Хмельницький національний університет
Факультет інформаційних технологій
Кафедра комп'ютерних наук

КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА

на тему *Метод нейромережевої ідентифікації гендерної дискримінації у соціальних медіа за мультимодальними даними*

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

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

Освітня програма *Комп'ютерні науки*
Назва освітньої програми

Виконав: *студент групи КН-22-1* (Група виконавця) _____ (Підпис) *Даніїл ШАШОК* (Ім'я, ПРІЗВИЩЕ)

Керівник: *д-р філ., ст.викл. 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)
18 червня 2026 р.

Хмельницький 2026

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
«*22*» *Січня* *2026* року

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ БАКАЛАВРА

1. Тема кваліфікаційної роботи бакалавра: *«Метод нейромережевої ідентифікації гендерної дискримінації у соціальних медіа за мультимодальними даними»*

2. Завдання видано студенту *Даниілу Шашку*
(Ім'я, прізвище)

3. Керівник роботи *д-р філ., ст.викл. каф. КН Марина МОЛЧАНОВА*
(посада, ім'я, прізвище)

4. Затверджено наказом університету від «*20*» *Січня* *2026* р. № *7*

5. Дата видачі завдання студенту: «*22*» *Січня* *2026* р.

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

Мета роботи – підвищення точності виявлення проявів гендерної дискримінації у мультимодальному контенті соціальних медіа шляхом розроблення методу нейромережевої ідентифікації сексизму та відповідної інтелектуальної системи. Для досягнення мети слід виконати такі задачі: провести аналіз предметної області виявлення проявів гендерної дискримінації у мультимодальному контенті соціальних медіа; розробити метод нейромережевої ідентифікації гендерної дискримінації у мультимодальному контенті соціальних медіа; здійснити програмну реалізацію інтелектуальної системи, що забезпечить роботу розробленого методу для аналізу мультимодального контенту на предмет виявлення гендерної дискримінації; здійснити дослідження розробленого методу ідентифікації проявів гендерної дискримінації з використанням створеної інтелектуальної системи за допомогою метрик якості.

7. Календарний план виконання кваліфікаційної роботи бакалавра:

№	Назва етапів (розділів) кваліфікаційної роботи бакалавра	Термін виконання	Примітка
1	*Вибір напряму дослідження та узгодження теми кваліфікаційної роботи з керівником, складання календарного графіка виконання*	*січень 2026*	виконано
2	*Ознайомлення з предметною областю, формулювання мети і задач дослідження, визначення об'єкта та предмета дослідження*	*лютий 2026*	виконано
3	*Проєктування методу розв'язання задачі, опис архітектурних рішень, розроблення математичних моделей та алгоритмів.*	*березень 2026*	виконано
4	*Обґрунтування інструментарію розробки, програмна реалізація розробленого методу, проведення експериментального тестування та оцінювання ефективності.*	*квітень 2026*	виконано
5	*Написання тексту кваліфікаційної роботи, урахування зауважень керівника, оформлення згідно з вимогами*	*травень 2026*	виконано
6	*Розроблення презентаційних матеріалів та попередній захист кваліфікаційної роботи*	*травень 2026*	виконано
7	*Отримання відгуку керівника, рецензії, перевірка тексту кваліфікаційної роботи на плагіат, нормоконтроль*	*червень 2026*	виконано
8	*Підготовка до захисту та захист кваліфікаційної роботи*	*червень 2026*	виконано

Виконавець: *студент групи КН-22-1* _____ *Данііл ШАШОК*
Група виконавця — Підпис — Ім'я, ПРІЗВИЩЕ

Керівник: *д-р філ., ст.викл. каф. КН* _____ *Марина МОЛЧАНОВА*
Науковий ступінь, посада — Підпис — Ім'я, ПРІЗВИЩЕ

Анотація

Тема кваліфікаційної роботи бакалавра: *«Метод нейромережевої ідентифікації гендерної дискримінації у соціальних медіа за мультимодальними даними»*

Виконавець кваліфікаційної роботи бакалавра: *студент групи КН-22-1 Даниіл Шашок*

Керівник кваліфікаційної роботи бакалавра: *д.філ., ст.викл. каф. КН Марина Молчанова*

Кваліфікаційна робота бакалавра містить:

Пояснювальна записка				Кількість додатків
Сторінок	Рисунків	Таблиць	Джерел інформації	
69	24	3	42	2

Метою кваліфікаційної роботи є підвищення точності виявлення проявів гендерної дискримінації у мультимодальному контенті соціальних медіа шляхом розроблення методу нейромережевої ідентифікації сексизму. Для програмної реалізації інтелектуальної системи використано мову програмування Python, фреймворк PyTorch і бібліотеку Transformers, на базі яких реалізовано нейромережеві архітектурні моделі BiLSTM та RoBERTa для класифікації текстових даних. Метод орієнтований на використання у системах модерації контенту, тоді як інтелектуальна система – на модераторів соціальних платформ.

Напрямами практичного використання розробленої інтелектуальної системи є модерація користувацьких повідомлень та моніторинг мультимодального контенту на предмет наявності гендерно дискримінаційних висловлювань.

Ключові слова: точність, гендерна дискримінація, BiLSTM, RoBERTa, мультимодальність, соціальні медіа.

Виконавець: *студент групи КН-22-1* (Група виконавця) _____ (Підпис) *Даниіл ШАШОК* (Ім'я, ПРІЗВИЩЕ)

Зміст

Перелік скорочень

Скорочення, термін, позначення	Пояснення
КРБ	Кваліфікаційна робота бакалавра
ШІ	Штучний інтелект
ІТ	Інформаційні технології
МН	Машинне навчання
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
BiLSTM	Bidirectional Long Short-Term Memory
BERT	Bidirectional Encoder Representations from Transformers
RoBERTa	Robustly Optimized BERT Pretraining Approach
NLP	Natural Language Processing
API	Application Programming Interface
SDET	Sexism Detection in English Text
SSMB	Sexism Social Media Balanced
LLM	Large Language Model
ASR	Automatic Speech Recognition
VS Code	Visual Studio Code

Вступ

Кваліфікаційна робота присвячена підвищенню точності виявлення проявів гендерної дискримінації у мультимодальному контенті соціальних медіа за рахунок розроблення методу нейромережевої ідентифікації гендерної дискримінації.

Актуальність. Стрімкий розвиток цифрових платформ спричинив значне зростання обсягів мультимодального контенту, який поєднує текст, аудіо та відео. Соціальні мережі стали важливим середовищем комунікації та поширення інформації, однак разом із цим створили умови для активного розповсюдження дискримінаційних висловлювань, зокрема проявів гендерної дискримінації. Ручна модерація таких матеріалів є трудомісткою та недостатньо масштабованою в умовах постійного зростання обсягів контенту. У зв'язку з цим підвищення точності виявлення проявів гендерної дискримінації у мультимодальному контенті соціальних медіа є актуальним науково-практичним завданням у сфері ІТ.

Технології обробки природної мови та моделі глибокого навчання демонструють значний прогрес у задачах аналізу тексту, зокрема у виявленні мови ворожнечі, тональності та семантичних патернів дискримінаційного характеру. Більшу ефективність у задачах класифікації текстових послідовностей демонструють рекурентні нейронні мережі та трансформерні архітектури, що враховують контекст висловлювань та виявляють приховані форми сексизму, виражені через сарказм, стереотипні узагальнення або непрямі дискримінаційні конструкції. Це зумовлює потребу в розробленні методу для підвищення точності ідентифікації таких проявів.

Об'єкт дослідження – процес нейромережевої ідентифікації гендерної дискримінації у мультимодальному контенті соціальних медіа.

Предмет дослідження – нейромережеві засоби, зокрема рекурентні й трансформерні нейромережеві архітектури, для аналізу та класифікації текстових представлень мультимодальних даних.

Метою кваліфікаційної роботи бакалавра є підвищення точності виявлення проявів гендерної дискримінації у мультимодальному контенті

соціальних медіа за рахунок розроблення методу нейромережевої ідентифікації гендерної дискримінації, а також відповідної інтелектуальної системи.

Завдання кваліфікаційної роботи бакалавра – провести аналіз предметної області виявлення проявів гендерної дискримінації у мультимодальному контенті соціальних медіа; розробити метод нейромережевої ідентифікації гендерної дискримінації у мультимодальному контенті соціальних медіа; здійснити програмну реалізацію інтелектуальної системи, що забезпечить роботу розробленого методу для аналізу мультимодального контенту на предмет виявлення гендерної дискримінації; здійснити дослідження розробленого методу ідентифікації проявів гендерної дискримінації з використанням створеної інтелектуальної системи за допомогою метрик якості.

Розділ 1 Характеристика предметної області: аналіз моделей, методів та реалізацій

1.1 Аналіз інформаційних моделей

Сучасні соціальні медіа стали одним із ключових середовищ формування суспільної думки, міжособистісної комунікації та публічного обговорення соціально значущих тем [1]. Платформи для обміну текстовими повідомленнями, зображеннями та відео, зокрема Telegram, Instagram та інші популярні соціальні сервіси, щодня генерують мільярди публікацій і забезпечують безпрецедентну швидкість поширення інформації в глобальному масштабі [2]. Разом із позитивними аспектами відкритості та доступності такі середовища створюють умови для поширення мови ворожнечі, дискримінаційних висловлювань і упереджених наративів. Однією з найбільш поширених форм дискримінаційного контенту є гендерна дискримінація, що проявляється у приниженні, знеціненні або стереотипізації осіб за ознакою статі чи гендерної ідентичності.

Поняття гендерної дискримінації визначається як обмеження прав або можливостей людини на підставі її належності до певної статі чи гендерної групи [3]. У міжнародних документах і правових актах наголошується на необхідності запобігання таким проявам у всіх сферах суспільного життя, включно з інформаційним простором. У цифровому середовищі дискримінація часто набуває форми сексизму – системи поглядів і практик, що поєднує відкрито негативні установки стосовно жінок (ворожий сексизм) та нібито позитивні, але фактично шкідливі форми ставлення (доброзичливий сексизм) [4]. До проявів сексизму належать образливі висловлювання, приписування негативних якостей за гендерною ознакою, об'єктивація, заклики до обмеження прав або соціальної ролі певної групи.

Соціальні мережі та платформи обміну контентом функціонують як багатокористувацькі інформаційні системи. У їх межах можна виокремити кілька основних груп учасників: звичайні користувачі, що створюють та публікують контент; модератори, які здійснюють перевірку матеріалів на

відповідність правилам спільноти; адміністратори, відповідальні за технічну підтримку та політику безпеки; а також аналітики або дослідники, що вивчають інформаційні потоки та поведінкові моделі. Основні функції таких систем включають створення, зберігання та поширення повідомлень, коментування, реакції, обмін мультимедійними файлами, а також механізми скарг на неприйнятний контент. Масштаб сучасних платформ робить ручну перевірку всього контенту практично неможливою, що зумовлює потребу в засобах ІТ для автоматизованої підтримки модераційних процесів [5].

У контексті досліджуваної теми ключовим об'єктом аналізу виступає користувацький контент. Він може бути представлений у текстовій формі (пости, коментарі, повідомлення), в аудіоформаті (подкасти, голосові повідомлення) або як відеофайл. Важливою характеристикою сучасного інформаційного простору є мультимодальність – поєднання різних форматів подання інформації (тексту, аудіо, відео) в межах одного комунікаційного середовища. Кожен із цих форматів може виступати самостійним носієм змісту та потребує специфічних ІТ-методів обробки й аналізу. Така особливість ускладнює виявлення дискримінаційних проявів. Разом із тим гендерна дискримінація має специфічні лінгвістичні та семантичні маркери. Дослідження [6] свідчить, що сексизм в онлайн-середовищі охоплює широкий спектр проявів – від відвертих форм ненависті до значно тонших, прихованих форм упередженості, зокрема через узагальнюючі негативні судження щодо певної статі, знецінення соціальних ролей та об'єктивацію. Часто такі висловлювання подаються у прихованій формі – через іронію, сарказм або стереотипні шаблони, що ускладнює їх однозначну інтерпретацію.

До основних процесів, пов'язаних із функціонуванням інформаційного середовища, належать: створення контенту, його публікація, поширення та модерація. На етапі модерації застосовуються як ручні методи перевірки, так і автоматизовані ІТ-інструменти фільтрації [7]. Ручна модерація передбачає залучення персоналу, який аналізує повідомлення відповідно до внутрішніх політик платформи. Такий підхід забезпечує гнучкість і врахування контексту,

проте потребує значних ресурсів і не гарантує оперативності реагування при великих обсягах даних. Автоматизовані системи обробляють великі масиви інформації в режимі реального часу, однак потребують чітко визначених критеріїв і якісних навчальних вибірок.

Особливістю гендерної дискримінації в онлайн-середовищі є її масштабованість і швидкість поширення. Один дискримінаційний допис може бути реплікований тисячами користувачів протягом короткого часу, що посилює негативний вплив на цільові групи. Дослідження [8] доводить, що сексистський контент у соціальних медіа безпосередньо знижує психологічне благополуччя жінок та їхню готовність до участі в онлайн-дискурсі, а ефективність ІТ інструментів платформ для нейтралізації цього впливу залишається недостатньо вивченою. У зв'язку з цим своєчасне виявлення та обмеження поширення такого контенту набуває пріоритетного значення.

Аналіз предметної області свідчить про наявність кількох актуальних проблем. По-перше, складно однозначно формалізувати мовні маркери дискримінації через багатозначність природної мови. По-друге, необхідно обробляти великі обсяги мультимедійних даних, які постійно оновлюються. По-третє, відсутні універсальні ІТ-підходи до оцінювання рівня ризику та класифікації контенту. Подолання цих труднощів потребує поєднання різноманітних підходів із використанням сучасних методів ШІ та ІТ.

Отже, предметна область охоплює створення й поширення мультимодального контенту в соціальних медіа, а також механізми його модерації та виявлення проявів гендерної дискримінації. Актуальність дослідження зумовлена масштабністю проблеми та потребою в точних методах аналізу на основі сучасних ІТ і ШІ.

1.2 Огляд теоретичних підходів до розв’язку подібних задач

Стрімке зростання обсягів користувацького контенту в соціальних медіа спричинило активний розвиток ІТ інструментарію для автоматизованого аналізу

текстових і мультимедійних даних [9]. У задачах виявлення дискримінаційного або образливого контенту поєднуються підходи обробки природної мови, МН та статистичного аналізу даних. Актуальними напрямами застосування технологій ІТ у цій сфері є автоматична класифікація тексту, розпізнавання мовлення та оцінювання ризиків поширення дискримінаційного контенту.

Центральне місце серед зазначених підходів займають нейронні мережі як інструмент побудови моделей, здатних виявляти складні залежності у даних. Нейронна мережа – це математична модель, що складається з взаємопов'язаних обчислювальних елементів (нейронів), організованих у шари. Машинне навчання передбачає побудову таких моделей, здатних виявляти закономірності в даних на основі навчальної вибірки без явного програмування правил. Залежно від архітектури виділяють повнозв'язні мережі, згорткові нейронні мережі (CNN), рекурентні нейронні мережі (RNN) та трансформерні архітектури; у задачах бінарної класифікації тексту застосовуються як класичні алгоритми (логістична регресія, метод опорних векторів), так і нейромережеві підходи [10].

Для аналізу текстових послідовностей особливого поширення в задачах ІТ-систем набули рекурентні нейронні мережі та їх модифікації. Архітектура LSTM дозволяє зберігати інформацію про попередні елементи послідовності протягом тривалого контексту, що є критичним для розуміння складних мовних конструкцій. Використання двонаправленої архітектури BiLSTM дає змогу аналізувати текст одночасно у прямому та зворотному напрямках, що підвищує повноту врахування контексту та якість класифікації [11].

Альтернативним підходом до моделювання мовних залежностей є трансформерні моделі, засновані на механізмі самоуваги (Self-Attention). На відміну від RNN, вони формують контекстне представлення кожного слова з урахуванням усіх інших слів у послідовності одночасно, що забезпечує вищу точність у задачах семантичного аналізу. Модель BERT забезпечує глибоке двонаправлене контекстне подання тексту, а RoBERTa вдосконалює процедуру попереднього навчання через динамічне маскування токенів і ширші тренувальні корпуси. Дослідження [12] підтверджує, що трансформерні моделі, зокрема

RoBERTa, демонструють вищу точність порівняно з класичними методами при виявленні мови ворожнечі та дискримінаційного контенту в соціальних медіа, хоча й характеризуються значними вимогами до обчислювальних ресурсів.

Оскільки контент у соціальних медіа представлений не лише у текстовій формі, актуальним є застосування IT-систем автоматичного розпізнавання мовлення ASR. Такі системи базуються на нейронних архітектурах глибинного навчання та забезпечують перетворення аудіосигналу в текстову форму для подальшого аналізу. Інтеграція ASR забезпечить автоматичне перетворення мовлення у текстове представлення, що дозволить здійснювати аналіз мультимодального контенту, зокрема відеоматеріалів [13].

Отже, для реалізації методу планується використання нейромережевого підходу на основі порівняння рекурентної архітектури BiLSTM та трансформерної моделі RoBERTa для бінарної класифікації тексту за ознакою наявності гендерної дискримінації. Для роботи з аудіо та відео передбачено застосування IT-системи автоматичного розпізнавання мовлення з метою перетворення усного мовлення у текстову форму.

1.3 Аналіз існуючих програмних засобів та наукових рішень

У межах дослідження предметної області доцільним є аналіз існуючих програмних рішень та інтелектуальних систем, призначених для автоматичного виявлення мови ворожнечі та інших форм дискримінаційного контенту. Вивчення функціональних можливостей і технологічних особливостей таких сервісів допомагає визначити сучасний стан розвитку засобів автоматизованої модерації, виявити їхні переваги та недоліки, а також окреслити напрями подальшого вдосконалення. З огляду на тематику роботи, особливої уваги потребують веб-орієнтовані інструменти, що здійснюють аналіз текстових даних із використанням методів МН та ШІ, оскільки їх функціональні та технологічні рішення є найбільш релевантними для розробки власної системи.

Також важливим джерелом теоретичних і методологічних даних є наукові публікації, у яких досліджуються прояви гендерної дискримінації в цифровому середовищі, механізми її поширення та соціальні наслідки. Аналіз наукових статей допомагає глибше зрозуміти природу дискримінаційних практик, визначити їх типологію та обґрунтувати доцільність створення автоматизованих засобів їх виявлення.

Одним із інструментів для автоматичної ідентифікації мови ворожнечі є модель Cardiff NLP Twitter-RoBERTa Hate Speech, розроблена дослідницькою групою CardiffNLP та розміщена на платформі Hugging Face. Модель є донавченою версією трансформера RoBERTa, натренованого на текстах із Twitter. Вона навчалася на 13 англомовних датасетах мови ворожнечі, що покращило її узагальнювальну здатність і стійкість до кросдатасетного зміщення [14]. Платформа Hugging Face надає інтерфейс для введення тексту та отримання результату бінарної класифікації (рисунок 1.1).

