Хмельницький національний університет
Факультет інформаційних технологій
Кафедра комп'ютерних наук

КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА

на тему *Метод анотування україномовних художніх творів засобами машинного навчання*

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 КНс-21-1* _____ *Михайло ПРОСВІТЛЮК*
Група виконавця | Підпис | Ім'я, ПРІЗВИЩЕ

Керівник: *викладач каф. КН* _____ *Валерія КЛІМЕНКО*
Науковий ступінь, посада | Підпис | Ім'я, ПРІЗВИЩЕ

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

До захисту допускаю:
зав. кафедри КН, д.т.н., професор _____ *Олександр БАРМАК*
Підпис | Ім'я, ПРІЗВИЩЕ

20 06 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 анотування україномовних художніх творів засобами машинного навчання»*

2. Завдання видано студенту *Михайлу ПРОСВІТЛЮКУ*
(Ім'я, прізвище)

3. Керівник роботи *викладач кафедри КН Валерія КЛІМЕНКО*
(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 роботи – спрощення створення анотацій україномовних художніх творів шляхом автоматизації анотування засобами машинного навчання. Вихідними даними є україномовна анотація, та її оцінки за метриками. Для досягнення мети слід вирішити такі завдання: дослідити предметну область анотування художніх творів; обрати підхід до автоматизованого анотування художніх творів серед методів машинного навчання; створити метод анотування україномовних художніх творів засобами машинного навчання; виконати програмну реалізацію інтелектуальної системи анотування; провести тестування створеної програми та виконати дослідження ефективності запропонованого методу анотування україномовних художніх творів.

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 КНс-21-1* _____ *Михайло ПРОСВІТЛЮК*
Група виконавця | Підпис | Ім'я, ПРІЗВИЩЕ

Керівник: *викладач каф. КН* _____ *Валерія КЛІМЕНКО*
Науковий ступінь, посада | Підпис | Ім'я, ПРІЗВИЩЕ

Анотація

Тема кваліфікаційної роботи бакалавра: *«Метод анотування україномовних художніх творів засобами машинного навчання»*

Виконавець кваліфікаційної роботи бакалавра: *студент групи КНс-21-1 Михайло ПРОСВІТЛЮК*

Керівник кваліфікаційної роботи бакалавра: *викладач кафедри КН Валерія КЛІМЕНКО*

Кваліфікаційна робота бакалавра містить:

Пояснювальна записка				Кількість додатків
Сторінок	Рисунків	Таблиць	Джерел інформації	
68	24	16	28	5

Метою кваліфікаційної роботи бакалавра є спрощення створення анотацій україномовних художніх творів шляхом автоматизації анотування засобами машинного навчання. Для розробки інформаційної системи було використано мову програмування Python, середовище програмування PyCharm, а також систему керування базами даних SQLite.

Розроблена система призначена для авторів, викладачів, дослідників або літературних аналітиків, що шукають швидкий і ефективний спосіб аналізу літературних творів українською мовою.

Напрямками практичного використання розробленої інформаційної системи визначено автоматизовану генерацію україномовної анотації для заданого твору, а також її оцінку за метриками.

Ключові слова: анотування україномовних художніх творів, BARD, інформаційна система, метрики.

Виконавець: *студент групи КНс-21-1* — *Михайло ПРОСВІТЛЮК*

Група виконавця | Підпис | Ім'я, ПРІЗВИЩЕ

Зміст

Перелік скорочень

Скорочення, термін, позначення	Пояснення
RNN	Рекурентні нейронні мережі
CNN	Згорткові нейронні мережі
БД	База даних
ІС	Інформаційна система
КРБ	Кваліфікаційна робота бакалавра
НМ	Нейронна мережа
СКБД	Система керування базами даних
ХНУ	Хмельницький національний університет.

Вступ

Кваліфікаційна робота бакалавра присвячена спрощенню створення анотацій україномовних художніх творів шляхом автоматизації анотування засобами машинного навчання, для чого виконувалась розробка методу анотування україномовних художніх творів засобами машинного навчання та відповідного застосунку, що дозволяє за текстовим контентом виконати анотування художнього твору.

Актуальність. В умовах розвитку глобальної мережі, так само і невпинно зростає кількість художніх творів, які автори щоденно завантажують. Автоматичне анотування дозволяє швидко та ефективно створювати метадані для великих обсягів літературних текстів, що розширює наявні бази даних і полегшує доступ до них.

На основі анотованих даних є можливість розробки систем рекомендацій, які враховують індивідуальні інтереси користувачів, а також засоби для автоматичного аналізу змісту та семантики текстів. Ще анотування літературних текстів може служити як основа для розробки освітніх програм з цифрової грамотності та літературознавства, допомагаючи учням аналізувати та розуміти текстову інформацію, а також допомагає зберегти та розповсюдити літературну спадщину, забезпечуючи доступ до цінних творів для наступних поколінь та дослідників..

Отже, анотування україномовних художніх творів засобами машинного навчання сприяє розвитку літературознавства, полегшує доступ до літературних ресурсів та розвиває технології обробки природної мови.

Об’єкт дослідження – процес анотування україномовних художніх творів засобами машинного навчання.

Предмет дослідження – методи та засоби машинного навчання для роботи з текстовою інформацією.

Мета кваліфікаційної роботи бакалавра – спрощення створення анотацій україномовних художніх творів шляхом автоматизації анотування засобами машинного навчання.

Завдання кваліфікаційної роботи бакалавра – виконати дослідження предметної області анотування художніх творів засобами машинного навчання; виконати огляд теоретичних підходів до вирішення подібних задач, обрати підхід до автоматизованого анотування художніх творів серед методів машинного навчання; створити метод анотування україномовних художніх творів засобами машинного навчання; створити інформаційну структуру системи автоматизованого анотування користувацьких художніх текстів; виконати програмну реалізацію інформаційної системи; провести тестування інформаційної системи автоматизованого анотування користувацьких художніх текстів; виконати дослідження ефективності методу анотування україномовних художніх творів засобами машинного навчання.

Розділ 1 Характеристика предметної області анотування художніх творів засобами машинного навчання

1.1 Аналіз інформаційних моделей

Завдання анотування та реферування текстової інформації здійснюється з моменту виникнення області обробки природної мови. Це важливі завдання в контексті досліджень у галузі обробки природної мови та штучного інтелекту. Обробка природної мови відноситься до дисциплін інформаційних технологій і займається аналізом та синтезом текстової інформації, яку люди використовують для комунікації [1].

Анотування є процесом аналізу та структурування інформації з метою створення короткого опису або резюме текстового документу, такого як книга, стаття або наукова публікація. Головна мета анотування – надати коротку характеристику змісту документа, розкриваючи його основну ідею, ключові аспекти та структуру, але не розголошуючи повністю всю інформацію [2].

В анотаціях зазвичай включається бібліографічний опис, що містить дані про автора, назву, дату публікації тощо, а також сам текст анотації, який висвітлює основні пункти змісту документа. Цей метод дозволяє отримати швидкий огляд важливих аспектів документа, спростити пошук та систематизацію інформації, а також швидко оцінити його релевантність та значимість для власних потреб. Анотації допомагають зекономити час, оскільки дозволяють швидко засвоїти ключову інформацію з документа без необхідності читати його повністю. Вони є стандартним жанром наукового письма, що відрізняється від інших вторинних текстів своєю максимальною компактністю (до 500 символів) і лаконічністю. Надають загальне уявлення про зміст первинного джерела, не вдаючись у подробиці [3].

Анотації виконують дві ключові функції: сигнальну та пошукову. Сигнальна функція полягає в наданні важливої інформацію про документ, яка допомагає читачам встановити основний зміст і призначення документа. Це дозволяє вирішити, чи є необхідність у зверненні до повного тексту праці.

Пошукова функція анотації використовуються у інформаційно-пошукових системах, в тому числі автоматизованих, для пошуку конкретних документів. Вони допомагають користувачам знаходити необхідну інформацію шляхом зазначення ключових аспектів документа [4].

Анотація складається з двох основних частин: бібліографічного опису та самого тексту анотації. Вона не розкриває повністю зміст наукового джерела, а лише надає коротку інформацію про його зміст та характер. Анотація допомагає користувачам отримати достатньо об'єктивне уявлення про наукову публікацію та допомагає у пошуку, відборі та систематизації необхідної інформації.

Є доволі широка шкала класифікації анотацій, нижче буде наведено основні.

Анотації поділяються на описові та реферативні залежно від їх обсягу та глибини [5].

Описові анотації надають загальний огляд змісту первинного документа, узагальнюючи основні теми, що в ньому відображені. Описові анотації відповідають на питання: "Про що йдеться у документі?"

Реферативні анотації не лише перераховують основні теми, а й детально розкривають їх зміст. Реферативні анотації відповідають на два питання: "Про що йдеться в основному документі?" та "Що саме з цього приводу повідомляється?"

За функціональним призначенням анотації поділяються на довідкові та рекомендаційні.

Довідкова анотація є анотацією, яка уточнює тему документа, надає короткі відомості про автора, жанр твору, читацьку адресу та інші особливості публікації, які можуть бути короткими або розгорнутими. Цей тип анотації використовується в наукових, навчально-методичних та довідкових виданнях.

Рекомендаційна анотація є анотацією з дидактичною та просувальною спрямованістю, яка зосереджується на новизні ідей, фактах та інших перевагах публікації. Вона пропагує документ як корисне джерело інформації для певної

аудиторії та заохочує до його читання. Цей тип анотації переважно використовується в літературно-художніх та науково-популярних виданнях [6].

Видавнича анотація, як вторинний документ, повинна відповідати наступним вимогам:

– Доступність. Бути зрозумілою для широкого кола читачів без необхідності звертатися до первинного документа.

– Змістовність. Містити достатньо інформації, щоб читач міг утворити уявлення про первинний документ, включаючи тему, фактаж, структуру, художнє оформлення, цінність публікації та інше.

– Лаконічність. Мати значно менший обсяг, ніж первинний документ (зазвичай до 500 знаків), акцентуючи увагу на головних характеристиках і не деталізуючи додатково. Вона не повинна містити сторонніх відомостей.

– Стилістична відповідність. Тягнутися до безособових синтаксичних конструкцій, фраз-кліше, загальноприйнятих скорочень і уникати складних формулювань або цитат.

Отже, розробка алгоритмів та програмних засобів для автоматичного створення анотацій та рефератів дозволить значно полегшити процес обробки та аналізу художніх україномовних текстів, тому автоматизація процесу анотування є актуальною задачею обробки природної мови.

1.2 Огляд теоретичних підходів до задач автоматизованого анотування художніх творів

На сьогодні існує декілька підходів до автоматичного анотування, які можна розділити на дві основні групи: методи складання витягів (витягувальні алгоритми) і формування короткого викладу (генерувальні алгоритми) [7].

Витягувальні алгоритми формують анотацію, використовуючи текстові фрагменти вхідного документа. Вони виділяють блоки тексту з найбільшою лексичною та статистичною значущістю і об'єднують їх у складану анотацію. Цей підхід простий у реалізації, не вимагає великих обчислювальних ресурсів,

але може не забезпечити достатньої якості через відсутність семантичного аналізу тексту.

Генерувальні алгоритми аналізують вхідний документ для пошуку інформації, на основі якої формується текст анотації. Вони здатні урахувати семантичні зв'язки у тексті, уникнути дублювання інформації між основним текстом та анотацією, і забезпечити повноту анотації. Цей підхід вимагає більшого обсягу обчислювальних ресурсів, але може забезпечити вищу якість анотування.

Рисунок 1.1 – Класифікація методів автоматичного анотування [7]

Генерувальні алгоритми мають ряд переваг перед витягувальними, адже можуть генерувати анотації, які описують не лише фактичний зміст тексту, але й його емоційний та естетичний вплив, можуть бути використані для анотування текстів будь-якої складності, а також можуть бути адаптовані до різних жанрів та стилів текстів.

Ще до генерувальних алгоритмів належить ряд алгоритмів, таких як: алгоритми на основі правил, алгоритми машинного навчання, алгоритми на основі графів, алгоритми на основі нейронних мереж.

Алгоритми на основі правил генерують анотації, використовуючи набір правил, що описують, як анотувати певні типи тексту. Ці алгоритми можуть використовуватися для виявлення та анотування таких елементів, як персонажі, події, місця, теми, символи та метафори. Правила можуть бути написані вручну або вивчені з набору даних з анотованими текстами. Однак, ці алгоритми можуть бути ефективними, але вони також мають свої обмеження. Наприклад, вони можуть не завжди точно визначити ключові фрагменти або не враховувати контекст. Тому в деяких випадках комбінація алгоритмів на основі правил та методів машинного навчання може давати кращі результати.

Алгоритми машинного навчання генерують анотації, використовуючи модель, навчену на наборі даних з анотованими текстами. Ці алгоритми можуть використовуватися для класифікації текстів за жанром, стилем або настроєм. Також можуть використовуватися для виявлення та анотування емоцій, сентиментів та інших суб'єктивних характеристик тексту [8].

Алгоритми на основі графів генерують анотації, аналізуючи структуру тексту та виявляючи важливі мотиви та теми.

Алгоритми на основі нейронних мереж генерують анотації, використовуючи нейронну мережу, навчену на наборі даних з анотованими текстами. Ці алгоритми можуть використовуватися для генерування анотацій, які описують не лише фактичний зміст тексту, але й його емоційний та естетичний вплив.

Одним з методів розв'язання цієї задачі на основі нейронних мереж є використання рекурентних нейронних мереж. RNN включають один або кілька шарів нейронів, які зв'язані з попереднім шаром. Це базова архітектура, розроблена у 1980-х роках. Кожен нейрон мережі може змінювати свій поріг активації з часом, і має змінні ваги. Нейрони поділяються на вхідні, вихідні та приховані. RNN ефективно працюють з послідовними даними, такими як

речення або слова. RNN мають пам'ять і використовують її для аналізу кожного наступного елемента вхідних даних.

Однак у RNN є обмеження, пов'язане з короткою пам'яттю. Інформація у пам'яті змішується з новою на кожному кроці, тому контекст може бути втрачений. Наприклад, для речення "я пішов в магазин по ковбасу" достатньо контексту, щоб зрозуміти, що "я" пішов саме в "магазин". Але для речення "Я виріс в Україні. Я пишаюся тим, що я українець" потрібен контекст попереднього речення. Тут LSTM (Long Short-Term Memory) RNN вирішує цю проблему, оскільки вона зберігає інформацію на тривалий термін, не "забуваючи" її з часом. Вона використовує спеціальні структури, щоб контролювати, коли значення пам'яті має змінитися [9].

Seq2seq мережа, яка базується на архітектурі RNN (рисунок 1.2), використовується для перекладу однієї послідовності на іншу. Вона містить два LSTM RNN: кодувальник і декодувальник. Кодувальник генерує вектор, який представляє вхідну послідовність, який потім передається декодувальнику, який відновлює цільову послідовність. Кожен вихід використовується для оновлення прихованого стану, що допомагає у визначенні важливості елементів тексту при ранжуванні N-грам тексту.

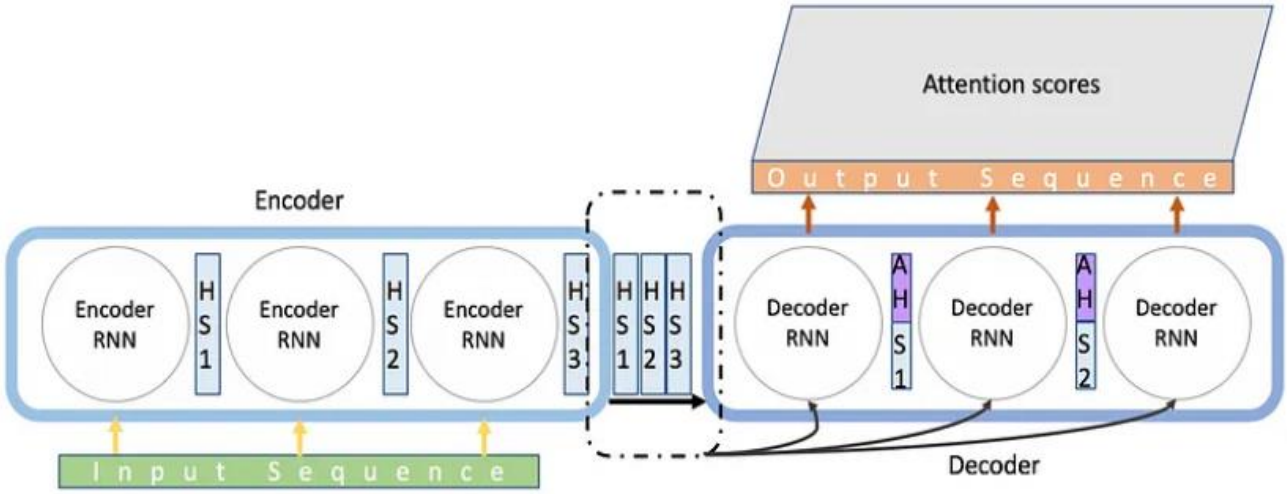


Рисунок 1.2 – Базова модель Seq2seq [10]

Ще однією моделлю машинного навчання є Google Bard що є передовою моделлю нейронної мережі, розроблена Google для завдань обробки природної мови. Вона представляє значний прогрес у можливостях штучного інтелекту, дозволяючи машинам розуміти, інтерпретувати та генерувати текст, подібний до людського, з високим ступенем узгодженості та контекстуальності. Google Bard демонструє високі навички в написанні віршів, оповідань та участі в змістовних бесідах. Його здатність створювати тексти поставила його на передній план інновацій штучного інтелекту, переосмислюючи межі машинно-генерованого контенту [11].

Bard від Google побудований на складній архітектурі нейронної мережі, навченої на великому масиві різноманітних текстових даних, включаючи літературу, історичні записи та сучасне мовлення. Завдяки передовим методам глибокого навчання Bard обробляє вхідні підказки та контекст для створення чіткого та тематично узгодженого тексту, демонструючи розуміння мовних нюансів і зв’язність оповіді. Його функціонування включає взаємодію мовних токенів, семантичних вбудовувань і прогнозного моделювання, що дозволяє генерувати текст із плавністю та виразністю, схожими на людські.

Отже, з виконаного огляду теоретичних підходів буде використано генерувальний підхід, адже можуть генерувати анотації, які описують не лише фактичний зміст тексту, але й його емоційний та естетичний вплив, що є суттєвим для анотування художніх творів. Для реалізації генерувального підходу буде використано нейромережеві засоби, а саме BARD на основі Seq2seq адже таким чином можна враховувати контекст та семантику україномовних художніх творів для створення анотації.

1.3 Аналіз існуючих програмних засобів та наукових рішень щодо художніх творів засобами машинного навчання

Текстові анотації є важливим елементом підходів до обробки природної мови. Процес анотування, який виконують люди вручну, має різні недоліки, такі

як суб’єктивність, повільність, втома та, можливо, недбалість. Крім того, анотатори можуть коментувати неоднозначні дані. Тому авторами розроблено концепцію автоматизованих анотацій, щоб отримати найкращі анотації за допомогою кількох підходів машинного навчання. Запропонований підхід базується на ансамблевому алгоритмі методів метанавчання та метавекторизатора. Цей підхід використовує техніку напівконтрольованого навчання для автоматизованих анотацій для виявлення мови ненависті. Це передбачає використання різних алгоритмів машинного навчання, зокрема Support Vector Machine (SVM), Decision Tree (DT), K-Nearest Neighbors (KNN) і Naive Bayes (NB), у поєднанні з методами вилучення тексту Word2Vec і TF-IDF. Процес анотації виконується з використанням 13 169 індонезійських даних коментарів YouTube. У запропонованій моделі використовувався підхід Стеммінга з використанням даних із Sastrawi та нових даних із 2245 слів. Напівконтрольоване навчання використовує 5%, 10% і 20% позначених даних порівняно з виконанням позначення на основі 80% наборів даних. У напівконтрольованому навчанні модель вивчає дані з мітками, які надають явну інформацію, і дані без міток, які пропонують неявні ідеї. Цей гібридний підхід дозволяє моделі узагальнювати та робити обґрунтовані прогнози, навіть коли доступні обмежені позначені дані (на основі самонавчання). Зрештою, це покращує його здатність працювати зі сценаріями реального світу з дефіцитною анотованою інформацією. Крім того, запропонований метод використовує різноманітні порогові значення для відповідності слів, позначених мовою ворожнечі, у діапазоні від 0,6, 0,7, 0,8 до 0,9. Експерименти показали, що модель DT-TF-IDF має найкраще значення точності 97,1% зі сценарієм 5%:80%:0,9. Однак кілька інших методів мають точність понад 90%, наприклад SVM (TF-IDF і Word2Vec) і KNN (Word2Vec), засновані на обох методах вилучення тексту в кількох сценаріях тестування [12].

У статті досліджуються особливості реферату та анотації як вторинних документів. Звертається увага на різні способи перекладу цих термінів, вказуються їхні спільні та відмінні характеристики. Основна увага зосереджена

на процесі створення професійно спрямованих документів, де важливо ретельно працювати над змістом первинного тексту для написання логічних та чітко структурованих вторинних висловлювань. У статті пропонуються етапи та алгоритм організації роботи з анотування та реферування тексту. Наголошується, що реферат та анотація мають відповідати певним вимогам і не просто бути скороченим варіантом первинного тексту. Розглядається система завдань, які сприяють розвитку навичок роботи з текстом, таких як узагальнення та перефразування, а також розвиток навичок компресії тексту. Важливим елементом є розуміння змістової структури абзацу та використання такого поняття, як key sentence або topic sentence. Наголошується на значенні пошуку ключових слів та використанні mind maps для логічного аналізу тексту. Вивчення питань навчання анотування та реферування текстів у студентів нелінгвістичних спеціальностей акцентується як важливий аспект формування їх професійної компетенції. Важливість навичок написання вторинних текстів для вивчення мови, оптимізації оцінювання та подальшої професійної діяльності студентів підкреслюється. Реферування та анотування розглядаються як додатковий засіб оптимізації процесу навчання іноземної мови в умовах дистанційної освіти [13].

