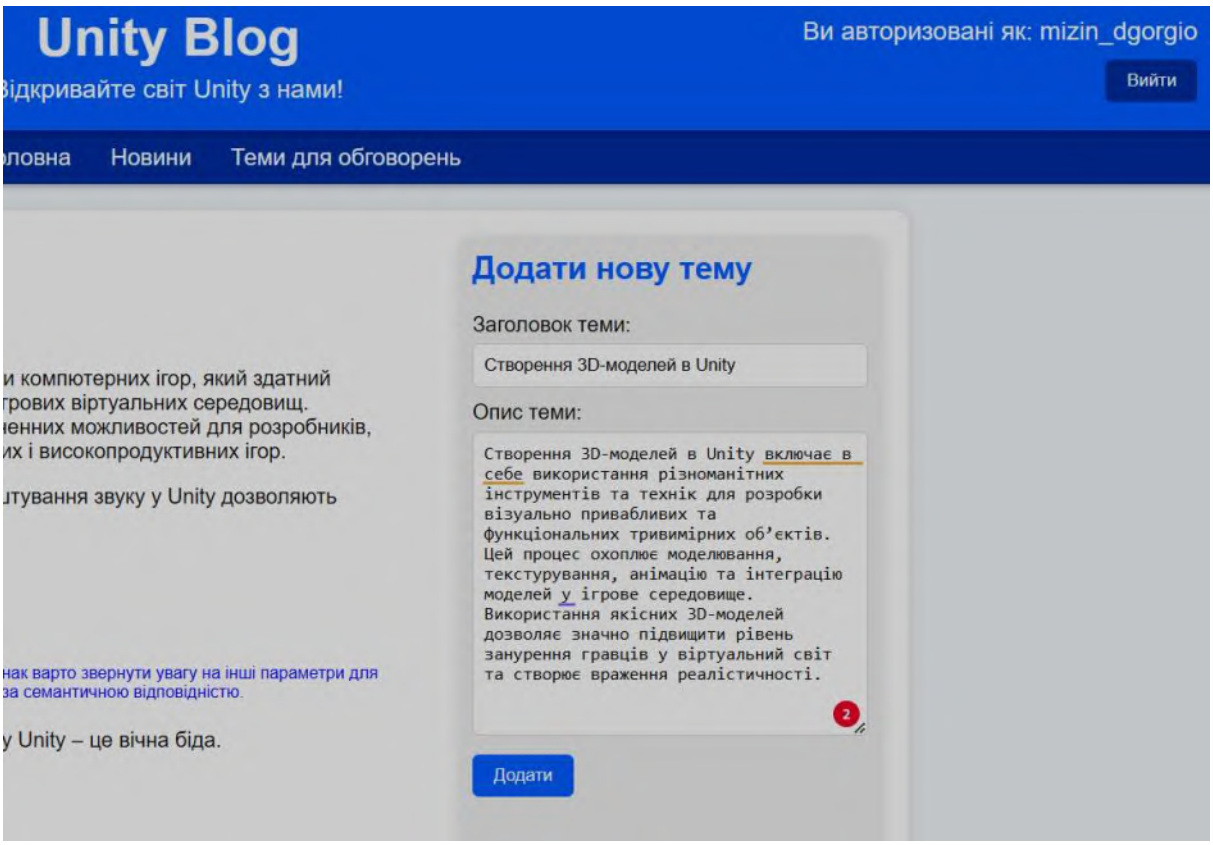


Рисунок 3.11 – Додавання теми для обговорення

Після натискання кнопки додати тему буде додано в БД та відображено на сторінці «Теми для обговорення» (рисунок 3.12).

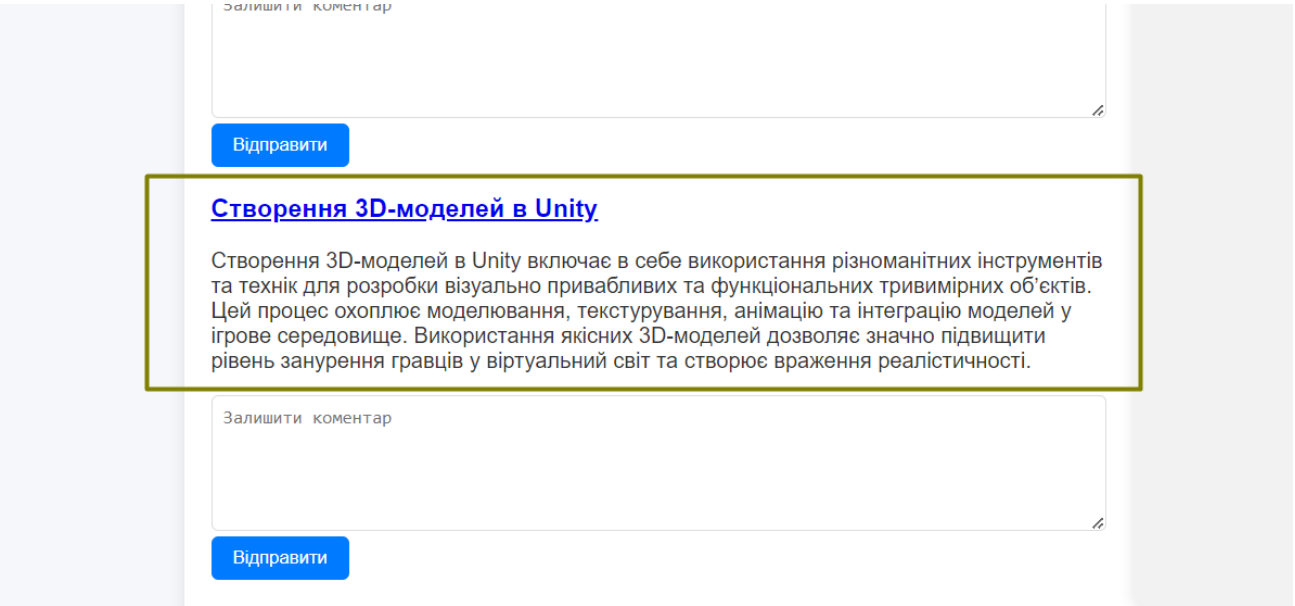


Рисунок 3.12 – Відображення доданої теми на екрані користувача

Для залишення коментаря та проведенні його контент-аналізу необхідно написати коментар у полі «Залишити коментар», та натиснути на кнопку «Відправити». Результат наведено на рисунку 3.13.

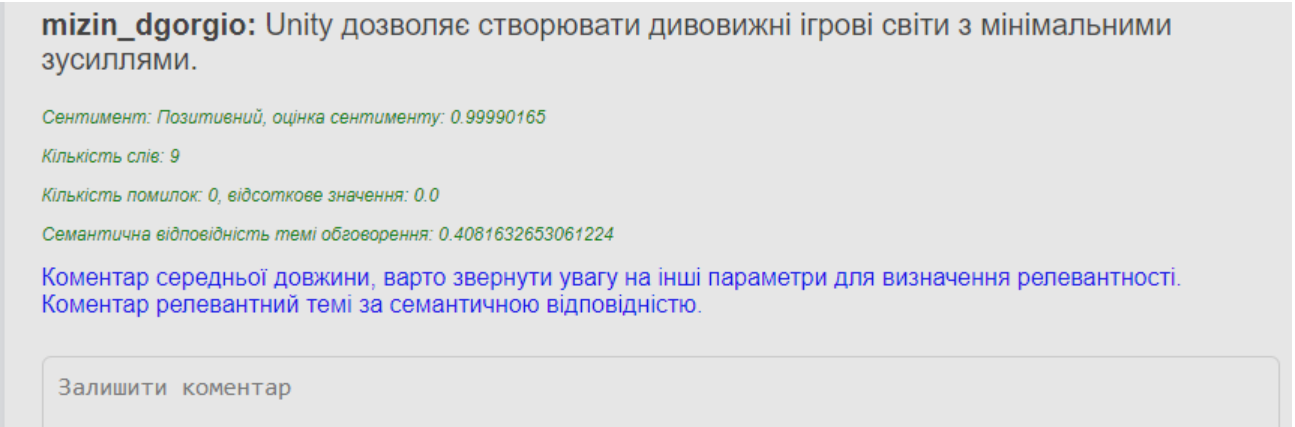


Рисунок 3.13 – Коментар та його контент-аналіз

Отже, проведено аналіз функціональності інформаційної системи інтелектуального аналізу коментарів для IT блогу розробників платформи Unity, який показав шляхи ефективного використання розробленого програмного комплексу.

3.7 Результати досліджень

Дослідження ефективності буде проведено спершу для ключових складових інформаційної системи у вигляді IT блогу. Першим буде проведено дослідження ефективності нейромережі GRU, що відповідає за визначення сентименту користувацького коментаря. Данні з експерименту впливу параметрів нейромережі на якість навчання наведено в таблиці 3.3

Таблиця 3.3 – Дослідження впливу параметрів нейромережі на якість навчання

Параметри	Accuracy	Loss
gru_units 64, к-сть епох 30, embedding_dim 30	0.9205	0.460
gru_units 128, к-сть епох 30, embedding_dim 40	0.924	0.3671
gru_units 256, к-сть епох 20, embedding_dim 40	0.8939	0.66
gru_units 64, к-сть епох 15, embedding_dim 50	0.9245	0.29

Графік навчання моделі з найкращим показником наведено на рисунку 3.14.

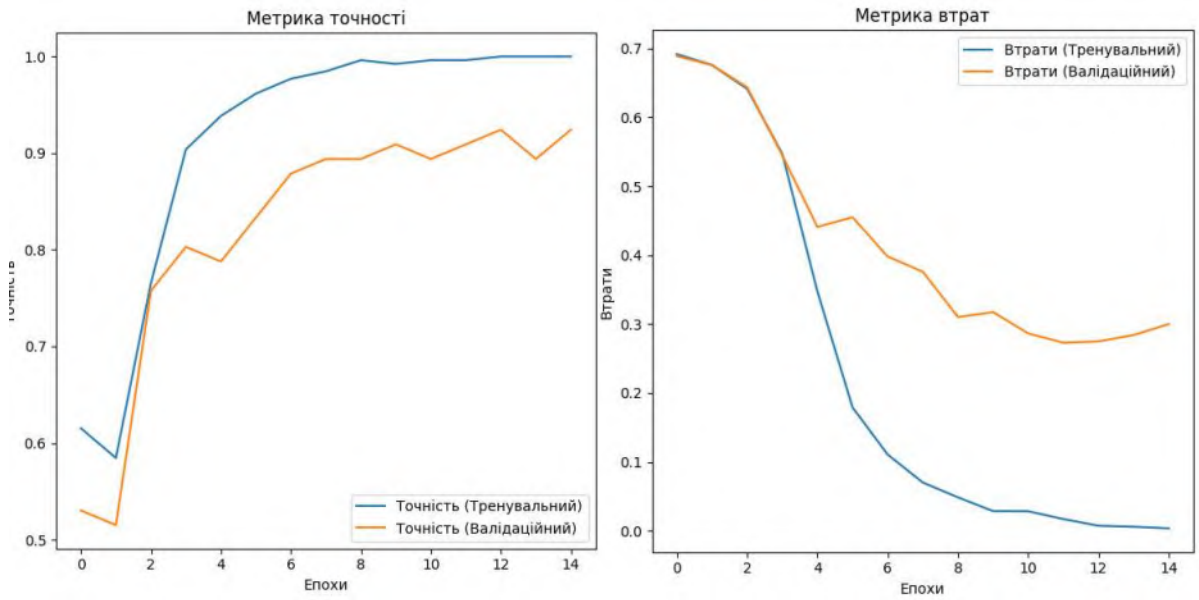


Рисунок 3.14 – Графік функції втрат та точності при параметрах gru_units 64, кількості епох 15, embedding_dim 50

Як видно з таблиці 3.3 та рисунку 3.14 – найкращі результати досягаються при розмірі шару gru у 64 нейрони, кількості епох 15, а розмір вектору збудувань – 50.

