Хмельницький національний університет
Факультет інформаційних технологій
Кафедра комп'ютерних наук

КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА

на тему *Метод виявлення проявів етнічної ворожнечі у текстових повідомленнях соціальних інтернет-мереж NLP-засобами*

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* (Група виконавця) _____ (Підпис) *Ілля БОЯРЧУК* (Ім'я, ПРІЗВИЩЕ)

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

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

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

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 проявів етнічної ворожнечі у текстових повідомленнях соціальних інтернет-мереж NLP-засобами»*

2. Завдання видано студенту *Іллі БОЯРЧУКУ*
(Ім'я, прізвище)

3. Керівник роботи *викладач кафедри КН Марина МОЛЧАНОВА*
(посада, ім'я, прізвище)

4. Затверджено наказом університету від «*16*» *02* *2024* р. № *8*

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

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

Мета роботи – спрощення експертизи виявлення проявів етнічної ворожнечі за рахунок автоматизованого її виявлення у текстових повідомленнях соціальних інтернет-мереж. Для досягнення мети слід виконати такі задачі: виконати дослідження предметної області для задачі виявлення проявів етнічної ворожнечі у текстових повідомленнях; розробити метод виявлення проявів етнічної ворожнечі у текстових повідомленнях; здійснити програмну реалізацію інформаційної системи ідентифікації етнічної ворожнечі за текстовим представленням та провести дослідження ефективності розробленого 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* (Група виконавця) _____ (Підпис) *Ілля БОЯРЧУК* (Ім'я, ПРІЗВИЩЕ)

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

Анотація

Тема кваліфікаційної роботи бакалавра: *«Метод виявлення проявів етнічної ворожнечі у текстових повідомленнях соціальних інтернет-мереж NLP-засобами»*

Виконавець кваліфікаційної роботи бакалавра: *студент групи КНс-21-1 Ілля БОЯРЧУК*

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

Кваліфікаційна робота бакалавра містить:

Пояснювальна записка				Кількість додатків
Сторінок	Рисунків	Таблиць	Джерел інформації	
66	34	7	33	4

Метою кваліфікаційної роботи бакалавра є спрощення експертизи виявлення проявів етнічної ворожнечі за рахунок автоматизованого її виявлення у текстових повідомленнях соціальних інтернет-мереж. Для розробки прикладної інформаційної системи було використано мову програмування C#, а також спеціалізоване програмне розширення ML.NET для навчання та використання класифікатора. Розроблена система призначена для використання органами правопорядку, дослідниками етнічних конфліктів, аналітиками та соціологами, які цікавляться виявленням та аналізом етнічної ворожнечі у текстових повідомленнях соціальних мереж.

Напрямками практичного використання розробленої інформаційної системи є аналіз та виявлених випадків етнічної ворожнечі.

Ключові слова: FastForest, NLP, виявлення проявів етнічної ворожнечі.

Виконавець: *студент групи КНс-21-1* _____ *Ілля БОЯРЧУК*

Група виконавця | Підпис | Ім'я, ПРІЗВИЩЕ

Зміст

Перелік скорочень

Скорочення, термін, позначення	Пояснення
NLP	Natural Language Processing
SVM	Support Vector Machine
RNN	Recurrent Neural Network
BiLSTM	Bidirectional Long Short-Term Memory
BiGRU	Bidirectional Gated Recurrent Unit
GRU	Gated Recurrent Unit
API	Application Programming Interface
AI	Artificial Intelligence
PTLM	Transfer Learning from Pretrained Language Models
BERT	Bidirectional Encoder Representations from Transformers
XLM-R	Cross-lingual Language Model - RoBERTa
ML.NET	Machine Learning .NET
КН	Комп’ютерні науки
ПЗ	Пояснювальна записка
ПП	Програмний продукт
ХНУ	Хмельницький національний університет.
MS	Microsoft

Вступ

Метою кваліфікаційної роботи бакалавра є спрощення експертизи виявлення проявів етнічної ворожнечі за рахунок автоматизованого її виявлення у текстових повідомленнях соціальних інтернет-мереж, для чого проводилась розробка методу виявлення проявів етнічної ворожнечі у текстових повідомленнях соціальних інтернет-мереж NLP-засобами, а також відповідної інформаційної системи ідентифікації етнічної ворожнечі за текстовим представленням, яка використала створений метод.

Актуальність. За останні роки спостерігається стрімке зростання популярності соціальних мереж серед користувачів з усього світу. Вони стали важливим каналом для вираження думок, поглядів та емоцій. В той же час, у світі нерідко виникають конфлікти та напруженість між різними етнічними групами. Часто такі ситуації знаходять відображення у висловлюваннях у соціальних мережах, що створює потребу у виявленні та аналізі таких виразів.

Виявлення проявів етнічної ворожнечі є важливою проблемою для суспільства, оскільки вона може призвести до серйозних наслідків, включаючи конфлікти та розбрати між різними етнічними групами.

Засоби обробки природної мови наразі набувають великої популярності у сфері аналізу текстів. Вони дозволяють автоматизувати та полегшити аналіз великих обсягів даних, включаючи тексти, що публікуються в соціальних мережах.

Об’єкт дослідження – процес виявлення проявів етнічної ворожнечі у текстових повідомленнях соціальних інтернет-мереж NLP-засобами.

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

Мета кваліфікаційної роботи бакалавра полягає в спрощенні експертизи виявлення проявів етнічної ворожнечі за рахунок автоматизованого її виявлення у текстових повідомленнях соціальних інтернет-мереж..

Завдання кваліфікаційної роботи бакалавра – виконати дослідження предметної області для задачі виявлення проявів етнічної ворожнечі у текстових повідомленнях; в рамках дослідження предметної області виконати огляд теоретичних підходів щодо виявлення проявів етнічної ворожнечі у текстових повідомленнях; виконати аналіз існуючих програмних рішень в області виявлення проявів етнічної ворожнечі у текстових повідомленнях; розробити метод виявлення проявів етнічної ворожнечі у текстових повідомленнях; на основі розробленого методу виконати проєктування інформаційної структури системи ідентифікації етнічної ворожнечі за текстовим представленням; виконати підготовку навчальних даних для тренування класифікатора; здійснити вибір засобів розробки для створення інформаційної системи; здійснити програмну реалізацію інформаційної системи ідентифікації етнічної ворожнечі за текстовим представленням; провести тестування розробленої програмної реалізації; здійснити дослідження ефективності розробленого методу методу виявлення проявів етнічної ворожнечі у текстових повідомленнях соціальних інтернет-мереж NLP-засобами з використанням розробленої програмної реалізації.

Розділ 1 Характеристика предметної області щодо виявлення проявів етнічної ворожнечі у текстових повідомленнях

1.1 Аналіз інформаційних моделей виявлення проявів етнічної ворожнечі

Сучасні соціальні мережі відіграють важливу роль у формуванні суспільного діалогу та обговоренні різноманітних питань. Це великі платформи, які дозволяють мільйонам користувачів обмінюватися ідеями, інформацією та поглядами з усього світу. Глобальна комунікація, що відбувається в соціальних мережах, дозволяє обговорювати події та питання на глобальному рівні [1].

Ці платформи стали публічним простором для різних дебатів, що охоплюють політику, науку, культуру та інші аспекти життя. Люди можуть об'єднуватися для обговорення конкретних питань, мобілізуючи громадську думку та впливаючи на суспільні зміни. Соціальні мережі сприяють збільшенню свідомості про різні питання, включаючи соціальні, екологічні та гуманітарні проблеми. Різноманіття голосів та поглядів, яке представлено в соціальних мережах, збагачує дискусії та сприяє різнобічному розгляду питань [2].

Однак важливо враховувати, що у соціальних мережах має місце поширення дезінформації, кібербулінг, порушення приватності та інші проблеми, які вимагають уваги та регулювання. Таким чином, роль соціальних мереж у сучасному суспільстві є багатогранною та вимагає уважного розгляду всіх аспектів їх впливу [3].

Сучасні соціальні мережі є динамічним та різнобарвним віртуальним простором, населеним різноманітними типами користувачів. Кожен з них вносить свій унікальний внесок у формування спільнот та динаміку в цифровому середовищі. Звичайні користувачі, які обмінюються думками та враженнями, модератори, що стежать за порядком, адміністратори, які визначають стратегії розвитку, та інші учасники створюють мозаїку різних ролей, які взаємодіють у цьому цифровому віртуальному світі [4].

Проблеми, такі як поширення дезінформації, кібербулінгу, етнічної ворожнечі стають об'єктом уваги та регулювання з боку різних сторін, включаючи самі соціальні мережі, урядові органи, організації громадського контролю та користувачів. Соціальні мережі активно розробляють та впроваджують заходи для вирішення проблем дезінформації, кібербулінгу, етнічної ворожнечі, ставлячи перед собою завдання створити безпечне та позитивне середовище для своїх користувачів [5].

У боротьбі з дезінформацією, соціальні мережі використовують алгоритми машинного навчання для автоматичного виявлення та фільтрації дезінформації, кібербулінгу, етнічної ворожнечі. Ці алгоритми аналізують величезний обсяг контенту, використовуючи певні ознаки, такі як заголовки, ключові слова, чутливість до контексту тощо. Застосування цих технологій дозволяє виявляти та блокувати потенційно шкідливий вміст [6].

Однак, крім автоматизованих інструментів, соціальні мережі співпрацюють із фактчекерськими агентствами та експертами для ручної перевірки контенту. Позначення контенту як неправдивого та виведення інформації від фактчекерів може допомагати користувачам розрізняти достовірні джерела в інтернеті [7].

У сфері протидії кібербулінгу, соціальні мережі вживають заходів для ідентифікації та припинення агресивного чи образливого вмісту. Модератори відіграють важливу роль у моніторингу та видаленні такого контенту, але важливо підкреслити, що ручна модерація може бути обмеженою через великий обсяг інформації. Тому соціальні мережі також розробляють і вдосконалюють алгоритми для автоматичного виявлення порушень правил спільноти, що спрощує та прискорює реакцію на інциденти кібербулінгу [8].

Загальний наголос робиться на підвищенні свідомості користувачів та розвитку їхньої критичної обізнаності. Соціальні мережі надають освітні матеріали, оголошення та рекомендації, які допомагають користувачам впоратися з дезінформацією та розпізнавати прояви кібербулінгу. Взаємодія зі

спільнотою та включення користувачів у процес розробки політик також важливі для підвищення відповідальності та розуміння проблем в масштабі спільності.

Особливо гострою також є проблема етнічної ворожнечі в соціальних мережах. Етнічна ворожнеча (також відома як етнічний конфлікт, міжетнічна ворожнеча або етнічні конфлікти) – це форма конфлікту, що виникає між представниками різних етнічних груп [9]. Ця ворожнеча може мати різні форми, включаючи соціальне, економічне та політичне підґрунтя. Етнічна ворожнеча може виникати між користувачами соціальних мереж з різних причин [10]:

– конфлікт виникає через різниці в етнічних ідентичностях та самосвідомості різних груп;

– етнічні конфлікти часто пов'язані з претензіями на певні території чи ресурси;

– нерівні умови життя, обмежений доступ до ресурсів, а також дискримінація можуть сприяти виникненню етнічної ворожнечі;

– події з історії, такі як конфлікти або пригнічення, можуть вливатися в сучасні етнічні відносини;

– конкуренція за ресурси та можливості може призводити до економічних аспектів етнічної ворожнечі;

– влада, політика та розподіл ресурсів можуть стати причиною етнічних конфліктів;

– релігійні відмінності можуть впливати на етнічні взаємини та викликати конфлікти.

Етнічна ворожнеча в соціальних мережах може служити інструментом для різних груп та індивідів, маючи різні мотивації та цілі. Однією з таких мотивацій є політичні цілі, де розпалювання етнічної ворожнечі може бути використане для дестабілізації конкурентів чи політичних опонентів, спрямовуючи національне або етнічне розгортання проти них [11].

Крім того, релігійні та культурні конфлікти можуть викликати етнічну ворожнечу як засіб об'єднання та зміцнення впливу певних груп. Особливо це

стосується випадків, коли етнічна або релігійна ідентичність використовується як інструмент для мобілізації та розвитку власної спільноти.

Ознаки етнічної ворожнечі в текстових повідомленнях можуть виявлятися через вживання специфічної мови та виразів, які вказують на негативне ставлення до певної етнічної групи чи виявляють ворожнечу. Далі розглянуто кілька ознак [11]:

– вживання різних расистських слів, образливих термінів або етнічних образливих слів;

– висловлення стереотипів та генералізації про певну етнічну групу;

– тексти, які містять загрози чи призиви до насильства проти представників певної етнічної групи;

– тексти з використанням агресивної або негативної тональності, а також образливі словесні звороти;

– розповсюдження або підтримка дискримінаційних міфів та перекручень;

– вживання мови або виразів, що вказують на негативні враження чи ворожнечу до певної етнічної групи;

– за допомогою текстових повідомлень може поширюватися фейкові новини чи дезінформація про певну етнічну групу;

– коментарі, що виявляють агресію, ворожнечу або викликають конфлікти.

Виявлення проявів етнічної ворожнечі в соціальних мережах має важливе соціокультурне та етичне значення, оскільки стосується як безпеки, так і динаміки онлайн-спільнот. Перш за все, етнічна ворожнеча може стати джерелом реальних загроз для безпеки користувачів. Виявлення та вчасна реакція на образливий вміст може запобігти поширенню ненависті, уникненню конфліктів та захисту прав та гідності користувачів. Також виявлення етнічної ворожнечі важливе з погляду підтримання позитивного та відкритого інтернет-середовища. Розпізнавання та припинення негативних взаємодій допомагає формувати безпечний простір для спілкування, обміну ідеями та взаємодії між

користувачами різних культур та етнічностей. Це сприяє розвитку гармонійних міжкультурних відносин та взаєморозуміння [14].

Крім того, виявлення етнічної ворожнечі в соціальних мережах є важливим для запобігання поширенню дезінформації та фейкових новин, які часто використовуються для підкріплення стереотипів та формування негативного ставлення до певних етнічних груп.

З точки зору етики, робота над виявленням та припиненням етнічної ворожнечі відображає зобов'язання соціальних мереж працювати на користь спільноти та створювати платформи, де різноманіття вітається, а діалог є конструктивним. В цілому, виявлення проявів етнічної ворожнечі сприяє формуванню позитивного онлайн-середовища, що сприяє взаєморозумінню, толерантності та відкритому обміну ідеями. Тому в рамках роботи доцільно автоматизувати процеси виявлення етнічної ворожнечі у соціальних інтернет-мережах, використовуючи засоби інформаційних технологій.

1.2 Огляд теоретичних підходів до розв’язку задачі виявлення проявів етнічної ворожнечі у текстовому контенті

В сучасному аналізі текстового контенту соціальних інтернет-мереж для виявлення проявів етнічної ворожнечі використовуються різні алгоритми, методи та моделі штучного інтелекту. Одним із ключових підходів є застосування класифікаторів, які можуть автоматично визначати категорію тексту на основі навчання на позначених даних. Класифікатори, такі як метод опорних векторів, логістична регресія, дерева рішень та їхні ансамблі, можуть визначати, чи містить текст ознаки етнічної ворожнечі.

SVM – це вид алгоритмів навчання з учителем, які використовуються для завдань класифікації та регресії. Вони широко використовуються в різних галузях, включаючи розпізнавання образів, аналіз зображень та обробку природної мови. Основною метою алгоритму машинного навчання опорних

векторів є визначення гіперплощини у N-вимірному просторі (де N представляє кількість ознак), яка чітко визначає класифікацію точок даних [14].

Логістична регресія використовується для розв'язання задач двійкової класифікації. У цьому методі використовується сигмоїдна функція, яка обробляє незалежні змінні вхідних даних, генеруючи ймовірності в межах від 0 до 1. Наприклад, якщо у нас є два класи (клас 0 і клас 1), то, якщо значення логістичної функції для вхідних даних перевищує 0,5 (порогове значення), воно відноситься до класу 1, інакше воно відноситься до класу 0 [15].

Алгоритм Gradient Boosting є широко використовуваним методом у машинному навчанні, призначеним для розв'язання завдань класифікації та регресії. Це відомий метод ансамблевого навчання, що використовує послідовне навчання моделей, при чому кожна нова модель спрямована на виправлення попередньої. Основна ідея полягає в об'єднанні декількох слабких моделей для створення потужної та ефективної [16].

В області обробки природної мови використовуються різні архітектури нейронних мереж для розв'язання завдань, пов'язаних з текстовою інформацією, наприклад рекурентні нейронні мережі (RNN), довга короткочасна пам'ять, багатокласові складні мережі (BiLSTM, BiGRU, тощо).

RNN є однією з перших архітектур, використовуваних у NLP. Вони дозволяють моделі враховувати контекст і послідовність даних, що є важливим для багатьох завдань, таких як машинний переклад та аналіз відношень [17].

LSTM є розширенням RNN та дозволяє ефективно управляти проблемою зниклого градієнту. Вони часто використовуються для розв'язання завдань, де важливий контекст має довгострокове значення, наприклад, у генерації тексту [18].

Багатокласові рекурентні мережі використовуються для одночасної обробки тексту в обох напрямках. Наприклад, BiLSTM (багатонапрямкова LSTM) та BiGRU (багатонапрямкова GRU) дозволяють моделі аналізувати контекст як зліва, так і справа від поточного слова.

Архітектура типової BiLSTM-моделі наведена на рисунку 1.1.

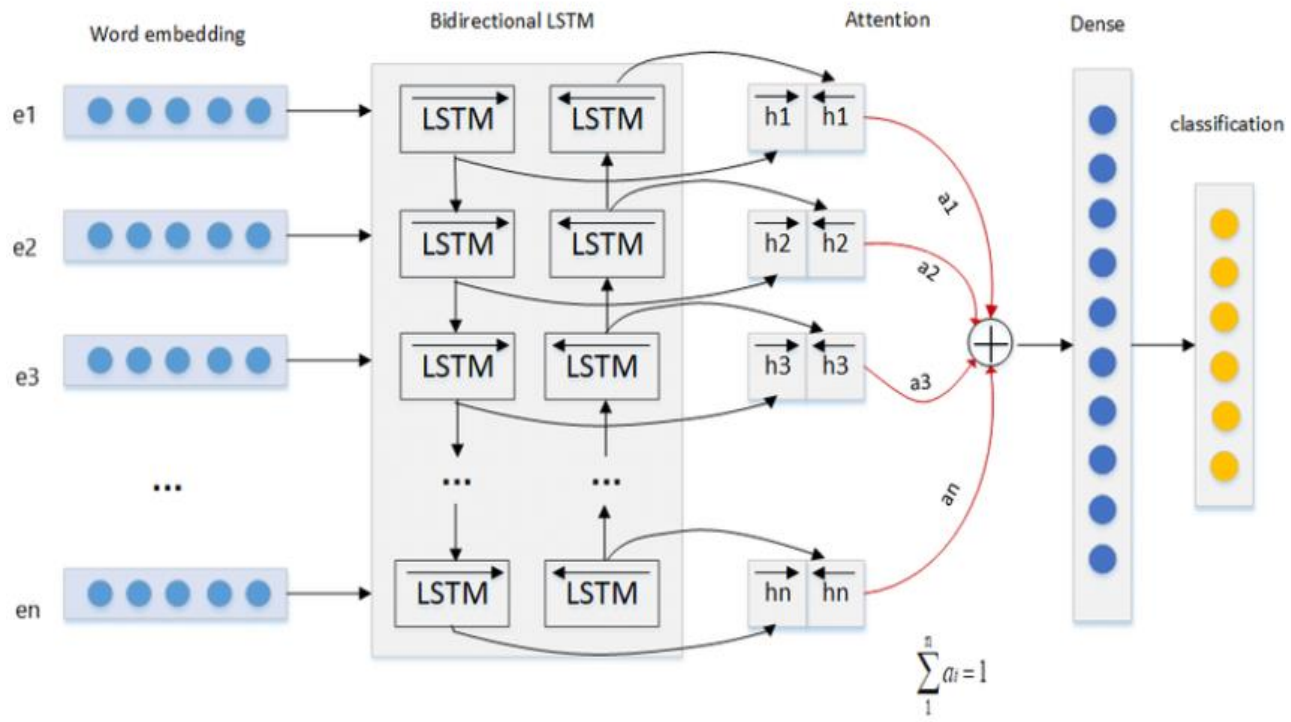


Рисунок 1.1 – Архітектура BiLSTM [18]

Також часто використовуються ансамблі нейронних мереж. Ансамблі нейронних мереж – це підхід, при якому кілька моделей об'єднуються, щоб вирішити задачу. Вони можуть бути ефективними у різних випадках, і їхня ефективність може залежати від конкретного завдання та даних. Ансамблі дозволяють зменшити ризик перенавчання, оскільки кожна модель може виокремлювати різні аспекти даних. Ансамблі можуть підвищити стійкість та точність моделі, оскільки різні моделі можуть вирішувати ту саму задачу з різних поглядів [19].

Ансамблевий алгоритм FastForest є алгоритмом машинного навчання, який використовується для задач класифікації та регресії. Він схожий на алгоритм випадкового лісу (Random Forest), але має ряд оптимізацій, що роблять його значно швидшим, не жертвуючи при цьому точністю [20].

В даному алгоритмі суббеггінг (Subbagging) використовується для випадкового вибору підмножини даних для тренування кожного дерева в лісі, що робить процес тренування швидшим і може допомогти запобігти

перенавчанню. А логарифмічне вибирання точок розбиття використовується для більш ефективного вибору точок розбиття для кожного дерева..

Загалом, використання різноманітних алгоритмів та моделей штучного інтелекту дозволяє автоматизовано та ефективно виявляти прояви етнічної ворожнечі у текстових повідомленнях соціальних інтернет-мереж, сприяючи покращенню безпеки та комфорту користувачів у цьому цифровому середовищі. У контексті виявлення етнічної ворожнечі в текстах соціальних мереж застосування NLP-засобів виявляється ключовим. Для цієї задачі використовуватимуться NLP-засоби, які підтримують аналіз текстового контенту, виявлення тону, класифікацію текстів та розпізнавання емоцій, а саме методи ансамблевого навчання.

1.3 Аналіз існуючих програмних засобів та наукових рішень

Виявлення етнічної ворожнечі у текстових повідомленнях стає все більше актуальним завданням у сучасному інформаційному середовищі. Це важливе завдання з погляду забезпечення безпеки та відсіювання потенційно шкідливого контенту в соціальних мережах та онлайн-спільнотах. На сьогоднішній день існує ряд API, які розробники можуть використовувати для імплементації власних рішень у своїх проєктах.

Google пропонує два API, які можуть виявляти етнічну ворожнечу у текстових повідомленнях – це Hate Speech Detection API та Perspective API.

Hate Speech Detection API – це інструмент від Google Cloud Platform, який використовує машинне навчання для виявлення мови ворожнечі, включаючи етнічну ворожнечу, у тексті [21]. Він може аналізувати текст на 43 мовах, включаючи українську. Hate Speech Detection API пропонує три рівні аналізу:

– швидкий – рівень дає уявлення про те, чи містить текст мову ворожнечі;

– детальний – рівень надає більш детальну інформацію про тип мови ворожнечі, яка була виявлена, а також про те, на які групи людей вона спрямована;

– класифікація – рівень дає вам змогу класифікувати текст за категоріями, такими як «ненависть», «образа» або «нейтральний».

Perspective API – це інструмент від Google Cloud Platform, який використовує машинне навчання для оцінювання токсичності тексту, включаючи етнічну ворожнечу [22]. Він може аналізувати текст на 103 мовах, включаючи українську. Perspective API дає оцінку ймовірності того, що текст буде сприйнятий як токсичний. Він також надає інформацію про те, які аспекти тексту роблять його токсичним (рисунок 1.2).

