Хмельницький національний університет
Факультет інформаційних технологій
Кафедра комп'ютерних наук

КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА

на тему *Метод виявлення гендерної приналежності за дописами соціальних інтернет-мереж засобами NLP*

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*		*Павло СУПРУН*
	Група виконавця	Підпис	Ім'я, ПРІЗВИЩЕ
Керівник:	*викладач каф. КН*		*Марина МОЛЧАНОВА*
	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 — Підпис — *Олександр БАРМАК* (Ім'я, ПРІЗВИЩЕ)
21 червня 2024 р.

Хмельницький 2024

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

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

ЗАТВЕРДЖУЮ
Завідувач кафедри *комп'ютерних наук*
_____
(підпис)
д.т.н., професор Олександр БАРМАК
«16» 02 *2024* року

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ БАКАЛАВРА

1. Тема кваліфікаційної роботи бакалавра: *«Метод виявлення гендерної приналежності за дописами соціальних інтернет-мереж засобами NLP»*

2. Завдання видано студенту *Павлові СУПРУНУ*
(Ім'я, прізвище)

3. Керівник роботи *викладач кафедри КН Марина МОЛЧАНОВА*
(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 – покращення виявлення гендерної приналежності у дописах соціальних інтернет-мереж. Для досягнення поставленої мети слід вирішити такі задачі: аналіз предметної області виявлення гендерної приналежності за текстовими даними; огляд теоретичних підходів щодо можливості виявлення гендерної приналежності за дописами; виконати огляд існуючих наукових надбань та програмних рішень; створити метод виявлення гендерної приналежності за дописами соціальних інтернет-мереж засобами NLP та описати його кроки; спроектувати інформаційну систему для реалізації методу виявлення гендерної приналежності за дописами соціальних інтернет-мереж засобами NLP; створити програмну реалізацію; тестування створеної програмної реалізації та дослідження ефективності створеного методу.

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* _____ *Павло СУПРУН*
Група виконавця — Підпис — Ім'я, ПРІЗВИЩЕ

Керівник: *викладач каф. КН* _____ *Марина МОЛЧАНОВА*
Науковий ступінь, посада — Підпис — Ім'я, ПРІЗВИЩЕ

Анотація

Тема кваліфікаційної роботи бакалавра: *«Метод виявлення гендерної приналежності за дописами соціальних інтернет-мереж засобами NLP»*

Виконавець кваліфікаційної роботи бакалавра: *студент групи КНс-21-1 Павло СУПРУН*

Керівник кваліфікаційної роботи бакалавра: *викладач кафедри КН Марина МОЛЧАНОВА*

Кваліфікаційна робота бакалавра містить:

Пояснювальна записка				Кількість додатків
Сторінок	Рисунків	Таблиць	Джерел інформації	
68	37	18	25	5

Метою кваліфікаційної роботи бакалавра є покращення виявлення гендерної приналежності у дописах соціальних інтернет-мереж шляхом розробки методу виявлення гендерної приналежності за дописами соціальних інтернет-мереж засобами NLP, а також відповідної програмної реалізації. Для розробки інформаційної системи було використано мову програмування Python, а для створення та взаємодії з базою даних мову запитів SQL.

Розроблена система призначена для виявлення гендерної приналежності авторів на основі їхніх текстових дописів у соціальних мережах за допомогою методів NLP. Може використовуватись науковцями та маркетологами.

Напрямками практичного використання розробленої інформаційної системи визначено здійснення автоматизованого аналізу текстів з метою встановлення гендерної приналежність авторів для вивчення розподілу гендерних ролей та стереотипів у мережевому середовищі.

Ключові слова: гендери, види гендерів, SVM, інформаційна система.

Виконавець: *студент групи КНс-21-1* _____ *Павло СУПРУН*

Група виконавця | Підпис | Ім'я, ПРІЗВИЩЕ

Зміст

Перелік скорочень

Скорочення, термін, позначення	Пояснення
NLP	Обробка природної мови
TF-IDF	Term Frequency – Inverse Document Frequency
TF	Term Frequency
IDF	Inverse Document Frequency
КРБ	Кваліфікаційна робота бакалавра
SVM	Support Vector Machines
SGD	Класифікатор стохастичного градієнтного спуску
НМ	Нейронна мережа
ПЗ	Пояснювальна записка
ПП	Програмний продукт
LSA	Latent Semantic Analysis
API	Application Programming Interface
ХНУ	Хмельницький національний університет.

Вступ

Кваліфікаційна робота бакалавра присвячена покращенню виявлення гендерної приналежності у дописах соціальних інтернет-мереж шляхом розробки методу виявлення гендерної приналежності за дописами соціальних інтернет-мереж засобами NLP, а також відповідної програмної реалізації, що буде використовувати розроблений метод.

Актуальність. Зростання використання соціальних мереж і стійка популярність онлайн-комунікацій робить завдання виявлення гендерної приналежності за дописами необхідним для різноманітних застосувань, включаючи маркетингові дослідження, аналіз громадської думки, персоналізовану рекламу, політичні дослідження та багато іншого.

Завдяки розвитку методів машинного навчання та обробки природної мови, можливості виявлення гендерної приналежності зростають. Ці дані можуть бути використані для аналізу та розуміння поведінки користувачів в інтернеті, виявлення та прогнозування тенденцій у споживчому ринку, а також для створення більш ефективних стратегій спілкування та взаємодії з аудиторією. Використання таких методів дозволяє уникнути стереотипів та прихованих асоціацій, а також забезпечити репрезентативність аналізу за гендерними критеріями.

Зростаюча увага до питань рівності та різноманітності також робить цей напрямок актуальним, оскільки він може використовуватися для аналізу та виявлення можливих стереотипів, дискримінації або нерівності, що можуть існувати в онлайн-спільнотах.

Отже, виявлення гендерної приналежності за допомогою NLP має великий потенціал у багатьох сферах, що робить напрям дослідження вкрай актуальним в сучасному інформаційному суспільстві.

Об’єкт дослідження – процес виявлення гендерної приналежності за дописами соціальних інтернет-мереж засобами NLP.

Предмет дослідження – моделі, методи та засоби машинного навчання, які дозволяють визначити гендерну приналежність за дописами соціальних інтернет-мереж.

Мета кваліфікаційної роботи бакалавра – покращення виявлення гендерної приналежності у дописах соціальних інтернет-мереж.

Завдання кваліфікаційної роботи бакалавра. Провести аналіз предметної області виявлення гендерної приналежності за текстовими даними. Виконати огляд теоретичних підходів щодо можливості виявлення гендерної приналежності за дописами соціальних інтернет-мереж, обрати підхід для подальшої реалізації. Виконати огляд існуючих наукових надбань та програмних рішень. Створити метод виявлення гендерної приналежності за дописами соціальних інтернет-мереж засобами NLP та описати його кроки. Спроектувати інформаційну систему для реалізації методу виявлення гендерної приналежності за дописами соціальних інтернет-мереж засобами NLP. Створити програмну реалізацію за спроєктованою структурою інформаційної системи. Виконати тестування створеної програмної реалізації. Виконати дослідження ефективності методу виявлення гендерної приналежності за дописами соціальних інтернет-мереж засобами NLP на основі створеного ПЗ.

Розділ 1 Характеристика предметної області виявлення гендерної приналежності за текстовими даними

1.1 Аналіз ознак у тексті для виявлення гендерних відмінностей

Зважаючи на те, що мова є важливим засобом вираження особистості, існує кілька характеристик тексту, які можуть вказувати на можливі гендерні відмінності в способі спілкування.

Лінгвістична гендерологія розглядає мову як ключовий засіб для розуміння гендеру як окремої дисципліни та вивчає взаємозв'язок між мовою та гендером як складовими соціокультурного контексту [1]. Це дозволяє вивчати, як сама мова відтворює та формує стереотипи, ролі та ідеї про гендер, а також розглядати мовленнєву поведінку як важливий аспект гендерної ідентичності.

Наука гендерна лінгвістика вивчає мову та комунікацію з урахуванням гендерних аспектів з метою продукції нових знань у сфері лінгвістики. Це включає дослідження гендерних відмінностей у вживанні мови, аналіз гендерного впливу на мовленнєві структури та практики, а також розробку та застосування гендерно-чутливих методів дослідження [2].

На сьогоднішній день термін «гендер» використовується в різних сферах суспільства. У лінгвістиці, вперше, слово «гендер» описувало граматичну відмінність за ознаками чоловічої, жіночої та середньої статей в англійській мові. У науковому контексті термін «гендер» використовується для позначення соціальної статі [3].

Жіноча мова має деякі відмінності від чоловічої, нижче наведені ключові з них [4]:

– Використання «порожніх» оціночних прикметників. Жіноча мова може частіше містити слова, які виражають емоційне ставлення або оцінку, такі як «чудово», «гарний», «чудовий» і т. д. Ці слова можуть використовуватися для створення позитивного або співчутливого тону у спілкуванні.

– Використання питальних форм там, де чоловіки використовують позитивні. Жіноча мова може включати більше питальних речень або виразів, які

виражають сумнів або невпевненість. Наприклад, «Чи не можна мені це зробити?» замість «Я можу це зробити».

– Використання ввічливих форм. Жіноча мова може бути більш ввічливою та уважною до співрозмовника, використовуючи ввічливі вирази та форми. Наприклад, «Чи не могли б ви, будь ласка, допомогти мені?».

– Використання форм, що виражають невпевненість. Жіноча мова може частіше містити вирази, які виражають невпевненість або обережність у висловленнях. Наприклад, «Можливо, це не зовсім правильно, але...».

– Використання підсилювачів. Жіноча мова може містити більше використання підсилювачів, які допомагають підкреслити емоційність або важливість висловлювань. Наприклад, «Дуже дякую!».

– Використання гіперкорректної граматики. Жіноча мова може бути більш схильною до використання граматично правильних, але можливо зайвих конструкцій у мові. Наприклад, уникання використання розмовного стилю та підтримка формальної граматичної правильності.

Гендерна ідентичність – це внутрішнє сприйняття статевої сутності та вираження гендерних характеристик через одяг, поведінку та зовнішній вигляд. Це психосоціальний конструкт, який виникає у ранньому віці і може відрізнятися від статевої приналежності, присвоєної при народженні. Хоча гендер традиційно розглядали як бінарний, сучасна розуміння визнає багато різновидів гендерної ідентичності, включаючи трансгендерність, небінарність та інші форми [5].

Різноманітність гендерів охоплює широкий спектр термінів та понять, які використовуються для характеризації гендерної ідентичності людей. До них відносяться:

– цисгендерність – стан, коли гендерна ідентичність співпадає з призначеною статтю або гендером тіла;

– трансгендерність – гендерна ідентичність, яка відрізняється від біологічної статі при народженні;

– гендерквір (або небінарний) – гендерна ідентичність, що не вписується в традиційні бінарні уявлення про стать, або може поєднувати елементи різних гендерних ідентичностей.

Ці терміни розглядають гендер як більш складне поняття, ніж просто «чоловік» або «жінка», визнавши наявність широкого спектру гендерних ідентичностей, які можуть знаходитися як всередині, так і поза бінарним спектром.

Отже, перераховані вище різноманітні гендери мають певні риси на письмі, які сприйматися як характеристики, що сприяють відокремленню гендерів і враховуються при автоматизованому аналізі текстів з метою виявлення гендерних відмінностей за допомогою методів NLP.

1.2 Огляд теоретичних підходів до розв'язку задачі виявлення гендерної приналежності за дописами

Одним з методів є аналіз лінгвістичних ознак, таких як вживання слів, фраз, структури речень тощо. Наприклад, деякі дослідження спрямовані на виявлення стилістичних відмінностей між чоловічими та жіночими текстами. Інші методи використовують машинне навчання для класифікації дописів за гендерною приналежністю.

Класифікація текстів – це процес аналізу текстових даних з метою їхнього розподілу на категорії або класи в залежності від їх змісту. Цей процес є ключовим у багатьох системах штучного інтелекту та природної обробки мови , оскільки NLP дозволяє автоматично виконувати різноманітні завдання, такі як фільтрація спаму, аналіз настроїв, аналіз гендерної приналежності або робота чат-ботів [6].

Методи класифікації тексту на основі правил використовують чітко визначені лінгвістичні правила для аналізу текстових даних і приведення їх до певних категорій. У цих системах інженери створюють правила, які дозволяють визначити клас, до якого належить кожний фрагмент тексту, шукаючи в них

семантично відповідні елементи. Кожне правило має певний шаблон, за яким текст визначається як належний до певної категорії.

Наприклад, якщо потрібно розробити класифікатор тексту для відрізнення різних тем розмов, таких як погода, фільми або їжа, можна створити список ключових слів, фраз та інших шаблонів для кожної теми. Наприклад, для класифікації тексту про погоду можуть використовуватися слова «вітер», «дощ», «сонце», «сніг» або «хмара». Система аналізує вхідний текст, підраховує кількість входжень цих ключових слів та відносить текст до певного класу, визначеного цими правилами.

Перевагою методів класифікації на основі правил є їхня передбачуваність та інтерпретованість людьми, що дозволяє покращувати їх за допомогою ручного втручання. Однак вони також можуть бути крихкими, оскільки вони обмежені лише попередньо визначеними шаблонами та можуть мало пристосовуватися до нових ситуацій без значного втручання інженерів.

Системи машинного навчання працюють за допомогою алгоритмів, які аналізують набори даних для виявлення шаблонів, що вказують на певний клас. Ці алгоритми використовують попередньо класифіковані екземпляри для навчання, що дозволяє їм вивчати шаблони, пов'язані з різними класами. Потім вони застосовують ці навчальні дані для класифікації нових екземплярів, позбавляючи їх міток та призначаючи їм власні мітки. Як результат, точність класифікації оцінюється порівнянням призначених міток з вихідними, щоб визначити, наскільки добре модель навчилася розпізнавати шаблони, що вказують на різні класи.

Перш ніж використовувати текстові дані у моделі машинного навчання, їх потрібно перетворити у числовий формат, що відбувається через процес вилучення ознак. Вилучення ознак в тексті означає перетворення тексту на числові значення або функції, які можна використовувати для аналізу. Одним із методів вилучення ознак є TF-IDF [7].

TF-IDF (Term Frequency-Inverse Document Frequency) – це статистичний метод вилучення ознак, який оцінює важливість термінів у документах. TF-IDF

обчислюється шляхом множення двох компонентів: частоти терміну, TF, яка показує, наскільки часто термін зустрічається у документі, і зворотної частоти документа, IDF, яка показує, наскільки унікальним є термін серед всіх документів у корпусі. Цей метод дозволяє виділити ключові слова у тексті, які є важливими для розрізнення документів у корпусі.

Наївний Байесовий класифікатор (Naive Bayes Classifier) – це проста, але ефективна модель машинного навчання, яка використовує теорему Байєса для класифікації об'єктів на основі ймовірностей входження певних ознак у кожен клас. Основна передумова наївного байєсового класифікатора – усі ознаки у вхідних даних вважаються незалежними одна від одної, хоча це може бути спрощенням у реальних ситуаціях. Тем не менш, наївний байєсовий класифікатор може бути дуже ефективним у багатьох завданнях, таких як фільтрація спаму, аналіз текстів або класифікація документів [8].

Метод опорних векторів (Support Vector Machines) – це потужний метод для класифікації та регресії, який працює, шукаючи оптимальну гіперплощину (або гіперплощини), яка найкраще розділяє точки даних різних класів у просторі ознак (рисунок 1.1).

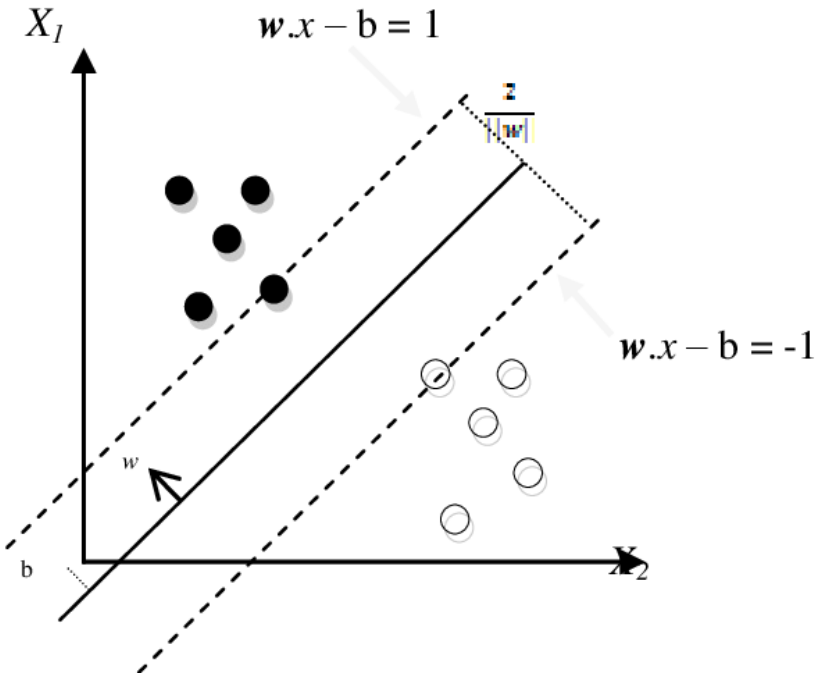


Рисунок 1.1 – Гіперплощина для SVM [9]

Головна ідея полягає в тому, щоб знайти оптимальну гіперплощину, яка максимізує відстань між найближчими точками кожного класу, відомих як опорні вектори. SVM може працювати як з лінійними, так і з нелінійними наборами даних, за допомогою ядерної функції, яка дозволяє виконувати складніші розділення. SVM є популярним методом в багатьох областях, таких як розпізнавання образів, біоінформатика, аналіз тексту та інші завдання машинного навчання.

З урахування описаних теоретичних підходів, для виявлення гендерної приналежності за дописами соціальних інтернет-мереж буде використано SVM.

1.3 Аналіз існуючих програмних засобів та наукових рішень

У сфері NLP автори підійшли до завдання профілювання автора з точки зору гендеру як до проблеми класифікації тексту [10]. На основі стилю письма автора можна передбачити соціально-демографічну інформацію, таку як гендер, вік або рідна мова автора. Експоненціальне зростання даних, що кожен день створюється користувачами, і розвиток методів машинного навчання призвели до значного прогресу в автоматичному визначенні гендеру. На жаль, моделі визначення гендеру часто стають чорними ящиками з точки зору інтерпретації. У цій статті автори пропонують деревоподібну обчислювальну модель для визначення гендеру, що складається з 198 ознак. На відміну від попередніх робіт [11] з визначення гендеру, автори розподілили ознаки з лінгвістичної точки зору на шість категорій: орфографічні, морфологічні, лексичні, синтаксичні, цифрові та прагматично-дискурсивні. Автори запровадили класифікатор Decision-Tree для оцінки ефективності всіх комбінацій функцій, і експерименти показали, що в середньому точність класифікації зросла до 3,25% із додаванням наборів функцій. Максимальної точності класифікації досягла трирівнева модель, яка поєднувала лексичні, синтаксичні та цифрові ознаки. У дослідження представлені найбільш релевантні функції для визначення гендеру відповідно до дерев, згенерованих класифікатором, і контекстуалізовані значення результатів

обчислень з лінгвістичними моделями, визначеними попередніми дослідженнями щодо гендеру.

У дослідженні авторів [12] вивчається проблема побудови моделі для характеристики користувачів Pinterest за двома демографічними змінними, віком і гендером, використовуючи їх текстову інформацію в мережі. Для цього автори представили набір даних, сформований текстами англійською мовою з 548 761 пінів, що відповідають 264 користувачам. Цей набір даних є незбалансованим і відображає фактичний розподіл соціальної мережі за гендером та віком, з домінуючою присутністю жінок над чоловіками та осіб середнього віку над молоддю. З цим набором даних авторами проведено експерименти з різноманітними моделями машинного навчання, різними функціями та врахували набір показників продуктивності.

Авторами публікації [13] представляємо підхід до завдання з профілювання авторів PAN 2019. Завдання розділено на дві підпроблеми, бот і визначення гендеру, для двох різних мов: англійської та іспанської. Для кожного випадку проблеми та кожної мови автори вирішують проблему по-різному. Запропоновано використовувати архітектуру ансамблю для вирішення проблеми виявлення ботів для облікових записів, які пишуть англійською мовою, і один SVM для тих, хто пише іспанською. Для визначення гендеру запропоновано використання єдиної архітектури SVM для обох мов, але попередньо оброблення твітів відбувається іншим способом. Остаточні моделі авторів досягають точності понад 90% у завданні виявлення ботів, а для визначення гендеру – 84,17% та 77,61% відповідно для англійської та іспанської мов.

Авторами [14] повідомляється про участь у завданні «Боти та гендерне профілювання» на PAN 2019. Запропонована методологія виявлення ботів використовує переваги текстової та стилістичної інформації твітів. Для завдання виявлення ботів дослідниками запропоновано використати лінійний класифікатор Support Vector Machine, навчений на словах і грамах символів, а також на додаткових функціях, які представляють варіації у використанні різних стилістичних особливостей. Для гендерної ідентифікації запропоновано

використати класифікатор стохастичного градієнтного спуску, SGD навчений на словах і грамах символів, настроях твітів і поточковій взаємній інформації (PMI) термінів. Функції PMI представляють важливість термінів для кожної статі. Авторам вдалося досягти середніх показників точності 0,881 і 0,7105 для завдань бота та гендерної ідентифікації відповідно.

У роботі авторів [15] представлено підхід до завдання з профілювання авторів PAN 2019 на основі архітектури ансамблів для вирішення проблеми виявлення ботів для облікових записів, які пишуть англійською мовою, і один SVM для тих, хто пише іспанською. Для визначення гендеру автори використовують єдину архітектуру SVM для обох мов, але попередньо обробляють твіти іншим способом. Остаточні моделі досягають точності понад 90% у завданні виявлення ботів, а для визначення гендеру – 84,17% та 77,61% відповідно для англійської та іспанської мов.

Окрім уваги науковців, є також деякі програмні рішення. Наприклад, Genderize.io, сервіс, який надає API для визначення гендера за іменем або текстом. Він дозволяє автоматично визначати гендер для великої кількості текстів безпосередньо через програмний інтерфейс. Основна функціональність Genderize.io полягає у визначенні гендера за іменами, але він також може бути використаний для аналізу тексту та визначення гендерних характеристик на основі змісту тексту. Вигляд сайту з API наведено на рисунку 1.2.

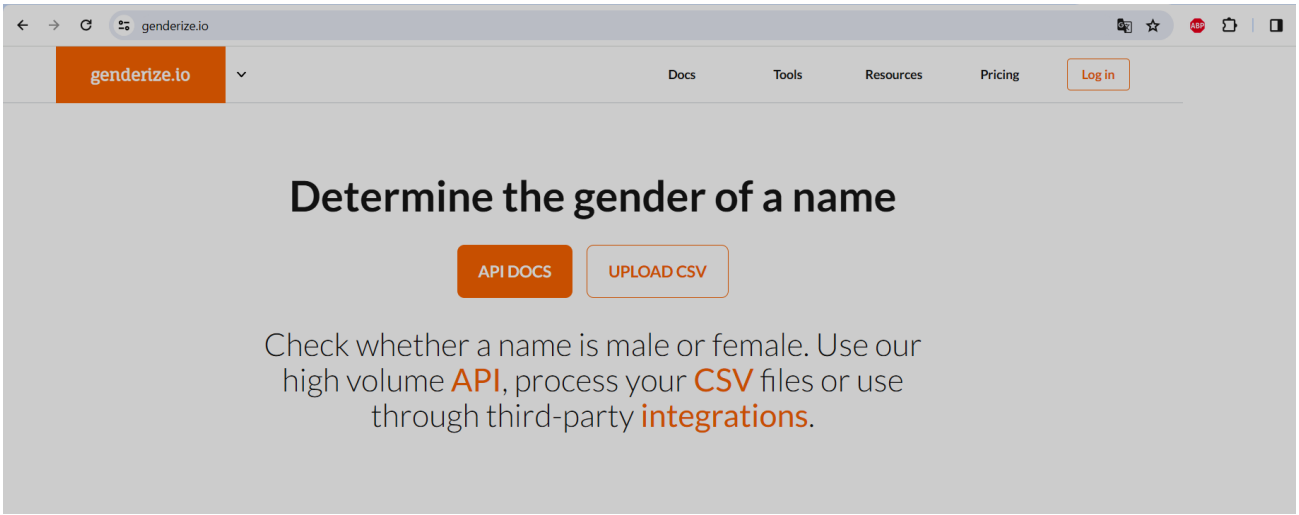


Рисунок 1.2 – Сервіс Genderize.io [16]

Робота з Genderize.io зазвичай виконується через HTTP-запити до їхнього API, де ви передаєте ім'я або текст, для якого потрібно визначити гендер. У відповідь сервіс повертає інформацію про статистичний розподіл гендера для поданого імені або тексту, вказуючи ймовірності належності до чоловічого або жіночого гендерів.