Аналізуючи дані з таблиці, при кількості епох понад 15 нейромережа перенавчається. На рисунку 3.15 наведено результати при параметрах gru_units 64, кількості епох 30, embedding_dim 30.

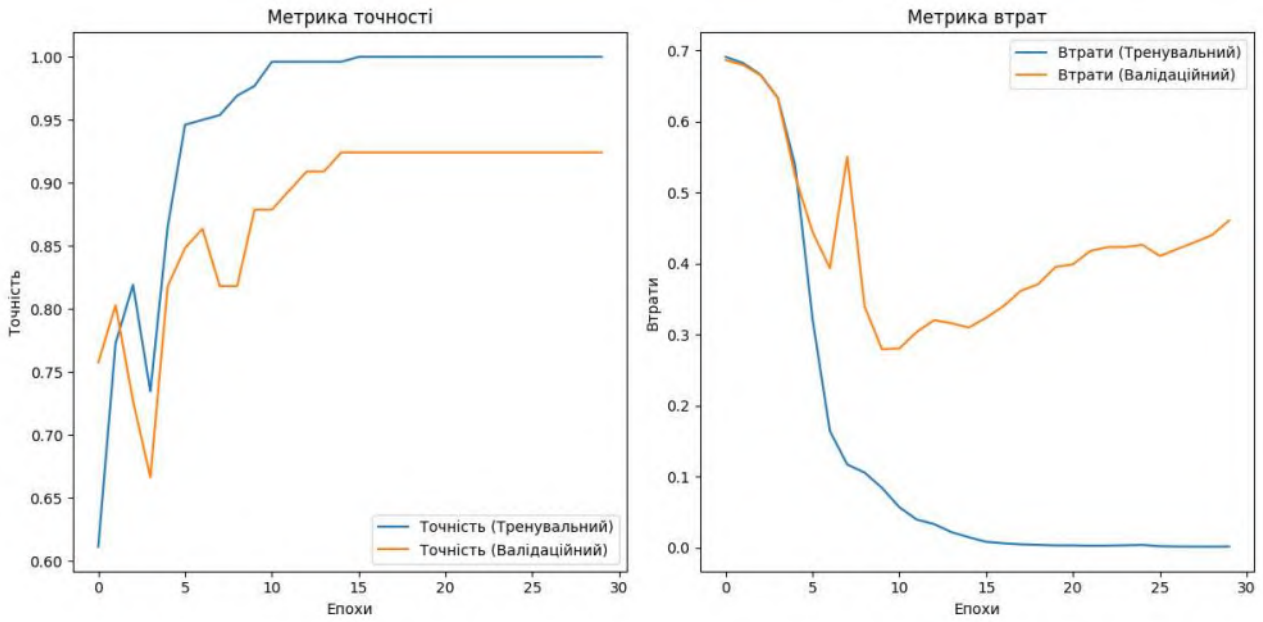


Рисунок 3.15 – Графік функції втрат та точності при параметрах gru_units 64, кількості епох 30, embedding_dim 30

Як видно з графіку на рисунку 3.15, після 14-ї епохи нейромережа не змогла покращити свій результат, а функція втрат взагалі стала зростати. Отже, зважаючи на специфіку роботи з українською мовою та коротким текстовим представленням коментарів, результат для метрики очність 0.9245 є високим та спроможним виконувати якісний аналіз сентименту відгуків.

Наступним буде досліджено ефективність методу контент-аналізу коментарів засобами інтелектуального аналізу даних. Для цього експерту буде подано 2 теми до кожної з яких подано по 5 коментарів, необхідно визначити:

– оцінку семантичної відповідність темі обговорення по шкалі від 0 до 1, де 1 – повністю відповідає, 0 – повністю не відповідає темі обговорення;

– оцінку сентименту коментаря за шкалою від 0 до 1, де 1 – повністю позитивні оцінка, 0 – негативний коментар.

Результати експерименту для теми 1 наведено в таблиці 3.4.

Таблиця 3.4 – Аналіз відповідності коментарів до теми 1, «Поради щодо оптимізації»

	Оцінка сентименту		**Оцінка семантичної відповідність темі обговорення**	
Коментарі	Експерт	Метод	Експерт	Метод
Оптимізація продуктивності в Unity – це великий крок вперед.	0.8	0.78	0.9	0.68
Можливості оптимізації проектів в Unity безперечні.	0.8	0.96	0.9	0.74
Можна працювати з різними мовами програмування.	0.6	0.55	0.7	0.2
Не можу знайти реальні шляхи для оптимізації ігор у Unity.	0.5	0.31	0.9	0.88
Проблеми з оптимізацією у Unity – це стало нормою.	0.4	0.2	0.8	0.69

Діаграма порівняння проведеної експертизи для оцінки сентименту наведено на рисунку 3.16.

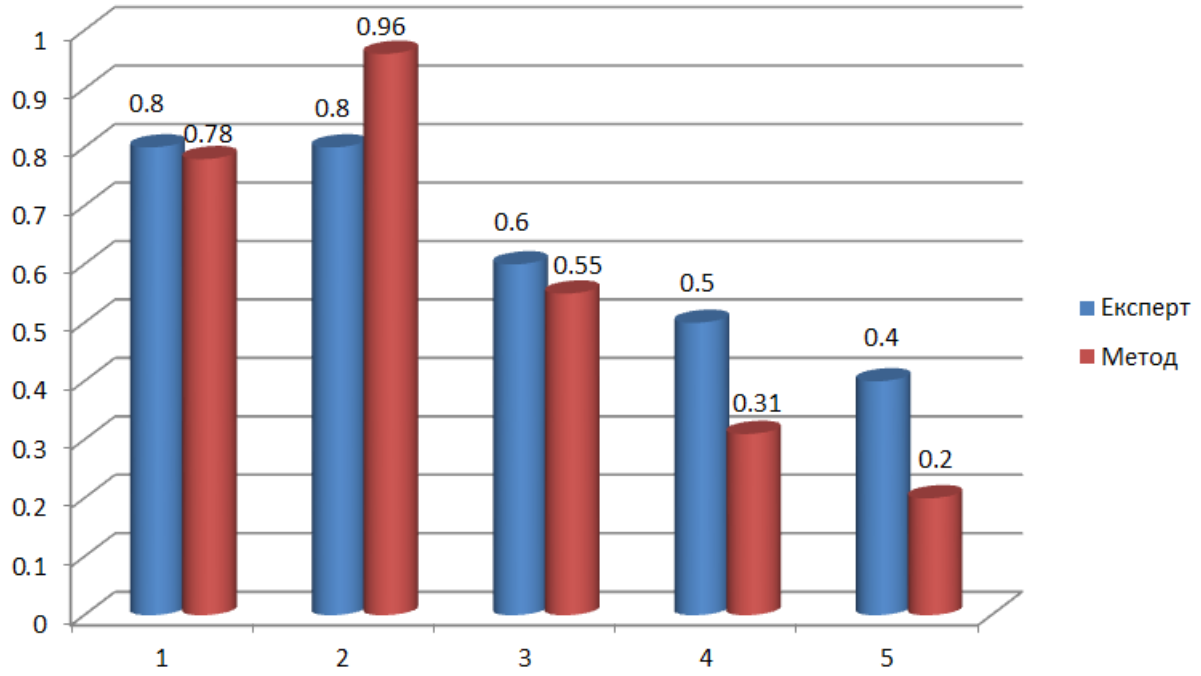


Рисунок 3.16 – Діаграма порівняння відповідей експерта та методу

Діаграма порівняння проведеної експертизи для оцінки семантичної відповідність темі обговорення наведено на рисунку 3.17.

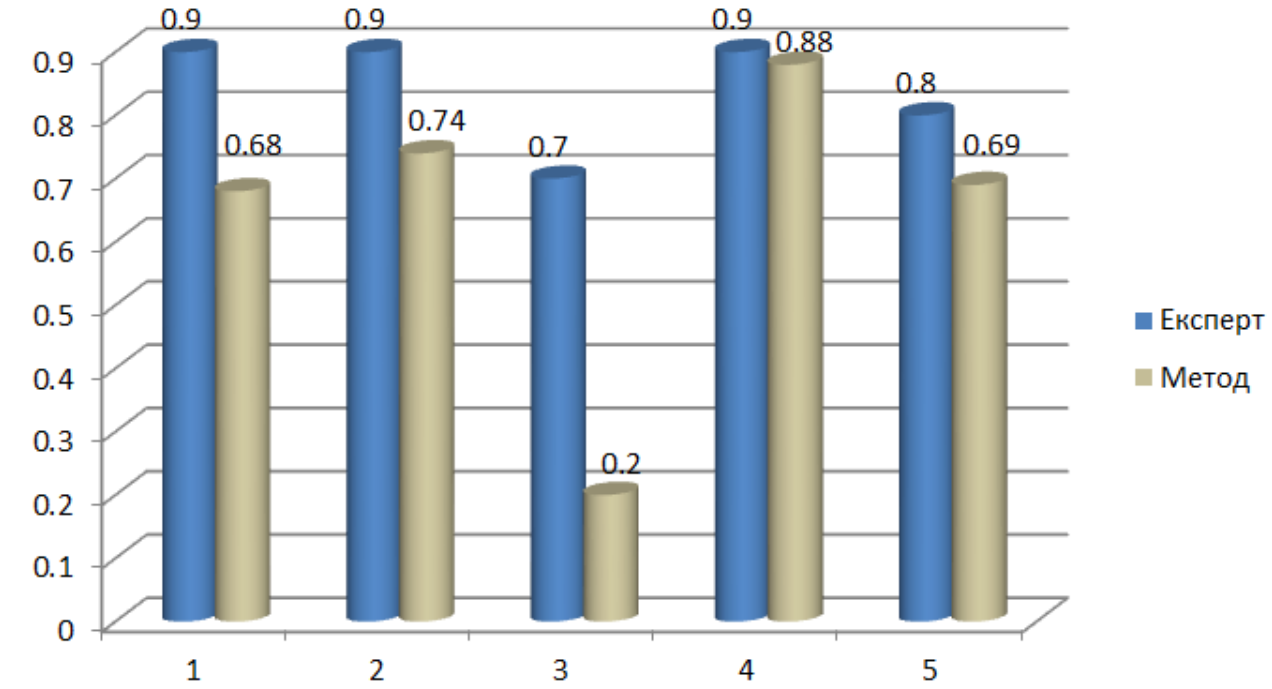


Рисунок 3.17 – Діаграма порівняння відповідей експерта та методу для оцінки семантичної відповідність темі обговорення

Як видно з діаграм рисунків 3.16 та 3.17 – значних відхилень роботи методу від експертної оцінки немає. Однак, як видно на графіку 3.17, для коментаря «Можна працювати з різними мовами програмування.» оцінка методу щодо семантичної відповідність темі обговорення є значно нижчою. Проте даний коментар є неоднозначним, і може стосуватись не тільки платформи Unity. Тому подавши даний коментар моделі GPT 3.5, було отримано таку відповідь: «*Оцінюючи за шкалою від 0 до 1, де 0 означає повну невідповідність, а 1 - повну відповідність, я б оцінив цей коментар на 0.1. Коментар віддалено стосується програмування, але не вносить жодного внеску до конкретної теми обговорення щодо оптимізації продуктивності в Unity*». Тобто розроблений метод дав ближчий до GPT 3.5 результат ніж оцінка експерта.