Так як завдання реферування належить до задачі генерації текстів, нижче будуть розглянуті програмні продукти, що можуть виконувати генерацію текстів та узагальнення. Одна з таких програм – Smodin. AI Writer і генератор тексту Smodin. Цей безкоштовний редактор AI дозволяє створювати есе та статті. Він може бути корисним для написання художніх текстів.

Для узагальнення тексту необхідно ввести текст або завантажити текстовий файл, який потрібно узагальнити, та по обраній довжині буде сформовано текст. Smodin автоматично генерує стисле резюме, яке можна редагувати та налаштовувати. Можна зберігати узагальнені тексти у Smodin для подальшого використання, а також ділитись ними з оточуючими. Вигляд програми наведено на рисунку 1.3

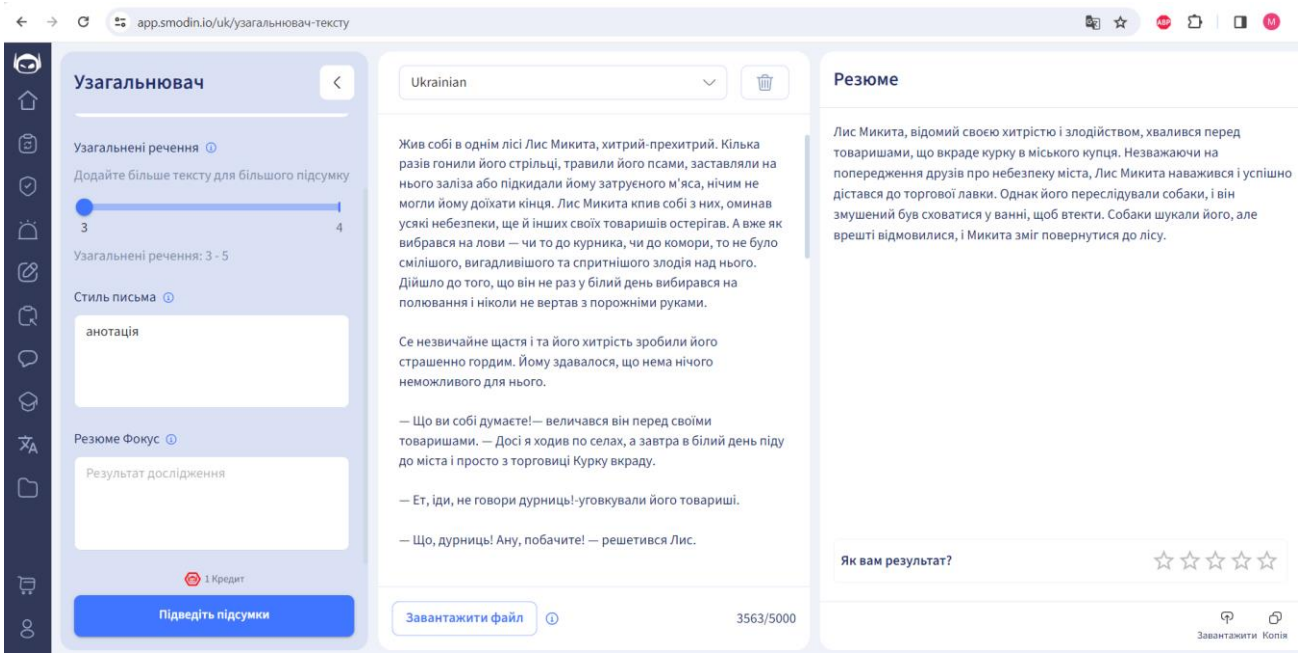


Рисунок 1.3 – Інтерфейс програми Smodin [14]

Smodin може допомогти узагальнювати статті, дослідницькі роботи, книги, електронні листи та інші текстові документи.

Простіше кажучи, підсумовування штучного інтелекту – це використання технологій штучного інтелекту для дистиляції тексту, документів або вмісту в короткий і легкозасвоюваний формат. Наприклад, підведення підсумків штучним інтелектом може використовувати обробку природної мови або розуміння, щоб стиснути довгий PDF-файл і повторити його найважливіші висновки лише в кількох реченнях.

Найкращий штучний інтелект для узагальнення залежить від цілей користувача. Google Bard може допомогти узагальнити текст, код, сценарії, музичні твори, електронні листи, листи тощо для особистого користування. Для більш розширеного резюмування, в тому числі для цілей дослідження та бізнес-аналітики, Vertex AI PaLM API може витягти короткий виклад найважливішої інформації з тексту за допомогою підказок резюмування [15].

Отож, було виконано аналіз предметної області, в рамках якого зазначено, що розробка алгоритмів та програмних засобів для автоматичного створення анотацій та рефератів дозволить значно полегшити процес обробки та аналізу художніх україномовних текстів, тому автоматизація процесу анотування

є актуальною задачею обробки природної мови. З виконаного огляду теоретичних підходів було обрано для використання генерувальний підхід, адже можуть генерувати анотації, які описують не лише фактичний зміст тексту, але й його емоційний та естетичний вплив, що є суттєвим для анотування художніх творів. Для реалізації генерувального підходу буде використано нейромережеві засоби, а саме Seq2seq мережа, яка належить класу рекурентних нейронних мереж, адже вона може враховуючи контекст та семантику україномовних художніх творів.

З огляду наукових праць та існуючих програмних рішень, напрям є актуальним. Відповідно, для ефективного вирішення задачі анотування художніх україномовних творів необхідним є створення методу анотування україномовних художніх творів засобами машинного навчання, та програмного забезпечення, що буде використовувати даний метод.

1.4 Мета, задачі та вимоги до реалізації інформаційної системи

Метою кваліфікаційної роботи бакалавра є спрощення створення анотацій україномовних художніх творів шляхом автоматизації анотування засобами машинного навчання.

Для досягнення мети необхідно розробити відповідний метод та створити відповідну програмну реалізацію у вигляді віконного застосунку, що призначений для автоматизованого анотування україномовних художніх творів за користувацьким текстом, а також вирішити наступні задачі:

– дослідження предметної області анотування художніх творів засобами машинного навчання;

– виконати огляд теоретичних підходів до вирішення подібних задач, обрати підхід до автоматизованого анотування художніх творів серед методів машинного навчання;

– створити метод анотування україномовних художніх творів засобами машинного навчання;

– створити інформаційну структуру системи автоматизованого анотування користувацьких художніх текстів;

– виконати програмну реалізацію інформаційної системи;

– провести тестування інформаційної системи автоматизованого анотування користувацьких художніх текстів;

– виконати дослідження ефективності методу анотування україномовних художніх творів засобами машинного навчання.

Розділ 2 Розробка методу анотування україномовних художніх творів засобами машинного навчання

2.1 Модель анотування україномовних художніх творів

При здійсненні автоматичної обробки природно-мовних текстів виникають труднощі у процесі формалізації завдань. Наявність різних видів анотацій впливає на спосіб формалізації, фактичний результат роботи системи та змінює підхід до їх побудови. Загалом, автоматичне анотування передбачає для даного тексту T створення іншого тексту A (анотації), що містить короткий виклад основних питань, висвітлених у T [7].

Дискретна природа текстів дозволяє зручно розглядати їх як скінченні множини $T = \{t_1, t_2, \dots t_n\}$ і $A = \{a, a_2, \dots a_m\}$. Елементами множини T можуть бути різні лексичні одиниці (речення, абзаци, параграфи тощо) залежно від розміру тексту, а елементами множини A – лише речення (через обмеженість розміру тексту).

Основною проблемою в задачах автоматичного анотування є забезпечення того, щоб основний зміст тексту T відповідав змісту анотації A, а також знаходження цього основного змісту.

Для досягнення цих цілей використовуються різноманітні підходи та методи. Один з найпоширеніших підходів – використання методів машинного навчання для автоматичного визначення ключових речень або фраз, які найкраще описують основний зміст тексту.

Отже, наведено математичну модель процесу анотування текстової інформації при здійсненні автоматичної обробки природно-мовних текстів, що буде застосована до процесу анотування україномовних художніх творів.

2.2 Схема методу анотування україномовних художніх творів засобами машинного навчання

Схема та кроки методу анотування україномовних художніх творів засобами машинного навчання наведені на рисунку 2.1.

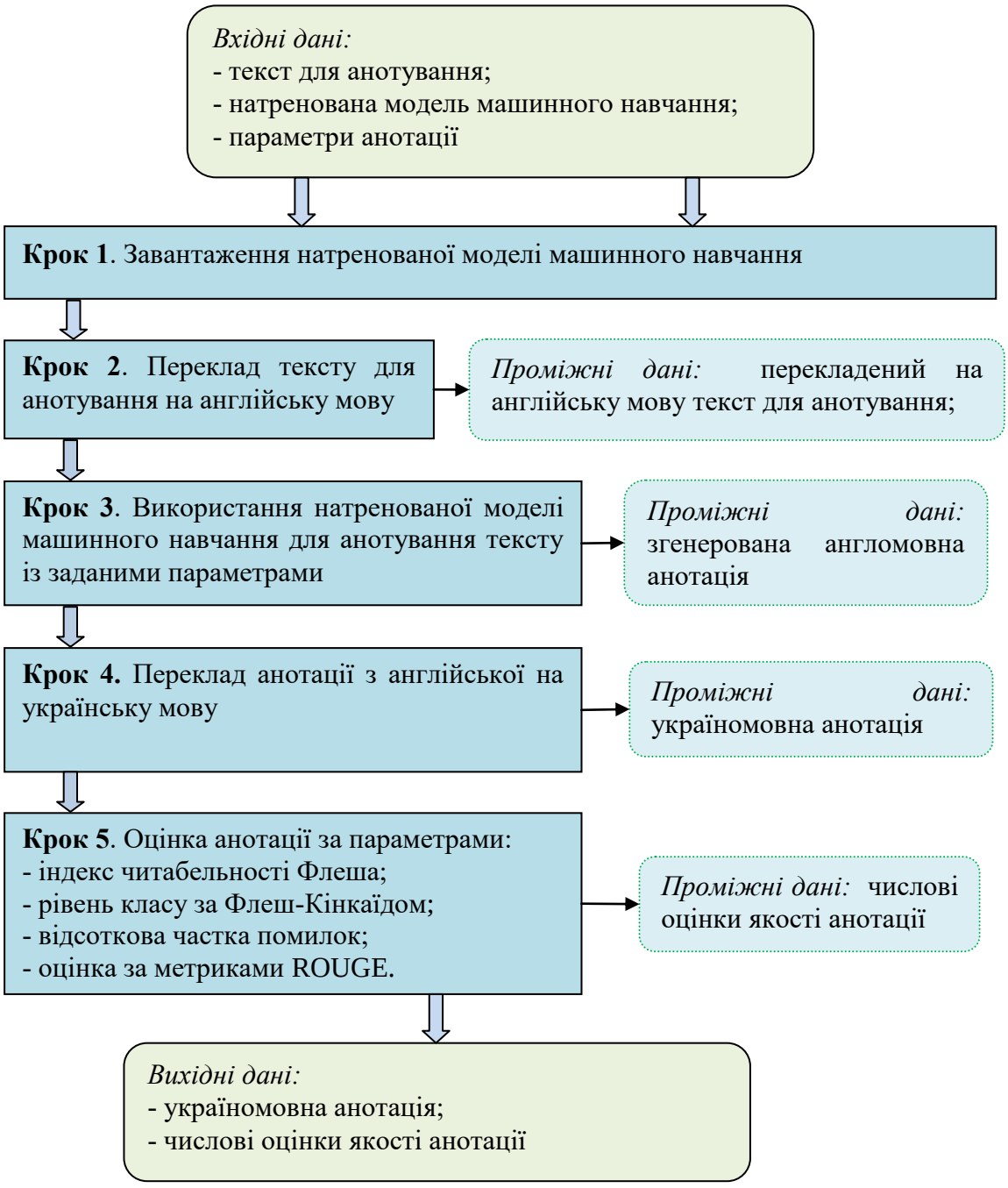


Рисунок 2.1 – Кроки методу анотування україномовних художніх творів засобами машинного навчання

Метод анотування україномовних художніх творів засобами машинного навчання призначений для автоматизованого створення анотацій, що у свою чергу спрощує процес опису змісту, ключових тем, персонажів та інших аспектів твору.

Відповідно, метод анотування україномовних художніх творів працює шляхом перетворення вхідних даних у вигляді тексту для анотування художнього твору, натренованої моделі машинного навчання та бажаних параметрів анотації у вихідні дані у вигляді україномовної анотації та числової оцінки якості анотації.

Першим кроком методу є завантаження натренованої моделі машинного навчання, яка є вхідними даними методу. Модель в базовому форматі є англомовною.

Наступним кроком є автоматизований переклад тексту для анотування на англійську мову, після якого утворюються проміжні дані у вигляді художнього твору на англійській мові.

Далі здійснюється крок використання натренованої моделі машинного навчання для анотування тексту із заданими параметрами. Параметрами є бажані розміри анотації (мінімальний та максимальний). Також особливістю виконання даного кроку є алгоритм розбиття тексту на частини, оскільки натренована модель має обмеження на вхідну кількість токенів, що становить 1024.

Тому береться весь обсяг тексту, та ділиться на фрагменти по реченнях розміром не більше 1024 (до крапки). В залежності від кількості отриманих частин, на кожну частину накладаються обмеження стосовно величини анотації. Результуюча анотація є комплексною, що складається із частин поданих на англійській мові.

Наступним кроком є переклад анотації з англійської на українську мову. Переклад відбувається аналогічним чином, як і з української на англійську.

Останнім кроком є оцінка анотації за параметрами індексу читабельності Флеша, рівеню класу за Флеш-Кінкаїдом, відсоткової частки помилок та оцінки за метриками ROUGE.

Вихідними даними запропонованого методу є отримана україномовна анотація, наближеної величини до заданих параметрів, а також числові оцінки якості анотації.

Ілюстрація проходження кроків методу наведена на рисунку 2.2.

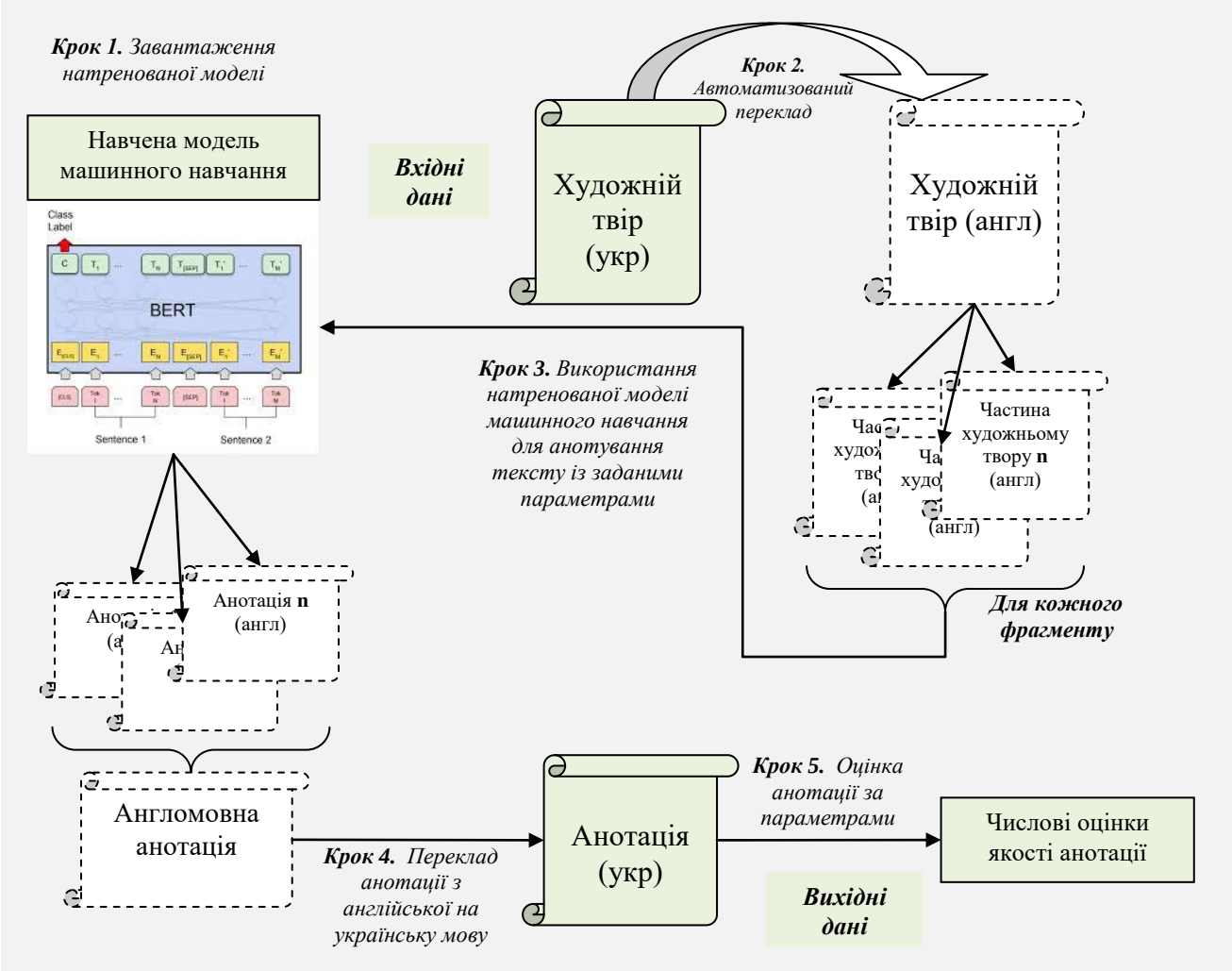


Рисунок 2.2 – Схема послідовності виконання кроків методу анотування україномовних художніх творів

Таким чином, було описано схему методу анотування україномовних художніх творів, що працює шляхом перетворення вхідних даних у вигляді тексту для анотування художнього твору, натренованої моделі машинного навчання та бажаних параметрів анотації у вихідні дані у вигляді україномовної анотації та числової оцінки якості анотації та призначений для автоматизованого створення анотацій.

2.3 Архітектура використаної моделі машинного навчання

Вхідними даними методу анотування україномовних художніх творів є навчена модель машинного навчання. Тому буде використано нейромережеву модель машинного навчання «Bart-large-cnn» [16], схематична архітектура якої наведена на рисунку 2.3. Повна архітектура використаної нейромережі наведена у додатку В.

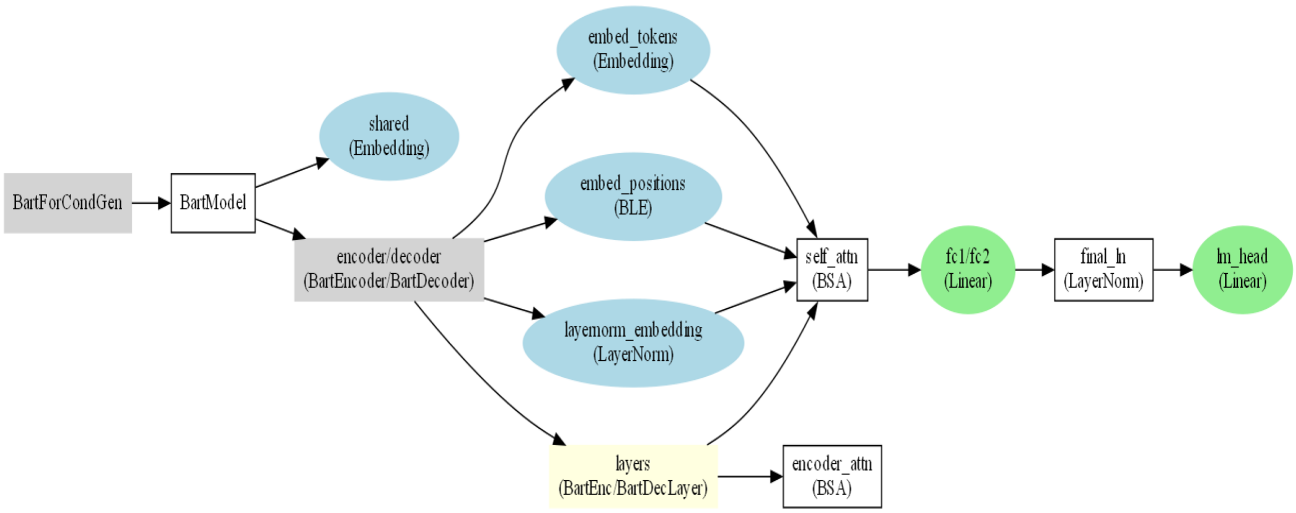


Рисунок 2.3 – Архітектура моделі машинного навчання «BART-Large-CNN» для анотування україномовних художніх творів

BART – це модель типу трасформер кодера-кодера (seq2seq) із двонаправленим (BERT-подібним) кодувальником і авторегресійним (GPT-подібним) декодером. BART попередньо навчається шляхом спотворення тексту за допомогою довільної функції шуму та навчання моделі для реконструкції вихідного тексту [17].

Ця нейромережева модель особливо ефективна, якщо налаштована для створення тексту, однак також добре працює для завдань розуміння (наприклад, класифікація тексту, відповідь на запитання). Дану модель машинного навчання було налаштовано на CNN Daily Mail, великій колекції пар текст-резюме.

Основні складові шарів моделі машинного навчання.

Шар типу shared (Embedding) використовується для кодування та декодування вхідного тексту, представлення слів у векторному просторі.

Шар типу encoder (BartEncoder) приймає вхідний текст та виконує його кодування. Включає в себе шари для обробки вхідного тексту та генерації його внутрішнього представлення.