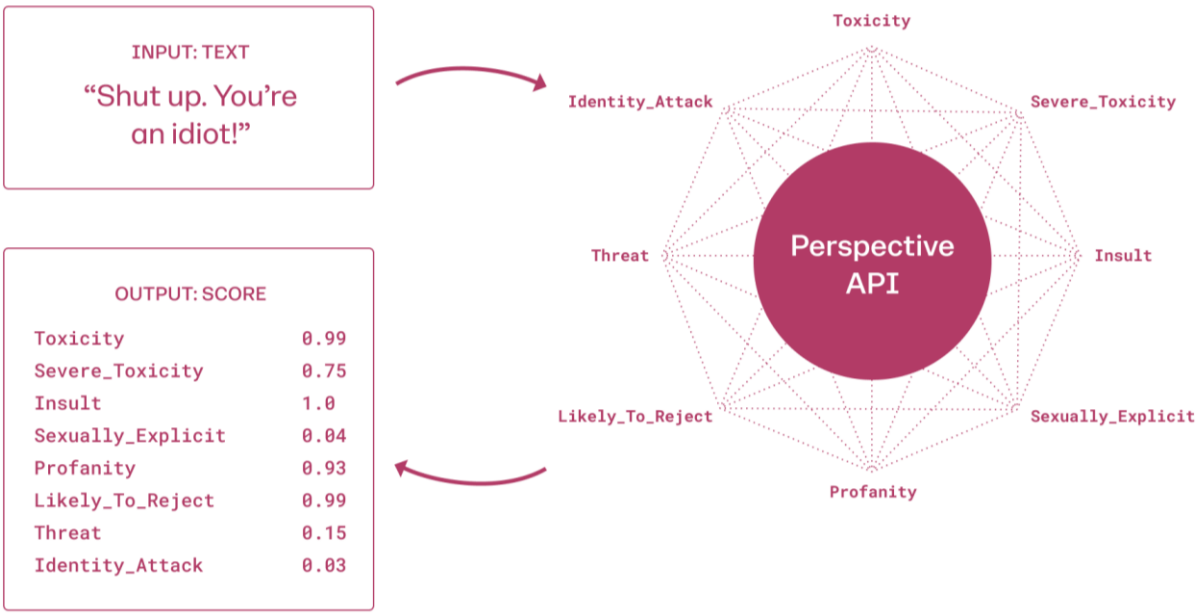


Рисунок 1.2 – Perspective API [22]

Для модерації неприйнятного контенту Azure AI Services пропонує Azure AI Content Safety – це платформа модерації вмісту, яка використовує штучний інтелект для захисту програмних продуктів (рисунок 1.3). Потужні моделі AI, швидко й ефективно виявляють образливий або неприйнятний вміст у тексті та зображеннях. Класифікатори штучного інтелекту визначають вміст контету що

містять ознаки насильства, ненависті та самоушкодження з високим рівнем деталізації [23].

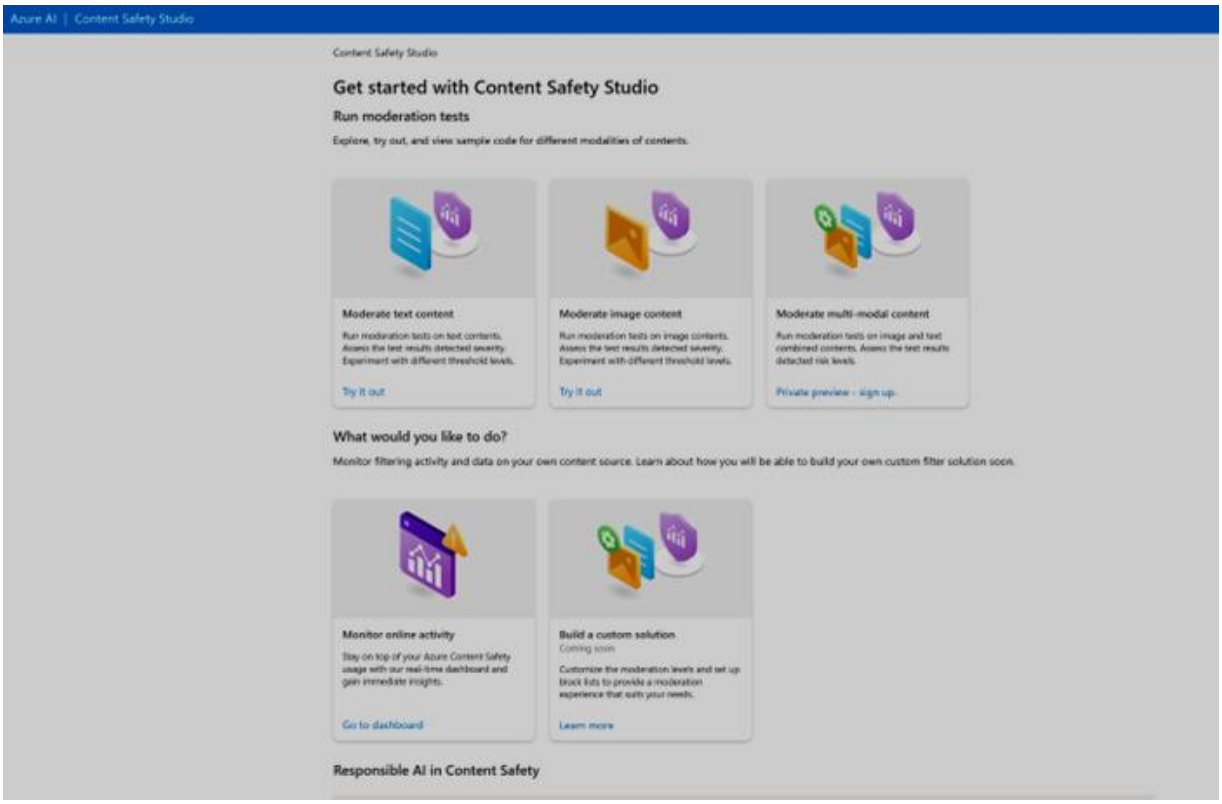


Рисунок 1.3 – AI Content Safety [23]

Amazon Comprehend – це сервіс від Amazon Web Services (AWS), який надає можливості для обробки природної мови. Цей інструмент створений для розуміння та аналізу текстової інформації, що дозволяє підприємствам та розробникам отримувати цінні інсайти з текстових даних. Одними із ключових можливостей в контексті виявлення етнічної ворожнечі у текстових повідомленнях, є визначення настрою тексту, такого як позитивний, негативний чи нейтральний, що допомагає зрозуміти емоційний відтінок, а також витягування важливих іменованих сутностей, таких як імена людей, організації, місця тощо. Amazon Comprehend використовує передові алгоритми машинного навчання, щоб автоматично визначати мовні конструкції та витягувати ключову інформацію з текстів [24].

Окрім існуючих API, активно проводяться наукові дослідження у напрямку розробки більш ефективних методів виявлення мови ворожнечі у

текстах соціальних мереж. Ці дослідження фокусуються на розробці нових моделей та алгоритмів, які не лише забезпечують високу точність виявлення загальної ворожнечі, але і спеціально звертають увагу на аспекти етнічної ворожнечі.

Одним з напрямків є використання глибокого навчання та нейронних мереж для аналізу текстів і виявлення субтильних сигналів, які можуть вказувати на етнічно спрямовану ворожнечу. Субтильні сигнали в контексті виявлення мови ворожнечі у текстових повідомленнях вказують на тонкі та неочевидні елементи, які можуть свідчити про наявність ворожнечі або агресивного ставлення в тексті. Ці сигнали можуть бути вираженими через мовні конструкції, образливі висловлювання, використання стереотипів або ключові слова, що можуть бути пов'язані з етнічністю [25].

Наприклад, у статті [26] розглядається проблема онлайн-ворожнечі, зокрема hate speech, яка є шкідливим вмістом в Інтернеті, що напряму або опосередковано атакує або пропагує ненависть до групи або окремого члена на основі їхньої реальної чи уявної ідентичності, такої як етнічність, релігія тощо. За останній час спостерігається зростання інтересу до автоматичного виявлення онлайн-ворожнечі в рамках завдань обробки природної мови. Однак останні дослідження показали, що існуючі моделі слабко узагальнюються на невидимі дані. Стаття пропонує узагальнену оглядову роботу, метою якої є аналіз того, наскільки ефективно існуючі моделі виявлення ворожнечі узагальнюються та причини, з яких вони мають труднощі з узагальненням.

У статті [27] обговорюється зростаюча поширеність онлайн-ворожнечі на соціальних мережах. Дослідження має на меті розробити ефективний фреймворк для виявлення онлайн-ворожнечі та образливих висловлювань для арабського тексту з метою вирішення цього серйозного питання. Автори побудували надійний арабський текстовий корпус, використовуючи дані з Twitter та чотири надійні стратегії вилучення, спрямовані на чотири типи ворожнечі: релігійну, етнічну, національну та за статевою ознакою. У заключній частині проведено інтенсивний експеримент для оцінки продуктивності різних навчених моделей та

вивчення помилок класифікації. Результати продуктивності є дуже обнадійливими порівняно з попередніми дослідженнями ворожнечі та образливих висловлювань, проведеними для арабської та інших мов.

У статті [28] обговорюється проблема поширення онлайн-ворожнечі на соціальних мережах у зв'язку зі зростанням їх впливу по всьому світу. Автори пропонують новий метод виявлення ворожнечі у текстах за допомогою крос-мовного навчання. Їхній підхід використовує трансферне навчання з попередньо навчених мовних моделей на великих корпусах для вирішення проблем у мовах з меншими ресурсами для конкретної задачі. Методологія включає чотири етапи: отримання корпусів, визначення PTLM, стратегії навчання та оцінку. В експериментах використовувались попередньо навчені мовні моделі (BERT та XLM-R) для англійської, італійської та португальської мов. Джерелом текстів слугували корпуси англійської (WH) та італійської (Evalita 2018) мов, а мовою призначення була португальська (корпус OffComBr-2). Результати експериментів свідчать про те, що запропонована методологія є перспективною: для корпусу OffComBr-2 отримано кращий результат серед сучасних (F1-вимір = 92%).

Отже, враховуючи різноманітне API для виявлення етнічної ворожнечі разом із дослідженнями, проведеними у галузі виявлення hate speech та етнічної ворожнечі в текстових повідомленнях соціальних мереж, можна зробити висновок, що існує активний інтерес і різнобічні підходи до вирішення проблеми етнічної ворожнечі в онлайн-середовищі. Однак аналіз проведених досліджень вказує на те, що варто робити акцент на особливості мов, для яких планується використовувати розроблений метод, а саме мовні конструкції, образливі висловлювання, використання стереотипів або ключові слова, що можуть бути пов'язані з етнічною ворожнечею для конкретної мови. Тому подальші дослідження у цій області є актуальними.

1.4 Мета, задачі та вимоги до реалізації інформаційної системи

Метою кваліфікаційної роботи бакалавра є спрощення експертизи виявлення проявів етнічної ворожнечі за рахунок автоматизованого її виявлення у текстових повідомленнях соціальних інтернет-мереж, для чого потрібно розробити метод виявлення проявів етнічної ворожнечі у текстових повідомленнях соціальних інтернет-мереж NLP-засобами, а також відповідної програмної реалізації, яка буде використовувати створений метод.

Для досягнення мети потрібно виконати наступні задачі:

– виконати дослідження предметної області для задачі виявлення проявів етнічної ворожнечі у текстових повідомленнях;

– в рамках дослідження предметної області виконати огляд теоретичних підходів щодо виявлення проявів етнічної ворожнечі у текстових повідомленнях;

– виконати аналіз існуючих програмних рішень в області виявлення проявів етнічної ворожнечі у текстових повідомленнях;

– розробити метод виявлення проявів етнічної ворожнечі у текстових повідомленнях;

– на основі розробленого методу виконати проєктування інформаційної структури системи ідентифікації етнічної ворожнечі за текстовим представленням;

– виконати підготовку навчальних даних для тренування класифікатора;

– здійснити вибір засобів розробки для створення інформаційної системи;

– здійснити програмну реалізацію інформаційної системи ідентифікації етнічної ворожнечі за текстовим представленням;

– провести тестування розробленої програмної реалізації;

– здійснити дослідження ефективності розробленого методу виявлення проявів етнічної ворожнечі у текстових повідомленнях соціальних інтернет-мереж NLP-засобами з використанням розробленої програмної реалізації.

Розділ 2 Розробка методу виявлення проявів етнічної ворожнечі у текстових повідомленнях соціальних інтернет-мереж

2.1 Схема та кроки методу виявлення проявів етнічної ворожнечі у текстових повідомленнях

Метод виявлення проявів етнічної ворожнечі у текстових повідомленнях соціальних інтернет-мереж NLP-засобами призначений для автоматизованого аналізу текстів, що публікуються у соціальних мережах з метою виявлення ознак ворожнечі або конфлікту між представниками різних етнічних груп. Метод використовує техніки обробки природної мови, а саме підхід на основі ансамблів, і здійснює перетворення вхідних даних у вигляді навченого класифікатора FastForest та вхідного текстового повідомлення у вихідні дані у вигляді відсотка прояву етнічної ворожнечі у тестовому повідомленні соціальних інтернет-мереж. Схема та кроки методу наведена на рисунку 2.1.

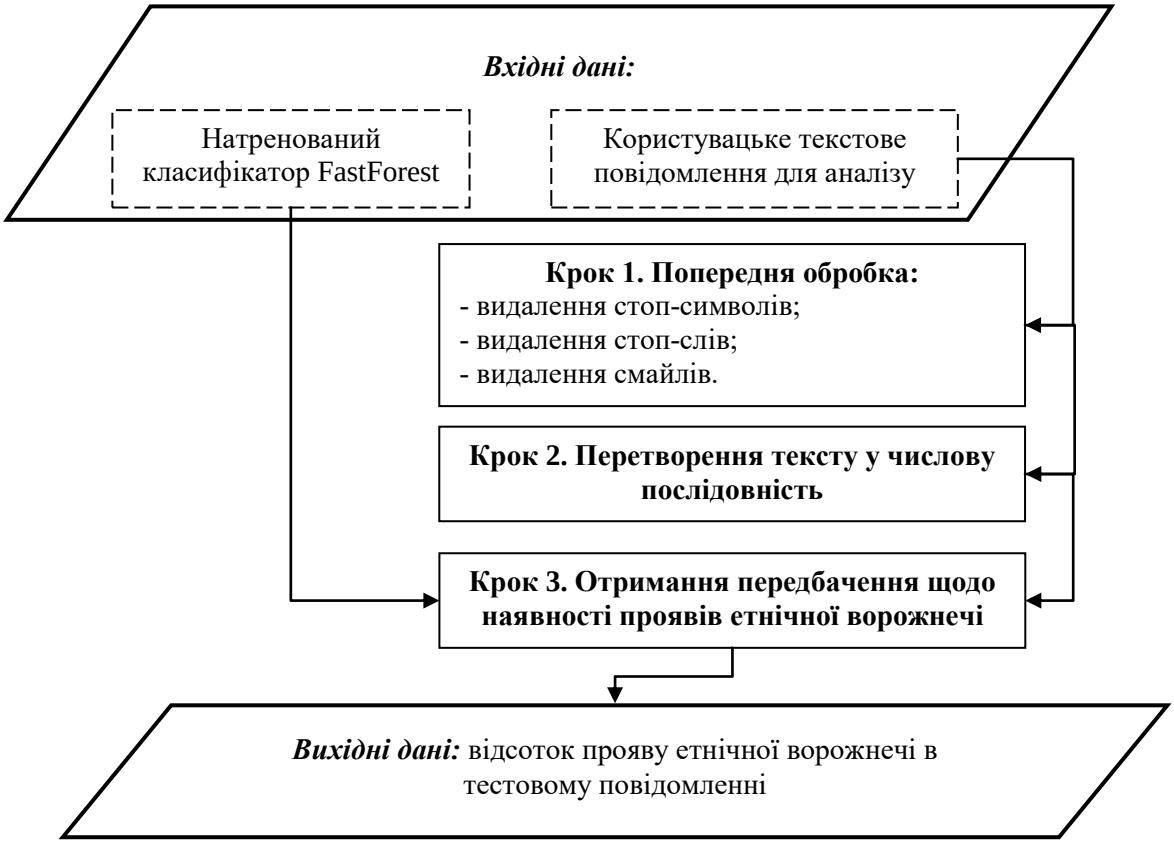


Рисунок 2.1 – Схема та кроки методу виявлення проявів етнічної ворожнечі у текстових повідомленнях соціальних інтернет-мереж

Вхідними даними методу виявлення проявів етнічної ворожнечі у текстових повідомленнях є натренований класифікатор FastForest та користувацьке текстове повідомлення для аналізу.

Робота методу ділиться на 2 основні етапи – попередня обробка користувацького повідомлення та виконання передбачення класифікатором FastForest. Першим кроком є крок попередньої обробки, який включає в себе видалення стоп-символів, видалення стоп-слів та видалення смайлів.

На другому кроці здійснюється перетворення тексту у числову послідовність, яка буде вхідними даними для навченої моделі машинного навчання FastForest.

Наступним кроком є отримання передбачення щодо наявності проявів етнічної ворожнечі, на основі моделі машинного навчання FastForest, в яку завантажено числову послідовність з попереднього кроку.

Вихідними даними методу є відсоток прояву етнічної ворожнечі в тестовому повідомленні.

Отже, було наведено схему та основні кроки методу виявлення проявів етнічної ворожнечі у текстових повідомленнях соціальних інтернет-мереж NLP-засобами, що призначений для автоматизованого аналізу текстів, які публікуються у соціальних мережах з метою виявлення ознак ворожнечі або конфлікту між представниками різних етнічних груп. Завдяки створеному методу досягається ефект спрощення експертизи виявлення проявів етнічної ворожнечі за рахунок автоматизованого її виявлення у текстових повідомленнях соціальних інтернет-мереж.

2.2 Функціональна структура інформаційної системи ідентифікації етнічної ворожнечі за текстовим представленням

Для створення функціональної структури інформаційної системи потрібно виконати етап проєктування майбутніх інтерфейс них форм. Схема

навігації між інтерфейсними формами інформаційної системи виявлення проявів етнічної ворожнечі наведено на ристунку 2.2.

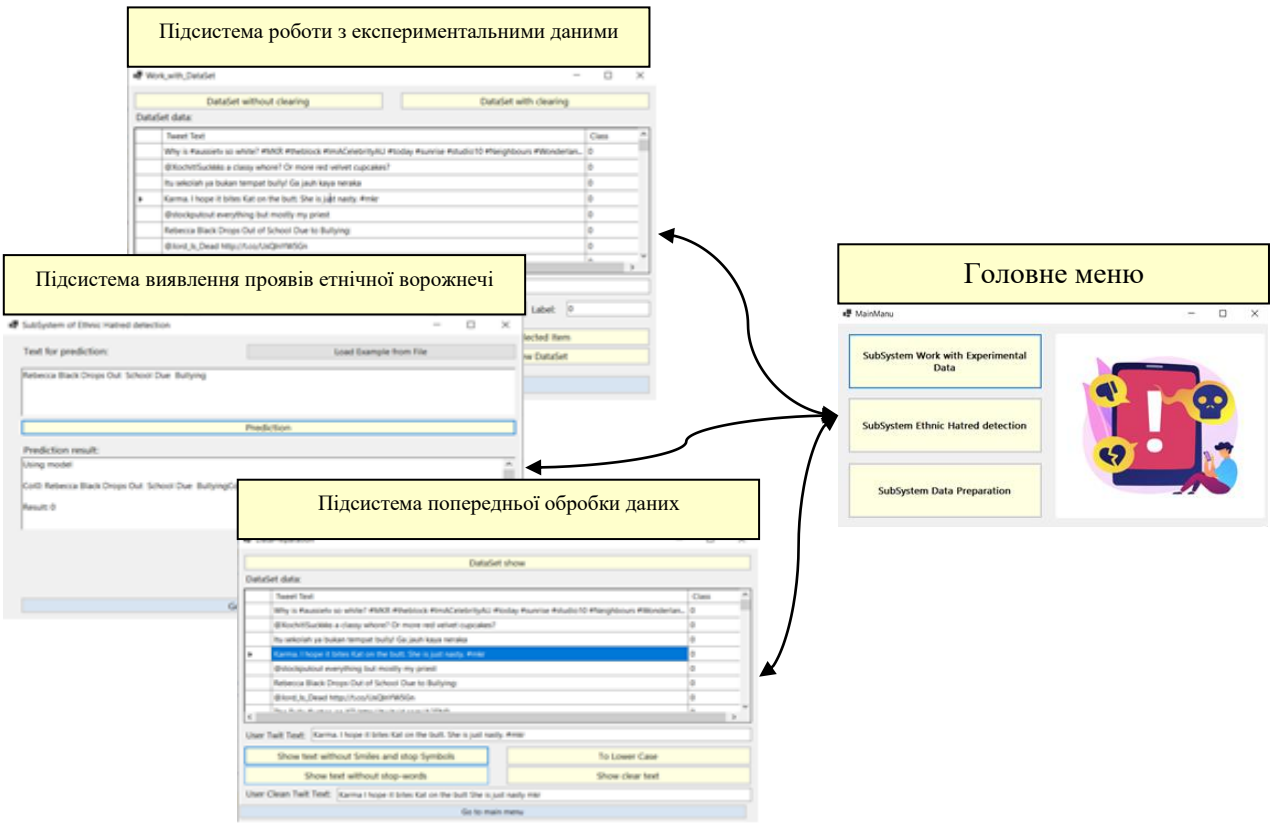


Рисунок 2.2 – Схема навігації між інтерфейсними формами інформаційної системи ідентифікації етнічної ворожнечі за текстовим представленням

Інформаційна структура системи буде мати у складі 4 інтерфейсні форми: «Головне меню», «Підсистема роботи з експериментальними даними», «Підсистема виявлення проявів етнічної ворожнечі», «Підсистема попередньої обробки даних».

Взаємодія між інтерфейсними формами здійснюється через головне меню, яке має тригери переходів по натисненні на відповідні кнопки з назвами підсистем.

Кожна підсистема покликана виконувати ряд функцій, покладених на неї. Підсистема роботи з експериментальними даними призначена для виконання таких функцій:

– завантаження неочищеного набору даних (зі стоп-словами, смайлами тощо) та відображення його в таблицю;

– завантаження очищеного набору даних (без стоп-слів, смайлів тощо) та відображення його в таблицю;

– деталізація обраного текстового повідомлення з таблиці;

– зміна деталізованого текстового повідомлення;

– додавання нового текстового повідомлення;

– видалення обраного текстового повідомлення;

– збереження змін в поточному наборі даних;

– збереження змін, як нового набору даних.

Підсистема виявлення проявів етнічної ворожнечі є головною підсистемою, і призначена для автоматизованого аналізу текстів, що публікуються у соціальних мережах з метою виявлення ознак ворожнечі або конфлікту між представниками різних етнічних груп. На дану підсистему покладені такі основні функції:

– завантаження тестового тексту з файлу;

– написання тестового тексту у текстове поле;

– виявлення відсоток прояву етнічної ворожнечі в тестовому повідомленні та виведення аналізу користувачу;

– виведення статистики по метрикам використаної моделі машинного навчання.

Підсистема попередньої обробки даних призначена для маніпуляцій з даними, які підготують їх для подачі у натренований класифікатор FastForest. Підсистема виконує такі функції:

– виведення неочищеного датасету у таблицю;

– деталізація обраного твіта;

– виведення тексту без смайлів та стоп-символів;

– виведення тексту без стоп-слів;

– виведення тексту у нижньому реєстрі;

– виведення тексту із застосуванням усіх фільтрів одночасно.

Узагальнено можливості всієї інформаційної системи для користувача наведені на рисунку 2.3 у вигляді діаграми варіантів використання.

Рисунок 2.3 – Use-Case діаграма інформаційної системи ідентифікації етнічної ворожнечі за текстовим представленням

Отже, було наведено функціональну структуру інформаційної системи ідентифікації етнічної ворожнечі за текстовим представленням, та виконано етап проєктування майбутніх інтерфейсних форм. Наведена схема навігації між інтерфейсними формами, наведені основні групи функцій підсистем.

2.3 Проєктування пайплайну моделі машинного навчання FastForest

Метод виявлення проявів етнічної ворожнечі у текстових повідомленнях соціальних інтернет-мереж використовує ансамблевий алгоритм машинного навчання FastForest. Для його застосування необхідно побудувати відповідний пайплайн, що є послідовністю кроків обробки даних та моделювання, які

виконуються для побудови моделі машинного навчання. Пайплайн складається з різних етапів, наведених на рисунку 2.4.