Цей сервіс може бути корисним у різних сферах, де потрібно автоматично визначати гендерні характеристики для великих обсягів даних, таких як аналіз соціальних мереж, дослідження аудиторії або категоризація текстів за гендером. Такий інструмент дозволяє ефективно обробляти великі обсяги даних і швидко отримувати результати визначення гендера для кожного тексту або імені.

Ще одним сервісом є Hacker Factor: Gender Guesser. Цей інструмент намагається визначити гендер автора на основі використаних слів. Він аналізує поданий текст за двома типами письма: формальним та неформальним. Зображення сайту наведено на рисунку 1.3.

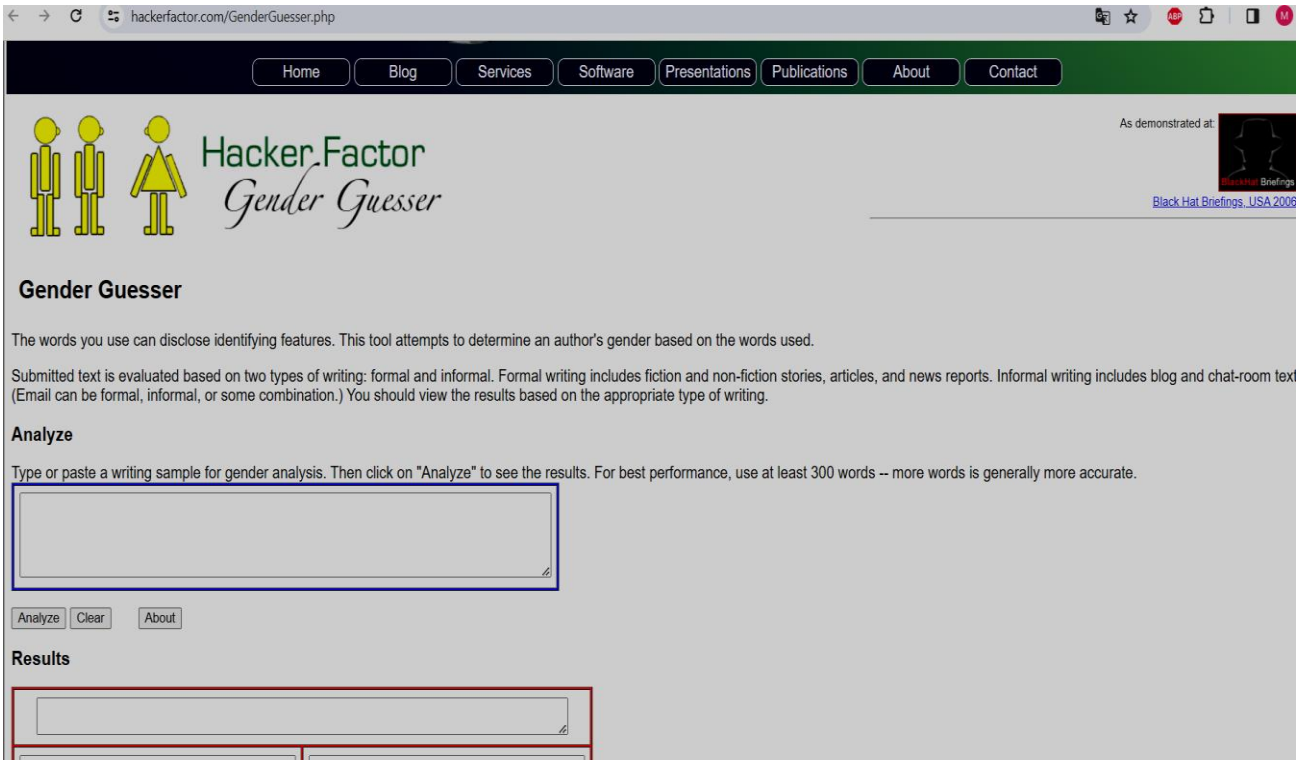


Рисунок 1.3 – Інтерфейс Hacker Factor: Gender Guesser [17]

Формальне письмо включає художню та наукову літературу, статті та новини. Неформальне письмо включає тексти блогів та чат-кімнат 1. Цей метод базується на роботі вчених з Іллінойського технологічного інституту та Бар-Іланського університету, які розробили метод оцінки гендеру на основі використання слів. Вони враховували вагові частоти слів та частин мови, щоб визначити гендер автора. Їхні дослідження показали, що навіть у різних жанрах письма (наприклад, художня та наукова література, блоги) є специфічні частоти слів для кожної статі.

З урахуванням надбань науковців у галузі виявлення гендерної приналежності за дописами соціальних інтернет-мереж, здебільшого науковцями використовується обраний вище підхід на основі SVM. Зважаючи на значну увагу науковців, напрямок актуальний та потребує продовження досліджень.

1.4 Мета та задачі кваліфікаційної роботи бакалавра

Мета кваліфікаційної роботи бакалавра полягає в розробці методу виявлення гендерної приналежності за дописами соціальних інтернет-мереж засобами NLP, та відповідного програмного забезпечення.

Для досягнення поставленої мети слід виконати такі задачі:

– виконати аналіз предметної області виявлення гендерної приналежності за текстовими даними;

– виконати огляд теоретичних підходів щодо можливості виявлення гендерної приналежності за дописами соціальних інтернет-мереж, обрати підхід для подальшої реалізації;

– виконати огляд існуючих наукових надбань та програмних рішень;

– створити метод виявлення гендерної приналежності за дописами соціальних інтернет-мереж засобами NLP та описати його кроки;

– спроектувати інформаційну систему для реалізації методу виявлення гендерної приналежності за дописами соціальних інтернет-мереж засобами NLP;

– створити програмну реалізацію за спроєктованою структурою інформаційної системи;

– виконати тестування створеної програмної реалізації;

– виконати дослідження ефективності методу виявлення гендерної приналежності за дописами соціальних інтернет-мереж засобами NLP на основі створеної інформаційної системи.

Розділ 2 Розробка методу виявлення гендерної приналежності за дописами соціальних інтернет-мереж засобами NLP

2.1 Схема та кроки методу виявлення гендерної приналежності за дописами соціальних інтернет-мереж

Метод виявлення гендерної приналежності за дописами соціальних інтернет-мереж засобами NLP призначений для аналізу та класифікації текстових даних з метою визначення гендерної приналежності, яка здійснюється на основі особливостей мови та стилю письма, що є характерними для певних груп користувачів. Схема та кроки методу наведені на рисунку 2.1.

Рисунок 2.1 – Схема та кроки методу виявлення гендерної приналежності за дописами соціальних інтернет-мереж

Метод виявлення гендерної приналежності за дописами соціальних інтернет-мереж працює шляхом перетворення вхідних даних у вигляді набору натренованих класифікаторів SVM та векторизаторів, що використовувались під час навчання та допису соціальних інтернет-мереж для аналізу у вихідні дані у вигляді оцінки приналежності допису для аналізу до визначеного гендеру.

Вхідними даними методу є набір навчених на англомовних даних класифікаторів SVM та відповідних векторизаторів, що використовувались під час навчання та допис соціальних інтернет-мереж для аналізу.

Нульовим кроком є вибір виду гендера для аналізу та завантаження відповідної моделі класифікатора SVM з векторизатором.

Першим кроком є попередня обробка допису соціальних інтернет-мереж, що включає в себе видалення стоп-символів та стоп-слів, а також перевірку мови написання. Якщо мова допису не англійська, відбувається автоматизований переклад на англійську.

Наступним кроком здійснюється векторизація попередньо обробленого допису соціальних інтернет-мереж, після чого відбувається крок класифікації натренованим SVM-класифікатором.

Четвертим кроком є формування висновків користувачу мовою оригіналу допису соціальних інтернет-мереж, оскільки сам класифікатор працює з англомовними даними.

Вихідними даними методу є оцінка приналежності допису для аналізу до визначеного гендеру.

Отже, вище наведено та описано схему та кроки методу виявлення гендерної приналежності за дописами соціальних інтернет-мереж засобами NLP, який своїм призначенням має класифікацію вхідних текстових даних з метою визначення гендерної приналежності, яка здійснюється шляхом перетворення вхідних даних у вигляді набору натренованих класифікаторів SVM та відповідних векторизаторів, що використовувались під час навчання та допису

соціальних інтернет-мереж для аналізу у вихідні дані у вигляді оцінки приналежності допису для аналізу до визначеного гендеру.

2.2 Схема навчання типового класифікатора

В якості вхідних даних методу виявлення гендерної приналежності за дописами соціальних інтернет-мереж засобами NLP є набір натренованих класифікаторів та відповідних векторизаторів, які використовувались при навчанні. Відповідно, нижче наведено схему та кроки навчання типового класифікатора SVM (рисунок 2.2).

Рисунок 2.2 – Схема та кроки навчання типового класифікатора

Вхідними даними є датасет та набори параметри SVM, такі як тип ядра, параметр регуляризації C та параметр ступеню впливу одного тренувального зразка. Для підбору оптимальних параметрів за метрикою точності, будуть зібрані декілька комбінацій.

Першим кроком є попередня обробка датасету, що включає в себе визначення мови кожного запису з метою відбору тільки англомовних зразків, видалення стоп-символів та стоп-слів.

Наступним кроком є поділ датасету на тренувальну та тестову вибірки. Вибірки поділяються у співвідношенні 80% на 20%, де 80% це тренувальні дані, а 20% – тестові.

Після поділу, вибірки підлягають процесу векторизації методом TF-IDF. Векторизовані дані використовуються для навчання класифікатора із застосуванням перехресної перевіркидля підбору найкращих гіперпараметрів моделі SVM.

Наступним кроком є крок збереження класифікатора із підібраними гіперпараметрами, які показали найвищі показники метрики точності. Також окрім самого класифікатора зберігається і векторизатор.

Вихідними даними є векторизатор та навчена модель SVM, яка знайдена шляхом перехресної перевіркидля підбору найкращих гіперпараметрів.

Отже, наведено схему та кроки навчання типового класифікатора SVM, який є вхідними даними методу виявлення гендерної приналежності за дописами соціальних інтернет-мереж засобами NLP.

2.3 Аналіз та автоматизація обробки потоків даних

Ще одним важливим етапом проєктування інформаційних систем є автоматизація обробки потоків даних. Схема навігації між інтерфейсними формами інформаційної системи наведена на рисунку 2.3.

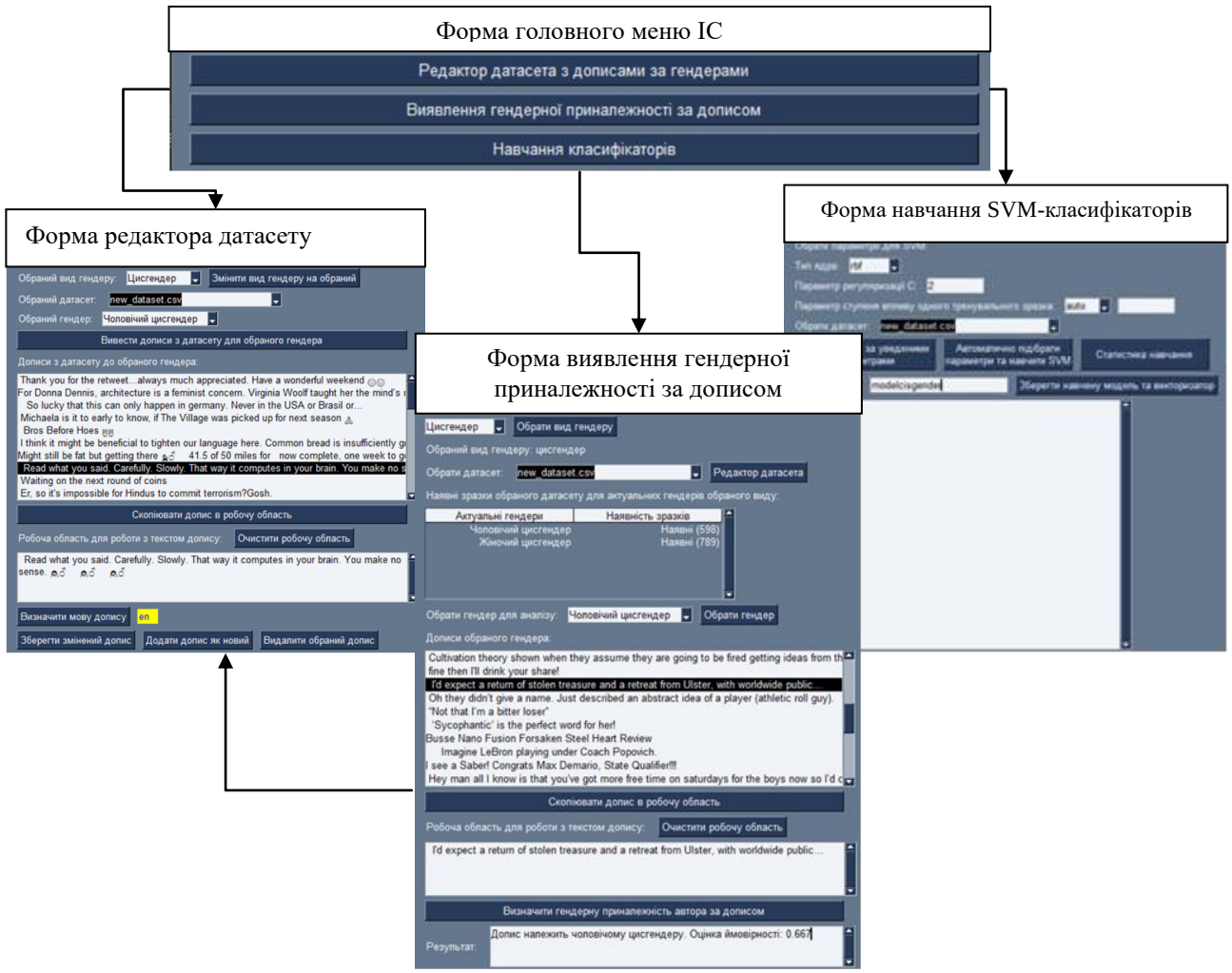


Рисунок 2.3 – Схема навігації між інтерфейсними формами інформаційної системи

Інформаційна система буде складатись із 4-х інтерфейс них форм: «Головне меню», «Редактор датасету», «Виявлення гендерної приналежності за дописом», «Навчання SVM-класифікаторів».

З форми головного меню повинно бути організовано переходи на решту форм інформаційної системи. Спрацьовувати переходи повинні по подіям натиснення відповідних контролів з назвами підсистем.

Також повинен бути реалізований перехід з форми виявлення гендерної приналежності за дописом на форму редактора датасету, що повинен спрацьовувати по події натиснення на контрол «Редактор датасету» форми виявлення гендерної приналежності за дописом.

Отже, таким чином описано та наведено схему навігації між інтерфейсними формами інформаційної системи, що буде складатись із 4-х інтерфейсних форм: «Головне меню», «Редактор датасету», «Виявлення гендерної приналежності за дописом», «Навчання SVM-класифікаторів».

2.4 Проектна архітектура інформаційної системи виявлення гендерної приналежності за дописами та взаємозв'язок компонентів

На рисунку 2.4 наведено проектну архітектуру інформаційної системи та взаємозв'язок компонентів. Наведена ІС складається із 3-х підсистем та бази даних.

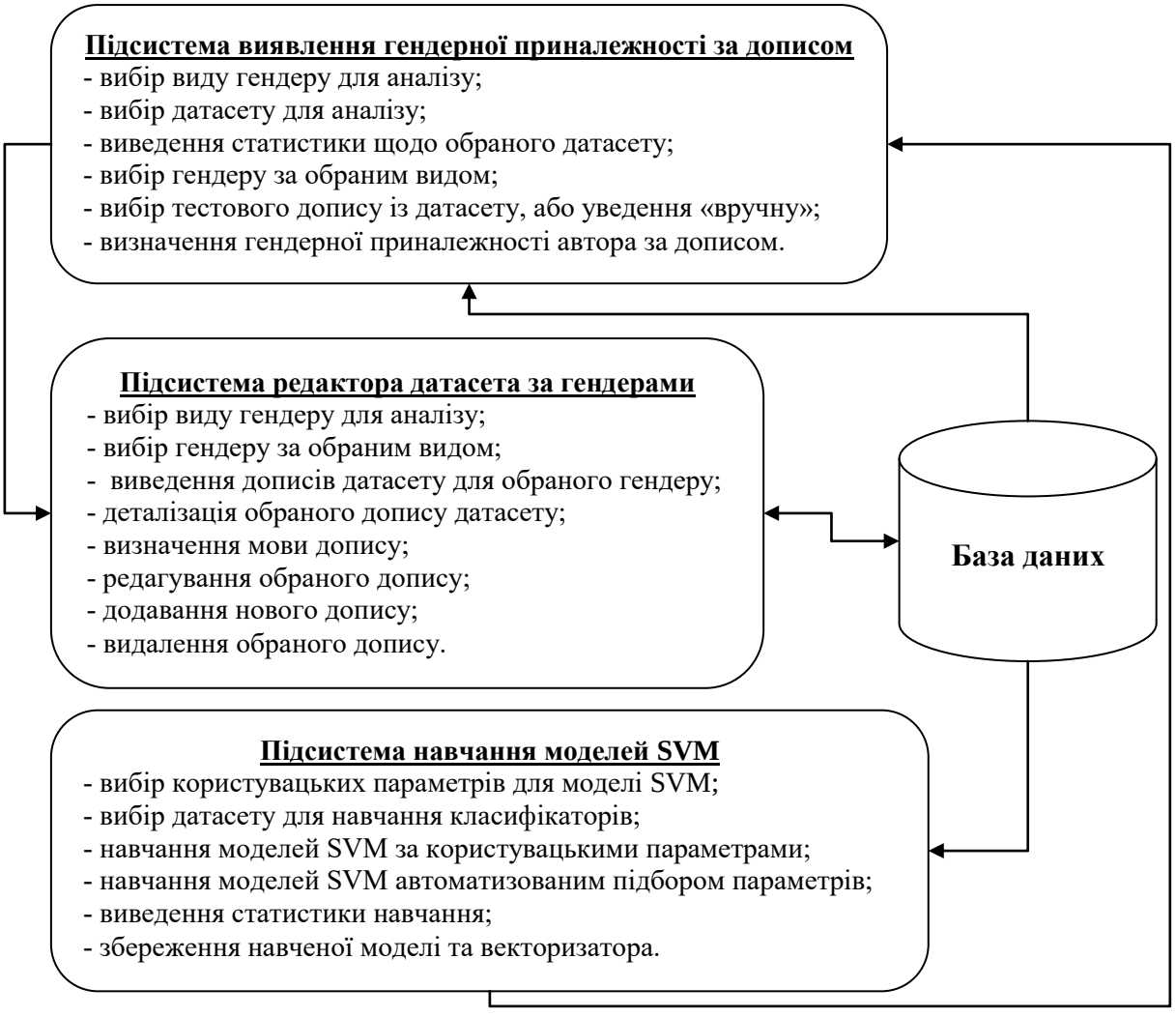


Рисунок 2.4 – Схема інформаційної системи виявлення гендерної приналежності за дописами

Підсистема виявлення гендерної приналежності за дописом є головною підсистемою інформаційної системи, та призначена для виявлення гендеру за користувацьким текстовим дописом. Також включає в себе можливості вибору виду гендеру для аналізу, вибір датасету для аналізу, виведення статистики щодо обраного датасету, вибір гендеру за обраним видом, вибір тестового допису із датасету або уведення «вручну», визначення гендерної приналежності автора за дописом. Аналіз допису для виявлення гендеру здійснюється на підставі класифікатора SVM, попередньо натренованого засобами підсистеми навчання моделей SVM.

Підсистема навчання моделей SVM призначена для тренування SVM-моделей, які призначені для виявлення гендерів. Для кожного виду гендеру повинен бути натренований свій класифікатор та відповідний векторизатор. Також дана підсистема включає в себе такі можливості, як вибір користувацьких параметрів для моделі SVM, вибір датасету для навчання класифікаторів, навчання моделей SVM за користувацькими параметрами, навчання моделей SVM автоматизованим підбором параметрів, виведення статистики навчання, збереження навченої моделі та векторизатора.

Підсистема редактора датасета за гендерами призначена для редагування датасетів, що використовуються для навчання та тестування класифікаторів. Вмісти датасетів та решти параметрів класифікаторів знаходяться в базі даних, тому дана підсистема є також редактором вмісту бази даних. Також виконує такі функції: вибір виду гендеру для аналізу, вибір гендеру за обраним видом, виведення дописів датасету для обраного гендеру, деталізація обраного допису датасету, визначення мови допису, редагування обраного допису, додавання нового допису, видалення обраного допису.

На рисунку 2.5 зображено схему бази даних інформаційної системи виявлення гендерної приналежності за дописами, яка візуально демонструє структуру бази даних, включаючи її таблиці, поля та зв'язки між ними.

Модель бази даних включає такі таблиці: genders, subject_types, webservices, subjects, posts_for_analysis, post_analysis, posts_for_training, subject_gender_analysis, gender_analysis_for_1_subject, types_of_genders.

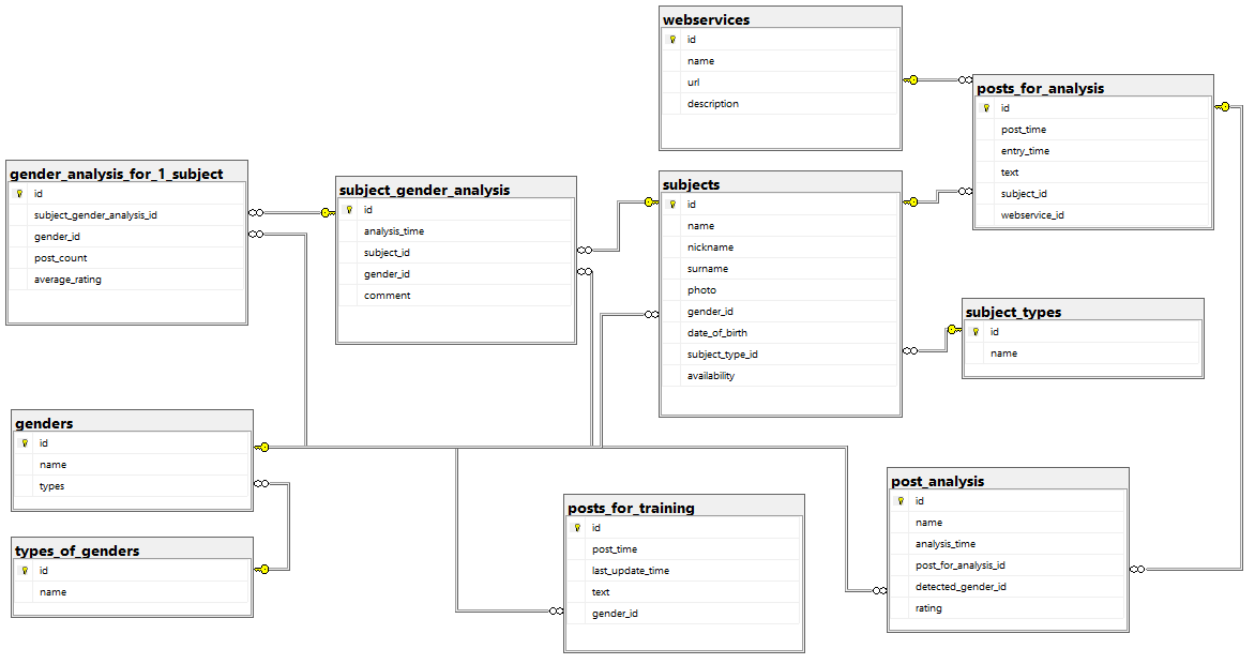


Рисунок 2.5 – Схема бази даних інформаційної системи виявлення гендерної приналежності за дописами

Так, таблиця 2.1 містить інформацію про гендери, що використовуються в системі.