Шар типу decoder (BartDecoder) використовує внутрішнє представлення вхідного тексту, отримане від енкодера, для генерації вихідного тексту. Включає в себе шари для генерації тексту з урахуванням контексту та попередньої частини згенерованого тексту.

Шар типу lm_head (Linear) є останнім шаром моделі, який використовується для передбачення наступного токена в тексті. Розмір вихідного простору цього шару відповідає кількості унікальних токенів у вхідному тексті.

Кожен з енкодерів та декодерів має 12 шарів, які включають в себе механізми уваги, лінійні трансформації та нормалізацію. Механізми уваги дозволяють моделі зосередитися на важливих частинах вхідного тексту під час кодування та декодування. Кожен шар також використовує функцію активації GELU для нелінійності. Розмір вбудовування (embedding) та внутрішній розмірності текстових представлень у цій моделі – 1024.

Отже, наведено нейромережеву архітектуру моделі машинного навчання, що є вхідними даними запропонованого методу анотування україномовних художніх творів. Дана нейромережева модель належить до типу «трансформери» та є на сьогодні однією із найпотужніших моделей генерації текстів.

2.4 Проектна архітектура системи та взаємозв’язок компонентів

Інформаційна система анотування україномовних художніх творів призначена для автоматичної обробки природно-мовних текстів з метою генерації анотації за вказаними параметрами, а також для оцінки її якості. Складається інтелектуальна система із бази даних, підсистеми оцінювання якості

анотацій художніх творів, підсистеми інтелектуальної генерації художньої анотації та електронної бібліотеки художніх творів. Проектна архітектура системи та взаємозв'язок компонентів наведено на рисунку 2.4.

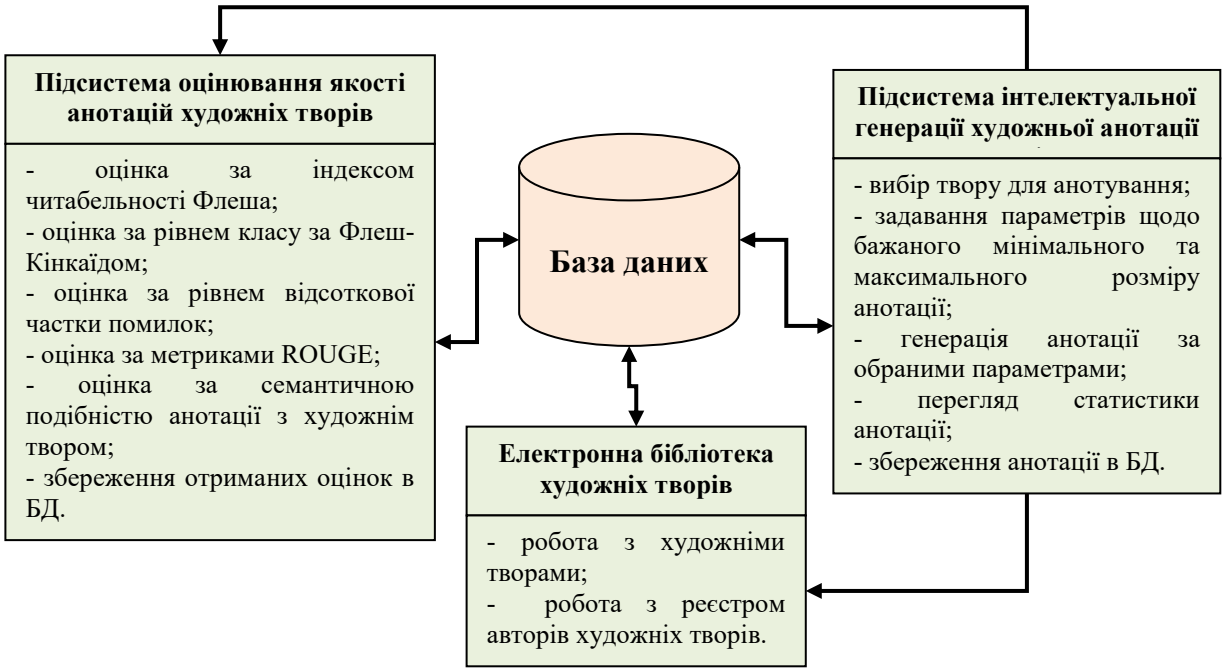


Рисунок 2.4 – Схема інформаційної системи анотування україномовних художніх творів

Підсистема інтелектуальної генерації художньої анотації є головною підсистемою, яка призначена безпосередньо для генерації анотації за вказаним розміром (або близьким до нього) та виконує такі основні функції, як: вибір твору для анотування, задавання параметрів щодо бажаного мінімального та максимального розміру анотації, генерація анотації за обраними параметрами, перегляд статистики анотації, збереження анотації в БД. Дана підсистема взаємодіє з підсистемою оцінювання якості анотацій художніх творів, а також з електронною бібліотекою художніх творів.

Підсистема оцінювання якості анотацій художніх творів є допоміжною підсистемою для підсистеми інтелектуальної генерації художньої анотації, яка дозволяє оцінити якість згенерованої анотації, та виконує такі функції: оцінка за індексом читабельності Флеша, оцінка за рівнем класу за Флеш-Кінкаїдом,

оцінка за рівнем відсоткової частки помилок, оцінка за метриками ROUGE, оцінка за семантичною подібністю анотації з художнім твором, збереження отриманих оцінок в БД.

Електронна бібліотека художніх творів призначена для перегляду та редагування художніх творів та інформації про авторів. В рамках роботи з художніми творами наявні функції перегляду обраного твору, редагування інформації по ньому, перехід на роботу з автором твору. Щодо роботи з реєстром авторів, присутні функції перегляду переліку художніх творів, деталі перегляду обраного художнього твору, додавання нового художнього твору. Дана підсистема є допоміжною для підсистеми інтелектуальної генерації художньої анотації.

База даних дозволяє організовано та впорядковано зберігати інформацію щодо роботи з художніми творами та анотаціями. База даних взаємодіє зі всіма підсистемами інформаційної системи.

Отже, було виконано проектування архітектури інтелектуальної системи анотування україномовних художніх творів, що призначена для автоматичної обробки природно-мовних текстів з метою генерації анотації за вказаними параметрами. Спроектована інтелектуальна система складається із бази даних, підсистеми оцінювання якості анотацій художніх творів, підсистеми інтелектуальної генерації художньої анотації та електронної бібліотеки художніх творів.

2.5 Проєктування бази даних програмної системи

Створення бази даних для методу анотування україномовних художніх творів засобами машинного навчання є важливим кроком у розвитку сучасних технологій обробки природної мови та збагачення україномовних цифрових ресурсів. Така база даних слугує фундаментом для розробки та вдосконалення алгоритмів, які можуть автоматично аналізувати, класифікувати та створювати анотації для художніх текстів.

Рисунок 2.5 – Даталогічна модель бази даних інтелектуальної системи анотування україномовних художніх творів

Таблиця 2.1 – Атрибути таблиці «Authors»

№ п/п	Назва	Тип даних	Опис
1.	ID	Integer	Первинний ключ, унікальний ідентифікатор запису таблиці
2.	Pseudonym	String	Псевдонім автора. Текстове поле, яке містить альтернативне ім'я, під яким автор відомий у літературному світія
3.	FirstName	String	Текстове поле, яке містить справжнє ім'я автора
4.	LastName	String	Текстове поле, яке містить справжнє прізвище автора
5.	Patronym	String	Текстове поле, яке містить по батькові автора, якщо воно використовується в культурі або країні автора.
6.	Photo	String	Поле містить шлях до файлу або URL зображення, що представляє автора
7.	FK_Country	Integer	Вторинний ключ до співставленням із відповідним записом таблиці «Countries»
8.	BirthDate	Date	Поле типу дата, яке вказує день, місяць і рік народження автора
9.	DeathDate	Date	Поле типу дата, яке вказує день, місяць і рік смерті автора, якщо він помер.

На рисунку 2.5 наведено даталогічну модель бази даних інтелектуальної системи анотування україномовних художніх творів.

Розроблена модель бази даних складається із 10 таблиць, у яких охоплено усі аспекти інтелектуальної системи анотування україномовних художніх творів.

Таблиця 2.2 – Атрибути таблиці «LiteraryWorks»

№	Назва	Тип даних	Опис
1.	ID	Integer	Первинний ключ, ID запису таблиці
2.	Title	String	Назва художнього твору
3.	Description	Text	Короткий опис твору
4.	Content	Memo	Повний текст твору
5.	FK_Author	Integer	Вторинний ключ до співставленням із відповідним записом таблиці «Authors»
6.	FK_Genre	Integer	Вторинний ключ до співставленням із відповідним записом таблиці «Genres»
7.	FK_Style	Integer	Вторинний ключ до співставленням із відповідним записом таблиці «Styles»
8.	FK_OriginalLanguage	Integer	Вторинний ключ до співставленням із відповідним записом таблиці «Languages»
9.	WritingDate	Date	Дата написання твору
10.	LoadingDate	Date	Дата завантаження твору в базу даних
11.	Accessibility	Boolean	Статус доступності твору
12.	WordCount	Integer	Кількість слів у творі
13.	NumberOfSymbols	Integer	Кількість символів у творі
14.	AVGnumberOf-WordsPerSentence	Integer	Середня кількість слів у реченні

Таблиця «Authors» (таблиця 2.1) створена для збереження інформацію про авторів художніх творів. Таблиця містить поля для збереження основної інформації: псевдоніму (або ПІБ), фото, країну походження автора, тощо.

Таблиця «LiteraryWorks» (таблиця 2.2) призначена для збереження інформації про художні твори. Ця таблиця містить інформацію про художні твори, включаючи їхні назви, описи, повний текст, авторів, жанри, стилі, дати написання та завантаження, доступність, кількість слів і символів, а також середню кількість слів у реченні.

Таблиця «Genres» (таблиця 2.3) створена для збереження інформації щодо літературних художніх жанрів та їх назв.

Таблиця 2.3 – Атрибути таблиці «Genres»

№ п/п	**Назва**	**Тип даних**	**Опис**
1.	ID	Integer	Первинний ключ, унікальний ідентифікатор запису таблиці
2.	name	String	Назва літературного художнього жанру

Таблиця «Styles» (таблиця 2.4) призначена задля збереження інформації щодо назв художніх літературних стилів.

Таблиця 2.4 – Атрибути таблиці «Styles»

№ п/п	**Назва**	**Тип даних**	**Опис**
1.	ID	Integer	Первинний ключ, унікальний ідентифікатор запису таблиці
2.	name	String	Назва літературного художнього стилю

Таблиця 2.5 – Атрибути таблиці «Styles»

№	Назва	Тип даних	Опис
1.	ID	Integer	Первинний ключ, унікальний ідентифікатор запису таблиці
2.	Name	String	Назва країни

Таблиця «Countries» (таблиця 2.5) зберігатиме назви країн походження авторів.

Таблиця «TypesOfVolumeEstimates» (таблиця 2.6) зберігатиме назви видів оцінок обсягу літературних художніх творів.

Таблиця 2.6 – Атрибути таблиці «TypesOfVolumeEstimates»

№	Назва	Тип даних	Опис
1.	ID	Integer	Первинний ключ, унікальний ідентифікатор запису таблиці
2.	name	String	Назва оцінки обсягу

Таблиця «AssessmentLevelsOfArtistiSkills» (таблиця 2.7) містить назви рівнів оцінки художньої майстерності тексту.

Таблиця 2.7 – Атрибути таблиці «AssessmentLevelsOfArtistiSkills»

№	Назва	Тип даних	Опис
1.	ID	Integer	Первинний ключ, унікальний ідентифікатор запису таблиці
2.	name	String	Назва рівня оцінки художньої майстерності тексту

Таблиця «ArtisticWorksAnnotations» (таблиця 2.8) призначена для збереження анотацій художніх творів. Таблиця зберігатиме наступні поля: посилання на відповідний художній твір, зміст анотації, дата й час генерації

анотації, максимальну та мінімальну кількість слів анотації, кількість слів анотації, кількість речень анотації, значення середньої довжини речення в словах та символах.

Таблиця 2.8 – Атрибути таблиці «ArtisticWorksAnnotations»

№ п/п	Назва	Тип даних	Опис
1.	ID	Integer	Первинний ключ, унікальний ідентифікатор запису таблиці
2.	FK_Article	Integer	Назва рівня оцінки художньої майстерності тексту
3.	GenerationTime	DateTime	Дата й час генерування анотації
4.	AnnotationMaxWords	Integer	Максимальна кількість слів в анотації
5.	AnnotationMinWords	Integer	Мінімальна кількість слів в анотації
6.	AnnotationMinWords	Integer	Кількість слів в анотації
7.	SentenceAmount	Integer	Кількість речень в анотації
8.	AVGSentenceWordLenght	Integer	Середня кількість слів в реченнях
9.	AVGSentenceCharLenght	Integer	Середня кількість символів в реченнях

Таблиця «SkillAssessmentResult» (таблиця 2.9) створена для збереження результатів оцінювання художньої майстерності тексту та містить поля для посилання на відповідну анотацію, запис назви рівня оцінки художньої майстерності тексту та значення індексу читабельності FRES.

Таблиця 2.9 – Атрибути таблиці «SkillAssessmentResult»

№ п/п	Назва	Тип даних	Опис
1.	ID	Integer	Первинний ключ, унікальний ідентифікатор запису таблиці
2.	FK_Article	Integer	Назва рівня оцінки художньої майстерності тексту
3.	FK_ LevelOfArtistiSkillst	Integer	Дата й час генерування анотації
4.	FRES	Integer	Максимальна кількість слів в анотації

Таблиця «AnnotationVolumeAssessmentResult» (таблиця 2.10) створена для збереження інформації щодо результатів оцінювання обсягів анотації та містить поля для посилання на відповідні записи щодо анотації та виду оцінки обсягу.

Таблиця 2.10 – Атрибути таблиці «AnnotationVolumeAssessmentResult»

№ п/п	Назва	Тип даних	Опис
1.	ID	Integer	Первинний ключ, унікальний ідентифікатор запису таблиці
2.	FK_Annotation	Integer	Вторинний ключ, посилання на відповідний запис таблиці «Annotations» для співставленням із відповідною анотацією
3.	FK_VolumeEstimateType	Integer	Вторинний ключ, посилання на відповідний запис таблиці «TypesOfVolumeEstimates» для співставленням із відповідною видом оцінки обсягу
4.	Rating	Integer	Отриманий показник

Таким чином, було реалізовано базу даних для інтелектуальної системи на базі методу анотування україномовних художніх творів засобами машинного навчання. Створено таблиці, налагоджено зв’язки між ними, БД заповнено базовими даними.

2.6 Особливості використання спеціалізованих програмних компонентів

Для спрощення процесу розробки програмного забезпечення, доцільно використовувати спеціалізовані програмні компоненти.

Бібліотека «Transformers» є Python-пакетом, що містить реалізації моделей-трансформерів з відкритим вихідним кодом для текстових, графічних та аудіо задач [18]. Ця бібліотека сумісна з основними фреймворками глибокого навчання, такими як PyTorch, TensorFlow і JAX, що дозволяє інтегрувати її в різні проєкти. У ній містяться реалізації популярних моделей, таких як BERT, GPT, T5 та інших. Спочатку бібліотека називалася «pytorch-pretrained-bert», потім була перейменована на «pytorch-transformers», і зрештою отримала назву «transformers». Завдяки своїй гнучкості та підтримці широкого спектру моделей, бібліотека «Transformers» стала основним інструментом для дослідників і розробників у сфері обробки природної мови та інших задач, що потребують потужних моделей трансформерів. В роботі планується її використання для роботи з моделлю машинного навчання для генерації англомовної анотації.

Бібліотека «Translate» це простий, але потужний інструмент перекладу, написаний мовою Python із підтримкою кількох постачальників перекладу [19]. Наразі він пропонує інтеграцію з Microsoft Translation API, Translated MyMemory API, LibreTranslate та безкоштовними та професійними API DeepL. Дану бібліотеку буде використано для перекладу тектсу художнього твору на англійську, та для перекладу результатів анотації назад на українську.

«LanguageTool» є інструментом перевірки граматики з відкритим вихідним кодом, відомий також як засіб перевірки правопису для OpenOffice

[20]. Ця бібліотека дозволяє виявляти граматичні та орфографічні помилки за допомогою сценарію на Python або через інтерфейс командного рядка. За замовчуванням, бібліотека language_tool_python завантажує сервер LanguageTool у форматі .jar і запускає його у фоновому режимі для локальної перевірки граматики. Окрім локальної перевірки, «LanguageTool» пропонує публічний HTTP API для перевірки тексту, хоча існують обмеження на кількість запитів через цей API. Бібліотека підтримує кілька мов, серед яких є українська, і надає можливості для налаштування правил перевірки, що робить її гнучким інструментом для поліпшення якості тексту в різних додатках і системах. Буде використано з метою перевірки граматики автоматично згенерованої анотації.

Бібліотека «ROUGE» для Python є інструментом, що дозволяє автоматично оцінювати якість текстових узагальнень та перекладів за допомогою метрики ROUGE [21]. Ця метрика орієнтована на повноту та точність порівняння результатів автоматичного узагальнення або перекладу з еталонними зведеннями, створеними людиною. Використовуючи різні варіанти n-грамів (одноgрам, біграм, триграм тощо), метрика ROUGE дозволяє оцінити, наскільки добре автоматично згенеровані тексти відповідають референсним текстам, аналізуючи збіги n-грамів між ними.

Основна функціональність бібліотеки включає обчислення ROUGE-N, ROUGE-L, ROUGE-W та інших варіантів метрики, де кожен з них враховує певні аспекти текстової подібності. ROUGE-N вимірює збіги n-грамів, ROUGE-L враховує довжину найдовшої спільної підпослідовності, а ROUGE-W зважує такі підпослідовності за допомогою їхньої довжини. Ця гнучкість дозволяє користувачам налаштовувати оцінку відповідно до специфічних потреб їхніх досліджень.

Бібліотека «ROUGE» для Python забезпечує точні та надійні методи оцінки, що робить її важливим інструментом для дослідників і розробників у галузі обробки природної мови. Дана бібліотека буде використана для оцінки якості згенерованої анотації.

«SentenceTransformer» є бібліотекою для Python, призначеною для отримання семантично значущих представлень текстових даних шляхом використання моделей-трансформерів [22]. Розроблена для обробки різноманітних завдань обробки природної мови (NLP), таких як семантичний пошук, класифікація текстів, кластеризація, ідентифікація дублюючих питань, ідентифікація текстових подібностей, ця бібліотека дозволяє ефективно конвертувати текстові дані у векторні представлення.

Основою «SentenceTransformer» є попередньо навчені моделі трансформерів, такі як BERT, RoBERTa, та інші. Ці моделі навчені на великих корпусах текстів і здатні вловлювати складні семантичні зв’язки між словами та реченнями. Бібліотека дозволяє тонко налаштовувати ці моделі для специфічних завдань користувача за допомогою додаткового навчання на спеціалізованих наборах даних.

Однією з ключових особливостей «SentenceTransformer» є можливість створення високоякісних ембеддінгів речень, які будуть використовуватись для порівняння семантичної подібності згенерованої анотації та базового художнього твору.

Бібліотека «PyQt» представляє собою потужний інструмент для розробки графічних інтерфейсів, що поєднує гнучкість і зручність Python з багатими можливостями та стабільністю інструментарію Qt [23]. Це робить її хорошим засобом для розробників, які прагнуть створювати якісні та функціональні програмні продукти.

Основою «PyQt» є об'єктно-орієнтовані принципи, що дозволяють створювати та керувати графічними елементами, такими як вікна, кнопки, текстові поля та інші компоненти GUI. Бібліотека забезпечує підтримку різноманітних функцій, включаючи обробку подій, інтеграцію з базами даних, розробку мережевих застосунків і обробку мультимедіа. Це робить її потужним інструментом для створення як простих утиліт, так і складних програмних комплексів.

«PyQt» також відзначається високим рівнем абстракції, що дозволяє зосередитися на логіці застосунку, не витрачаючи багато часу на реалізацію низькорівневих деталей. Бібліотека включає в себе розширені можливості стилізації інтерфейсу, що дозволяє створювати естетично привабливі та інтуїтивно зрозумілі інтерфейси, саме з такою метою вона буде використана у кваліфікаційній роботі бакалавра.

Отже, в роботі планується використати бібліотеку «Transformers» для роботи з моделлю машинного навчання для генерації англомовної анотації, бібліотеку «Translate» для перекладу тектсу художнього твору на англійську, та для перекладу результатів анотації назад на українську, бібліотеку «LanguageTool» з метою перевірки граматики автоматично згенерованої анотації, «ROUGE» для оцінки якості згенерованої анотації, «SentenceTransformer» буде використовуватись для порівняння семантичної подібності згенерованої анотації та базового художнього твору, а «PyQt» буде використано для побудови інтерфейсу користувача.

2.8 Висновки до розділу 2

В рамках написання другого розділу наведено математичну модель процесу анотовування текстової інформації при здійсненні автоматичної обробки природно-мовних текстів, а також створено метод анотування україномовних художніх творів засобами машинного навчання.

Описано схему методу анотування україномовних художніх творів, що працює шляхом перетворення вхідних даних у вигляді тексту для анотування художнього твору, натренованої моделі машинного навчання та бажаних параметрів анотації у вихідні дані у вигляді україномовної анотації та числової оцінки якості анотації та призначений для автоматизованого створення анотацій.

Наведено нейромережеву архітектуру моделі машинного навчання, що є вхідними даними запропонованого методу анотування україномовних художніх

творів. Дана нейромережева модель належить до типу «трансформери» та є на сьогодні однією із найпотужніших моделей генерації текстів.

Виконано проектування архітектури інтелектуальної системи анотування україномовних художніх творів, що призначена для автоматичної обробки природно-мовних текстів з метою генерації анотації за вказаними параметрами та є прямою реалізацією розробленого методі. Спроектована інтелектуальна система складається із бази даних, підсистеми оцінювання якості анотацій художніх творів, підсистеми інтелектуальної генерації художньої анотації та електронної бібліотеки художніх творів.