Отже, досліджено ефективність розробленого методу контент-аналізу коментарів засобами інтелектуального аналізу даних, що дозволяє за проведеним аналізом коментарів визначати їх релевантність. З проведеного дослідження видно, що розроблений метод, реалізований на базі вебсистеми повністю виконує поставлені завдання. Вдалось досягти точності 0.9245 для оцінки сентименту коментаря, а також в порівнянні з оцінками експерта метод показав лише незначні відхилення від думки експерта, що свідчить про його спроможність до проведення контент-аналізу коментарів.

3.8 Висновки до розділу 3

Визначено шляхи дослідження та засоби створення програмного забезпечення, для розробки інформаційної системи інтелектуального аналізу коментарів для IT блогу розробників платформи Unity було обрано мову програмування Python, фреймворк Flask для створення веб-інтерфейсу, середовище розробки PyCharm, сервіс Google Colab для тренування нейромережі, мова запитів SQL для взаємодії з базою даних, СКБД MySQL, яка відповідає за зберігання та керування даними.

Наведено структуру та описано функціональне призначення програмних складових інформаційної системи інтелектуального аналізу коментарів для IT блогу розробників платформи Unity. Структура майбутнього вебзастосунку представлена у вигляді діаграми класів, кожен з яких описано згідно його функціонального призначення.

Описано особливості реалізації програмних складових інформаційної системи інтелектуального аналізу коментарів для IT блогу розробників платформи Unity. Виконано тестування, під час якого некоректних функцій не виявлено. Весь функціонал працює згідно заявлених функцій, та сприяє покращенню користувацького досвіду, підвищує релевантність обговорень та забезпечує надання більш якісного контенту, що, у свою чергу, підтримує активну та конструктивну взаємодію всередині спільноти розробників Unity

Проведено аналіз функціональності інформаційної системи інтелектуального аналізу коментарів для IT блогу розробників платформи Unity, який показав шляхи ефективного використання розробленого програмного комплексу.

Досліджено ефективність розробленого методу контент-аналізу коментарів засобами інтелектуального аналізу даних, що дозволяє за проведеним аналізом коментарів визначати їх релевантність. З проведеного дослідження видно, що розроблений метод, реалізований на базі вебсистеми повністю виконує поставлені завдання. Вдалось досягти точності 0.9245 для оцінки сентименту коментаря, а також в порівнянні з оцінками експерта метод показав лише незначні відхилення від думки експерта, що свідчить про його спроможність до підвищення якості контент-аналізу коментарів IT блогу розробників платформи Unity засобами інтелектуального аналізу даних.

Загальні висновки

Метою кваліфікаційної роботи бакалавра було підвищення якості контент-аналізу коментарів IT блогу розробників платформи Unity засобами інтелектуального аналізу даних.

Для досягнення мети, були поставлені та вирішені такі задачі:

– виконано аналіз інформаційних моделей в області контент-аналізу;

– розглянуто засоби інтелектуального аналізу даних області контент-аналізу коротких текстових даних, та обрано підхід для реалізації;

– виконано аналіз існуючих програмних засобів та наукових рішень;

– створено метод контент-аналізу коментарів IT блогу розробників платформи Unity засобами інтелектуального аналізу даних;

– виконано розробку архітектури нейромережі для визначення сентименту коментаря;

– створено проектну архітектуру інформаційної системи інтелектуального аналізу коментарів для IT блогу розробників платформи Unity;

– виконано проєктування бази даних інформаційної системи інтелектуального аналізу коментарів для IT блогу розробників платформи Unity;

– виконано вибір та підготовку робочих вхідних даних методу контент-аналізу коментарів IT блогу;

– розглянуто особливості використання спеціалізованих програмних компонентів для спрощення програмної розробки;

– виконано вибір засобів програмної реалізації методу контент-аналізу коментарів IT блогу розробників платформи Unity засобами інтелектуального аналізу даних;

– виконано програмну реалізацію створеного методу;

– виконано тестування створеного IT блогу розробників платформи Unity та застосунку для тренування нейромереж, що показало відсутність некоректно працюючих функцій;

– виконано дослідження ефективності створеного методу з використанням розробленого ПЗ.

Досліджена ефективність розробленого методу контент-аналізу коментарів засобами інтелектуального аналізу даних, що дозволяє за проведеним аналізом коментарів визначати їх релевантність. З проведеного дослідження видно, що розроблений метод, реалізований на базі вебсистеми ІТ блогу розробників платформи Unity повністю виконує поставлені завдання. Вдалось досягти точності 0.9245 для оцінки сентименту коментаря, а також в порівнянні з оцінками експерта метод показав лише незначні відхилення від думки експерта, що свідчить про його спроможність до підвищення якості контент-аналізу коментарів ІТ блогу розробників платформи Unity засобами інтелектуального аналізу даних.

Актуальним напрямком для подальшої роботи в даному напрямку є розширення переліку факторів, що впливають на виконання контент-аналізу, з метою покращення якості контент-аналізу коментарів блогу засобами інтелектуального аналізу даних.

Перелік посилань

1. AIContentfy. Content analysis is becoming more and more relevant. URL: https://aicontentfy.com/en/blog/future-of-content-analysis-with-ai

2. Questionpro. Content Analysis: What is it in Qualitative Studies? URL: https://www.questionpro.com/blog/content-analysis/

3. Delve. Is Content Analysis Qualitative or Quantitative? URL: https://delvetool.com/blog/is-content-analysis-qualitative-or-quantitative

4. Delvetool. The Practical Guide to Qualitative Content Analysis. URL: https://delvetool.com/blog/guide-qualitative-content-analysis

5. Llinkedin. What is lexical analysis in NLP? URL: https://www.linkedin.com/pulse/what-lexical-analysis-nlp-rahul-sharma/

6. Monkeylearn. Sentiment Analysis: A Definitive Guide. URL: https://monkeylearn.com/sentiment-analysis/

7. Lawinsider. Relevant Discussions definition. URL: https://www.lawinsider.com/dictionary/relevant-discussions

8. Smartcomment. Leveraging Natural Language Processing for Efficient Comment Analysis on SmartComment.com. URL: https://www.smartcomment.com/resources/leveraging-natural-language-processing-for-efficient-comment-analysis-on-smartcomment-com/

9. J. Kasperiuniene, M. Briediene, V. Zydziunaite. Automatic Content Analysis of Social Media Short Texts: Scoping Review of Methods and Tools. pp. 89-101. URL: https://www.researchgate.net/publication/335868877_Automatic_Content_Analysis_of_Social_Media_Short_Texts_Scoping_Review_of_Methods_and_Tools

10. Baeldung. Algorithms for Determining Text Sentiment. URL: https://www.baeldung.com/cs/sentiment-analysis-practical

11. Medium. Understanding Gated Recurrent Unit (GRU) in Deep Learning. URL: https://medium.com/@anishnama20/understanding-gated-recurrent-unit-gru-in-deep-learning-2e54923f3e2

12. Turing. A Guide on Word Embeddings in NLP. URL: https://www.turing.com/kb/guide-on-word-embeddings-in-nlp

13. Medium. Understanding Vector Similarity for Machine Learning. URL: https://medium.com/advanced-deep-learning/understanding-vector-similarity-b9c10f7506de

14. Qsrinternational. About NVivo. URL: https://help-nv.qsrinternational.com/14/win/Content/about-nvivo/about-nvivo.htm

15. Lumivero. NVivo 14 - Leading Qualitative Data Analysis Software with AI Solution. URL: https://lumivero.com/products/nvivo

16. IBM. Infuse your product with artificial intelligence from IBM https://dsce.ibm.com/wizard/try?page=embed-nlp

17. Cloud academy. Google Cloud Natural Language Processing API, First Steps. URL: https://cloudacademy.com/blog/google-cloud-natural-language-processing-api/

18. R. Haque, N. Islam, M. Tasneem, A. K. Das. Multi-class sentiment classification on Bengali social media comments using machine learning. International Journal of Cognitive Computing in Engineering.Volume 4 (2023), pp 21-35. URL: https://doi.org/10.1016/j.ijcce.2023.01.001.

19. T. Chalke, V. Dhumal and H. Kanakia. Sentiment Analysis of Social Media Comments. 2024 IEEE 9th International Conference for Convergence in Technology (I2CT). Pune. India (2024). pp. 1-5. URL: https://ieeexplore.ieee.org/abstract/document/10544131

20. J. H. Setiawan, C. Caroline, D. Muharman. Content analysis of readers'comments on media aggregator as feedback and form of public opinion about Covid-19. Aspiration Journal 2 (1). pp. 50–69. URL: https://doi.org/10.56353/aspiration.v2i1.22

21. kaggle.com. IMDB Dataset of 50K Movie Reviews. URL: https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews/data

22. kaggle.com. Welcome page. URL: https://www.kaggle.com/

23. python.org. Бібліотека Pickle. URL: https://docs.python.org/3/library/pickle.html

24. flask.palletsprojects.com. Бібліотека Flask. URL: https://flask.palletsprojects.com/

25. keras.io. Бібліотека Keras. URL: https://keras.io/

26. tensorflow.org. Бібліотека Keras. URL: https://www.tensorflow.org/guide/keras

27. nltk.org. Бібліотека NLTK. URL: https://www.nltk.org

28. stanfordnlp.github.io. Бібліотека Stanza. URL: https://stanfordnlp.github.io/stanza/

29. requests.readthedocs.io. Бібліотека Requests. URL: https://requests.readthedocs.io/

30. docs.python.org. Бібліотека Datetime. URL: https://docs.python.org/3/library/datetime.html

31. Genius. Де використовується Python і чому вам потрібно знати цю мову. URL: https://genius.space/lab/de-vikoristovuyetsya-python-i-chomu-vam-potribno-znati-tsyu-movu/