Таблиця 2.1 – Атрибути таблиці «Genders»

№ п/п	Назва	Тип даних	Опис
1	id	INT	Унікальний ідентифікатор гендеру
2	name	VARCHAR	Назва гендеру
3	Types_id	INT	Вид гендеру

Таблиця 2.2 описує атрибути, які використовуються для зберігання інформації про суб'єктів, що підлягатимуть аналітичній обробці. Включає його ім'я, нікнейм, прізвище, фотографію, ідентифікатор гендеру, дату народження,

тип суб'єкта та його доступність. Детальний опис атрибутів цієї таблиці БД подано в таблиці 2.2.

Таблиця 2.2 – Атрибути таблиці «Subjects»

№ п/п	Назва	Тип даних	Опис
1	id	INT	Унікальний ідентифікатор суб'єкта
2	name	VARCHAR	Ім'я суб'єкта
3	nickname	VARCHAR	Нікнейм суб'єкта
4	surname	VARCHAR	Прізвище суб'єкта
5	photo	VARCHAR	Фотографія суб'єкта
6	gender_id	INT	Гендер
7	date_of_birth	DATE	Дата народження суб'єкта
8	subject_type_id	INT	Тип суб’єкта
9	availability	BIT	Доступність суб'єкта

Таблиця 2.3 містить структуровані дані про публікації, які підлягають аналітичній обробці. До неї включено такі атрибути: дату та час публікації, дату та час внесення публікації в базу даних, текст публікації, суб'єкт, який є власником публікації, та вебсервіс розміщення публікації.

Таблиця 2.3 – Атрибути таблиці «Posts_for_analysis»

№ п/п	Назва	Тип даних	Опис
1	id	INT	Унікальний ідентифікатор публікації
2	post_time	DATETIME	Дата і час публікації
3	entry_time	DATETIME	Дата і час внесення публікації в базу даних
4	text	TEXT	Текст публікації
5	subject_id	INT	Власник публікації
6	webservice_id	INT	Вебсервіс розміщення публікації

Таблиця 2.4 описує дані про вебсервіси. Включає назву вебсервісу, гіперпосилання на нього та його опис.

Таблиця 2.4 – Атрибути таблиці «Webservices»

№ п/п	Назва	Тип даних	Опис
1	id	INT	Унікальний ідентифікатор вебсервісу
2	name	VARCHAR	Назва вебсервісу
3	url	VARCHAR	Гіперпосилання на вебсервіс
4	description	TEXT	Опис вебсервісу

Таблиця 2.5 описує атрибути, які використовуються для зберігання інформації про результати аналізу публікацій: назву аналізу, його дату та час, публікацію для аналізу, гендер власника публікації, та її оцінку.

Таблиця 2.5 – Атрибути таблиці «Post_analysis»

№ п/п	Назва	Тип даних	Опис
1	id	INT	Унікальний ідентифікатор аналізу публікації
2	name	VARCHAR	Назва аналізу публікації
3	analysis_time	DATETIME	Дата і час аналізу
4	post_for_analysis_id	INT	Публікація для аналізу
5	detected_gender_id	INT	Гендер власника публікації
6	rating	FLOAT	Оцінка публікації

Таблиця 2.6 надає пояснення атрибутів, що використовуються для опису публікацій, які застосовуються для тренування моделі: дату та час публікації, дату та час останнього оновлення публікації, текст публікації та гендер власника публікації.

Таблиця 2.6 – Атрибути таблиці «Posts_for_training»

№ п/п	Назва	Тип даних	Опис
1	id	INT	Унікальний ідентифікатор публікації для навчання
2	post_time	DATETIME	Дата і час публікації
3	last_update_time	DATETIME	Дата і час останнього оновлення публікації
4	text	TEXT	Текст публікації
5	gender_id	INT	Гендер власника публікації

Таблиця 2.7 містить інформацію про аналітичні процедури, що стосуються гендерної ідентифікації суб'єктів. Включаючи дату та час підрахунку, суб’єкт та його гендер, а також коментар до аналізу.

Таблиця 2.7 – Атрибути таблиці «Subject_gender_analysis»

№ п/п	Назва	Тип даних	Опис
1	id	INT	Унікальний ідентифікатор аналізу приналежності суб'єкта до гендеру
2	analysis_time	DATETIME	Дата і час підрахунку
3	subject_id	INT	Суб’єкт
4	gender_id	INT	Гендер
5	comment	VARCHAR	Коментар до аналізу

Таблиця 2.8 містить узагальнену інформацію про результати аналізу гендерної ідентифікації суб'єктів. Вона включає в собі аналіз приналежності суб'єкта до гендеру, гендер суб’єкта, кількість публікацій, що вказують на цей гендер, та середню оцінку всіх аналізів, що вказують на цей гендер.

Таблиця 2.8 – Атрибути таблиці «Gender_analysis_for_1_ subject»

№ п/п	Назва	Тип даних	Опис
1	id	INT	Унікальний ідентифікатор аналізу приналежності суб'єкта до одного гендеру
2	subject_gender_analysis_id	INT	Аналіз приналежності субєкта до гендеру
3	gender_id	INT	Гендер суб’єкта
4	post_count	INT	Кількість публікацій, що вказують на цей гендер
5	average_rating	FLOAT	Середня оцінка всіх аналізів, що вказують на цей гендер

Таблиця 2.9 описує категорії типів суб'єктів.

Таблиця 2.9 – Атрибути таблиці «Subject_types»

№ п/п	Назва	Тип даних	Опис
1	id	INT	Унікальний ідентифікатор типу суб'єкта
2	name	VARCHAR	Назва типу суб'єкта

Таблиця 2.10 описує атрибути, які використовуються для зберігання інформації про гендерні категорії.

Таблиця 2.10 – Атрибути таблиці «Types_of_genders»

№ п/п	Назва	Тип даних	Опис
1	id	INT	Унікальний ідентифікатор виду гендера
2	Name	VARCHAR	Назва виду гендера

Таким чином було наведено проектну архітектуру інформаційної системи виявлення гендерної приналежності за дописами та описано взаємозв’язок компонентів. Наведена інформаційна система виявлення гендерної приналежності за дописами складається із 3-х підсистем: «Підсистеми виявлення гендерної приналежності за дописом», Підсистеми навчання моделей SVM», «Підсистеми редактора датасета за гендерами» та бази даних.

2.5 Підготовка робочих вхідних даних для інформаційної системи виявлення гендерної приналежності за дописами

Зважаючи на специфіку предметної області, однозначного датасету, який би покрив усю множину тендерів на сьогоднішній день не існує. Тому в рамках кваліфікаційної роботи бакалавра буде використано набір даних для виду гендеру «цисгендер» , що має назву «Tweet Files for Gender Guessing» [18], також для небінарного виду гендеру було сформовано датасет «NONEBinaryData» засобами використання мовної моделі штучного інтелекту ChatGPT 3.5 [19].

Набір даних містить тексти твітів, завантажених через Twitter API між 2019-05-21 і 2019-06-01, з ідентифікатором користувача та міткою часу для ідентифікації кожного твіту, а також індикатором того, чи користувач (на основі відображуваного імені) може бути чоловіком або жінкою. Приклад даних наведено на рисунку 2.6

Твіти обмежуються оригінальними твітами англійською мовою (тобто не ретвітами), пов’язаними з користувачами, чиї імена за допомогою пакета визначення статі можна визначити як чоловічі чи жіночі. Твіти поділяються на навчальні, перевірочні та тестові набори за часом і ідентифікатором користувача. Твіти в кожному файлі вибрано так, щоб вони були рівномірно збалансовані між гендерами. Твіти попередньо оброблено (за допомогою сценарію, скопійованого з Тімоті Реннера), щоб видалити хештеги, згадки, URL-адреси, медіа, і символи.

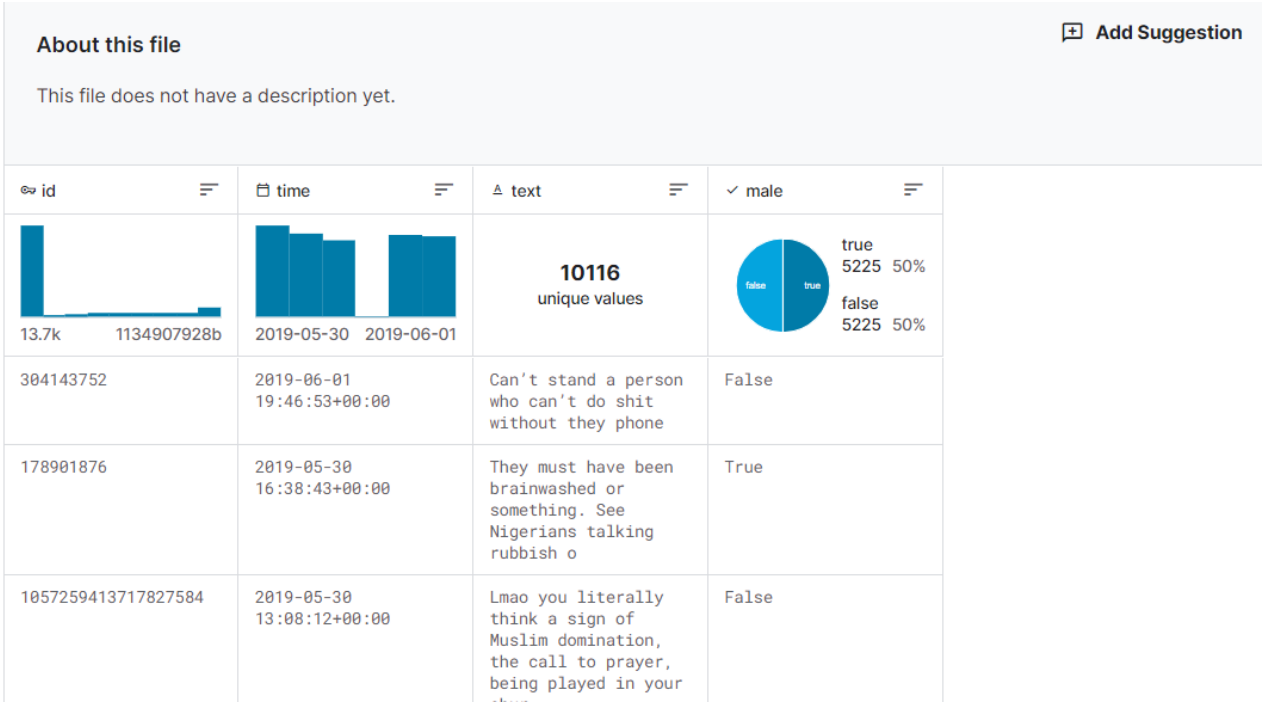


Рисунок 2.6 – Приклад даних датасету для інформаційної системи виявлення гендерної приналежності за дописами

Для навчання класифікатора SVM буде взято колонки «text» та «male». Тренувальний набір налічує близько 33 000 записів, рівномірно розподілених на 2 класи (чоловічі та жіночі записи). Статистика за довжиною записів наведена у таблиці 2.11.

Таблиця 2.11 – Статистика розподілу даних за довжиною

	Чоловічий цисгендер	Жіночий цисгендер
Середня довжина твіта в символах	60.357	59.4127
Мінімальна довжина твіта в символах	1	4
Максимальна довжина твіта в символах	148	153

Також було досліджено розподіл твітів за їх довжинами відносно цисгендеру, що наведено на рисунках 2.7 та 2.8 відповідно.

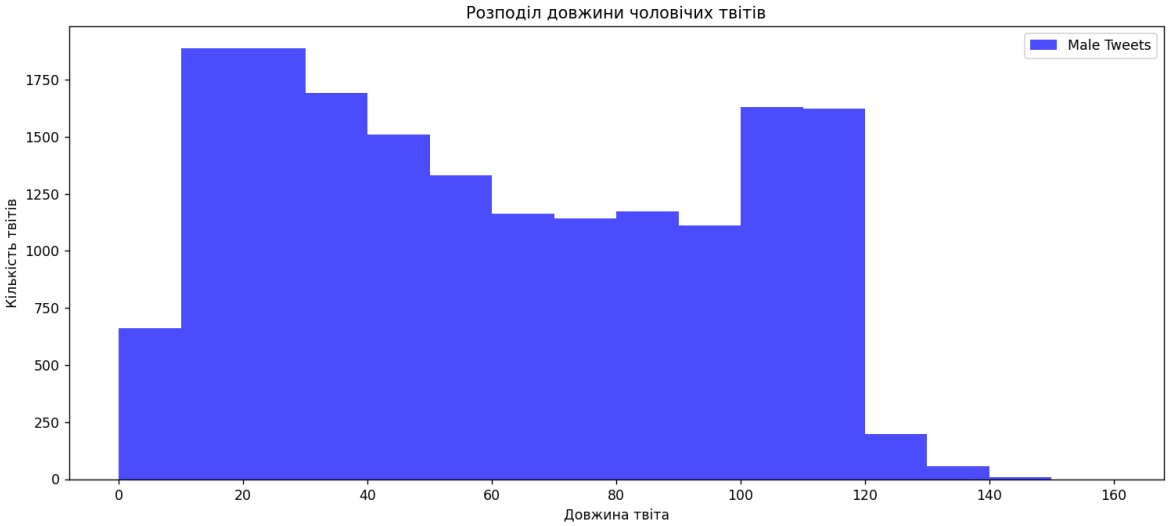


Рисунок 2.7 – Розподіл твітів чоловічого цисгендеру за довжиною

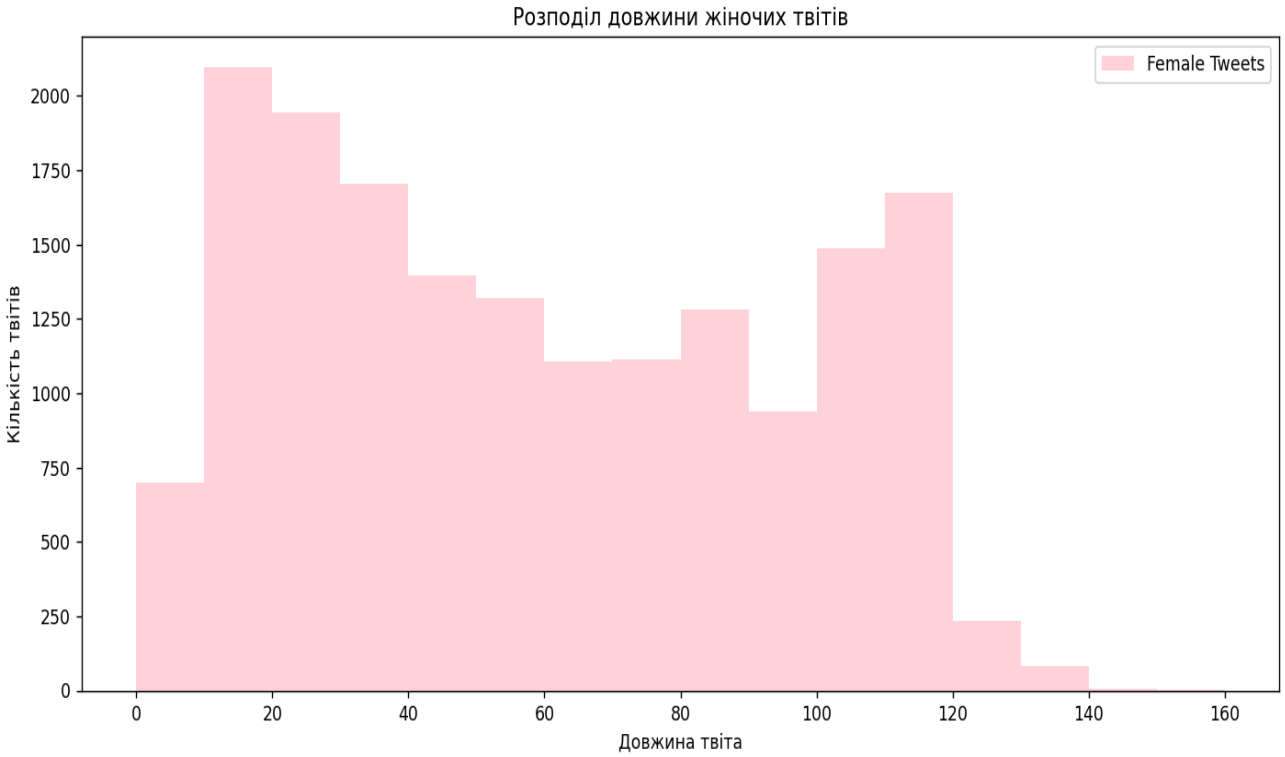


Рисунок 2.8 – Розподіл твітів жіночого цисгендеру за довжиною

В сукупності, середня довжина твіта в символах становить 59.885, і суттєвої різниці між довжиною записів чоловічого та жіночого цисгендеру у даному наборі даних немає.

Згідно до рисунків 2.7 та 2.8 – розподіл записів теж досить схожий з незначними відмінностями, в жінок спостерігається дещо більша кількість

записів у діапазоні від 10 до 40 символів, в той час як у чоловіків розподіл більш рівномірний.

Отже, для навчання та тестування класифікатора SVM буде використано набір даних «Tweet Files for Gender Guessing», що налічує близько 33 000 записів, а також власне створений набір даних «NONEBinaryData», що налічує 350 записів, що рівномірно розподілені між такими гендерами, як: агендер, андроген, бігендер, гендерфлюїд, гендерний вигнанець, омнігендер, полігендер та пангендер.

2.6 Особливості використання спеціалізованих програмних компонентів

Для спрощення процесу розробки спроектованої інформаційної системи буде використано ряд спеціалізованих програмних компонентів.

Бібліотека Scikit-learn є однією з найпопулярніших бібліотек машинного навчання у середовищі Python. Вона надає широкий спектр інструментів для класифікації, регресії, кластеризації, а також інших задач аналізу даних та машинного навчання. Ця бібліотека базується на ідеї простоти використання та консистентності інтерфейсів, що робить її дуже привабливою для початківців та досвідчених дослідників [20].

Основними компонентами Scikit-learn є алгоритми машинного навчання, які лежать в основі різноманітних методів. Ці алгоритми охоплюють широкий спектр тематик, включаючи навчання з учителем (наприклад, класифікація та регресія), навчання без учителя (наприклад, кластеризація та знайдення асоціативних правил) та навчання з підгляду (наприклад, активне навчання та виявлення аномалій).

У складі бібліотеки також присутні інструменти для попередньої обробки даних, включаючи відбір ознак, стандартизацію та нормалізацію даних. Крім того, Scikit-learn надає засоби для оцінки моделей, включаючи різноманітні метрики ефективності, перехресну перевірку та оптимізацію гіперпараметрів.

Ще однією важливою характеристикою Scikit-learn є його активне спільнота та розширюваність. Багато алгоритмів та інструментів розроблені у вигляді модулів, які можна легко інтегрувати з іншими бібліотеками Python, такими як NumPy, SciPy та Pandas. Це робить Scikit-learn потужним і гнучким інструментом для вирішення різноманітних завдань у сфері машинного навчання та аналізу даних. Дана бібліотека буде використана для векторизації текстових даних, навчання класифікаторів SVM, та оцінок якостей навчених моделей шляхом застосування метрик.

Бібліотека Pandas є однією з найбільш популярних і потужних бібліотек для обробки та аналізу даних у середовищі Python. Вона надає зручний і ефективний інтерфейс для роботи з даними у вигляді табличних структур, що дозволяє легко завантажувати, обробляти та аналізувати дані з різних джерел [21].

Головними об'єктами у бібліотеці Pandas є DataFrame і Series. DataFrame – це двовимірна таблична структура даних, схожа на таблицю бази даних, яка містить рядки та стовпці. Series – це одновимірна маркована структура даних, що містить однорідні дані.

Pandas надає широкий спектр функцій для роботи з даними, включаючи завантаження даних з різних джерел (таких як CSV файли, бази даних SQL, Excel таблиці), очищення та підготовку даних (видалення дублікатів, обробка пропущених значень, перетворення типів даних), аналіз та візуалізацію даних (групування, сортування, обчислення статистик, побудова діаграм, графіків тощо). Дана бібліотека буде використана в роботі для обробки даних датасетів (зчитування та формування навчальної та валідаційної вибірок).

Бібліотека Langdetect є інструментом для визначення мови тексту у середовищі Python. Вона забезпечує швидке та ефективне визначення мови тексту на основі статистичних методів та аналізу символів у тексті.

Основним функціоналом бібліотеки Langdetect є можливість визначати мову тексту за допомогою вбудованих алгоритмів, які аналізують частоту вживання різних символів та слів у тексті. Ці алгоритми базуються на

статистичних моделях, які навчаються на великих корпусах текстів у різних мовах [22].

Крім того, бібліотека Langdetect підтримує роботу з текстами різної складності та довжини, що робить її універсальним інструментом для визначення мови текстів у різних сценаріях застосування. Буде використана в роботі для визначення мови допису.

PySimpleGUI – це бібліотека для створення графічного інтерфейсу користувача у Python, яка відома своєю простотою використання та широким спектром можливостей. Ця бібліотека дозволяє створювати інтерфейси для десктопних застосунків швидко та легко, навіть для початківців у програмуванні [23].

Основними особливостями PySimpleGUI є його простий та зрозумілий синтаксис, який дозволяє швидко створювати різноманітні вікна, кнопки, поля вводу та інші елементи інтерфейсу. Вона також має можливості для організації розташування елементів у вікні, обробки подій користувача та взаємодії з іншими бібліотеками Python.

Крім того, PySimpleGUI підтримує крос-платформеність, що означає, що інтерфейси, створені за допомогою цієї бібліотеки, можуть працювати на різних операційних системах, таких як Windows, macOS та Linux, без змін в коді. Дана бібліотека буде використана для створення інтерфейсу користувача.

Отже, з вищерозглянутих бібліотек Scikit-learn буде використана для векторизації текстових даних, навчання класифікаторів SVM, та оцінок якостей навчених моделей шляхом застосування метрик; Pandas буде використана в роботі для обробки даних датасетів (зчитування та формування навчальної та валідаційної вибірок); Langdetect буде використана в роботі для визначення мови допису, а PySimpleGUI буде використана для створення інтерфейсу користувача.

2.7 Висновки до розділу 2

Під час виконання другого розділу було створено метод виявлення гендерної приналежності за дописами соціальних інтернет-мереж засобами NLP, а також наведено схему та описані основні його кроки. Метод виявлення гендерної приналежності за дописами соціальних інтернет-мереж засобами NLP призначений для аналізу та класифікації текстових даних з метою визначення гендерної приналежності, яка здійснюється шляхом перетворення вхідних даних у вигляді набору натренованих класифікаторів SVM та відповідних векторизаторів, що використовувались під час навчання та допису соціальних інтернет-мереж для аналізу у вихідні дані у вигляді оцінки приналежності допису для аналізу до визначеного гендеру.

Також окремим компонентом наведено схему та кроки навчання типового класифікатора SVM, який є вхідними даними методу виявлення гендерної приналежності за дописами соціальних інтернет-мереж засобами NLP.

Проведено аналіз та автоматизацію обробки потоків даних, в рамках якого наведено схему навігації між інтерфейсними формами інформаційної системи виявлення гендерної приналежності за дописами що буде складатись із 4-х інтерфейсних форм: «Головне меню», «Редактор датасету», «Виявлення гендерної приналежності за дописом», «Навчання SVM-класифікаторів».

Наведено проектну архітектуру інформаційної системи виявлення гендерної приналежності за дописами та описано взаємозв’язок компонентів. Наведена інформаційна система виявлення гендерної приналежності за дописами складається із 3-х підсистем: «Підсистеми виявлення гендерної приналежності за дописом», Підсистеми навчання моделей SVM», «Підсистеми редактора датасета за гендерами» та бази даних.