Спроектовано базу даних, що є складовою частиною інтелектуальної системи генерації анотації.

Обрано спеціалізовані програмні компоненти, вирішено в роботі планується використати бібліотеку «Transformers» для роботи з моделлю машинного навчання для генерації англомовної анотації, бібліотеку «Translate» для перекладу тектсу художнього твору на англійську, та для перекладу результатів анотації назад на українську, бібліотеку «LanguageTool» з метою перевірки граматики автоматично згенерованої анотації, «ROUGE» для оцінки якості згенерованої анотації, «SentenceTransformer» буде використовуватись для порівняння семантичної подібності згенерованої анотації та базового художнього твору, а «PyQt» буде використано для побудови інтерфейсу користувача.

За розробленим методом та спроектованою архітектурою інтелектуальної системи анотування україномовних художніх творів необхідно реалізувати застосунок, за допомогою якого буде досліджено ефективність розробленого методу. Розроблений застосунок також необхідно протестувати на предмет коректного виконання функцій.

Розділ 3 Експериментальне дослідження методу анотування україномовних художніх творів

3.1 Визначення шляхів дослідження та засобів створення інтелектуальної системи анотування україномовних художніх творів

Для реалізації спроєктованої інтелектуальної системи анотування україномовних художніх творів буде створено віконний застосунок, який буде виконувати такі основні функції:

– генерація анотації за вказаними параметрами;

– перегляд статистики анотації;

– оцінка анотації за: індексом читабельності Флеша, рівнем класу за Флеш-Кінкаїдом, рівнем відсоткової частки помилок, метриками ROUGE, семантичною подібністю анотації з художнім твором;

– робота з художніми творами;

– робота з реєстром авторів художніх творів;

– збереження отриманих оцінок в БД;

– збереження анотації в БД.

Зазначені функції потрібно протестувати з використанням тест-кейсів та засобами функціонального тестування.

Якість згенерованої анотації доцільно досліджувати з використанням метрик:

– індекс читабельності Флеша;

– рівень класу за Флеш-Кінкаїдом;

– відсоткова частка помилок;

– оцінка за метриками ROUGE;

– оцінка семантичної подібності.

Індекс Флеша, також відомий як індекс легкості читання Флеша (Flesch Reading Ease), є метрикою для оцінки доступності тексту, що базується на структурних характеристиках тексту, таких як довжина речень і кількість складів у словах [24]. Він розраховується за допомогою формули, що враховує середню

кількість слів у реченні та середню кількість складів на слово. Результат виражається числом на шкалі від 0 до 100, де вищі значення вказують на легший для читання текст. Наприклад, текст з високим індексом Флеша підходить для широкої аудиторії, включаючи дітей та людей з базовим рівнем освіти, тоді як нижчі значення вказують на більш складний текст, який може вимагати вищого рівня освіти для розуміння.

Індекс Флеша-Кінкаїда (Flesch-Kincaid Grade Level) є оцінкою складності тексту, що вказує на рівень освіти, необхідний для розуміння тексту. Він заснований на тих самих принципах, що й індекс легкості читання Флеша, але результат виражається у вигляді освітнього рівня у системі освіти США. Формула для розрахунку включає середню довжину речень (у словах) та середню кількість складів на слово. Наприклад, оцінка 8.0 вказує на те, що текст зрозумілий для учнів восьмого класу. Цей показник широко використовується в освіті, видавничій діяльності та інших галузях для визначення придатності текстів для цільової аудиторії.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) є оціночною метрикою, широко використовуваною для оцінки якості завдань обробки природної мови, таких як узагальнення тексту та машинний переклад [25]. На відміну від BLEU, ROUGE використовує як метрики точності, так і метрики повноти (відкликання), щоб порівняти результати зведення, відомі як кандидати, з еталонними зведеннями, створеними людиною, відомими як посилання. Ця метрика вимірює, яка частка n-грамів із посилань присутня в прогнозованих кандидатах.

Обчислення повноти (відкликання) здійснюється за формулою:

$$Recall = True\ Positives / (True\ Positives + False\ Negatives)$$

Тут *True Positives* (справжні позитивні результати) – це n-грами, що збігаються між посиланнями та кандидатами. *False Negatives* (хибні негативні результати) можна розглядати як n-грами, які присутні у посиланнях, але

відсутні у кандидатах. *False Positives* (хибні позитивні результати) – це n-грами, що містяться в кандидатах, але відсутні у посиланнях.

Таким чином, ROUGE оцінює ступінь відповідності між автоматично згенерованими текстами та еталонними текстами, надаючи змогу якісно оцінити ефективність моделей узагальнення або перекладу.

Отже, наведено шляхи дослідження запропонованого методу, окреслено основні функції майбутньої програмної реалізації та запропоновано засоби для перевірки коректного виконання зазначених функцій.

3.2 Вибір засобів розробки інформаційної системи

Для програмної реалізації інтелектуальної системи буде використано середовище програмування PyCharm, мову програмування Python, СКБД SQLite та мову запитів SQL.

PyCharm є інтегрованим середовищем розробки для мови програмування Python, розроблене компанією JetBrains [26]. Відоме своєю потужною функціональністю та багатим набором інструментів, PyCharm значно спрощує процес написання, налагодження та тестування коду. Підтримує широкий спектр функцій, включаючи автодоповнення коду, рефакторинг, інтеграцію з системами контролю версій (наприклад, Git), та потужний дебагер.

Однією з ключових переваг PyCharm є його здатність забезпечувати зручне і ефективне середовище для розробки, особливо для великих проєктів. PyCharm підтримує різноманітні веб-фреймворки, такі як Django, Flask та Pyramid, а також надає інструменти для роботи з базами даних, інтеграцію з інструментами для тестування та підтримку наукових обчислень (наприклад, через інтеграцію з Jupyter Notebook). Це робить його універсальним інструментом для розробників, що працюють у різних галузях від веброзробки до наукових досліджень.

Python є високорівневою мовою програмування загального призначення, створена Гвідо ван Россумом і випущена в 1991 році. Вона відома своєю

простотою та зрозумілістю синтаксису, що дозволяє розробникам швидко і ефективно писати читабельний код. Python підтримує кілька парадигм програмування, включаючи об'єктно-орієнтоване, процедурне і функціональне програмування. Завдяки динамічній типізації та автоматичному управлінню пам'яттю, Python забезпечує високу продуктивність розробки.

Python широко використовується в різних галузях, включаючи веб-розробку, наукові дослідження, аналіз даних, штучний інтелект і машинне навчання. Мова має велику стандартну бібліотеку та активну спільноту, що сприяє розвитку та підтримці численних сторонніх пакетів і фреймворків, таких як Django для веб-розробки, NumPy і SciPy для наукових обчислень, та TensorFlow для машинного навчання. Ці характеристики роблять Python універсальним інструментом, який підходить як для початківців, так і для досвідчених розробників.

SQLite є легкою вбудованою реляційною системою керування базами даних, яка широко використовується завдяки своїй простоті та ефективності [27]. Вона не потребує налаштування серверного програмного забезпечення, оскільки всі дані зберігаються в одному файлі на диску, що робить її ідеальним вибором для мобільних додатків, вбудованих систем і невеликих проєктів. SQLite підтримує більшість стандартних SQL-команд, що дозволяє розробникам використовувати знайомий синтаксис для маніпуляції даними.

Основною перевагою SQLite є її відповідність ACID-властивостям (атомарність, узгодженість, ізольованість та довговічність), що забезпечує надійність зберігання даних. Завдяки цим характеристикам, а також можливості інтеграції з багатьма мовами програмування, SQLite знайшла своє застосування в багатьох відомих програмах і системах, таких як браузер Firefox, поштовий клієнт Thunderbird, медіацентр VLC, а також в операційних системах Android та iOS.

Отже, для розробки програмного забезпечення буде використано такий набір засобів: середовище програмування PyCharm, мова програмування Python, СКБД SQLite та мова запитів SQL.

3.3 Структура та функціональне призначення програмних складових інтелектуальної системи анотування україномовних художніх творів

Структура програмних складових інтелектуальної системи анотування україномовних художніх творів у вигляді діаграми класів наведено на рисунку 3.1.

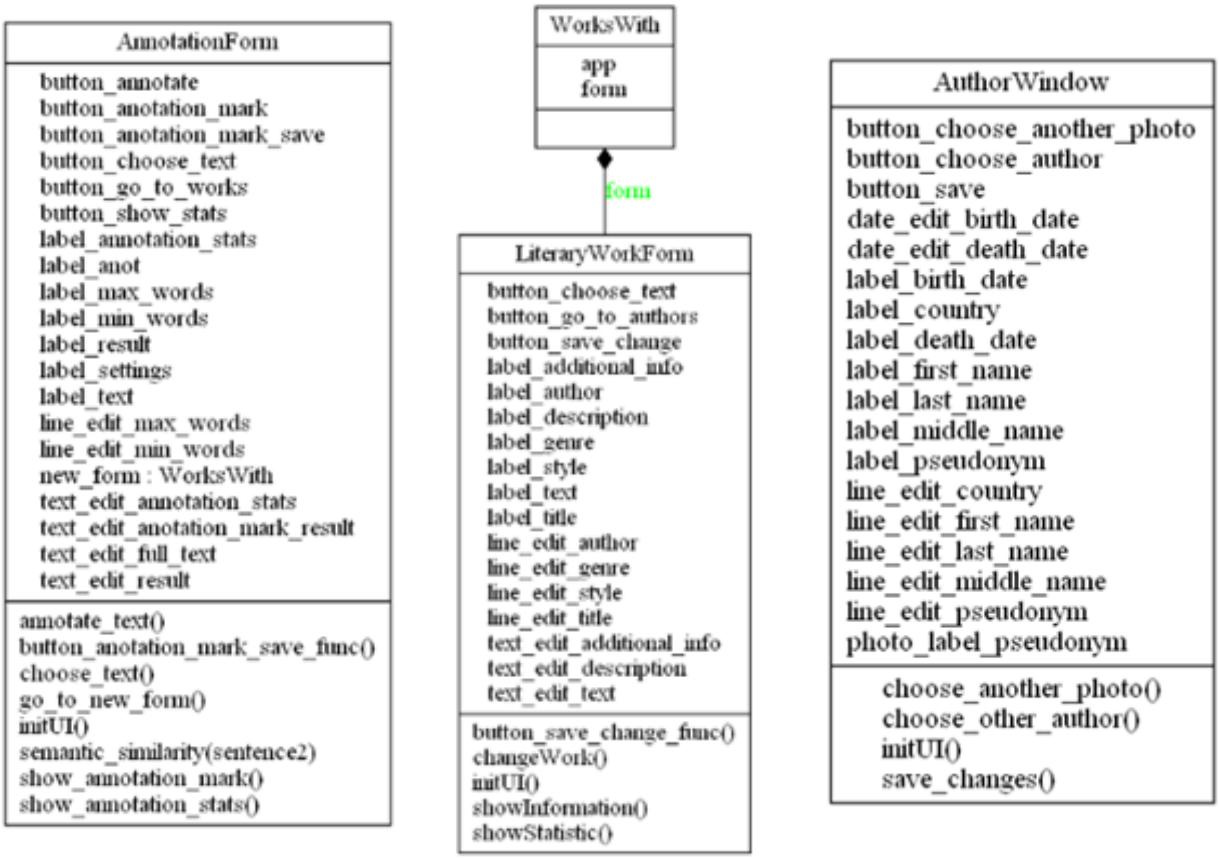


Рисунок 3.1 – Діаграма класів інтелектуальної системи анотування україномовних художніх творів

Клас «AnnotationForm» є головним класом програмного застосунку, який реалізовує функціональність 2-х підсистем: підсистеми інтелектуальної генерації художньої анотації та підсистеми оцінювання якості анотацій художніх творів.

Метод initUI() ініціалізує користувацький інтерфейс, додає всі елементи до макету та задає їх властивості. Метод annotate_text() виконує анотування

тексту, використовуючи модель BART. Спочатку перекладає вхідний текст на англійську, створює анотацію, а потім перекладає результат назад на українську. Метод show_annotation_stats() обчислює та відображає статистику для згенерованої анотації, включаючи кількість слів та речень, середню довжину речень. Метод choose_text() дозволяє вибрати інший текст для анотування. Метод show_annotation_mark() оцінює якість анотації, використовуючи індекси читабельності Флеша-Кінкаїда та кількість граматичних помилок за допомогою language_tool_python. Також обчислює метрику ROUGE для оцінки відповідності анотації вихідному тексту. Метод semantic_similarity() обчислює семантичну подібність між двома реченнями, використовуючи WordNet.

У свою чергу класи «AuthorWindow» та «LiteraryWorkForm» реалізовують функціонал підсистеми «Електронна бібліотека художніх творів». Їх функціонал призначений для роботи з авторами та художніми творами.

Метод initUI(self) ініціалізує інтерфейс користувача, створюючи всі необхідні віджети та компоновки. Встановлює заголовок вікна, додає кнопки, текстові поля, мітки, та поля вибору дати. Підключає кнопки до відповідних методів для обробки подій натискання. Метод choose_other_author(self) обробляє натискання кнопки «Обрати іншого автора». Метод choose_another_photo(self) обробляє натискання кнопки «Обрати інше фото». Метод save_changes(self) обробляє натискання кнопки «Зберегти зміни». Збирає дані з полів вводу (псевдонім, прізвище, ім’я, по батькові, країна, дата народження, дата смерті) та записує їх у БД.

Отже, таким чином наведено діаграму класів програмного застосунку та описано основні призначення програмних складових інтелектуальної системи анотування україномовних художніх творів.

3.4 Особливості реалізації програмних складових системи

Після створення структури та опису основних програмних частин здійснюється їх програмна реалізація. Найважливішою частиною

інтелектуальної системи анотування україномовних художніх творів є генерація анотації на базі обраного художнього твору.

Метод annotate_text() призначений саме для генерації анотації, і його реалізація здійснюється таким чином: спершу здійснюється отримання тексту для анотування, команда toPlainText() витягує текст з віджета QTextEdit, де користувач вводить повний текст твору, або текст твору вибирається із БД. Наступним етапом відбувається переклад тексту з української на англійську мову, для чого використовується функція translate з параметрами української ("uk") на англійську ("en"). Далі шукається довжина тексту, і якщо довжина перевищує 1024 токени, текст розділяється на менші фрагменти.

Наступним кроком є токенізація та підготовка до анотування за допомогою BART. Текст перекладається в формат, зрозумілий для моделі BART. До тексту додається префікс "summarize: ", щоб модель розуміла, що потрібно зробити анотацію. Метод tokenizer.encode() перетворює текст в токени та повертає тензор PyTorch (return_tensors="pt"), обмежуючи максимальну довжину тексту до 1024 токенів (max_length=1024) і обрізаючи його, якщо він перевищує цей ліміт. Далі для кожного текстового фрагменту відбувається генерація анотації, модель генерує анотацію з заданими параметрами: max_length (максимальна довжина анотації), min_length (мінімальна довжина анотації), length_penalty (штраф за довжину для контролю довжини вихідного тексту), num_beams=4 (використовується 4 промені в пошуку променів для кращої генерації тексту) та з параметром early_stopping=True для раньої зупинки генерації, коли всі промені досягнуть кінця. Далі відбувається декодування анотації з токенів у текст, пропускаючи спеціальні токени (skip_special_tokens=True).

Після чого здійснюється переклад англійської анотації на українську мову для кожного згенерованого фрагменту, після чого анотація відображається у віджеті QTextEdit, який показує результат.

Також в цьому методі передбачено виклик методу для показу статистики анотації. Після генерування та відображення анотації викликається метод

show_annotation_stats, який обчислює та показує статистику для анотації (його логіка передбачає обчислення кількості слів, речень, середньої довжини речень тощо).

Результат виконання методу генерації анотації наведено на рисунку 3.2.

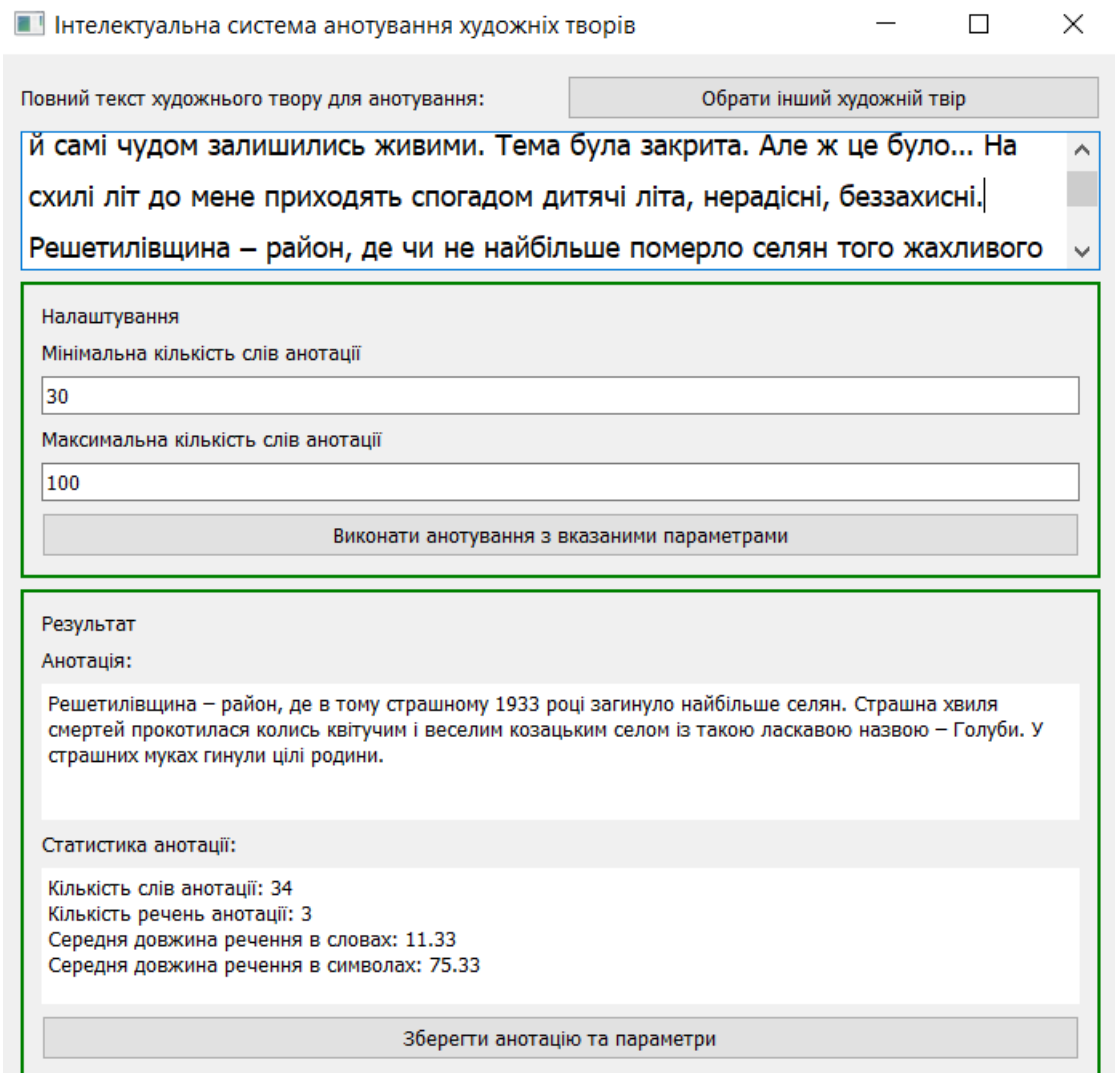


Рисунок 3.2 – Приклад генерації анотації

Метод show_annotation_mark() виконує комплексну оцінку анотації, включаючи читабельність, граматичні помилки та схожість з повним текстом, і виводить результати у відповідне текстове поле.

Першим кроком в роботі методу є отримання текстів анотації та повного тексту: annotation_text витягується текст з віджета text_edit_result, де міститься

згенерована анотація, а full_Text витягується з віджета text_edit_full_text, де міститься повний текст твору.

Наступним кроком є оцінка читабельності тексту анотації readability_score – розраховуєтьсяя індекс читабельності Флеша для тексту анотації. Цей індекс визначає, наскільки легким для читання є текст, а також за допомогою flesch_kincaid_grade розраховується рівень читабельності за шкалою Флеша-Кінкаїда, який відповідає класу шкільного рівня, необхідному для розуміння тексту. Обраховані дані виводяться до віджета text_edit_anotation_mark_result, який відображає результати оцінки анотації.

Наступним кроком є перевірка граматики з використанням бібліотеки LanguageTool. Виводиться загальна кількість знайдених помилок, а також для кожної помилки виводиться її ідентифікатор (ruleId), опис (message) та пропозиції щодо виправлення (replacements).

Наступним кроком здійснюється оцінка за метрикою ROUGE для оцінки схожості текстів. Обчислюються оцінки ROUGE між анотацією та повним текстом. Результати обчислень додаються до віджета text_edit_anotation_mark_result.

Результат виконання методу оцінки згенерованої анотації наведено на рисунку 3.3.

І останнім кроком здійснюється пошук семантичної схожості з використанням моделі SentenceTransformer, яка призначена для генерації векторних подань текстів. Завантажується модель paraphrase-multilingual-MiniLM-L12-v2, яка навчена на багатьох мовах та призначена для вирішення задач парафразування. Векторне представлення текстів дозволяє порівнювати їх семантичну схожість за допомогою косинусної схожості. Використовуючи метод cos_sim() з модуля util бібліотеки sentence_transformers, обчислюється косинусна схожість між векторами, які представляють вхідний текст та анотований текст.

Значення косинусної схожості, виводиться на екран. Косинусна схожість близька до 1 вказує на високу схожість текстів, тоді як значення близьке до 0 вказує на низьку схожість.