32. Dou. Безсерверні веб-застосунки на Python з використанням Lambda і Flask. URL: https://dou.ua/lenta/articles/serverless-python/

33. Foxminded. PyCharm як найкраща IDE для розробки ПЗ на Python. URL: https://foxminded.ua/pycharm-tse/

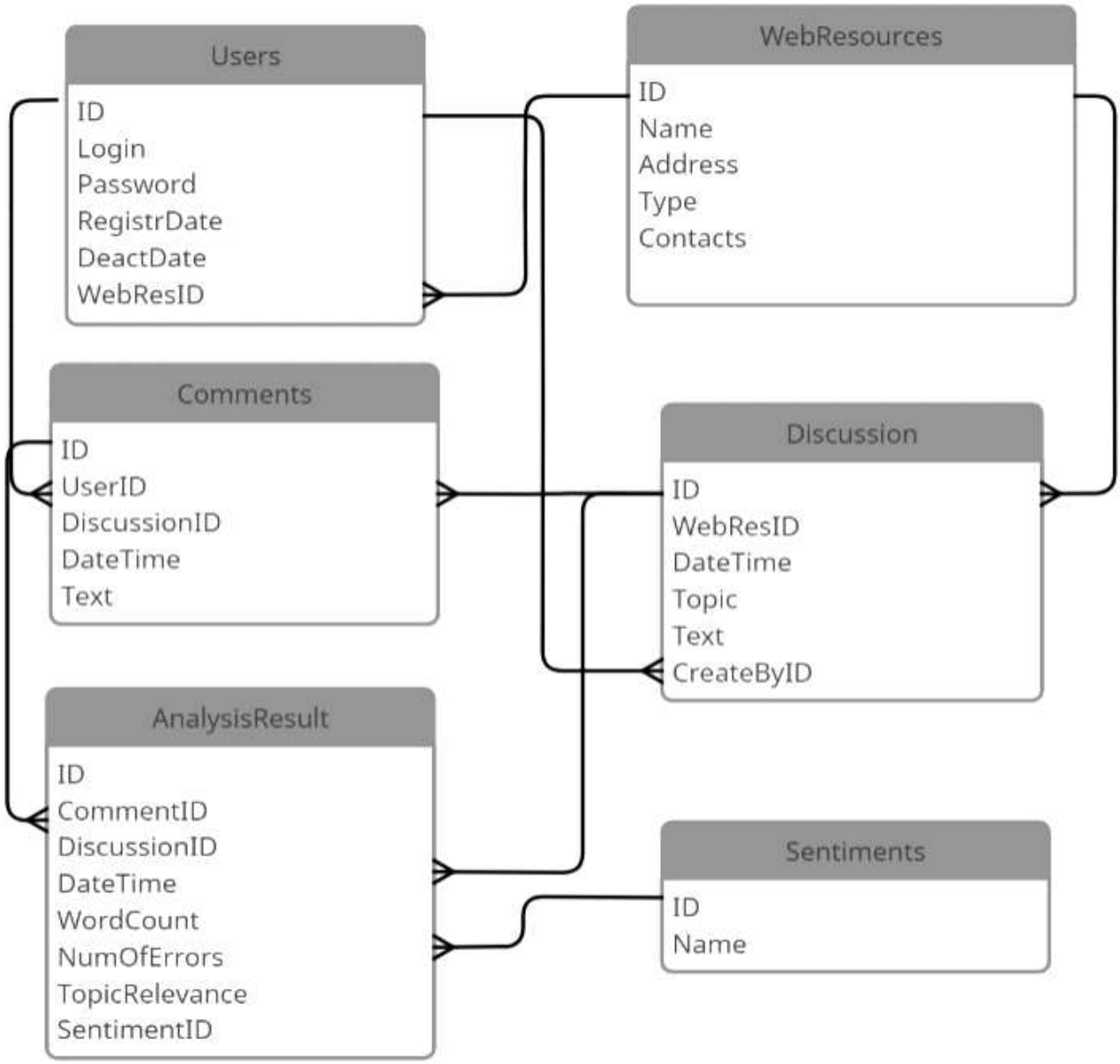
34. w3schools. SQL Tutorial. URL: https://www.w3schools.com/sql/

35. w3schools. MySQL Tutorial. URL: https://www.w3schools.com/MySQL/default.asp

ДОДАТКИ

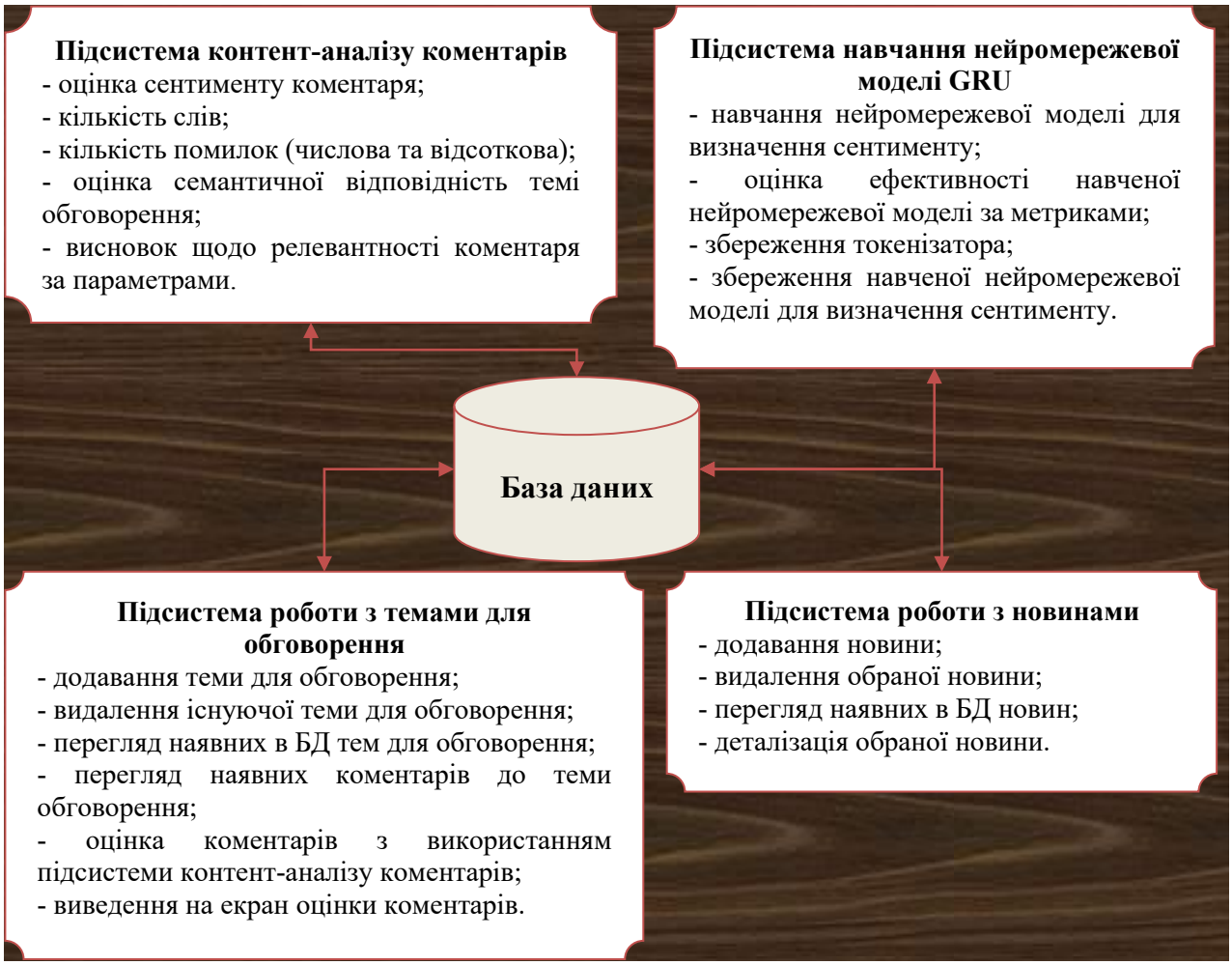
Додаток А

Структура бази даних інформаційної системи інтелектуального аналізу коментарів для IT блогу розробників платформи Unity



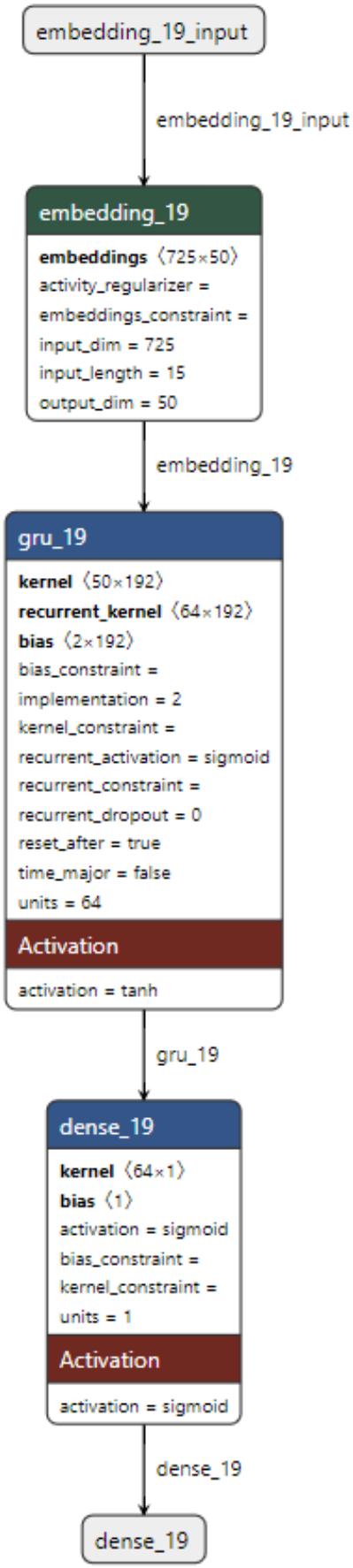
Додаток Б

Проектна архітектура та взаємозв'язок компонентів інформаційної системи інтелектуального аналізу коментарів для IT блогу