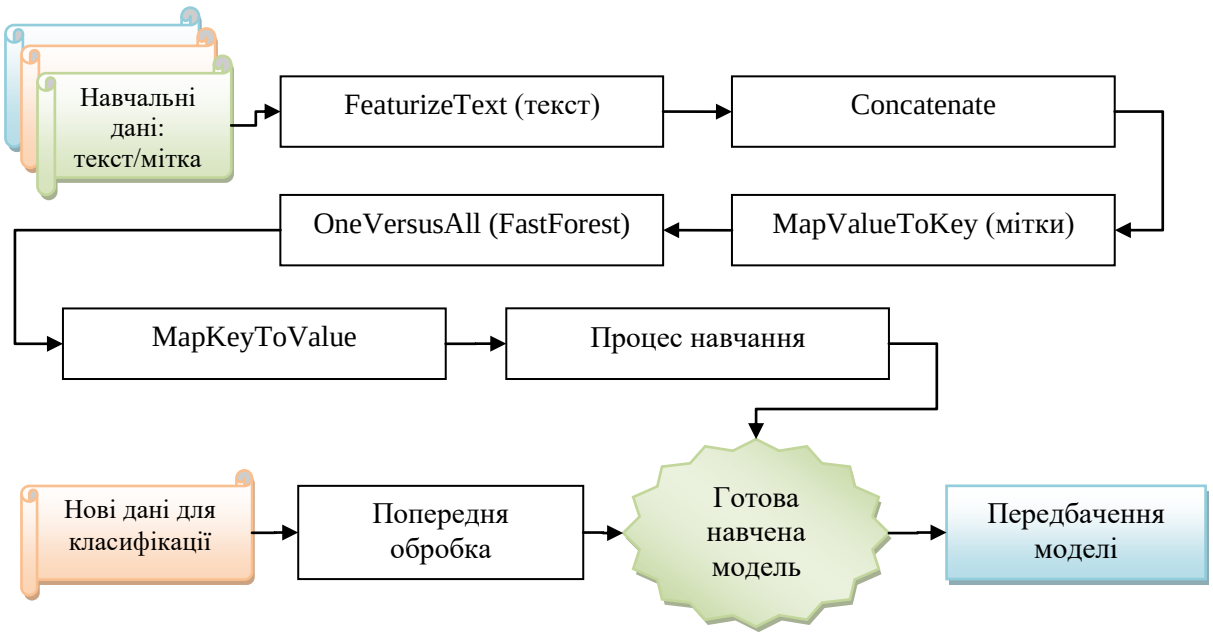


Рисунок 2.4 – Побудова пайплайну FastForest

Кожен блок в схемі рисунку 2.4 представляє один крок пайплайна. Блок «FeaturizeText» відповідає за перетворення текстових даних з колонки «текст» у числові ознаки, які можна використовувати для машинного навчання.

Блок «Concatenate» об’єднує ознаки в один вектор ознак під назвою «Features», шляхом об'єднання даних з двох стовпців в один (текст/мітка). Це корисно для об'єднання характеристик, вилучених з текстових даних на попередньому кроці.

Блок «MapValueToKey» перетворює значення цільової змінної «Мітка» у числові ключі.

Блок «OneVersusAll» застосовує алгоритм бінарної класифікації FastForest до кожного класу проти всіх інших, використовуючи параметри, такі як кількість дерев і листя. Для даного пайплайну кількість дерев рішень, що використовуються в алгоритмі буде взято 4. Це дозволить збільшити точність моделі за рахунок комбінування прогнозів з декількох дерев, але при тому зменшити перенавчання, оскільки кожне дерево навчається на підмножині

даних. Максимальну кількість листків, дозволених в дереві буде встановлено в 4, що зменшує складність моделі, роблячи її більш прозорою.

Блок «MapKeyToValue» перетворює прогнозовані числові ключі назад у вихідні мітки.

Блок «Процес навчання» відповідає за тренування даного пайплайну на наборі даних для навчання. Після завершення тренування зберігається готова натренована модель, яку можна використовувати для визначення проявів етнічної ворожнечі у текстових повідомленнях соціальних інтернет-мереж, яких немає у навчальних даних. Ці дані після проходження попередньої обробки тексту подаються як вхідні дані навченій моделі, яка надасть передбачення, стосовно наявності у тестовому контенті визначення проявів етнічної ворожнечі.

Отже, таким чином наведено пайплайн для ансамблевого алгоритму машинного навчання FastForest, який є складовою методу виявлення проявів етнічної ворожнечі у текстових повідомленнях соціальних інтернет-мереж.

2.4 Підготовка робочих вхідних даних для методу виявлення проявів етнічної ворожнечі

Для навчання класифікатора FastForest був використаний набір даних «Cyberbullying Classification» [29]. Даний датасет в загальному містить понад 47000 твітів, розмічених за класами видів кібербулінгу. Датасет налічує такі категорії кіберзалякувань, як:

- Age;
- Ethnicity;
- Gender;
- Religion;
- Other type of cyberbullying;
- Not cyberbullying.

Набір даних є англомовним, приклад завантажених даних датасету, відкритих у MS Excell наведено на рисунку 2.5.

Оскільки в рамках роботи є потреба виявлення проявів етнічної ворожнечі, для використання цього набору даних його буде дещо видозмінено. В рамках дослідження буде використано 2 категорії: «Ethnicity» та «Not cyberbullying».

A1 tweet_text,cyberbullying_type

	A
40861	11 в™Ў hii!! first of all, ure so gorgeous omggg i hope no one tell u otherwise!! i luv all of ur love memes hehe and also ur layout is amazing!! i love how classy u are when it comes to
40862	the niggers i fuck with aint no dumb niggers. http://www.youtube.com/user/BIGBKELtheGOD#p/u/0/F71wgPUCMQ4,ethnicity
40863	!!!вЂњ@PURDYdidit: Lol he must b smarter than romney..cuz tha dumb ass CRACKER a nobody lik u now..RT @tayyoung_: FUCK OBAMA, dumb ass niggerвЂќ,ethnicity
40864	@HuffingtonPost being black and being bullied in America is nothing new to the majority of black people #dontfetit,ethnicity
40865	RACISM is when u select ur white cloth to wash first before the colored n dyed ones. Say NO to racism... Wash them all 2gether!,ethnicity
40866	THATS RACIST DUMB FUCK "@goshawty123: black people who still think sayin nigger is racist , get a fucken life .",ethnicity
40867	@HeyIts_Jayla @ryan_temple goodone dumb nigger ur ugly as fuck so stfu #BJWON #getoverit,ethnicity
40868	U know the backlash I get for being darkskin with waffle colored children? Inner racism is worse than others!,ethnicity
40869	Just saw a skinhead call a black man in the street "a fat ugly nigger". Disgusting behaviour.,ethnicity
40870	I been done through enough bullshit so no one can tell me shit! I been see dumb ass niggers and dumb ass bitches up in this shit so fuck!,ethnicity
40871	we are talking about how institutional racism affects health care in bio therefore id like to aknowledge that my BIO TEACHER used the term вЂњcoloredвЂќ fuck u nancy,ethnicity
40872	@XXLtina: thank god Romney is winning! the blacks must be eating watermelon frantically! #ew #gross #getajob #niggers ignorant dumb fuck!,ethnicity
40873	Talk about your People you sorry ass dumb-ass Nigger bitch. Leave My Peo0le the fuck alone! You dumb black Fuck!,ethnicity
40874	WOW lol "@CowboyQuan: fuck education. imma dumb ignoarant nigger",ethnicity
40875	@tayyoung_: FUCK OBAMA, dumb ass nigger that "nigger" won tho...So suck a nigger dick!,ethnicity

Рисунок 2.5 – Приклад даних датасету категорії «Ethnicity»

Розмір навчальної вибірки буде становити 7961 твіт категорії «Ethnicity» та 7945 твітів категорії «Not cyberbullying». Розподіл твітів по категоріям за довжиною наведено на рисунку 2.6.

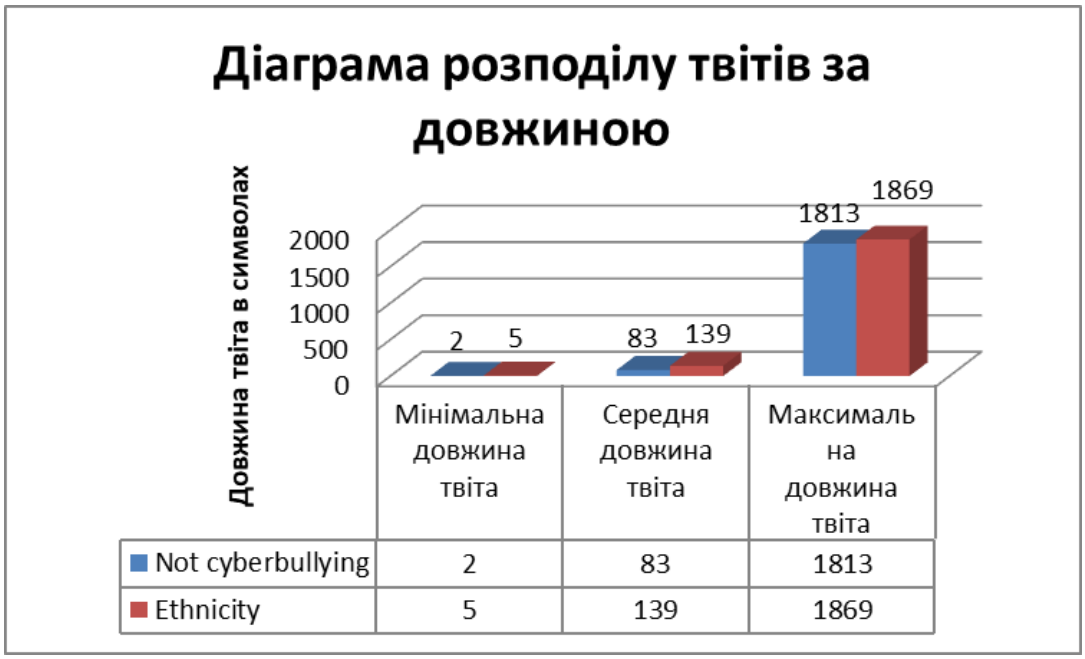


Рисунок 2.6 – Розподіл твітів по категоріям за довжиною

Як видно з рисунку 2.6, дані збалансовані і за кількістю і за довжиною. В подальшому перед початком навчання ще необхідно буде дослідити кількість твітів з мінімальним значенням та з максимальним більш детально з використанням програмних засобів, оскільки за визначенням, твітом називається повідомлення у твіттері довжиною до 280 символів [30]. Тобто можна висунути припущення, що там або декілька твітів зклеєні в купу, або датасет доповнено іншими даними, які не є твітами.

Отже, було виконано підготовку робочих вхідних даних для навчання класифікатора FastForest, що призначений для виявлення проявів етнічної ворожнечі. У якості датасету обрано «Cyberbullying Classification», з якого взято дві категорії: «Ethnicity» та «Not cyberbullying». Розмір навчальної вибірки становить 7961 твіт категорії «Ethnicity» та 7945 твітів категорії «Not cyberbullying», що разом складає 15906 твітів. Варто зауважити, що набір ще потребує подальшої програмної обробки та проведення аналізу.

2.5 Проектна архітектура інформаційної системи ідентифікації етнічної ворожнечі та взаємозв'язок компонентів

Проектна архітектура інформаційної системи ідентифікації етнічної ворожнечі за текстовим представленням складається із 3-х основних підсистем: «Підсистеми роботи з експериментальними даними», «Підсистеми виявлення проявів етнічної ворожнечі», «Підсистеми попередньої обробки даних»; 2-х допоміжних підсистем: «Підсистема формування навчальної вибірки з датасету» та «Підсистеми навчання моделі машинного навчання», а також загального датасету та робочого набору даних (рисунок 2.7).

Підсистема формування навчальної вибірки з датасету для інформаційної системи ідентифікації етнічної ворожнечі за текстовим представленням є допоміжною підсистемою, призначеною для створення робочого набору даних з загального датасету шляхом фільтрації даних. Продуктом роботи даної

підсистеми є сформований набір робочих даних, який буде використаний для навчання класифікатора та роботи інших підсистем.

Підсистема навчання моделі машинного навчання є допоміжною підсистемою, яка відповідає за вибір параметрів навчання, навчання моделі машинного навчання та збереження навченої моделі. Її продуктом є навчена модель машинного навчання FastForest, яка спроможна виконувати виявлення проявів етнічної ворожнечі у текстових повідомленнях.

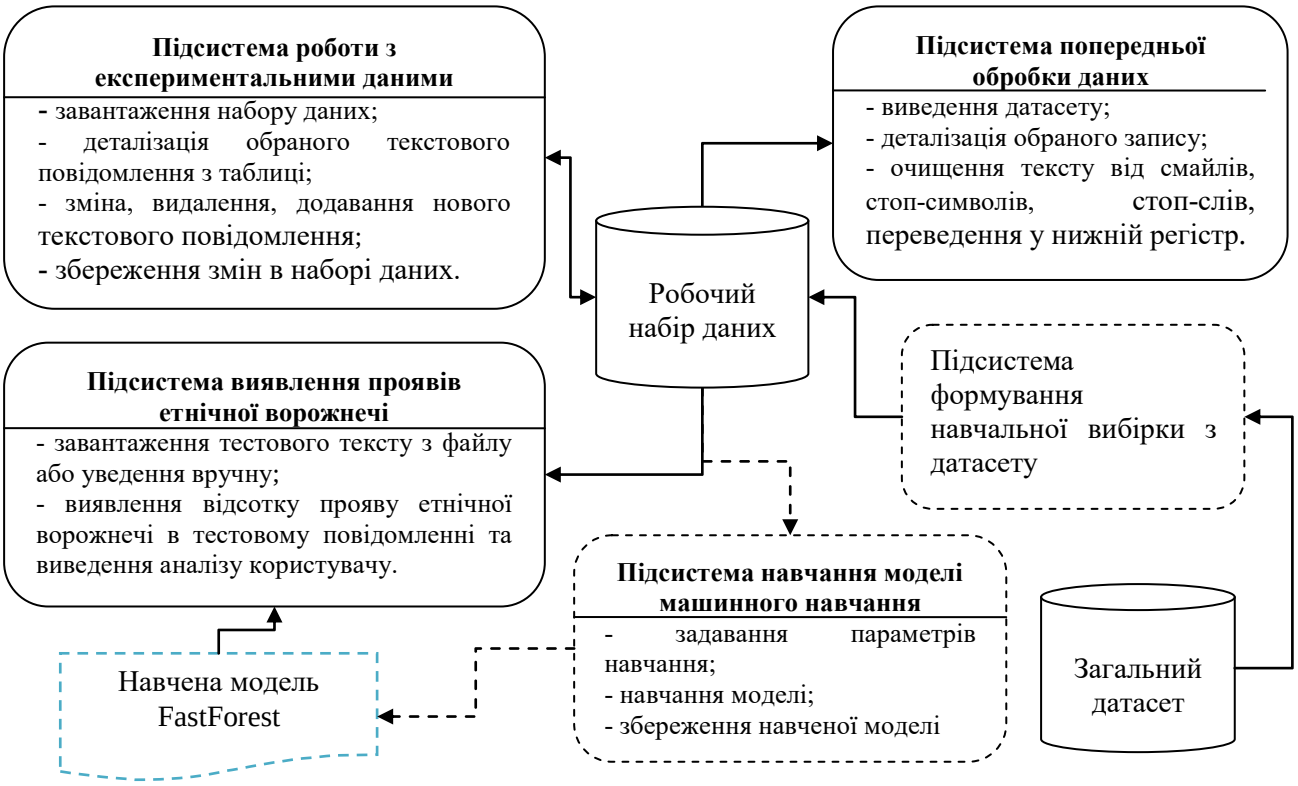


Рисунок 2.7 – Проектна архітектура інформаційної системи ідентифікації етнічної ворожнечі за текстовим представленням

Підсистема роботи з експериментальними даними є однією із основних підсистем інформаційної системи, та призначена для завантаження набору даних, деталізації обраного текстового повідомлення з таблиці, зміни, видалення, додавання нового текстового повідомлення, а також збереження змін в робочому наборі даних. Головна мета підсистеми – взаємодія з робочим набором даних.

Підсистема попередньої обробки даних є також однією із основних підсистем інформаційної системи для виявлення проявів етнічної ворожнечі і

призначена для вирішення таких завдань, як: виведення датасету, деталізація обраного запису, очищення тексту від смайлів, стоп-символів, стоп-слів, переведення у нижній регістр. У якості джерела даних використовує робочий набір даних.

Підсистема виявлення проявів етнічної ворожнечі є головною складовою інформаційної системи виявлення проявів етнічної ворожнечі, та призначена для безпосереднього виявлення проявів етнічної ворожнечі шляхом завантаження тестового тексту з файлу або уведення вручну та виявлення відсотку прояву етнічної ворожнечі в тестовому повідомленні за допомогою попередньо навченого класифікатора FastForest.

Отже, таким чином було наведено проектна архітектуру інформаційної системи для виявлення проявів етнічної ворожнечі, що складається із 3-х основних підсистем: «Підсистеми роботи з експериментальними даними», «Підсистеми виявлення проявів етнічної ворожнечі», «Підсистеми попередньої обробки даних»; 2-х допоміжних підсистем: «Підсистема формування навчальної вибірки з датасету» та «Підсистеми навчання моделі машинного навчання», а також загального датасету та робочого набору даних, що є продуктом роботи підсистеми формування навчальної вибірки з датасету.

2.6 Особливості використання спеціалізованих програмних компонентів

ML.NET є відкритою бібліотекою машинного навчання, розробленою компанією Microsoft. Вона забезпечує можливість реалізації алгоритмів машинного навчання та обробки даних у застосунках, що працюють на платформі .NET. Однією з головних переваг ML.NET є те, що вона інтегрована безпосередньо з екосистемою .NET, що робить її досить зручною для розробників, які вже працюють з цією платформою. Завдяки цьому, розробники можуть легко інтегрувати моделі машинного навчання в свої програми або вебсервіси без зайвих зусиль [31].

Бібліотека ML.NET підтримує широкий спектр завдань машинного навчання, включаючи класифікацію, регресію, кластеризацію, обробку тексту та рекомендації. Вона надає різноманітні алгоритми, такі як дерева рішень, лінійна регресія, метод опорних векторів (SVM), нейронні мережі та багато інших.

Однією зі суттєвих переваг ML.NET є його здатність до роботи з наборами даних різної складності та обсягу. Від невеликих наборів даних, що вміщуються в пам'ять, до великих обсягів даних, які можуть зберігатися на диску або опрацьовуватися паралельно. Це робить ML.NET привабливим вибором для різноманітних сценаріїв, від прототипування до впровадження великомасштабних систем.

Крім того, ML.NET має активну спільноту та добре документовану базу знань, що сприяє зручності використання. Регулярні оновлення та підтримка зі сторони Microsoft гарантують, що бібліотека залишається актуальною та здатною задовольнити потреби розробників у сфері машинного навчання, працюючих на платформі .NET.

Враховуючи описані можливості ML.NET, вона буде використана для навчання та роботи з моделлю машинного навчання FastForest, що буде використана для виявлення проявів етнічної ворожнечі у текстових повідомленнях соціальних інтернет-мереж.

Бібліотека System.Text.RegularExpressions в .NET надає потужний інструментарій для роботи з регулярними виразами в мові програмування C#. Ця бібліотека дозволяє виконувати пошук, заміну та аналіз тексту на основі заданих шаблонів. Вона включає в себе клас Regex, який є основним інструментом для роботи з регулярними виразами. За допомогою класу Regex можна визначити шаблон, за яким буде проводитися пошук, і виконати різноманітні операції з введеним текстом, такі як пошук підходящих фрагментів, вилучення збігів або заміна їх на інші значення [32].

Використання бібліотеки System.Text.RegularExpressions дозволяє вирішувати різноманітні завдання, пов'язані з обробкою текстових даних. Наприклад, вона дозволяє здійснювати перевірку валідності введених даних за

допомогою регулярних виразів, які визначають необхідні формати даних. Крім того, з її допомогою можна виконувати складні операції з рядками, такі як знаходження та вилучення певних фрагментів тексту, підрахунок кількості входжень певного шаблону або аналіз синтаксичних структур.

Ця бібліотека входить до стандартного складу .NET Framework і .NET Core, тому вона доступна для використання в будь-якому проекті на платформі .NET. Вона має широкий спектр функцій і можливостей, що робить її незамінним інструментом для роботи з текстовими даними у багатьох програмних проектах. Бібліотека System.Text.RegularExpressions також надає можливість оптимізації роботи з регулярними виразами шляхом компіляції їх в пам'ять, що дозволяє підвищити швидкодію обробки тексту в програмі.

Вищеописана бібліотека буде використана для попередньої обробки тексту для видалення стоп-символів та стоп-слів.

Отже, в рамках роботи для інформаційної системи виявлення проявів етнічної ворожнечі буде використано бібліотеку ML.NET, що буде використана для навчання та роботи з моделлю машинного навчання FastForest, та бібліотеку System.Text.RegularExpressions для попередньої обробки тексту для видалення стоп-символів та стоп-слів.

2.7 Висновки до розділу 2

Виконуючи другий розділ кваліфікаційної роботи бакалавра, було створено метод виявлення проявів етнічної ворожнечі у текстових повідомленнях соціальних інтернет-мереж NLP-засобами, який своїм призначенням має автоматизований аналіз текстів, що публікуються у соціальних мережах з метою виявлення ознак ворожнечі або конфлікту між представниками різних етнічних груп. Метод використовує техніки обробки природної мови, а саме підхід на основі ансамблів, і здійснює перетворення вхідних даних у вигляді навченого класифікатора FastForest та вхідного

текстового повідомлення у вихідні дані у вигляді відсотка прояву етнічної ворожнечі у тестовому повідомленні соціальних інтернет-мереж.

Наведено функціональну структуру інформаційної системи, та виконано етап проєктування майбутніх інтерфейсних форм, описані основні функції для інтерфейс них форм.

Сформовано пайплайн для ансамблевого алгоритму машинного навчання FastForest, який є послідовністю кроків обробки даних та моделювання, які виконуються для побудови моделі машинного навчання. Модель FastForest є складовою методу виявлення проявів етнічної ворожнечі у текстових повідомленнях соціальних інтернет-мереж.

Було виконано підготовку робочих вхідних даних для навчання класифікатора FastForest, у якості датасету обрано «Cyberbullying Classification», з якого взято дві категорії: «Ethnicity» та «Not cyberbullying». Розмір навчальної вибірки становить 7961 твіт категорії «Ethnicity» та 7945 твітів категорії «Not cyberbullying», що разом складає 15906 твітів.

Наведено проектна архітектуру інформаційної системи для виявлення проявів етнічної ворожнечі, що складається із 3-х основних підсистем: «Підсистеми роботи з експериментальними даними», «Підсистеми виявлення проявів етнічної ворожнечі», «Підсистеми попередньої обробки даних»; 2-х допоміжних підсистем: «Підсистема формування навчальної вибірки з датасету» та «Підсистеми навчання моделі машинного навчання», а також загального датасету та робочого набору даних, що є продуктом роботи підсистеми формування навчальної вибірки з датасету.

Для інформаційної системи ідентифікації етнічної ворожнечі за текстовим представленням було обрано спеціалізовані програмні компоненти. Буде використано бібліотеку ML.NET, що буде використана для навчання та роботи з моделлю машинного навчання FastForest, та бібліотеку System.Text.RegularExpressions для попередньої обробки тексту для видалення стоп-символів та стоп-слів.

Розроблений метод дозволяє досягти спрощення експертизи виявлення проявів етнічної ворожнечі за рахунок автоматизованого її виявлення у текстових повідомленнях соціальних інтернет-мереж.

За відповідними спроектованими складовими необхідно розробити застосунок, за допомогою якого провести дослідження ефективності методу виявлення проявів етнічної ворожнечі у текстових повідомленнях соціальних інтернет-мереж NLP-засобами. Для доведення коректності результатів його треба окремо функціонально дослідити й протестувати.

Розділ 3 Експериментальне дослідження методу виявлення проявів етнічної ворожнечі у текстових повідомленнях соціальних інтернет-мереж

3.1 Визначення шляхів дослідження та засобів створення інформаційної системи ідентифікації етнічної ворожнечі

В рамках дослідження ефективності методу виявлення проявів етнічної ворожнечі у текстових повідомленнях соціальних інтернет-мереж засобами NLP, що призначений для автоматизованого аналізу текстів, що публікуються у соціальних мережах з метою виявлення ознак ворожнечі або конфлікту між представниками різних етнічних груп необхідно створити програмну систему у формі віконного застосунку, що буде виконувати такі функції:

– завантаження неочищеного набору даних (зі стоп-словами, смайлами тощо) та відображення його в таблицю;

– завантаження очищеного набору даних (без стоп-слів, смайлів тощо) та відображення його в таблицю;

– деталізація обраного текстового повідомлення з таблиці;

– зміна деталізованого текстового повідомлення;

– виведення тексту без смайлів та стоп-символів;

– виведення тексту повідомлення без стоп-слів;

– виведення тексту повідомлення у нижньому реєстрі;

– виведення тексту повідомлення із застосуванням усіх фільтрів одночасно;

– додавання нового текстового повідомлення в робочий набір даних;

– видалення обраного текстового повідомлення з робочого набору даних;

– збереження змін в поточному робочому наборі даних;

– збереження змін, як нового робочого набору даних;

– виявлення відсотку прояву етнічної ворожнечі в тестовому повідомленні та виведення аналізу користувачу;

– виведення статистики по метрикам використаної моделі машинного навчання.