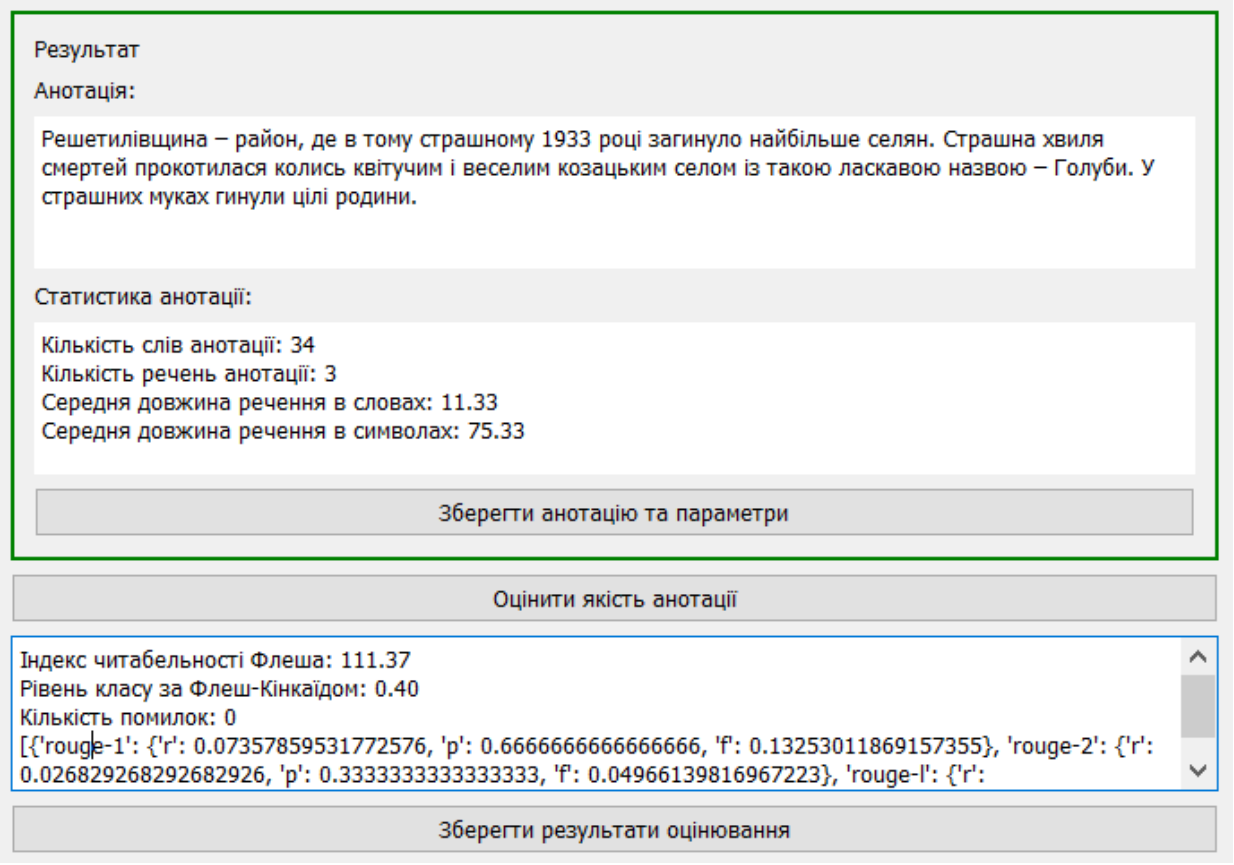


Рисунок 3.3 – Приклад оцінки якості анотації

Отже, описано основні моменти реалізації складових програмного застосунку інтелектуального анотування україномовних художніх творів.

3.5 Тестування інформаційної системи та вимоги до розгортання

Для перевірки наскільки ефективно може програмний продукт виконувати задані в постановці задачі функції буде задіяно тест-кейси. Перший тест-кейс (таблиця 3.1) призначений для перевірки генерації анотації підсистеми підсистеми інтелектуальної генерації художньої анотації.

Наступним тестовим випадком є перевірка виведення статистики для згенерованої анотації. Кроки тестового випадку наведені у таблиці 3.2

Таблиця 3.1 – Тест-кейс А-001

Тест-кейс ID: А-001	**Пріоритет: 1**	**Створено:25.04.2024**
Назва: Перевірка генерації анотації		
Кроки		**Очікуваний результат**
1. Запустити програмний застосунок; 2. Вставити досліджуваний текст для генерації анотації; 3. У полях для мінімальної та максимальної довжини анотації встановити 30 та 100 відповідно. 4. Натиснути кнопку «Виконати анотування з вказаними параметрами».		1. Запущено застосунок 2. Текст відображено у текстовому полі 3. Величина мінімальної та максимальної довжини уведена у текстові поля 4. Виконано генерацію анотації
Результат виконання тест-кейсу: пройдено успішно		

Результат успішної генерації анотації підтверджено на рисунку 3.4.

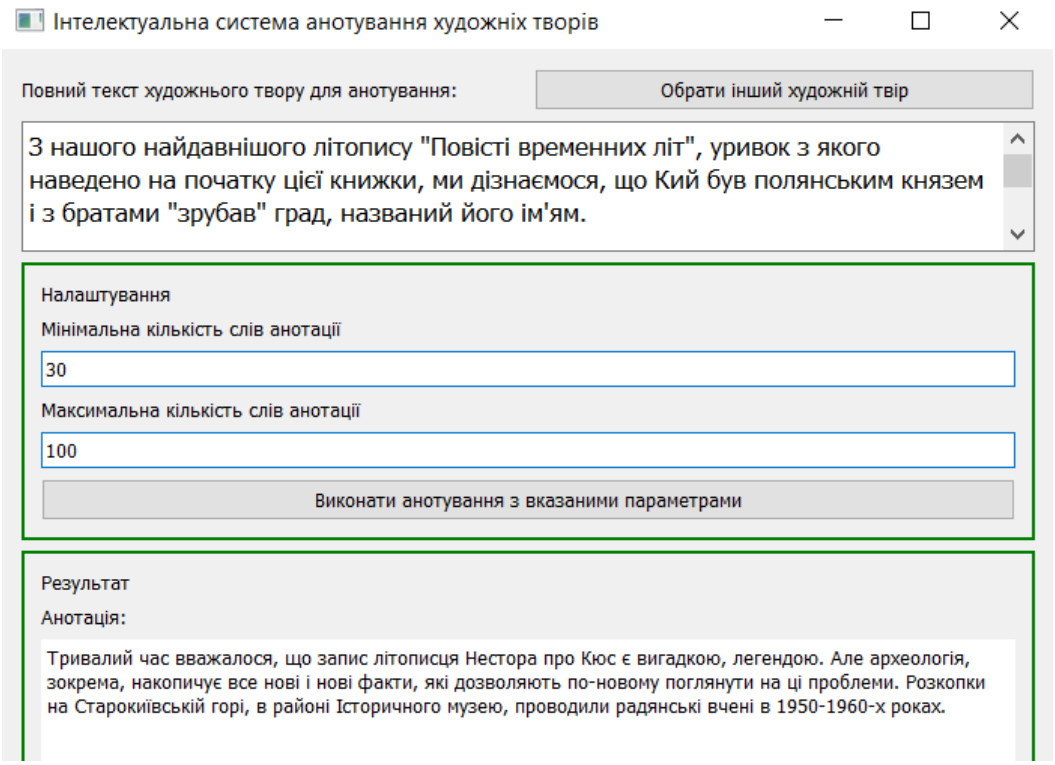
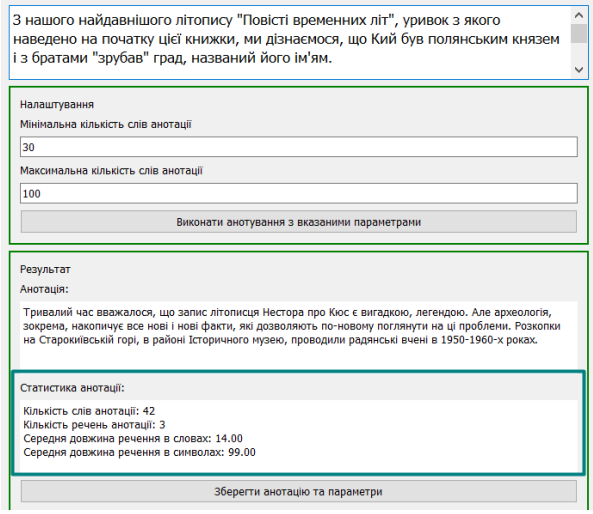


Рисунок 3.4 – Генерація анотації за вказаними параметрами

Таблиця 3.2 – Тест-кейс А-002

Тест-кейс ID: А-002	**Пріоритет: 1**	**Створено:27.04.2024**
Назва: Перевірка виведення статистики для згенерованої анотації		
Кроки		**Очікуваний результат**
1. Запустити програмний застосунок; 2. Вставити досліджуваний текст для генерації анотації; 3. У полях для мінімальної та максимальної довжини анотації встановити 30 та 100 відповідно. 4. Натиснути кнопку «Виконати анотування з вказаними параметрами». 5. Переглянути статистику анотації		1. Запущено застосунок 2. Текст відображено у текстовому полі 3. Величина мінімальної та максимальної довжини уведена у текстові поля 4. Виконано генерацію анотації 5. Виведено стаистику
Результат виконання тест-кейсу: пройдено успішно		



Рисунко 3.4 – Статистика анотації

Результат успішного виконання тестового випадку інтелектуальної системи анотування україномовних художніх творів наведено на рисунку 3.5.

Наступним тестовим випадком буде перевірка оцінки якості для згенерованої анотації. Кроки тестового випадку наведені у таблиці 3.3

Таблиця 3.3 – Тест-кейс А-003

Тест-кейс ID: А-003	**Пріоритет: 1**	**Створено:27.04.2024**
Назва: Перевірка оцінки якості для згенерованої анотації		
Кроки		**Очікуваний результат**
1. Запустити програмний застосунок; 2. Вставити досліджуваний текст для генерації анотації; 3. У полях для мінімальної та максимальної довжини анотації встановити 30 та 100 відповідно. 4. Натиснути кнопку «Виконати анотування з вказаними параметрами». 5. Переглянути статистику анотації. 6. Натиснути кнопку «Оцінити якість анотації»		1. Запущено застосунок 2. Текст відображено у текстовому полі 3. Величина мінімальної та максимальної довжини уведена у текстові поля 4. Виконано генерацію анотації 5. Виведено статистику. 6. Виведено оцінки якості анотації.
Результат виконання тест-кейсу: пройдено успішно		

Результат успішного виконання тестового випадку А-003 інтелектуальної системи анотування україномовних художніх творів наведено на рисунку 3.5.

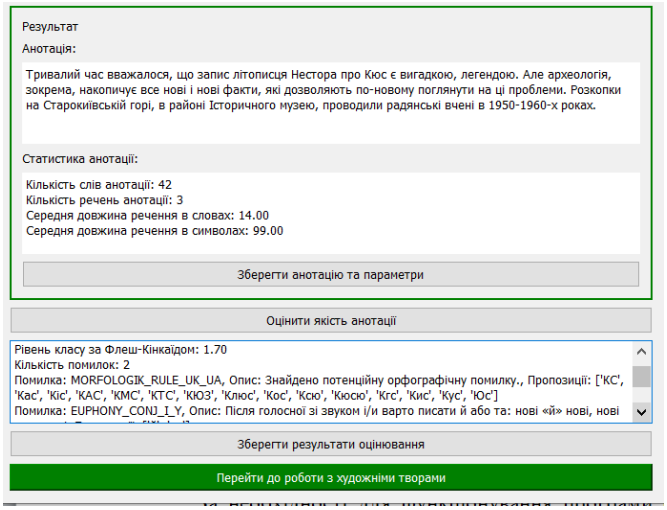


Рисунок 3.5 – Результат виконання тест-кейсу А-003

Отже, з проведеного тестування непрацюючих функцій не виявлено. Весь заявлений функціонал працює згідно до поставлених завдань та описаних функцій.

3.6 Аналіз функціональності системи

Наступним етапом є проведення аналізу функціональності інтелектуальної системи анотування україномовних художніх творів з метою розкрити можливий потенціал її використання.

Після запуску інтелектуальної системи анотування україномовних художніх творів відкриється головний екран з підсистемою «Інтелектуальної генерації художньої анотації». Вигляд головного екрану після запуску наведено на рисунку 3.6.

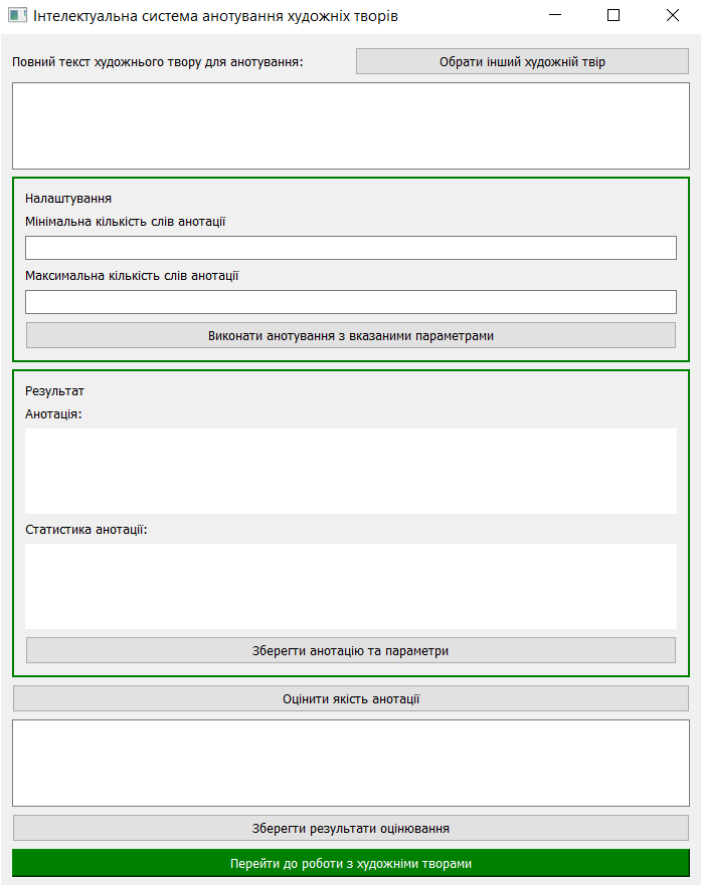


Рисунок 3.6 – Вигляд головного екрану інтелектуальної системи анотування україномовних художніх творів

Для обирання твору, що міститься в базі даних, необхідно натиснути кнопку «Обрати інший художній твір». Також можна вставити текст для формування анотації у текстове поле «Повний текст художнього твору для анотування». Наступним етапом для можливості сформувати анотацію є вказування параметрів для анотації (мінімальний та максимальний бажаний розмір). Якщо ці параметри не вказати, анотація буде створена з параметрами за замовчуванням (50 - 150). Для генерації анотації необхідно натиснути кнопку «Виконати анотування з вказаними параметрами». Результат виконання анотування наведено на рисунку 3.7.

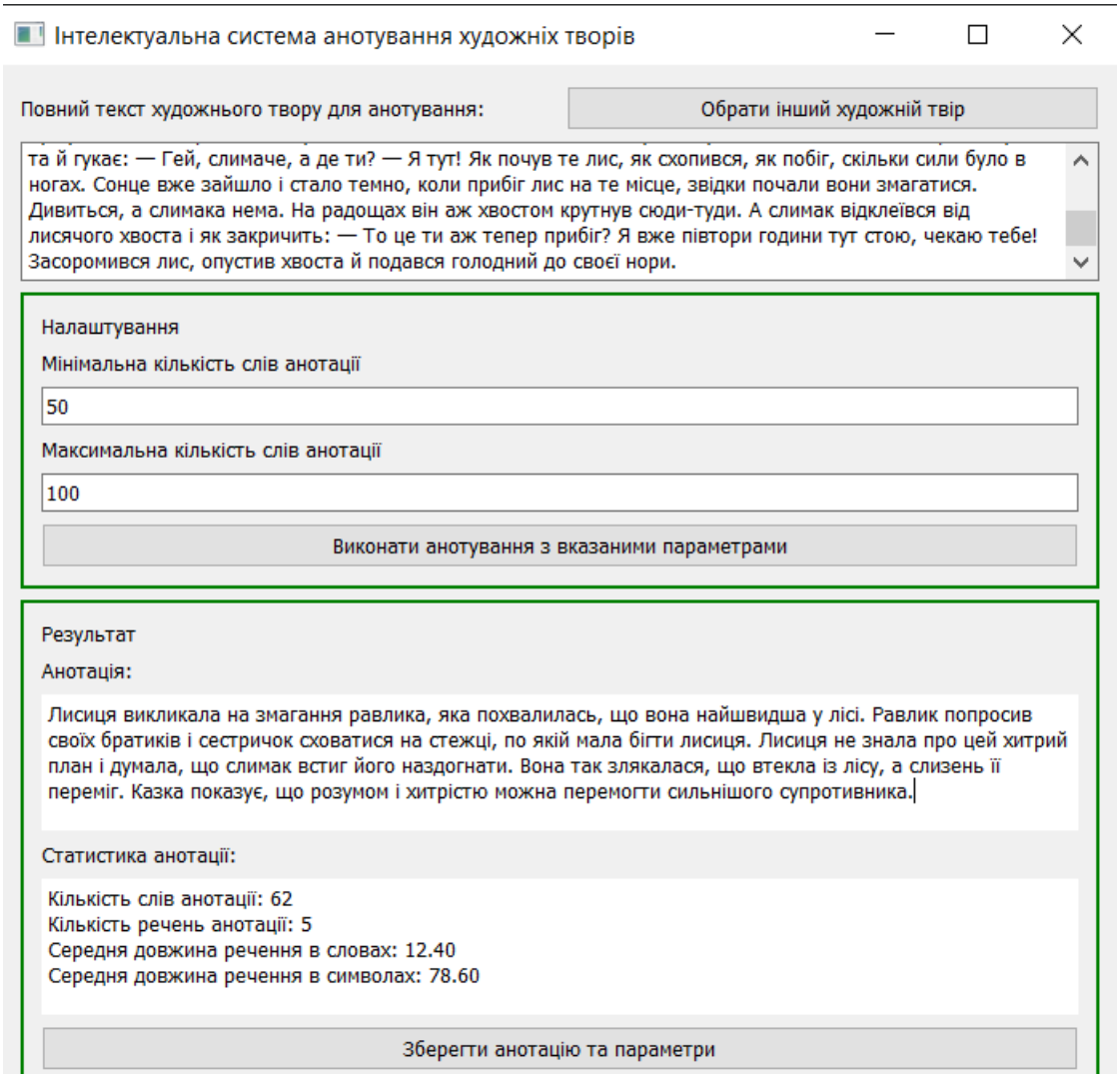


Рисунок 3.7 – Виконана анотація художнього твору

Разом з анотацією буде виведено її статистику у полі «Статистика анотації:». Статистика анотації містить дані про кількість слів в анотації, кількість речень, а також дані про середню довжину речення в словах та середню довжину речення в символах. Згенеровану анотацію разом із параметрами можна зберегти до бази даних, написнувши на кнопку «Зберегти анотацію та параметри». Отриману анотацію можна оцінити за різноманітними метриками, натиснувши на кнопку «Оцінити якість анотації». Приклад оцінки якості згенерованої анотації наведено на рисунку 3.8.

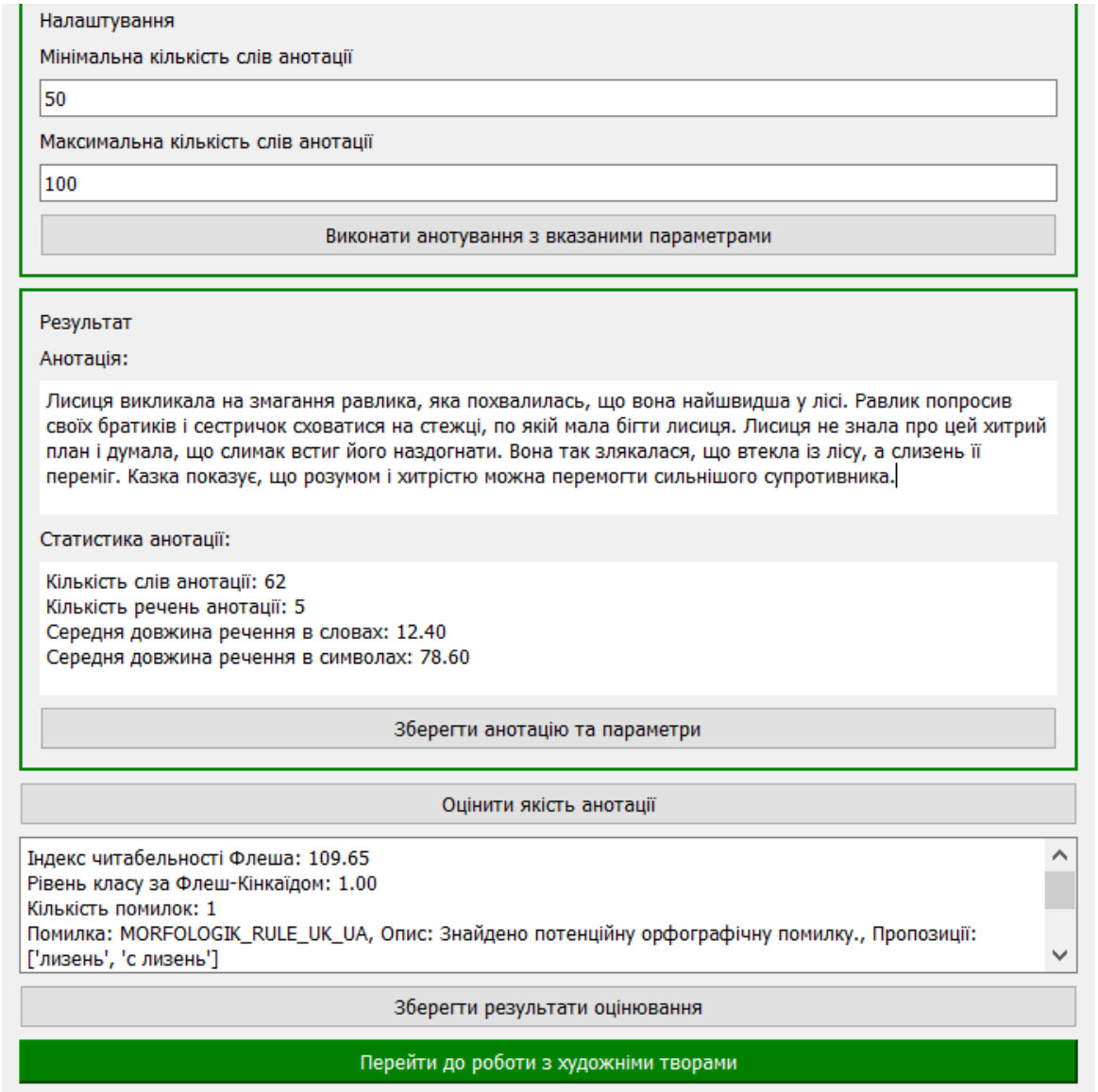


Рисунок 3.8 – Згенерована та оцінена за метриками анотація

Результати оцінювання також є можливість зберегти до бази даних, для цього необхідно натиснути на кнопку «Зберегти результати оцінювання». З головної підсистеми є можливість переходу на підсистему роботи з художніми творами. Для переходу до підсистеми роботи з художніми творами потрібно натиснути кнопку «Перейти до роботи з художніми творами».