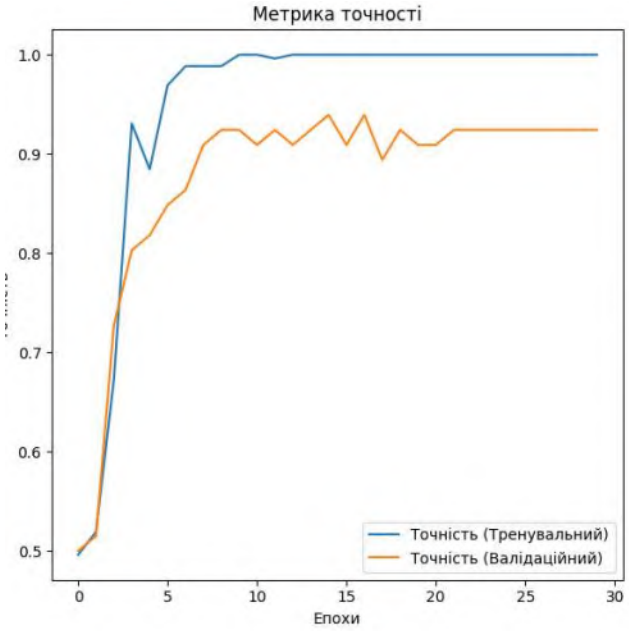
Додаток В

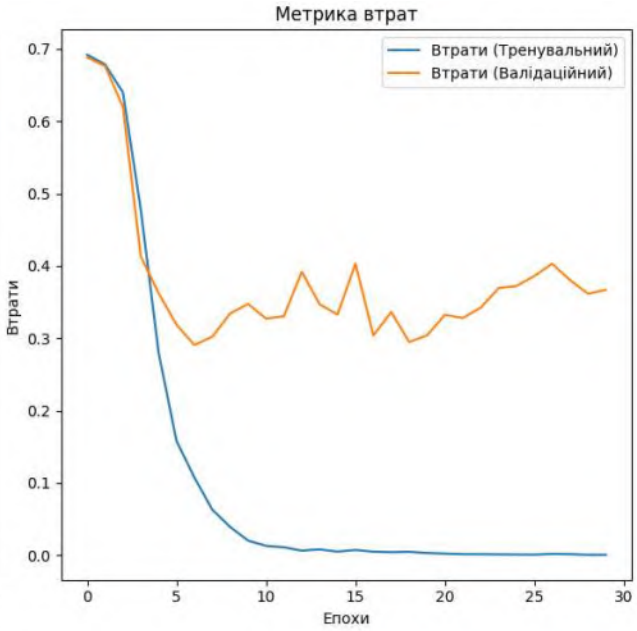
Розроблена архітектура нейромережі GRU



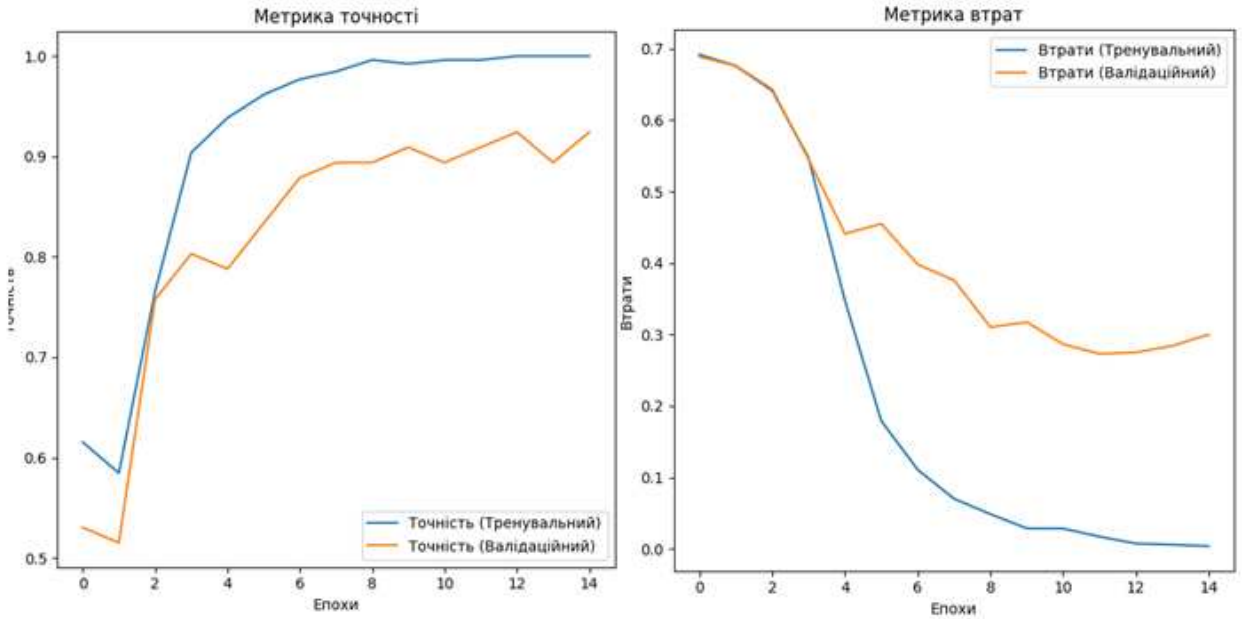
Додаток Г

Графіки експериментів та логи з навчання нейромереж GRU






```
9/9 [==============================] - 0s 39ms/step - loss: 0.0048 - accuracy: 1.0000 - val_loss: 0.4952 - val_accuracy: 0.9091
Epoch 14/30
9/9 [==============================] - 0s 34ms/step - loss: 0.0075 - accuracy: 0.9962 - val_loss: 0.4607 - val_accuracy: 0.9091
Epoch 15/30
9/9 [==============================] - 0s 37ms/step - loss: 0.0104 - accuracy: 0.9923 - val_loss: 0.4231 - val_accuracy: 0.8636
Epoch 16/30
9/9 [==============================] - 0s 36ms/step - loss: 0.0064 - accuracy: 1.0000 - val_loss: 0.3820 - val_accuracy: 0.8636
Epoch 17/30
9/9 [==============================] - 0s 34ms/step - loss: 0.0036 - accuracy: 1.0000 - val_loss: 0.3785 - val_accuracy: 0.8788
Epoch 18/30
9/9 [==============================] - 0s 34ms/step - loss: 0.0017 - accuracy: 1.0000 - val_loss: 0.3916 - val_accuracy: 0.8788
Epoch 19/30
9/9 [==============================] - 0s 47ms/step - loss: 0.0013 - accuracy: 1.0000 - val_loss: 0.4009 - val_accuracy: 0.9091
Epoch 20/30
9/9 [==============================] - 0s 34ms/step - loss: 9.9876e-04 - accuracy: 1.0000 - val_loss: 0.4121 - val_accuracy: 0.9091
Epoch 21/30
9/9 [==============================] - 0s 43ms/step - loss: 8.9508e-04 - accuracy: 1.0000 - val_loss: 0.4251 - val_accuracy: 0.9091
Epoch 22/30
9/9 [==============================] - 0s 40ms/step - loss: 7.3633e-04 - accuracy: 1.0000 - val_loss: 0.4434 - val_accuracy: 0.9242
Epoch 23/30
9/9 [==============================] - 0s 23ms/step - loss: 6.5033e-04 - accuracy: 1.0000 - val_loss: 0.4544 - val_accuracy: 0.9242
Epoch 24/30
9/9 [==============================] - 0s 21ms/step - loss: 5.9988e-04 - accuracy: 1.0000 - val_loss: 0.4749 - val_accuracy: 0.9242
Epoch 25/30
9/9 [==============================] - 0s 21ms/step - loss: 5.4718e-04 - accuracy: 1.0000 - val_loss: 0.4750 - val_accuracy: 0.9242
Epoch 26/30
9/9 [==============================] - 0s 26ms/step - loss: 4.9842e-04 - accuracy: 1.0000 - val_loss: 0.4783 - val_accuracy: 0.9242
Epoch 27/30
9/9 [==============================] - 0s 22ms/step - loss: 4.6067e-04 - accuracy: 1.0000 - val_loss: 0.4890 - val_accuracy: 0.9242
Epoch 28/30
9/9 [==============================] - 0s 20ms/step - loss: 4.2808e-04 - accuracy: 1.0000 - val_loss: 0.5044 - val_accuracy: 0.9242
Epoch 29/30
9/9 [==============================] - 0s 21ms/step - loss: 4.0508e-04 - accuracy: 1.0000 - val_loss: 0.5077 - val_accuracy: 0.9242
Epoch 30/30
9/9 [==============================] - 0s 22ms/step - loss: 3.7618e-04 - accuracy: 1.0000 - val_loss: 0.5201 - val_accuracy: 0.9242
Модель збережено успішно.
```




```
Epoch 1/15
9/9 [==============================] - 4s 84ms/step - loss: 0.6894 - accuracy: 0.6154 - val_loss: 0.6870 - val_accuracy: 0.5303
Epoch 2/15
9/9 [==============================] - 0s 27ms/step - loss: 0.6745 - accuracy: 0.5846 - val_loss: 0.6767 - val_accuracy: 0.5152
Epoch 3/15
9/9 [==============================] - 0s 24ms/step - loss: 0.6441 - accuracy: 0.7654 - val_loss: 0.6522 - val_accuracy: 0.7576
Epoch 4/15
9/9 [==============================] - 0s 24ms/step - loss: 0.5713 - accuracy: 0.9038 - val_loss: 0.5776 - val_accuracy: 0.8030
Epoch 5/15
9/9 [==============================] - 0s 25ms/step - loss: 0.3937 - accuracy: 0.9385 - val_loss: 0.4347 - val_accuracy: 0.7879
Epoch 6/15
9/9 [==============================] - 0s 24ms/step - loss: 0.1709 - accuracy: 0.9615 - val_loss: 0.4479 - val_accuracy: 0.8333
Epoch 7/15
9/9 [==============================] - 0s 26ms/step - loss: 0.1133 - accuracy: 0.9769 - val_loss: 0.3440 - val_accuracy: 0.8788
Epoch 8/15
9/9 [==============================] - 0s 22ms/step - loss: 0.0659 - accuracy: 0.9846 - val_loss: 0.3378 - val_accuracy: 0.8939
Epoch 9/15
9/9 [==============================] - 0s 24ms/step - loss: 0.0512 - accuracy: 0.9962 - val_loss: 0.3273 - val_accuracy: 0.8939
Epoch 10/15
9/9 [==============================] - 0s 24ms/step - loss: 0.0383 - accuracy: 0.9923 - val_loss: 0.3171 - val_accuracy: 0.9091
Epoch 11/15
9/9 [==============================] - 0s 23ms/step - loss: 0.0313 - accuracy: 0.9962 - val_loss: 0.3013 - val_accuracy: 0.8939
Epoch 12/15
9/9 [==============================] - 0s 24ms/step - loss: 0.0162 - accuracy: 0.9962 - val_loss: 0.3030 - val_accuracy: 0.9091
Epoch 13/15
9/9 [==============================] - 0s 22ms/step - loss: 0.0139 - accuracy: 1.0000 - val_loss: 0.3060 - val_accuracy: 0.9242
Epoch 14/15
9/9 [==============================] - 0s 25ms/step - loss: 0.0078 - accuracy: 1.0000 - val_loss: 0.3130 - val_accuracy: 0.8939
Epoch 15/15
9/9 [==============================] - 0s 25ms/step - loss: 0.0053 - accuracy: 1.0000 - val_loss: 0.3327 - val_accuracy: 0.9242
```

Додаток Д

Презентаційний матеріал

КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА

МЕТОД КОНТЕНТ-АНАЛІЗУ КОМЕНТАРІВ ЗАСОБАМИ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ ДЛЯ ІТ БЛОГУ РОЗРОБНИКІВ ПЛАТФОРМИ UNITY

Виконав:
студент групи КН-20-1
Джорджо МІЗИН

Керівник:
PhD, ст. викл. каф. КН
Павло РАДЮК

Актуальність

У сучасному світі блогові платформи є важливим джерелом інформації для розробників, які обмінюються досвідом, обговорюють нові технології та діляться своїми знаннями. У зв'язку з цим виникає необхідність автоматизованого аналізу великої кількості коментарів, що дозволяє визначити їх релевантність та корисність. Ручний аналіз коментарів займає багато часу та ресурсів, що робить його неефективним для великих платформ.