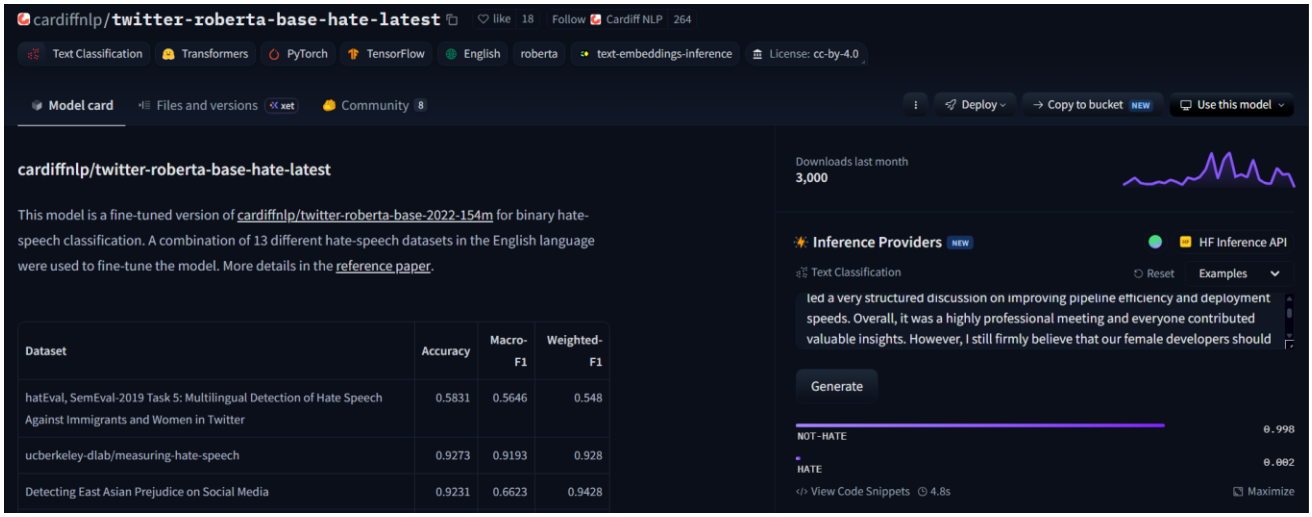


Рисунок 1.1 – Результат роботи моделі на платформі Hugging Face [14]

Переваги:

– навчання на 13 датасетах для кращого розуміння мови ворожнечі;

– відкритий доступ до моделі та можливість використання через API;

– інтерактивний інтерфейс для тестування на платформі Hugging Face.

Недоліки:

– відсутність підтримки мультимодального контенту;

– немає деталізованого аналізу по реченнях або фрагментах тексту.

Другим прикладом є Kimola Hate Speech Detector – платформа з готовими, навченими моделями для розпізнавання токсичного та дискримінаційного контенту, зокрема проявів сексизму, ксенофобії і мови ненависті. Сервіс аналізує текстові дані на основі алгоритмів МН, формуючи числову оцінку рівня ризику та визначаючи категорію за типом виявленого порушення. Отримані результати можуть використовуватися як для ручної модерації, так і для автоматизованих систем фільтрації контенту [16]. На рисунку 1.2 показано приклад роботи системи: введення текстового фрагмента, запуск процедури аналізу та відображення результату числовим значенням ймовірності токсичності разом із відповідною інтерпретацією.

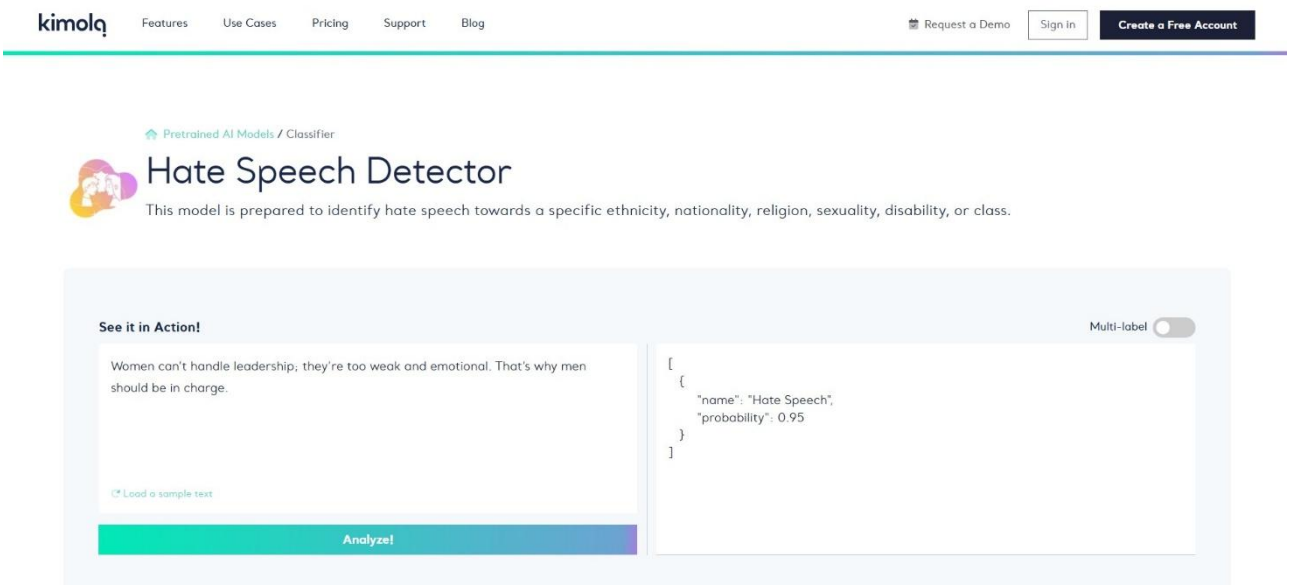


Рисунок 1.2 – Онлайн платформа Kimola Hate Speech Detector [17]

Переваги:

– готові навчені моделі для різних типів дискримінації;

– зручний веб-інтерфейс;

– можливість інтеграції у власні програми через API.

Недоліки:

– платформа обмежена у безкоштовному доступі;

– відсутня підтримка мультимодального контенту (аудіо, відео);

– не надає деталізованого аналізу за реченнями чи вікнами тексту.

Останнім прикладом є прототип Portuguese Hate Speech Detection від knowhate. Дана система спеціалізується на виявленні мови ворожнечі

португальською мовою та реалізована, як демонстраційний застосунок на платформі Hugging Face. Користувач має можливість ввести текстовий фрагмент і отримати результат бінарної класифікації щодо наявності або відсутності токсичних чи дискримінаційних елементів. Модель орієнтована на конкретну мовну вибірку, що дозволяє підвищити точність визначення контекстуальних маркерів мови ненависті [18]. На рисунку 1.3 продемонстровано роботу системи у середовищі Hugging Face: поле для введення тексту, кнопку запуску аналізу та відображення числової оцінки і відповідної категорії.

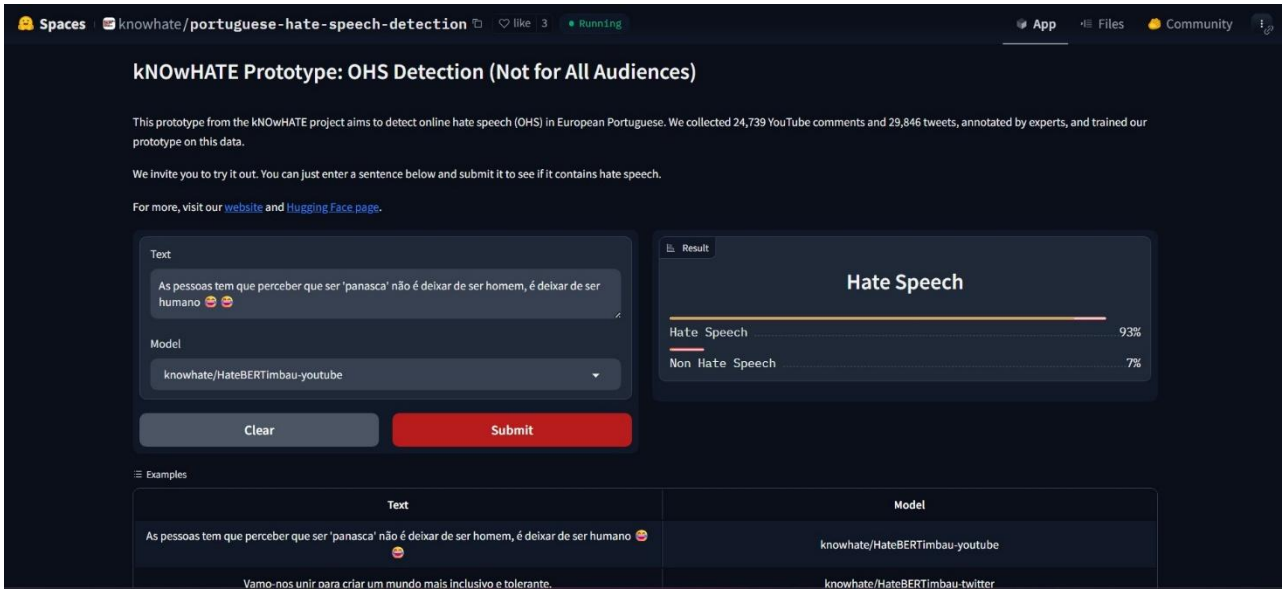


Рисунок 1.3 – Система PHS Detection у середовищі Hugging Face [19]

Переваги:

– орієнтація на конкретну мову для підвищення точності;
– простий та легкий у використанні інтерфейс;
– інтеграція з ресурсом Hugging Face для тестування моделей.

Недоліки:

– обмежена підтримка інших мов;
– відсутність підтримки мультимодального контенту;
– мінімальна деталізація результату.

Окрім розглянутих програмних рішень, важливе значення для формування теоретико-методологічної основи дослідження мають наукові публікації, у яких аналізуються соціальні, психологічні, комунікативні та технічні аспекти гендерної дискримінації в цифровому середовищі. Аналіз таких

праць дає змогу глибше зрозуміти природу сексизму в соціальних медіа, особливості його прояву та сучасні підходи до автоматизованого виявлення дискримінаційного контенту.

У науковій публікації [**Помилка! Джерело посилання не знайдено.**] досліджується вплив цифрових технологій та соціальних платформ на поширення гендерної нерівності у професійному середовищі. Автори аналізують механізми онлайн-комунікації та репрезентації користувачів, зосереджуючись на досвіді жінок у сферах роздрібної торгівлі, цифрових послуг та адміністративної діяльності. Дослідження виконано з використанням описового якісного підходу на основі глибинних інтерв'ю, спостережень і аналізу цифрових документів. У роботі показано, що соціальні медіа стали новими каналами гендерної дискримінації через нав'язливі повідомлення, сексистські коментарі та несанкціоноване поширення особистих матеріалів. Автори також наголошують на необхідності впровадження цифрової етики у корпоративні політики задля створення безпечного середовища для жінок.

У статті [21] розглядається проблема гендерної дискримінації в середовищі нових медіа на прикладі платформи TikTok. Автори аналізують контент і механізми цифрового дискурсу, демонструючи, що декларована відкритість соціальних мереж не усуває структурні прояви гендерної нерівності. У результаті дослідження встановлено, що жінки часто стикаються зі стигматизацією через інтернет-сленг, об'єктивацією зовнішності та комерціалізацією образу. Окрему увагу приділено ролі анонімності користувачів і алгоритмів рекомендацій у підтримці та поширенні гендерних стереотипів.

Комплексний медіапсихологічний аналіз гендерних репрезентацій, упереджень і дискримінації в соціальних мережах представлено у публікації [22]. Робота складається з кількох взаємопов'язаних досліджень, у межах яких аналізується вплив медійних образів на гендерну соціалізацію, поширення прихованих форм сексизму, зокрема доброзичливого сексизму, а також використання гендерно маркованої мови в YouTube-контенті. Окремо

досліджується представленість гендерних меншин у соціальних медіа та вплив цифрової репрезентації на суспільне сприйняття. Результати роботи підкреслюють необхідність підвищення різноманітності гендерних образів і розробки інструментів зменшення дискримінаційних практик.

Значний інтерес для розробки систем модерації становлять дослідження, присвячені застосуванню сучасних моделей машинного навчання для автоматичного виявлення гендерної дискримінації. У науковій публікації [23] розглядається проблема стрімкого зростання сексистського контенту в соціальних медіа та його негативного психологічного впливу на жінок. Автори пропонують автоматизовану платформу, засновану на моделях машинного та глибокого навчання, яка дозволяє виявляти дискримінаційні висловлювання з високою точністю. Особливу увагу приділено використанню методів пояснювального штучного інтелекту, що дає змогу не лише класифікувати текст, а й інтерпретувати причини прийняття рішення моделлю.

Перспективним напрямом є також масштабовані підходи до автоматизованої модерації контенту на великих онлайн-платформах. У статті [24] обґрунтовується необхідність застосування мультимодальних систем фільтрації, здатних аналізувати як текстову, так і візуальну інформацію. Автори підкреслюють, що образливий контент часто поширюється у комбінованій формі, поєднуючи зображення, меми та текстові коментарі, що ускладнює виявлення за допомогою однотипного аналізу. Запропонований підхід на основі трансформерних архітектур демонструє високу точність у задачах масштабованої модерації.

Важливим аспектом оцінки сучасних систем автоматизованої модерації є аналіз алгоритмічних упереджень моделей. У дослідженні [**Помилка! Джерело посилання не знайдено.**] виявлено, що навіть моделі з високими показниками точності, зокрема BERT і RoBERTa, можуть демонструвати приховану гендерну упередженість при класифікації семантично подібних висловлювань із різними гендерними маркерами. Автори наголошують, що такі системні відхилення знижують надійність інструментів автоматичного аналізу контенту та

потребують впровадження додаткових процедур аудиту й контролю справедливості моделей.

Таким чином, дослідження наявних програмних засобів, демонстраційних прототипів та наукових праць підтверджує актуальність підвищення точності ідентифікації дискримінаційного контенту в соціальних медіа. Розглянуті системи здатні виявляти явні прояви мови ворожнечі у текстових даних, проте не підтримують аналіз аудіо- чи відеоматеріалів та не надають деталізованого контекстуального розбору текстових фрагментів по реченнях. Складний і часто прихований характер гендерної дискримінації, зокрема доброзичливий сексизм, вимагає врахування контекстуальних зв'язків між реченнями та обробки мультимодального контенту. Це зумовлює необхідність розроблення методу нейромережевої класифікації та відповідної інтелектуальної системи для комплексного мультимодального аналізу.

1.4 Мета, задачі та вимоги до реалізації інтелектуальної системи

Метою кваліфікаційної роботи є підвищення точності виявлення проявів гендерної дискримінації у мультимодальному контенті соціальних медіа шляхом розроблення методу нейромережевої ідентифікації сексизму на основі архітектур глибокого навчання, а також програмної реалізації інтелектуальної системи для кількісного та якісного оцінювання досягнутих показників точності.

Для досягнення поставленої мети необхідно виконати такі задачі:

– провести аналіз предметної області виявлення проявів гендерної дискримінації у мультимодальному контенті соціальних медіа;

– розробити метод нейромережевої ідентифікації гендерної дискримінації у мультимодальному контенті соціальних медіа;

– здійснити програмну реалізацію інтелектуальної системи, що забезпечить роботу розробленого методу для аналізу мультимодального контенту на предмет виявлення гендерної дискримінації;

– здійснити дослідження розробленого методу ідентифікації проявів гендерної дискримінації з використанням створеної інтелектуальної системи за допомогою метрик якості.

Розроблювана інтелектуальна система повинна забезпечувати аналіз тексту, аудіо та відео, здійснювати бінарну класифікацію контенту за ознакою наявності гендерної дискримінації, підтримувати обробку великих текстових фрагментів, а також надавати інтерпретовані результати аналізу.

1.5 Висновки до розділу 1

У розділі було виконано дослідження предметної області виявлення проявів гендерної дискримінації у мультимодальному контенті соціальних медіа. Охарактеризовано основні форми сексизму та визначено особливості їхнього прояву в онлайн-середовищі. Підтверджено актуальність підвищення точності аналізу такого контенту із використанням МН та ШІ. Також проаналізовано сучасні підходи до виявлення дискримінаційного контенту та засоби автоматичного розпізнавання мовлення у контексті мультимодальної обробки даних. За результатами аналізу існуючих програмних рішень визначено їхні функціональні переваги та недоліки. Розглянуті наукові роботи вказують на різноманітність проявів сексизму в соціальних медіа та доцільність використання автоматизованих засобів його виявлення на основі сучасних моделей машинного навчання.

На основі проведеного аналізу сформульовано вимоги до інтелектуальної системи та встановлено доцільність розроблення методу нейромережевої ідентифікації гендерної дискримінації у мультимодальному контенті соціальних медіа.

Розділ 2 Метод нейромережевої ідентифікації гендерної дискримінації у соціальних медіа за мультимодальними даними

2.1 Математична формалізація задачі нейромережевої ідентифікації гендерної дискримінації за мультимодальними даними

Ідентифікація гендерної дискримінації в цифровому середовищі є нетривіальною задачею з огляду на варіативність, неструктурованість та мультимодальну природу контенту соціальних медіа. Явище сексизму може проявлятися у текстових повідомленнях, відеозаписах та аудіофайлах, що ускладнює застосування уніфікованих підходів до його виявлення. Для вирішення цієї проблеми планується проєктування методу $\mathcal{M}$ нейромережевої ідентифікації гендерної дискримінації. Метод буде формалізовано як послідовність перетворень, що забезпечуватимуть відображення вхідних мультимодальних даних D у семантичну мітку рівня дискримінації L:

$$\mathcal{M}: D \xrightarrow{\mathcal{T}} X \xrightarrow{f} P \xrightarrow{Aggr} L, \tag{2.1}$$

де $D = \{D_{txt}, D_{aud}, D_{vid}\}$ – множина вхідних даних (текст, аудіо, відео); $\mathcal{T}$ є оператором уніфікації модальностей, що забезпечуватиме транскрибування та нормалізацію вхідних даних до єдиного текстового представлення X; f – функція нейромережевої класифікації текстових послідовностей; P – множина локальних оцінок ризику; $Aggr$ – оператор статистичної агрегації локальних оцінок ризику. Таким чином, нейромережева класифікація виступатиме центральним етапом методу, тоді як оператор $\mathcal{T}$ та блок $Aggr$ забезпечуватимуть адаптацію до специфіки мультимодальних даних.

Оскільки базовою одиницею семантичного аналізу виступає текстова послідовність, вхідні аудіо- та відеодані попередньо проходитимуть етап автоматичного розпізнавання мовлення, після чого задача зводитиметься до бінарної класифікації тексту. На етапі класифікації текстове представлення X подаватиметься на функцію нейромережевої класифікації f, яка відображатиме простір вхідних текстів у бінарну множину міток класів:

$$f: X \rightarrow \{0, 1\}, \tag{2.2}$$

де 1 відповідає класу «сексизм», а 0 – класу «не сексизм». Нейромережева модель наближатиме цю функцію, повертаючи проміжну ймовірність $p \in [0, 1]$, яка за допомогою порогового правила перетворюватиметься у дискретну мітку класу. Подання (2.1) формує загальну структуру методу та визначає послідовність основних етапів обробки даних.

Важливою особливістю аналізу текстів соціальних медіа є те, що дискримінаційний контент часто має локалізований характер: сексистське висловлювання може займати лише незначну частину документа (наприклад, одне речення у великому абзаці). Пряме застосування класифікатора до всього об'єму тексту призвело б до ефекту «розмиття» семантичних ознак і критичного заниження оцінки ризику. Для вирішення цього буде спроєктовано техніку декомпозиції контексту, що базуватиметься на алгоритмі ковзного вікна [26]. Ця техніка виконуватиме роль структурного фільтра для формування послідовності локальних вхідних блоків для нейромережі f.

Згідно з цим алгоритмом, уніфікований текст $x \in X$ попередньо сегментуватиметься на впорядковану множину речень S:

$$S = \{s_1, s_2, \dots, s_M\}, \tag{2.3}$$

де M – загальна кількість виділених речень у тексті. Сегментація виконуватиметься за розділовими знаками і символами нового рядка.