Зважаючи на широкий фукціонал інформаційної системи ідентифікації етнічної ворожнечі за текстовим представленням, її працездатність необхідно перевірити засобами тест-кесів. Щодо дослідження ефективності методу, його планується виконати з використанням розробленого програмного забезпечення, шляхом порівняння отриманих відповідей з валідаційним набором, а також необхідно виконати оцінку навченої моделі машинного навчання FastForest, за допомогою використання метрик MicroAccuracy, MacroAccuracy, LogLoss, ConfusionMatrix, f1-міри та Recall.

3.2 Вибір засобів розробки інформаційної системи ідентифікації етнічної ворожнечі за текстовим представленням

Для написання програмної реалізації застосунка для дослідження ефективності методу виявлення проявів етнічної ворожнечі у текстових повідомленнях соціальних інтернет-мереж буде використано платформу .NET 7.0, середовище програмування Visual Studio, та мову програмування C#.

Платформа .NET 7.0 є однією з найновіших версій платформи розробки програмного забезпечення .NET від Microsoft. Ця версія представляє собою значний крок уперед у розвитку екосистеми .NET, принісши численні покращення та нові функції для розробників. Одним із ключових напрямків розвитку .NET 7.0 є підвищення продуктивності та оптимізація швидкодії, що робить розробку програм більш ефективною.

Visual Studio є інтегрованим середовище розробки, що розроблене компанією Microsoft, яке надає засоби для створення програмного забезпечення для різних платформ, включаючи Windows, Linux та мобільні пристрої. Це одне з найпопулярніших інтегрованих середовищ розробки серед професійних розробників програмного забезпечення, які працюють у різних сферах індустрії. IDE Visual Studio має широкий набір інструментів, які допомагають

автоматизувати створення графічних інтерфейсів користувача та реалізовувати машинне навчання.

Мова програмування C# є однією з найбільш популярних мов у розробці програмного забезпечення, особливо в середовищі Visual Studio. У контексті навчання моделей штучного інтелекту, ця комбінація є дуже потужною, оскільки Visual Studio надає широкий набір інструментів та можливостей для створення, навчання та експериментів з моделями машинного навчання на основі мови C#.

Отже, у якості засобів розробки інформаційної системи ідентифікації етнічної ворожнечі за текстовим представленням буде використано платформу .NET 7.0, середовище програмування Visual Studio, та мову програмування C#.

3.3 Структура та функціональне призначення програмних складових інформаційної системи ідентифікації етнічної ворожнечі

Для забезпечення функціональних можливостей інформаційної системи ідентифікації етнічної ворожнечі за текстовим представленням, яка складається із 3-х основних підсистем: «Підсистеми роботи з експериментальними даними», «Підсистеми виявлення проявів етнічної ворожнечі», «Підсистеми попередньої обробки даних»; 2-х допоміжних підсистем: «Підсистема формування навчальної вибірки з датасету» та «Підсистеми навчання моделі машинного навчання», було спроєктовано відповідну діаграму класів (рисунок 3.1).

Підсистеми навчання моделі машинного навчання реалізована відповідним класом «BoyarchukModel1», що має методи для побудови пайплайну, навчання нейромережі та збереження результату. Результатом підсистеми є навчена модель машинного навчання FastForest.

Клас «FormatedDataSet» призначений для реалізації допоміжної підсистеми формування навчальної вибірки з датасету та виведення статистики по наявним категоріям в датасеті. Результатом роботи підсистеми є сформований робочий набір даних, що буде використовуватись як вхідні дані для інших підсистем.

Підсистеми роботи з експериментальними даними реалізована за допомогою класів «Work_with_DataSet» та «DataSetOperation», який містить методи ReadDataFromDataSet() для читання даних з датасета WriteToDataGrid() для запису даних в таблицю, WriteNewDataToDataSet() для запису нових даних в датасет, метод SaveChanges() призначений для збереження змін після роботи з робочими даними, а RemoveSelectedItem() для видалення обраного запису з робочих даних.

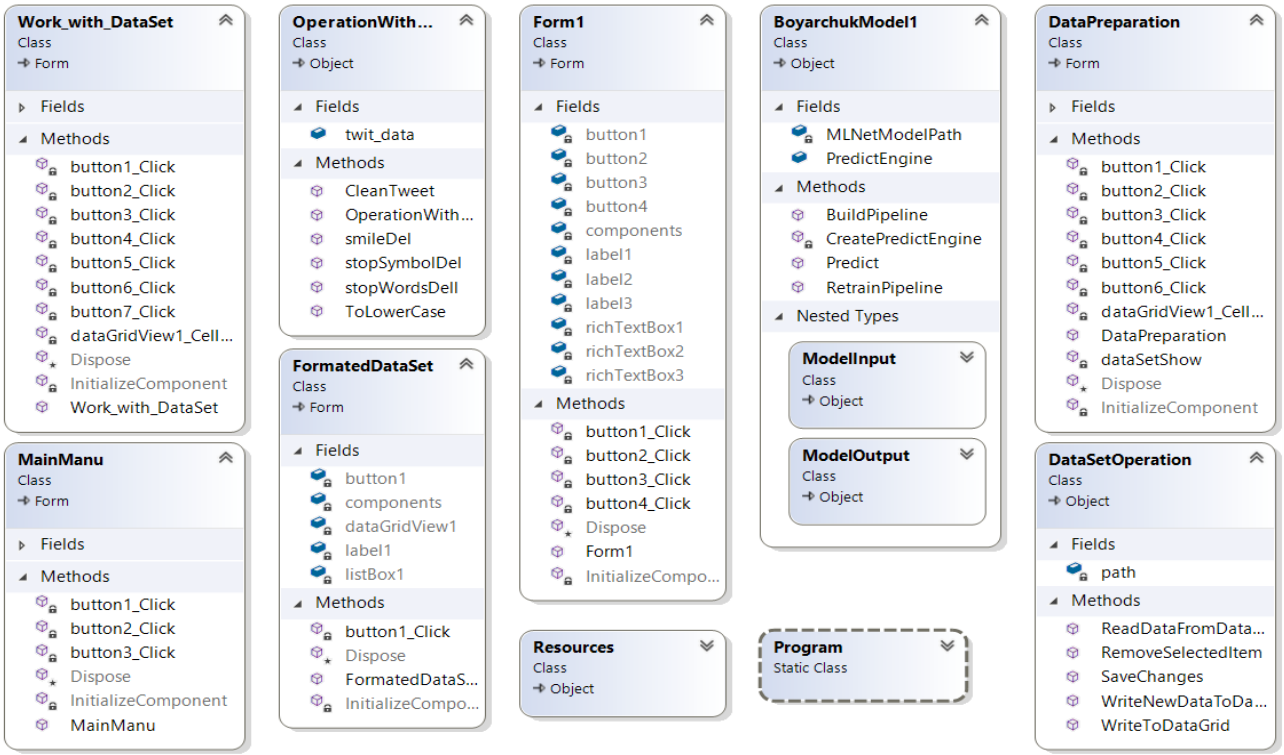


Рисунок 3.1 – Діаграма класів інформаційної системи ідентифікації етнічної ворожнечі за текстовим представленням

Підсистеми попередньої обробки даних реалізована функціоналом класу «DataPreparation» та «OperationWithData». Метод dataSetShow() призначений для відображення робочого набору даних в перелік, в методах-обробниках натиснення на кнопки відбуваються виклики методів класу «OperationWithData». Метод smileDel() призначений для видалення смайлів. Метод stopWordsDell() призначений для видалення стоп-слів. Метод stopSymbolDel() призначений для

видалення стоп-символів. Метод CleanTweet() призначений для очистки тексту одразу від всіх видів інформаційних шумів.

Підсистеми виявлення проявів етнічної ворожнечі реалізована класом «Form1». Основний функціонал підсистеми реалізований в обробниках подій натиснення на кнопки. Вхідними даними використовує навчену модель FastForest, і є головною підсистемою проектованого застосунку.

Отже, таким чином описано структуру та функціональне призначення програмних складових інформаційної системи ідентифікації етнічної ворожнечі за текстовим представленням. Система складається із 12 класів, які виконують весь описаний функціонал.

3.4 Особливості реалізації програмних складових інформаційної системи ідентифікації етнічної ворожнечі

Для реалізації основних підсистем інформаційної системи ідентифікації етнічної ворожнечі за текстовим представленням, необхідно спершу виконати аналіз датасету, та сформувати робочий набір даних для навчання класифікатора та подальшої роботи підсистем. Оскільки обраний набір даних містить 6 категорій, а для коректної роботи необхідно використати тільки 2 з них, було створено відповідний програмний модуль для аналізу та збереження зміненого датасету у форматі, доступному для навчання через бібліотеку ML.NET.

Тому перш за все за обраним шляхом до завантаженого з Kaggle набору даних інформація виводиться в таблицю для перегляду категорій та вмісту. Далі виконується обробка даних з таблиці dataTable, яка містить інформацію про твіти та їх класифікацію за категорією "cyberbullying_type".

Визначається довжини твітів, кожен твіт аналізується, і для кожного запису обчислюється його довжина у символах. Виконується групування за категорією «cyberbullying_type». Виконується розрахунок статистики довжини твітів. Для кожної категорії обчислюються мінімальна, середня та максимальна довжина твітів, та виводяться результати у список listBox, який

використовується для відображення статистики, а потім для кожної категорії виводиться інформація про мінімальну, середню та максимальну довжину твітів у цій категорії. Кожен запис відображається в форматі: «Категорія: Мінімальна довжина - [MinLength], Середня довжина - [AvgLength], Максимальна довжина - [MaxLength]». Останнім етапом здійснюється вибір всіх записів з категоріями «not_cyberbullying» та «ethnicity» з позначками через знач табуляції 0 та 1 відповідно. Дані записуються у текстовий файл з розширенням .txt. Вигляд статистики по записам дата сетів на рисунку 3.2.

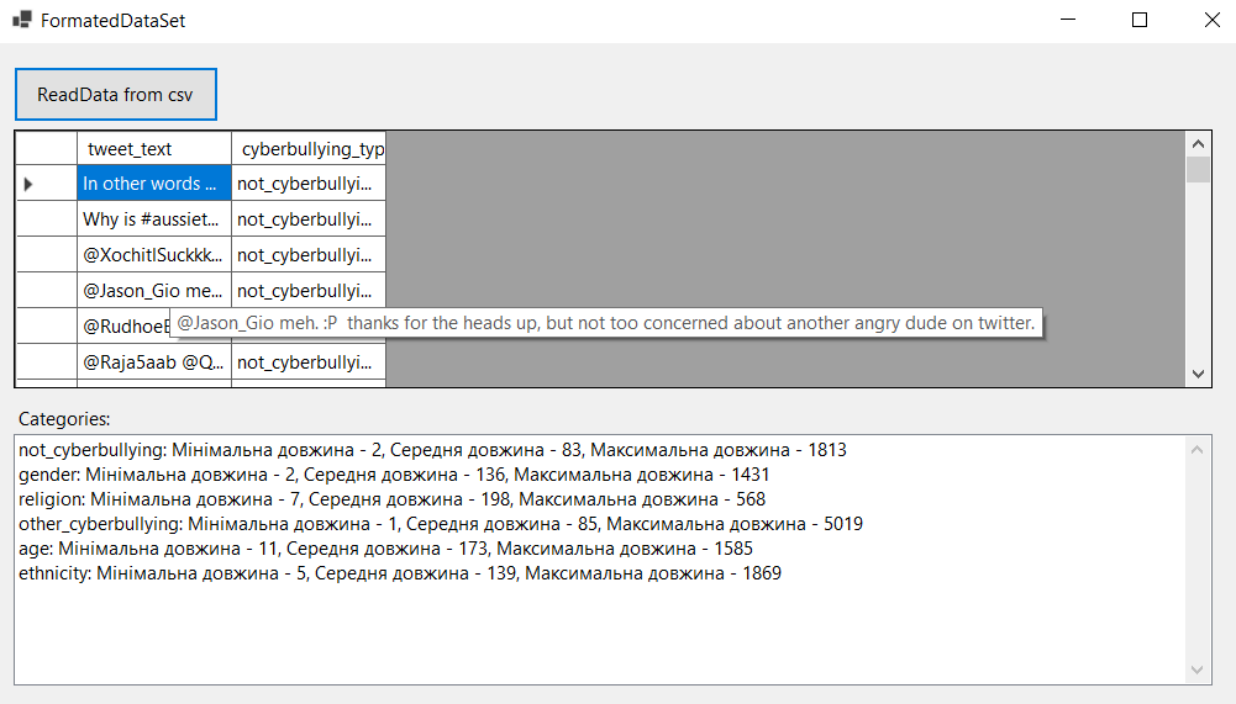


Рисунок 3.2 – Допоміжна підсистема для формування робочого набору даних

Основною підсистемою є підсистема виявлення проявів етнічної ворожнечі, що використовує навчену модель машинного навчання FastForest. Для навчання моделі метод BuildPipeline() приймає об'єкт MLContext як параметр і повертає конвеєр для обробки даних. В конвеєрі визначається послідовність перетворень даних, які будуть застосовані до вхідних даних перед навчанням моделі. Першим кроком в цьому конвеєрі є FeaturizeText(), який перетворює текстові дані у числовий формат. Потім використовується метод Concatenate(), який об'єднує вектори ознак у новий вектор, який

використовується для навчання моделі. Після цього використовується MapValueToKey(), який перетворює значення колонки з назвами класів в числові індекси. Далі модель навчається за допомогою OneVersusAll, який використовує багатокласовий класифікатор на основі алгоритму бінарного класифікатора «FastForest». Цей класифікатор створюється з параметрами, такими як кількість дерев (4), кількість листків (4), вибірка ознак і назви колонок міток та ознак. На останок, використовується MapKeyToValue(), який перетворює числові індекси навченої моделі назад у їхні оригінальні значення. Навчена модель зберігається для подальшого використання (рисунок 3.3).

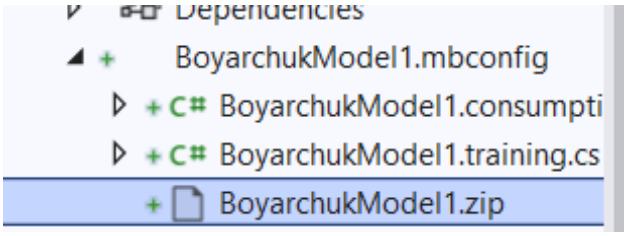


Рисунок 3.3 – Збережена модель «FastForest»

Для визначення наявності проявів етнічної ворожнечі створюється екземпляр sampleData класу «BoyarchukModel1.ModelInput», який представляє вхідні дані для моделі. Значення для аналізу береться з тексту, що міститься у текстовому полі.

Застосовується метод Predict з класу «BoyarchukModel1», щоб зробити прогноз на основі вхідних даних sampleData. Результат прогнозу зберігається у змінній predictionResult. Далі перевіряється значення PredictedLabel у predictionResult. Якщо воно дорівнює 0, то текст не містить проявів етнічної ворожнечі, і відповідний результат додається до текстового поля результату. У протилежному випадку, додається повідомлення про те, що текст містить прояви етнічної ворожнечі. Крім того, виводиться відсоток проявів етнічної ворожнечі з predictionResult.Score[1]. Виконання описаного функціоналу наведено на рисунку 3.4.

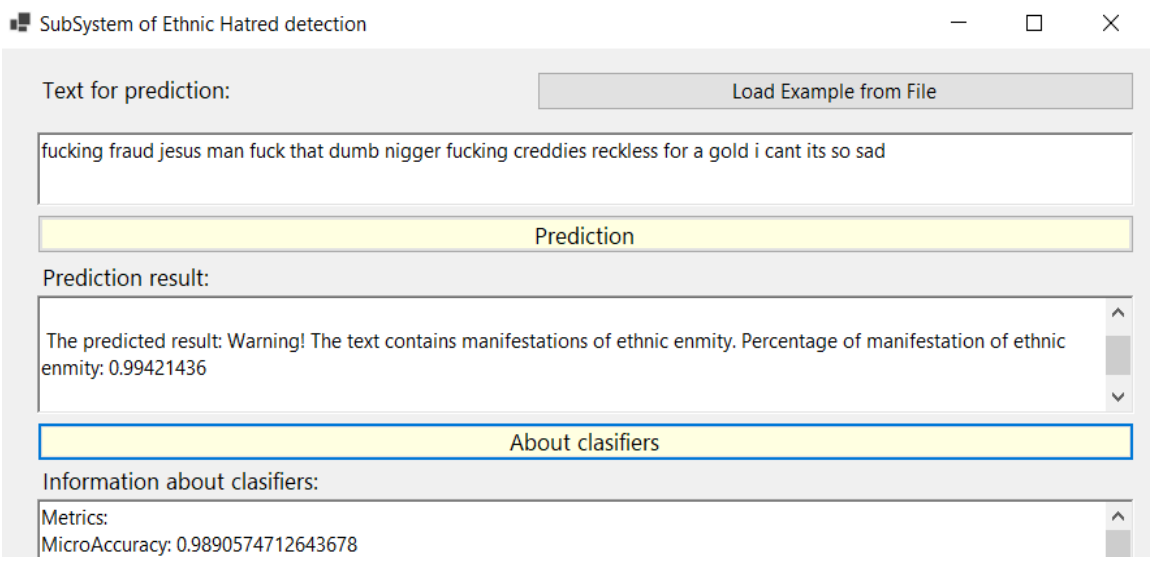


Рисунок 3.4 – Приклад виявлення проявів етнічної ворожнечі

Для розуміння, наскільки добре працює навчена модель FastForest для класифікації даних використовуються метрики для оцінки її ефективності. Для цього завантажуються тестові дані з файлу з використанням методу LoadFromTextFile() класу «MLContext.Data», після чого виконується прогнозування на тестових даних за допомогою навченої моделі. Це виконується за допомогою методу Transform(), який застосовує модель до вхідних даних і повертає прогнозовані значення, на базі яких проводиться оцінка прогнозів і отримання метрик ефективності моделі. Використовується мультикласифікаційна оцінка, що дозволяє отримати різноманітні метрики ефективності, такі як точність, макро-точність, логарифмічна втрата тощо. Для цього використовується метод MulticlassClassification.Evaluate(), де labelColumnName вказує на назву стовпця, що містить мітки класів.

Останнім етапом є виведення отримані метрики ефективності. У цьому випадку виводяться мікро-точність, макро-точність, логарифмічна втрата та матриця сплутування. Форматована матриця сплутування виводиться за допомогою методу GetFormattedConfusionTable(), що повертає читабельне представлення матриці сплутування у текстовому вигляді (рисунок 3.5).

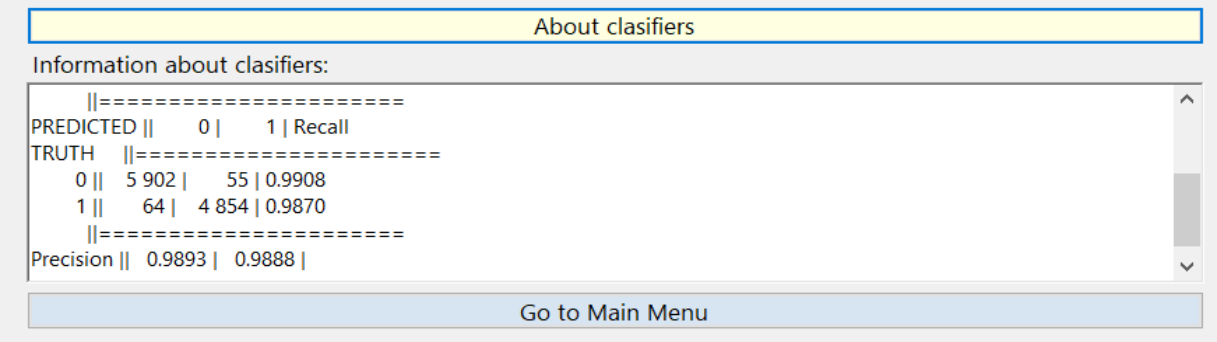


Рисунок 3.5 – Виведення матриці сплутування

Отже, описано особливості реалізації програмних складових інформаційної системи ідентифікації етнічної ворожнечі за текстовим представленням, призначеної для автоматизованого аналізу текстів, що публікуються у соціальних мережах з метою виявлення ознак ворожнечі або конфлікту між представниками різних етнічних груп.

3.5 Тестування інформаційної системи ідентифікації етнічної ворожнечі за текстовим представленням та вимоги до розгортання

Необхідно перевірити, чи може й наскільки ефективно може створена програмна реалізація виконувати задані функції. Відповідно до цього, є потреба дослідити програмну реалізацію засобами текст-кейсів.

Таблиця 3.1 – Тест-кейс 00001

Тест-кейс ID: 00001	**Приорітет:** 1	**Створено:** 15.03.2024, Ілля БОЯРЧУК
Назва: Перевірка переходів на підсистеми застосунку з головного меню		
Кроки		**Очікуваний результат**
1. Відкрити програмний застосунок. 2. Натиснути кнопку «SubSystem Ethnic Harted detection».		Відкрито програмний застосунок Відкрилась підсистема виявлення проявів етнічної ворожнечі
Результат виконання тест-кейсу: перевірку пройдено успішно.		

Першим тестовим випадком буде перевірка переходів на підсистеми з головного меню. Кроки перевірки коректності переходів на підсистеми наведено в таблиці 3.1. Після виконання вказаних в таблиці 3.1 кроків, тест-кейс 00001 вважається успішно пройденим. Результат виконання тест-кейсу наведено на рисунку 3.6.

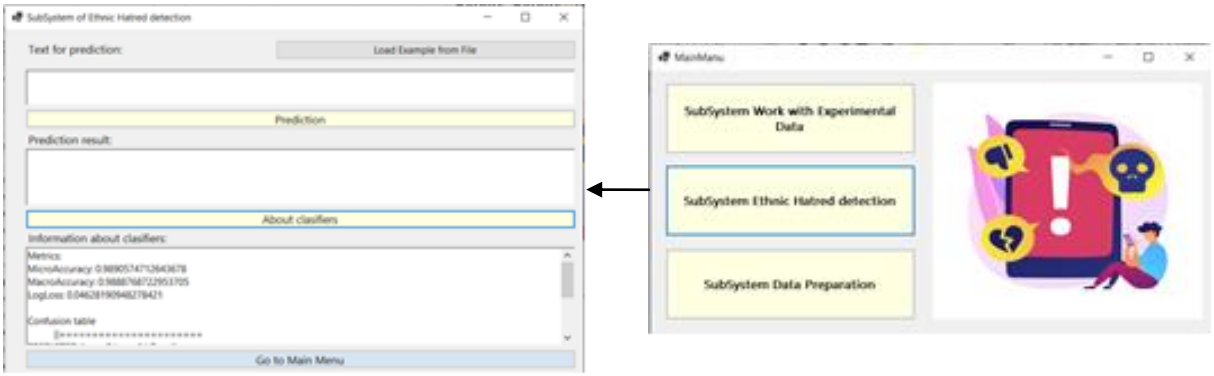


Рисунок 3.6 – Перехід з головного меню на головну підсистему

Наступним тестовим випадком буде перевірка завантаження очищеного набору даних (без стоп-слів, смайлів тощо) та відображення його в таблицю підсистеми попередньої обробки даних. Кроки тест-кейсу наведені у таблиці 3.2.

Таблиця 3.2 – Тест-кейс 00002

Тест-кейс ID: 00002	**Приорітет:** 1	**Створено:** 18.03.2024, Ілля БОЯРЧУК
Назва: Перевірка завантаження очищеного набору даних (без стоп-слів, смайлів тощо) та відображення його в таблицю		
Кроки		**Очікуваний результат**
1. Відкрити програмний застосунок. 2. Натиснути кнопку «SubSystem Work with Experimental Data». 3. Натиснути кнопку «DataSet with Clearing»		Відкрито програмний застосунок (головне меню) Відкрилась підсистема роботи з робочою вибіркою. У таблицю завантажено очищений вміст датасету.
Результат виконання тест-кейсу: перевірку пройдено успішно.		

Після виконання вказаних в таблиці 3.2 кроків, тест-кейс 00002 вважається успішно пройденим. Результат виконання тест-кейсу наведено на рисунку 3.7.