Використання інтелектуального аналізу даних для обробки коментарів в IT блогах надає можливість підвищити якість обговорень та сприяти створенню цінного контенту. Це особливо важливо для платформи Unity, яка є однією з провідних у сфері розробки ігор та інших інтерактивних додатків. Завдяки автоматизації процесу аналізу коментарів можна швидко виявляти нерелевантні або шкідливі відгуки, тим самим підтримуючи високий рівень дискусій та зберігаючи професійність блогу.

Інтелектуальний аналіз коментарів також сприяє покращенню взаємодії між користувачами платформи. Завдяки аналізу тональності коментарів можна краще розуміти настрої та потреби користувачів, що дозволяє адміністраторам блогу оперативніше реагувати на запити та покращувати загальний досвід користувачів. Це, в свою чергу, підвищує лояльність користувачів до платформи та сприяє її подальшому розвитку.

В умовах постійного зростання кількості користувачів та обсягу інформації в інтернеті, створення автоматизованої системи аналізу коментарів стає критично важливим завданням. **Впровадження таких систем дозволяє ефективніше управляти інформаційним простором, підтримувати високу якість контенту та забезпечувати корисність і релевантність інформації для всіх користувачів платформи.**

Мета і задачі роботи

Метою кваліфікаційної роботи бакалавра є підвищення якості контент-аналізу коментарів IT блогу розробників платформи Unity засобами інтелектуального аналізу даних

Для досягнення поставленої мети слід вирішити такі **завдання**:

- виконати аналіз інформаційних моделей в області контент-аналізу;
- розглянути засоби інтелектуального аналізу даних області контент-аналізу коротких текстових даних, та обрати підхід для реалізації;
- виконати аналіз існуючих програмних засобів та наукових рішень;
- створити метод контент-аналізу коментарів IT блогу розробників платформи Unity засобами інтелектуального аналізу даних;
- виконати розробку архітектури нейромережі для визначення сентименту коментаря;
- створити проектну архітектуру інформаційної системи IT блогу розробників платформи Unity;
- виконати проєктування бази даних;
- виконати вибір та підготовку робочих вхідних даних методу контент-аналізу коментарів IT блогу;
- розглянути особливості використання спеціалізованих програмних компонентів для спрощення програмної розробки;
- виконати вибір засобів програмної реалізації методу контент-аналізу коментарів IT блогу розробників платформи Unity засобами інтелектуального аналізу даних;
- виконати програмну реалізацію створеного методу;
- виконати тестування створеного IT блогу розробників платформи Unity та застосунку для тренування нейромереж;
- виконати дослідження ефективності створеного методу з використанням розробленої програмної реалізації.

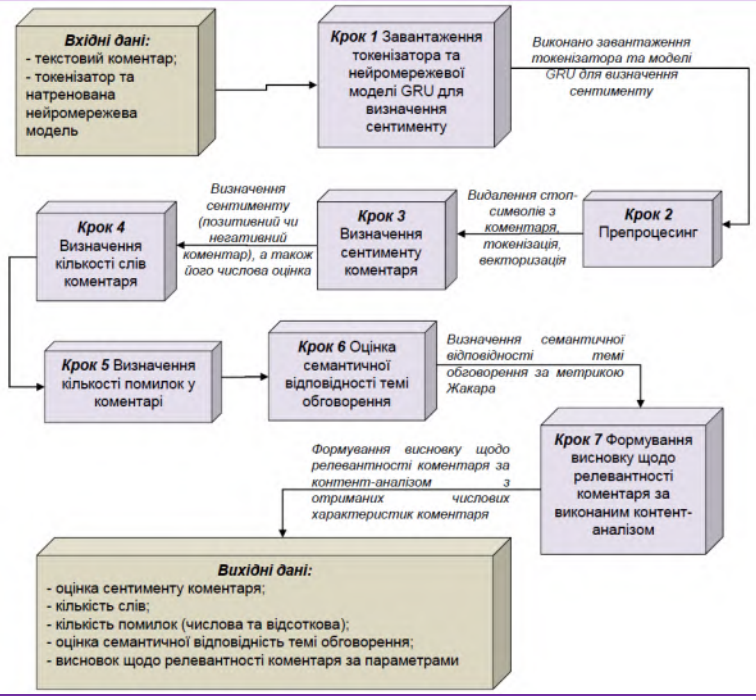


Схема та кроки методу контент-аналізу коментарів засобами інтелектуального аналізу даних

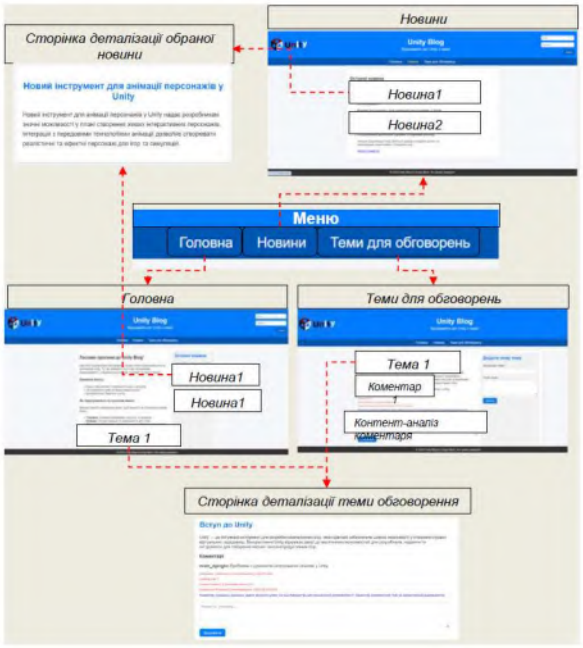
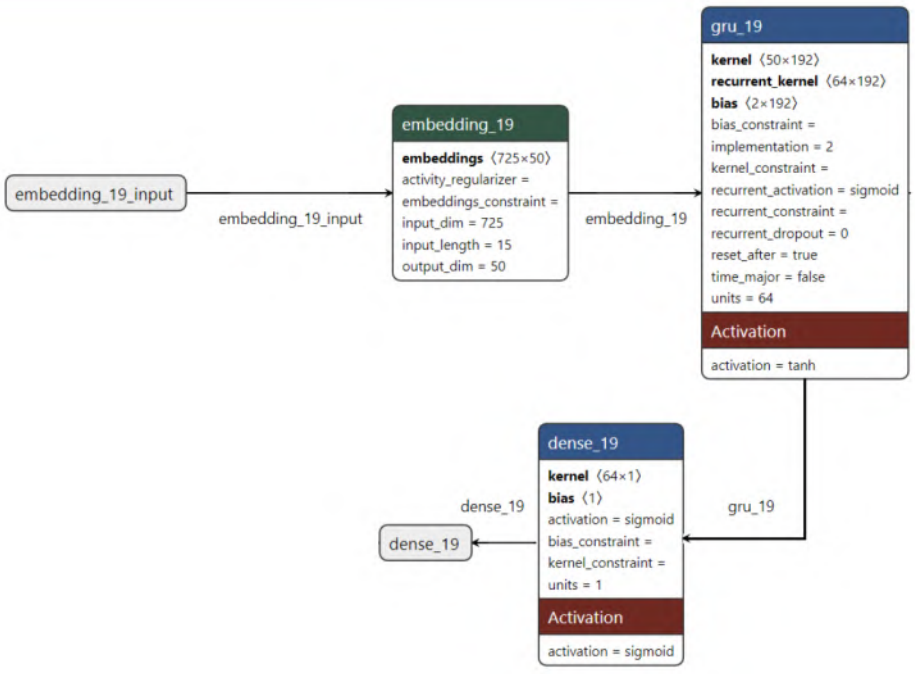


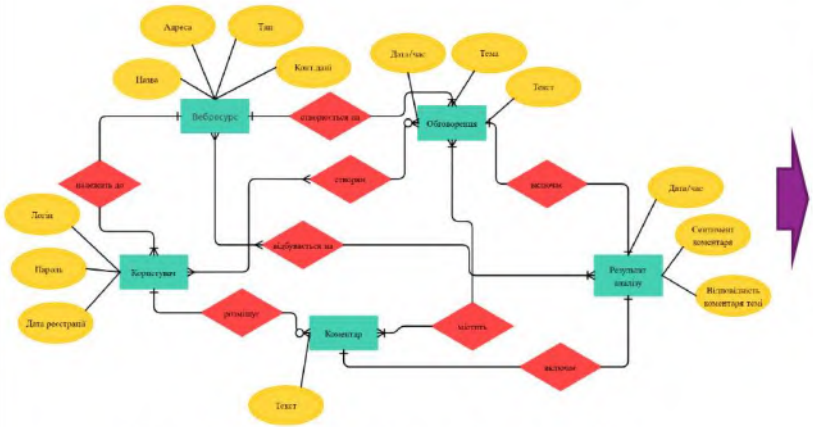
Схема навігації між сторінками IT блогу розробників платформи Unity



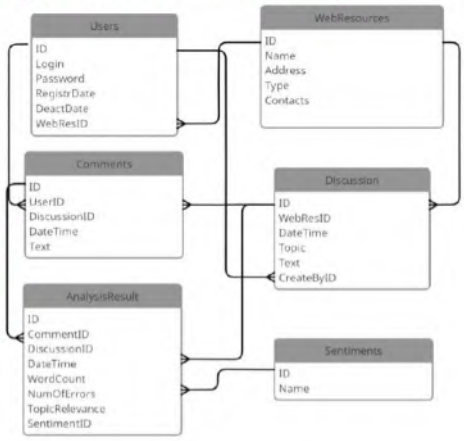
Спроєктована архітектура нейромережі GRU

Архітектура інформаційної системи інтелектуального аналізу коментарів для IT блогу розробників платформи Unity

Проєктування бази даних інформаційної системи інтелектуального аналізу коментарів для IT блогу