Над множиною S визначатиметься множина текстових вікон, кожне з яких утворюватиметься шляхом поєднання w сусідніх речень. Вікно зсуватиметься по тексту з заданим кроком d:

$$W_k = \mathrm{s_k} \oplus \mathrm{s_{\{k+1\}}} \oplus \cdots \oplus \mathrm{s_{\{k+w-1\}}}, \tag{2.4}$$

де w – розмір вікна (кількість речень), d – крок зсуву, k – індекс початкового речення вікна. Значення $w = 3$ та $d = 1$ будуть використовуватися, як параметри за замовчуванням, проте ці параметри планується зробити налаштовуваними. Таке перекриття вікон $(d < w)$ гарантуватиме, що жодне речення не буде аналізуватися лише у власному ізольованому контексті.

$$k \in \{1, 1 + d, 1 + 2d, \dots, M - w + 1\} \tag{2.5}$$

Для кожного сформованого вікна W_k класифікатор f незалежно обчислюватиме локальну ймовірність наявності дискримінаційного змісту p_k:

$$p_k = f(W_k) \tag{2.6}$$

Результатом застосування класифікатора до всіх вікон буде масив локальних оцінок $\{p_k\}$, що надалі передаватиметься до блоку $Aggr$ для формування глобального вердикту.

Для врахування контекстуального характеру деяких висловлювань обчислюватиметься показник стабільності ризику r – частка речень, що перевищили поріг ризику 0.5, відносно загального обсягу тексту:

$$r = \frac{|\{j: f(s_j) \geq 0.5\}|}{M}, \tag{2.7}$$

де j – порядковий номер речення з подання (2.3).

Низьке значення r при одночасно помірній максимальній оцінці вказуватиме на поодинокий характер висловлювання, що може бути цитатою або фразою, вирваною з контексту, а не систематичним сексистським контентом. Для математичного моделювання цього явища застосовуватиметься коефіцієнт ізоляції α. Це штрафна функція від показника r та максимального значення $\max(p_k)$:

$$\alpha = \begin{cases} 0.4 + 0.6 * \dfrac{r}{0.15} : \ r < 0.15, \max(p_k) < 0.85 \\ 0.7 + 0.3 * \dfrac{r}{0.30} : \ r < 0.30, \max(p_k) < 0.70 \\ 1.0 : \text{в іншому випадку} \end{cases} \tag{2.8}$$

Коефіцієнт α набуватиме значень від 0.4 (мінімальний штраф при вкрай ізольованому ризику) до 1.0 (відсутність штрафу при системному ризику), забезпечуючи плавне, а не дискретне коригування фінальної оцінки.

Інтегральна оцінка ризику $Score$ для всього документа формуватиметься як зважена комбінація дев'яностого перцентиля P_{90} усього масиву $\{p_k\}$ та максимального значення, скоригованого на коефіцієнт ізоляції:

$$Score = 0.6 * P_{90}(\{p_k\}) + 0.4 * \max(p_k) * \ \alpha \tag{2.9}$$

Застосування P_{90} замість середнього арифметичного або абсолютного максимуму буде доцільним вибором: він стійкий до статистичних викидів (одиничних шумових спрацьовувань), але водночас чутливий до стабільних

ризикованих тенденцій, присутніх у значній частині тексту. Оцінка $Score$ унормується до відрізку $[0, 1]$ і буде кількісним індикатором рівня дискримінаційної загрози у вхідному контенті.

Послідовність основних етапів обробки мультимодальних даних у межах методу нейромережевої ідентифікації гендерної дискримінації формалізовано у вигляді псевдокоду, опис якого наведено в Алгоритмі 2.1.

Алгоритм 2.1 – Математичний псевдокод методу ідентифікації гендерної дискримінації

Вхідні дані: Мультимодальний об'єкт D; навчена модель класифікації f; оператор транскрибування $\mathcal{T}$; параметри вікна w, d.

Вихідні дані: Інтегральна оцінка ризику $Score$; текстова мітка класу L; масив локальних оцінок P.

---

1. Ініціалізувати порожній масив локальних оцінок $P \leftarrow \emptyset$
2. Якщо $D \in D_{vid}$:
   - Вилучити аудіосигнал $D \leftarrow ExtractAudio(D)$
3. Якщо $D \in D_{aud}$:
   - Отримати текстове подання $X \leftarrow \mathcal{T}(D)$
4. Інакше:
   - $X \leftarrow D$
5. Виконати сегментацію тексту X на множину речень S згідно подання (2.3)
6. Для кожного вікна W_k з кроком d (подання 2.4 та 2.5), виконувати:
   - Обчислити локальну ймовірність ризику $p_k \leftarrow f(W_k)$
   - Додати p_k до масиву P
   - Кінець циклу
7. Розрахувати показник стабільності ризику r за поданням (2.7)
8. Обчислити коефіцієнт ізоляції α, як функцію від r та $max\,(P)$ згідно з поданням (2.8)
9. Обчислити інтегральну оцінку $Score$ за формулою (2.9)
10. Визначити мітку L на основі порогових значень параметра $Score$
11. Повернути $Score, L, P$

Взаємодію ключових етапів обробки мультимодальних даних у межах методу, починаючи з надходження вхідного контенту та завершуючи формуванням оцінки ризику, наочно ілюструє рисунок 2.1.

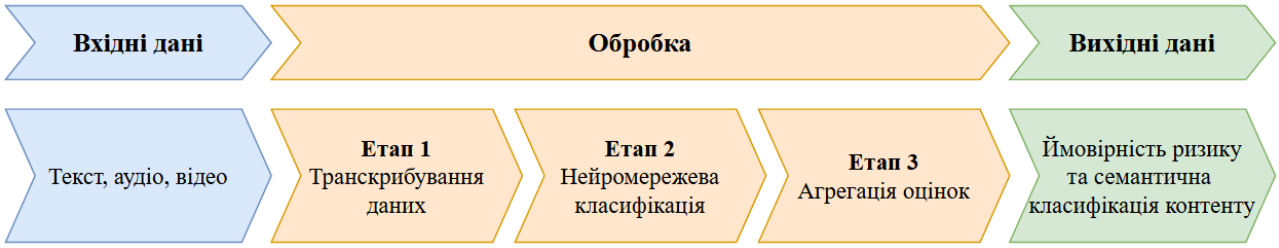


Рисунок 2.1 – Схема обробки мультимодальних даних у методі нейромережевої ідентифікації гендерної дискримінації

Отже, математична формалізація задачі описує метод нейромережевої ідентифікації гендерної дискримінації, як послідовність етапів обробки мультимодальних даних, їх уніфікації до текстового представлення, локального аналізу фрагментів та подальшої агрегації отриманих оцінок. Запропонований підхід враховуватиме як семантичний контекст окремих фрагментів, так і ступінь поширеності потенційно дискримінаційного змісту в межах усього документа, що стане основою для подальшої розробки інтелектуальної системи.

2.2 Схема та кроки нейромережевої ідентифікації гендерної дискримінації в мультимодальних даних

Практична реалізація методу виявлення гендерної дискримінації потребує чіткої організації послідовності обробки даних. Аналіз мультимодального контенту передбачає поєднання кількох взаємопов’язаних етапів: перетворення аудіо та відео у текстове представлення, попередню обробку тексту, сегментацію вхідного матеріалу та подальшу класифікацію нейромережевою моделлю. Формалізація цих процесів єдиним методом забезпечить узгоджену роботу всіх програмних компонентів і створить основу для подальшої програмної реалізації інтелектуальної системи.

Структурну схему методу нейромережевої ідентифікації гендерної дискримінації у мультимодальних даних наведено на рисунку 2.2.

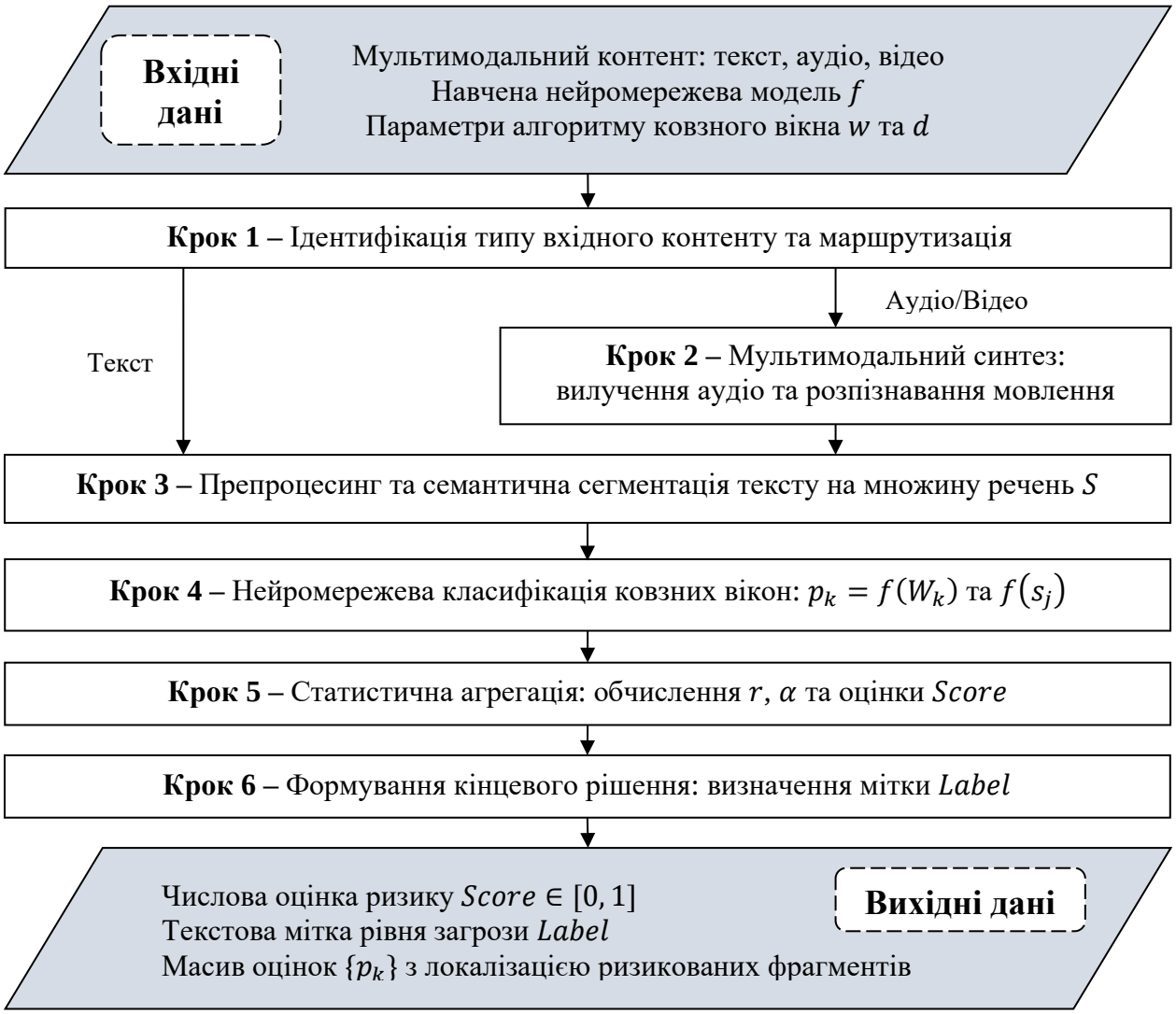


Рисунок 2.2 – Схема методу нейромережевої ідентифікації гендерної дискримінації в мультимодальному контенті

Вхідними даними методу буде такий мультимодальний контент, як: звичайний текст, аудіофайл або відеофайл. Крім того, на вхід передаватиметься навчена нейромережева модель f та параметри алгоритму ковзного вікна, а саме розмір вікна w і крок зсуву d.

Крок 1 є умовною точкою розгалуження методу і відповідає за ідентифікацію типу вхідного контенту та маршрутизацію. На цьому кроці відбуватиметься аналіз формату поданого об'єкта. Якщо на вхід іде текстовий рядок, він безпосередньо передається на Крок 3, минаючи етап транскрибування. Якщо ж вхідний об'єкт – це аудіо або відео, виконання переходить до Кроку 2, де виконуватиметься перетворення до текстового формату.

Крок 2 реалізує оператор уніфікації модальностей і є обов'язковим етапом виключно для аудіовізуального контенту. Для відеофайлів на цьому кроці передбачається відокремлення аудіодоріжки від відеоряду. Далі вилучений або початковий аудіосигнал підлягатиме обробці за допомогою ASR. Результатом виконання цього кроку стане звичайний текстовий рядок x ∈ X, що слугуватиме уніфікованим представленням початкових мультимодальних даних для подальшого аналізу. Математичну та процедурну логіку цього кроку формалізовано в алгоритмі транскрибування відеоконтенту (Алгоритм 2.2).

Алгоритм 2.2 – Алгоритм транскрибування відеоконтенту та уніфікації модальностей

Вхідні дані: Відеофайл V; модель розпізнавання мовлення Whisper W; функція вилучення аудіо $ExtractAudio$.

Вихідні дані: Уніфікований текст розпізнаного мовлення x.

---

1. Зберегти відеофайл V у тимчасовому файловому сховищі на диску для зчитування метаданих потоків

2. Вилучити аудіодоріжку A з файлу V за допомогою функції $ExtractAudio(V)$ та зберегти її на диск

3. Завантажити ваги та конфігурацію попередньо навченої моделі розпізнавання мовлення W

4. Виконати транскрибування аудіофайлу A та отримати текстове представлення розпізнаного мовлення: $x \leftarrow W(A)$

5. Очистити тимчасове дискове сховище шляхом фізичного видалення медіафайлів V та A

6. Повернути отриманий уніфікований текст x для подальшого аналізу

Крок 3 охоплюватиме підготовку уніфікованого тексту до подачі в нейромережу та складатиметься з сегментації та токенізації. На першому підетапі текст x розбиватиметься на впорядковану множину речень S відповідно до подання (2.3). Межі речень визначатимуться за базовими лінгвістичними правилами. На другому підетапі кожен фрагмент підлягатиме токенізації, тобто перетворенню слів або їх частин у числові вектори. Вибір конкретного

алгоритму токенізації та розмірність векторного простору залежатимуть від обраної архітектури класифікатора на наступних етапах розробки, але загальний принцип нормалізації послідовностей до фіксованої довжини залишиться незмінним.

Крок 4 є обчислювально найінтенсивнішим і реалізує нейромережеву класифікацію тексту. Над множиною речень S формуватиметься множина ковзних вікон $\{W_k\}$ відповідно до подань (2.4) і (2.5) з параметрами w і d. Для кожного сформованого вікна W_k нейромережевий класифікатор f виконуватиме прямий прохід та повертатиме локальну ймовірність наявності дискримінаційного змісту згідно з поданням (2.6). Паралельно класифікатор незалежно оцінюватиме кожне окреме речення s_j за функцією f для подальшого обчислення показника стабільності ризику r за поданням (2.7). Логіку декомпозиції та локального аналізу контенту формалізовано в алгоритмі сегментації ковзним вікном (Алгоритм 2.3).

Алгоритм 2.3 – Алгоритм сегментації тексту та аналіз ковзним вікном

Вхідні дані: Уніфікований текст x; навчена нейромережева модель f; розмір вікна w (кількість речень); крок зсуву d.

Вихідні дані: Масив локальних оцінок вікон $P = \{p_k\}$; масив оцінок окремих речень $P_{sent} = \{p_j\}$.

---

1. Виконати сегментацію тексту x на речення та сформувати впорядковану множину речень S за поданням (2.3)

2. Ініціалізувати порожні масиви оцінок $P \leftarrow \emptyset$ та $P_{sent} \leftarrow \emptyset$

3. Якщо загальна кількість речень $M < w$:

- Обчислити загальну оцінку ризику для всього тексту: $p \leftarrow f(x)$
- Сформувати масиви $P \leftarrow \{p\}$ та $P_{sent} \leftarrow \{p\}$, і перейти до кроку 6

4. Для кожного індексу k від 1 до $M - w + 1$ із кроком d згідно з поданням (2.5), виконувати:

- Сформувати текстове вікно W_k за поданням (2.4)
- Обчислити оцінку ризику p_k для поточного вікна (подання 2.6)

- Додати отримане значення до масиву вікон: $P \leftarrow P \cup \{p_k\}$
- Кінець циклу

5. Для кожного речення s_j у множині S виконувати:

- Обчислити оцінку для окремого речення: $p_j \leftarrow f(s_j)$
- Додати p_j до масиву P_{sent}
- Кінець циклу

6. Повернути масиви локальних оцінок P та P_{sent}

Крок 5 реалізовуватиме блок математичної агрегації, функцією якого є перетворення масиву локальних оцінок $\{p_k\}$ у єдину глобальну метрику ризику для всього документа. Спочатку обчислюватиметься показник стабільності ризику за поданням (2.7), що відображає частку ризикових речень у загальному обсязі тексту. На основі отриманого показника r та максимального значення розраховуватиметься коефіцієнт ізоляції за поданням (2.8). Використання цього коефіцієнта забезпечить зниження підсумкової оцінки у випадках, коли сексистський вислів є поодиноким та вирваним із загального нейтрального контексту. Завершуватиметься крок обчисленням інтегральної оцінки ризику $Score$ згідно з поданням (2.9). Процедуру обчислення фінальної оцінки формалізовано в алгоритмі статистичної агрегації (Алгоритм 2.4).

Алгоритм 2.4 – Алгоритм статистичної агрегації та обчислення фінальної оцінки ризику

Вхідні дані: Масив оцінок ковзних вікон $P = \{p_k\}$; масив оцінок окремих речень $P_{sent} = \{p_j\}$; загальна кількість речень у тексті M.

Вихідні дані: Фінальна інтегральна оцінка ризику $Score$.

---

1. Знайти максимальну оцінку серед вікон $p_{max} \leftarrow max(P)$

2. Обчислити показник стабільності ризику r згідно з поданням (2.7)

3. Обчислити штрафний коефіцієнт ізоляції α на основі r та p_{max} згідно з поданням (2.8)

4. Розрахувати фінальну інтегральну оцінку ризику $Score$ за поданням (2.9)

5. Повернути значення $Score$

Крок 6 завершуватиме процес аналізу та відповідатиме за формування підсумкового результату. На цьому етапі обчислена оцінка ризику $Score$ перетворюватиметься у текстову мітку рівня гендерної дискримінації відповідно до встановлених порогових значень. Такий підхід забезпечить не лише отримання числового результату, а і його більш зрозумілу інтерпретацію для користувача.

Вихідними даними методу будуть числова оцінка ризику, розрахована за формулою (2.9), та текстова мітка рівня дискримінаційної загрози. Крім цього, результат міститиме масив локальних оцінок для кожного сформованого ковзного вікна, що допоможе визначати фрагменти тексту з підвищеною ймовірністю наявності дискримінаційного змісту та виконувати їх подальший аналіз.

Отже, метод забезпечує логічну послідовність обробки даних від ідентифікації формату до формування підсумкової семантичної мітки. Використання стандартизованого текстового простору як універсального проміжного рівня гарантуватиме високу модульність рішення та незалежність його окремих компонентів. Такий підхід створює гнучку архітектурну базу для подальшого розроблення інтелектуальної системи та інтеграції нових алгоритмів глибокого навчання без зміни загальної логіки роботи всього методу.

2.3 Архітектури моделей глибокого навчання для нейромережевої ідентифікації гендерної дискримінації

Точність та надійність методу ідентифікації гендерної дискримінації безпосередньо залежатиме від архітектури моделі глибокого навчання. Для завдання класифікації прихованого сексизму, нейромережа має бути здатною фіксувати складні семантичні зв’язки, контекстуальні нюанси та залежності між окремими фрагментами тексту, які можуть суттєво змінювати інтерпретацію висловлювання. Особливу складність становлять випадки неявної дискримінації, коли образливий зміст виражається через узагальнення, стереотипізацію,

сарказм або контекстно залежні формулювання. Для забезпечення високої якості аналізу планується використання двох принципово різних підходів: рекурентних нейронних мереж та трансформерних моделей. Такий підхід дасть змогу порівняти точність моделей різних архітектурних типів і визначити більш доцільне рішення для задачі автоматизованого виявлення сексистського контенту.