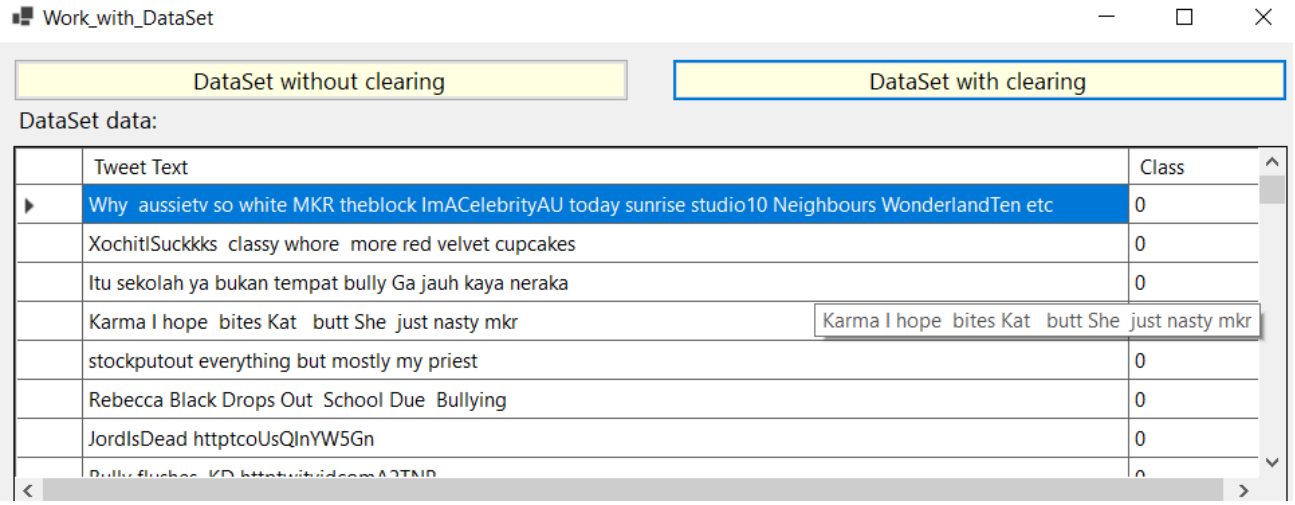


Рисунок 3.7 – Результат виконання тест-кейсу 00002

Таблиця 3.3 – Тест-кейс 00003

Тест-кейс ID: 00003	**Приорітет:** 1	**Створено:** 18.03.2024, Ілля БОЯРЧУК
Назва: Перевірка деталізації обраного текстового повідомлення з таблиці		
Кроки		**Очікуваний результат**
1. Відкрити програмний застосунок. 2. Натиснути кнопку «SubSystem Work with Experimental Data». 3. Натиснути кнопку «DataSet with Clearing». 4. Натиснути мишкою на текстовому повідомленні для деталізації.		Відкрито програмний застосунок (головне меню) Відкрилась підсистема роботи з робочою вибіркою. У таблицю завантажено очищений вміст датасету. У відповідних текстових полях відобразився деталізований зміст.
Результат виконання тест-кейсу: перевірку пройдено успішно.		

Наступним тестовим випадком буде перевірка деталізації обраного текстового повідомлення з таблиці підсистеми попередньої обробки даних. Кроки тест-кейсу наведені у таблиці 3.3.

Результат виконання тест-кейсу наведено на рисунку 3.8.

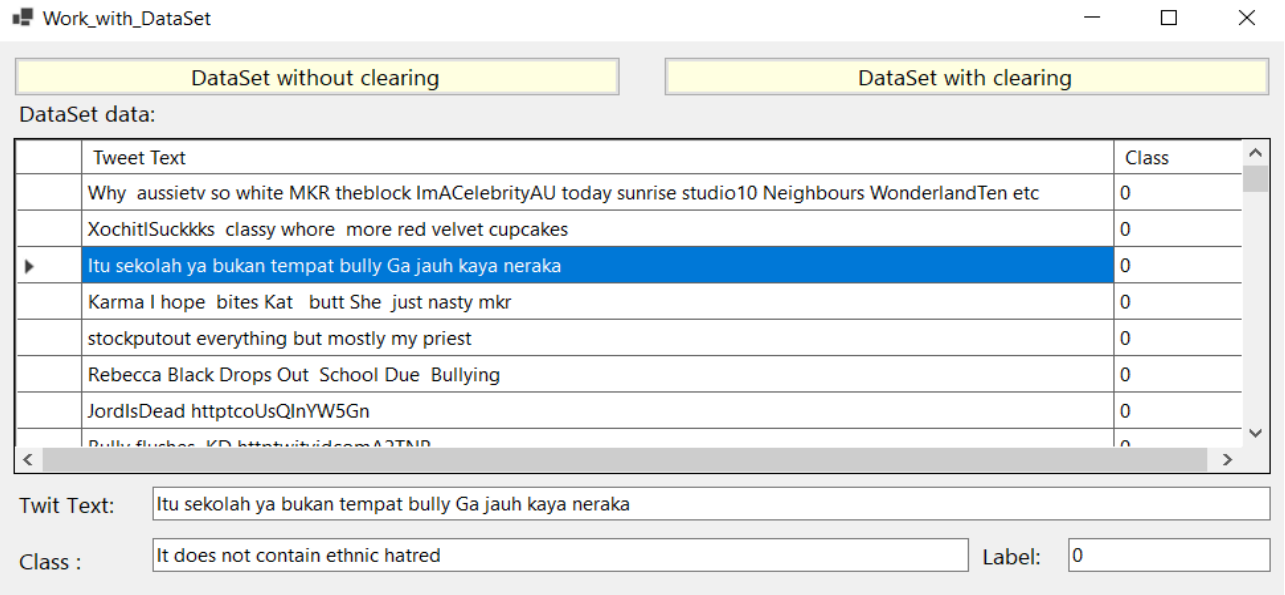


Рисунок 3.8 – Результат виконання тест-кейсу 00003

Після виконання вказаних в таблиці 3.3 кроків, тест-кейс 00003 вважається успішно пройденим. Обране текстове повідомлення з таблиці деталізоване, та виведене у відповідні текстові поля.

Наступним тестовим випадком буде перевірка виявлення відсотку прояву етнічної ворожнечі в тестовому повідомленні та виведення аналізу користувачу підсистеми виявлення проявів етнічної ворожнечі. Кроки тест-кейсу наведені у таблиці 3.4.

Після виконання вказаних в таблиці 3.4 кроків, тест-кейс 00004 вважається успішно пройденим. Виявлення відсотку прояву етнічної ворожнечі в обраному з файлової системи тестовому повідомленні коректно виконано та відображено на екрані. Результат успішного виконання тест-кейсу наведено на рисунку 3.9.

Таблиця 3.4 – Тест-кейс 00004

Тест-кейс ID: 00004	**Приорітет:** 1	**Створено:** 19.03.2024, Ілля БОЯРЧУК
Назва: Перевірка виявлення відсотку прояву етнічної ворожнечі в тестовому повідомленні		
Кроки		**Очікуваний результат**
1. Відкрити програмний застосунок. 2. Натиснути кнопку «SubSystem Ethnic and Hatred Detection 3. Натиснути кнопку «Load Example from File». 4. Обрати «example4.txt». 5. Натиснути кнопку «Prediction»		Відкрито програмний застосунок (головне меню) Відкрилась підсистема виявлення проявів етнічної ворожнечі. Відкрилось діалогове вікно для вибору файлу. Вміст файлу відобразився у текстовому полі «Text for Prediction» У текстовому полі «Prediction result» відображено «*Warning! The text contains manifestations of ethnic enmity. Percentage of manifestation of ethnic enmity: 0.99421436*»
Результат виконання тест-кейсу: перевірку пройдено успішно.		

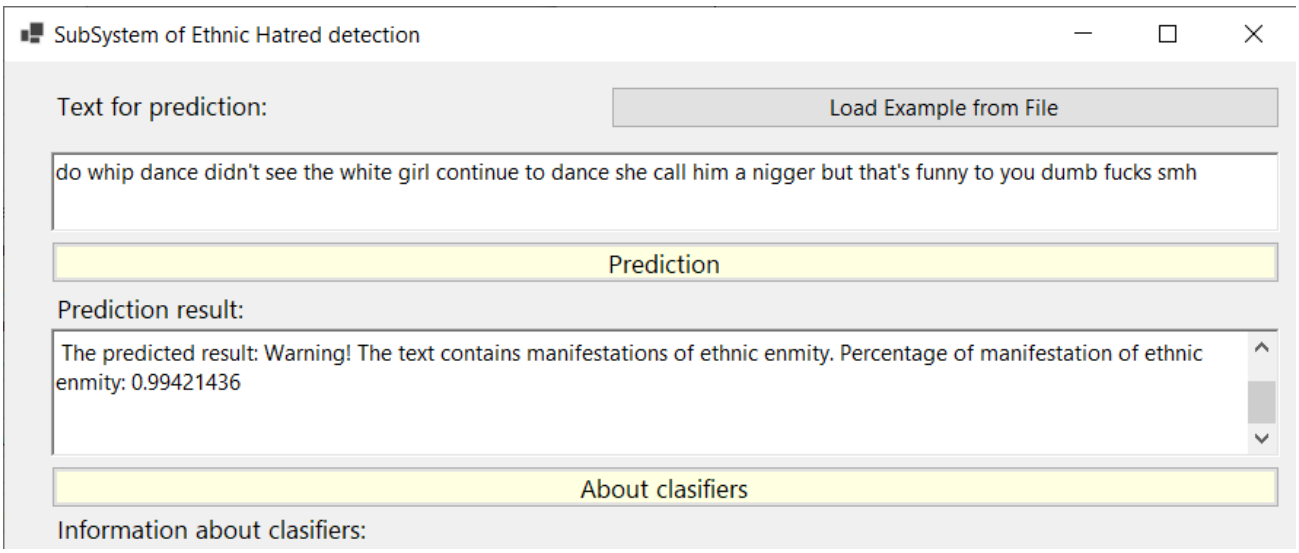


Рисунок 3.9 – Результат виконання тест-кейсу 00004

Для розгортання розробленої інформаційної системи є певні вимоги. Щодо вимог до апаратних засобів, то перелік рекомендованих вимог наведено нижче:

– процесор Intel Core i5 або еквівалентний;
– оперативна пам'ять 8 ГБ;
– вільний дисковий простір від 50 Гб.

Рекомендована операційна система – Windows 10.

Отже, було здійснено тестування інформаційної системи ідентифікації етнічної ворожнечі за текстовим представленням засобами NLP. У ході перевірки функцій інформаційної системи ідентифікації етнічної ворожнечі за текстовим представленням, непрацюючого функціоналу не виявлено, всі заявлені функції працюють коректно згідно до поставленого завдання.

3.6 Аналіз функціональності інформаційної системи ідентифікації етнічної ворожнечі за текстовим представленням

Для використання інформаційної системи ідентифікації етнічної ворожнечі за текстовим представленням необхідно запустити застосунок, та обрати одну з 3-х підсистем, використовуючи для цього кнопки переходу з головного меню. Після запуску застосунку користувач побачить головне меню (рисунок 3.10).

З головного меню є можливість скористатись функціями таких підсистем, як: «Підсистеми роботи з експериментальними даними», «Підсистеми виявлення проявів етнічної ворожнечі», «Підсистеми попередньої обробки даних». Для переходу на підсистему роботи з експериментальними даними необхідно натиснути кнопку «SubSystem Work with Experimental Data». Після чого відкриється форма підсистеми роботи з експериментальними даними (рисунок 3.11).

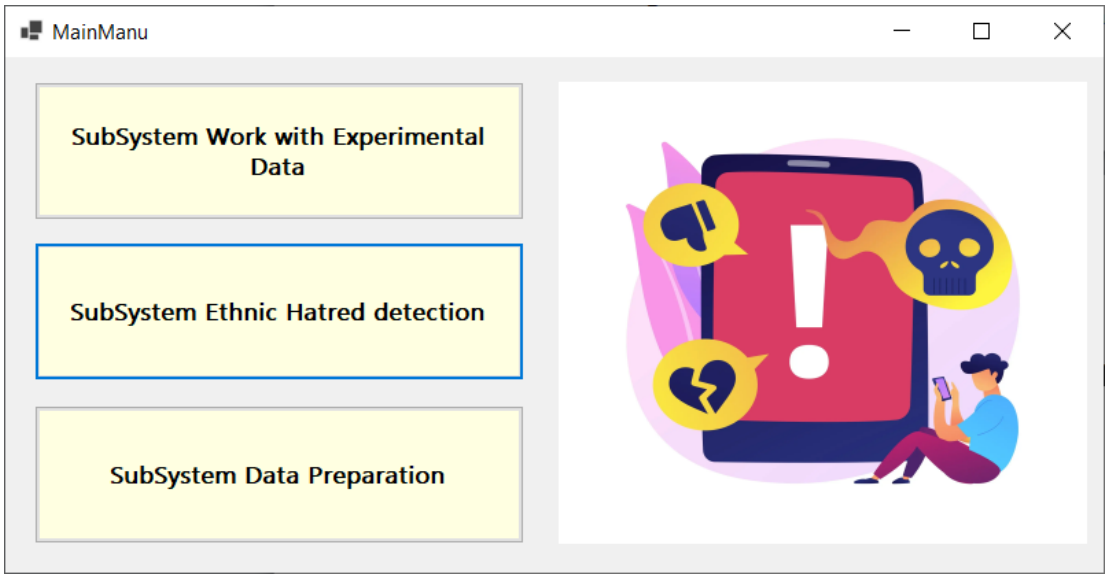


Рисунок 3.10 – Вигляд головного меню інформаційної системи ідентифікації етнічної ворожнечі за текстовим представленням

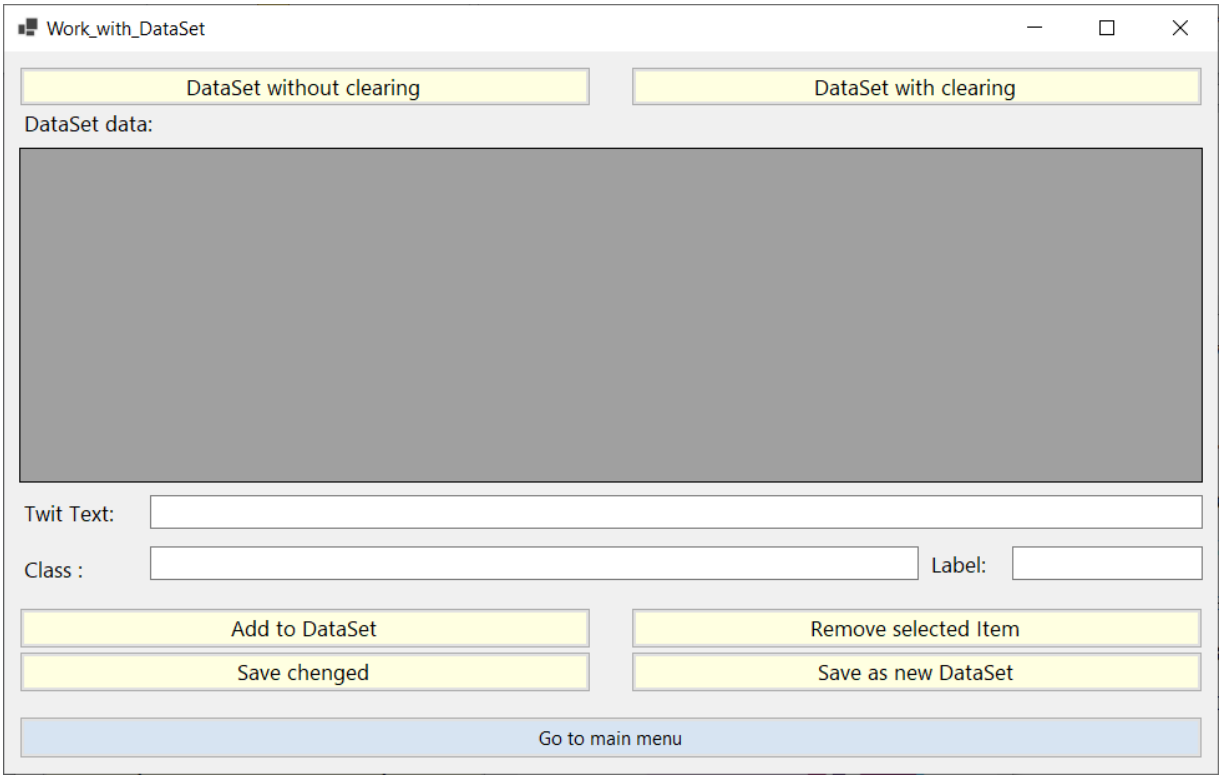


Рисунок 3.11 – Вигляд форми підсистеми роботи з експериментальними даними

Для завантаження неочищеного робочого набору даних необхідно натиснути на кнопку «DataSet without cleaning». Результат виконання наведено на рисунку 3.12.

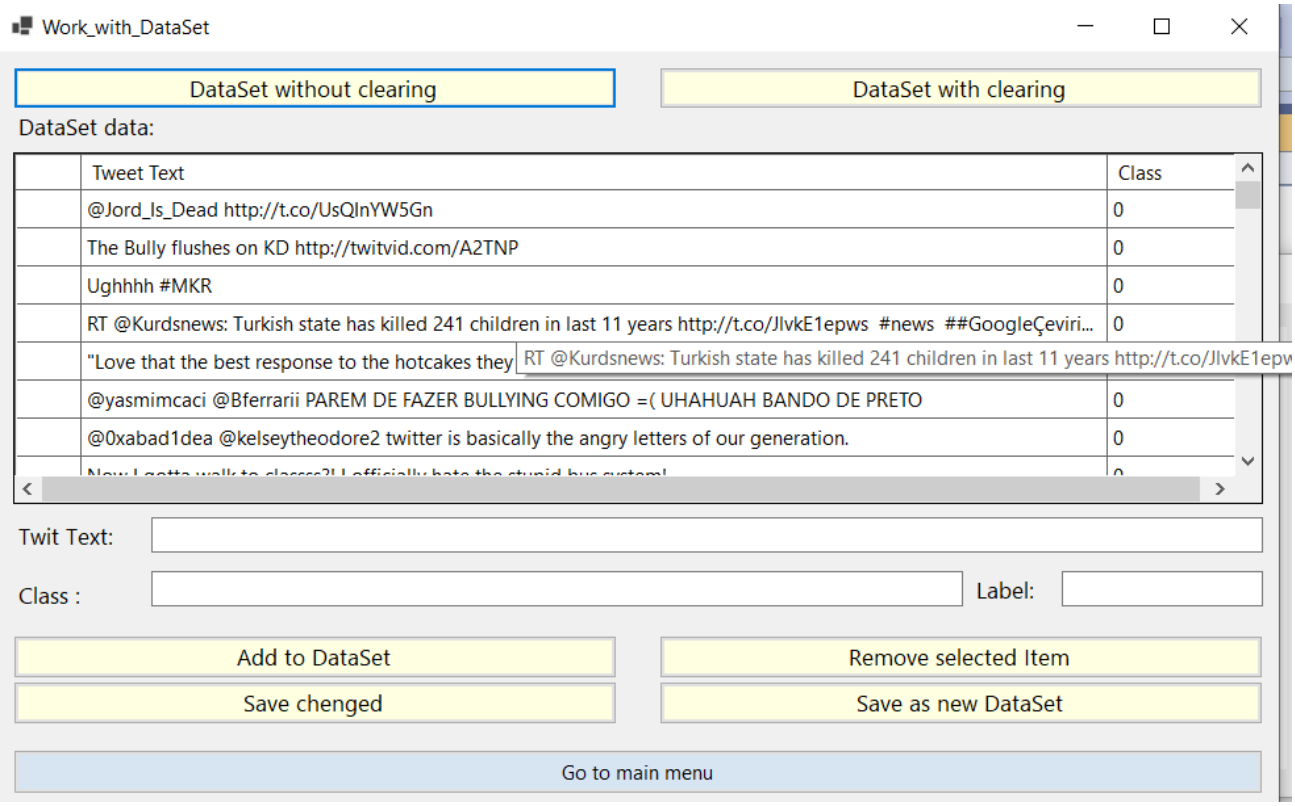


Рисунок 3.12 – Неочищений набір даних

Для деталізації обраного запису з датасету необхідно натиснути маніпулятором-миша на запис, який є потреба деталізувати. Приклад деталізації запису наведено на рисунку 3.13.

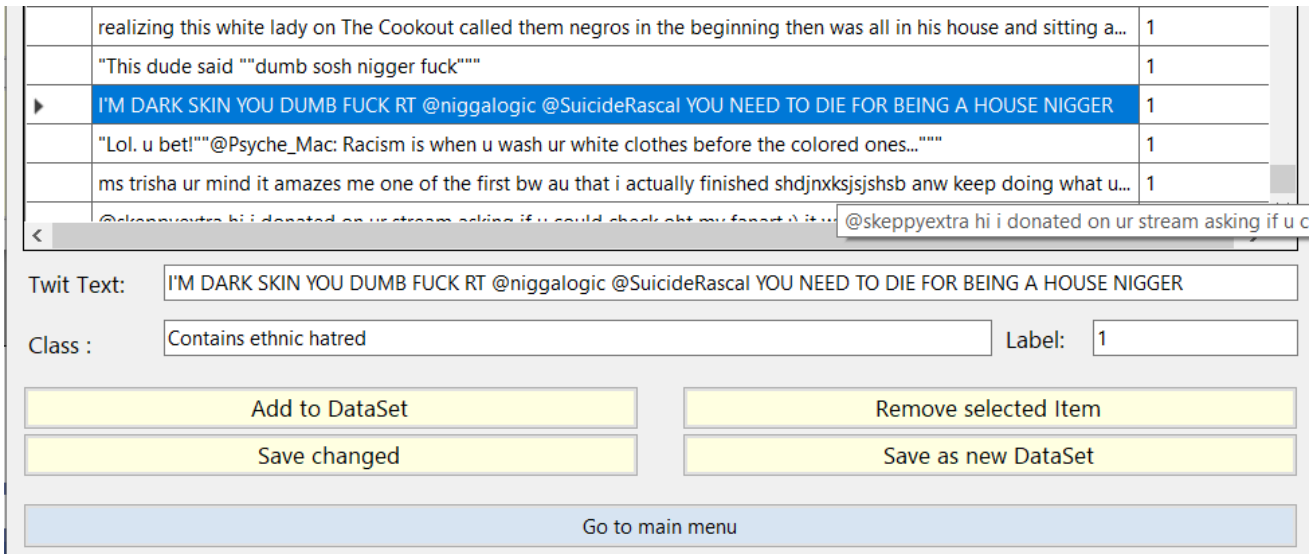


Рисунок 3.13 – Деталізація обраного запису

Для додавання нового запису в набір даних необхідно в поле «Twit Text» ввести необхідний текст, в поле «Class» ввести один з двох класів, в поле «Label» ввести 1, якщо є прояви етнічної ворожнечі, та 0 – якщо немає.

Для збереження змін необхідно натиснути кнопку «Save changed». Для видалення обраного рядка необхідно натиснути кнопку «Remove selected Item». Для збереження даних як нового датасета, необхідно натиснути кнопку «save as new DataSet», після чого відкриється діалогове вікно для вибору місця, куди зберегти датасет.

Ті самі дії можна робити і з очищеним датасетом. Для перегляду неочищеного датасету необхідно натиснути кнопку «DataSet with cleaning». При натисненні на кнопку «Go to main menu» відбудеться перехід до головного меню.

Для роботи з підсистемою попередньої обробки даних, в головному меню необхідно натиснути кнопку «SubSystem Data Preparation». Вигляд форми наведено на рисунку 3.14.