Здійснено підготовку навчальних даних для інформаційної системи виявлення гендерної приналежності за дописами. Для навчання та тестування класифікатора SVM буде використано набір даних «Tweet Files for Gender

Guessing», що налічує близько 33 000 записів, а також власне створений набір даних «NONEBinaryData», що налічує 350 записів, що рівномірно розподілені між такими гендерами, як: агендер, андроген, бігендер, гендерфлюїд, гендерний вигнанець, омнігендер, полігендер та пангендер.

За розробленим методом необхідно створити програмний застосунок, який буде призначений для дослідження ефективності запропонованого методу. Також для доведення коректності результатів його треба окремо функціонально дослідити й протестувати.

Розділ 3 Експериментальне дослідження методу виявлення гендерної приналежності за дописами соціальних інтернет-мереж

3.1 Визначення шляхів дослідження та засобів створення інформаційної системи виявлення гендерної приналежності

Для дослідження методу виявлення гендерної приналежності за дописами соціальних інтернет-мереж необхідно створити програмне забезпечення у вигляді інформаційної системи виявлення гендерної приналежності за дописами, що зможе забезпечити виконання таких основних функцій:

- аналіз допису з бази дописів за обраним гендером;
- вибір датасету для аналізу;
- виведення статистики щодо обраного датасету;
- визначення гендерної приналежності автора за дописом.
- виведення дописів датасету для обраного гендеру;
- визначення мови допису;
- редагування обраного допису;
- додавання нового допису;
- видалення обраного допису;
- вибір користувацьких параметрів для моделі SVM;
- вибір датасету для навчання класифікаторів;
- навчання моделей SVM за користувацькими параметрами;
- навчання моделей SVM автоматизованим підбором параметрів;
- виведення статистики навчання;
- збереження навченої моделі та векторизатора.

Коректність виконання заявлених функцій інформаційної системи виявлення гендерної приналежності за дописами необхідно перевірити з використанням тест-кейсів.

Коректність навчання моделей SVM буде досліджено із застосуванням метрики accuracy. Дослідження ефективності планується виконати шляхом порівняння з розглянутим вище аналогом – «Gender Guesser».

3.2 Вибір засобів розробки інформаційної системи виявлення гендерної приналежності за дописами

Для розробки інформаційної системи виявлення гендерної приналежності за дописами для дослідження ефективності запропонованого методу буде використано інтегроване середовище програмування PyCharm, мову програмування Python та мову запитів SQL.

PyCharm представляє собою інтегроване середовище програмування, призначене для розробки програм на мові програмування Python [23]. Воно має ряд зручних функцій, таких як редагування коду з підсвічуванням синтаксису, автодоповнення коду, інтегровану підтримку систем контролю версій, відлагоджувач, інструменти для тестування коду та багато іншого. PyCharm надає зручний та продуктивний інтерфейс для роботи з Python-проектами будь-якої складності. Буде використано для створення застосунка.

Мова Python відома своєю простотою та легкістю використання, а також великою кількістю бібліотек для машинного навчання та обробки даних [24]. Такі бібліотеки, як scikit-learn, надають зручний інтерфейс для роботи з алгоритмами машинного навчання, включаючи метод опорних векторів. Python також є популярним серед дослідників та розробників у галузі машинного навчання та штучного інтелекту, що сприяє широкому сприйняттю та підтримці спільнотою. Зважаючи на попередній вибір спеціалізованих програмних розширень, та великий потенціал для програмування систем зі штучним інтелектом, для розробки застосунку буде використано саме цю мову програмування.

Мова SQL є стандартизованою мовою програмування, що призначена для взаємодії з реляційними базами даних [25]. За допомогою SQL можна

створювати, модифікувати та керувати базами даних, виконуючи різноманітні операції, такі як створення таблиць, вставка, оновлення та видалення даних, а також виконання складних запитів для отримання необхідної інформації. SQL має різні команди, такі як SELECT, INSERT, UPDATE, DELETE, які використовуються для роботи з даними в базі даних. Вона є основним інструментом для управління даними в реляційних базах даних, і буде використана для створення бази даних та взаємодії з нею.

Отже, буде використано інтегроване середовище програмування PyCharm, мову програмування Python та мову запитів SQL. Даний набір має якісну взаємодію між собою, та задовольняє використання спеціалізованих програмних розширень.

3.3 Структура та функціональне призначення програмних складових інформаційної системи виявлення гендерної приналежності

Програмні складові інформаційної системи виявлення гендерної приналежності за дописами відповідають розробленій вище проектній архітектурі системи. Схема діаграми класів наведена на рисунку 3.1.

Підсистема виявлення гендерної приналежності за дописом. Клас «GenderDetectionSystem» представляє основну підсистему, яка координує процес виявлення гендерної приналежності. Він містить інші класи, необхідні для виконання цього завдання: «GenderAnalyzer», «GenderClassifier», «Statistics».

Підсистема редактора датасета за гендерами. Клас «DataSetEditor» відповідає за редагування та управління датасетом. Він містить методи для додавання, редагування та видалення дописів, а також для визначення мови допису, такі як: editPost(), addPost(), deletePost(), determinePostLanguage().

Клас «PostDetails» відображає деталізовану інформацію про окремий допис, таку як текст та інші атрибути: displayPostDetails().

Підсистема навчання моделей SVM. Клас «SVMModelTrainer» відповідає за навчання моделей SVM та управління параметрами навчання. Він містить

методи для вибору параметрів, навчання моделей та збереження навчених моделей: selectParameters(), trainSVMModels(), saveModels().

Клас «ParameterSelector» допомагає вибрати параметри для навчання моделей selectParameters().

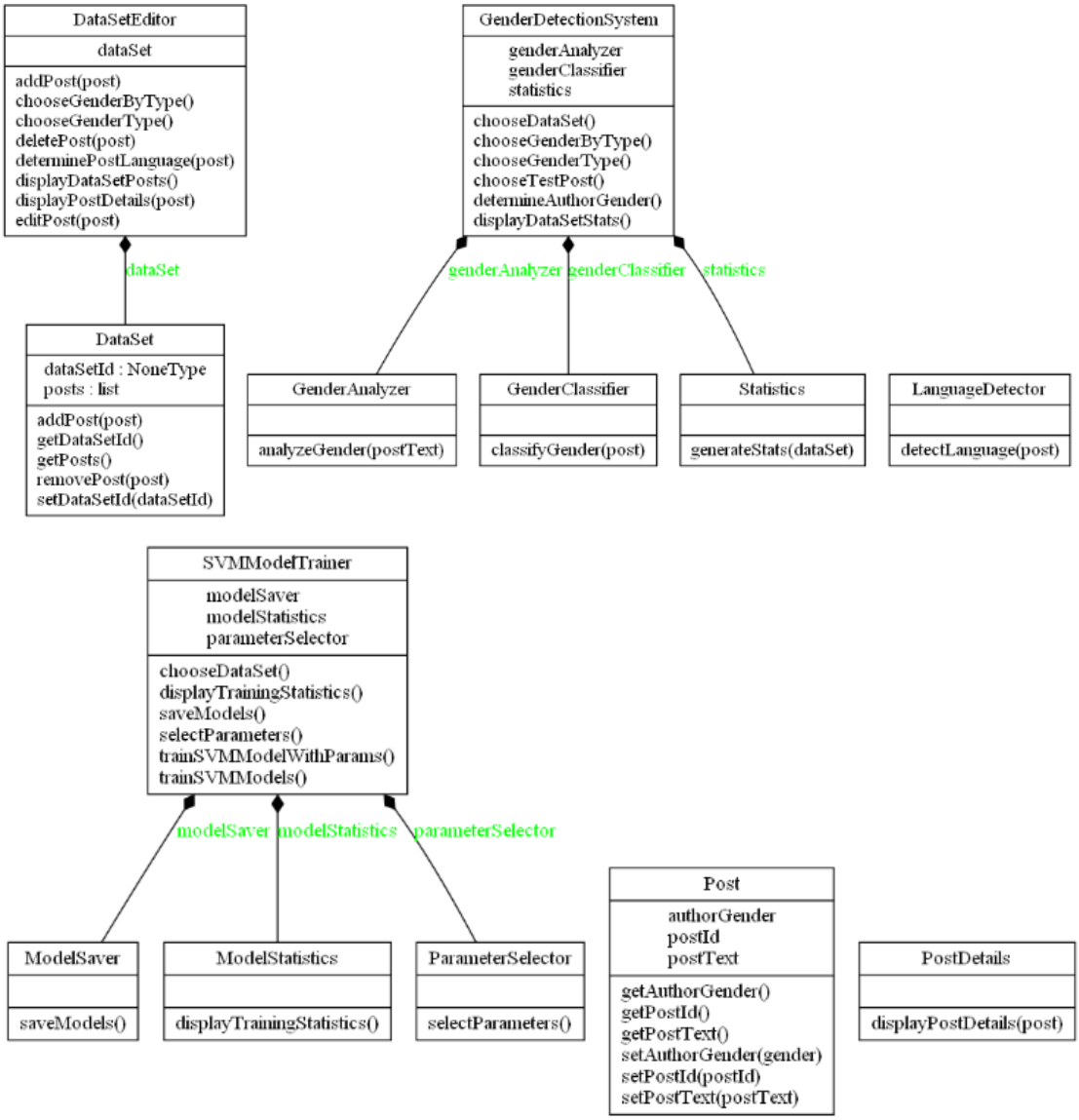


Рисунок 3.1 – Діаграма класів інформаційної системи виявлення гендерної приналежності за дописами

Клас «ModelStatistics» відображає статистику навчання, використовує метод displayTrainingStatistics().

Клас «ModelSaver» зберігає навчені моделі. Використовує метод saveModels().

Отже, було описано структуру та функціональне призначення програмних складових інформаційної системи виявлення гендерної приналежності за дописами соціальних інтернет-мереж.

3.4 Особливості реалізації програмних складових інформаційної системи виявлення гендерної приналежності за дописами

Програмні складові системи реалізовувались згідно описаної вище структури та функціонального призначення її складових. Оскільки застосунок містить модуль базу даних, то її спершу необхідно створити.

База даних створювалась з використанням бібліотеки SQLAlchemy, що призначена для роботи з базами даних у мові програмування Python. Вона надає зручний спосіб взаємодії з реляційними базами даних, дозволяючи створювати, зчитувати, оновлювати та видаляти дані з бази даних за допомогою об'єктно-реляційного відображення. Після створення БД, буде видно перелік існуючих таблиць (рисунок 3.2).

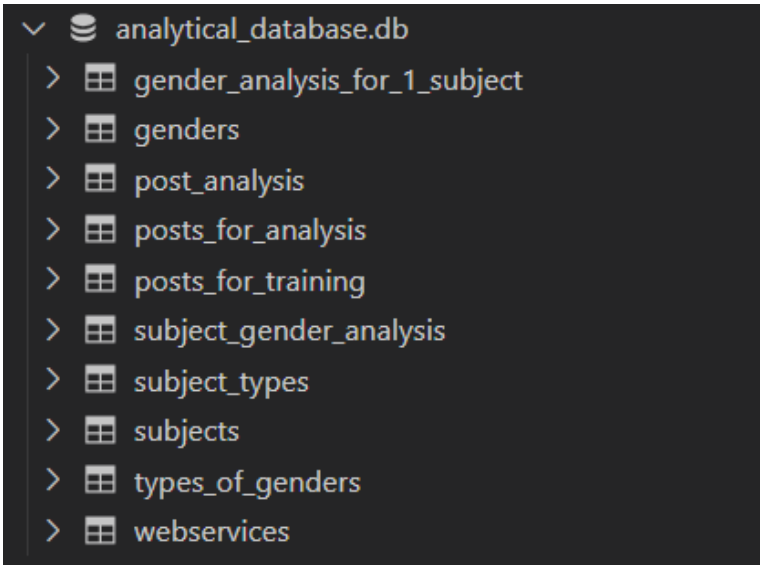
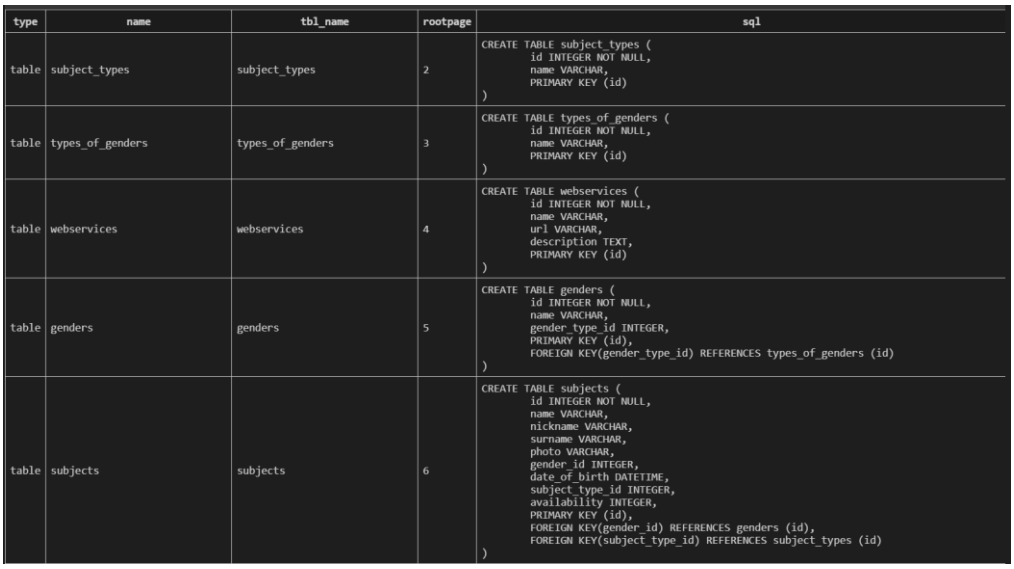


Рисунок 3.2 – Створення БД інформаційної системи виявлення гендерної приналежності за дописами

Сам процес створення наведено на рисунку 3.3.



type	name	tbl_name	rootpage	sql
table	subject_types	subject_types	2	CREATE TABLE subject_types (id INTEGER NOT NULL, name VARCHAR, PRIMARY KEY (id))
table	types_of_genders	types_of_genders	3	CREATE TABLE types_of_genders (id INTEGER NOT NULL, name VARCHAR, PRIMARY KEY (id))
table	webservices	webservices	4	CREATE TABLE webservices (id INTEGER NOT NULL, name VARCHAR, url VARCHAR, description TEXT, PRIMARY KEY (id))
table	genders	genders	5	CREATE TABLE genders (id INTEGER NOT NULL, name VARCHAR, gender_type_id INTEGER, PRIMARY KEY (id), FOREIGN KEY(gender_type_id) REFERENCES types_of_genders (id))
table	subjects	subjects	6	CREATE TABLE subjects (id INTEGER NOT NULL, name VARCHAR, nickname VARCHAR, surname VARCHAR, photo VARCHAR, gender_id INTEGER, date_of_birth DATETIME, subject_type_id INTEGER, availability INTEGER, PRIMARY KEY (id), FOREIGN KEY(gender_id) REFERENCES genders (id), FOREIGN KEY(subject_type_id) REFERENCES subject_types (id))

Рисунок 3.3 – Проміжний етап створення таблиць

Після створення БД її можна використовувати у застосунку в вигляді інформаційної системи виявлення гендерної приналежності за дописами. Сам застосунок складається із головного меню та трьох інтерфейсних форм, що призначені для реалізації функцій інформаційної системи.

Підсистема навчання класифікаторів у реалізації має 2 різних підходи: підхід із заданими користувацькими параметрами, та підхід із автоматичним підбором параметрів SVM.

Цікавим є підхід автоматичного підбору гіперпараметрів. Для його використання створюється об'єкт моделі SVM за допомогою конструктора SVC() з модуля svm, після чого відбувається створення словника параметрів для перебору. Для параметру ядра додаються значення 'linear', 'rbf', 'poly', а для параметру регуляризації C досліджуваний набір значень містив діапазон від 0.1 до 10.

Далі здійснюється пошук найкращих параметрів за допомогою перехресної перевірки (крос-валідація). Об'єкт GridSearchCV перебирає

комбінації параметрів, обчислюючи метрику точності за допомогою крос-валідації.

Виводяться найкращі параметри, знайдені під час перехресної перевірки, а також середнє значення метрики точності на кожному з розбиттів крос-валідації. Приклад виконання пошуку найкращих параметрів на визначеному наборі даних наведено на рисунку 3.4.

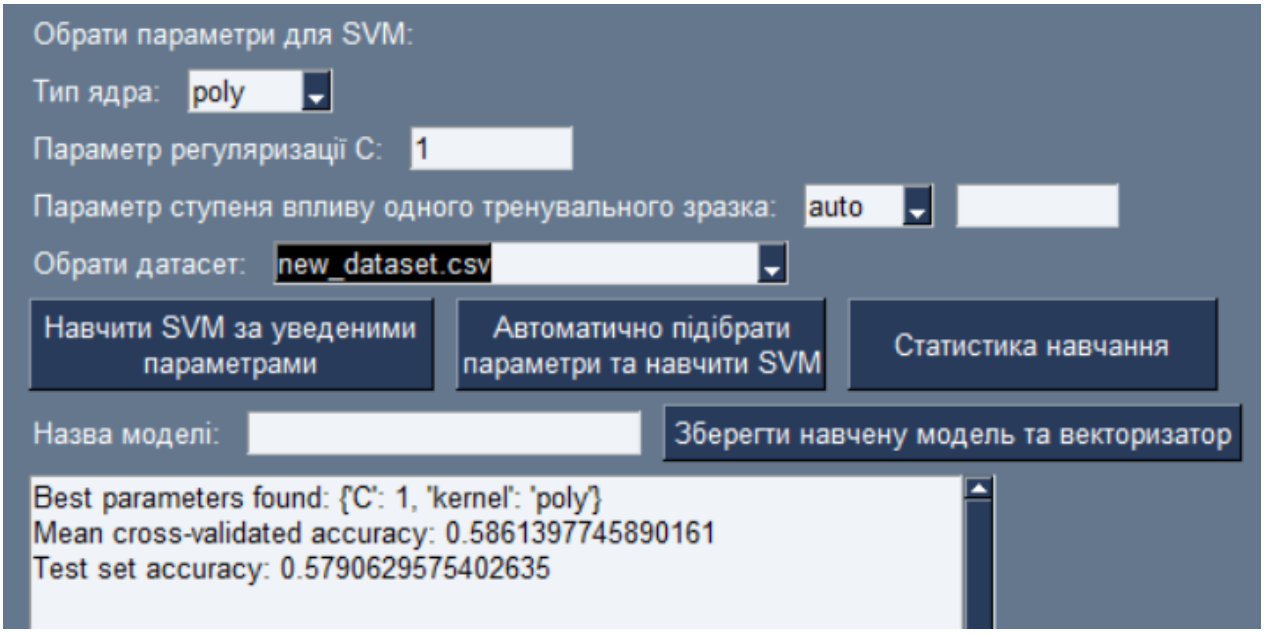


Рисунок 3.4 – Приклад виконання пошуку найкращих параметрів

В результаті, навчені класифікатори можна зберегти для подальшого використання підсистемою виявлення гендеру. Завантаження класифікатора та токенізатора відбувається за принципом найвищої оцінки точності для обраного виду гендера. Розроблена інформаційна система наведена на рисунку 3.5.

Рисунок 3.5 – Підсистема визначення гендеру за дописом

Отже, таким чином описано особливості реалізації програмних складових інформаційної системи виявлення гендеру за дописом, що складається із головного меню та трьох інтерфейсних форм, що призначені для реалізації функцій інформаційної системи.

3.5 Тестування інформаційної системи виявлення гендерної приналежності за дописами та вимоги до розгортання

Наступним кроком є перевірка ефективності використання програмного продукту щодо заданих функцій. Також необхідно визначити на скільки відповідає розроблена програмна реалізація поставленим задачам.

Для перевірки функціоналу будуть використані тест-кейси. Першим тестовим випадком буде перевірка коректності виведення дописів обраного гендера підсистеми «Редактор датасета з дописами за гендерами». Кроки тестового випадку наведені у таблиці 3.1.

Таблиця 3.1 – Тест-кейс 00001

<table>
<tr><td>Тест-кейс ID: 00001</td><td>Приорітет: 1</td><td>Створено: 15.04.2024, Павло СУПРУН</td></tr>
<tr><td colspan="3">Назва: перевірка коректності виведення дописів обраного гендера підсистеми «Редактор датасета з дописами за гендерами»</td></tr>
<tr><td colspan="2">Кроки</td><td>Очікуваний результат</td></tr>
<tr><td colspan="2">1. Відкрити застосунок.
2. Перейти на підсистему «Редактор датасета з дописами за гендерами», натиснувши відповідну кнопку в головному меню.
3. Обрати з випадаючого переліку вид гендеру «Цисгендер»
4. Обрати гендер «Чоловічий цисгендер».
5. Натиснути кнопку «Вивести дописи з датасету для обраного гендера»</td><td>Відкрилась форма головного меню
Відкрилась підсистема «Редактор датасета з дописами за гендерами»
Відображення постів датасету, які розмічені як пости чоловічого цисгендеру.</td></tr>
<tr><td colspan="3">Результат виконання тест-кейсу: перевірку пройдено успішно.</td></tr>
</table>

Результат виконання кроків з таблиці 3.1 наведено на рисунку 3.6.

Наступним тестовим випадком буде перевірка коректності визначення мови обраного допису обраного гендера підсистеми «Редактор датасета з дописами за гендерами». Кроки тестового випадку наведені у таблиці 3.2.

Таблиця 3.2 – Тест-кейс 00002

<table>
<tr><td>Тест-кейс ID: 00002</td><td>Приорітет: 1</td><td colspan="2">Створено: 16.04.2024, Павло СУПРУН</td></tr>
<tr><td colspan="4">Назва: перевірка коректності визначення мови обраного допису обраного гендера підсистеми «Редактор датасета з дописами за гендерами»</td></tr>
<tr><td colspan="3">Кроки</td><td>Очікуваний результат</td></tr>
<tr><td colspan="3">1. Відкрити застосунок.
2. Перейти на підсистему «Редактор датасета з дописами за гендерами», натиснувши відповідну кнопку в головному меню.
3. Обрати з випадаючого переліку вид гендеру «Цисгендер»
4. Обрати гендер «Чоловічий цисгендер».
5. Натиснути кнопку «Вивести дописи з датасету для обраного гендера»
6. Обрати пост для деталізації.
7. Натиснути кнопку «Визначити мову допису».</td><td>Відкрилась форма головного меню
Відкрилась підсистема «Редактор датасета з дописами за гендерами»
Відображення постів датасету, які розмічені як пости чоловічого цисгендеру.
Обраний пост відобразився в робочій області.
Визначилась мова допису.</td></tr>
<tr><td colspan="4">Результат виконання тест-кейсу: перевірку пройдено успішно.</td></tr>
</table>

Після відкриття застосунку, як і у попередньому тестовому випадку, необхідно перейти на підсистему «Редактор датасета з дописами за гендерами», після чого обрати з випадаючого переліку вид гендеру «Цисгендер». Тест-кейс перевіряє визначення мови на прикладі допису чоловічого цисгендеру. Результат виконання кроків з таблиці 3.2 наведено на рисунку 3.7.