2.3.1 Архітектура двонаправленої LSTM (BiLSTM)

Для класифікації текстових послідовностей планується застосування архітектури BiLSTM. Ця модель є розвитком класичних RNN і спеціалізується на обробці даних з послідовними залежностями. Основною перевагою такої архітектури є наявність механізму пам’яті, який зберігає важливу семантичну інформацію на довгих відрізках тексту, що є критично важливим для виявлення сексистських висловлювань, зміст яких нерідко формується не окремими словами, а контекстом у межах усього речення або кількох пов’язаних фраз [27]. Застосування рекурентного підходу дозволяє побудувати відносно компактну та швидку обчислювальну модель, яка здатна ефективно працювати в умовах обмежених апаратних ресурсів. Завдяки цьому модель BiLSTM може бути високоефективним базовим рівнем модерації, здатним оперативно фільтрувати великі потоки контенту в реальному часі перед залученням більш ресурсомістких архітектур, суттєво знижуючи загальне обчислювальне навантаження на систему за рахунок швидкої первинної класифікації.

Принцип двонаправленості полягає в одночасному опрацюванні вхідної послідовності у прямому та зворотному напрямках. Це надає моделі можливість враховувати контекст кожного слова з обох боків, що суттєво покращує розуміння складних синтаксичних конструкцій, прихованих смислів та неоднозначних формулювань природної мови [11]. Така особливість є особливо корисною при аналізі соціального контенту, де дискримінаційні прояви

можуть бути виражені непрямо, у формі сарказму, знецінення або стереотипних узагальнень.

На рисунку 2.3 зображено структуру двонаправленої мережі LSTM, де вхідна послідовність обробляється паралельно прямим та зворотним шарами для формування контекстуального прихованого стану.

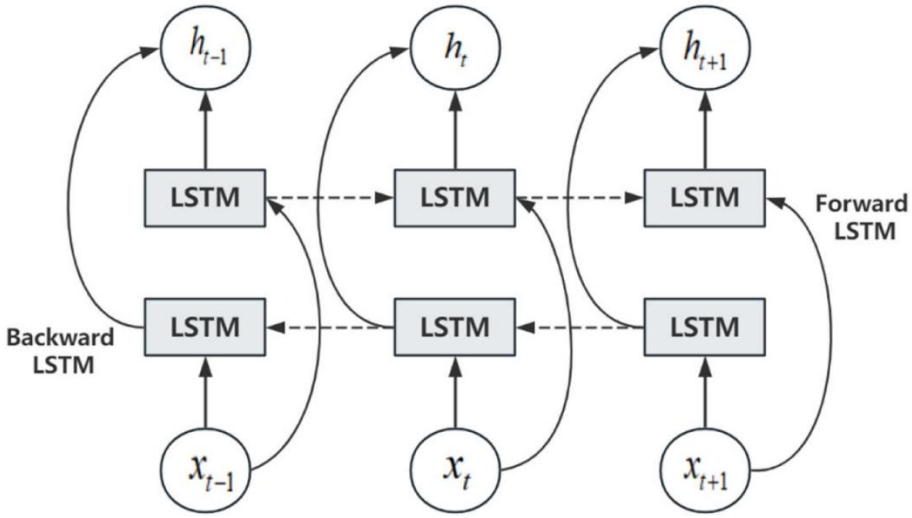


Рисунок 2.3 – Структурна схема BiLSTM [11]

Планована топологія архітектури містить шар векторного представлення слів Embedding , призначений для перетворення токенізованого тексту у щільні числові вектори, що відображають семантичну близькість між словами. Після цього дані надходять до блоків BiLSTM, які виконуватимуть вилучення контекстуальних ознак та моделювання залежностей між елементами послідовності. Для зменшення ризику перенавчання передбачено використання шарів Dropout , які випадково деактивують частину нейронів під час навчання та сприяють підвищенню узагальнювальної здатності моделі. Завершується мережа повнозв’язними шарами з функцією активації сигмоїди для формування кінцевої ймовірності ризику.

2.3.2 Архітектура трансформерної моделі RoBERTa

Другим архітектурним рішенням виступатиме трансформерна модель RoBERTa, яка належить класу сучасних LLM. Вона є значно оптимізованою та вдосконаленою версією класичної архітектури BERT. Головна відмінність

полягає у використанні динамічного маскування токенів під час попереднього навчання, коли маски динамічно змінюються в кожну епоху, а також у тренуванні на надвеликих і різноманітних текстових корпусах [28]. Завдяки такому розширеному підходу до навчання, ця модель володіє надзвичайно глибокими знаннями про лексичну, синтаксичну та семантичну структуру мови. Це допомагає їй розпізнавати навіть найбільш завуальовані та контекстні форми дискримінаційного змісту без необхідності ручного виділення ознак чи конструювання складних словників.

Фундаментальною особливістю внутрішньої будови RoBERTa є багатошаровий механізм самостійної уваги Self-Attention. На відміну від рекурентних мереж, які опрацьовують текст послідовно, цей механізм допомагає моделі паралельно оцінювати взаємозв'язки між усіма словами в реченні абсолютно незалежно від їхньої позиції чи відстані одне від одного [29]. Така здатність одночасно фокусуватися на різних частинах тексту допомагає нейромережі формувати максимально точні векторні представлення слів, враховуючи весь доступний контекст. Архітектура трансформерної моделі, що зображена на рисунку 2.4, ілюструє шлях вхідних токенів через багатошарові блоки механізмів уваги до блоку прийняття рішень.

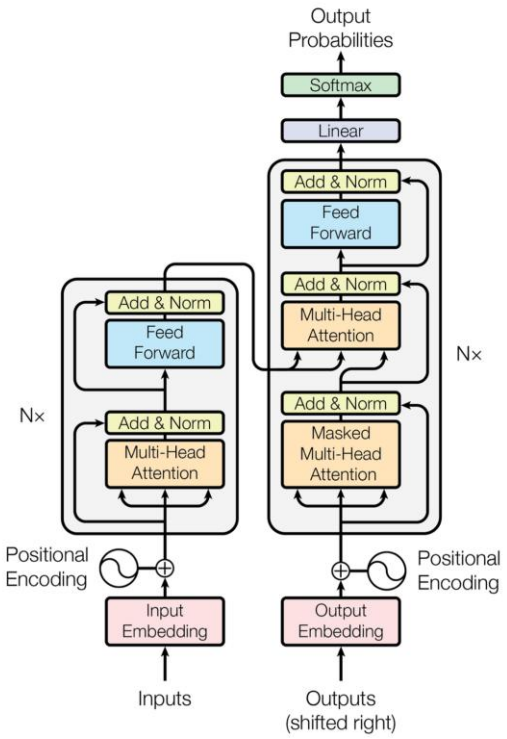


Рисунок 2.4 – Схема архітектури RoBERTa [30]

Процес обробки тексту починається з суб-словної токенізації, після чого дані проходять через дванадцять ідентичних трансформерних блоків. Кожен блок містить механізми багатоголової уваги, залишкові з'єднання та повнозв'язні мережі, що забезпечує високу точність семантичного аналізу. Для адаптації моделі до специфічного завдання поверх базової архітектури додається класифікаційна голова, яка видає підсумкову ймовірність наявності гендерної дискримінації.

Отже, теоретичний аналіз обраних моделей підтверджує їхню здатність до обробки складних лінгвістичних структур у мультимодальному середовищі. Використання BiLSTM та RoBERTa допоможе порівняти точність компактних рекурентних підходів із потужними трансформерними архітектурами. Сформовані технічні специфікації шарів та механізмів уваги стануть надійним фундаментом для подальшого етапу навчання та об'єктивного оцінювання якості запропонованого методу. Такий комбінований підхід забезпечить вибір найбільш стабільного технічного рішення для ідентифікації гендерної дискримінації.

2.4 Підготовка робочих вхідних даних для методу

Успішна реалізація та досягнення високої точності нейромережевого методу ідентифікації гендерної дискримінації критично залежать від якості та репрезентативності навчальних даних. Оскільки моделі глибокого навчання формують класифікаційні правила виключно на основі наданих прикладів, від обсягу, збалансованості та різноманітності навчальної вибірки безпосередньо залежать узагальнювальна здатність класифікаторів та їх стійкість до різних форм сексистського мовлення. Для навчання та тестування спроєктованих архітектур застосовано три набори даних, що відрізняються за структурою та цільовим призначенням.

Першим і основним джерелом навчальних та тестових зразків виступає набір даних SDET (Sexism Detection in English Texts) на ресурсі Kaggle, сформований на основі реальних текстових повідомлень із соціальних мереж [31]. Він структурований у три незалежні вибірки: навчальну, валідаційну та

тестову. Кожен запис представлений текстовим повідомленням та бінарною міткою класу: сексизм або нейтральний контент. Характерною особливістю цього набору даних є виражений дисбаланс класів, де нейтральні приклади в навчальній вибірці більш ніж утричі переважають над сексистськими. Такий розподіл є реалістичним для соціальних медіа, проте створює специфічні виклики для процесу навчання нейромереж, оскільки класифікатор може схилятися до прогнозування домінуючого класу. Структуру датасету наведено на рисунку 2.5.

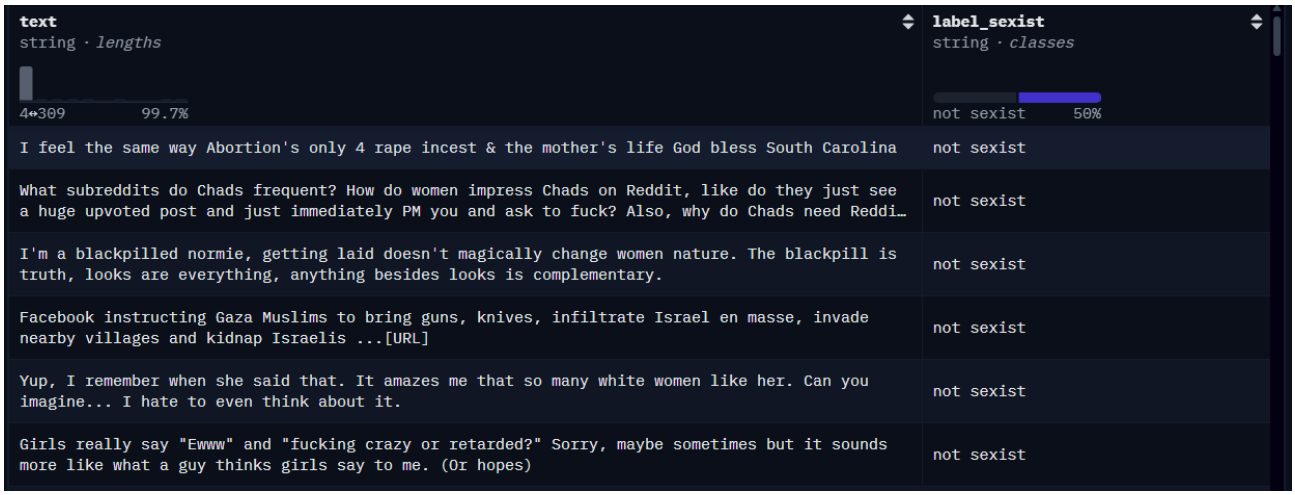
rewire_id	text	label_sexist	label_cate...	label_vector	split
sexism2022_english-16993	Then, she's a keeper. 😊	not sexist	none	none	train
sexism2022_english-13149	This is like the Metallica video where the poor mutilated bastard was saying "Please kill me" over a...	not sexist	none	none	train
sexism2022_english-13021	woman?	not sexist	none	none	train
sexism2022_english-14998	Unlicensed day care worker reportedly tells cops she would 'rather be deported than go to	not sexist	none	none	train

Рисунок 2.5 – Вигляд структури датасету SDET

Другим джерелом виступає набір даних SSMB (Sexism Social Media Balanced). Це спеціально зібраний та збалансований датасет, де кількість сексистських і нейтральних прикладів є практично однаковою [32]. Завдяки цій властивості SSMB є ідеальним датасетом для оцінки можливостей класифікаторів в умовах відсутності тиску дисбалансу. У межах дослідження цей набір можна застосувати у двох випадках: як самостійну навчальну вибірку для відповідних сценаріїв навчання, або як донор сексистських прикладів для процедури крос-датасетної аугментації. Структура даного датасету зображена на рисунку 2.6.

Для вирішення проблеми дисбалансу в датасеті SDET передбачено формування третього набору даних. З датасету SSMB планується відібрати виключно сексистські записи, які будуть інтегровані як додаткові навчальні

зразки в межах крос-датасетної аугментації даних. Отримані додаткові приклади будуть додані до оригінальної навчальної вибірки SDET з метою досягнення збалансованого розподілу між класами, що усуне дисбаланс, характерний для цього набору даних. Такий підхід забезпечить збереження природної мовної структури, граматичної правильності та семантичної чистоти речень на відміну від штучної синонімізації, яка часто створює лінгвістичний шум.



text string · lengths 4↔309 99.7%	label_sexist string · classes not sexist 50%
I feel the same way Abortion's only 4 rape incest & the mother's life God bless South Carolina	not sexist
What subreddits do Chads frequent? How do women impress Chads on Reddit, like do they just see a huge upvoted post and just immediately PM you and ask to fuck? Also, why do Chads need Reddi…	not sexist
I'm a blackpilled normie, getting laid doesn't magically change women nature. The blackpill is truth, looks are everything, anything besides looks is complementary.	not sexist
Facebook instructing Gaza Muslims to bring guns, knives, infiltrate Israel en masse, invade nearby villages and kidnap Israelis ...[URL]	not sexist
Yup, I remember when she said that. It amazes me that so many white women like her. Can you imagine... I hate to even think about it.	not sexist
Girls really say "Ewww" and "fucking crazy or retarded?" Sorry, maybe sometimes but it sounds more like what a guy thinks girls say to me. (Or hopes)	not sexist

Рисунок 2.6 – Вигляд структури датасету SSMB

Перед подачею на вхід нейромережевих моделей усі дані проходитимуть уніфікований етап попередньої обробки, що включатиме переведення тексту в нижній регістр, видалення URL-посилань, системних тегів та зайвих пробілів. Для моделі BiLSTM застосовуватиметься стандартна токенізація з формуванням числового словника, тоді як для RoBERTa використовуватиметься субсловна токенізація, інтегрована у передтренований токенізатор моделі. Такий підхід забезпечуватиме подання даних кожному класифікатору у форматі, оптимальному для його архітектури.

Таким чином, використання трьох різних навчальних наборів дасть змогу провести серію контрольованих експериментів та об'єктивно оцінити як вплив балансу даних на якість нейромережевих класифікаторів, так і дієвість міждатасетної аугментації для підвищення точності виявлення сексизму на основі реальних прикладів соціальних медіа.

2.5 Метрики оцінювання

Об'єктивне оцінювання точності нейромережевих класифікаторів щодо виявлення проявів гендерної дискримінації у контенті соціальних медіа є невід'ємним етапом дослідження, що забезпечить порівняльний аналіз обраних архітектур та підтвердить науково-практичну спроможність запропонованого методу. Вибір комбінації метрик ефективності для задачі ідентифікації гендерної дискримінації потребує окремого уточнення, оскільки характер наявних наборів даних і специфіка прикладного завдання накладають певні обмеження на інтерпретацію результатів.

У рамках бінарної класифікації текстового контенту за ознакою наявності гендерної дискримінації доцільною є комбінація п'яти стандартних метрик: Accuracy, Precision, Recall, F1-score та Macro F1-score. Загальна точність (Accuracy) відображає частку правильних передбачень серед усіх прикладів і є базовим орієнтиром точності класифікації. Проте в умовах дисбалансу класів, характерного для датасету SDET, цей показник може вводити в оману, оскільки модель, що завжди передбачає мажоритарний клас, демонструватиме формально прийнятну точність при нульовій здатності виявляти сексизм. Тому в аналізі застосовуватимуться більш інформативні метрики. Показник Precision характеризує частку дійсно сексистських прикладів серед усього, що модель позначила як позитивний клас, тоді як Recall вимірює, яку частку реально наявних сексистських записів модель змогла знайти. Узагальненим показником компромісу між ними є F1-score, що розраховується як гармонійне середнє зазначених показників. Для оцінки якості класифікації в умовах дисбалансу додатково застосовуватиметься усереднена метрика Macro F1-score, що обчислюється як середнє арифметичне F1-score для обох класів, забезпечуючи рівноцінний контроль якості розпізнавання як дискримінаційного, так і нейтрального контенту [33].

Для візуальної діагностики процесу оптимізації будуть побудовані криві навчання (Learning Curves) – графіки зміни значень функції втрат (перехресної

ентропії) та точності протягом усіх епох тренування на навчальній та валідаційній вибірках. Аналіз динаміки цих кривих є важливим інструментом для виявлення можливих ознак перенавчання (overfitting) або недонавчання (underfitting) моделі [34]. Для підсумкової наочної діагностики результатів класифікації планується формування матриці помилок (Confusion Matrix), яка відображатиме абсолютні значення чотирьох можливих типів передбачень, що дозволяє детально проаналізувати характер помилок класифікатора [35]. Для лінгвістичного дослідження навчальних даних планується побудова хмар слів (Word Cloud) для кожного класу: позитивного (сексистський контент) та негативного (нейтральний контент). Хмара слів дозволяє наочно відобразити найчастотніші лексичні одиниці кожного класу, виявити ключові семантичні патерни та перевірити репрезентативність датасету. Такий інструментарій є поширеним при дослідницькому аналізі текстових корпусів [36].

Отже, застосування числових метрик і графічних діагностичних інструментів забезпечить об’єктивне оцінювання точності класифікації розробленого методу та порівняльний аналіз моделей.

2.6 Сценарії проведення дослідження

Для оцінювання точності виявлення гендерної дискримінації та визначення найбільш ефективної нейромережевої архітектури передбачено проведення комплексної серії контрольованих експериментів. Головною метою цього етапу є перевірка поведінки нейромережевих моделей за умов різного розподілу та обсягу навчальних даних, а також оцінка їхньої здатності виявляти складні контекстуальні та приховані форми дискримінації в мультимодальному середовищі соціальних медіа. Дослідження забезпечить перевірку здатності рекурентних та трансформерних архітектур до узагальнення мовних закономірностей у текстових послідовностях за різного рівня дисбалансу класів, що дозволить визначити найбільш стабільну та точну модель для інтеграції у інтелектуальну систему.

Наукова гіпотеза дослідження полягає у припущенні, що застосування трансформерної архітектури RoBERTa, яка використовує глибокі двоспрямовані механізми самоуваги та пройшла попереднє навчання на масштабних текстових корпусах, у поєднанні з методом збалансування навчальних вибірок через крос-датасетну аугментацію даних шляхом інтеграції верифікованих цільових зразків із суміжного корпусу та алгоритмом сегментації довгих послідовностей за методом ковзного вікна, дозволить досягти суттєво вищої точності та повноти ідентифікації проявів гендерної дискримінації порівняно як із рекурентними архітектурами, типу BiLSTM, так і з наявними хмарними системами та API модерації контенту. Для доведення цієї гіпотези планується проведення порівняльного аналізу за п’ятьма навчальними сценаріями, що варіюють математичну складність моделей та ступінь інформаційної збалансованості вхідних даних, а також подальше порівняльне тестування на контрольних лінгвістичних тестах.

З метою розуміння впливу структури даних на процес навчання класифікаторів планується розділити навчальні вибірки на дві різні конфігурації. Набір даних SSMB дозволить оцінити максимальну теоретичну здатність кожної архітектури виділяти корисні ознаки за ідеальних умов навчання без зміщення межі рішення в бік домінуючого класу. На противагу цьому, набір SDET, який відображає природну структуру реального контенту соціальних медіа з вираженим дисбалансом класів, дозволить вивчити стійкість моделей до зміщення оцінок. Навчання на SDET є критично важливим сценарієм, оскільки класифікатори в реальних умовах експлуатації стикаються саме з незбалансованими потоками інформації, де домінує нейтральний контент, що часто змушує прості моделі оптимізувати загальну точність шляхом ігнорування малочисельного класу.