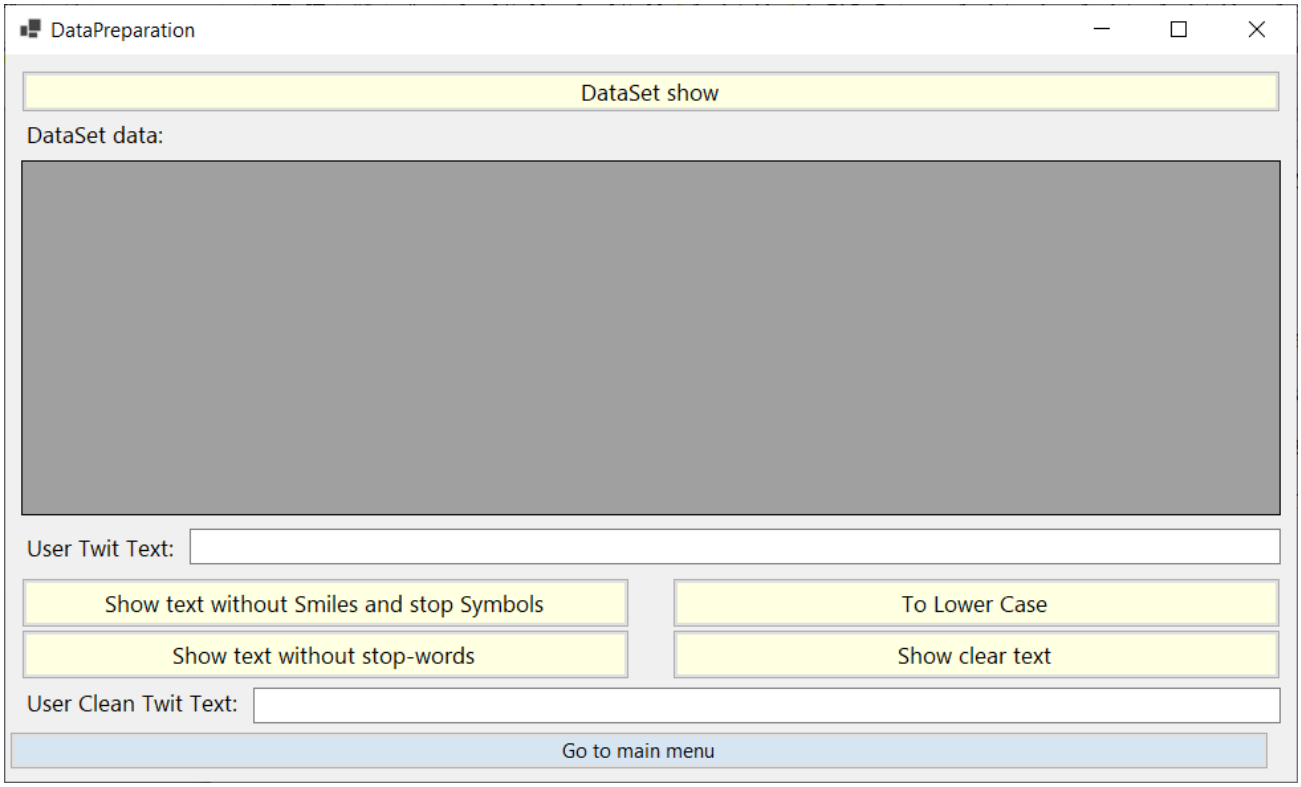


Рисунок 3.14 – Вигляд підсистеми попередньої обробки даних

Для перегляду датасету потрібно натиснути на кнопку «DataSet show», після чого датасет буде відображено у таблиці «DataSet Data» (рисунок 3.15).

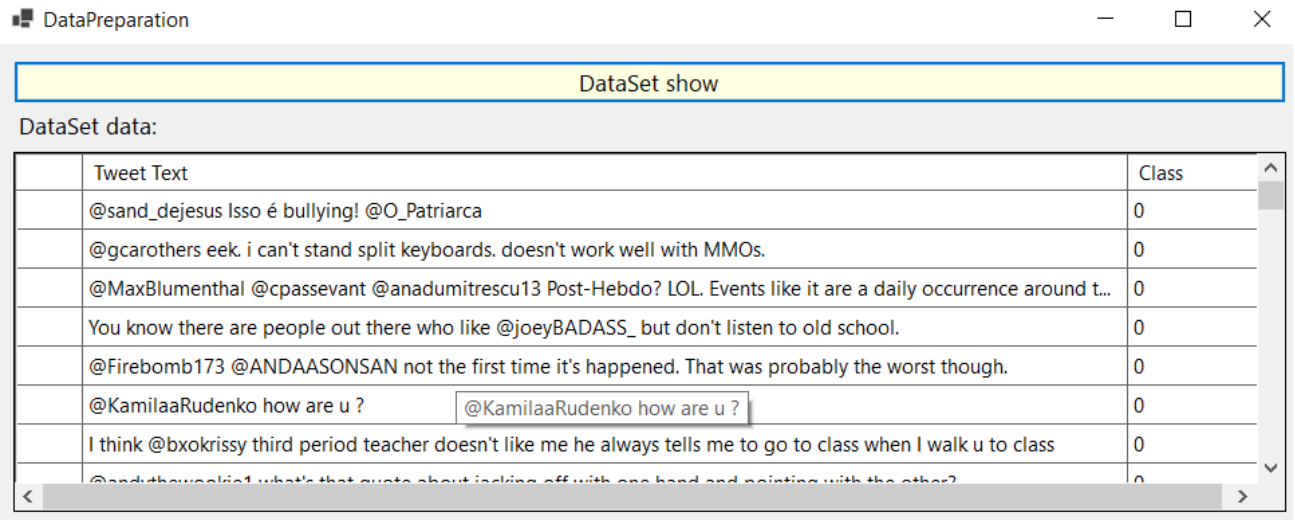


Рисунок 3.15 – Завантаження датасету

Попередню обробку тексту є можливість робити як для обраного текстового повідомлення, так і до уведеного вручну. Щоб виконати операції попередньої обробки обраного повідомлення потрібно натиснути на нього мишкою, і воно відобразиться у текстовому полі «User Tweet Text» (рисунок 3.16).

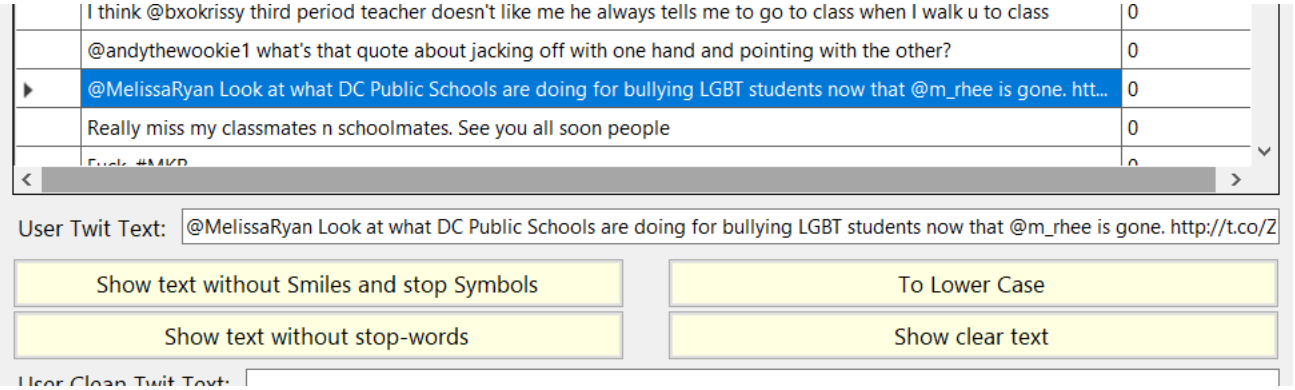


Рисунок 3.16 – Відображення тексту твіта

Для очистки тексту повідомлення від смайлів та стоп-символів необхідно натиснути кнопку «Show text without Smiles and stop Symbols». Очищений від

смайлів і стоп-символів текст буде виведено в полі «User Clean Twit Text:» (рисунок 3.17).

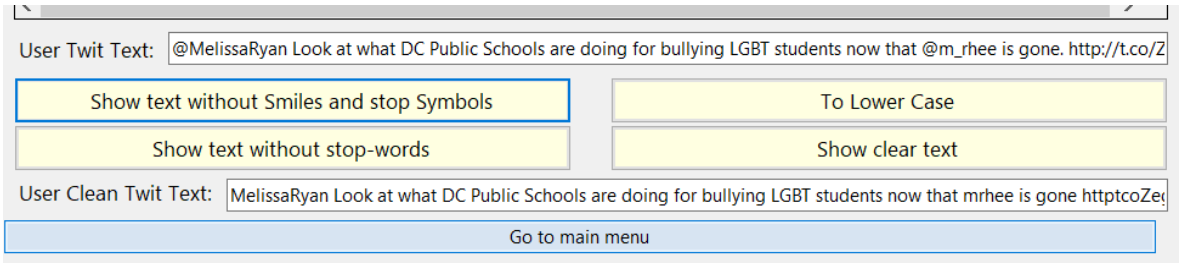


Рисунок 3.17 – Очистка від стоп-символів та смайлів

Для очистки тексту від стоп-слів необхідно натиснути на кнопку «Show text without stop-words». Приклад виконання очищення від стоп-слів наведено на рисунку 3.18.

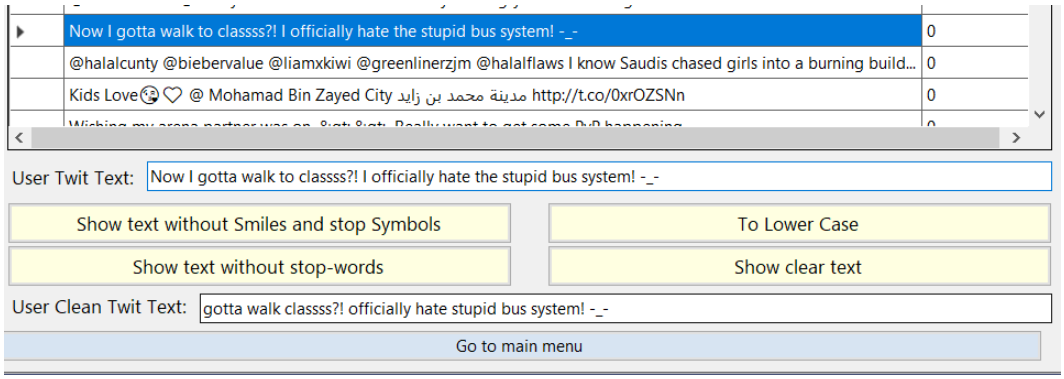


Рисунок 3.18 – Очистка від стоп-слів

Для переведення тексту у нижній реєстр необхідно натиснути кнопку «To Lower Case», а для застосування усіх фільтрів одразу необхідно натиснути кнопку «Show clear text». Приклад переведення тексту у нижній реєстр наведено на рисунку 3.19.

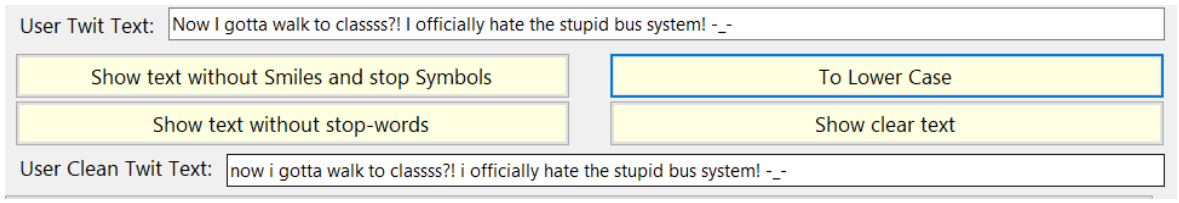


Рисунок 3.19 – Перетворення у нижній регістр

Для переходу назад у головне меню необхідно натиснути кнопку «Go to main menu». Для використання функціоналу головної підсистеми виявлення проявів етнічної ворожнечі необхідно натиснути кнопку «SubSystem Ethnic Hatred detection» в головному меню. Вигляд головної підсистеми наведено на рисунку 3.20.

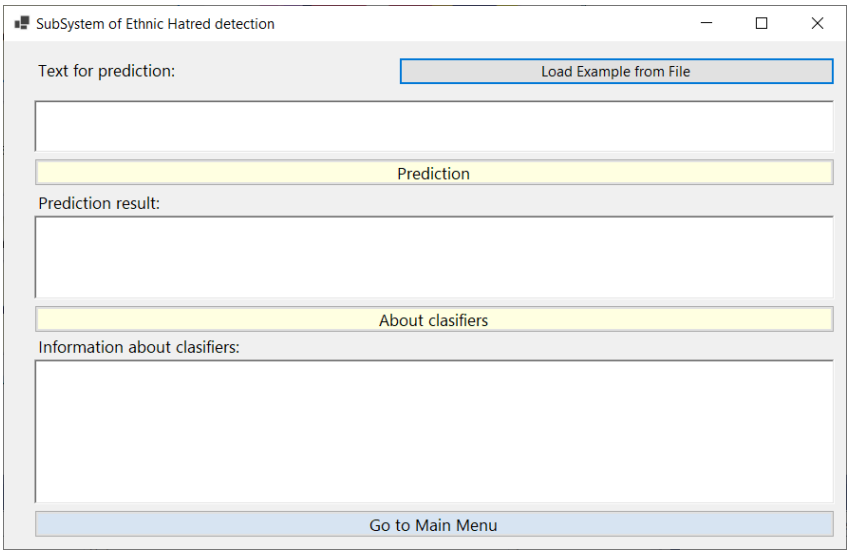
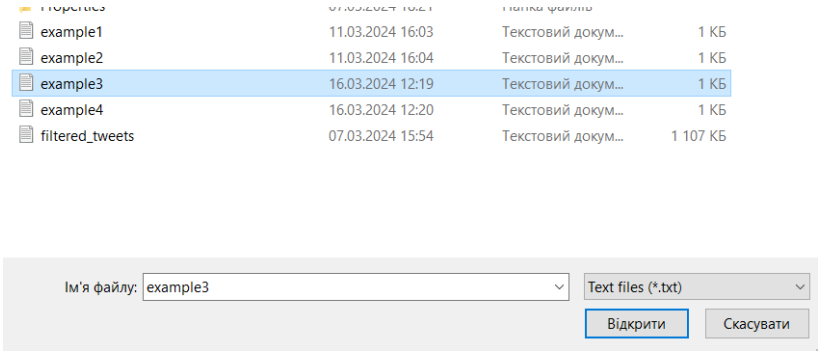


Рисунок 3.20 – Підсистема виявлення проявів етнічної ворожнечі

Є можливість як досліджувати уведені текстові повідомлення, так і використовувати уже існуючі зразки з файлів. Для перевірки тексту написаного власноруч, текст необхідно вписати у текстове поле «Text for prediction». Для перевірки тексту з файлу необхідно натиснути кнопку «Load Example from File», після чого буде відкрито діалогове вікно з можливістю обрати файл для аналізу (рисунок 3.21).



Рсунок 3.21 – Вибір документа для аналізу

Для виявлення вмісту етнічної ворожнечі у повідомленні необхідно натиснути кнопку «Prediction», і результат аналізу буде наведено у текстовому полі «Prediction result:» (рисунок 3.22).

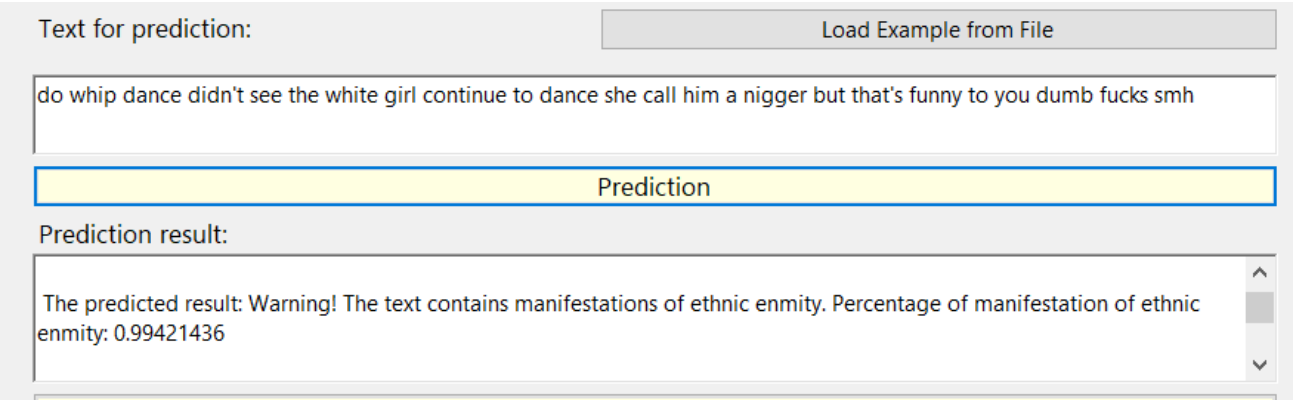


Рисунок 3.22 – Приклад виявлення етнічної ворожнечі

Для перегляду даних класифікатора необхідно натиснути кнопку «About Classifiers», де буде виведено метрики та матрицю сплутувань класифікатора FastForest (рисунок 3.23).

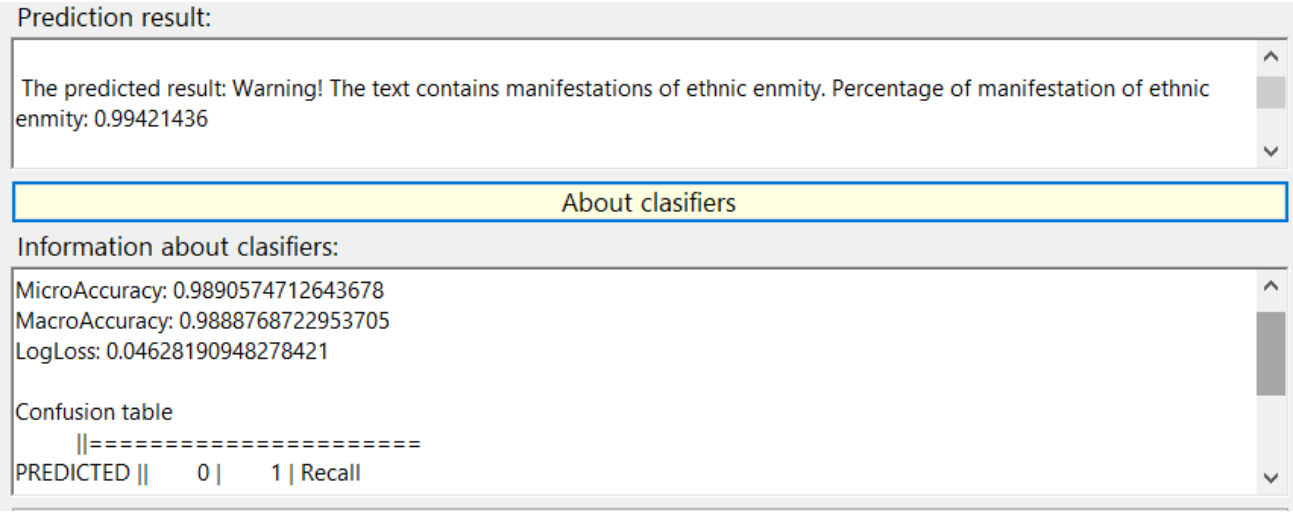


Рисунок 3.23 – Статистичні дані метрик

Отже, було виконано аналіз функціональності інформаційної системи ідентифікації етнічної ворожнечі за текстовим представленням, що складається із 3-х основних підсистем: «Підсистеми роботи з експериментальними даними», «Підсистеми виявлення проявів етнічної ворожнечі», «Підсистеми попередньої

обробки даних». Для кожної підсистеми детально описано особливості їх використання.

3.7 Результати досліджень

Дослідження ефективності методу виконувалось з використанням розробленого програмного забезпечення, шляхом порівняння отриманих відповідей з валідаційним набором, а також виконано оцінку навченої моделі машинного навчання FastForest, за допомогою використання метрик MicroAccuracy, MacroAccuracy, LogLoss, ConfusionMatrix, f1-міри та Recall.

Без зміни робочої навчальної вибірки, значення метрик були такими, як таведено в таблиці 3.5.

Таблиця 3.5 – Метрики за базовим навчальним набором

Метрика:	**MicroAccuracy**	**MacroAccuracy**	**LogLoss**
Значення:	0.9890	0.9889	0.0463

Тим часом, матриця сплутувань та метрики Precision і Recall мали значення наведені в таблиці 3.6.

Таблиця 3.6 – Матриця сплутувань та метрики Precision і Recall за базовим навчальним набором

Predicted:	**Не містить проявів етнічної вороженчі**	**Містить прояви етнічної ворожнечі**	**Recall**
Не містить проявів етнічної вороженчі	5 902	55	0.9908
Містить прояви етнічної ворожнечі	64	4854	0.9870
Precision	0.9893	0.9888	

Значення метрик Precision і Recall проілюстровані на графіку (рисунок 3.24).

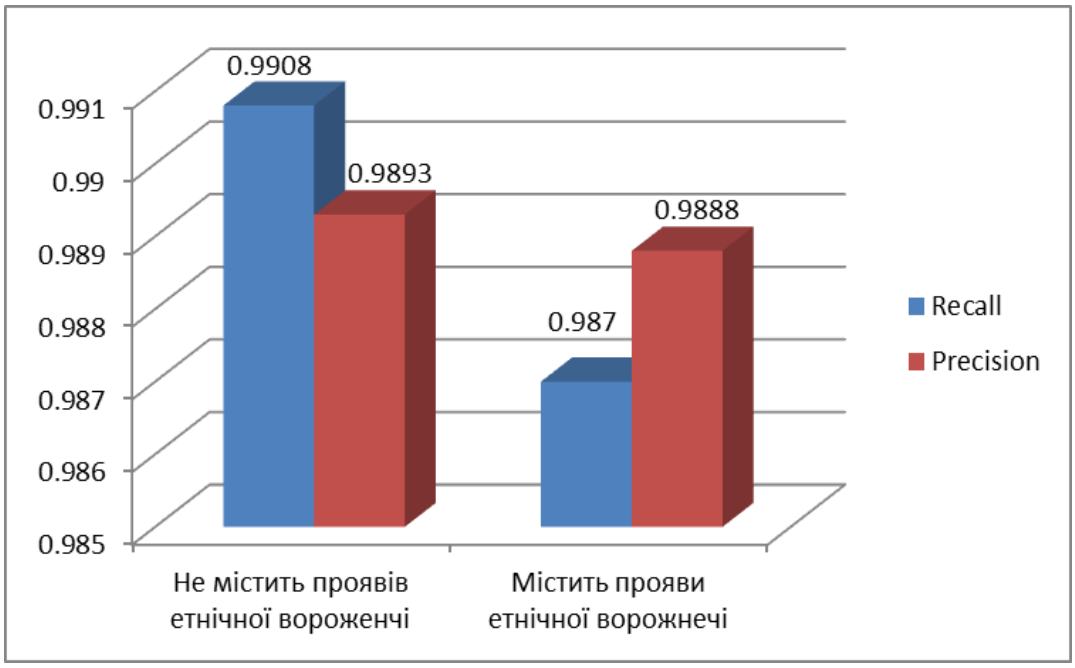


Рисунок 3.24 – Значення метрик Precision і Recall

Метрика Precision для класу «Не містить проявів етнічної вороженчі» рівна 0.9893, а для класу «Містить прояви етнічної ворожнечі» - 0.9888. Метрика вимірює, наскільки точно модель визначає позитивні випадки з усіх випадків, які вона визначає як позитивні. Висока точність вказує на те, що модель має добре відокремлені класи.

Метрика Recall для класу «Не містить проявів етнічної вороженчі» 0.9908, а для класу «Містить прояви етнічної ворожнечі» становить 0.9870. Дана метрика вимірює, наскільки ефективно модель впізнає всі істинні позитивні випадки. Ці результати свідчать про те, що модель має високу точність і чутливість для обох класів, що робить її дуже ефективною для цільових задач виявлення проявів кібербулінгу.

Також був виконаний аналіз текстів, що були некоректно ідентифіковані. Приклад текстів з розміткою в наборі даних «Не містять проявів етнічної ворожнечі», які були некоректно ідентифіковані програмою наведено в таблиці 3.7.

Як видно з таблиці 3.7 – тексти містять прояв агресії, а останній текст напряму відноситься до проявів етнічної ворожнечі. Також можливі помилки через мову.

Таблиця 3.7 – Приклад некоректно ідентифікованих текстів з розміткою в наборі даних «Не містять проявів етнічної ворожнечі»

Текст	**Переклад**
@XochitlSuckkks a classy whore? Or more red velvet cupcakes? (англійська)	@XochitlSuckkks класна повія? Або ще кекси з червоного оксамиту?
Itu sekolah ya bukan tempat bully! Ga jauh kaya neraka (індонезійська)	Це школа, а не місце для знущань! Недалеко від пекла
Karma. I hope it bites Kat on the butt. She is just nasty. #mkr (англійська)	Карма. Сподіваюся, це вкусить Кет за дупу. Вона просто противна. #мкр
You look like a gypsy. Sorry I'm not sorry.	Ви схожі на циганку. Вибачте, я не шкодую.