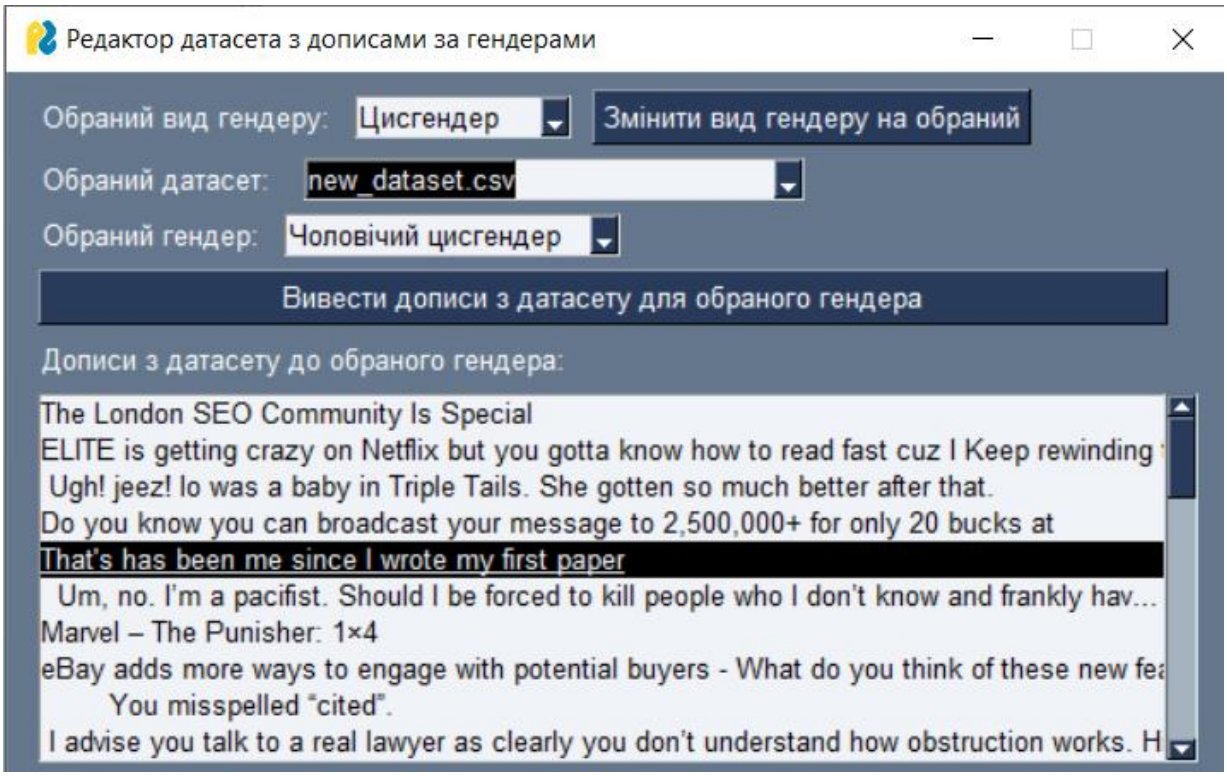


Рисунок 3.6 – Виконання тест-кейсу 00001

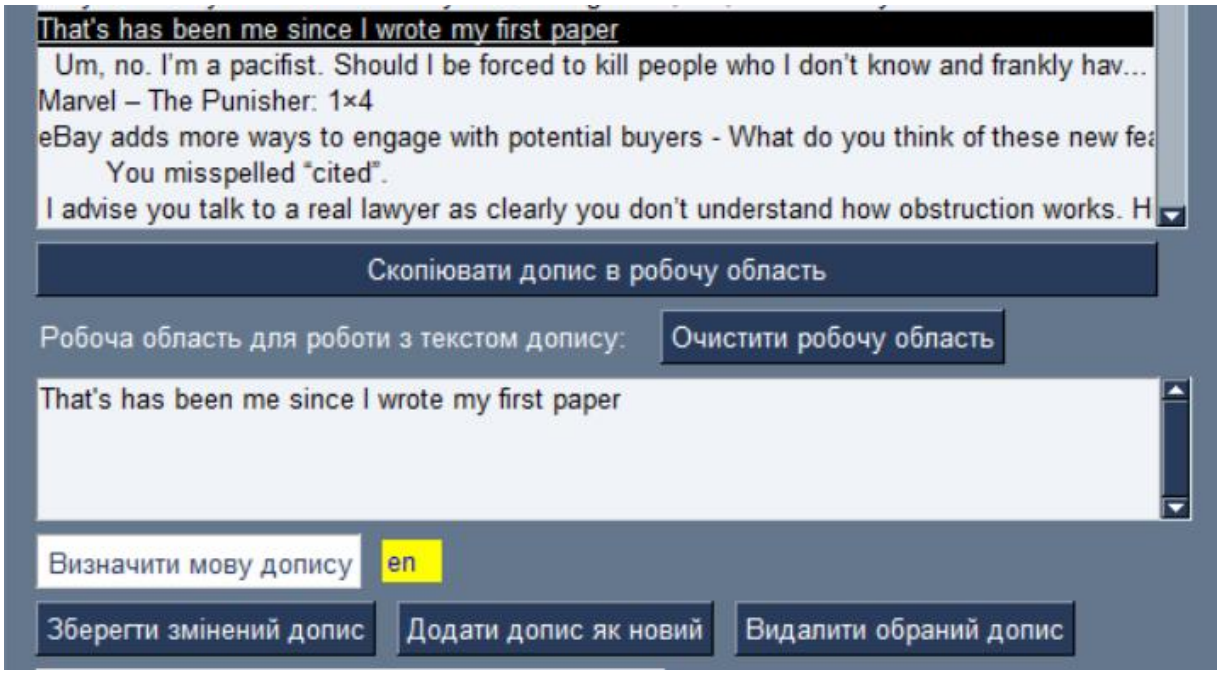


Рисунок 3.7 – Виконання тест-кейсу 00002

Як видно з рисунка 3.7 – тестування пройдено успішно. Наступним тестовим випадком буде перевірка коректності виявлення гендеру за дописом. Кроки наведені в таблиці 3.3. Для даного тестового випадку буде використано функцію уведення допису вручну, замість обирання з переліку наявних. Даний

тест-кейс перевіряє підсистему «Виявлення гендерної приналежності за дописом», яка є головною підсистемою застосунку.

Результат виконання тест-кейсу 00001 наведено на рисунку 3.8, тест-кейс пройдено успішно.

Таблиця 3.3 – Тест-кейс 00003

Тест-кейс ID: 00003	**Приорітет:** 1	**Створено:** 16.04.2024, Павло СУПРУН
Назва: перевірка коректності коректності виявлення гендеру за дописом		
Кроки		**Очікуваний результат**
1. Відкрити застосунок. 2. Перейти на підсистему «Виявлення гендерної приналежності за дописом», натиснувши відповідну кнопку в головному меню. 3. Обрати з випадаючого переліку вид гендеру «Цисгендер» 4. Написати в робочій області текст англійською мовою. 5. Натиснути кнопку «Визначити гендерну приналежність автора за дописом»		Відкрилась форма головного меню Відкрилась підсистема «Виявлення гендерної приналежності за дописом». У полі «Результат» наведено отриманий цисгендер.
Результат виконання тест-кейсу: перевірку пройдено успішно.		

Нижче наведено вимоги до розгортання інформаційної системи виявлення гендерної приналежності автора за дописом.

Апаратне забезпечення:

– мінімальні вимоги до процесора: Intel Core i5 або еквівалентний;

– рекомендована кількість оперативної пам'яті: не менше 8 ГБ;

– мінімальний обсяг вільного місця на диску: 20 ГБ для збереження моделей, даних та системних файлів.

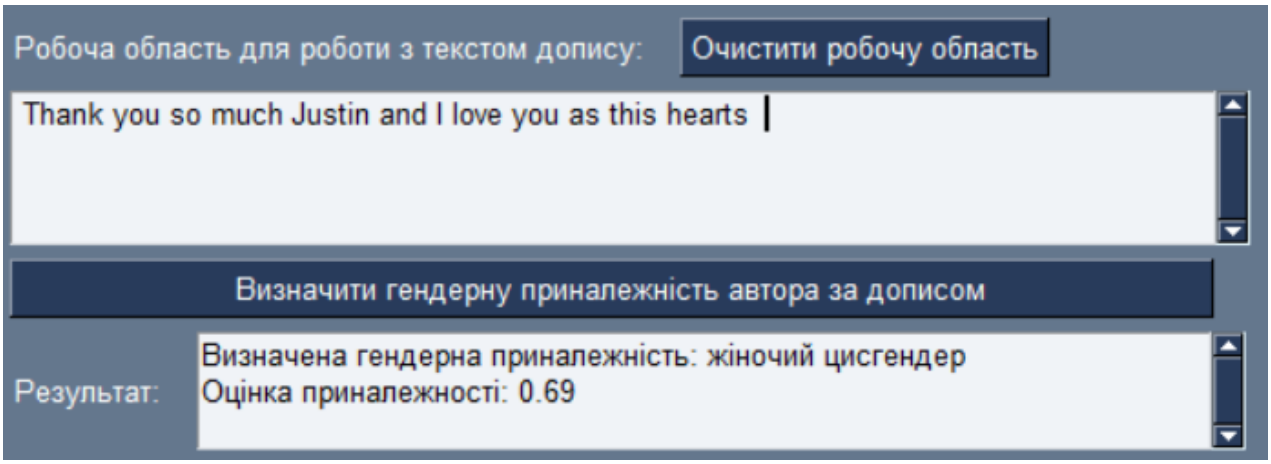


Рисунок 3.8 – Результат виконання тест-кейсу 00003

Операційна система:

– підтримувані операційні системи: Windows 10, macOS 10.13+, Linux;

– наявність встановленого Python 3.7+ та відповідних пакетів керування пакетами.

Отже, проведено тестування інформаційної системи виявлення гендерної приналежності за дописами. В ході виконаного тестування всі функції працюють коректно, відповідно до заявлених. Програмний продукт повністю відповідає поставленим завданням. Також наведено основні вимоги щодо розгортання створеної інформаційної системи виявлення гендерної приналежності за дописами.

3.6 Аналіз функціональності інформаційної системи виявлення гендерної приналежності за дописами

Використання інформаційної системи виявлення гендерної приналежності за дописами соціальних інтернет-мереж розпочинається із головного меню (рисунок 3.9).

З головного меню у користувача є можливість перейти на такі підсистеми: «Редактор датасета з дописами за гендерами», «Виявлення гендерної приналежності за дописом» та «Навчання класифікаторів».

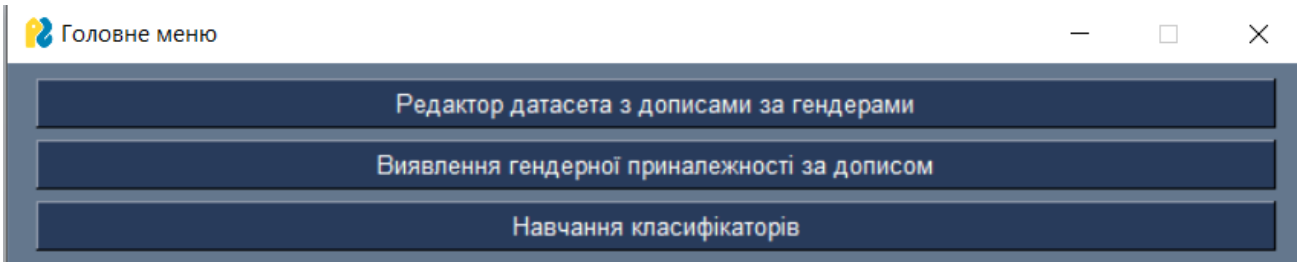


Рисунок 3.9 – Головне меню інформаційної системи виявлення гендерної приналежності за дописами соціальних інтернет-мереж

Для використання підсистеми «Редактор датасета з дописами за гендерами» необхідно натиснути однойменну кнопку головного меню. Відкриється форма редактора датасета з дописами за гендерами (рисунок 3.10).

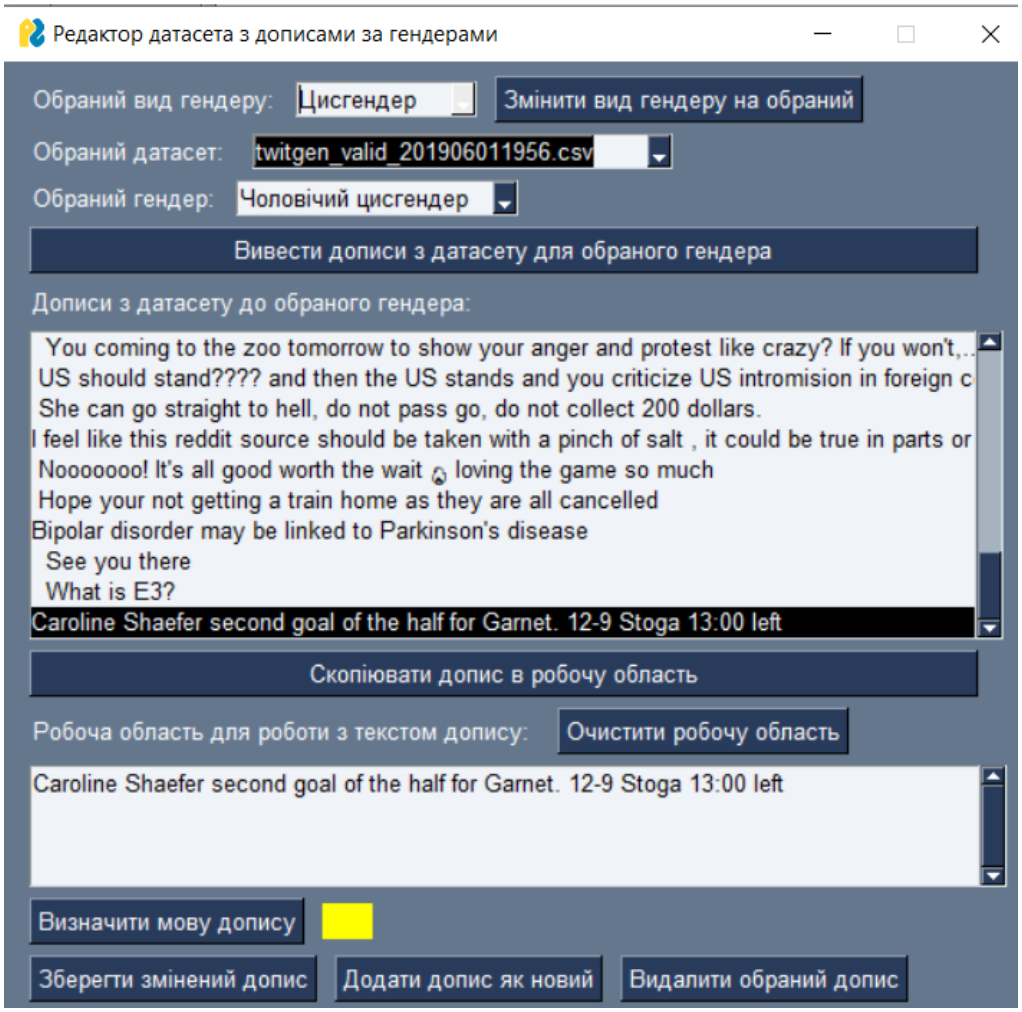


Рисунок 3.10 – Форма редактора датасета з дописами за гендерами

Перш за все необхідно обрати вид гендеру з випадаючого переліку, з яким буде працювати користувач (рисунок 3.11)

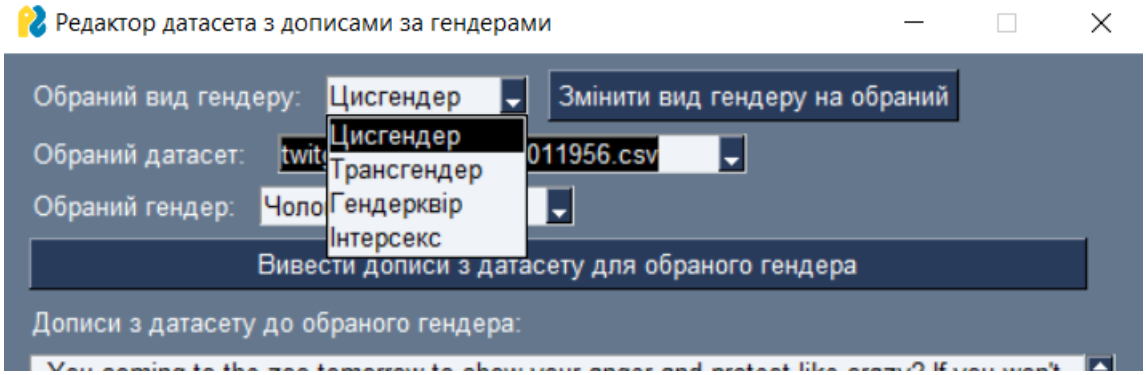


Рисунок 3.11 – Вибір виду гендеру для редагування

Далі за обраним видом гендеру обрати доступний датасет та гендер з відповідних випадаючих переліків (рисунок 3.12).

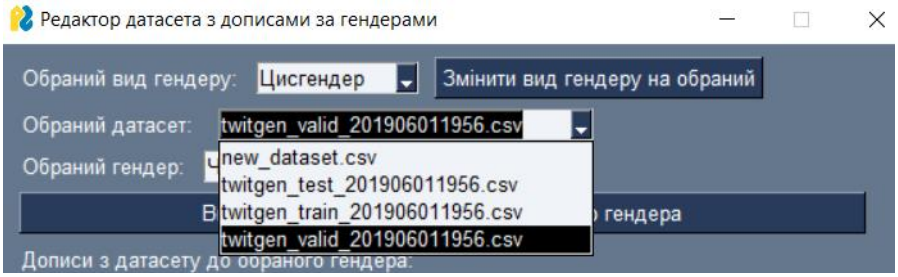


Рисунок 3.12 – Вибір датасету за обраним видом гендеру

Наступним кроком є вибір гендеру, що представлений в датасеті (рисунок 3.13).

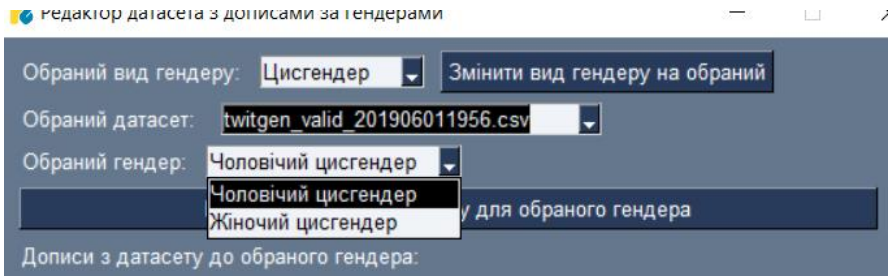


Рисунок 3.13 – Вибір гендеру за обраним видом гендеру

Для перегляду дописів обраного гендеру необхідно натиснути кнопку «Вивести дописи з датасету для обраного гендера». Дописи будуть відображені у полі «Дописи з датасету до обраного гендера» (рисунок 3.14).

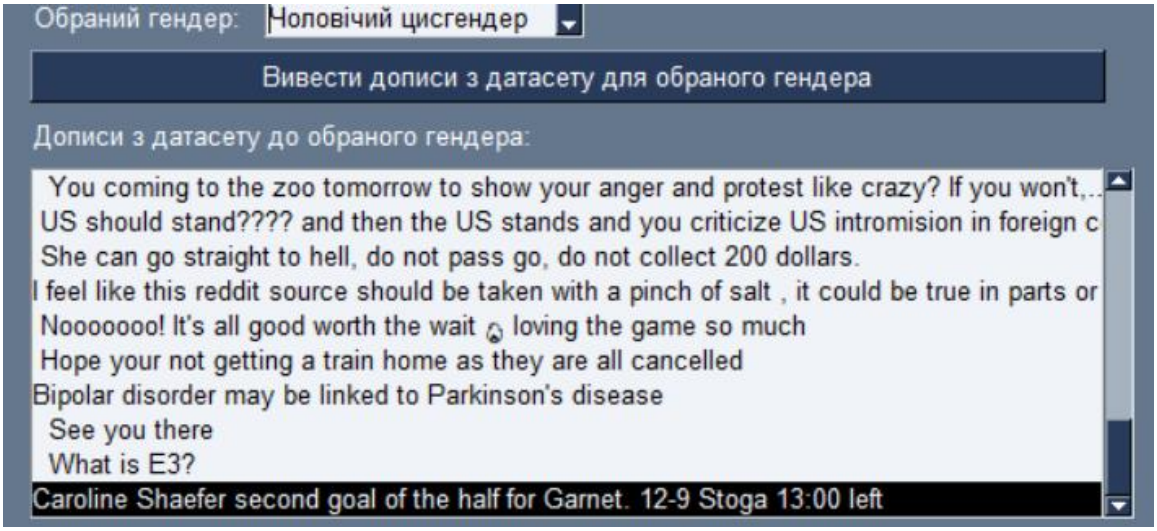


Рисунок 3.14 – Виведення дописів до обраного гендера

Для деталізації і внесення змін в існуючі дописи, потрібно виділити допис для аналізу та натиснути кнопку «Скопіювати допис в робочу область», після чого допис буде відображено у робочій області (рисунок 3.15).

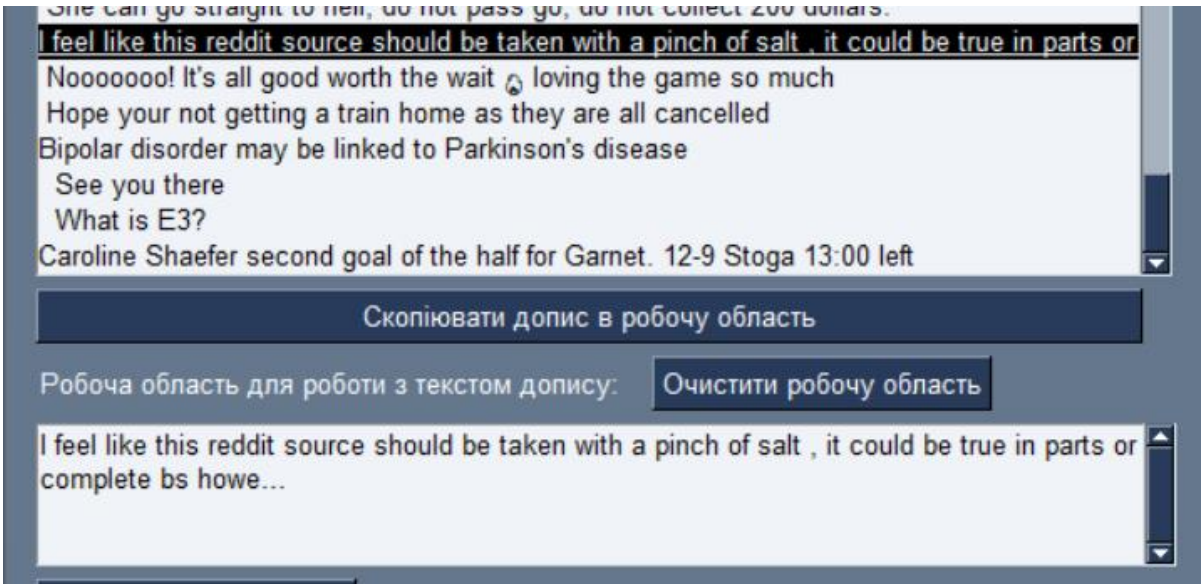


Рисунок 3.15 – Відображення допису у робочій області

Для очищення робочої області необхідно натиснути кнопку «Очистити робочу область». Для визначення мови обраного допису необхідно натиснути кнопку «Визначити мову допису» (рисунок 3.16)

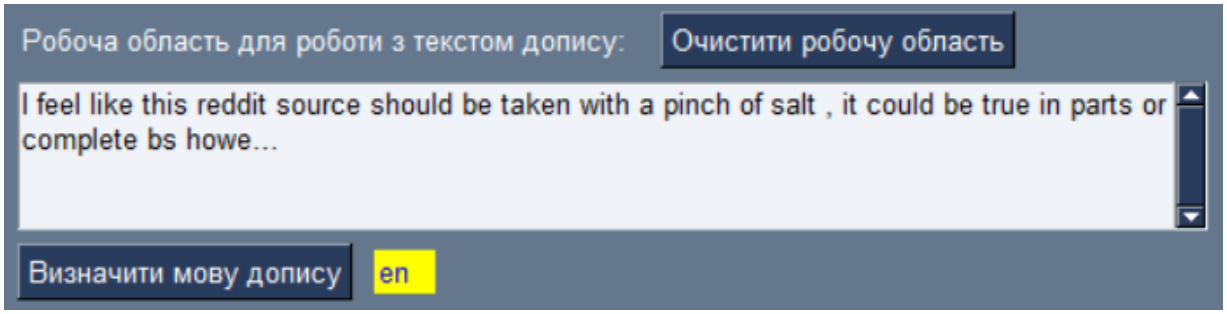


Рисунок 3.16 – Визначення мови допису

Рисунок 3.17 – Підсистема виявлення гендерної приналежності за дописом

Для збереження змін у коригованому дописі необхідно натиснути кнопку «Зберегти змінений допис», а для того щоб додати допис як новий, необхідно

натиснути кнопку «Додати допис як новий». Для видалення обраного допису необхідно натиснути кнопку «Видалити обраний допис».