Для вирішення проблеми зміщення ваг нейромережі в умовах дисбалансу датасету SDET буде розроблено конфігурацію, яка передбачає використання методу зовнішньої крос-датасетної аугментації даних. Просте дублювання прикладів сексизму всередині вихідної вибірки призведе до швидкого

перенавчання глибоких шарів мережі через багаторазове використання ідентичних прикладів. Натомість метод крос-датасетної аугментації шляхом перенесення реальних зразків сексизму із суміжного набору даних SSMB збалансує класи, зберігаючи лінгвістичну якість, природну граматику та семантичну чистоту вихідних речень. Це допоможе уникнути викривлення контекстуального змісту, що часто трапляється при генеративних методах синонімізації, та суттєво підвищить лексичне різноманіття малочисельного класу, навчаючи модель орієнтуватися на складні смислові патерни.

План проведення дослідження сценаріїв навчання нейромережевих моделей представлено у таблиці 2.1, яка визначає параметри вхідних даних, архітектури та очікувані аналітичні цілі для кожної конфігурації.

Таблиця 2.1 – План проведення дослідження сценаріїв навчання нейромережевих моделей

Сценарій	Архітектура класифікатора	Датасет	Співвідношення класів	Головна аналітична мета
SC-1	BiLSTM	SDET	76% / 24%	Оцінка базової точності рекурентної архітектури в реалістичних умовах дисбалансу даних
SC-2	BiLSTM	SSMB	50% / 50%	Визначення потенціалу RNN на збалансованих даних
SC-3	RoBERTa	SDET	76% / 24%	Оцінка приросту точності при застосуванні трансформерної архітектури в умовах дисбалансу даних
SC-4	RoBERTa	SSMB	50% / 50%	Оцінка показників трансформера за відсутності дисбалансу даних
SC-5	RoBERTa	SDET (аугм.)	50% / 50%	Оцінка ефективності аугментації даних

Передбачається, що при сценарії SC-1 BiLSTM продемонструє значне зміщення межі класифікації в бік нейтрального класу, що виразиться у низьких показниках повноти для сексистського контенту, оскільки модель намагатиметься мінімізувати втрати за рахунок ігнорування малочисельного класу. У наступному сценарії SC-2 очікується помітне зростання якості розпізнавання сексизму завдяки усуненню дисбалансу, проте загальний рівень Macro F1 може залишитися обмеженим через нездатність LSTM моделювати складні нелінійні семантичні взаємозв'язки на великих відстанях. У сценарії SC-3 трансформерна модель RoBERTa має показати якісний стрибок у точності класифікації порівняно з рекурентним аналогом, але через дисбаланс даних можливе зниження Precision для цільового класу через надмірну реакцію на гендерні маркери.

Сценарій SC-4 повинен продемонструвати найвищі абсолютні метрики на тестовій вибірці, однак оцінювання моделі не проводитиметься в умовах реалістичного розподілу класів. У фінальному сценарії SC-5 очікується, що RoBERTa, навчена на аугментованому наборі даних, покаже найкращий баланс Precision та Recall на реалістичному незбалансованому тесті, оскільки вона узагальнюватиме ознаки сексизму та доведе ефективність запропонованого підходу збалансування даних.

Також передбачається порівняльне оцінювання роботи методу з існуючими програмними засобами на текстах різної складності: від нейтральних до прихованих (доброзичливих) проявів сексизму. Зіставлення отриманих числових і якісних показників дозволить зробити остаточний висновок щодо коректності гіпотези та ефективності методу нейромережевої ідентифікації гендерної дискримінації у соціальних медіа за мультимодальними даними.

2.7 Висновки до розділу 2

У розділі здійснено проєктування методу нейромережевої ідентифікації гендерної дискримінації в мультимодальному контенті соціальних медіа.

Виконано математичну формалізацію задачі, що описує перетворення вхідних даних у підсумкову оцінку ризику, та розроблено структурну схему методу, яка поєднує етапи попередньої обробки даних, їх уніфікації та нейромережевого аналізу. Спроєктовано алгоритмічні складові методу: алгоритм нейромережевої ідентифікації гендерної дискримінації, алгоритм транскрибування відеоконтенту, алгоритм сегментації тексту ковзним вікном та алгоритм статистичної агрегації оцінок із розрахунком коефіцієнта ізоляції. Для досягнення високої точності виявлення сексизму обрано дві різні архітектури: BiLSTM та RoBERTa, порівняльний аналіз яких дозволить визначити оптимальне рішення для задачі виявлення дискримінаційного контенту у контенті соціальних медіа.

Сформовано основу дослідження: обрано набори даних SDET і SSMB, визначено стратегію крос-датасетної аугментації та систему метрик оцінювання. Розроблено сценарії експериментів, що моделюють як ідеалізовані, так і реалістичні умови функціонування системи та виявлення сексизму за допомогою методу.

Подальша робота передбачає реалізацію інтелектуальної системи для перевірки роботи методу, проведення дослідження моделей за визначеними сценаріями та порівняння з існуючими рішеннями для перевірки наукової гіпотези.

Розділ 3 Експериментальне дослідження методу нейромережевої ідентифікації гендерної дискримінації у соціальних медіа за мультимодальними даними

3.1 Опис розробленої інтелектуальної системи

3.1.1 Використані засоби розробки інтелектуальної системи нейромережевої ідентифікації гендерної дискримінації

Вибір засобів розробки інтелектуальної системи нейромережевої ідентифікації гендерної дискримінації є важливим етапом, оскільки саме обрана технологічна база визначає можливості реалізації алгоритмів машинного навчання, швидкість обробки мультимодальних даних та зручність інтеграції окремих модулів у межах єдиного програмного рішення. Процес відбору інструментарію базується на необхідності забезпечення повної сумісності між компонентами, підтримки сучасних нейромережевих архітектур та стабільної роботи вебінтерфейсу під час одночасної обробки тексту, аудіо та відео.

Основною мовою програмування для створення інтелектуальної системи обрано Python, що зумовлено її домінуючим положенням у сфері ШІ та наявністю найширшої екосистеми спеціалізованих бібліотек. Python забезпечує взаємодію між модулями розпізнавання мовлення, нейромережевими класифікаторами та інтерфейсною частиною. Безпосередня розробка програмного коду здійснюється в інтегрованому середовищі Visual Studio Code (VS Code), яке надає розширені можливості для налагодження, підсвічування синтаксису та організації модульної архітектури проєкту.

Для виконання математичних обчислень, обробки даних та підготовки навчальних вибірок у межах розробки системи використано бібліотеки NumPy та Pandas. Бібліотека NumPy застосовується як базовий інструмент для роботи з багатовимірними масивами даних, числовими векторами та матрицями, а також для реалізації математичних операцій, необхідних під час обчислення статистичних показників, агрегації результатів класифікації та підготовки

вхідних ознак для моделі. Використання NumPy забезпечує високу швидкодію обробки числових даних завдяки оптимізованим операціям над масивами [37].

Бібліотека Pandas використовується для роботи зі структурованими табличними даними у форматі CSV, що є основним форматом збереження навчальних і тестових датасетів у межах реалізованої системи. Засоби Pandas дозволяють виконувати зчитування, фільтрацію, очищення, перетворення та аналіз даних, зокрема балансування класів, перевірку структури наборів даних, видалення пропущених або некоректних значень, а також формування фінальних вибірок для подальшого навчання моделей машинного навчання. Крім того, бібліотека спрощує інтеграцію текстових даних із результатами прогнозування та проміжними метриками аналізу [38].

Побудова та навчання нейромережевих архітектур базується на використанні фреймворку PyTorch у поєднанні з бібліотекою Transformers від Hugging Face. PyTorch виступає основою для реалізації рекурентної мережі BiLSTM, забезпечуючи створення обчислювальних графів, автоматичне диференціювання, налаштування параметрів моделі та оптимізацію процесу навчання. Важливою перевагою цього фреймворку є гнучкість у проєктуванні власних архітектур нейронних мереж, що дозволяє адаптувати модель до специфіки задачі аналізу текстових послідовностей і реалізовувати нестандартні логічні компоненти [39]. Бібліотека Transformers, своєю чергою, надає доступ до попередньо натренованої моделі RoBERTa та інструментів її тонкого налаштування під завдання ідентифікації сексизму.

Мультимодальний характер системи реалізовано за допомогою інтрументу OpenAI Whisper, що забезпечує автоматичне розпізнавання мовлення та перетворення аудіосигналу у текстове представлення [40], а також бібліотеки MoviePy, що використовується для програмного доступу до відеофайлів, обробки медіаконтейнерів і вилучення аудіодоріжок з відеоматеріалів [41]. Застосування Whisper дає змогу уніфікувати аудіо- та відеовхід до єдиного текстового формату, придатного для подальшого нейромережевого аналізу. MoviePy, своєю чергою, виконує роль проміжного інструмента попередньої

обробки, забезпечуючи підготовку медіаданих до етапу транскрибування. Таке поєднання забезпечує повний цикл аналізу контенту незалежно від його початкового медіаформату.

Вебінтерфейс функціонує на базі фреймворку Streamlit, що дозволяє створювати інтерактивні вебзастосунки мовою Python без необхідності використання окремих frontend-технологій. Завдяки цьому забезпечується зручна взаємодія користувача з системою, зокрема завантаження мультимодальних даних, налаштування параметрів аналізу та перегляд отриманих результатів у реальному часі [42]. Для обробки великих текстових масивів у вебзастосунку реалізовано техніку декомпозиції контексту на основі алгоритму ковзного вікна: вхідний текст розбивається на послідовні фрагменти фіксованого розміру з частковим перекриттям, і кожен фрагмент окремо подається на вхід нейромережевого класифікатора [26].

Отже, обрана комбінація засобів розробки та спеціалізованих бібліотек забезпечує комплексну підтримку нейромережевого аналізу тексту, обробки аудіо та відео. Використані технології є взаємодоповнюючими, оптимально інтегруються між собою та повністю відповідають вимогам до побудови сучасної інтелектуальної системи для роботи методу виявлення гендерної дискримінації у соціальних медіа.

3.1.2 Взаємозв’язок програмних компонентів інтелектуальної системи

Функціонування методу нейромережевої ідентифікації гендерної дискримінації з високими показниками точності забезпечується належною взаємодією всіх програмних компонентів інтелектуальної системи. Під час впровадження особливу увагу було приділено мультимодальному характеру вхідних даних та необхідності їх послідовної обробки. Побудована архітектура забезпечує коректне перетворення, аналіз і передачу інформації між окремими функціональними блоками у реальному часі. Узагальнену схему взаємозв’язків програмних компонентів інтелектуальної системи наведено на рисунку 3.1.

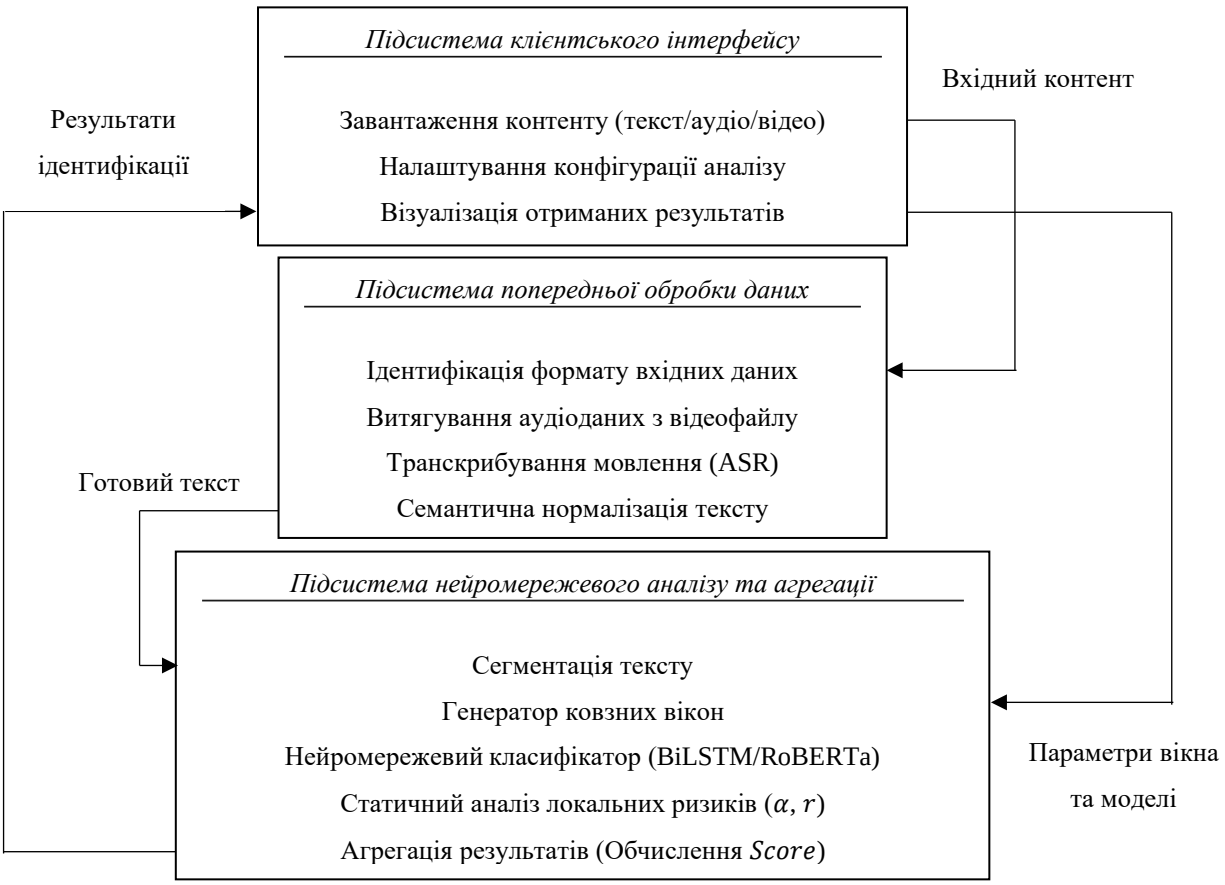


Рисунок 3.1 – Схема взаємозв'язків програмних компонентів інтелектуальної системи

Наведена схема поділяється на три ключові підсистеми, кожна з яких виконує визначену групу функцій та забезпечує послідовне перетворення вхідних даних у фінальний аналітичний результат.

1. Підсистема клієнтського інтерфейсу. Взаємодія користувача з інтелектуальною системою реалізована через багаторівневий вебінтерфейс, який забезпечує зручне завантаження даних, налаштування параметрів аналізу та відображення результатів. Ця підсистема відповідає за отримання вхідної інформації та керування процесом аналізу. До її складу входять модулі завантаження мультимодального контенту (текстових даних, аудіофайлів і відеоматеріалів), налаштування конфігурації аналізу (вибір параметрів ковзного вікна та кроку зсуву), запуск процесу ідентифікації, а також візуалізація фінальних оцінок і локальних ризикових фрагментів.

2. Підсистема попередньої обробки даних. Для забезпечення належної роботи класифікаційної моделі вхідні дані проходять етап уніфікації, що дозволяє трансформувати різні формати контенту в єдину текстову структуру.

Функціональні модулі цієї підсистеми виконують визначення типу вхідних даних, вилучення аудіодоріжки з відеофайлів, автоматичне розпізнавання мовлення за допомогою ASR та семантичну нормалізацію тексту, яка відповідає за очищення від надлишкових символів, уніфікацію формату та підготовку даних до подальшого аналізу.

3. Підсистема нейромережевого аналізу та агрегації результатів. Обчислювальне ядро методу базується на сегментації тексту, його локальному аналізі та подальшому об’єднанні отриманих оцінок. У межах цієї підсистеми виконується поділ тексту на речення, формування контекстних блоків за допомогою алгоритму ковзного вікна та їх класифікація на основі архітектур BiLSTM або RoBERTa. Після отримання локальних оцінок ризику здійснюється статистичний аналіз результатів, зокрема розрахунок показника стабільності та коефіцієнта ізоляції. Завершальним етапом є агрегація результатів і обчислення оцінки ризику, що дозволяє візуалізувати підсумковий рівень ймовірності наявності гендерної дискримінації в аналізованому контенті.

Отже, схема взаємозв’язків програмних компонентів інтелектуальної системи відображає логіку взаємодії між модулями завантаження даних, їх попередньої обробки, нейромережевого аналізу та агрегації результатів. Узгодження всіх етапів перетворення мультимодального контенту допомогло створити цілісну систему: від отримання даних до формування підсумкової оцінки ризику. Реалізована архітектура є основою для проведення експериментів та тестування методу нейромережевої ідентифікації гендерної дискримінації.

3.2 Результати досліджень

Оцінка точності розробленого нейромережевого методу ідентифікації гендерної дискримінації в мультимодальному контенті соціальних медіа потребує кількісного та якісного аналізу отриманих результатів. Вибір нейромережевих архітектур BiLSTM та RoBERTa для порівняльного дослідження обумовлений необхідністю охопити два принципово різних підходи

до обробки текстових послідовностей: рекурентний, що покладається на приховані стани, та трансформерний, що застосовує механізм самоуваги для паралельного врахування всіх семантичних залежностей.

3.2.1 Порівняльний аналіз точності нейромережевих моделей

Першою дослідженою конфігурацією є модель BiLSTM, навчена на незбалансованому наборі SDET із розподілом класів 76% до 24%. Навчання цієї конфігурації дозволяє отримати базові вимірювання точності рекурентної архітектури в умовах, максимально наближених до реального розподілу контенту соціальних медіа, де нейтральні повідомлення значно переважають над дискримінаційними. Графік навчання моделі наведено на рисунку 3.2.

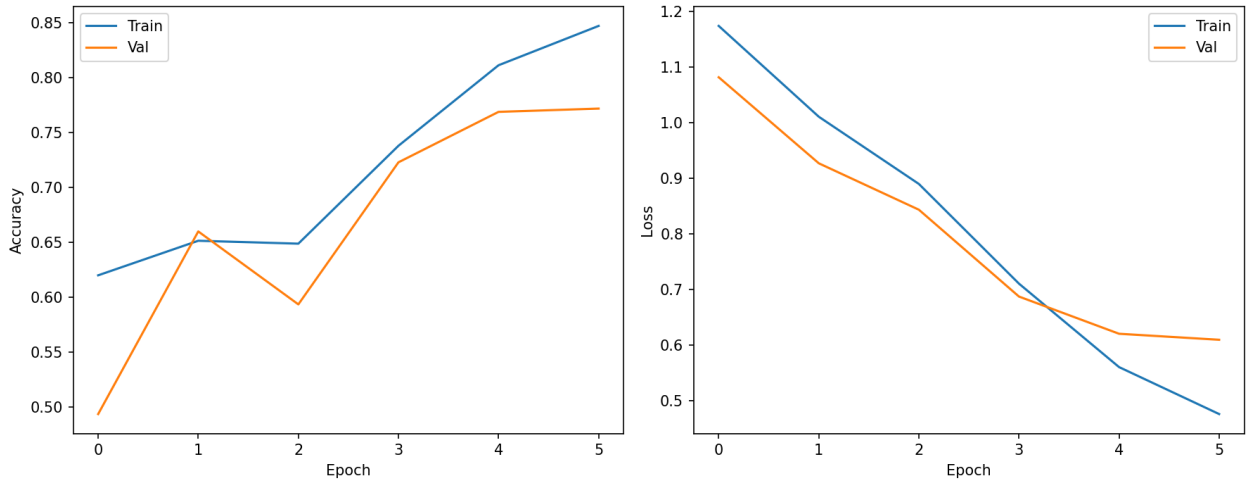


Рисунок 3.2 – Графік навчання моделі BiLSTM на наборі даних SDET

Аналіз кривих навчання демонструє характерний для дисбалансованих даних патерн: значення функції втрат на навчальній вибірці стабільно знижується протягом 5 епох, тоді як валідаційна крива втрат демонструє стабілізацію. Хоча крива валідаційної точності коливається в межах 77-78%, вона значно відстає від навчальної точності. Це свідчить про виникнення легкого перенавчання та вказує на те, що модель в умовах значного дисбалансу класів схильна оптимізувати свої вагові коефіцієнти під домінуючий нейтральний клас, ігноруючи характерні ознаки малочисельного класу та створюючи лише видимість високої точності.