ER-діаграма БД для інформаційної системи



Даталогічна модель бази даних інформаційної системи

Датасет

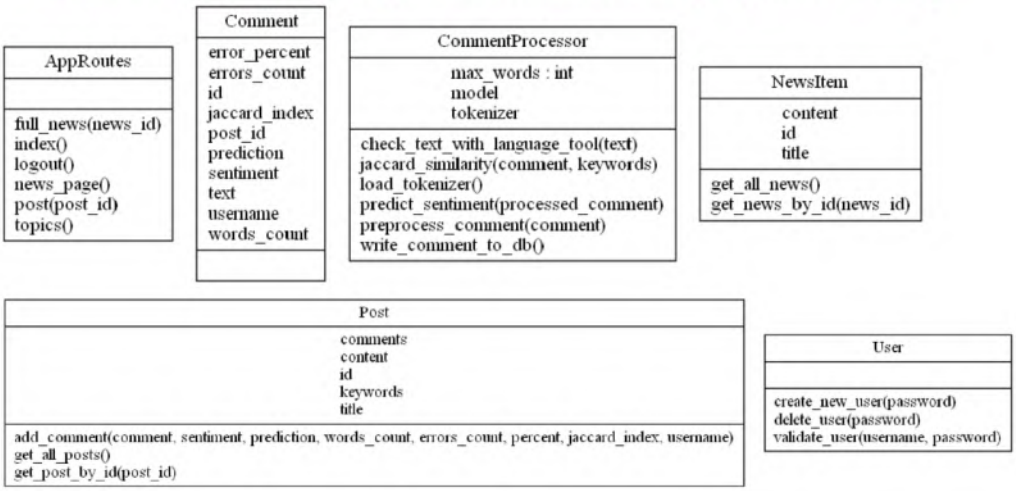
	review	sentiment
0	One of the other reviewers has mentioned that ...	positive
1	A wonderful little production. The...	positive
2	I thought this was a wonderful way to spend ti...	positive
3	Basically there's a family where a little boy ...	negative
4	Petter Mattei's "Love in the Time of Money" is...	positive

Також було проведено статистичне дослідження вмісту датасету:

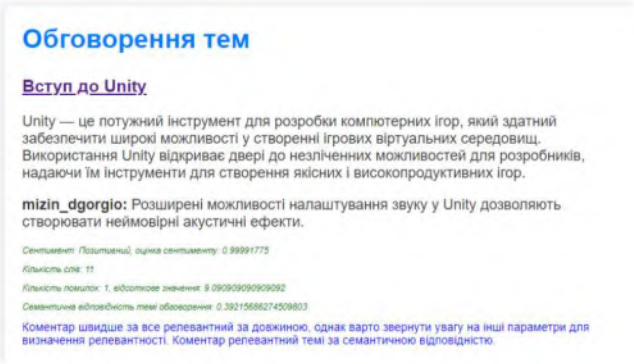
- кількість унікальних відгуків: 49,582;
- кількість унікальних відгуків: 2 (позитивний і негативний);
- загальна кількість відгуків: 50 000;
- позитивні: 25 000 відгуків;
- негативні: 25 000 відгуків.

Для реалізації методу контент-аналізу коментарів засобами інтелектуального аналізу даних для IT блогу розробників платформи Unity необхідно підготувати робочих вхідних даних для системи. Для цього було обрано «IMDB Dataset of 50K Movie Reviews»

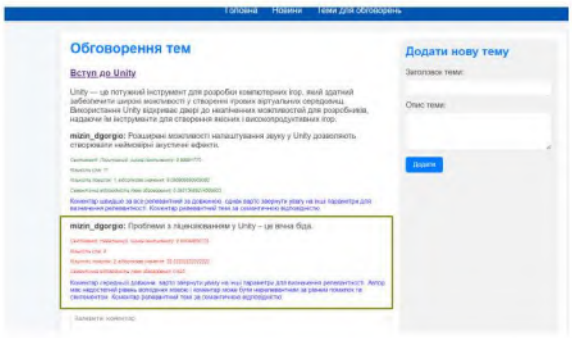
Діаграма класів інформаційної системи інтелектуального аналізу коментарів для IT блогу розробників платформи Unity



Інформаційна система інтелектуального аналізу коментарів для IT блогу



Приклад роботи підсистем контент-аналізу коментарів та роботи з темами для обговорення



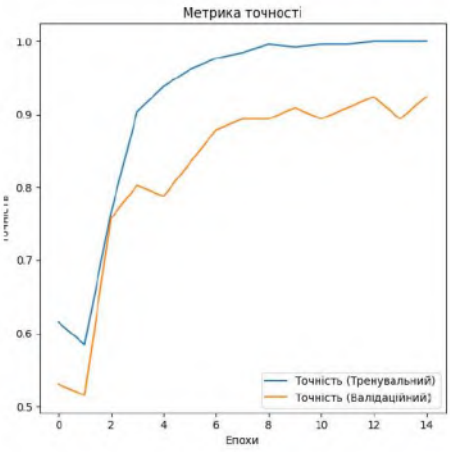
Приклад додавання негативного коментаря та його контент-аналіз

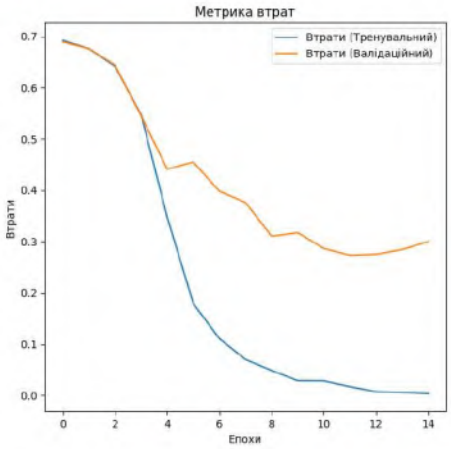
Результати досліджень

Дослідження ефективності було проведено спершу для ключових складових інформаційної системи у вигляді IT блогу. Першим було проведено дослідження ефективності нейромережі GRU, що відповідає за визначення сентименту користувацького коментаря. Данні з експерименту впливу параметрів нейромережі на якість навчання наведено в таблиці

Параметри	Accuracy	Loss
gru_units 64, к-сть епох 30, embedding_dim 30	0.9205	0.460
gru_units 128, к-сть епох 30, embedding_dim 40	0.924	0.3671
gru_units 256, к-сть епох 20, embedding_dim 40	0.8939	0.66
gru_units 64, к-сть епох 15, embedding_dim 50	0.9245	0.29

Результати досліджень

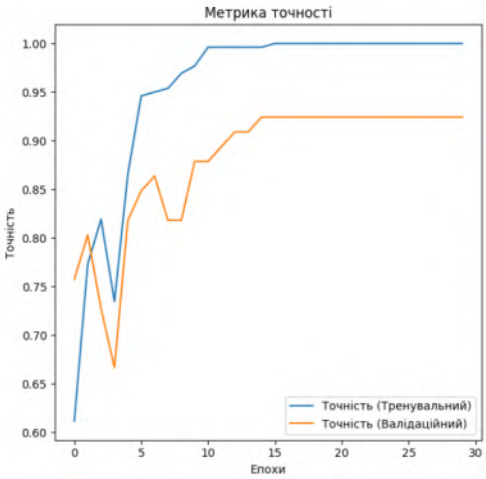


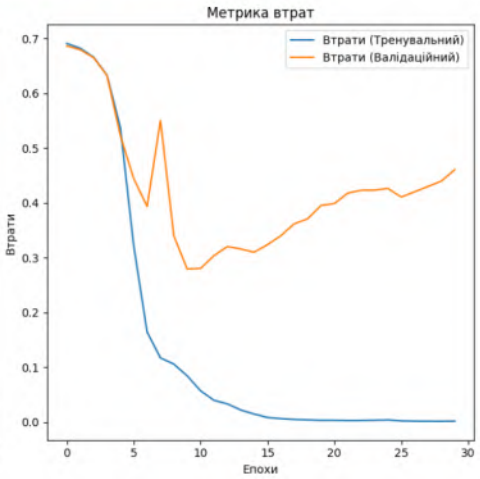


Графік функції втрат та точності при параметрах
gru_units 64, кількості epox 15, embedding_dim 50

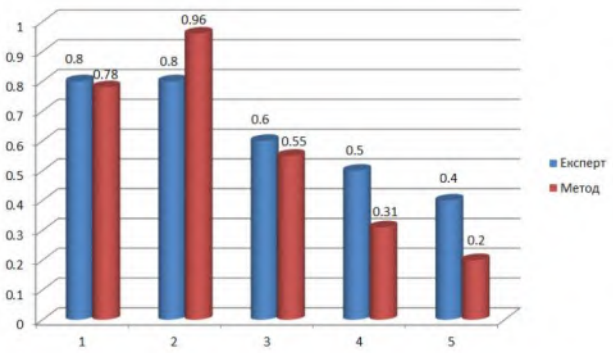
Результати досліджень

Графік функції втрат та точності при параметрах
gru_units 64, кількості epox 30, embedding_dim 30

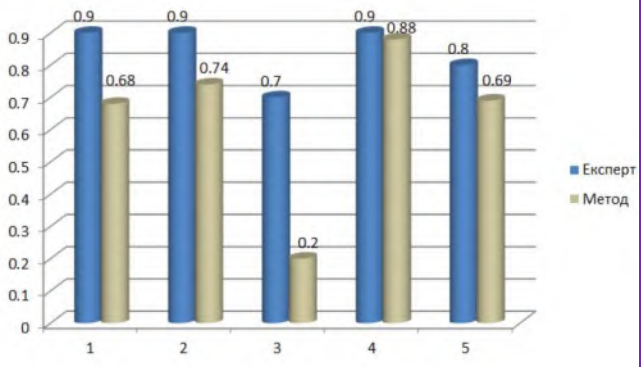




Результати досліджень



Діаграма порівняння відповідей експерта та методу



Діаграма порівняння відповідей експерта та методу для оцінки семантичної відповідність темі обговорення

Висновки

Було досягнуто мету кваліфікаційної роботи бакалавра – підвищення якості контент-аналізу коментарів IT блогу розробників платформи Unity засобами інтелектуального аналізу даних.

Для досягнення поставленої мети було поставлено та вирішено такі завдання:

- виконано аналіз інформаційних моделей в області контент-аналізу;
- розглянуто засоби інтелектуального аналізу даних області контент-аналізу коротких текстових даних, та обрано підхід для реалізації;
- виконано аналіз існуючих програмних засобів та наукових рішень;
- створено метод контент-аналізу коментарів IT блогу розробників платформи Unity засобами інтелектуального аналізу даних;
- виконано розробку архітектури нейромережі для визначення сентименту коментаря;
- створено проектну архітектуру інформаційної системи інтелектуального аналізу коментарів для IT блогу розробників платформи Unity;
- виконано проєктування бази даних інформаційної системи інтелектуального аналізу коментарів для IT блогу розробників платформи Unity;
- виконано вибір та підготовку робочих вхідних даних методу контент-аналізу коментарів IT блогу;
- розглянуто особливості використання спеціалізованих програмних компонентів для спрощення програмної розробки;
- виконано вибір засобів програмної реалізації методу контент-аналізу коментарів IT блогу розробників платформи Unity засобами інтелектуального аналізу даних;
- виконано програмну реалізацію створеного методу;
- виконано тестування створеного IT блогу розробників платформи Unity та застосунку для тренування нейромереж, що показало відсутність некоректно працюючих функцій;
- виконано дослідження ефективності створеного методу з використанням розробленого ПЗ.