Для роботи з підсистемою «Виявлення гендерної приналежності за дописом» необхідно натиснути на однойменну кнопку в головному меню. Вигляд форми «Виявлення гендерної приналежності за дописом» наведено на рисунку 3.17.

Аналогічно як і у першій підсистемі, робота розпочинається із вибору виду гендера для аналізу. За обраним видом гендеру для аналізу буде здійснюватись процес класифікації. Наприклад, для виду гендеру «Цисгендер», буде здійснюватись бінарна класифікація тексту з поділом на чоловічий цисгендер або жіночий цисгендер.

Після вибору виду гендеру можна побачити статистику датасету за кількістю записів у ньому, а також можна як обрати запис з існуючих, скориставшись вибором допису для аналізу з поля «Дописи обраного гендера», та натиснувши кнопку «Скопіювати допис в робочу область» (рисунок 3.18)

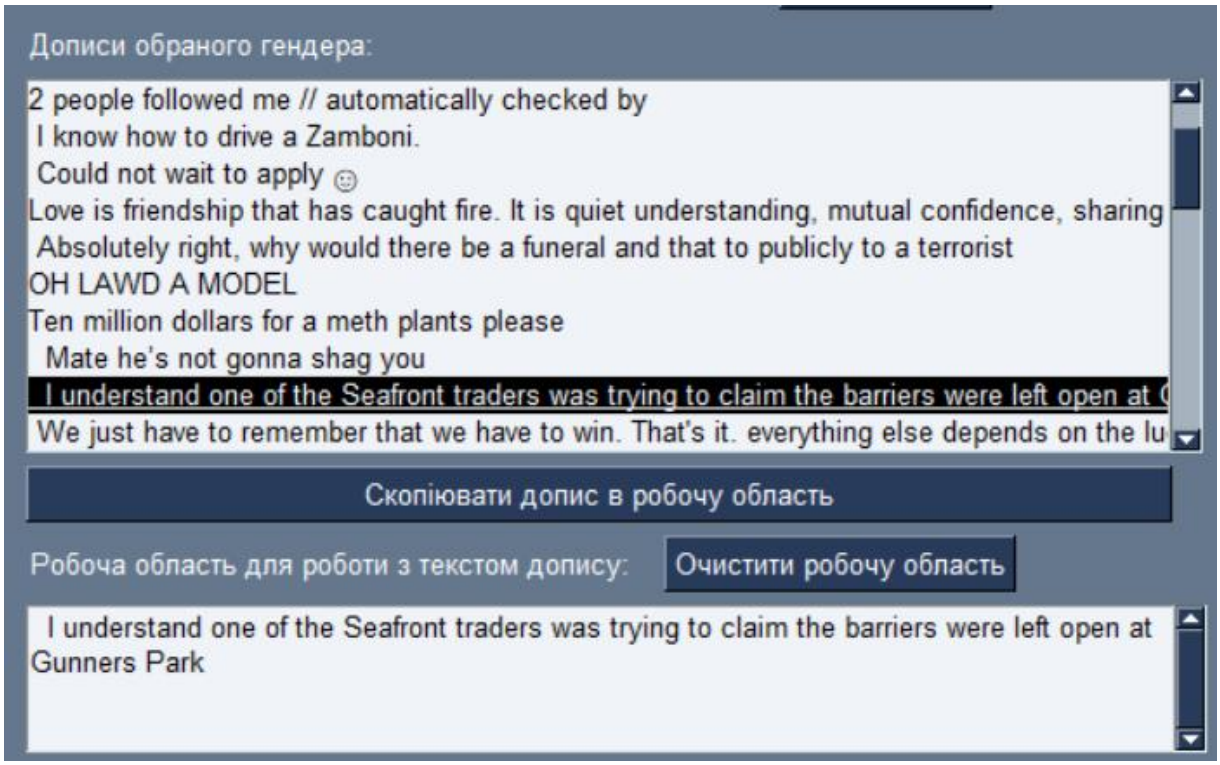


Рисунок 3.18 – Копіювання допису в робочу область

Також можна увести допис в робочу область вручну. Для очищення робочої області від наявного тексту необхідно натиснути кнопку «Очистити робочу область».

Для визначення гендерної приналежності за дописом в робочій області необхідно натиснути кнопку «Визначити гендерну приналежність допису автора за дописом» (рисунок 3.19). Результат визначення гендерної ідентичності наведено в полі «Результат».

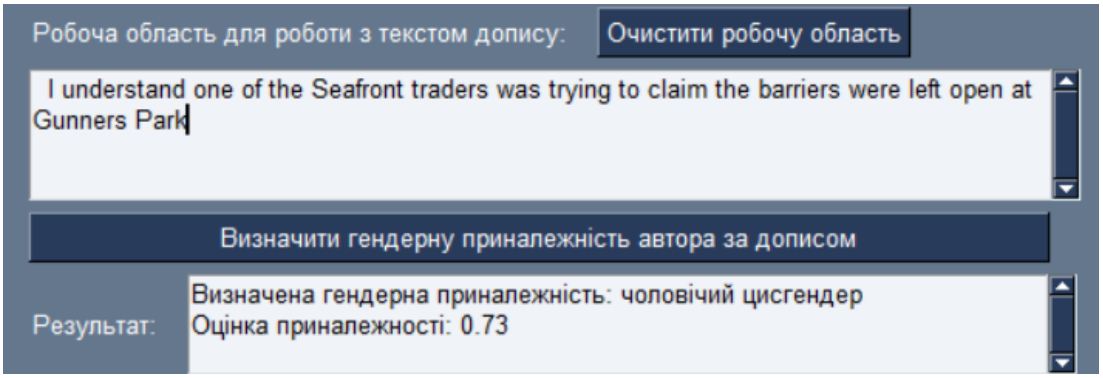


Рисунок 3.19 – Визначення гендерної приналежності за дописом

Для використання функціоналу модуля «Навчання класифікаторів», необхідно в головному меню натиснути однойменну кнопку. Вигляд підсистеми що дозволяє здійснювати навчання класифікаторів SVM наведено на рисунку 3.20.

Дана підсистема дозволяє як навчати класифікатор за обраними користувацькими параметрами типу ядра, параметра регуляризації C, параметру ступеня впливу одного тренувального зразка, так і здійснювати автоматизований підбір параметрів. Для вибору користувацьких параметрів необхідно скористатись випадаючими переліками (рисунок 3.21).

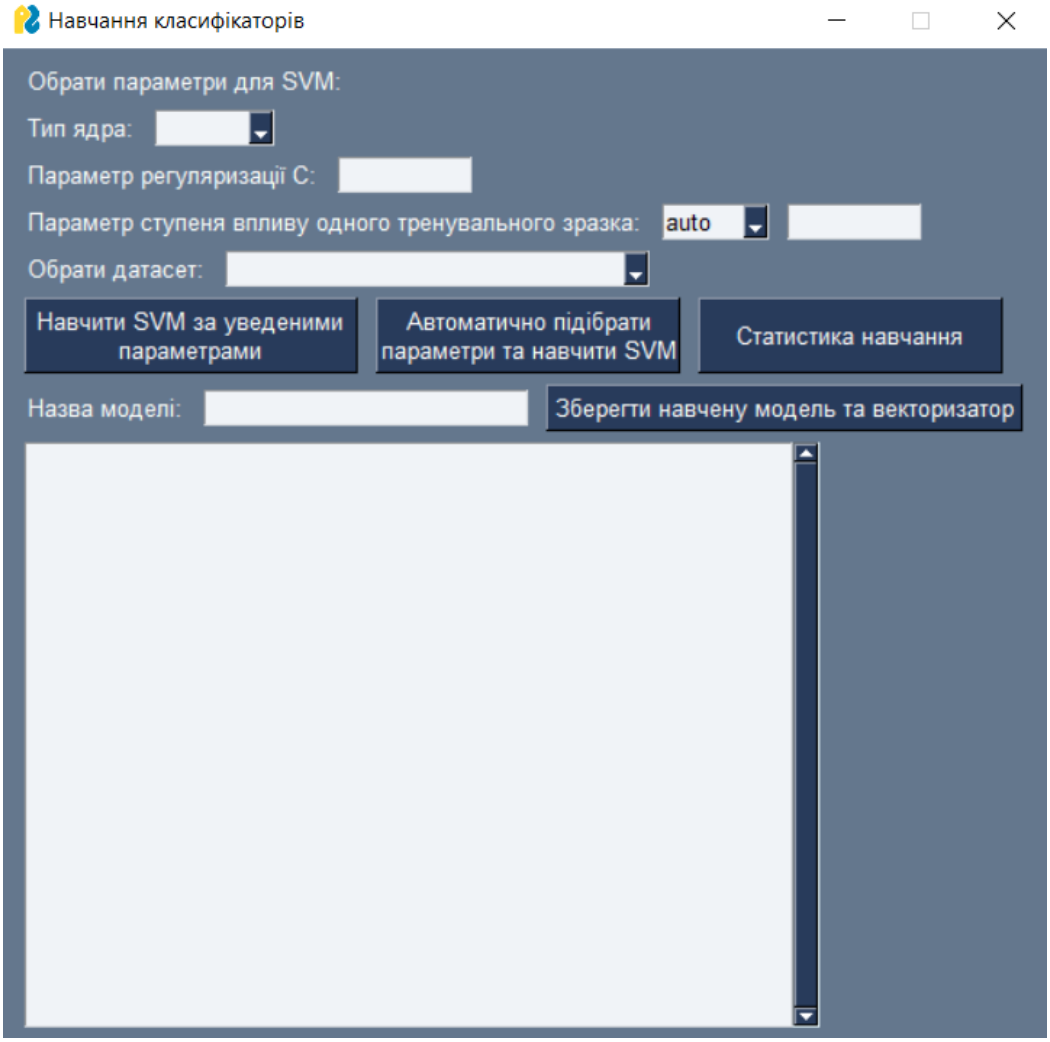


Рисунок 3.20 – Підсистема навчання класифікаторів

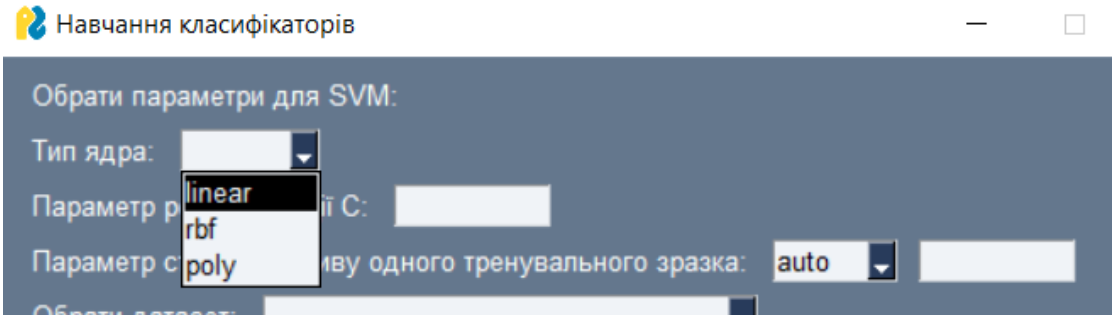


Рисунок 3.21 – Вибір параметрів «вручну»

Також необхідно обрати датасет, на основі якого буде здійснюватись тренування класифікатора (рисунок 3.22).

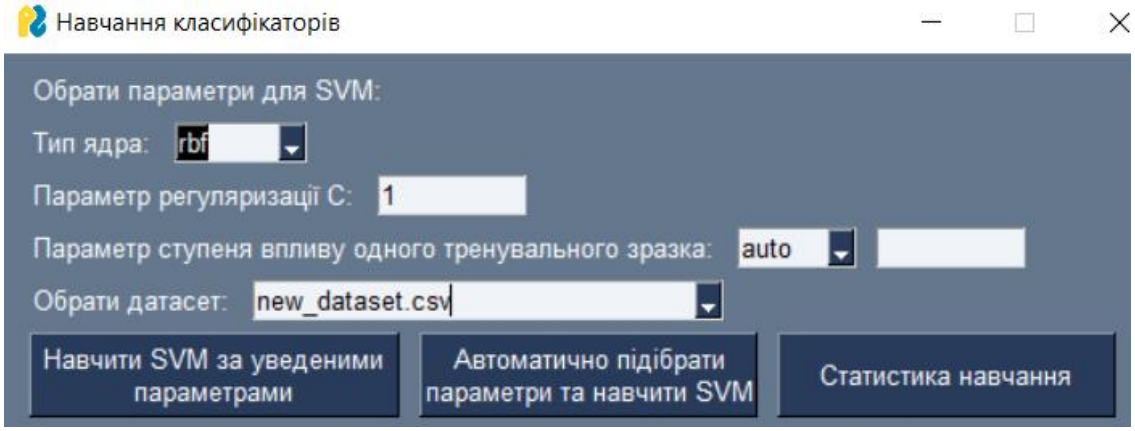


Рисунок 3.22 – Вибір датасету для навчання типового класифікатора SVM

Для початку навчання за користувацькими параметрами необхідно натиснути на кнопку «Навчити SVM за уведеними параметрами», а для автоматизованого підбору необхідно натиснути на кнопку «Автоматично підібрати параметри та навчити SVM».

Після завершення навчання для виведення статистики необхідно натиснути кнопку «Статистика навчання» (рисунок 3.23).

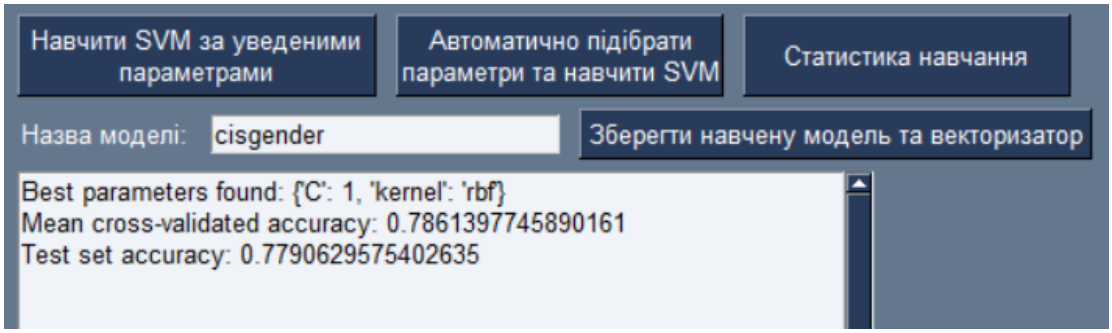


Рисунок 3.23 – Статистика навчання SVM автоматизованим шляхом

Для збереження моделі і векторизатора необхідно натиснути кнопку «Зберегти навчену модель та векторизатор».

Отже, таким чином було здійснено аналіз функціональності системи виявлення гендерної приналежності за дописами соціальних інтернет-мереж та наведено особливості її використання.

3.7 Результати досліджень

За створеною інформаційною системою виявлення гендерної приналежності за дописами буде досліджено ефективність запропонованого методу, яке буде відбуватись у два етапи. Першим етапом буде порівняння різних моделей навчених класифікаторів. Шляхом використання параметрів, заданих «вручну», та автоматичним пошуком. Досліджувані комбінації наведено в таблиці 3.4.

Таблиця 3.4 – Дослідження навчальних параметрів моделей SVM

Параметри:	***C***	***Ядро***	***GAMMA***	***Accuracy***
Набір 1	1	rbf	auto	*0.71*
Набір 2	10	rbf	auto	*0.68*
Набір 3	0.1	rbf	scale = 0.5	*0.66*
Набір 4	1	poly	auto	*0.67*
Набір 5	10	poly	auto	*0.65*
Набір 6	0.1	poly	scale = 0.5	*0.62*
Набір 7	1	linear	auto	*0.66*
Набір 8	10	linear	auto	*0.63*
Набір 9	0.1	linear	scale = 0.5	*0.62*

Дослідження наборів навчальних параметрів моделей SVM за метрикою точності наведено на рисунку 3.24.

Як видно з таблиці 3.4 та рисунку 3.24 найкращим набором параметрів виявилась комбінація C = 1, ядро rbf та GAMMA у значенні auto. Наступним експериментом буде порівняння роботи створеного програмного застосунку та існуючого рішення «Gender Guesser». «Gender Guesser» погано працює з текстами довжиною менше 300 слів і дозволяє на думку авторів виявити цисгендер з точністю близько 60-70%. Однак зважаючи на специфіку набору навчальних даних, тестування буде проведено на твітах. Для експерименту взято

50 твітів з валідаційної вибірки випадковим чином. Після вибору даних, 27 твітів належать категорії «чоловічий цисгендер» та 23 твіта категорії «жіночий цисгендер».

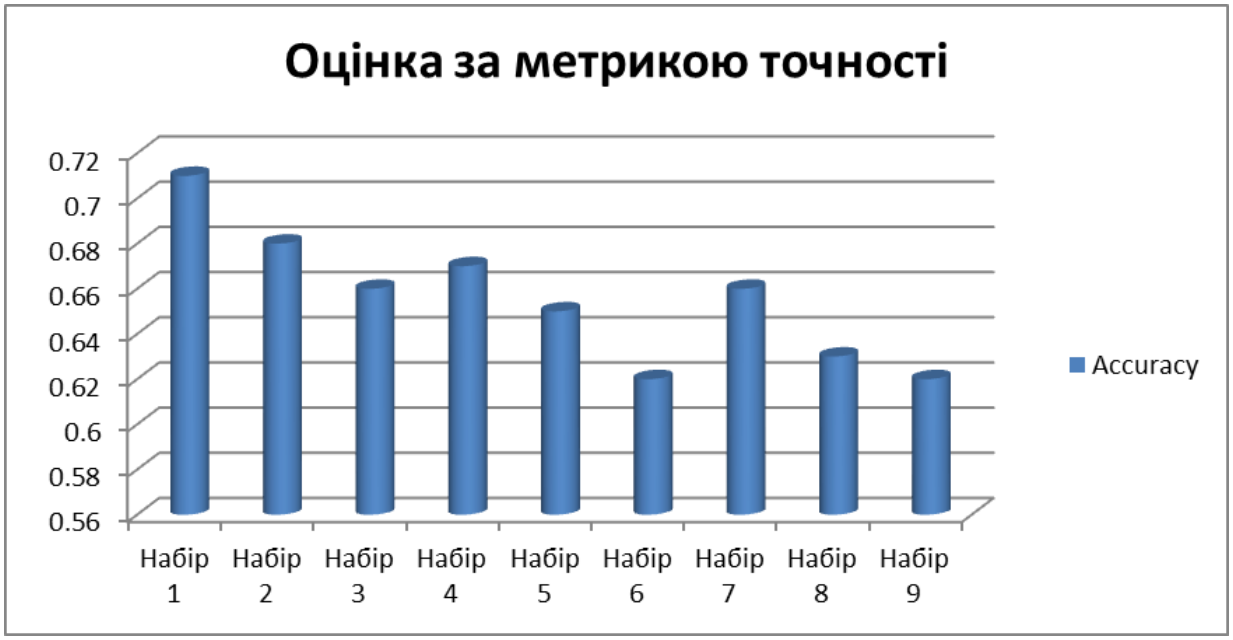


Рисунок 3.24 – Дослідження наборів параметрів

«Gender Guesser» надісланий текст оцінює за двома типами написання: формальним і неформальним. Офіційне написання включає в себе художні та науково-популярні історії, статті та новини. Неформальне письмо включає тексти в блогах і чатах. Отже, для експерименту буде використано колонку неформального результату. Приклад ідентифікації твіта чоловічого цисгендеру наведено на рисунку 3.25. Результат помилковий, хоча і в оцінюванні офіційного письма відрізняється.

Результати виявлення цисгендеру за допомогою програми «Gender Guesser» та створеного методу наведено в таблицях 3.5 та 3.6 відповідно.

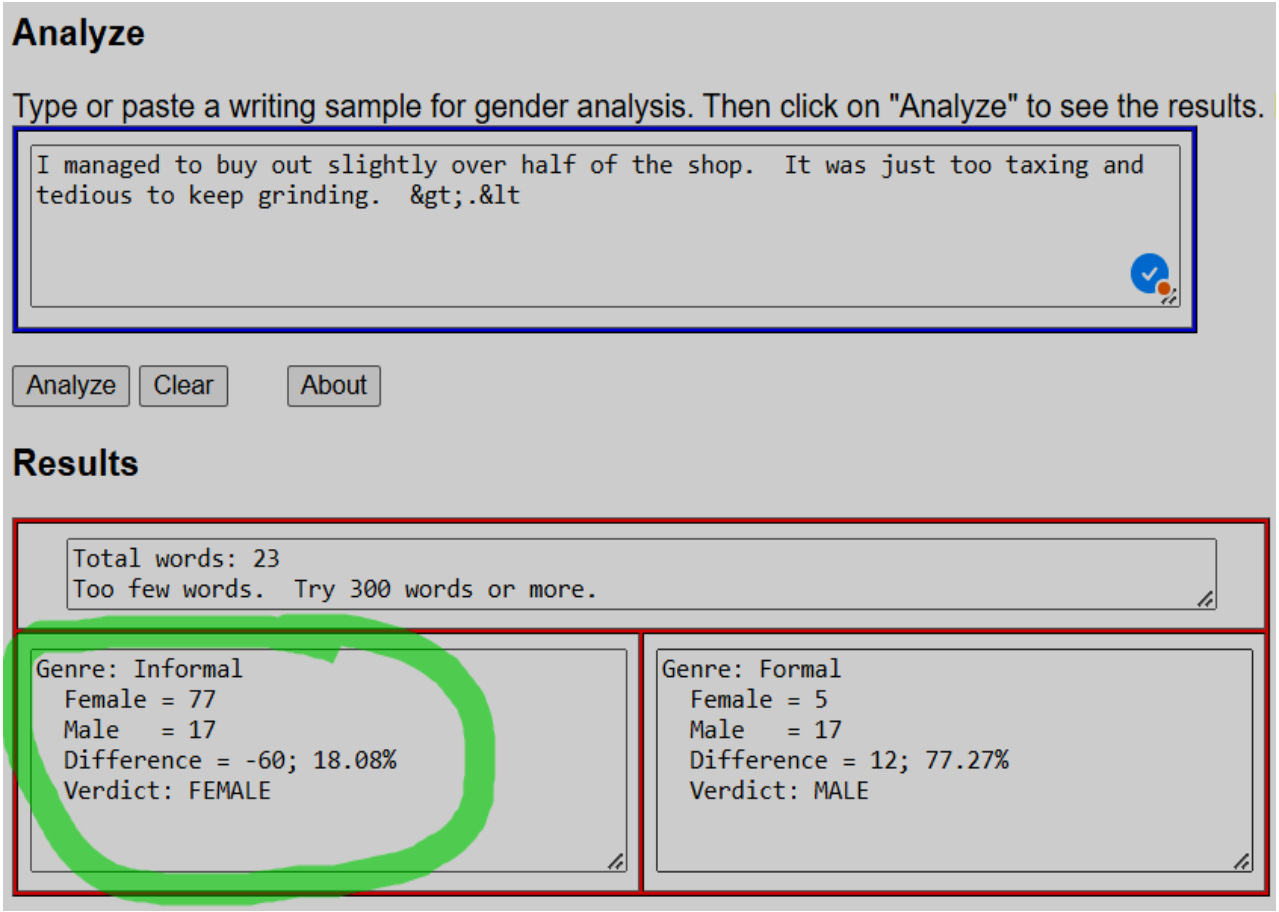


Рисунок 3.25 – Визначення гендеру за допомогою «Gender Guesser»

Таблиця 3.5 – Результати виявлення цисгендеру за допомогою «Gender Guesser»

	Чоловічий цисгендер	Жіночий цисгендер
Кількість коректно ідентифікованих дописів	15	12
Кількість некоректно ідентифікованих дописів	12	11
Разом дописів:	27	23

У таблиці 3.7 наведено спільні результати у відсотковому значенні, що також проілюстровані на діаграмі 3.26.