Відкриється головне вікно підсистеми роботи з художніми творами інтелектуальної системи анотування україномовних художніх творів, і якщо даний досліджуваний художній твір є у база даних, про нього виведеться інформація. Якщо ж даний твір не знаходиться у базі даних, то при переході в підсистемі буде відображено тільки сам текст твору. Вигляд підсистеми наведено на рисунку 3.9.

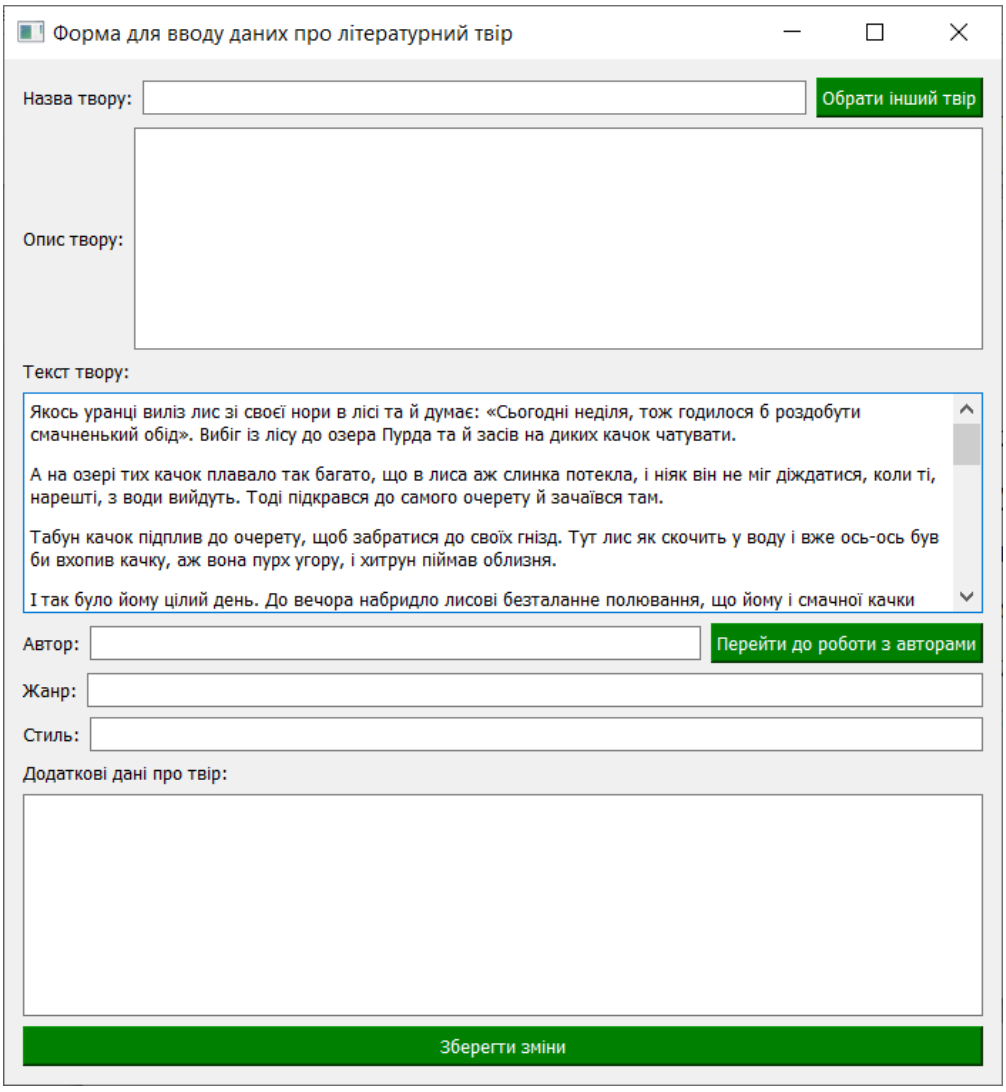


Рисунок 3.9 – Робота з художніми творами

Можна внести дані, та зберегти інформацію про цей твір в базі даних, натиснувши кнопку «Зберегти зміни». Також на цій підсистемі є можливість обрати з творів, які уже є в базі даних, для цього необхідно натиснути на кнопку «Обрати інший твір».

При виборі твору, який є у базі даних буде відображена вся інформація про нього. Приклад перегляду інформації про твір що міститься у базі наведено на рисунку 3.10.

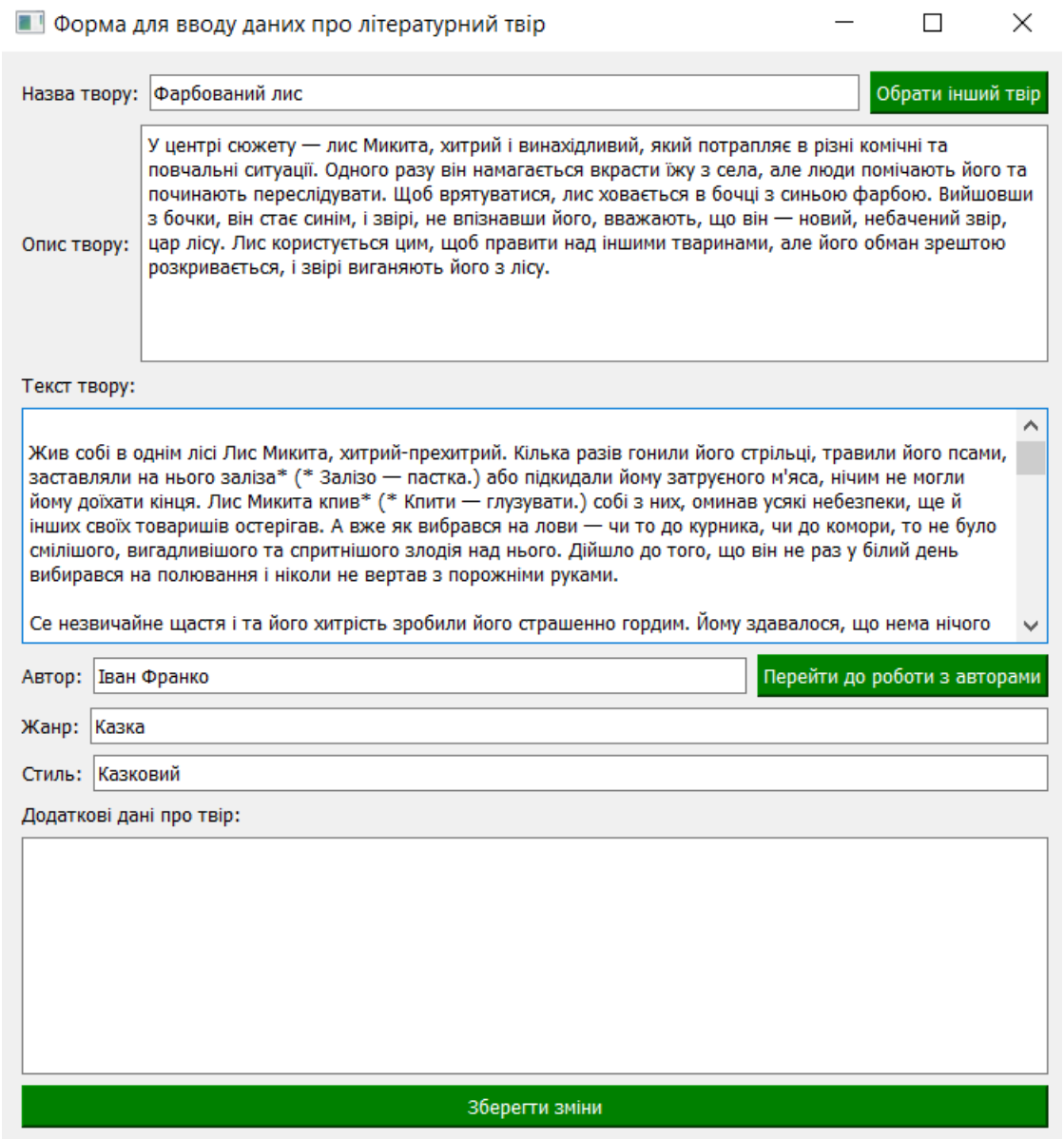


Рисунок 3.10 – Перегляд інформації про художній твір з БД

Як видно з рисунку 3.10 – про даний твір немає додаткових даних. Їх можна додати та відповідно натиснути на кнопку «Зберегти зміни». При таких діях дані про твір будуть збережені у базі даних. Також з цієї форми можна перейти на форму роботи з авторами (рисунок 3.11).

При переході на форму роботи з авторами буде підтягнуто автора, з яким твір якого переглядався в підсистемі роботи з літературними творами.

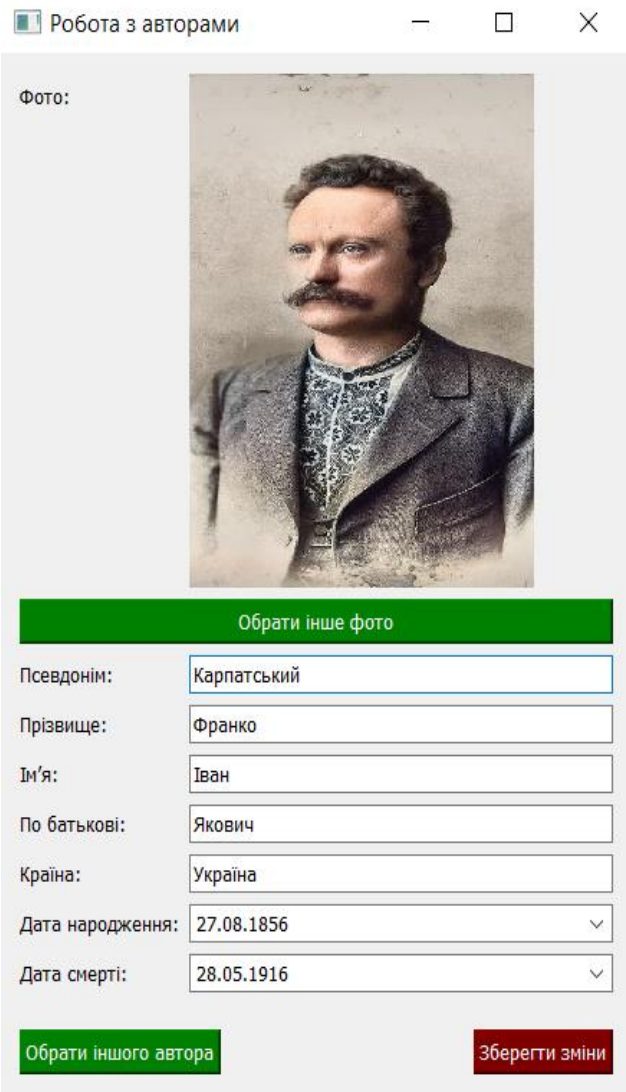


Рисунок 3.11 – Підсистема роботи з авторами

Для зміни фото необхідно натиснути на кнопку «Обрати інше фото». Відкриється діалогове вікно, з якого можна обрати інше фото з локального диска. Результат зміни фото автора наведено на рисунку 3.12.

Для внесення змін у дату народження чи дату смерті необхідно натиснути на календар в полі «Дата народження» або «Дата смерті» (рисунок 3.13).

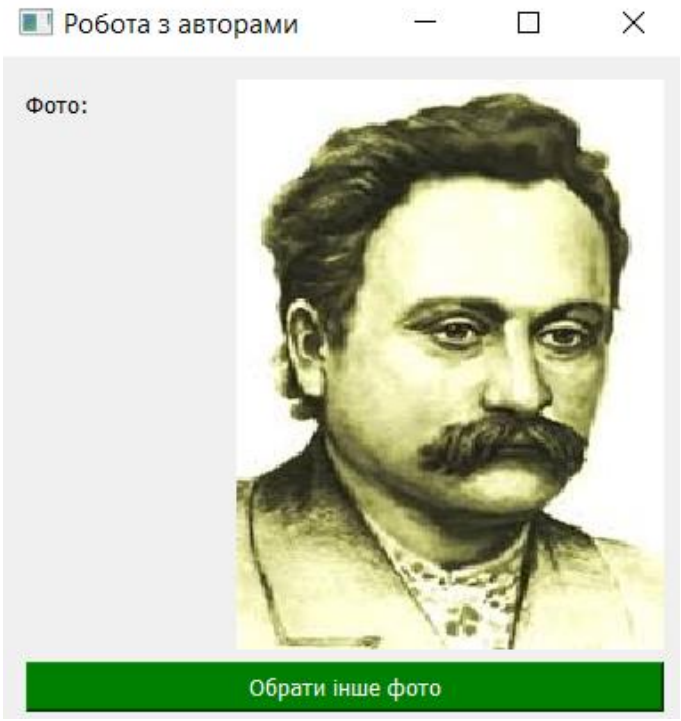


Рисунок 3.12 – Зміна фото автора

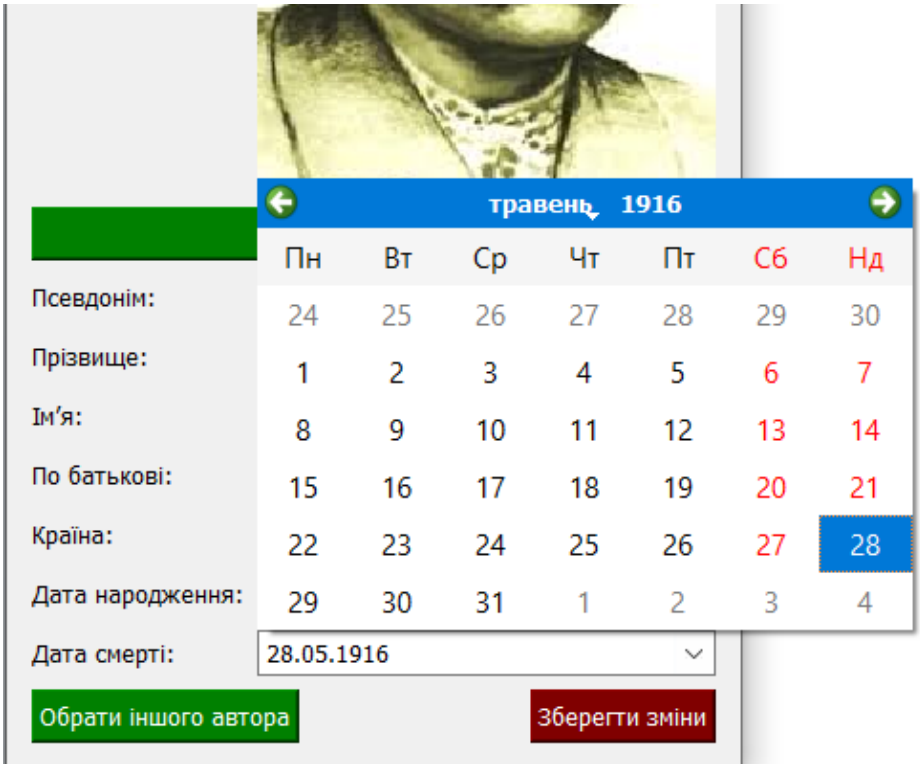


Рисунок 3.13 – Зміна дати смерті

Для збереження змін відомостей про автора необхідно натиснути кнопку «Зберегти зміни». Також у даній підсистемі інтелектуальної системи анотування україномовних художніх творів є можливість обрати іншого автора для роботи. Для цього необхідно натиснути кнопку «Обрати іншого автора».

Таким чином було виконано аналіз функціональності інтелектуальної системи анотування україномовних художніх творів засобами машинного навчання.

3.7 Результати досліджень

Для дослідження ефективності методу анотування україномовних художніх творів засобами машинного навчання буде використано створену програмну реалізацію у вигляді інтелектуальної системи анотування україномовних художніх творів. Також буде виконано порівняння з результатами мовної моделі GPT-3.5. У таблиці 3.4 наведено порівняння моделі GPT-3.5 з розробленим методом за семантичною подібністю згенерованих анотацій.

Таблиця 3.4 – Порівняння моделі GPT-3.5 з розробленим методом за семантичною подібністю

	Оцінка семантичної подібності для створеної моделлю GPT-3.5 анотації з оригіналом	Оцінка семантичної подібності для створеної запропонованим методом анотації з оригіналом
Анотація 1	0.48	0.75
Анотація 2	0.36	0.37
Анотація 3	0.75	0.68
Анотація 4	0.54	0.72
Анотація 5	0.5	0.63

Дані експерименту таблиці 3.4 також проілюстровано на діаграмі 3.14.

Рисунок 3.14 – Результати експерименту за семантичною подібністю

Дані в таблиці 3.4 та наведені на графіку 3.14 свідчать про те, що більша семантична подібність у анотацій, згенерованих запропонованим методом.

Також було проведено експериментальне дослідження щодо оцінки близькості анотації до заданих користувацьких параметрів за довжиною (параметри від 50 до 100).

Таблиця 3.5 – Порівняння моделі GPT-3.5 з розробленим методом за довжиною

	Оцінка довжини анотації моделлю GPT-3.5	Оцінка довжини анотації запропонованим методом
Анотація 1	54	62
Анотація 2	53	60
Анотація 3	42	56
Анотація 4	46	63
Анотація 5	59	58

Дані експерименту проілюстровано на рисунку 3.15.

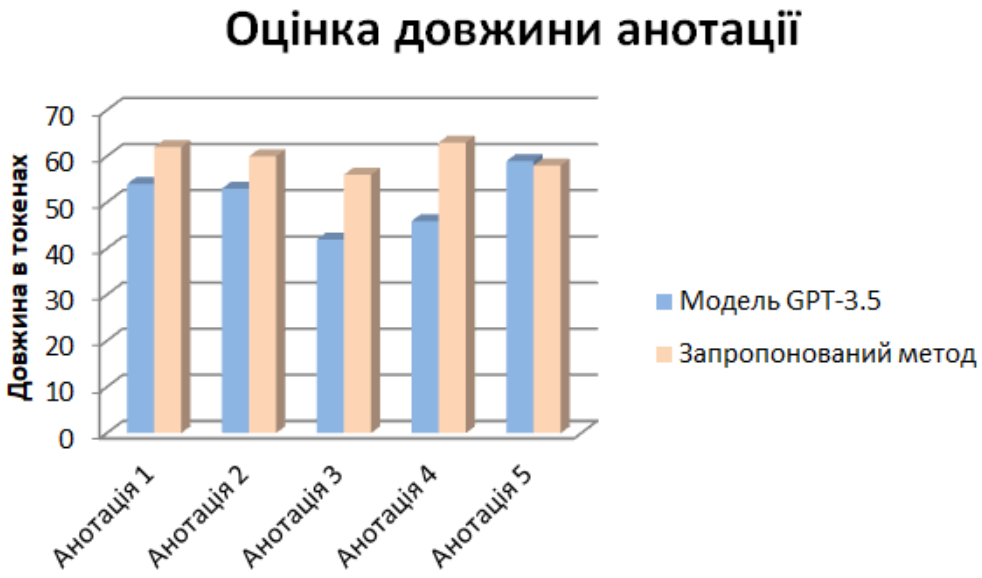


Рисунок 3.15 – Результат експерименту по довжині згенерованих анотацій

Таблиця 3.6 – Порівняння моделі GPT-3.5 з розробленим методом за метрикою ROUGE по уніграмам

	Оцінка анотації згенерованої моделлю GPT-3.5 за ROUGE	Оцінка анотації згенерованої запропонованим методом за ROUGE
Анотація 1	Recall: 0.042 Precision: 0.286 F1 Score: 0.073	Recall: 0.0755 Precision: 0.4167 F1-score: 0.1279
Анотація 2	Recall: 0.098 Precision: 0.44 F1 Score: 0.16	Recall: 0.17 Precision: 0.391 F1-score: 0.237
Анотація 3	Recall: 0.122 Precision: 0.372 F1 Score: 0.184	Recall: 0.243 Precision: 0.457 F1-score: 0.317
Анотація 4	Recall: 0.17 Precision: 0.347 F1 Score: 0.228	Recall: 0.21 Precision: 0.443 F1-score: 0.285
Анотація 5	Recall: 0.23 Precision: 0.287 F1 Score: 0.255	Recall: 0.176 Precision: 0.386 F1-score: 0.242

Дані експерименту свідчать про можливість обох моделей не відхилятись значною мірою від бажаного діапазону довжини згенерованих анотацій, однак розроблений метод показав кращий результат, згенерувавши всі анотації не виходячи з діапазону.