Так як для обробки навчальних даних було використано інструменти для роботи з англійською мовою, тексти на інших мовах могли бути опрацьовані некоректно. Тому є потреба в подальшому зробити ще один етап навчання класифікатора, видаливши перед тим ряд суперечливих даних, як до прикладу, в таблиці 3.7, а також видаливши всі тексти, які наведені не англійською мовою. Щодо порівняння отриманих програмою відповідей та ChatGpt 3.5, подавши останній твіт, чат видав таку відповідь «*Так, цей текст містить прояви етнічної ворожнечі. Він містить образливий термін, який стосується певної етнічної групи, та використання вибачення не компенсує його негативного змісту..*», що говорить про те, що ряд помилок не є помилками.

Напрямками можливого застосування методу виявлення проявів етнічної ворожнечі і програмної системи можуть бути платформи соціальних медіа та вебсайти, які можуть використовувати програмні системи для виявлення та фільтрації образливих або расистських коментарів. Це може сприяти створенню

більш безпечних та онлайн-середовищ для користувачів. Також отримані результати можуть використовуватись дослідниками та активістами, які можуть використовувати методи та програмні системи для аналізу та вивчення виявлених випадків етнічної ворожнечі. Це може допомогти в розумінні причин та наслідків таких проявів і розробці стратегій протидії.

Отже, було наведено дослідження ефективності методу виявлення проявів етнічної ворожнечі на базі створеної інформаційної системи. Виконано оцінку навченої моделі машинного навчання FastForest, за допомогою використання метрик MicroAccuracy, MacroAccuracy, LogLoss, ConfusionMatrix, f1-міри та Recall, що показали високі результати використання навченої моделі у спроможності виявлення проявів етнічної ворожнечі.

3.8 Висновки до розділу 3

У рамках розділу 3 було проведено експериментальне дослідження методу з використанням власної програмної реалізації, інформаційної системи ідентифікації етнічної ворожнечі за текстовим представленням. Для написання програмної реалізації застосунка на базі методу виявлення проявів етнічної ворожнечі у текстових повідомленнях соціальних інтернет-мереж було використано платформу .NET 7.0, середовище програмування Visual Studio, та мову програмування C#, що є дуже потужною комбінацією, оскільки Visual Studio надає широкий набір інструментів та можливостей для створення, навчання та експериментів з моделями машинного навчання на основі мови C#.

Для забезпечення функціональних можливостей інформаційної системи ідентифікації етнічної ворожнечі за текстовим представленням, яка складається із 3-х основних підсистем було спроєктовано відповідну діаграму класів, що складається із 12 класів, які виконують весь зазначений функціонал.

Описано особливості реалізації програмних складових інформаційної системи ідентифікації етнічної ворожнечі за текстовим представленням, призначеної для автоматизованого аналізу текстів, що публікуються у

соціальних мережах з метою виявлення ознак ворожнечі або конфлікту між представниками різних етнічних груп.

Виконано тестування інформаційної системи ідентифікації етнічної ворожнечі за текстовим представленням засобами NLP. У ході перевірки функцій системи, непрацюючого функціоналу не виявлено, всі заявлені функції працюють коректно згідно до поставленого завдання. Перевірка виконувалась засобами тест-кейсів.

Виконано аналіз функціональності інформаційної системи ідентифікації етнічної ворожнечі за текстовим представленням виявлення проявів етнічної ворожнечі, що складається із 3-х основних підсистем: «Підсистеми роботи з експериментальними даними», «Підсистеми виявлення проявів етнічної ворожнечі», «Підсистеми попередньої обробки даних». Для кожної підсистеми детально описано особливості їх використання.

Наведено дослідження ефективності методу виявлення проявів етнічної ворожнечі на базі створеної інформаційної системи. Виконано оцінку навченої моделі машинного навчання FastForest, за допомогою використання метрик MicroAccuracy, MacroAccuracy, LogLoss, ConfusionMatrix, f1-міри та Recall, що показали високі результати використання навченої моделі у спроможності виявлення проявів етнічної ворожнечі.

Загальні висновки

Метою кваліфікаційної роботи бакалавра було спрощення експертизи виявлення проявів етнічної ворожнечі за рахунок автоматизованого її виявлення у текстових повідомленнях соціальних інтернет-мереж, для чого проводилась розробка методу виявлення проявів етнічної ворожнечі у текстових повідомленнях соціальних інтернет-мереж NLP-засобами, а також відповідної інформаційної системи ідентифікації етнічної ворожнечі за текстовим представленням, яка використала створений метод.

Для досягнення мети поставлені та виконані наступні задачі:

– виконано дослідження предметної області для задачі виявлення проявів етнічної ворожнечі у текстових повідомленнях;

– в рамках дослідження предметної області виконано огляд теоретичних підходів щодо виявлення проявів етнічної ворожнечі у текстових повідомленнях, обрано методи ансамблевого навчання;

– виконано аналіз існуючих програмних рішень в області виявлення проявів етнічної ворожнечі у текстових повідомленнях;

– розроблено метод виявлення проявів етнічної ворожнечі у текстових повідомленнях, що використовує техніки обробки природної мови, а саме підхід на основі ансамблів, і здійснює перетворення вхідних даних у вигляді навченого класифікатора FastForest та вхідного текстового повідомлення у вихідні дані у вигляді відсотка прояву етнічної ворожнечі у тестовому повідомленні соціальних інтернет-мереж;

– на основі розробленого методу виконано проєктування інформаційної структури системи ідентифікації етнічної ворожнечі за текстовим представленням;

– виконано підготовку навчальних даних, у якості датасету обрано «Cyberbullying Classification», з якого взято дві категорії: «Ethnicity» та «Not cyberbullying», розмір навчальної вибірки становить 7961 твітів категорії

«Ethnicity» та 7945 твітів категорії «Not cyberbullying», що разом складає 15906 твітів;

– здійснено вибір засобів розробки для створення інформаційної системи;

– здійснено програмну реалізацію інформаційної системи ідентифікації етнічної ворожнечі за текстовим представленням;

– проведено тестування розробленої інформаційної системи ідентифікації етнічної ворожнечі за текстовим представленням;

– здійснено дослідження ефективності розробленого методу виявлення проявів етнічної ворожнечі у текстових повідомленнях соціальних інтернет-мереж NLP-засобами з використанням розробленої програмної реалізації.

Результат роботи відповідає поставленому завданню в повній мірі, про що свідчить проведене тестування інформаційної системи ідентифікації етнічної ворожнечі за текстовим представленням та дослідження ефективності методу.

Метрика Precision для класу «Не містить проявів етнічної вороженчі» показала значення 0.9893, а для класу «Містить прояви етнічної ворожнечі» - 0.9888. Висока точність вказує на те, що модель має добре відокремлені класи. Метрика Recall для класу «Не містить проявів етнічної вороженчі» показала 0.9908, а для класу «Містить прояви етнічної ворожнечі» - 0.9870. Це є доволі високими показниками з урахуванням роботи з текстами, які не були до кінця опрацьованими. Зустрічались тексти на інших мовах, а також були випадки, коли категорії були розмічені некоректно. Подальші дослідження спрямовані на роботу з навчальною вибіркою та очисткою даних з метою подальшого перенавчання класифікатора.

Основні наукові й практичні результати доповідалися у доповіді «Object-Oriented Approach for Ethnic Enmity Detection in Text Messages by NLP» на XXI Міжнародній науково-практичній конференції «Scientific Achievements and Innovations as a Way to Success» (May 1-3, 2024, Vilnius, Lithuania), за темою кваліфікаційної роботи бакалавра автором виконано наукову публікацію [33].

Виявлення проявів етнічної ворожнечі сприяє формуванню позитивного онлайн-середовища, що сприяє взаєморозумінню, толерантності та відкритому обміну ідеями.

Використання різноманітних алгоритмів та моделей штучного інтелекту дозволяє автоматизовано та ефективно виявляти прояви етнічної ворожнечі у текстових повідомленнях соціальних інтернет-мереж, сприяючи покращенню безпеки та комфорту користувачів у цьому цифровому середовищі. У контексті виявлення етнічної ворожнечі в текстах соціальних мереж застосування NLP-засобів виявляється ключовим. Для цієї задачі було використано NLP-засоби, які підтримують аналіз текстового контенту, виявлення тону, класифікацію текстів та розпізнавання емоцій, а саме ансамблевий підхід.

Перелік посилань

1. Study smarter. Social Networking. URL: https://www.studysmarter.co.uk/explanations/social-studies/social-institutions/social-networking/

2. Linkedin. Social Networks Definition: What are They and Why are They Important? URL: https://www.linkedin.com/pulse/social-networks-definition-what-why-important-affiliate-marketing/

3. Familylives. Bullying on social networks. URL: https://www.familylives.org.uk/advice/bullying/cyberbullying/what-to-do-if-you-re-being-bullied-on-a-social-network

4. Social Media Management. User roles. URL: https://social-media-management-help.brandwatch.com/hc/en-us/articles/4626232528029-User-Roles

5. Linkedin. Can social media companies curb cyberbullying. URL: https://www.linkedin.com/pulse/can-social-media-companies-curb-cyberbullying-soniya-roy/

6. Sudhakar M., Kaliyamurthie K.P.. Detection of fake news from social media using support vector machine learning algorithms. Measurement: Sensors, Volume 32. 2024. URL: https://doi.org/10.1016/j.measen.2024.101028

7. CSI Library. Misinformation and Disinformation: Thinking Critically about Information Sources. URL: https://library.csi.cuny.edu/c.php?g=619342&p=4310783

8. College of Engineering, Design and Computing. Cyberbullying Detection System. URL: https://engineering.ucdenver.edu/current-students/capstone-expo/archived-expos/spring-2020/computer-science/csci14-cyberbullying-detection-system

9. Britannica. Ethnic conflict. URL: https://www.britannica.com/topic/ethnic-conflict

10. Blagojevic B.. Causes of ethnic conflict: a conceptual framework. URL: https://peaceprognosis.files.wordpress.com/2014/01/causesofethnicconflict.pdf

11. Tahtinen T.. When Facebook Is the Internet: The Role of Social Media in Ethnic Conflict. URL: https://cadmus.eui.eu/bitstream/handle/1814/70858/ECO_2021_01.pdf?sequence=1&isAllowed=y

12. All together now. 10 signs of casual racism. URL: https://alltogethernow.org.au/10-signs-of-casual-racism/

13. Sambanis N., Shayo M.. Social Identification and Ethnic Conflict. American Political Science Review , Volume 107, Issue 2. 2013. pp. 294 – 325. URL: https://doi.org/10.1017/S0003055413000038

14. Towards datascience. Support Vector Machine – Introduction to Machine Learning Algorithms. URL: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

15. Geeksforgeeks. Logistic Regression in Machine Learning. URL: https://www.geeksforgeeks.org/understanding-logistic-regression/

16. Geeksforgeeks. Gradient Boosting in ML. URL: https://www.geeksforgeeks.org/ml-gradient-boosting/

17. Geeksforgeeks. Introduction to Recurrent Neural Network. URL: https://www.geeksforgeeks.org/introduction-to-recurrent-neural-network/

18. Medium. LSTM – Implementation, Advantages and Diadvantages. URL: https://medium.com/@prudhviraju.srivatsavaya/lstm-implementation-advantages-and-diadvantages-914a96fa0acb

19. Towards datascience. Ensembles: the only (almost) free Lunch in Machine Learning. URL: https://towardsdatascience.com/ensembles-the-almost-free-lunch-in-machine-learning-91af7ebe5090

20. FastForest: Increasing random forest processing speed while maintaining accuracy. URL: https://www.sciencedirect.com/science/article/abs/pii/S0020025520312330

21. Google Cloud. Improving Trust in AI and Online Communities with PaLM-based Moderation. URL: https://cloud.google.com/blog/products/ai-machine-learning/google-cloud-text-moderation

22. Perspective. Using machine learning to reduce toxicity online. URL: https://perspectiveapi.com

23. Azure. Azure AI Services. URL: https://azure.microsoft.com/en-us/products/ai-services

24. AWS. Start building on AWS today. URL: https://aws.amazon.com/

25. Ocampo N. B., Sviridova E., Cabrio E., Villata S.. An In-depth Analysis of Implicitand Subtle Hate Speech Messages. URL: https://hal.science/hal-04214094/document

26. Yin W., Zubiaga A. Towards generalisable hate speech detection: a review on obstacles and solutions. PeerJ Computer Science. 2021. URL: https://doi.org/10.7717/peerj-cs.598

27. Alsafari S., Sadaoui S., Mouhoub M.. Hate and offensive speech detection on Arabic social media. Online Social Networks and Media. Volume 19. 2020. URL: https://doi.org/10.1016/j.osnem.2020.100096

28. Anderson A. F., Baptista C. S., Paiva A. C.. Improving hate speech detection using Cross-Lingual Learning. Expert Systems with Applications. Volume 235. 2024. URL: https://doi.org/10.1016/j.eswa.2023.121115

29. Kaggle. Cyberbullying Classification. URL: https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification/data

30. Твіттер. URL: https://uk.wikipedia.org/wiki/Твіттер

31. ML.NET. URL: https://dotnet.microsoft.com/en-us/apps/machinelearning-ai/ml-dotnet

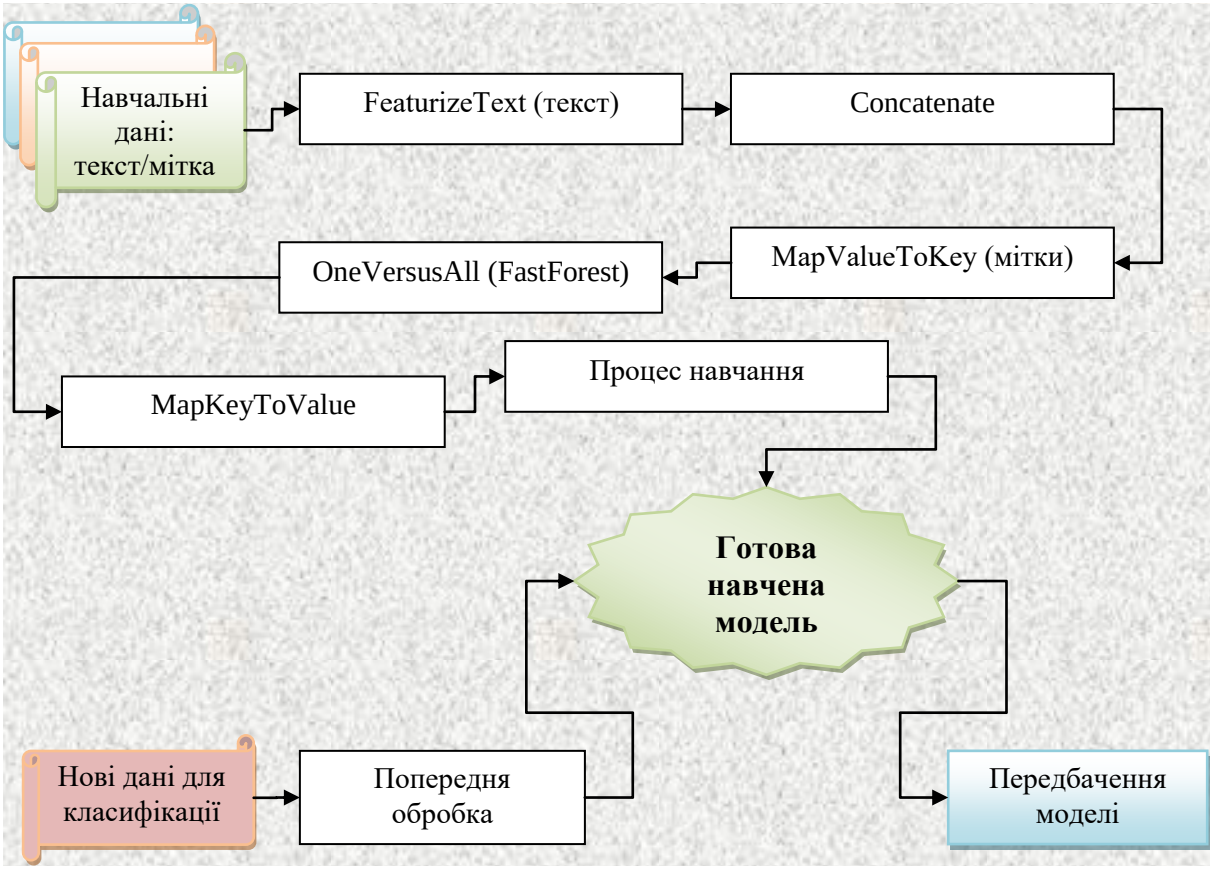
32. System.Text.RegularExpressions Namespace. URL: https://learn.microsoft.com/ru-ru/dotnet/api/system.text.regularexpressions?view=net-8.0

33. Molchanova M., Mazurets O., Sobko O., Boiarchuk I. Object-Oriented Approach for Ethnic Enmity Detection in Text Messages by NLP. Proceedings of XXI International Scientific and Practical Conference «Scientific Achievements and Innovations as a Way to Success». May 1-3, 2024. Vilnius, Lithuania. 2024. Pp. 73-77.

ДОДАТКИ

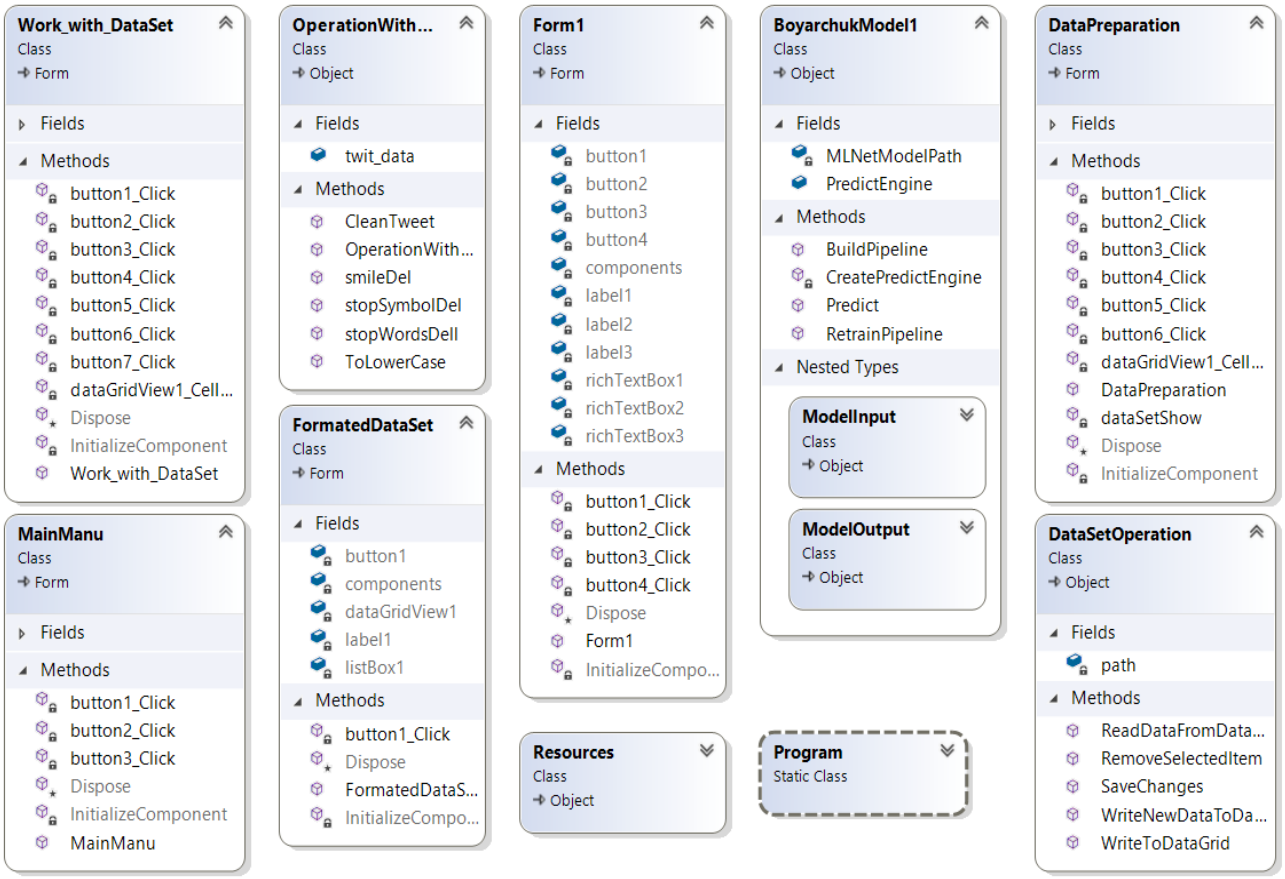
Додаток А

Проєктування пайплайну моделі машинного навчання для виявлення проявів етнічної ворожнечі



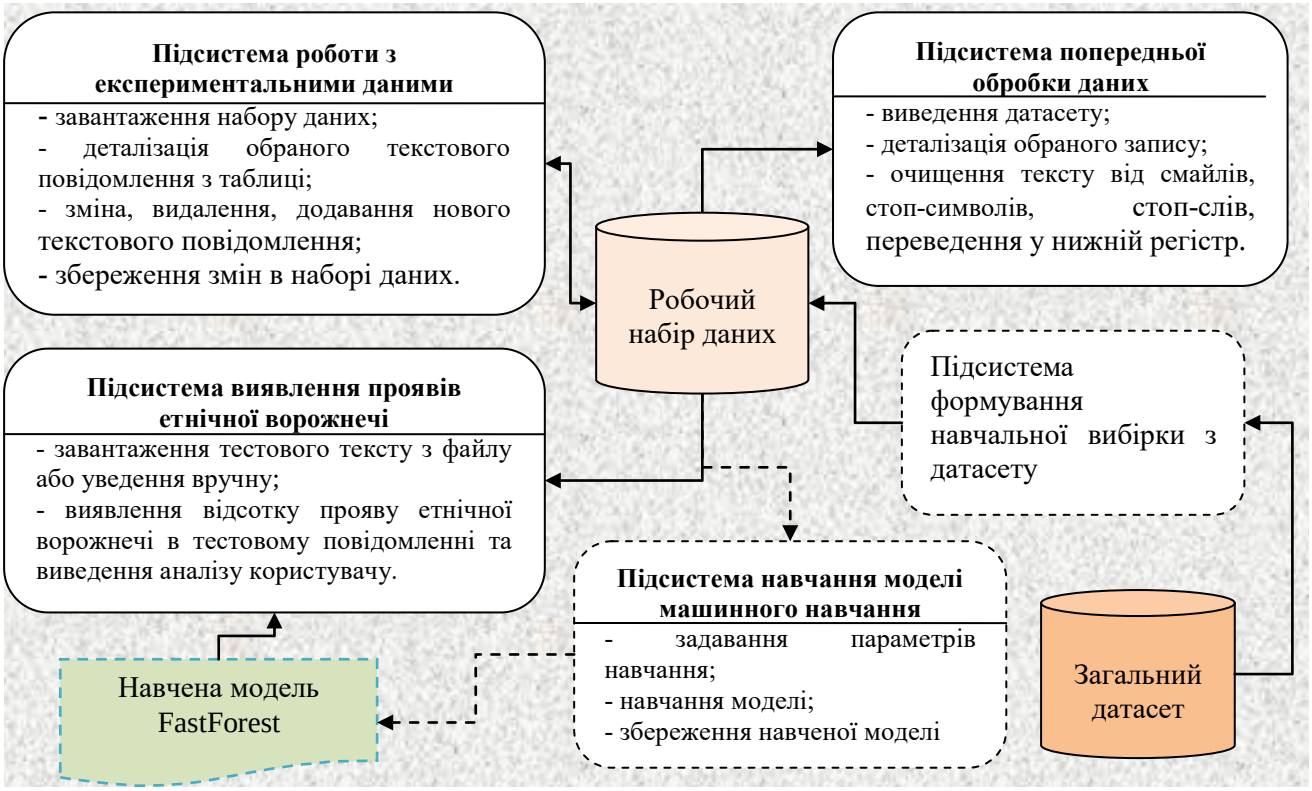
Додаток Б

Розгорнута структура класів інформаційної системи виявлення проявів етнічної ворожнечі



Додаток В

Проєктна архітектура інформаційної системи ідентифікації етнічної ворожнечі за текстовим представленням



Додаток Г

Презентаційний матеріал

КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА

МЕТОД ВИЯВЛЕННЯ ПРОЯВІВ ЕТНІЧНОЇ ВОРОЖНЕЧІ У ТЕКСТОВИХ ПОВІДОМЛЕННЯХ СОЦІАЛЬНИХ ІНТЕРНЕТ-МЕРЕЖ NLP-ЗАСОБАМИ

Виконав:
студент групи КНс-21-1
Ілля БОЯРЧУК

Керівник:
викладач кафедри КН
Марина МОЛЧАНОВА

Актуальність

За останні роки спостерігається стрімке зростання популярності соціальних мереж серед користувачів з усього світу. Вони стали важливим каналом для вираження думок, поглядів та емоцій. В той же час, у світі нерідко виникають конфлікти та напруженість між різними етнічними групами. Часто такі ситуації знаходять відображення у висловлюваннях у соціальних мережах, що створює потребу у виявленні та аналізі таких виразів.