Таблиця 3.6 – Результати виявлення цисгендеру за допомогою запропонованого методу

	Чоловічий цисгендер	Жіночий цисгендер
Кількість коректно ідентифікованих дописів	18	16
Кількість некоректно ідентифікованих дописів	9	7
Разом дописів:	27	23

Таблиця 3.7 – Порівняння точності розробленим методом та «Gender Guesser»

Коректно ідентифіковані дописи	Чоловічий цисгендер	Жіночий цисгендер
«Gender Guesser»	0.555555556	0.52173913
Розроблений метод	0.666666667	0.695652174

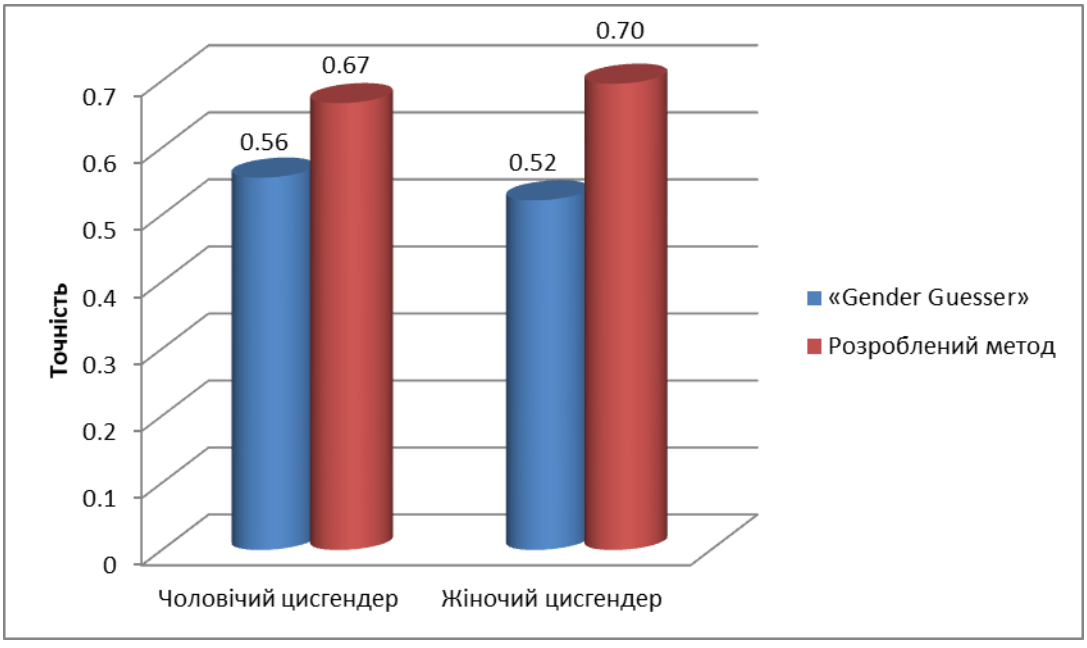


Рисунок 3.26 – Діаграма виявлення цисгендеру альтернативними підходами

Як видно з таблиці 3.7 та діаграми 3.26, розроблений метод показав вищу ефективність. Це може бути пов’язано з прикладами, на яких навчались

класифікатори, адже навчання SVM було здійснено а базі твітів, а в застосунку з яким відбувалось порівняння вказано, що для збільшення точності необхідно використовувати тексти понад 300 слів, а це є доволі великий обсяг.

Однак, точність розробленого методу також можна покращити. Вибір стилю письма часто залежить від різних факторів, таких як освіта, досвід, професія, вік та національність автора. Наприклад, представниця жіночого цисгендеру, яка працює у галузі, де переважають особи чоловічого цисгендеру може прийняти їх стиль письма. Автори з великим досвідом часто використовують професійні стилі письма, незалежно від гендерних особливостей. Для більшої точності класифікації необхідно більше даних для навчання і вони повинні бути більше збалансовані.

Отже, було виконано дослідження ефективності створеного методу виявлення гендеру за дописом. Розроблений метод показав високу ефективність, в порівнянні із аналогом його точність вища на 0.11. Також він може працювати з короткими текстами без втрати точності, такими як твіти. Подальші дослідження будуть спрямовані на виявлення додаткових ознак, а також на роботу з іншими видами гендерів.

3.8 Висновки до розділу 3

Визначено шляхи щодо дослідження ефективності, які полягають в створенні інформаційної системи виявлення гендерної приналежності за дописами на основі запропонованого методу, а коректність виконання заявлених функцій буде перевірено з використанням тест-кейсів. Коректність навчання SVM-моделей буде досліджено із застосуванням метрики accuracy.

Обрано засоби для розробки інформаційної системи виявлення гендерної приналежності за дописами, буде використано інтегроване середовище програмування PyCharm, мову програмування Python та мову запитів SQL. Даний набір має якісну взаємодію між собою, та задовольняє використання спеціалізованих програмних розширень.

Описано структуру та функціональне призначення програмних складових інформаційної системи виявлення гендерної приналежності за дописами соціальних інтернет-мереж. Наведено особливості реалізації програмних складових інформаційної системи виявлення гендеру за дописом, що складається із головного меню та трьох інтерфейсних форм, що призначені для реалізації функцій інформаційної системи.

Проведено тестування інформаційної системи виявлення гендерної приналежності за дописами, в ході виконаного тестування всі функції працюють коректно, відповідно до заявлених. Програмний продукт повністю відповідає поставленим завданням. Також наведено основні вимоги щодо розгортання створеного програмного забезпечення.

Здійснено аналіз функціональності інформаційної системи виявлення гендерної приналежності за дописами соціальних інтернет-мереж та наведено особливості її використання.

Було виконано дослідження ефективності створеного методу виявлення гендеру за дописом. Розроблений метод показав високу ефективність, в порівнянні із аналогом його точність у рамках проведеного експерименту вища на 0.11. Особливістю розробленого методу є можливість роботи з короткими текстами без втрати точності, такими як твіти. Подальші дослідження будуть спрямовані на виявлення додаткових ознак, а також на роботу з іншими видами гендерів.

Загальні висновки

Мета кваліфікаційної роботи бакалавра полягала в покращенні виявлення гендерної приналежності у дописах соціальних інтернет-мереж шляхом розробки методу виявлення гендерної приналежності за дописами соціальних інтернет-мереж засобами NLP, та відповідного програмного забезпечення.

Проведений аналіз наукових праць та наявних програмних реалізацій, свідчить про те, що напрям виявлення гендерної приналежності за дописами соціальних інтернет-мереж є актуальним. Здебільшого науковцями використовується обраний підхід на основі SVM. Проте, кількість прикладних програмних рішень та їх якість все ще недостатня, отже розробка програмного забезпечення у даному напрямі є актуальною та затребуваною.

Для досягнення поставленої мети були поставлені та виконані такі задачі:

– виконано аналіз предметної області виявлення гендерної приналежності за текстовими даними, де було помічено, що гендери мають певні риси на письмі, які сприйматися як характеристики, що сприяють відокремленню гендерів і враховуються при автоматизованому аналізі текстів з метою виявлення гендерних відмінностей;

– виконано огляд теоретичних підходів щодо можливості виявлення гендерної приналежності за дописами соціальних інтернет-мереж, обрано підхід для подальшої реалізації, в основі якого буде SVM-класифікатор;

– виконано огляд існуючих наукових надбань та програмних рішень;

– створено метод виявлення гендерної приналежності за дописами соціальних інтернет-мереж засобами NLP, що призначений для аналізу та класифікації текстових даних з метою визначення гендерної приналежності, яка здійснюється на основі особливостей мови та стилю письма, що є характерними для певних груп користувачів та описано його кроки;

– спроектовано інформаційну систему для реалізації методу виявлення гендерної приналежності за дописами соціальних інтернет-мереж засобами NLP;

– створено програмну реалізацію за спроєктованою структурою інформаційної системи;

– виконано тестування створеної інформаційної системи виявлення гендерної приналежності за дописами, яке некоректно працюючих функцій не виявило;

– виконано дослідження ефективності методу виявлення гендерної приналежності за дописами соціальних інтернет-мереж засобами NLP на основі створеного ПЗ, розроблений метод показав високу ефективність, в порівнянні із аналогом його точність у рамках проведеного експерименту вища на 0.11. Також він може працювати з короткими текстами без втрати точності, такими як твіти.

Результати виконаної кваліфікаційної роботи бакалавра відповідають очікуваним, всі поставлені завдання виконано. Подальші дослідження будуть спрямовані на виявлення додаткових ознак, а також на роботу з іншими видами гендерів.

Перелік посилань

1. Гендерна лінгвістика. URL: https://uk.wikipedia.org/wiki/Гендерна_лінгвістика

2. Гендерна лінгвістика як наука. URL: https://studfile.net/preview/7157868/page:5/

3. Прислів’я та приказки з ґендерним компонентом у сучасній німецькій мові. URL: https://ela.kpi.ua/server/api/core/bitstreams/6a0eb559-b7f3-4506-95d3-618c95329e8c/content

4. Понятійний апарат гендерних досліджень URL: https://learn.ztu.edu.ua/mod/resource/view.php?id=187636

5. Гендер. Поняття, суть, види та ознаки. URL: https://termin.in.ua/hender/#Vidi_ci_tipi_genderiv

6. Як працює класифікація тексту. URL: https://www.unite.ai/uk/як-працює-класифікація-тексту/.

7. Обробка природної мови (NLP) у Python з кодом. URL: https://oleg-dubetcky.medium.com/обробка-природної-мови-nlp-у-python-з-кодом-частина-2-класифікація-текстів-b168878ba32d

8. Text Classification: What it is And Why it Matters. URL: https://monkeylearn.com/text-classification/

9. Sentiment Analysis using SVM URL: https://medium.com/scrapehero/sentiment-analysis-using-svm-338d418e3ff1.

10. A White-Box Sociolinguistic Model for Gender Detection. URL: https://www.mdpi.com/2076-3417/12/5/2676

11. Aljohani, T.; Cristea, A.I. Learners Demographics Classification on MOOCs During the COVID-19: Author Profiling via Deep Learning Based on Semantic and Syntactic Representations. Front. Res. Metrics Anal. 2021, 6, 1–17

12. Age and Gender Identification in Unbalanced Social Media URL: https://ieeexplore.ieee.org/abstract/document/8673125

13. Bot and Gender Detection of Twitter Accounts Using Distortion and LSA. URL: https://iris.uniroma1.it/retrieve/handle/11573/1446354/1573368/Bacciu_Bot-and-gender_2019.pdf

14. Bot and Gender Detection using Textual and Stylistic Information. URL: https://downloads.webis.de/pan/publications/papers/giachanou_2019.pdf

15. Bot and gender detection of twitter accounts using distortion and LSA notebook for PAN at CLEF 2019. URL: https://iris.uniroma1.it/handle/11573/1446354

16. Determine the gender of a name. URL: https://genderize.io/

17. Gender Guesser. URL: https://www.hackerfactor.com/GenderGuesser.php

18. Tweet Files for Gender Guessing. URL: https://www.kaggle.com/datasets/aharless/tweet-files-for-gender-guessing

19. ChatGPT. URL: https://chat.openai.com/

20. Scikit-learn. URL: https://uk.wikipedia.org/wiki/Scikit-learn

21. Pandas. URL: https://uk.wikipedia.org/wiki/Pandas

22. langdetect. URL: https://pypi.org/project/langdetect/

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

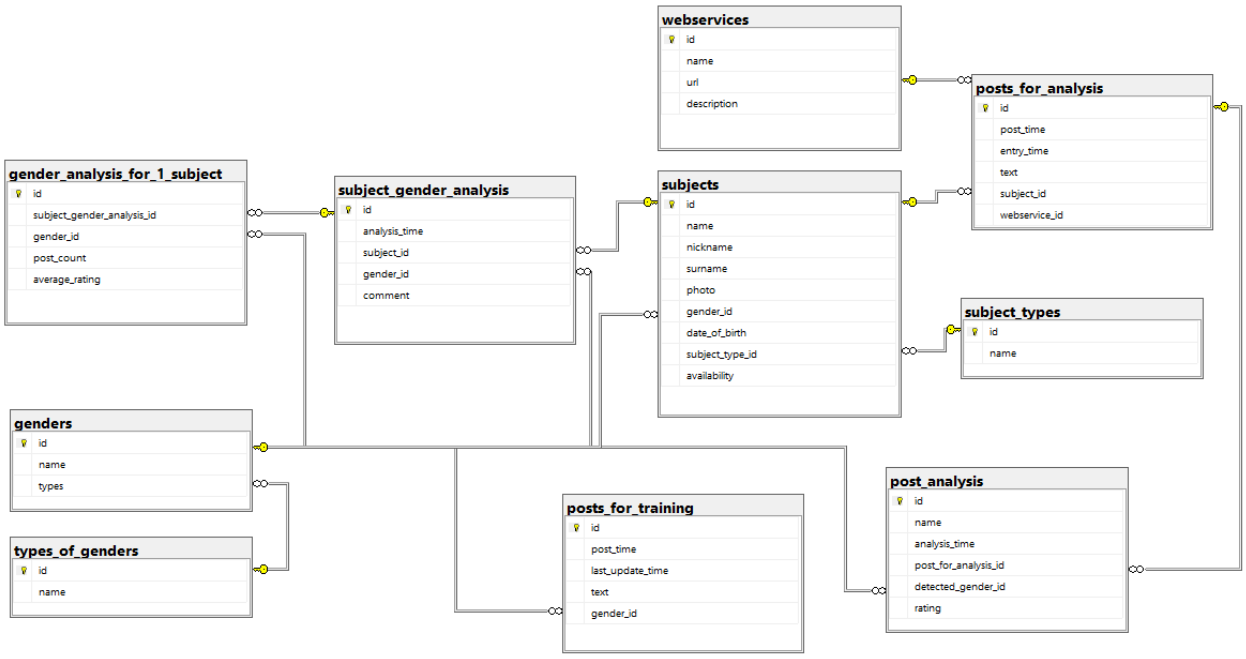
24. Python. URL: https://uk.wikipedia.org/wiki/Python

25. SQL. URL: https://uk.wikipedia.org/wiki/SQL

ДОДАТКИ

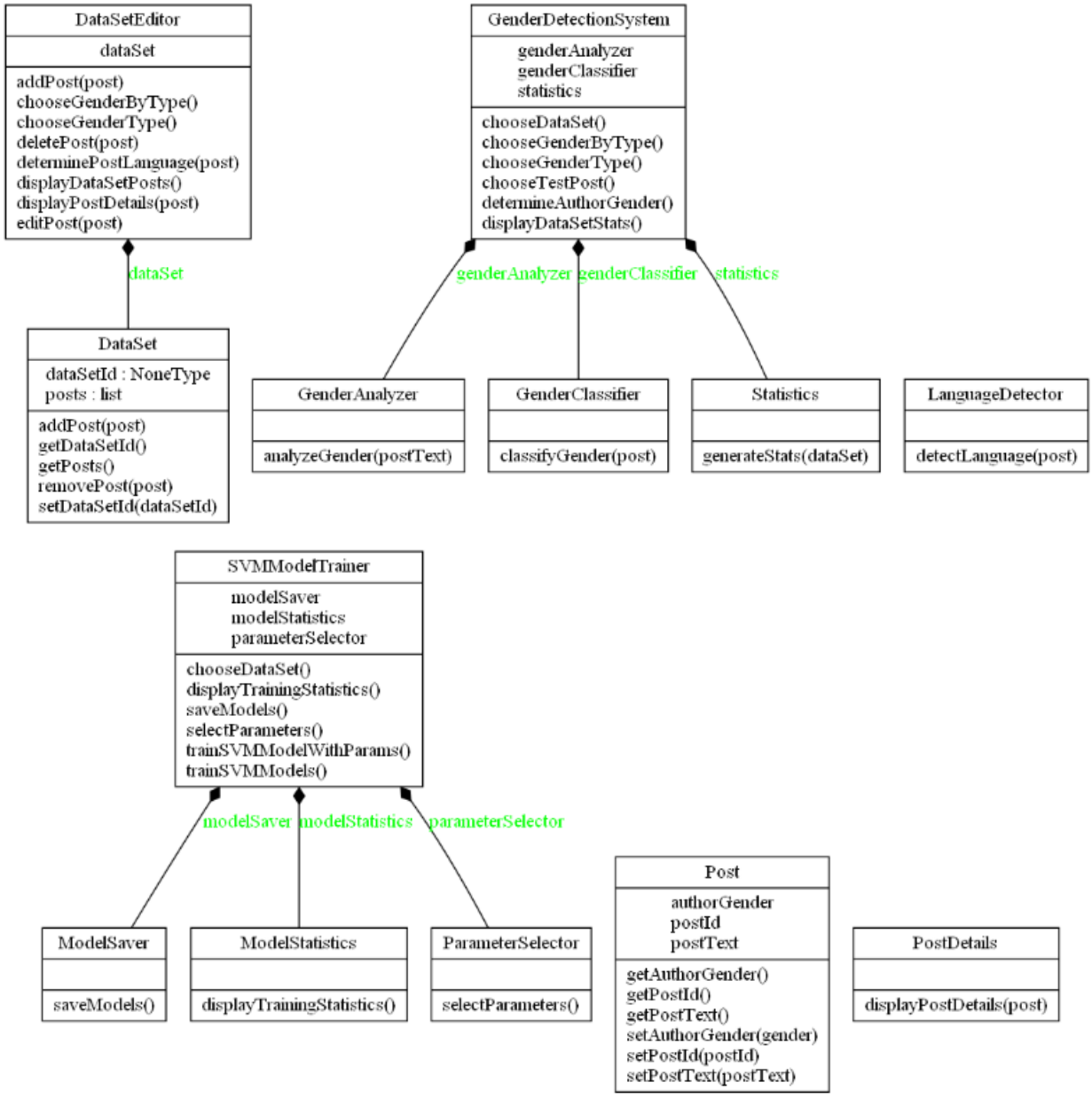
Додаток А

Структура бази даних інформаційної системи автоматизованого виявлення гендерної приналежності за дописами



Додаток Б

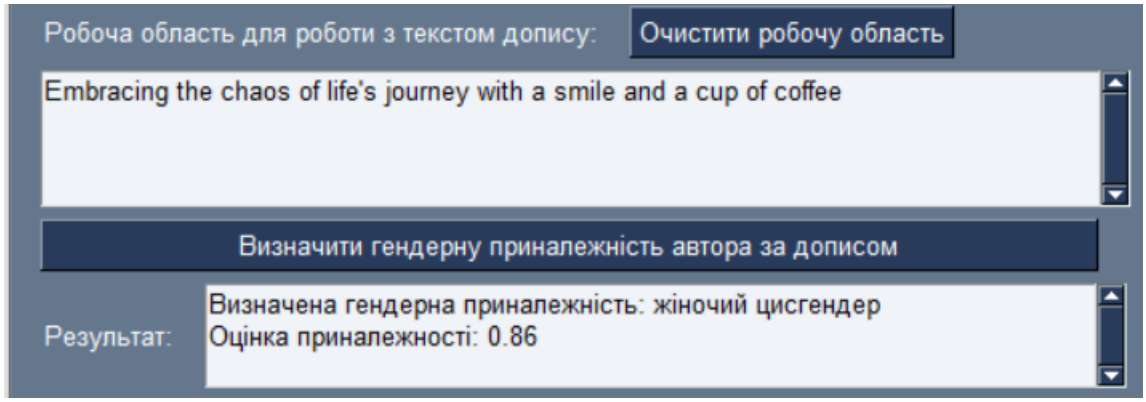
Розгорнута структура класів інформаційної системи автоматизованого виявлення гендерної приналежності за дописами

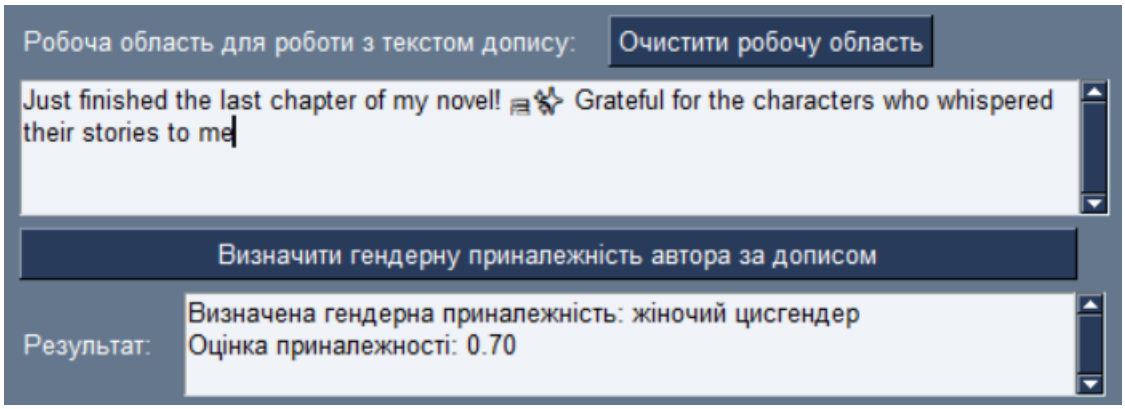


Додаток В

Ілюстрації роботи інформаційної системи виявлення гендерної приналежності за дописами

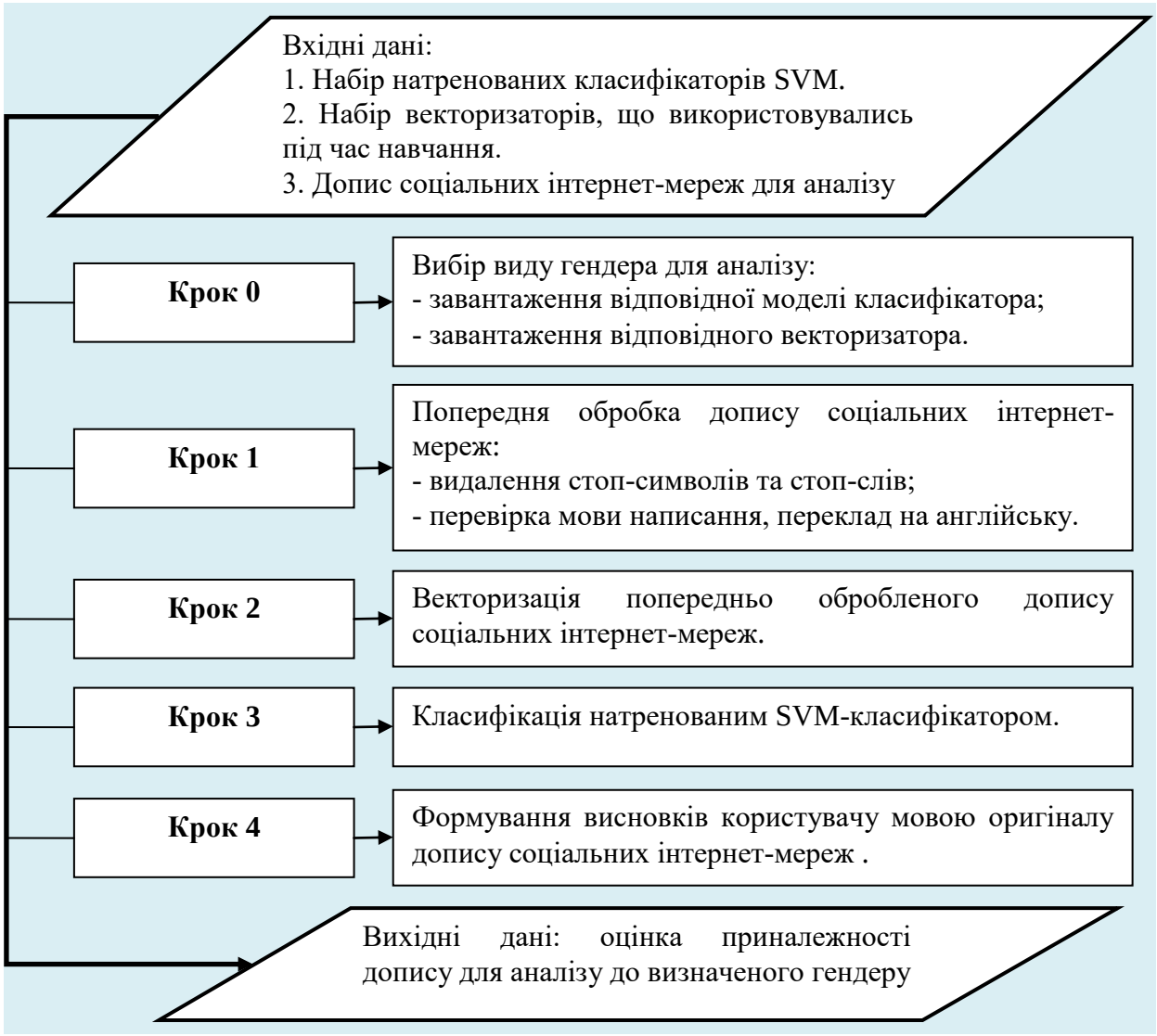
Вручну:


Робоча область для роботи з текстом допису:
Очистити робочу область
Embracing the chaos of life's journey with a smile and a cup of coffee
Визначити гендерну приналежність автора за дописом
Результат:
Визначена гендерна приналежність: жіночий цисгендер
Оцінка приналежності: 0.86


Робоча область для роботи з текстом допису:
Очистити робочу область
Just finished the last chapter of my novel! Grateful for the characters who whispered their stories to me
Визначити гендерну приналежність автора за дописом
Результат:
Визначена гендерна приналежність: жіночий цисгендер
Оцінка приналежності: 0.70

Додаток Г

Схема методу виявлення гендерної приналежності за дописами соціальних інтернет-мереж засобами NLP



Додаток Д

Презентаційний матеріал

КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА

МЕТОД ВИЯВЛЕННЯ ГЕНДЕРНОЇ ПРИНАЛЕЖНОСТІ ЗА ДОПИСАМИ СОЦІАЛЬНИХ ІНТЕРНЕТ-МЕРЕЖ ЗАСОБАМИ NLP

Виконав:
студент групи КНс-21-1
Павло СУПРУН

Керівник:
викладач кафедри КН
Марина МОЛЧАНОВА

Актуальність

Зростання використання соціальних мереж і стійка популярність онлайн-комунікацій робить завдання виявлення гендерної приналежності за дописами необхідним для різноманітних застосувань, включаючи маркетингові дослідження, аналіз громадської думки, персоналізовану рекламу, політичні дослідження та багато іншого.