Далі було проведено експериментальне дослідження щодо оцінки анотації за метрикою ROUGE. У таблиці 3.6 наведено статистику порівняння моделі GPT-3.5 з розробленим методом за метрикою ROUGE по уніграмам.

Дані з таблиці 3.6 проілюстровано на рисунку 3.16

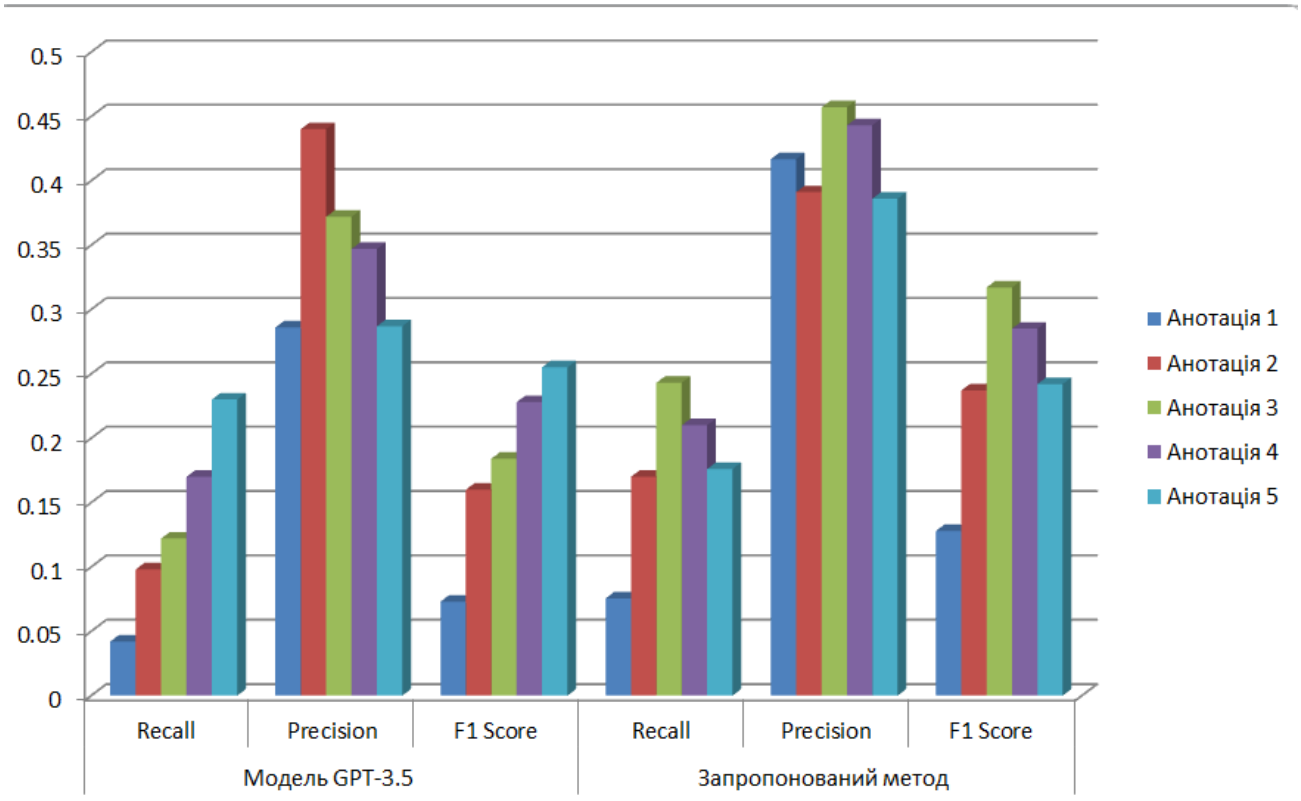


Рисунок 3.16 – Порівняння моделі GPT-3.5 з розробленим методом за метрикою ROUGE по уніграмам

Показник Recall є відношенням кількості унікальних співпадаючих одиничних (уніграмних) слов у згенерованому тексті до кількості уніграм у референсному тексті. Чим вище ця оцінка, тим більше згенерований текст покриває референтний.

Показник Precision показує відношення кількості унікальних співпадаючих одиничних слів у згенерованому тексті до загальної кількості

уніграм у згенерованому тексті. Висока точність вказує на те, що згенерований текст використовує більше унікальних слів, а не повторює одні й ті самі.

F1-score є середнім гармонійним значенням recall та precision і є загальною оцінкою якості збігу.

Зважаючи на отримані дані, обидва підходи спроможні генерувати якісні анотації. Розроблений метод за метриками ROUGE є семантично-ближчим до оригіналу та в одночас все ж таки відрізняється, що говорить про можливість розробленого методу генерувати новий текст, замінюючи деякі слова на оригінальні.

Що стосується оцінок за метриками індекс читабельності Флеша та рівню класу за Флеш-Кінкаїдом, то ці оцінки недоцільно використовувати для оцінок якості анотації, якщо її обсяг складає до 100 токенів.

Щодо показника помилок – він наближено ідентичний як у моделі GPT-3.5, так і у розробленому методі. Однак, GPT-3.5 має обмеження стосовно вхідної довжини оригінального тексту, а розроблений метод може приймати текст довільної довжини.

Отже, було проведено дослідження ефективності розробленого методу, що показало хороші результати за всіма досліджуваними параметрами. Розроблений метод формує анотації семантично-ближчі до оригіналу та дозволяє застосовувати користувацькі параметри у вигляді діапазону бажаної мінімальної та максимальної довжини.

3.8 Висновки до розділу 3

Отже, в рамках виконання третього розділу, наведено шляхи дослідження запропонованого методу, окреслено основні функції майбутньої інтелектуальної системи анотування україномовних художніх творів, а також запропоновано засоби для перевірки коректного виконання зазначених функцій.

Обрано засоби розробки інтелектуальної системи анотування україномовних художніх творів, буде використано такий набір засобів:

середовище програмування PyCharm, мова програмування Python, СКБД SQLite та мова запитів SQL.

Наведено діаграму класів програмного застосунку та описано основні призначення програмних складових інтелектуальної системи анотування україномовних художніх творів. Також описано основні моменти реалізації складових інтелектуальної системи анотування україномовних художніх творів.

Розроблене програмне забезпечення у вигляді інтелектуальної системи анотування україномовних художніх творів протестовано, з проведеного тестування непрацюючих функцій не виявлено. Весь заявлений функціонал інтелектуальної системи анотування україномовних художніх творів працює згідно до поставлених завдань та описаних функцій.

Виконано аналіз функціональності інтелектуальної системи анотування україномовних художніх творів, а також проведено дослідження розробленого методу з використанням розробленої програмної реалізації.

Загальні висновки

Метою кваліфікаційної роботи бакалавра було спрощення створення анотацій україномовних художніх творів шляхом автоматизації анотування засобами машинного навчання, для чого виконувалась розробка методу анотування україномовних художніх творів засобами машинного навчання. Також ставилась задача створити відповідну програмну реалізацію у вигляді віконного застосунку, що призначений для автоматизованого анотування україномовних художніх творів за користувацьким текстом, а також були сформулювані та вирішені завдання дослідження:

– досліджено предметну область анотування художніх творів засобами машинного навчання;

– виконано огляд теоретичних підходів до вирішення подібних задач, обрано підхід до автоматизованого анотування художніх творів серед методів машинного навчання, а саме використання нейромережевої моделі BARD;

– створено метод анотування україномовних художніх творів засобами машинного навчання;

– створено інформаційну структуру системи автоматизованого анотування користувацьких художніх текстів;

– виконано програмну реалізацію інформаційної системи;

– проведено тестування інформаційної системи автоматизованого анотування користувацьких художніх текстів;

– виконано дослідження ефективності методу анотування україномовних художніх творів засобами машинного навчання.

Обидва підходи демонструють здатність генерувати якісні анотації, але розроблений метод за метриками ROUGE показує семантичну близькість до оригіналу, одночасно зберігаючи свою унікальність. Це свідчить про можливість методу створювати новий текст, зберігаючи основний зміст і використовуючи оригінальні елементи. Оцінки за читабельністю та складністю тексту за індексами Флеша та Флеш-Кінкаїда непридатні для оцінки якості анотацій, коли

їх обсяг обмежений до 100 токенів. Щодо показника помилок, він майже ідентичний як у моделі GPT-3.5, так і у розробленому методі, проте розроблений метод може працювати з текстами будь-якої довжини, що є його перевагою порівняно з обмеженнями GPT-3.5.

Результат, отриманий в ході розробки КРБ, цілком відповідає поставленому завданню.

Наукові та практичні результати роботи публікувалися у матеріалах XXVII Міжнародної науково-практичної конференції «Prospects of Scientific Research in the Conditions of the Modern World» (12-14 червня 2024 року, Rotterdam, Netherlands), за темою кваліфікаційної роботи бакалавра автором виконано наукову публікацію «Practice Implementation of Neural Network Model BART-Large-CNN for Text Annotation» [28].

Перелік посилань

1. Алгоритмічно-програмний метод автоматизованого генерування зведених відгуків інтернет-користувачів. URL: https://ela.kpi.ua/bitstream/123456789/51088/1/Kazymyrov_magistr.pdf

2. Анотування – це що таке? Правила та методики, приклади. URL: https://presa.com.ua/navchannia/anotuvannya-tse-shcho-take-pravila-ta-metodiki-prikladi.html

3. Анотування і реферування наукових текстів. URL: https://studfile.net/preview/9812438/page:5/

4. Анотування і реферування наукових текстів. URL: https://pidru4niki.com/1513061640666/dokumentoznavstvo/anotuvannya_referuvannya_naukovih_tekstiv.

5. Анотування і реферування як вид інформаційної діяльності. URL: https://elartu.tntu.edu.ua/bitstream/lib/25242/2/MSNK_2018v2_Maksumchyk_K-Annotation_and_referencing_126.pdf

6. Анотація видавнича. URL: https://vue.gov.ua/Анотація_видавнича

7. Огляд методів автоматичного анотування текстів. URL: https://core.ac.uk/reader/52160900

8. Cahyana, N.H.; Saifullah, S.; Fauziah, Y.; Aribowo, A.S.; Drezewski, R. Semi-supervised Text Annotation for Hate Speech Detection using K-Nearest Neighbors and Term Frequency-Inverse Document Frequency. Int. J. Adv. Comput. Sci. Appl. 2022, 13, 147–151.

9. Довга короткочасна пам'ять. URL: https://uk.wikipedia.org/wiki/Довга_короткочасна_пам%27ять

10. Attention – Seq2Seq Models. URL: https://towardsdatascience.com/day-1-2-attention-seq2seq-models-65df3f49e263

11. Googles Bard. URL: https://www.larksuite.com/en_us/topics/ai-glossary/googles-bard#what-is-google's-bard?

12. Automated text annotation using a semi-supervised approach with meta vectorizer and machine learning algorithms for hate speech detection. URL: https://www.mdpi.com/2076-3417/14/3/1078

13. Навчання реферуванню та анотуванню фахових іншомовних текстів у підготовці студентів немовних спеціальностей. URL: https://www.philol.vernadskyjournals.in.ua/journals/2021/3_2021/part_1/18.pdf.

14. Smodin. URL: https://app.smodin.io/uk/письменник/ece

15. Summarize documents, text, and more with generative AI and LLMs. URL: https://cloud.google.com/use-cases/ai-summarization?

16. Huggingface. BART (large-sized model), fine-tuned on CNN Daily Mail URL: https://huggingface.co/facebook/bart-large-cnn

17. BART Model Architecture. URL: https://medium.com/@nadirapovey/bart-model-architecture-8ac1cea0e877

18. Hugging Face. URL: https://uk.wikipedia.org/wiki/Hugging_Face

19. The Python Programmer’s Toolkit: Essential Libraries for Translation. URL: https://python.plainenglish.io/the-python-programmers-toolkit-essential-libraries-for-translation-989b039979a8

20. LanguageTool: Grammar and Spell Checker in Python. URL: https://medium.com/swlh/languagetool-grammar-and-spell-checker-in-python-578ac4e94642

21. Python ROUGE Implementation. URL: https://pypi.org/project/rouge-score/

22. SentenceTransformers Documentation. URL: https://sbert.net/

23. PyQt library in Python. URL: https://www.javatpoint.com/pyqt-library-in-python

24. Професійний інструментарій сучасного pr-текстолога. URL: http://dspace.wunu.edu.ua/bitstream/316497/31425/1/74.PDF

25. ROUGE your NLP Results. URL: https://medium.com/@priyankads/rouge-your-nlp-results-b2feba61053a

26. PyCharm. URL: https://uk.wikipedia.org/wiki/PyCharm

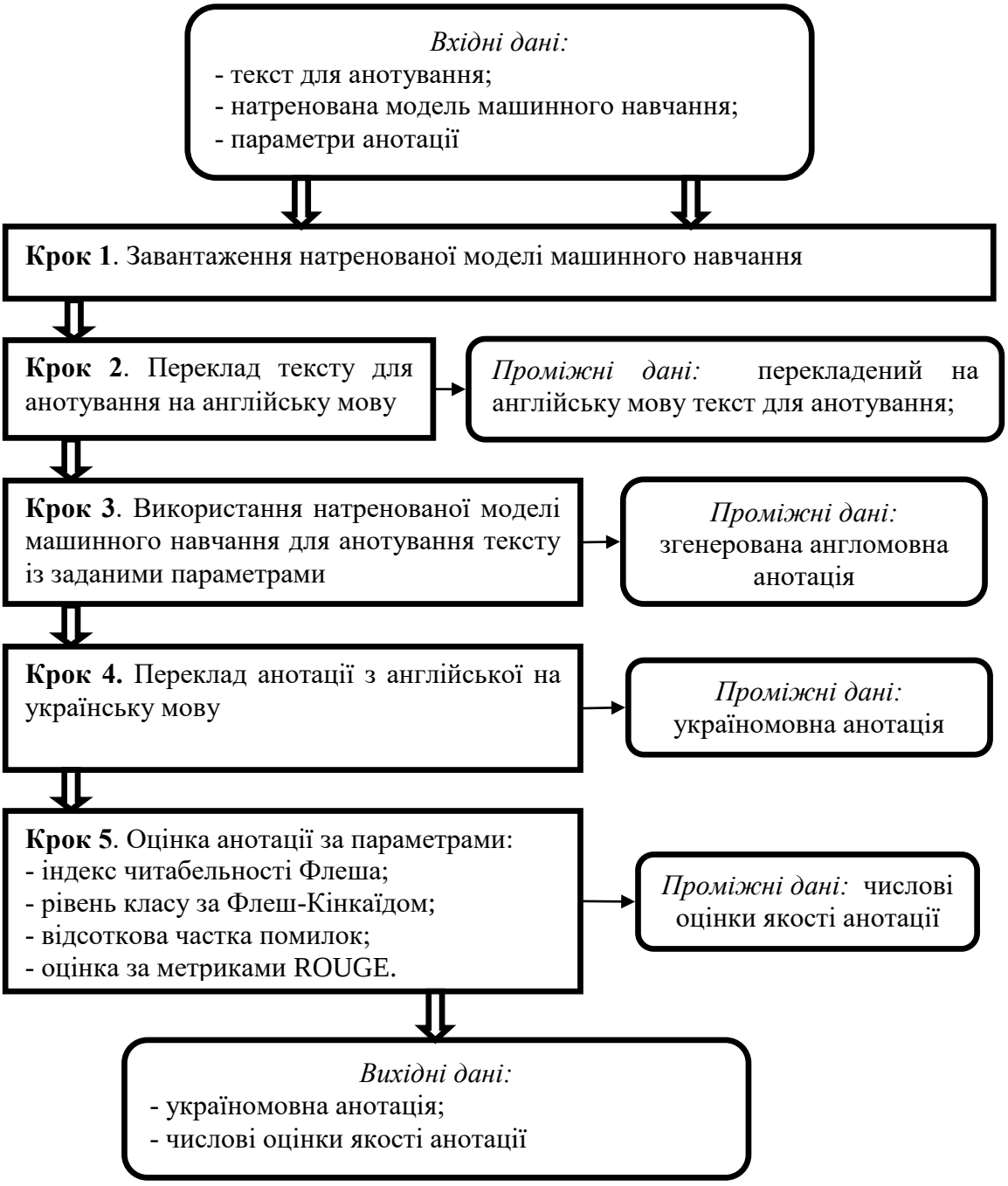
27. SQLite. URL: https://uk.wikipedia.org/wiki/SQLite

28. Mazurets O., Molchanova M., Klimenko V., Prosvitliuk M Practice Implementation of Neural Network Model BART-Large-CNN for Text Annotation. Prospects of Scientific Research in the Conditions of the Modern World. Proceedings of XXVII International scientific and practical conference. June 12-14, 2024. Rotterdam, Netherlands. 2024. Pp. 97-102.

ДОДАТКИ

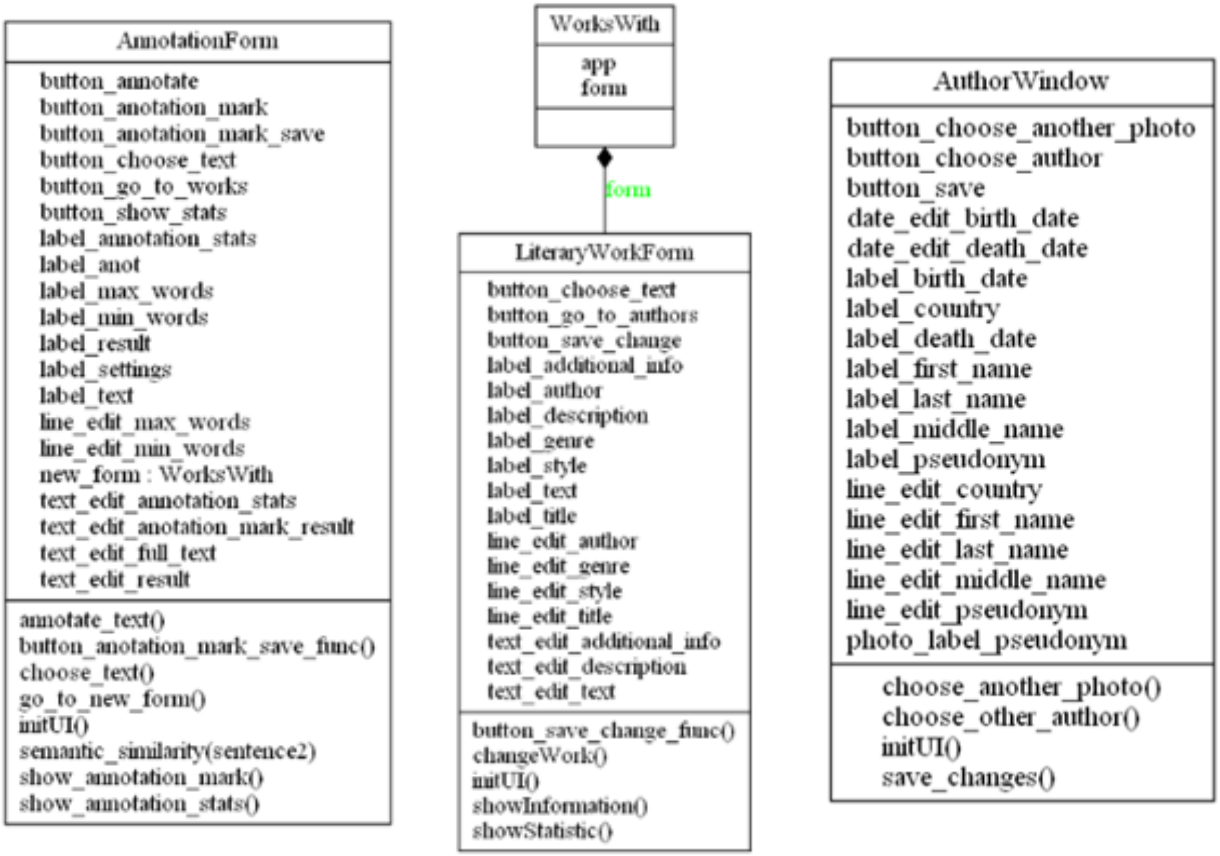
Додаток А

Кроки методу анотування україномовних художніх творів засобами машинного навчання



Додаток Б

Діаграма класів інтелектуальної системи анотування україномовних художніх творів



Додаток В

Розгорнута архітектура моделі машинного навчання для анотування україномовних художніх творів

Додаток Г

Презентаційний матеріал

КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА

МЕТОД АНОТУВАННЯ УКРАЇНОМОВНИХ ХУДОЖНІХ ТВОРІВ ЗАСОБАМИ МАШИННОГО НАВЧАННЯ

Виконав:
студент групи КНс-21-1
Михайло ПРОСВІТЛЮК

Керівник:
викладач кафедри КН
Валерія КЛІМЕНКО

Актуальність

В умовах розвитку глобальної мережі, так само і невпинно зростає кількість художніх творів, які автори щоденно завантажують. Автоматичне анотування дозволяє швидко та ефективно створювати метадані для великих обсягів літературних текстів, що розширює наявні бази даних і полегшує доступ до них.

На основі анотованих даних є можливість розробки систем рекомендацій, які враховують індивідуальні інтереси користувачів, а також засоби для автоматичного аналізу змісту та семантики текстів. Ще анотування літературних текстів може служити як основа для розробки освітніх програм з цифрової грамотності та літературознавства, допомагаючи учням аналізувати та розуміти текстову інформацію, а також допомагає зберегти та розповсюдити літературну спадщину, забезпечуючи доступ до цінних творів для наступних поколінь та дослідників..