Anti-Plagiarism v-15.257

Максимальне співпадіння з одним документом 3.0%

Словники перевірки: en_US, ru_RU, ua_UA. **Помилок в документах: 13%**

	Документ		Сумарний збіг по Базі Даних	
ID: 132213 Назва: КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА на тему Метод контент-аналізу коментарів засобами інтелектуального аналізу даних для IT блогу розробників платформи Unity Додано в БД: 2024-06-22 Автора: Джорджо МІЗИН Керівники: Павло РАДЮК Консультанти: Опоненти:	Символи	Лексеми	Символи	Лексеми
	75691	1086	4638 (6%)	75 (7%)

Джерело плагіату

ID	Опис	Наявність плагіату в документі	
		Символи	Лексеми

UNICHECK
by Turnitin

Ім'я користувача:
Кафедра КН

Дата перевірки:
22.06.2024 19:23:16 EEST

Дата звіту:
22.06.2024 19:25:16 EEST

ID перевірки:
1016382726

Тип перевірки:
Doc vs Internet + Library

ID користувача:
100005671

Назва документа: **КН-20-1 Мізин_ЗАПИСКА**

Кількість сторінок: 72 Кількість слів: 12057 Кількість символів: 96925 Розмір файлу: 2.57 MB ID файлу: 1016192612

8.57%
Схожість

Найбільша схожість: 3.77% з джерелом з Бібліотеки (ID файлу: 1016189018)

5.28% Джерела з Інтернету	477	Сторінка 74
6.06% Джерела з Бібліотеки	125	Сторінка 78

0% Цитат

Вилучення цитат вимкнене

Вилучення списку бібліографічних посилань вимкнене

0%
Вилучень

Немає вилучених джерел

Модифікації

Виявлено модифікації тексту. Детальна інформація доступна в онлайн-звіті.

Замінені символи	1

РІШЕННЯ ЕКСПЕРТНОЇ КОМІСІЇ КАФЕДРИ КОМП'ЮТЕРНИХ НАУК
ПРО ДОПУСК КВАЛІФІКАЦІЙНОЇ РОБОТИ ДО ЗАХИСТУ

Підтверджуємо ознайомлення з результатом звіту подібності щодо роботи, генерованого системою виявлення текстових збігів/ідентичності/схожості:

Назва: *Метод контент-аналізу коментарів засобами інтелектуального аналізу даних для IT блогу розробників платформи Unity*

Автор: *студент групи КН-20-1 Джорджо Мізин*

Спеціальність: *122 – Комп'ютерні науки*

Освітня програма: *освітньо-професійна*

Науковий керівник: *PhD, ст. викл. каф. КН Павло Радюк*

Після аналізу звіту подібності зроблено такий висновок:

№	Висновок	Позначка про відповідність
1	Запозичення, виявлені в роботі, є законними і не є плагіатом. Робота приймається до захисту.	***відповідає***
2	Виявлені запозичення не є плагіатом, розміщені в розділах, які не описують безпосередньо авторське дослідження, але кількість цитат перевищує обсяг, виправданий поставленою метою роботи Робота приймається до захисту, але має бути відкоригована. Відкоригований варіант має бути поданий на кафедру за 2 дні до захисту, разом із заявою щодо самостійності виконання письмової роботи та ідентичності друкованої та електронної версії роботи	
3	Виявлені запозичення не є плагіатом, але частково розміщені в розділах, які описують безпосередньо авторське дослідження, а кількість цитат перевищує обсяг, виправданий поставленою метою роботи. В зв'язку з цим мета роботи та поставлені завдання не були досягнені. Робота може бути допущена до захисту (наступного року) після того як буде відкоригована та допрацьована і успішно пройде повторну перевірку на академічний плагіат.	
4	Робота містить навмисні текстові спотворення, передбачувані спроби укриття запозичень або інші прояви академічного плагіату. Робота містить фабрикацію або фальсифікацію даних. Робота не допускається до захисту.	

Підтвердження:

Запозичення, виявлені в роботі Джорджо Мізина, не є плагіатом, оскільки: запозичення розміщені в розділі огляду існуючих підходів, не описують безпосередньо авторську роботу і не стосуються її результатів; усі запозичення фрагментарні; до запозичень входять фрагменти програмного коду, що не мають авторства і містять поширені конструкції; серед запозичень знаходяться загальновідомі терміни, скорочення та матеріали статей.

Обсяг запозичень, визначений системами виявлення збігів/ ідентичності/схожості, складає:

- за системою Anti-Plagiarism: 3 %;

- за системою Unicheck: 8.57 %.

Керівник роботи _____ *Павло РАДЮК*

Гарант ОП _____ *Олександр МАЗУРЕЦЬ*

Завідувач кафедри КН _____ *Олександр БАРМАК*

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
МОН УКРАЇНИ

Кафедра комп'ютерних наук

ВІДГУК НАУКОВОГО КЕРІВНИКА

на кваліфікаційну роботу бакалавра

студента *гр. КН-20-1 Мізина Джорджо Вадимовича*

за темою *Метод контент-аналізу коментарів засобами інтелектуального аналізу даних для IT блогу розробників платформи Unity*

1. Актуальність теми

Актуальність розробки методу контент-аналізу коментарів з використанням інтелектуального аналізу даних для IT блогу розробників платформи Unity зумовлена необхідністю ефективного моніторингу та обробки великого обсягу користувацьких відгуків, що дозволить швидко ідентифікувати основні потреби та проблеми користувачів, що, у свою чергу, сприятиме прискоренню процесу ухвалення рішень щодо удосконалення платформи і підвищенню рівня задоволеності користувачів.

2. Відповідність роботи предметній області Стандарту спеціальності 122 Комп'ютерні науки

За стандартом, а саме описом предметної області, об'єктом дослідження є процес інтелектуального аналізу коментарів IT блогу розробників платформи Unity засобами інтелектуального аналізу даних. Метою роботи є – підвищення якості контент-аналізу коментарів IT блогу розробників платформи Unity засобами інтелектуального аналізу даних.. При вирішенні поставленої задачі методи контент-аналізу для коментарів IT блогу. А, отже, результати виконання кваліфікаційної роботи бакалавра повністю відповідають стандарту бакалавра спеціальності 122 – Комп'ютерні науки.

3. Професійні та особистісні якості бакалавра

Студент Мізин Джорджо Вадимович під час виконання кваліфікаційної роботи бакалавра проявив компетентності передбачені стандартом бакалавра спеціальності 122 – Комп'ютерні науки, демонструючи розуміння теми та здатність аналізувати та інтерпретувати наукові дані з відповідними методами дослідження.

4. Ступінь самостійності під час виконання кваліфікаційної роботи

Під час виконання кваліфікаційної роботи студент продемонстрував достатньо високий рівень самостійності, що виявився у здатності самостійно визначати та формулювати завдання, обирати методи дослідження та аналізу отриманих даних.

5. Ступінь оволодіння методами дослідження

В процесі виконання кваліфікаційної роботи студент продемонстрував достатній рівень оволодіння методами дослідження, застосовуючи теоретичні знання для розробки методу та проведення експериментального дослідження.

6. Повнота та якість розкриття теми роботи

Повнота та якість розкриття теми роботи на достатньому рівні, що відображається у вирішенні поставлених завдань.

7. Логічність, послідовність, аргументованість, літературна грамотність викладення матеріалу

У кваліфікаційній роботі студент продемонстрував достатній рівень логічності, послідовності, аргументованості та літературної грамотності викладення матеріалу, що дозволило структурувати дослідження, уникнути непослідовностей і забезпечити наукову обґрунтованість.

8. Можливість практичного застосування кваліфікаційної роботи бакалавра, окремих її частин

Розроблена інформаційна система орієнтована на автоматизацію обробки коментарів і їх класифікацію, що дозволяє зменшити трудомісткість для модераторів блогу та підвищити ефективність в моніторингу обговорень.

9. Висновок про можливість допуску кваліфікаційної роботи бакалавра до захисту, на яку оцінку заслуговує робота

*Враховуючи рівень виконання та забезпечення усіх необхідних вимог, робота може бути допущена до захисту. Рекомендована оцінка «**задовільно**» .*

Керівник _____ *PhD, ст. викл. каф. КН Павло РАДЮК*

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
МОН УКРАЇНИ

Кафедра комп'ютерних наук

РЕЦЕНЗІЯ

на кваліфікаційну роботу бакалавра

студента *гр. КН-20-1 Мізина Джорджо Вадимовича*

за темою: *Метод контент-аналізу коментарів засобами інтелектуального аналізу даних для IT блогу розробників платформи Unity*

1. Актуальність обраної теми

Актуальність розробки методу контент-аналізу коментарів засобами інтелектуального аналізу даних для IT блогу розробників платформи Unity полягає у можливості виявлення ключових тенденцій та проблем, з якими стикаються користувачі, що сприятиме підвищенню якості продукту та покращенню взаємодії з аудиторією. Описаний в роботі підхід дозволить автоматизувати процес збирання зворотного зв'язку, оптимізувати підтримку клієнтів і сприяти розвитку спільноти розробників.

2. Повнота розкриття мети та завдань роботи

У своїй кваліфікаційній роботі студент повністю розкрив мету та завдання дослідження, показавши достатнє розуміння обраної теми та чітке формулювання наукових проблем. Він спланував та реалізував дослідження відповідно до визначених завдань, що свідчить про його компетентність.

3. Зміст кожного розділу роботи

Пояснювальна записка містить три розділи. В першому розділі наведено характеристику предметної області, а саме аналіз моделей, методів та реалізацій для контент-аналізу коментарів блогів. У другому розділі наведено та описано метод контент-аналізу коментарів засобами інтелектуального аналізу даних для IT блогу розробників платформи Unity, а також спроєктована архітектура та взаємозв'язок компонентів інформаційної системи. У третьому розділі проведено експериментальне дослідження методу контент-аналізу коментарів для IT блогу розробників платформи Unity.

4. Оцінка розробленої інформаційної системи, її практична цінність

Розроблена система спрямована на автоматизацію процесу аналізу коментарів IT блогу розробників платформи Unity, забезпечуючи високу якість контент-аналізу за допомогою інтелектуальних методів обробки даних. Використання цих методів дозволяє ефективно ідентифікувати ключові тенденції, проблеми та потреби користувачів, що сприяє покращенню взаємодії з аудиторією та вдосконаленню самої платформи.

5. Якість оформлення кваліфікаційної роботи бакалавра

Якість оформлення кваліфікаційної роботи бакалавра свідчить про достатній рівень академічної підготовки студента, який дотримався всіх встановлених стандартів і вимог до структури та стилю наукових досліджень. В результаті матеріал дослідження подано чітко і логічно.

6. Недоліки кваліфікаційної роботи бакалавра

В тексті роботі присутні граматичні та пунктуаційні помилки. Деякі елементи пояснювальної записки розміщені по тексту раніше ніж перше посилання на них. У дослідженні ефективності варто було використати декілька генеративних моделей у якості експерта.

7. Загальний висновок (допускається чи не допускається до захисту), та оцінка на яку заслуговує кваліфікаційна робота.

Враховуючи рівень виконання та забезпечення усіх необхідних вимог, робота може бути допущена до захисту. Рекомендована оцінка ***«добре»****.*

Рецензент Радивчук Галина Іванівна, канд. техн. наук, доцент кафедри ТПЗ ХНЕУ

21.06.2024 р.