Завдяки розвитку методів машинного навчання та обробки природної мови, можливості виявлення гендерної приналежності зростають. Ці дані можуть бути використані для аналізу та розуміння поведінки користувачів в інтернеті, виявлення та прогнозування тенденцій у споживчому ринку, а також для створення більш ефективних стратегій спілкування та взаємодії з аудиторією. Використання таких методів дозволяє уникнути стереотипів та прихованих асоціацій, а також забезпечити репрезентативність аналізу за гендерними критеріями.

Зростаюча увага до питань рівності та різноманітності також робить цей напрямок актуальним, оскільки він може використовуватися для аналізу та виявлення можливих стереотипів, дискримінації або нерівності, що можуть існувати в онлайн-спільнотах.

Мета і задачі роботи

Метою кваліфікаційної роботи бакалавра є покращення виявлення гендерної приналежності у дописах соціальних інтернет-мереж.шляхом автоматизації.

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

- виконати аналіз предметної області виявлення гендерної приналежності за текстовими даними;
- виконати огляд теоретичних підходів щодо можливості виявлення гендерної приналежності за дописами соціальних інтернет-мереж, обрати підхід для подальшої реалізації;
- виконати огляд існуючих наукових надбань та програмних рішень;
- створити метод виявлення гендерної приналежності за дописами соціальних інтернет-мереж засобами NLP та описати його кроки;
- спроектувати інформаційну систему для реалізації методу виявлення гендерної приналежності за дописами соціальних інтернет-мереж засобами NLP;
- створити програмну реалізацію за спроєктованою структурою інформаційної системи;
- виконати тестування створеної програмної реалізації;
- виконати дослідження ефективності методу виявлення гендерної приналежності за дописами соціальних інтернет-мереж засобами NLP на основі створеного ПЗ.

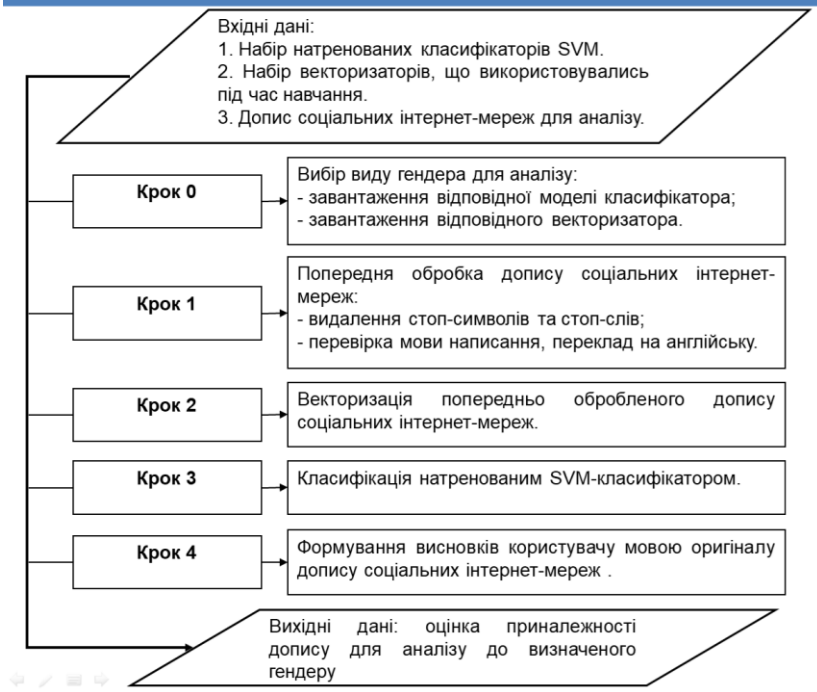


Схема та кроки методу виявлення гендерної приналежності за дописами соціальних інтернет-мереж

Схема та кроки навчання типового класифікатора

Схема навігації між формами інформаційної системи

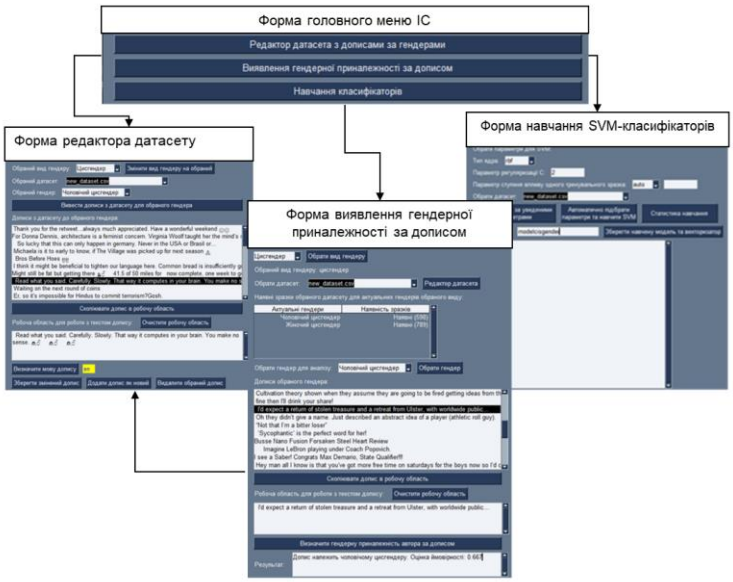
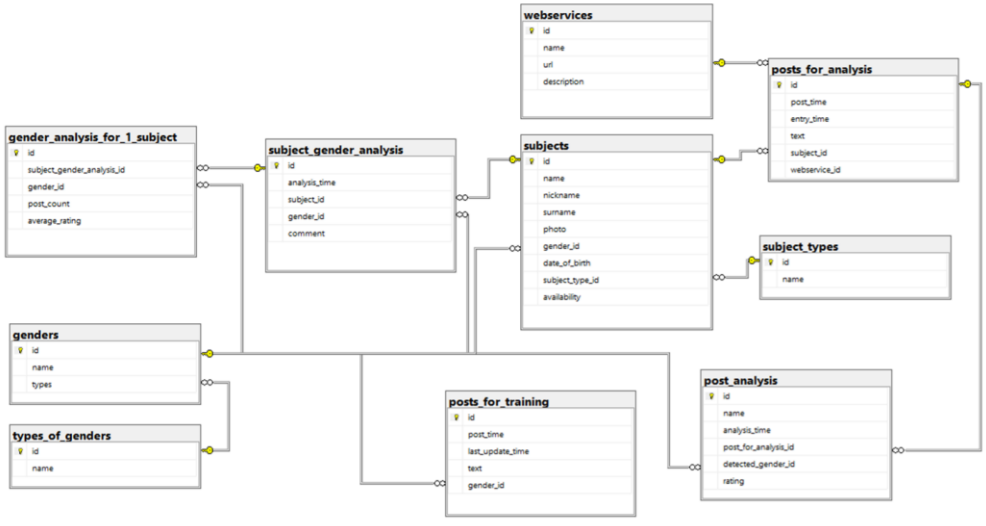


Схема інформаційної системи виявлення гендерної приналежності за дописами

Даталогічна модель бази даних



Підготовка робочих вхідних даних для системи

About this file

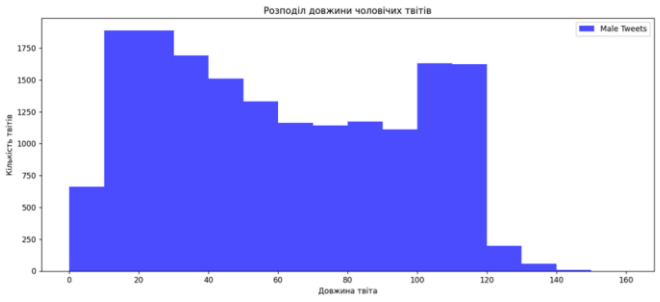
Add Suggestion

This file does not have a description yet.

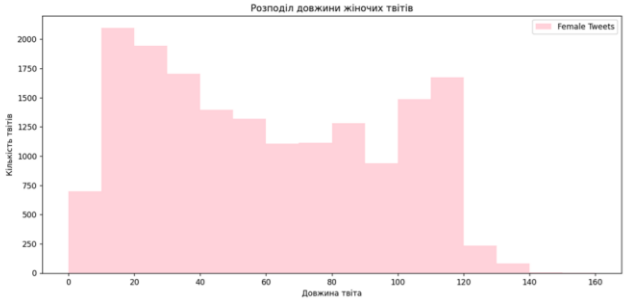
id	time	text	male
13.7k 1134907928b	2019-05-30 2019-06-01	10116 unique values	true 5225 50% false 5225 50%
304143752	2019-06-01 19:46:53+00:00	Can't stand a person who can't do shit without they phone	False
178901876	2019-05-30 16:38:43+00:00	They must have been brainwashed or something. See Nigerians talking rubbish o	True
1057259413717827584	2019-05-30 13:08:12+00:00	Lmao you literally think a sign of Muslim domination, the call to prayer, being played in your	False

Приклад даних датасету

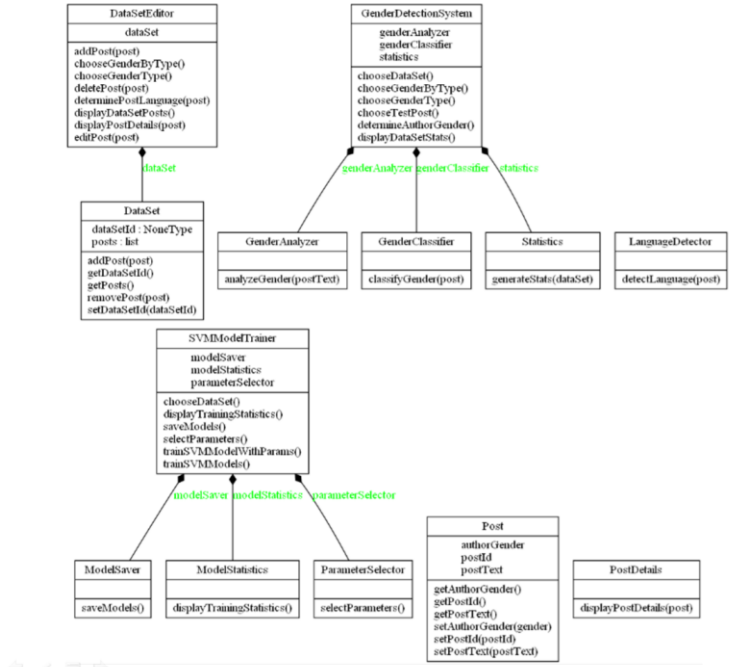
Підготовка робочих вхідних даних для системи



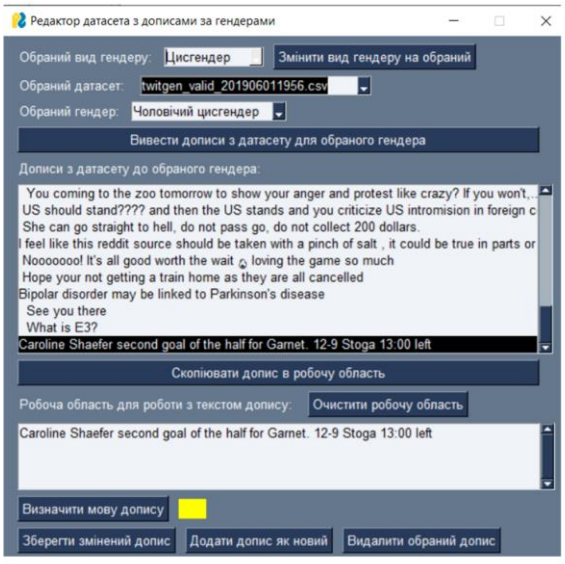
Розподіл твітів чоловічого цисгендеру за довжиною



Розподіл твітів жіночого цисгендеру за довжиною



Діаграма класів застосунку



Форма редактора датасета з дописами за гендерами

Інформаційна система виявлення гендерної приналежності за дописами соціальних інтернет-мереж

Підсистема виявлення гендерної приналежності за дописом

Інформаційна система виявлення гендерної приналежності за дописами соціальних інтернет-мереж

Результати досліджень

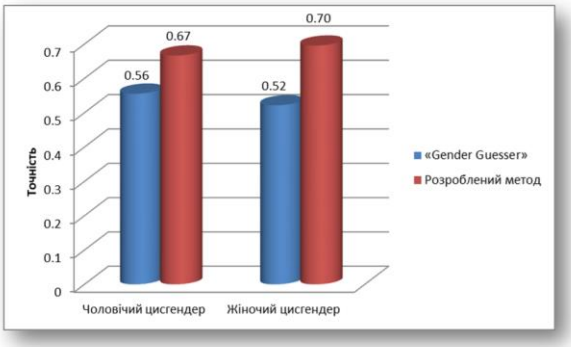
За створеним програмним за стосунком було досліджено ефективність запропонованого методу, яке буде відбуватись у два етапи. Першим етапом буде порівняння різних моделей навчених класифікаторів. Шляхом використання параметрів, заданих «вручну», та автоматичним пошуком

Параметри:	C	Ядро	GAMMA	Accuracy
Набір 1	1	rbf	auto	0.71
Набір 2	10	rbf	auto	0.68
Набір 3	0.1	rbf	scale = 0.5	0.66
Набір 4	1	poly	auto	0.67
Набір 5	10	poly	auto	0.65
Набір 6	0.1	poly	scale = 0.5	0.62
Набір 7	1	linear	auto	0.66
Набір 8	10	linear	auto	0.63
Набір 9	0.1	linear	scale = 0.5	0.62

Дослідження навчальних параметрів моделей SVM

Результати досліджень

Дослідження наборів параметрів



Діаграма виявлення цисгендеру альтернативними підходами

Висновки

Мета кваліфікаційної роботи бакалавра полягала в покращенні виявлення гендерної приналежності у дописах соціальних інтернет-мереж.шляхом автоматизації.

Для досягнення поставленої мети були поставлені та виконані такі задачі:

- виконано аналіз предметної області виявлення гендерної приналежності за текстовими даними;
- виконано огляд теоретичних підходів щодо можливості виявлення гендерної приналежності за дописами соціальних інтернет-мереж, обрано підхід для подальшої реалізації, в основі якого буде SVM-класифікатор;
- виконано огляд існуючих наукових надбань та програмних рішень;
- створено метод виявлення гендерної приналежності за дописами соціальних інтернет-мереж засобами NLP, що призначений для аналізу та класифікації текстових даних;
- спроектовано інформаційну систему для реалізації методу виявлення гендерної приналежності за дописами соціальних інтернет-мереж засобами NLP;
- створено програмну реалізацію за спроєктованою структурою інформаційної системи;
- виконано тестування створеної програмної реалізації, яке некоректно працюючих функцій не виявило;
- виконано дослідження ефективності методу виявлення гендерної приналежності за дописами соціальних інтернет-мереж засобами NLP на основі створеного ПЗ, розроблений метод показав високу ефективність, в порівнянні із аналогом його точність у рамках проведеного експерименту вища на 0.11. Також він може працювати з короткими текстами без втрати точності, такими як твіти.

UNICHECK
by Turnitin

Ім'я користувача:
Кафедра КН

ID перевірки:
1016379248

Дата перевірки:
20.06.2024 22:03:00 EEST

Тип перевірки:
Doc vs Internet + Library

Дата звіту:
20.06.2024 22:13:39 EEST

ID користувача:
100005671

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

Кількість сторінок: **71** Кількість слів: **10934** Кількість символів: **88257** Розмір файлу: **3.03 MB** ID файлу: **1016187986**

Виявлено модифікації тексту (можуть впливати на відсоток схожості)

10.9%
Схожість

Найбільша схожість: **5.09%** з джерелом з Бібліотеки (ID файлу: **1016181930**)

5.52% Джерела з Інтернету	426	Сторінка 73
9.37% Джерела з Бібліотеки	148	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	21 сторінка

Anti-Plagiarism v-15.257

Максимальне співпадіння з одним документом 4.0%

Словники перевірки: en_US, ru_RU, ua_UA. **Помилок в документах: 14%**

	Документ		Сумарний збіг по Базі Даних	
ID: 131882 Назва: КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА на тему Метод виявлення гендерної приналежності за дописами соціальних інтернет-мереж засобами NLP Додано в БД: 2024-06-20 Автора: Павло СУПРУН Керівники: Марина МОЛЧАНОВА Консультанти: Опоненти:	Символи	Лексеми	Символи	Лексеми
	68746	1002	5074 (7%)	79 (8%)

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 виявлення гендерної приналежності за дописами соціальних інтернет-мереж засобами NLP*

Автор: *студент групи КНс-21-1 Павло Супрун*

Спеціальність: *122 – Комп'ютерні науки*

Освітня програма: *освітньо-професійна*

Науковий керівник: *викладач Марина Молчанова*

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 Павла Супруна, 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.

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

- за системою Anti-Plagiarism: 4%;

- за системою Unicheck: 10.9%.

Керівник роботи _____ *Марина МОЛЧАНОВА*

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 Супруна Павла Костянтиновича*

за темою *Метод виявлення гендерної приналежності за дописами соціальних інтернет-мереж засобами NLP*

1. Актуальність теми

Визначення гендерної приналежності за дописами у соціальних інтернет-мережах є актуальним завданням з причини зростаючого значення соціальних мереж як платформи для взаємодії та вираження особистості. Метод дозволяє розуміти, як гендерна ідентичність впливає на поведінку, спілкування та сприйняття інформації в онлайн-середовищі, що є ключовим для культурологічних, психологічних та соціальних досліджень.

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

За стандартом, а саме описом предметної області, об'єктом в роботі є процес виявлення гендерної приналежності за дописами соціальних інтернет-мереж засобами NLP. Метою роботи покращення виявлення гендерної приналежності у дописах соціальних інтернет-мереж шляхом автоматизації. Для досягнення мети використано методи та засоби машинного навчання для роботи з текстовою інформацією. А, отже, результати виконання кваліфікаційної роботи бакалавра цілком відповідають стандарту бакалавра спеціальності 122 – Комп'ютерні науки.

3. Професійні та особистісні якості бакалавра

При роботі над кваліфікаційною роботою бакалавра Супрун Павло Костянтинович продемонстрував високий рівень самостійності. У процесі роботи студент виявив глибокі знання та розуміння предметної області, що дозволило йому ефективно розв'язувати поставлені завдання. Результати кваліфікаційної роботи свідчать про вміння студента креативно мислити та застосовувати актуальні методи і технології.

4. Ступінь самостійності під час виконання кваліфікаційної роботи

Студент відзначився високим рівнем самостійності та самодисципліни, виконуючи усі етапи кваліфікаційної роботи, що підтверджується успішною реалізацією роботи.

5. Ступінь оволодіння методами дослідження

Студент продемонстрував високий рівень компетенції у використанні методів дослідження. Його здатність застосовувати наукові підходи та методи показує глибоке розуміння теоретичних основ предметної області.

6. Повнота та якість розкриття теми роботи

Студент повністю розкрив тему своєї кваліфікаційної роботи бакалавра, шляхом детального аналізу предметної області та детальним описом кроків методу та його дослідження ефективності.

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

Матеріал роботи логічно структурований і послідовно викладений. Аргументи підтверджені науковими джерелами і прикладами. Викладення матеріалу літературно грамотне і зрозуміле.

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

Розроблений у роботі метод та його програмна реалізація може бути корисна для дослідження того, як гендерна ідентичність впливає на спосіб поведінки, комунікації та сприйняття інформації у віртуальному середовищі, що має важливе значення для культурологічних, психологічних та соціальних досліджень.

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 Марина МОЛЧАНОВА*

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
МОН УКРАЇНИ

Кафедра комп'ютерних наук

РЕЦЕНЗІЯ

на кваліфікаційну роботу бакалавра

студента *гр. КНс-21-1 Супруна Павла Костянтиновича*

за темою: *Метод виявлення гендерної приналежності за дописами соціальних інтернет-мереж засобами NLP*

1. Актуальність обраної теми

Актуальність методу виявлення гендерної приналежності за дописами соціальних інтернет-мереж засобами NLP визначається потребою у розвитку інструментів для аналізу інформації з великих обсягів даних. Метод дозволить автоматизовано визначати гендерну ідентичність користувачів за їхнім мовним виявом в онлайн-середовищі, що є важливим для різноманітних досліджень соціальних та психологічних аспектів поведінки в мережах.

2. Повнота розкриття мети та завдань роботи

У кваліфікаційній роботі бакалавра повністю було розкрито мету та завдання роботи, адже було глибоко проаналізовано предметну область, чітко та повною мірою наведено кроки методу та описано їх, а також проведено дослідження ефективності методу у розробленій інформаційній системі.

3. Зміст кожного розділу роботи

У роботі присутні три розділи, що присвячені різним етапам роботи. У розділі 1 наведено характеристику предметної області. У розділі 2 спроєктовано метод та інформаційну систему виявлення гендерної приналежності за дописами соціальних інтернет-мереж. У розділі 3 проведено експериментальне дослідження методу виявлення гендерної приналежності за дописами соціальних інтернет-мереж.

4. Оцінка розробленої інформаційної системи, її практична цінність

Розроблена інформаційна система спрямована на автоматизований аналіз текстів для визначення гендерної ідентифікації авторів, щоб досліджувати розподіл гендерних ролей і стереотипів у веб-середовищі, тому має практичну цінність.

5. Якість оформлення кваліфікаційної роботи бакалавра

Студент якісно оформив кваліфікаційну роботу бакалавра, адже матеріал викладено відповідно до структури записки. Також наведено посилання на сучасні джерела. Для наочного подання інформації використано таблиці, схеми та діаграми.

6. Недоліки кваліфікаційної роботи бакалавра

Робота виконана на високому рівні, проте має певні незначні недоліки. Пояснювальна записка містить незначні пунктуаційні та граматичні помилки. Деякі рисунки надто малі, що унеможливлює читання наявного на них тексту. Метод має природні обмеження в роботі, адже працює переважно з англомовними текстами.

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 к.т.н., доц. каф. КІС Кічмарук Л.О.