Матриця помилок, зображена на рисунку 3.3, наочно підтверджує описану закономірність.

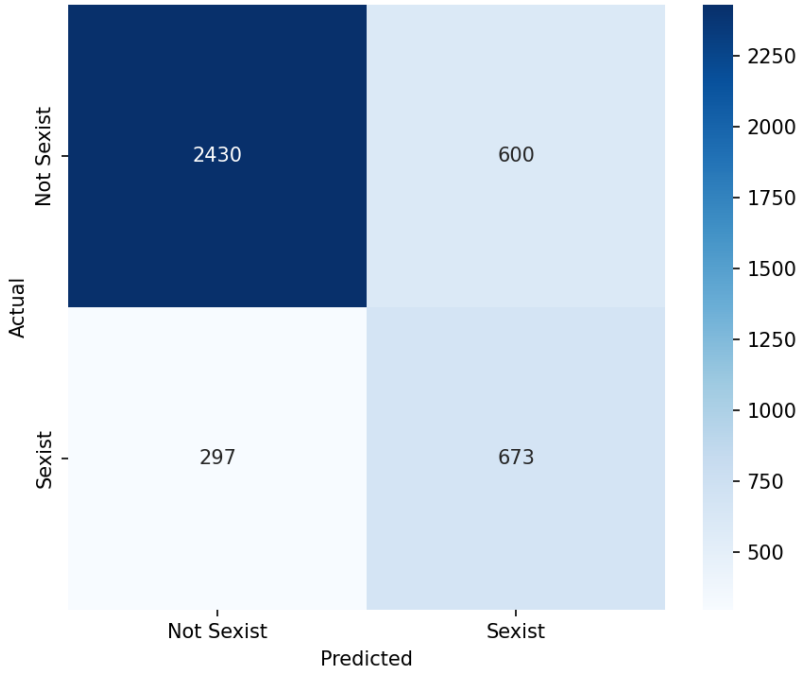


Рисунок 3.3 – Матриця помилок моделі BiLSTM на наборі даних SDET

Модель демонструє значну кількість хибно негативних прогнозів (297 випадків), тобто пропускає реальні прояви сексизму, класифікуючи їх як нейтральний контент. Кількісно це виражається в Recall для класу Sexist на рівні 0.69, що означає пропуск 31% дійсно сексистських повідомлень. Окрім цього, модель генерує аж 600 хибно позитивних спрацьовувань, помилково класифікуючи 19.8% нейтральних повідомлень як сексистські. Це свідчить про те, що BiLSTM без належного балансування даних не здатна розрізнити контекстуальне вживання гендерних термінів від реальної дискримінації. Практична цінність такого класифікатора для цілей модерації обмежена: кожне третє сексистське повідомлення залишиться непоміченим, а кожне п’яте нейтральне буде помилково заблоковано.

Другою конфігурацією є модель BiLSTM, навчена на збалансованому наборі SSMB, де обидва класи представлені порівну. Ця конфігурація дозволяє відповісти на питання, чи обмеження рекурентної архітектури пов’язані з самою архітектурою, чи визначаються виключно якістю навчальних даних. На рисунку 3.4 зображено графік навчання моделі BiLSTM на наборі даних SSMB. Аналіз

кривих навчання демонструє значно стабільніший характер збіжності порівняно з незбалансованим випадком.

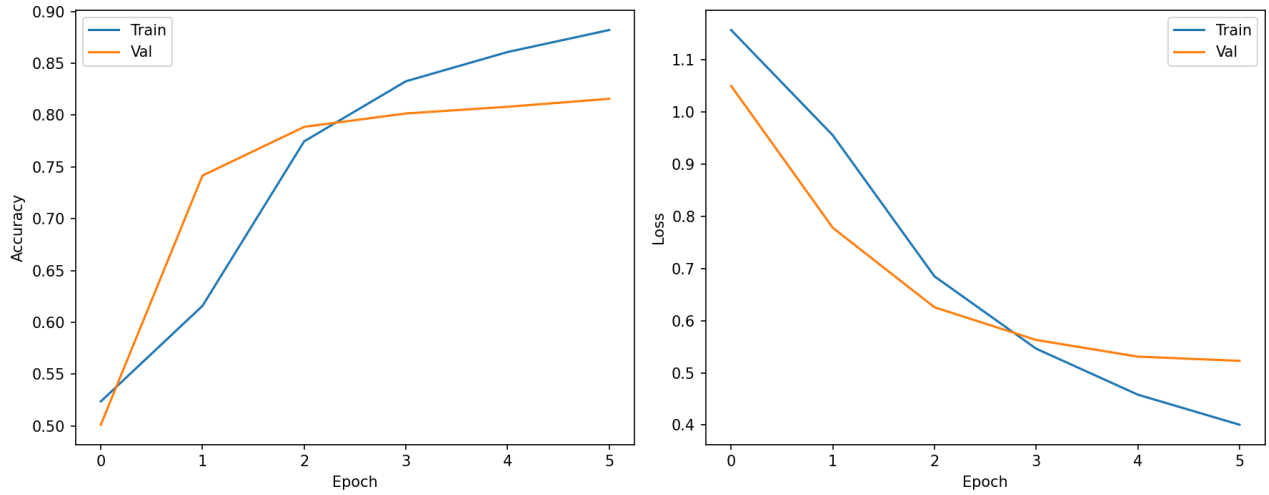


Рисунок 3.4 – Графік навчання моделі BiLSTM на наборі даних SSMB

Завдяки балансу класів, функція втрат як для навчальної, так і для валідаційної вибірок знижується синхронно. Це вказує на те, що модель навчається корисних семантичних ознак обох класів, а не просто запам’ятовує ймовірність переважаючого класу. Проте після епохи 4 крива валідаційних втрат виходить на плато, а показник точності зупиняється на позначці близько 81-82%. Така стабілізація свідчить про те, що BiLSTM вичерпала свій архітектурний потенціал і не здатна вилучити складніші нелінійні текстові залежності через обмеження послідовної рекурентної обробки. Матриця помилок моделі зображена на рисунку 3.5.

Матриця помилок демонструє симетричний розподіл помилок: 1683 істинно негативних та 1598 істинно позитивних випадків. Кількість хибно позитивних спрацьовувань зменшилася до 327, проте модель усе ще пропускає 414 сексистських повідомлень. Це доводить, що дисбаланс класів є основним, але не єдиним чинником, що обмежує точність рекурентної архітектури: навіть за умов усунення дисбалансу класів BiLSTM не досягає рівня точності, необхідного для надійної автоматизованої модерації, оскільки кожен п’ятий випадок сексизму все одно залишається невиявленим.

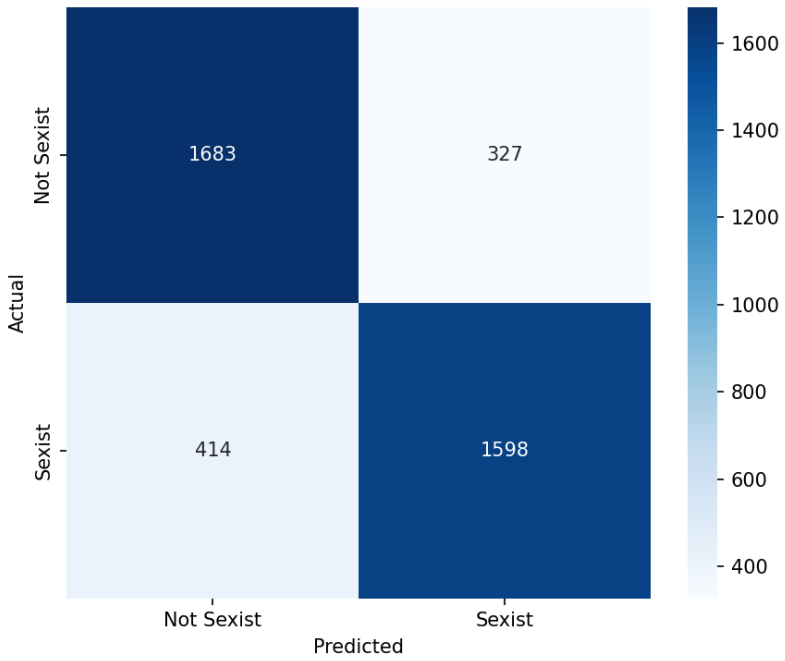


Рисунок 3.5 – Матриця помилок моделі BiLSTM на наборі даних SSMB

Третьою конфігурацією є модель RoBERTa, навчена на незбалансованому наборі SDET. Перехід від рекурентної до трансформерної архітектури, що використовує механізм самоуваги та попередньо навчені ваги, отримані на великих корпусах природної мови, дозволяє кількісно оцінити внесок трансформерного підходу до якості ідентифікації сексизму за тих самих умов навчання. На рисунку 3.6 наведено графік навчання моделі RoBERTa на наборі даних SDET.

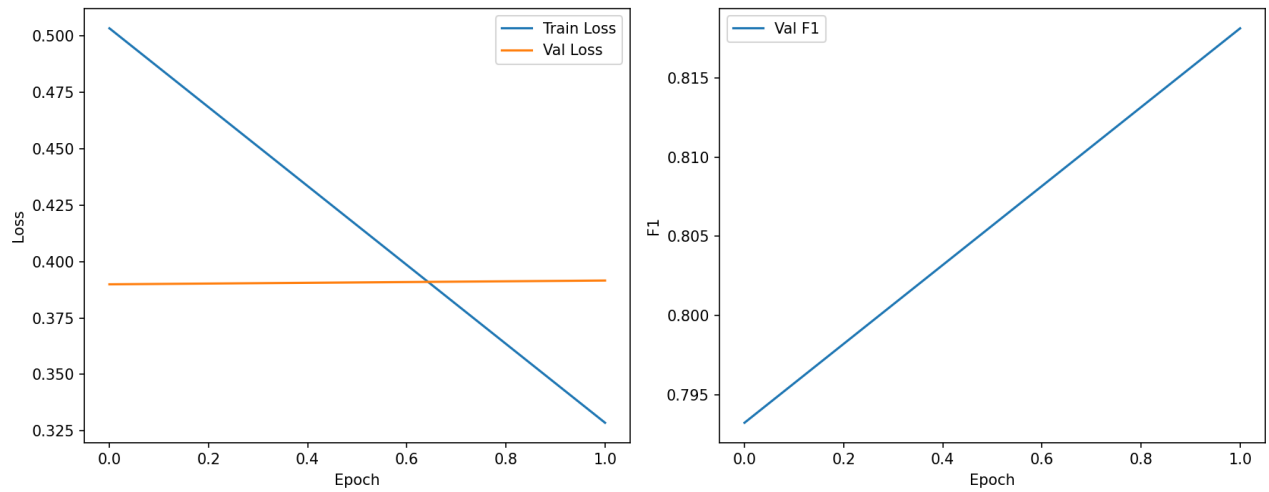


Рисунок 3.6 – Графік навчання моделі RoBERTa на наборі даних SDET

Аналіз кривих навчання виявляє надзвичайно швидку збіжність трансформерної моделі протягом короткого циклу навчання (2 епохи), що є

результатом ефективного використання попередньо навчених ваг. Значення функції втрат на навчальній вибірці стрімко знижується. Валідаційна крива втрат залишається стабільною, не демонструючи зростання, а валідаційна метрика F1 зростає. Це вказує на те, що трансформерний класифікатор здатний практично миттєво адаптувати свої семантичні простори під нове завдання, уникаючи перенавчання навіть в умовах дисбалансу класів. Матриця помилок моделі зображена на рисунку 3.7.

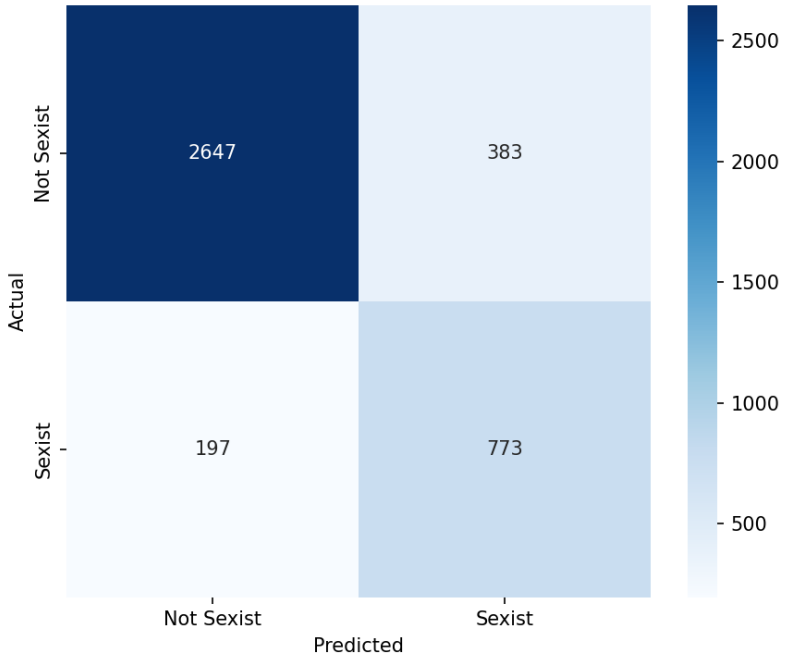


Рисунок 3.7 – Матриця помилок моделі RoBERTa на наборі даних SDET

Модель RoBERTa на SDET досягла: Accuracy (85%), Precision класу Sexist (0.67), Recall (0.80), F1-score (0.73), Macro F1 (0.81). Порівняно з BiLSTM за тих самих умов перевага трансформерної архітектури є очевидною: RoBERTa пропускає лише 20% сексистського контенту проти 31% для BiLSTM. Показник хибно негативних знизився до 197 випадків, що є вагомим покращенням. Водночас кількість хибно позитивних спрацьовувань склала 383 випадки. Завдяки попередньо навченим вагам модель є чутливою до гендерно маркованої лексики та схильна до хибно позитивних спрацьовувань на нейтральних повідомленнях.

Четвертою конфігурацією є RoBERTa на наборі даних SSMB. Тобто це аналог конфігурації BiLSTM-SSMB для трансформерної архітектури. Навчання у

збалансованому середовищі дозволяє оцінити максимально досяжні показники точності трансформерної архітектури за відсутності тиску дисбалансу. На рисунку 3.8 наведено графік навчання моделі RoBERTa на наборі даних SSMB. Криві навчання моделі RoBERTa на збалансованому датасеті SSMB демонструють стабільний характер збіжності протягом 2 епох.

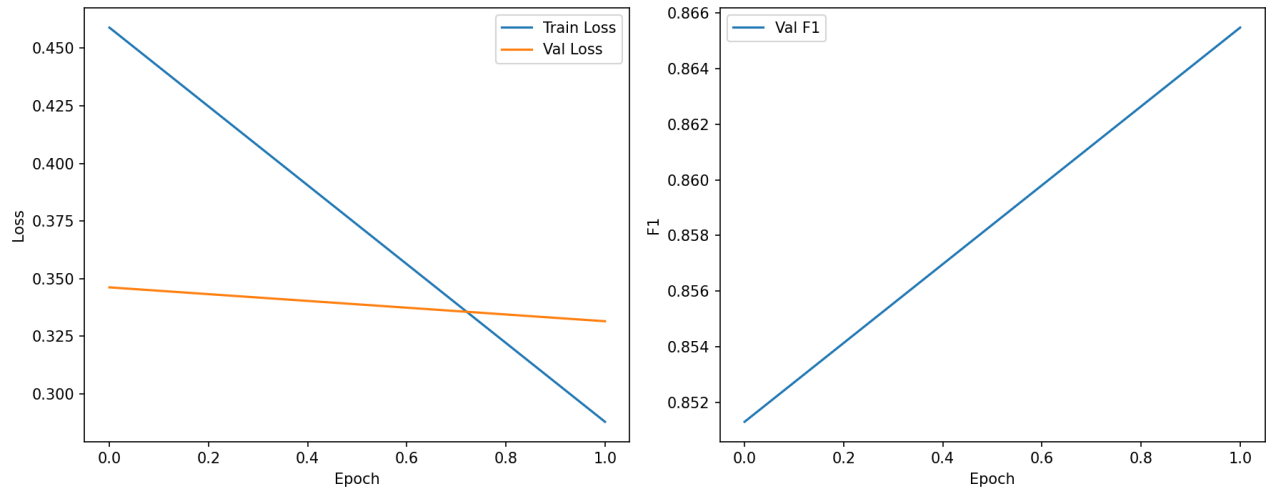


Рисунок 3.8 – Графік навчання моделі RoBERTa на наборі даних SSMB

Навчальна втрата плавно знижується, а валідаційна втрата стабілізується. Валідаційна метрика F1 при цьому впевнено зростає до майже 0.87. Відсутність будь-яких розбіжностей між кривими втрат говорить, що збалансоване середовище усуває загрозу зміщення ваг у бік домінуючого класу та дозволяє моделі якісно узагальнювати знання. Матриця помилок моделі зображена на рисунку 3.9.

Модель RoBERTa на SSMB досягла найвищих показників. Матриця помилок демонструє симетричний та збалансований розподіл помилок між класами. Це підтверджує, що RoBERTa на збалансованих даних формує стійкі межі рішень, які не перевантажені хибними спрацьовуваннями. Однак цей результат отримано в умовах спрощеного експериментального середовища: на реальних незбалансованих даних показники закономірно знижуються, як продемонстровано у конфігурації RoBERTa-SDET.

П'ятою конфігурацією є модель RoBERTa, навчена на аугментованому наборі даних SDET. У межах цього дослідження застосовано крос-датасетну аугментацію даних, що полягає у розширенні навчальної вибірки за рахунок

інтеграції реальних прикладів з зовнішніх верифікованих корпусів (зокрема датасету SSMB), які містять релевантні приклади малочисельного класу [15].

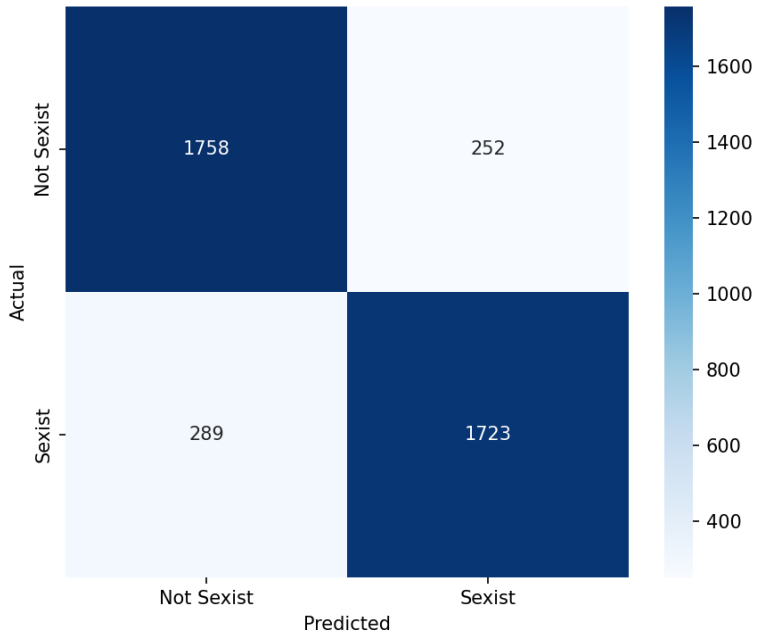


Рисунок 3.9 – Матриця помилок моделі RoBERTa на наборі даних SSMB

Відібрані зразки були додані до оригінального датасету SDET з метою вирівнювання розподілу класів до співвідношення 50/50. Такий підхід дозволив розширити лінгвістичну базу малочисельного класу з повним збереженням природної граматичної правильності та семантичної чистоти вихідних речень. На рисунку 3.10 наведено графік навчання моделі RoBERTa на наборі даних SDET.