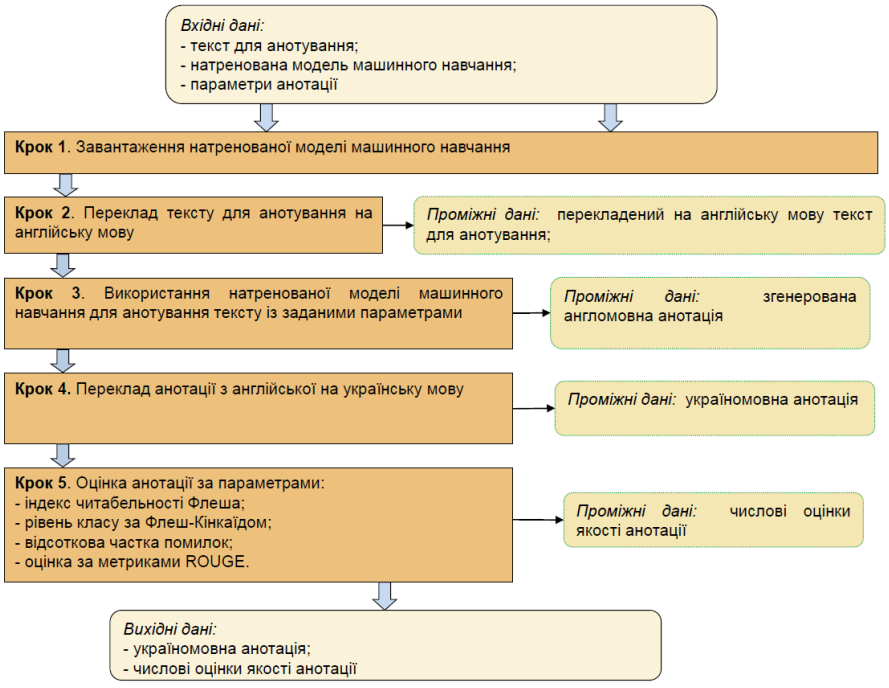
Отже, **анотування україномовних художніх творів засобами машинного навчання сприяє розвитку українського літературознавства, полегшує доступ до літературних ресурсів та розвиває прикладні технології обробки природної мови**.

Мета і задачі роботи

Метою кваліфікаційної роботи бакалавра є спрощення створення анотацій україномовних художніх творів шляхом автоматизації анотування засобами машинного навчання.

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

- виконати дослідження предметної області анотування художніх творів засобами машинного навчання;
- виконати огляд теоретичних підходів до вирішення подібних задач, обрати підхід до автоматизованого анотування художніх творів серед методів машинного навчання;
- створити метод анотування україномовних художніх творів засобами машинного навчання;
- створити інформаційну структуру системи автоматизованого анотування користувацьких художніх текстів; виконати програмну реалізацію інформаційної системи;
- провести тестування інформаційної системи автоматизованого анотування користувацьких художніх текстів;
- виконати дослідження ефективності методу анотування україномовних художніх творів засобами машинного навчання.



Кроки методу анотування україномовних художніх творів засобами машинного навчання

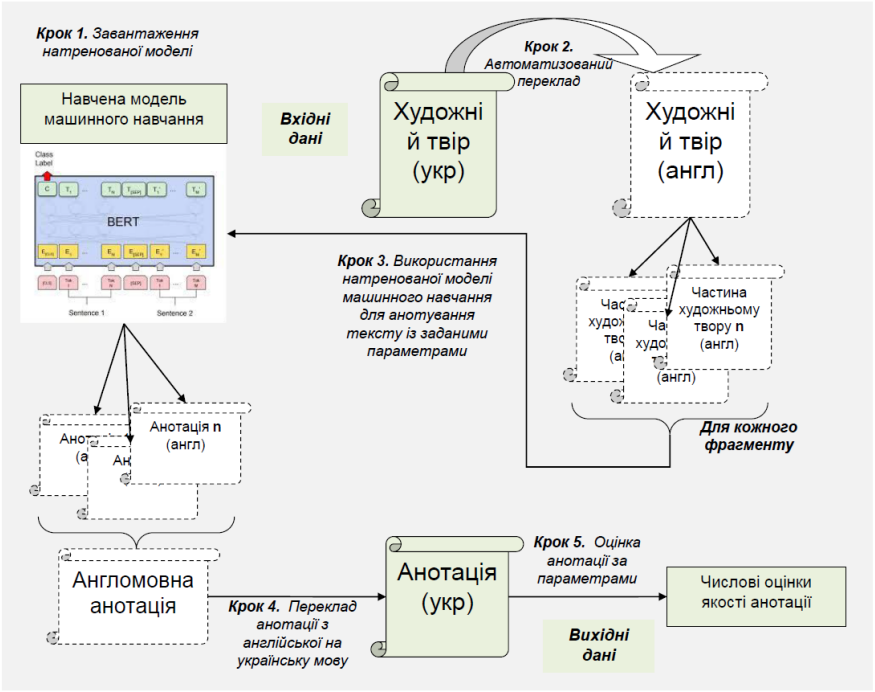


Схема послідовності виконання кроків методу анотування україномовних художніх творів

Архітектура моделі машинного навчання «BART-Large-CNN»

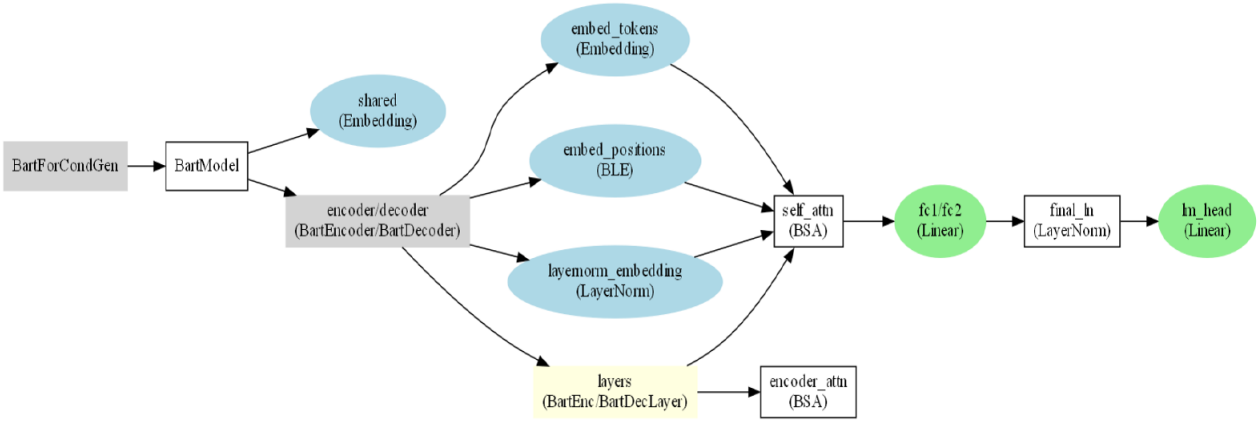
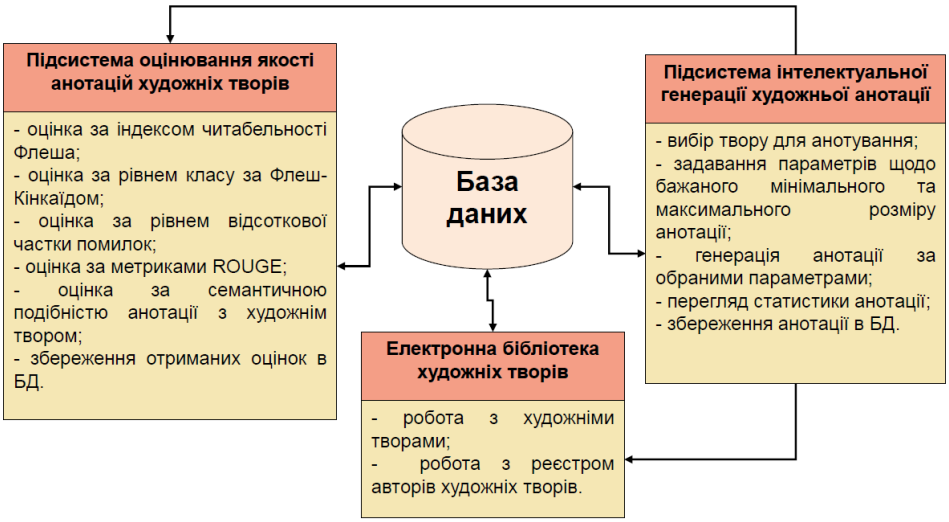
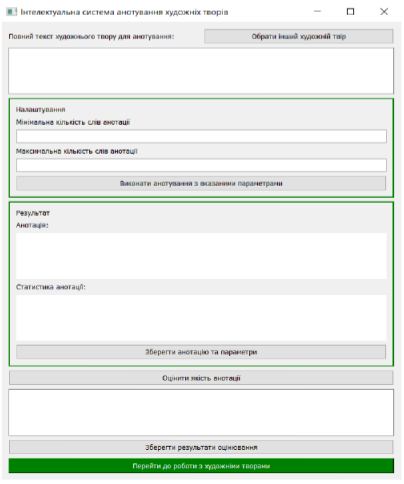


Схема підсистем інформаційної системи анотування україномовних художніх творів

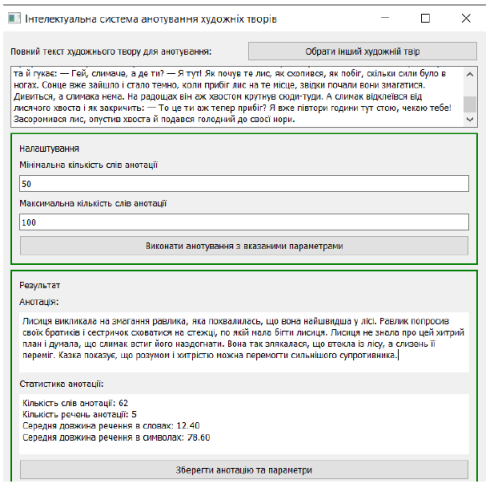


Даталогічна модель бази даних інтелектуальної системи анотування україномовних художніх творів

Інформаційна система анотування україномовних художніх творів



Вигляд головного екрану



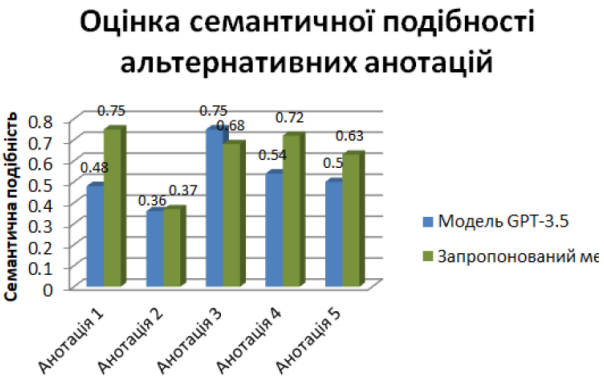
Виконана анотація художнього твору

Результати досліджень

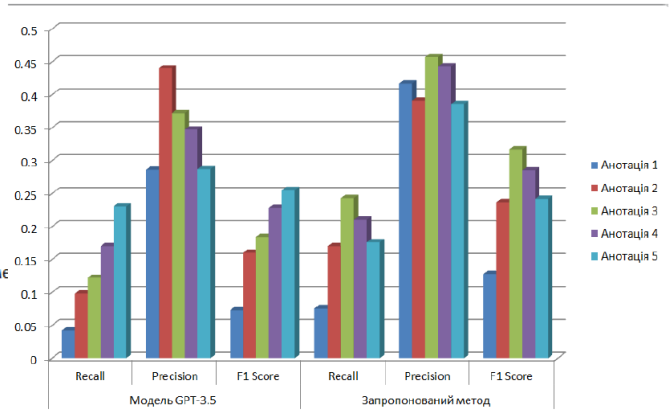
Для дослідження ефективності методу анотування україномовних художніх творів засобами машинного навчання буде використано створену програмну реалізацію. Також буде виконано порівняння з результатами мовної моделі GPT-3.5.

	Оцінка семантичної подібності для створеної моделлю GPT-3.5 анотації з оригіналом	**Оцінка семантичної подібності для створеної запропонованим методом анотації з оригіналом**
Анотація 1	0.48	0.75
Анотація 2	0.36	0.37
Анотація 3	0.75	0.68
Анотація 4	0.54	0.72
Анотація 5	0.5	0.63

Результати досліджень



Результати експерименту за семантичною подібністю



Порівняння моделі GPT-3.5 з розробленим методом за метрикою ROUGE по уніграмам

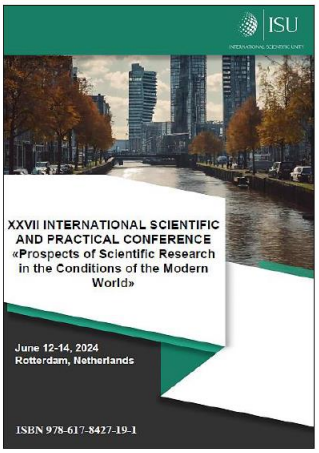
Висновки

Метою кваліфікаційної роботи бакалавра було спрощення створення анотацій україномовних художніх творів шляхом автоматизації анотування засобами машинного навчання.

Були вирішені завдання дослідження:

- досліджено предметну область анотування художніх творів засобами машинного навчання;
- виконано огляд теоретичних підходів до вирішення подібних задач, обрано підхід до автоматизованого анотування художніх творів серед методів машинного навчання, а саме використання нейромережевої моделі BARD;
- створено метод анотування україномовних художніх творів засобами машинного навчання;
- створено інформаційну структуру системи автоматизованого анотування користувацьких художніх текстів;
- виконано програмну реалізацію інформаційної системи;
- проведено тестування інформаційної системи автоматизованого анотування користувацьких художніх текстів;
- виконано дослідження ефективності методу анотування україномовних художніх творів засобами машинного навчання.

Наукові та практичні результати роботи публікувалися у матеріалах XXVII Міжнародної науково-практичної конференції «Prospects of Scientific Research in the Conditions of the Modern World» (12-14 червня 2024 року, Rotterdam, Netherlands), за темою кваліфікаційної роботи бакалавра автором виконано наукову публікацію «Practice Implementation of Neural Network Model BART-Large-CNN for Text Annotation».



UNICHECK
by Turnitin

Ім'я користувача:
Кафедра КН

ID перевірки:
1016374209

Дата перевірки:
19.06.2024 07:58:04 EEST

Тип перевірки:
Doc vs Internet + Library

Дата звіту:
19.06.2024 08:50:00 EEST

ID користувача:
100005671

Назва документа: **КНс-21-1 Просвітлюк_ЗАПИСКА**

Кількість сторінок: **72** Кількість слів: **11646** Кількість символів: **94381** Розмір файлу: **1.75 MB** ID файлу: **1016181931**

Виявлено модифікації тексту (можуть впливати на відсоток схожості)

8.28%
Схожість

Найбільша схожість: **3.59%** з джерелом з Бібліотеки (ID файлу: **1016181930**)

4.65% Джерела з Інтернету	353	Сторінка 74
5.74% Джерела з Бібліотеки	103	Сторінка 76

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	2
Підозріле форматування	19 сторінок

Wed Jun 19 07:54:03 EEST 2024, Петровський Сергій Степанович, Хмельницький національний університет, ХНУ

Anti-Plagiarism v-15.257

Максимальне співпадіння з одним документом 4.0%

Словники перевірки: en_US, ru_RU, ua_UA. **Помилок в документах: 10%**

ID: 131451 Назва: КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА на тему Метод анотування україномовних художніх творів засобами машинного навчання Додано в БД: 2024-06-19 Автора: Михайло ПРОСВІТЛЮК Керівники: Валерія КЛІМЕНКО Консультанти: Опоненти:	Документ		Сумарний збіг по Базі Даних	
	Символи	Лексеми	Символи	Лексеми
	73961	1097	3787 (5%)	54 (5%)

Джерело плагіату

ID	Опис	Наявність плагіату в документі	
		Символи	Лексеми

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 анотування україномовних художніх творів засобами машинного навчання*

Автор: *студент групи КНс-21-1 Михайло Просвітлюк*

Спеціальність: *122 – Комп'ютерні науки*

Освітня програма: *освітньо-професійна*

Науковий керівник: *викладач кафедри КН Валерія Кліменко*

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 Михайла Просвітлюка, 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.

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

- за системою Anti-Plagiarism: 4%;

- за системою Unicheck: 8.28 %.

Керівник роботи _____ *Валерія КЛІМЕНКО*

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. КНс-21-1 Просвітлюка Михайла Ігоровича*

за темою *Метод анотування україномовних художніх творів засобами машинного навчання*

1. Актуальність теми

Актуальним завданням, яке потребує аналізу і досліджується у даній роботі, є автоматизоване анотування україномовних художніх творів. Зростання обсягів літературного контенту вимагає ефективних інструментів для швидкого та якісного аналізу текстів. Для забезпечення цього необхідно розробити інформаційну систему, що генеруватиме анотації для творів та оцінюватиме їх за визначеними метриками. Такий підхід сприятиме збереженню та популяризації української літератури, полегшуючи доступ до її змісту для широкої аудиторії. Розробка методу анотування є важливим завданням у сфері комп'ютерних наук.

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

За стандартом, а саме описом предметної області, об'єктом кваліфікаційної роботи бакалавра є процес анотування україномовних художніх творів засобами машинного навчання. Метою роботи – спрощення створення анотацій україномовних художніх творів шляхом автоматизації анотування засобами машинного навчання. При вирішенні поставленої задачі використано методи та засоби машинного навчання для роботи з текстовою інформацією. Тому результати виконання кваліфікаційної роботи бакалавра відповідають стандарту бакалавра спеціальності 122 – Комп'ютерні науки.

3. Професійні та особистісні якості бакалавра

При виконанні кваліфікаційної роботи студент Просвітлюк Михайло Ігорович проявив себе як студент, який вчасно та повною мірою виконує поставлені задачі. У процесі розробки програмного забезпечення він продемонстрував високий рівень компетентності, підтвердивши свою кваліфікацію як фахівець, здатний вирішувати складні наукові та технічні задачі.

4. Ступінь самостійності під час виконання кваліфікаційної роботи

Отримані в роботі результати є наслідком самостійної діяльності студента, який особисто виконував всі поставлені завдання.

5. Ступінь оволодіння методами дослідження

Під час виконання кваліфікаційної роботи студент продемонстрував хороший рівень компетентності, а також володіння необхідними засобами, методами, методиками та технологіями в галузі комп'ютерних наук

6. Повнота та якість розкриття теми роботи

У роботі повністю розкрита тема, адже якісно проведено дослідження предметної області, що дозволило виконати поставлені завдання в повній мірі.

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

Матеріал у роботі подано логічно, послідовно та розгорнуто. Робота має чітку структуру та аргументовано усі наведені твердження. Літературна грамотність також на високому рівні.

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

Розроблений метод може бути застосований для автоматизованої генерації анотацій, що полегшує пошук і аналіз літературних творів, їх категоризацію та індексацію. Також, програмна реалізація цього методу може знайти застосування у бібліотечних системах, веб-платформах для культурного контенту або в навчальних середовищах для покращення доступу до україномовної літератури та її дослідження.

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

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

Керівник _____ *викладач каф. КН Валерія КЛІМЕНКО*

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
МОН УКРАЇНИ

Кафедра комп'ютерних наук

РЕЦЕНЗІЯ

на кваліфікаційну роботу бакалавра

студента *гр. КНс-21-1 Просвітлюка Михайла Ігоровича*

за темою: *Метод анотування україномовних художніх творів засобами машинного навчання*

1. Актуальність обраної теми

Актуальність анотування україномовних художніх творів зумовлена зростаючою потребою в автоматизації процесів обробки та аналізу літературних текстів. Анотації допомагають швидко орієнтуватися в змісті творів, сприяють їх популяризації та збереженню культурної спадщини. В умовах цифровізації інформації та розвитку електронних бібліотек, створення ефективних методів анотування україномовних текстів стає важливим завданням, яке сприяє підтримці і розвитку української літератури, полегшує доступ до її багатств для широкого кола читачів та дослідників.

2. Повнота розкриття мети та завдань роботи

В своїй кваліфікаційній роботі автор повною мірою розкрив мету та поставлені в роботі завдання, що видно з змісту кожного розділу пояснювальної записки.

3. Зміст кожного розділу роботи

Кваліфікаційна робота бакалавра містить три розділи, в яких розкрита мета та завдання роботи. У першому розділі розглянуто характеристику предметної області анотування художніх творів засобами машинного навчання. У другому розділі спроєктовано інформаційну систему анотування україномовних художніх творів .а саме подано модель анотування україномовних художніх творів, наведено схему метода, подано архітектуру використаної моделі та наведено проєктну архітектуру системи. У третьому розділі визначено шляхи дослідження методу та засобів створення програмного забезпечення, а також досліджено метод.

4. Оцінка розробленої інформаційної системи, її практична цінність

Напрямком практичного застосування розробленої інформаційної системи є автоматизована генерація україномовної анотації для заданого твору та її оцінку за відповідними метриками.

5. Якість оформлення кваліфікаційної роботи бакалавра

Кваліфікаційна робота бакалавра виконана достатньо якісно. Пояснювальна записка має чітку структуру У записці також використано графічний спосіб представлення інформації , а саме схеми, діаграми, що спрощує сприйняття результатів проведених досліджень.

6. Недоліки кваліфікаційної роботи бакалавра

У кінці кожного пункту записки надто короткі висновки. Деякі елементи пояснювальної записки (наприклад, таблиця 2.8) розміщені по тексту раніше ніж перше посилання на них. У дослідженні ефективності проведено порівняння ефективності роботи розробленого методу тільки з однією генеративною моделлю, а також з малою кількістю зразків.

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 ***«відмінно»****.*

Рецензент д.т.н., проф. Лисенко С.М.