Виявлення проявів етнічної ворожнечі є важливою проблемою для суспільства, оскільки вона може призвести до серйозних наслідків, включаючи конфлікти та розбрати між різними етнічними групами.

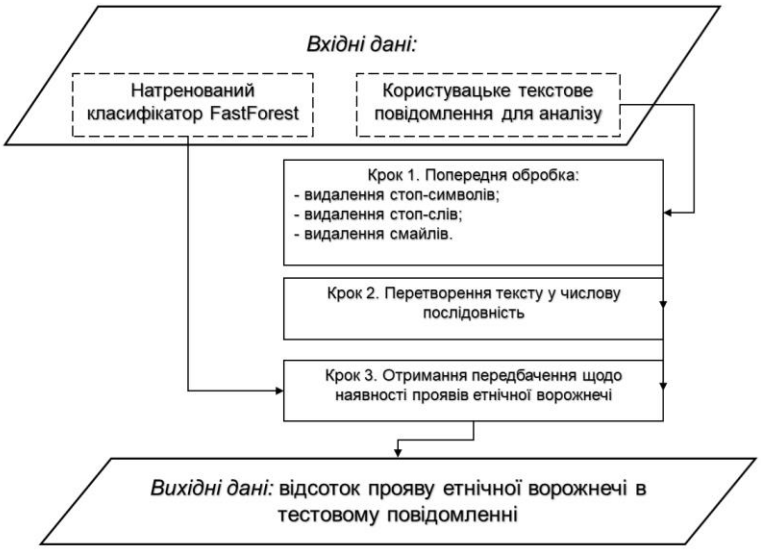
Засоби обробки природної мови наразі набувають великої популярності у сфері аналізу текстів. Вони дозволяють автоматизувати та полегшити аналіз великих обсягів даних, включаючи тексти, що публікуються в соціальних мережах.

Мета і задачі роботи

Мета кваліфікаційної роботи бакалавра – спрощення експертизи виявлення проявів етнічної ворожнечі за рахунок автоматизованого її виявлення у текстових повідомленнях соціальних інтернет-мереж.

Завдання кваліфікаційної роботи бакалавра – виконати дослідження предметної області для задачі виявлення проявів етнічної ворожнечі у текстових повідомленнях; в рамках дослідження предметної області виконати огляд теоретичних підходів щодо виявлення проявів етнічної ворожнечі у текстових повідомленнях; виконати аналіз існуючих програмних рішень в області виявлення проявів етнічної ворожнечі у текстових повідомленнях; розробити метод виявлення проявів етнічної ворожнечі у текстових повідомленнях; на основі розробленого методу виконати проєктування інформаційної структури системи ідентифікації етнічної ворожнечі за текстовим представленням; виконати підготовку навчальних даних для тренування класифікатора; здійснити вибір засобів розробки для створення інформаційної системи; здійснити програмну реалізацію інформаційної системи ідентифікації етнічної ворожнечі за текстовим представленням; провести тестування розробленої програмної реалізації; здійснити дослідження ефективності розробленого методу методу виявлення проявів етнічної ворожнечі у текстових повідомленнях соціальних інтернет-мереж NLP-засобами з використанням розробленої програмної реалізації..

Схема та кроки методу виявлення проявів етнічної ворожнечі у текстових повідомленнях



Проєктна архітектура системи виявлення проявів етнічної ворожнечі

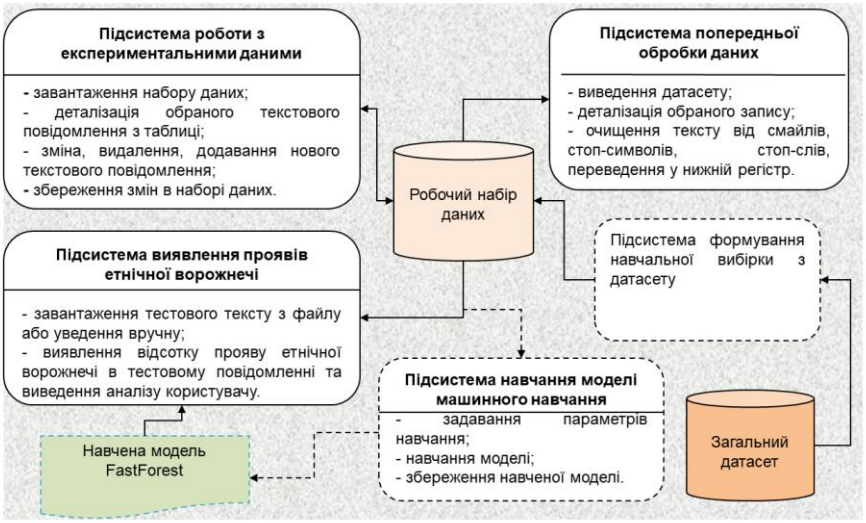
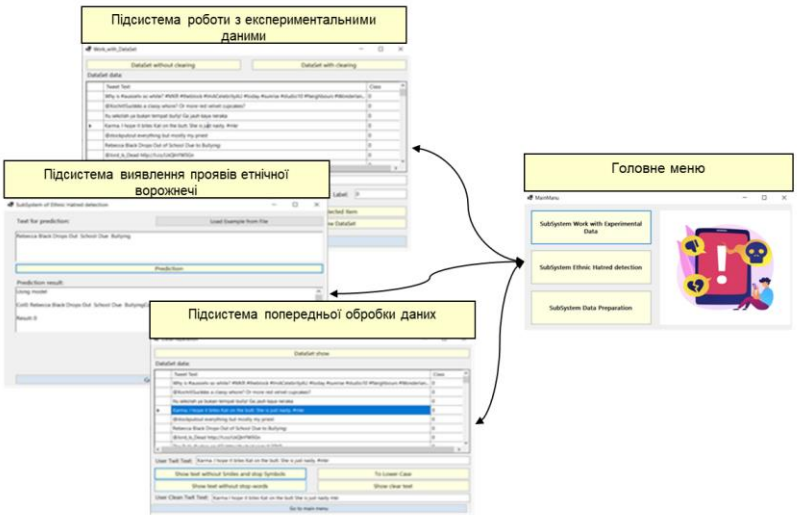


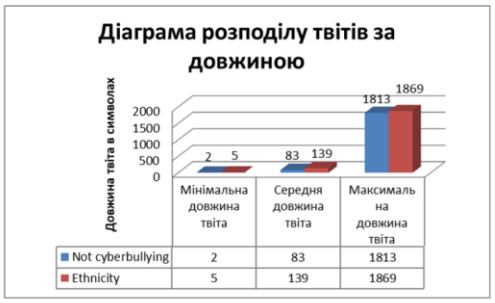
Схема навігації між інтерфейсними формами інформаційної системи



Підготовка робочих вхідних даних для системи

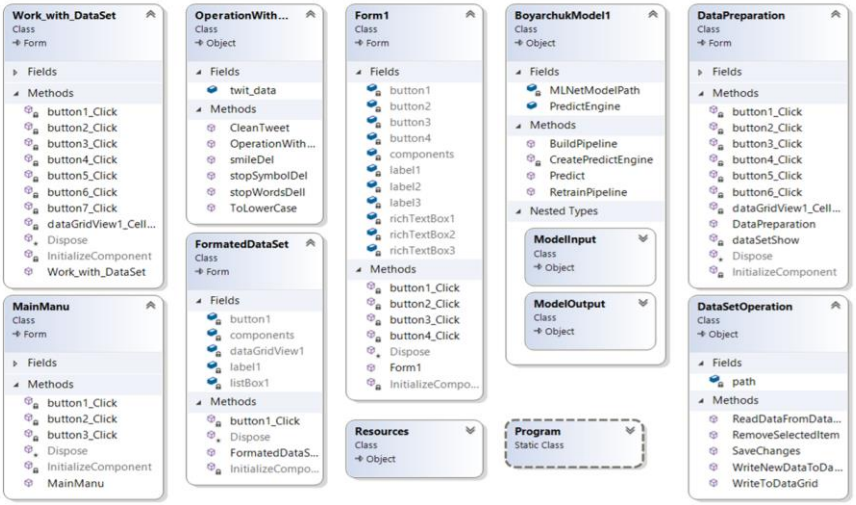
	tweet_text,cyberbullying_type
40861	11 в™Ў hi!! first of all, ure so gorgeous omggg i hope no one tell u otherwise!! i luv all of ur love memes hehe and also ur layout is amazing!! i love how classy u are when it comes to
40862	the niggers i fuck with aint no dumb niggers. http://www.youtube.com/user/BIGBKELtheGOD#p/u/0/F71wgPUCMQ4,ethnicity
40863	!!!аЂњ@PURDYdidit: Lol he must b smarter than romney...cuz tha dumb ass CRACKER a nobody lik u now..RT @tayyoung_: FUCK OBAMA, dumb ass niggeraЂ¦,ethnicity
40864	@HuffingtonPost being black and being bullied in America is nothing new to the majority of black people #dontfetit,ethnicity
40865	RACISM is when u select ur white cloth to wash first before the colored n dyed ones. Say NO to racism.... Wash them all 2gether!,ethnicity
40866	THATS RACIST DUMB FUCK "@goshawty123: black people who still think sayin nigger is racist , get a fucken life ."",ethnicity
40867	@Heylts_Jayla @ryan_temple goodone dumb nigger ur ugly as fuck so stfu #BJWON #getoverit,ethnicity
40868	U know the backlash I get for being darkskin with waffle colored children? Inner racism is worse than others!,ethnicity
40869	Just saw a skinhead call a black man in the street "a fat ugly nigger". Disgusting behaviour.,ethnicity
40870	I been done through enough bullshit so no one can tell me shit! I been see dumb ass niggers and dumb ass bitches up in this shit so fuck!,ethnicity
40871	we are talking about how institutional racism affects health care in bio therefore id like to aknowledge that my BIO TEACHER used the term aЂњcoloredaЂќ fuck u nancy,ethnicity
40872	@XXLtina: thank god Romney is winning! the blacks must be eating watermelon frantically! #ew #gross #getajob #niggers ignorant dumb fuck!,ethnicity
40873	Talk about your People you sorry ass dumb-ass Nigger bitch. Leave My Peo0le the fuck alone! You dumb black Fuck!,ethnicity
40874	WOW lol "@CowboyQuan: fuck education. imma dumb ignoarant nigger",ethnicity
40875	@tayyoung_: FUCK OBAMA, dumb ass nigger that "nigger" won tho...So suck a nigger dick!,ethnicity

Приклад даних датасету категорії «Ethnicity»

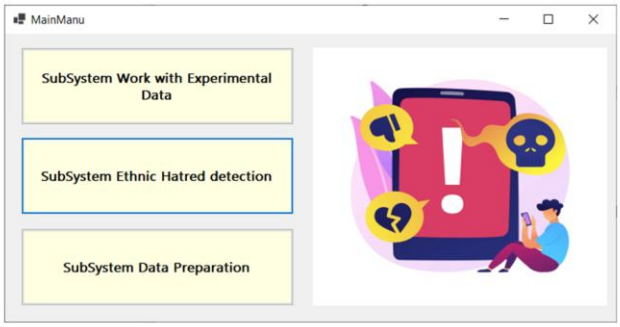


Розподіл твітів по категоріям за довжиною

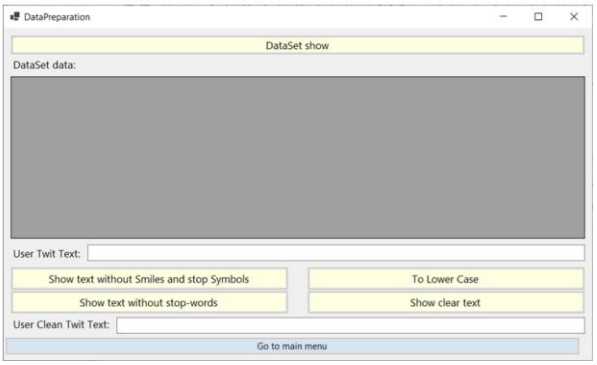
Діаграма класів застосунку



Інформаційна система виявлення проявів етнічної ворожнечі

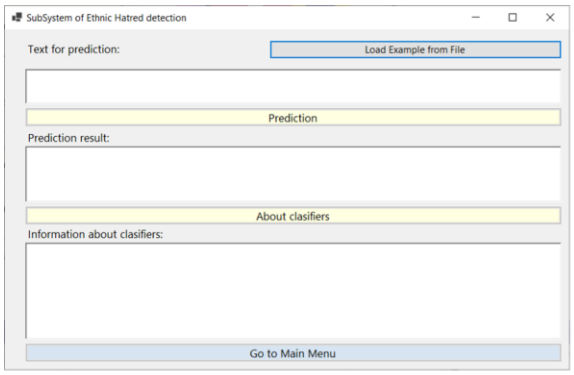


Вигляд головного меню застосунку

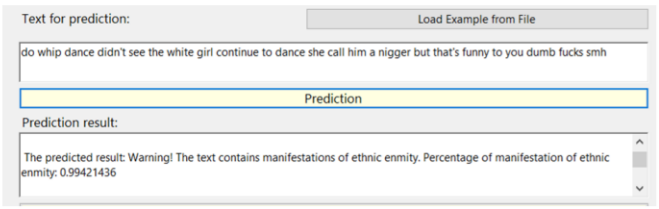


Вигляд підсистеми попередньої обробки даних

Інформаційна система виявлення проявів етнічної ворожнечі



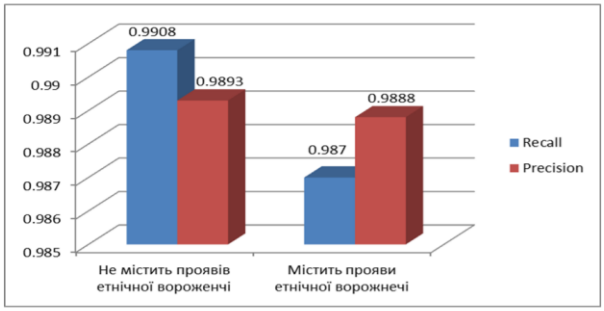
Підсистема виявлення проявів етнічної ворожнечі



Приклад виявлення етнічної ворожнечі

Результати досліджень

Predicted:	Не містить проявів етнічної ворожнечі	Містить прояви етнічної ворожнечі	Recall
Не містить проявів етнічної ворожнечі	5 902	55	0.9908
Містить прояви етнічної ворожнечі	64	4854	0.9870
Precision	0.9893	0.9888	



Висновки

Метою кваліфікаційної роботи бакалавра було спрощення експертизи виявлення проявів етнічної ворожнечі за рахунок автоматизованого її виявлення у текстових повідомленнях соціальних інтернет-мереж.

Для досягнення мети поставлені та виконані наступні задачі:

- виконано дослідження предметної області для задачі виявлення проявів етнічної ворожнечі у текстових повідомленнях;
- в рамках дослідження предметної області виконано огляд теоретичних підходів щодо виявлення проявів етнічної ворожнечі у текстових повідомленнях, обрано методи ансамблевого навчання;
- виконано аналіз існуючих програмних рішень в області виявлення проявів етнічної ворожнечі у текстових повідомленнях;
- розроблено метод виявлення проявів етнічної ворожнечі у текстових повідомленнях, що використовує техніки обробки природної мови;
- на основі розробленого методу виконано проєктування інформаційної структури системи ідентифікації етнічної ворожнечі за текстовим представленням;
- виконано підготовку навчальних даних;
- здійснено вибір засобів розробки для створення інформаційної системи;
- здійснено програмну реалізацію інформаційної системи ідентифікації етнічної ворожнечі за текстовим представленням;
- проведено тестування розробленої програмної реалізації;
- здійснено дослідження ефективності розробленого методу виявлення проявів етнічної ворожнечі у текстових повідомленнях соціальних інтернет-мереж NLP-засобами з використанням розробленої програмної реалізації.

UNICHECK
by Turnitin

Ім'я користувача:
Кафедра КН

ID перевірки:
1016379033

Дата перевірки:
20.06.2024 20:33:27 EEST

Тип перевірки:
Doc vs Internet + Library

Дата звіту:
20.06.2024 21:15:06 EEST

ID користувача:
100005671

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

Кількість сторінок: **70** Кількість слів: **11825** Кількість символів: **97326** Розмір файлу: **1.52 MB** ID файлу: **1016187706**

Виявлено модифікації тексту (можуть впливати на відсоток схожості)

10.7%
Схожість

Найбільша схожість: **5.4%** з джерелом з Бібліотеки (ID файлу: **1016177779**)

5.72% Джерела з Інтернету	612	Сторінка 72
7.7% Джерела з Бібліотеки	160	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.

Підозріле форматування	16 сторінок

Anti-Plagiarism v-15.257

Максимальне співпадіння з одним документом 4.0%

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

ID: 131876 Назва: КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА на тему Метод виявлення проявів етнічної ворожнечі у текстових повідомленнях соціальних інтернет-мереж NLP-засобами Додано в БД: 2024-06-20 Автора: Ілля БОЯРЧУК Керівники: Марина МОЛЧАНОВА Консультанти: Опоненти:	Документ		Сумарний збіг по Базі Даних	
	Символи	Лексеми	Символи	Лексеми
	78252	1122	5093 (7%)	76 (7%)

Джерело плагіату

ID	Опис	Наявність плагіату в документі	
		Символи	Лексеми

РІШЕННЯ ЕКСПЕРНОЇ 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 проявів етнічної ворожнечі у текстових повідомленнях соціальних інтернет-мереж NLP-засобами*

Автор: *студент гр. КНс-21-1 Боярчук Ілля Олександрович*

Спеціальність: *122 – Комп'ютерні науки*

Освітня програма: *освітньо-професійна*

Науковий керівник: *викладач кафедри КН 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 Боярчука Іллі, не є плагіатом, оскільки: запозичення розміщені в розділі огляду існуючих підходів, не описують безпосередньо авторську роботу і не стосуються її результатів; усі запозичення фрагментарні; серед запозичень знаходяться загальновідомі терміни, скорочення.

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

- за системою Anti-Plagiarism: 4%;

- за системою Unicheck: 10.7.

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 Боярчука Іллі Олександровича*

за темою *Метод виявлення проявів етнічної ворожнечі у текстових повідомленнях соціальних інтернет-мереж NLP-засобами*

1. Актуальність теми

Виявлення проявів етнічної ворожнечі є важливою суспільною проблемою, оскільки вона може призвести до серйозних наслідків, таких як конфлікти та суперечки між різними етнічними групами. Оскільки ці суперечки найчастіше виникають на інтернет-платформах, використання NLP-засобів для їх ідентифікації є вкрай актуальним.

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

За стандартом, а саме описом предметної області, об'єктом дослідження є процес виявлення проявів етнічної ворожнечі у текстових повідомленнях соціальних інтернет-мереж NLP-засобами. Метою роботи є спрощення експертизи виявлення проявів етнічної ворожнечі за рахунок автоматизованого її виявлення у текстових повідомленнях соціальних інтернет-мереж. При вирішенні поставленої задачі використано методи та технології штучного інтелекту для виявлення проявів етнічної ворожнечі. Отже, результати кваліфікаційної роботи бакалавра повністю відповідають стандарту бакалавра спеціальності 122 – Комп'ютерні науки.

3. Професійні та особистісні якості бакалавра

При виконанні кваліфікаційної роботи бакалавра Боярчук Ілля Олександрович проявив себе як дисциплінований студент з високим рівнем самостійності. Також показав достатні вміння та навики в розробці програмного забезпечення заданої теми, що дозволило отримати відмінні результати.

4. Ступінь самостійності під час виконання кваліфікаційної роботи

Студент самостійно виконував усі завдання під час виконання кваліфікаційної роботи, тому результати роботи є особистим надбанням студента.

5. Ступінь оволодіння методами дослідження

Під час роботи над кваліфікаційною роботою Боярчук Ілля Олександрович показав високий рівень володіння методами дослідження в області комп'ютерних наук та штучного інтелекта.

6. Повнота та якість розкриття теми роботи

Автор кваліфікаційної роботи бакалавра повністю розкрив мету та завдання, здійснив аналіз сучасного стану предметної області та огляд існуючих рішень і підходів. Це дозволило розробити метод виявлення проявів етнічної ворожнечі у текстових повідомленнях соціальних інтернет-мереж NLP-засобами, а також провести дослідження його ефективності з використанням створеного програмного забезпечення.

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

Кваліфікаційна робота бакалавра відзначається чіткою структурою, що сприяє послідовному та логічному викладенню матеріалу з наведеним аргументуванням. Також слід відзначити високий рівень наукової грамотності автора.

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

Розроблена інформаційна система для виявлення проявів етнічної ворожнечі в текстових повідомленнях у соціальних інтернет-мережах має значний потенціал для практичного застосування. Автоматизоване виявлення таких проявів може бути корисним для органів правопорядку, дослідників етнічних конфліктів, аналітиків та соціологів, які займаються вивченням і аналізом етнічної ворожнечі в інтернет-просторі. Крім того, система може сприяти вчасному реагуванню на такі прояви, запобігаючи ескалації конфліктів і забезпечуючи кращий моніторинг громадського настрою.

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 «**добре**» .*

Керівник _____ *викладач кафедри КН Марина Молчанова*

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
МОН УКРАЇНИ

Кафедра комп'ютерних наук

РЕЦЕНЗІЯ

на кваліфікаційну роботу бакалавра

студента *гр. КНс-21-1 Боярчук Ілля Олександрович*

за темою: *Метод виявлення проявів етнічної ворожнечі у текстових повідомленнях соціальних інтернет-мереж NLP-засобами*

1. Актуальність обраної теми

Виявлення проявів етнічної ворожнечі є важливою проблемою для суспільства, оскільки вона може призвести до серйозних наслідків, включаючи конфлікти та розбрати між різними етнічними групами. Такі розбрати найчастіше виникають у інтернет-сервісах, тому використання засобів NLP є доцільним.

2. Повнота розкриття мети та завдань роботи

Автор кваліфікаційної роботи бакалавра повністю розкрив мету та завдання. В роботі проаналізовано сучасний стан предметної області виявлення етнічної ворожнечі в соціальних мережах NLP-засобами, виконано огляд існуючих рішень та підходів, що дало змогу розробити метод виявлення проявів етнічної ворожнечі у текстових повідомленнях, а також провести дослідження ефективності з використанням розробленого програмного забезпечення.

3. Зміст кожного розділу роботи

Розділи роботи мають актуальну інформацію, що стосуються теми кваліфікаційної роботи бакалавра. У першому розділі надано характеристику предметної області, поставлено мету, що полягає в спрощенні експертизи виявлення проявів етнічної ворожнечі за рахунок автоматизованого її виявлення у текстових повідомленнях соціальних інтернет-мереж. Також сформовано задачі роботи. У другому розділі описано проєктування інформаційної системи виявлення проявів етнічної ворожнечі у текстових повідомленнях соціальних інтернет-мереж, створено відповідний метод. У третьому розділі проведено експериментальне дослідження методу виявлення проявів етнічної ворожнечі у текстових повідомленнях соціальних інтернет-мереж.

4. Оцінка розробленої інформаційної системи, її практична цінність

Розроблена інформаційна система виявлення проявів етнічної ворожнечі у текстових повідомленнях соціальних інтернет-мереж має високий потенціал застосування. Автоматизоване виявлення проявів етнічної ворожнечі може бути використане органами правопорядку, дослідниками етнічних конфліктів, аналітиками та соціологами, які цікавляться виявленням та аналізом етнічної ворожнечі у текстових повідомленнях соціальних мереж.

5. Якість оформлення кваліфікаційної роботи бакалавра

Автор якісно оформив кваліфікаційну роботу бакалавра. Робота містить необхідні розділи, ілюстративні матеріали для наочності, такі як таблиці та графіки. Використана література підтверджує обґрунтованість висновків, а послідовне викладення матеріалу забезпечує легке сприйняття тексту.

6. Недоліки кваліфікаційної роботи бакалавра

Суттєвих недоліків кваліфікаційна робота бакалавра немає. Було б доцільно у інтерфейсі користувача додати можливість локалізації для української мови. По тексту пояснювальної записки виявлено незначну кількість пунктуаційних помилок. Втім, наведене вище не впливає на загальну якість роботи та отримані результати.

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

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

Рецензент к.т.н., доц. каф. КІС Ніченорук П.О. _____