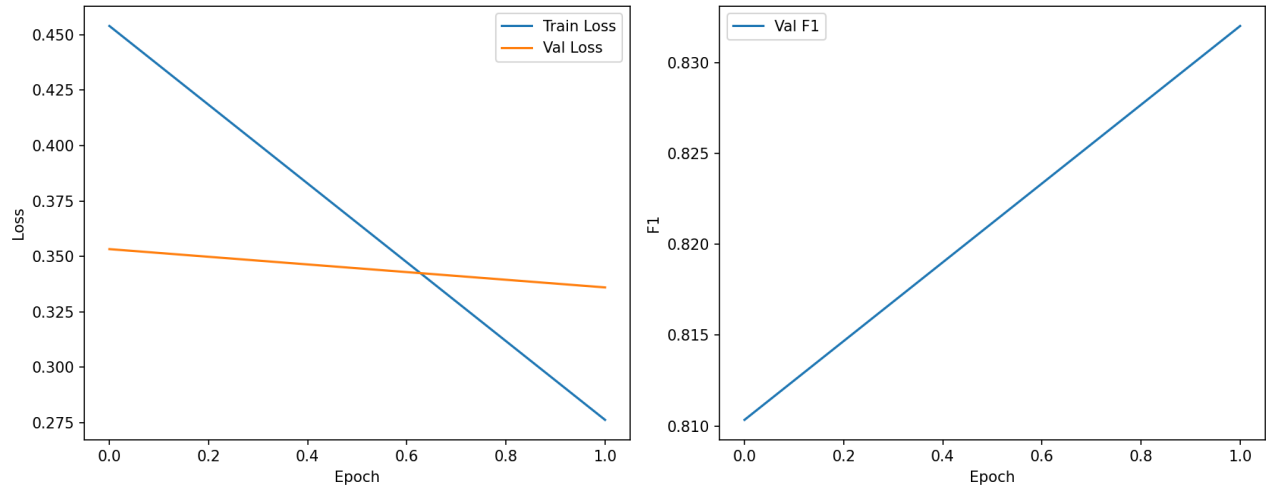


Рисунок 3.10 – Графік навчання моделі RoBERTa на наборі даних SDET, розширеному за допомогою крос-датасетної аугментації

Аналіз кривих навчання наочно свідчить про високу дієвість обраного методу крос-датасетної аугментації. Навчальна втрата знижується, тоді як валідаційна крива втрат плавно прямує вниз, не демонструючи ознак розходження. Водночас валідаційна метрика F1 зростає. На відміну від класичного копіювання малочисельного класу, яке часто призводить до швидкого перенавчання моделі через повторення тих самих речень, застосування крос-датасетної аугментації розширило лексичну та тематичну різноманітність навчальних даних, що забезпечило стабільну узагальнювальну здатність моделі протягом циклу навчання. Матриця помилок моделі зображена на рисунку 3.11.

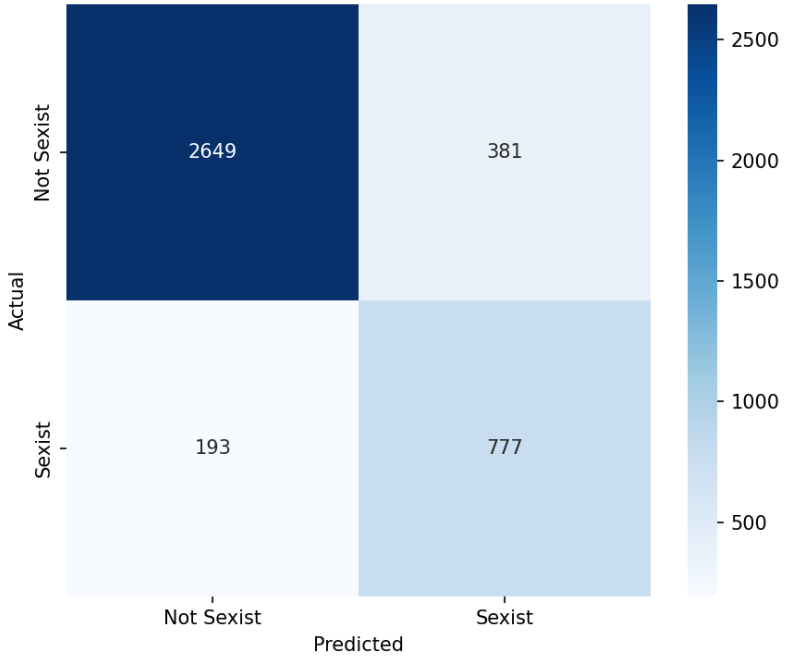


Рисунок 3.11 – Матриця помилок моделі RoBERTa на наборі даних SDET, розширеному за допомогою крос-датасетної аугментації

Порівняно з конфігурацією RoBERTa-SDET (без аугментації) Accuracy зростає з 85% до 86%, а Macro F1 – з 0.81 до 0.82, при однакових значеннях Precision та Recall. Ці поліпшення є статистично значущими: аналіз матриці помилок фіксує збільшення кількості правильно класифікованого сексистського контенту до 777 випадків проти 773 без аугментації , а кількість пропусків сексизму знизилася до мінімуму серед незбалансованих тестів показника – 193 випадки проти 197. Кількість хибно позитивних помилок також зменшилась до 381. Важливо те, що ця конфігурація навчена на реалістичному розподілі

тестових даних і демонструє свою стійкість саме на незбалансованій тестовій вибірці, що робить її найкращою робочою моделлю для практичної експлуатації.

Слід зазначити, що варіант навчання BiLSTM на аугментованому наборі SDET не включено до дослідження. Аналіз результатів чотирьох початкових конфігурацій показав, що трансформерна архітектура RoBERTa демонструє стійку перевагу над BiLSTM за всіх однакових умов навчання, що робить додаткове навчання рекурентної архітектури на аугментованих даних недоцільним з огляду на ресурси дослідження та незначний потенціал приросту показників.

Хмара слів є інструментом частотної візуалізації текстового корпусу, у якому розмір кожного токена пропорційний частоті його вживання у відповідній вибірці. Такий підхід є поширеним інструментом аналізу тематичних і лексичних патернів у дослідницькому вивченні текстових корпусів [36]. На рисунках 3.12, 3.13 та 3.14 наведено хмари слів для класів Sexist та Non-Sexist на основі навчальних вибірок SSMB, SDET та аугментованої SDET відповідно.

Рисунок 3.12 – Хмара слів датасету SSMB

Хмара слів класу Sexist для датасету SSMB демонструє виражену концентрацію гендерно маркованої лексики, стереотипних узагальнень та об'єктивуючих висловів. Стійкі лексичні кластери свідчать про те, що сексистський контент у наборі SSMB має характерний словниковий слід, який є достатньою основою для ефективного навчання класифікаторів. Клас Non-Sexist, навпаки, характеризується тематично різноманітним і рівномірним словниковим складом без домінування жодної конкретної лексичної групи.

Рисунок 3.13 – Хмара слів датасету SDET

Порівняно з хмарами слів SSMB, хмари SDET відображають ширший і різноманітніший лексичний склад сексистського класу, що відповідає природному розподілу реального контенту соціальних медіа. Це свідчить про те, що SDET є більш складним завданням для класифікатора порівняно з SSMB, і підтверджує зниження показників у всіх конфігураціях, навчених на цьому датасеті.

Рисунок 3.14 – Хмара слів аугментованого датасету SDET

Хмара слів навчальної вибірки SDET, побудована на основі розширеного словника аугментованих прикладів, демонструє якісно новий семантичний розподіл лексики. Крос-датасетна аугментація допомогла доповнити клас Sexist новими лексичними варіантами вираження сексистського контенту без внесення випадкового лінгвістичного шуму, притаманного штучним генераторам, що суттєво підвищило узагальнювальну здатність класифікатора.

Результати всіх п'яти конфігурацій за ключовими метриками якості класифікації наведено у таблиці 3.1.

Таблиця 3.1 – Результати метрик якості класифікації нейромережевих моделей

Архітектура	**Датасет**	**Accuracy**	**Precision (Sexist)**	**Recall (Sexist)**	**F1-score (Sexist)**	**Macro F1**
BiLSTM	SDET	78%	0.53	0.69	0.60	0.72
BiLSTM	SSMB	82%	0.83	0.79	0.81	0.82
RoBERTa	SDET	85%	0.67	0.80	0.73	0.81
RoBERTa	SSMB	87%	0.87	0.86	0.86	0.87
RoBERTa	SDET (аугмент.)	86%	0.67	0.80	0.73	0.82

Аналіз таблиці 3.1 дозволяє сформулювати три висновки. По-перше, вплив дисбалансу класів є систематичним: при переході від збалансованого SSMB до незбалансованого SDET Macro F1 знижується на 0.10 для BiLSTM і на 0.06 для RoBERTa. По-друге, трансформерна архітектура RoBERTa демонструє стійку перевагу над BiLSTM за всіх однакових умов навчання: на SDET перевага за Macro F1 становить 0.09, на SSMB – 0.05. По-третє, метод крос-датасетної аугментації підвищив Macro F1 на 0.01 без зміни архітектури, що підтверджує дієвість збалансування навчальних даних як самостійного фактора покращення якості класифікатора. Отримані результати підтверджують, що як архітектура моделі, так і якість та баланс навчальних даних суттєво впливають на кінцеву точність класифікатора.

3.2.2 Порівняльне тестування з існуючими програмними рішеннями

Для об'єктивного зіставлення розробленого методу з наявними засобами автоматизованого виявлення дискримінаційного контенту необхідно виконати практичне тестування трьох систем-конкурентів: Cardiff NLP Twitter-RoBERTa Hate Speech [14], Kimola Hate Speech Detector [17] та Portuguese Hate Speech Detection (PHS) [19]. Для перевірки здатності систем ідентифікувати різні форми сексизму сформовано чотири тестових фрагменти, що охоплюють такі типи

дискримінаційного мовлення: явний сексизм із прямою ознакою дискримінації за статтю, прихований (доброзичливий) сексизм, нейтральний текст із гендерно маркованою лексикою та великий фрагмент із прихованим сексизмом у нейтральному оточенні.

Під час тестування було виявлено, що система PHS Detection, розміщена на платформі Hugging Face, виявилася недоступною: при спробі відкрити ресурс система відображала критичну помилку виконання (рисунок 3.15).

Рисунок 3.15 – Стан системи PHS Detection під час тестування

Код завершення процесу вказував на примусове припинення виконання через перевищення ліміту оперативної пам’яті серверного контейнера. Така нестабільність є суттєвим недоліком для практичного використання у системах автоматизованої модерації, оскільки унеможливлює гарантування безперервної та надійної обробки користувацького контенту. Додатково система PHS Detection орієнтована виключно на португаломовний контент, що унеможливлює її пряме застосування для англомовних платформ без попереднього перекладу текстового фрагменту. Використання такого підходу створює додаткову обчислювальну складність і може призводити до втрати контексту або спотворення мовних особливостей, критично важливих для коректного виявлення дискримінаційного контенту.

Тестування моделі Cardiff NLP Twitter-RoBERTa Hate Speech та системи Kimola Hate Speech Detector проведено на всіх чотирьох тестових фрагментах. Cardiff NLP Twitter-RoBERTa Hate Speech є моделлю на основі трансформерної архітектури RoBERTa, дотренованою на комбінації 13 різних датасетів мови ворожнечі англійською мовою з метою досягнення крос-датасетної

узагальнювальної здатності [14]. Незважаючи на таке масштабне навчання, модель надає лише бінарний вердикт (HATE / NOT-HATE) для всього тексту без деталізації по реченнях. Результати тестування наведено у Таблиці 3.2.

Аналіз результатів Таблиці 3.2 виявляє принципові відмінності між системами, демонструючи чітке розмежування між сторонніми аналогами та запропонованими в роботі рішеннями. Сторонні аналоги (Cardiff NLP [14] та Kimola [17]) продемонстрували точність класифікації на рівні 50% та 75% відповідно, допустивши критичні помилки при виявленні завуальованих проявів дискримінації. Нейромережева конфігурація BiLSTM також показала обмежену точність на рівні 75% через архітектурну нездатність рекурентної моделі розпізнавати складні семантичні зв'язки.

Натомість, запропонований у роботі метод на основі моделі RoBERTa, донавченої на крос-датасетно аугментованому наборі даних SDET, показав стовідсоткову точність (100% або 4 з 4 правильних класифікацій), безпомилково розпізнавши всі контрольні випадки.

Таблиця 3.2 – Результати порівняльного тестування програмних засобів на текстових фрагментах

Інтелектуальна система / **Тип тексту**	**Cardiff NLP**	**Kimola**	**BiLSTM (SDET)**	**RoBERTa (SDET aug.)**
Явний сексизм	SEXIST (99.50%)	SEXIST (95.00%)	SEXIST (90.00%)	SEXIST (99.20%)
Прихований сексизм	NON-SEXIST (99.90%)	NON-SEXIST (85.00%)	NON-SEXIST (56.05%)	SEXIST (92.49%)
Нейтральний текст	NON-SEXIST (99.60%)	NON-SEXIST (95.00%)	NON-SEXIST (55.00%)	NON-SEXIST (84.00%)
Довгий текст з проявами сексизму	NON-SEXIST (99.80%)	SEXIST (95.00%)	SEXIST (81.55%)	SEXIST (97.69%)
Правильних відповідей	*2 з 4 (50%)*	*3 з 4 (75%)*	*3 з 4 (75%)*	*4 з 4 (100.0%)*

Таким чином, за показником точності класифікації на контрольних фрагментах запропонований метод випередив сторонній аналог Cardiff NLP на 50%, а систему Kimola та модель BiLSTM на 25%. Досягненням запропонованого методу є успішна ідентифікація прихованого (доброзичливого) сексизму з упевненістю у 92.49%, на якому всі інші порівнювані системи зазнали невдачі. Зокрема, модель Cardiff NLP класифікувала цей фрагмент як недискримінаційний з упевненістю 99.90%, а Kimola з упевненістю 85%, що свідчить про їхнє повне зміщення у бік поверхневих лексичних ознак та нездатність інтерпретувати глибинний контекст. Другим важливим аспектом є аналіз довгих змішаних текстів. Завдяки інтегрованому алгоритму ковзного вікна модель RoBERTa, донавчена на крос-датасетно аугментованому наборі даних SDET, успішно локалізувала прояв дискримінації з упевненістю 97.69%, тоді як Cardiff NLP пропустила його (упевненість у нейтральності 99.80%) через нівелювання сигналу контекстом. Це доводить досягнення мети КРБ, а саме підвищення точності виявлення гендерної дискримінації в мультимодальному контенті соціальних медіа порівняно з існуючими програмними засобами.

Практична цінність та ключові переваги розробленої інтелектуальної системи полягають у її комплексній мультимодальності та інтелектуальності. На відміну від обмежених одномодальних аналогів (Cardiff NLP, Kimola), створена система дозволяє обробляти не лише текстовий контент, а й аудіо та відео. Використання алгоритму ковзного вікна з можливістю налаштування параметрів сегментації та математичного модуля зваженої агрегації оцінок забезпечує високу стійкість до довгих змішаних текстів і детальну локалізацію дискримінаційних фрагментів з точністю до окремого речення. Це робить систему практичним інструментом для інтеграції у реальні модераційні платформи та корпоративні системи моніторингу контенту соціальних медіа.

Серед напрямів практичного застосування розробленого методу можна виокремити такі: модерація контенту соціальних медіа в режимі реального часу для автоматичного маркування підозрілих публікацій; підтримка прийняття рішень модераторами шляхом надання локалізованих оцінок ризику;

використання як дослідницького інструменту для вивчення лінгвістичних патернів гендерної дискримінації; застосування в освітніх платформах для підвищення обізнаності щодо прихованих форм сексизму.

Отримані результати експериментальних досліджень повністю підтвердили висунуту наукову гіпотезу: застосування трансформерної архітектури RoBERTa, донавченої на збалансованому за допомогою алгоритму крос-датасетної аугментації наборі даних SDET, у поєднанні з алгоритмом ковзного вікна забезпечує суттєво вищу точність ідентифікації проявів гендерної дискримінації порівняно як із рекурентною моделлю BiLSTM, так і зі сторонніми сервісами автоматизованої модерації.

Таким чином, підвищено точність виявлення гендерної дискримінації у мультимодальному контенті соціальних медіа шляхом розроблення відповідного нейромережевого методу та інтелектуальної системи. Розроблена система характеризується стійкістю до довгих змішаних текстів, здатністю до локалізованого аналізу дискримінаційних фрагментів і практичною придатністю до інтеграції у сучасні системи автоматизованої модерації контенту.

3.3 Обмеження методу та перспективи подальших досліджень

Розроблений метод нейромережевої ідентифікації гендерної дискримінації у мультимодальному контенті соціальних медіа забезпечує автоматизоване виявлення сексистських висловлювань на основі текстового представлення даних, однак має певні обмеження, які необхідно враховувати під час інтерпретації результатів експериментального дослідження.

Одним із суттєвих обмежень є орієнтація методу переважно на англомовний контент, оскільки навчання та тестування моделей виконувалося на англомовних наборах даних. Це ускладнює безпосереднє застосування системи до україномовних або багатомовних соціальних медіа без додаткового донавчання моделей. Ще одним обмеженням є те, що мультимодальність у межах реалізованої системи фактично зводиться до перетворення аудіо- та

відеоданих у текстову форму. Тому візуальні ознаки відео або зображень безпосередньо не аналізуються, що може знижувати ефективність методу у випадках, коли дискримінаційний зміст передається через меми, жести, візуальні символи або контекст зображення.

На результати роботи системи також впливає якість автоматичного розпізнавання мовлення. Помилки транскрибування, фоновий шум, нечітка вимова або наявність кількох мовців можуть призводити до втрати частини семантичної інформації та впливати на точність подальшої класифікації. Додатковим обмеженням є бінарний характер класифікації, за якого контент поділяється лише на класи «сексизм» та «не сексизм». Такий підхід не дає змоги детально розрізняти типи гендерної дискримінації, рівень її інтенсивності або форму прояву.

Перспективами подальших досліджень є розширення методу для підтримки української та інших мов, формування спеціалізованих багатомовних корпусів дискримінаційного контенту, а також донавчання трансформерних моделей на даних із соціальних медіа. Доцільним напрямом розвитку є перехід до повноцінного мультимодального аналізу, що поєднуватиме текстові, аудіальні та візуальні ознаки. Це дозволить підвищити точність виявлення прихованих або контекстно залежних проявів гендерної дискримінації.

Подальше вдосконалення системи може передбачати впровадження засобів пояснюваного штучного інтелекту для виділення фрагментів тексту, які найбільше вплинули на рішення моделі, а також механізмів аудиту упередженості самої нейромережевої моделі. Практично важливим напрямом є інтеграція системи з інструментами модерації соціальних платформ та реалізація механізму зворотного зв’язку від модератора для поступового покращення якості класифікації.

Отже, незважаючи на наявні обмеження, розроблений метод є придатною основою для подальшого розвитку інтелектуальних систем автоматизованої модерації мультимодального контенту. Його подальше вдосконалення має бути спрямоване на розширення мовної підтримки, поглиблення мультимодального

аналізу, підвищення інтерпретованості результатів та адаптацію до реальних умов функціонування соціальних медіа.

3.4 Висновки до розділу 3

У третьому розділі здійснено програмну реалізацію інтелектуальної системи нейромережевої ідентифікації гендерної дискримінації у мультимодальному контенті соціальних медіа та проведено експериментальне дослідження розробленого методу. Реалізована система забезпечує приймання текстових, аудіо- та відеоданих, їх уніфікацію до текстового представлення, сегментацію, нейромережеву класифікацію та формування підсумкової оцінки ризику.

Проведене порівняння моделей BiLSTM та RoBERTa підтвердило перевагу трансформерної архітектури для задачі виявлення гендерно дискримінаційного контенту. Найвищі показники якості отримано для моделі RoBERTa на збалансованому наборі SSMB: Accuracy становить 87%, Macro F1 – 0,87. Для незбалансованого набору SDET модель RoBERTa також перевищила BiLSTM за основними метриками, що свідчить про кращу здатність трансформерної архітектури враховувати контекстні та семантичні зв’язки у тексті. Додаткова крос-датасетна аугментація забезпечила незначне, але стабільне покращення Macro F1, що підтверджує вплив якості та збалансованості навчальних даних на результат класифікації.

Порівняльне тестування з існуючими програмними рішеннями показало, що запропонований метод ефективніше розпізнає не лише явні, а й приховані прояви сексизму. На контрольних текстових фрагментах конфігурація RoBERTa із крос-датасетною аугментацією забезпечила 100% правильних класифікацій, тоді як сторонні аналоги та модель BiLSTM припускалися помилок під час аналізу завуальованих або контекстно залежних висловлювань.

Загальні висновки

Було досягнуто мету кваліфікаційної роботи – підвищення точності виявлення проявів гендерної дискримінації у мультимодальному контенті соціальних медіа шляхом розроблення методу нейромережевої ідентифікації сексизму на основі архітектур глибокого навчання, а також програмної реалізації інтелектуальної системи для оцінки його ефективності.

Для досягнення поставленої мети були виконані такі задачі:

– проведено аналіз предметної області виявлення проявів гендерної дискримінації у мультимодальному контенті соціальних медіа;

– розроблено метод нейромережевої ідентифікації гендерної дискримінації у мультимодальному контенті соціальних медіа;

– здійснено програмну реалізацію інтелектуальної системи, що забезпечує роботу розробленого методу для аналізу мультимодального контенту на предмет виявлення гендерної дискримінації;

– здійснено дослідження розробленого методу ідентифікації проявів гендерної дискримінації з використанням створеної інтелектуальної системи за допомогою метрик якості та порівняльного аналізу.

Розроблена інтелектуальна система забезпечує аналіз тексту, аудіо та відео, здійснює бінарну класифікацію контенту за ознакою наявності гендерної дискримінації, підтримує обробку великих текстових фрагментів за допомогою алгоритму ковзного вікна та надає інтерпретовані результати аналізу з візуалізацією локальних оцінок ризику.

Результати дослідження точності підтвердили працездатність методу: модель RoBERTa з крос-датасетною аугментацією виявилася найбільш стійкою. На відміну від конфігурації на SSMB, що показувала високу точність лише в ідеалізованих збалансованих умовах, розроблена модель зберегла якість класифікації на реалістичних незбалансованих даних. Порівняльне тестування підтвердило перевагу розробленого рішення: безпомилково розпізнано усі контрольні типи контенту, включаючи прихований сексизм та довгі змішані

тексти, на яких сторонні аналоги (Cardiff NLP, Kimola, PHS Detection) допустили помилки.

Перспективи подальшого вдосконалення розробленого методу та інтелектуальної системи включають розширення мовної підтримки за межі англомовного контенту, інтеграцію мультикласової класифікації для класифікації типів сексизму, збільшення навчальної вибірки за рахунок нових джерел даних, а також оптимізацію швидкодії для застосування в системах модерації контенту в режимі реального часу.

Перелік посилань

1. Данько-Сліпцова А. А., Коваленко Н. А., Жорнокуй У. В. Вплив соціальних мереж на формування громадської думки під час кризових ситуацій: соціологічний аспект. *Scientific notes of V. I. Vernadsky Taurida National University. Series: Philology. Journalism*. 2024. Т. 2, № 3. С. 218–226. URL: https://doi.org/10.32782/2710-4656/2024.3.2/35 (дата звернення: 20.04.2026).

2. Kemp S. Digital 2023: global overview report. *DataReportal*. 2023. URL: https://datareportal.com/reports/digital-2023-global-overview-report (дата звернення: 20.04.2026).

3. Форми гендерної дискримінації. *Північно-Східне міжрегіональне управління Державної служби з питань праці*. URL: https://pns.dsp.gov.ua/news/formy-hendernoi-dyskryminatsii/ (дата звернення: 20.04.2026).

4. Barreto M., Doyle D. M. Benevolent and hostile sexism in a shifting global context. *Nature Reviews Psychology*. 2023. Т. 2, № 2. С. 98–111. URL: https://doi.org/10.1038/s44159-022-00136-x (дата звернення: 20.04.2026).

5. Evaluating the effectiveness of deplatforming as a moderation strategy on Twitter / S. Jhaver та ін. *Proceedings of the ACM on Human-Computer Interaction*. 2021. Т. 5, CSCW2. С. 1–30. URL: https://doi.org/10.1145/3479525 (дата звернення: 20.04.2026).

6. Rodríguez-Sánchez F., Carrillo-de-Albornoz J., Plaza L. Detecting sexism in social media: an empirical analysis of linguistic patterns and strategies. *Applied Intelligence*. 2024. URL: https://doi.org/10.1007/s10489-024-05795-2 (дата звернення: 20.04.2026).

7. Gorwa R., Binns R., Katzenbach C. Algorithmic content moderation: technical and political challenges in the automation of platform governance. *Big Data & Society*. 2020. Т. 7, № 1. URL: https://doi.org/10.1177/2053951719897945 (дата звернення: 20.04.2026).

8. Sasse J., Grossklags J. Breaking the silence: investigating which types of moderation reduce negative effects of sexist social media content. *Proceedings of the ACM on Human-Computer Interaction*. 2023. Т. 7, CSCW2. С. 1–26. URL: https://doi.org/10.1145/3610176 (дата звернення: 20.04.2026).

9. Празднiков В. О., Сугоняк І. І. Моделі та методи машинного навчання для розпізнавання фейкового контенту. *Технічна інженерія*. 2023. № 2(92). С. 131–136. URL: https://doi.org/10.26642/ten-2023-2(92)-131-136 (дата звернення: 20.04.2026).

10. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions / L. Alzubaidi та ін. *Journal of Big Data*. 2021. Т. 8, № 1. URL: https://doi.org/10.1186/s40537-021-00444-8 (дата звернення: 20.04.2026).

11. Xiong Y., Chen G., Cao J. Research on public service request text classification based on BERT-BiLSTM-CNN feature fusion. *Applied Sciences*. 2024. Т. 14, № 14. С. 6282. URL: https://doi.org/10.3390/app14146282 (дата звернення: 20.04.2026).

12. Mukherjee S., Das S. Application of transformer-based language models to detect hate speech in social media. *Journal of Computational and Cognitive Engineering*. 2023. Т. 2, № 4. С. 278–286. URL: https://doi.org/10.47852/bonviewJCCE2022010102 (дата звернення: 20.04.2026).

13. Robust speech recognition via large-scale weak supervision / A. Radford та ін. *Proceedings of the 40th International Conference on Machine Learning (ICML)*. 2023. С. 28492–28518. URL: https://proceedings.mlr.press/v202/radford23a.html (дата звернення: 20.04.2026).

14. Cardiffnlp/twitter-roberta-base-hate-latest. *Hugging Face – The AI community building the future.* URL: https://huggingface.co/cardiffnlp/twitter-roberta-base-hate-latest (дата звернення: 18.05.2026).

15. Leon-Gomez E. A., Álvarez-Meza A. M., Castellanos-Dominguez G. Cross-Dataset data augmentation using UMAP for deep learning-based wind speed prediction. *Computers*. 2025. Т. 14, № 4. С. 123. URL: https://doi.org/10.3390/computers14040123 (дата звернення: 18.05.2026).

16. How to detect hate speech with machine learning? *Kimola*. 2023. URL: https://kimola.com/blog/how-to-detect-hate-speech-with-machine-learning (дата звернення: 23.04.2026).

17. Hate Speech Detector. *Kimola*. URL: https://kimola.com/pretrained-ai-models/hate-speech-detector (дата звернення: 23.04.2026).

18. Try out kNOwHATE prototype for detecting online hate speech. *KnowHate*. 2024. URL: https://knowhate.eu/2024/07/05/try-out-knowhate-prototype-for-detecting-online-hate-speech/ (дата звернення: 23.04.2026).

19. Portuguese hate speech detection prototype. *Hugging Face*. URL: https://huggingface.co/spaces/knowhate/portuguese-hate-speech-detection (дата звернення: 23.04.2026).

20. Sarie D. E., Yuzarni S. The impact of technology and social media on discrimination and harassment in the workplace. *Journal Discrimination and Injustice*. 2025. С. 15–25. URL: https://doi.org/10.70992/yva9c438 (дата звернення: 23.04.2026).

21. Yin Q., Abdullah K. B. B. Analysis of gender discourse bias and gender discrimination in social media: a case study of the TikTok platform. *Journal of Intercultural Communication*. 2024. Т. 24, № 2. С. 93–102. URL: https://doi.org/10.36923/jicc.v24i2.802 (дата звернення: 23.04.2026).

22. Hale M. L. Gender based biases and discrimination – new developments in social media contexts. *ORBilu – University of Luxembourg*. 2023. URL: https://orbilu.uni.lu/handle/10993/58871 (дата звернення: 23.04.2026).

23. Interpretable sexism detection with explainable transformers / S. Rayhana та ін. *CEUR Workshop Proceedings*. 2025. Т. 4017. URL: https://ceur-ws.org/Vol-4017/paper_20.pdf (дата звернення: 23.04.2026).

24. AI model for automated content moderation: integrating Hugging Face model for text and image processing / C. Naik та ін. *Proceedings of the 9th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*. 2025. URL: https://doi.org/10.1109/ICECA66444.2025.11382809 (дата звернення: 23.04.2026).

25. Muntasir F., Noor J. Explainable AI discloses gender bias in sexism detection algorithm. *NSysS '24: 11th International Conference on Networking, Systems, and Security*. New York, NY, USA, 2024. С. 120–127. URL: https://doi.org/10.1145/3704522.3704524 (дата звернення: 23.04.2026).

26. Long text classification model based on transformer sliding window and threshold optimization / J. Pan та ін. *Journal of Internet Technology*. 2025. Т. 26, № 2. С. 231–240. URL: https://doi.org/10.70003/160792642025032602008 (дата звернення: 27.04.2026).

27. Research on sentiment classification for netizens based on the BERT-BiLSTM-TextCNN model / X. Jiang та ін. *PeerJ Computer Science*. 2022. Т. 8. Ст. e1005. URL: https://doi.org/10.7717/peerj-cs.1005 (дата звернення: 05.05.2026).

28. Kula S., Kozik R., Choraś M. Implementation of the BERT-derived architectures to tackle disinformation challenges. *Neural Computing and Applications*. 2021. URL: https://doi.org/10.1007/s00521-021-06276-0 (дата звернення: 05.05.2026).

29. A survey of transformers / T. Lin та ін. *AI Open*. 2022. URL: https://doi.org/10.1016/j.aiopen.2022.10.001 (дата звернення: 05.05.2026).

30. Attention is all you need / A. Vaswani та ін. *Advances in Neural Information Processing Systems*. 2017. URL: https://arxiv.org/abs/1706.03762 (дата звернення: 05.05.2026).

31. Sexism detection in English texts. *Kaggle*. URL: https://www.kaggle.com/datasets/aadyasingh55/sexism-detection-in-english-texts (дата звернення: 05.05.2026).

32. Tum-nlp/sexism-socialmedia-balanced dataset. *Hugging Face*. URL: https://huggingface.co/datasets/tum-nlp/sexism-socialmedia-balanced (дата звернення: 05.05.2026).

33. Grandini M., Bagli E., Visani G. Metrics for multi-class classification: an overview. *arXiv preprint arXiv:2008.05756*. 2020. URL: https://doi.org/10.48550/arXiv.2008.05756 (дата звернення: 05.05.2026).

34. Viering T., Loog M. The shape of learning curves: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2022. С. 1–20. URL: https://doi.org/10.1109/tpami.2022.3220744 (дата звернення: 05.05.2026).

35. Opitz J. A closer look at classification evaluation metrics and a critical reflection of common evaluation practice. *Transactions of the Association for Computational Linguistics*. 2024. Т. 12. С. 820–836. URL: https://doi.org/10.1162/tacl_a_00675 (дата звернення: 05.05.2026).

36. Cooshna-Naik D. Exploring the use of tweets and word clouds as strategies in educational research. *Journal of Learning for Development*. 2022. Т. 9, № 1. С. 89–103. URL: https://doi.org/10.56059/jl4d.v9i1.541 (дата звернення: 05.05.2026).

37. Array programming with NumPy / C. R. Harris та ін. *Nature*. 2020. Т. 585, № 7825. С. 357–362. URL: https://doi.org/10.1038/s41586-020-2649-2 (дата звернення: 08.05.2026).

38. Gupta P., Bagchi A. Introduction to pandas. *Essentials of Python for Artificial Intelligence and Machine Learning*. Cham, 2024. С. 161–196. URL: https://doi.org/10.1007/978-3-031-43725-0_5 (дата звернення: 08.05.2026).

39. PyTorch: an imperative style, high-performance deep learning library / A. Paszke та ін. *Advances in Neural Information Processing Systems*. 2019. Т. 32. С. 8024–8035. URL: https://arxiv.org/abs/1912.01703 (дата звернення: 08.05.2026).

40. Whisper. *OpenAI*. URL: https://openai.com/uk-UA/index/whisper/ (дата звернення: 08.05.2026).

41. MoviePy. *PyPI*. URL: https://pypi.org/project/moviepy/ (дата звернення: 08.05.2026).

42. Streamlit documentation. *Streamlit*. URL: https://docs.streamlit.io/ (дата звернення: 08.05.2026).

ДОДАТКИ

Додаток А

Програмні коди

Програмний код, використаний у дослідженні, доступний у відкритому репозиторії GitHub: https://github.com/Twinker25/ShashokDA_Multimodal-gender-discrimination-detection/tree/main (дата звернення: 22.05.2026).

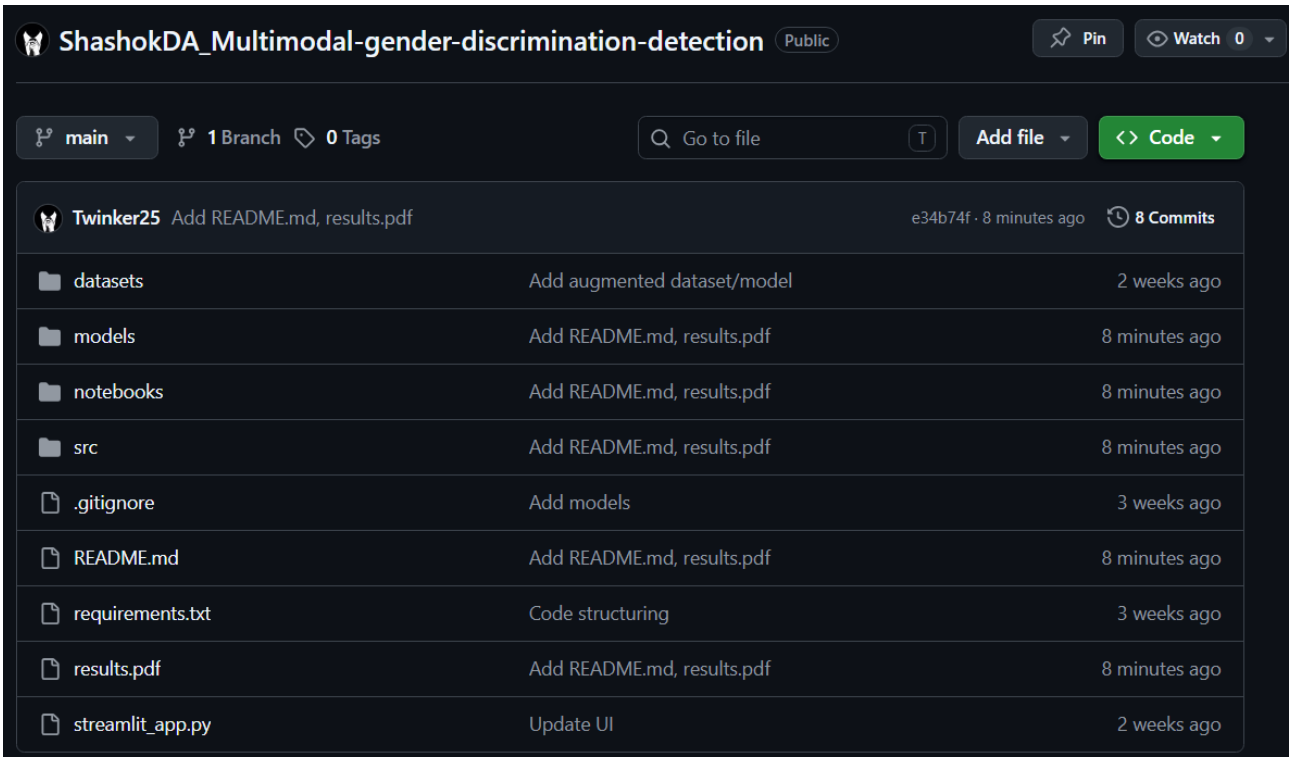


Рисунок А.1 – Головна сторінка репозиторію

Структура репозиторію наступна:

– Головний файл вебінтерфейсу (*streamlit_app.py*), який забезпечує запуск та взаємодію користувача із інтелектуальною системою на базі фреймворку Streamlit;

– Каталог з наборами даних (*datasets/*). Містить навчальні та тестові вибірки (SDET, SSMB) для підготовки й перевірки якості нейромережевих моделей;

– Каталог навчених нейромережевих моделей (*models/*). Зберігає навчені ваги моделей (BiLSTM, RoBERTa), конфігураційні файли та результати оцінювання для подальшого використання системи без повторного навчання;

– Каталог програмного коду (*src/*). Містить модулі для завантаження моделей і обробки тексту, транскрибування аудіо й відео, а також функції для реалізації алгоритму ковзного вікна та розрахунку показників ризику;

– Каталог блокнотів (*notebooks/*). Містить програмний код для навчання нейромережевих моделей, візуалізації процесу навчання.

Додаток Б

Презентаційний матеріал

КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА

МЕТОД НЕЙРОМЕРЕЖЕВОЇ ІДЕНТИФІКАЦІЇ ГЕНДЕРНОЇ ДИСКРИМІНАЦІЇ У СОЦІАЛЬНИХ МЕДІА ЗА МУЛЬТИМОДАЛЬНИМИ ДАНИМИ

Виконав:

студент 4 курсу, група КН-22-1

Даниіл **ШАШОК**

Керівник:

д.філ., ст.викл. кафедри КН

Марина **МОЛЧАНОВА**

Актуальність

Стрімкий розвиток цифрових платформ спричинив значне зростання обсягів мультимодального контенту, що об'єднує текст, аудіо та відео, і водночас створив умови для масового поширення дискримінаційних висловлювань.

Гендерна дискримінація є однією з найпоширеніших форм шкідливого контенту в соціальних медіа. Її прояви нерідко виражаються через іронію, стереотипи та доброзичливий (прихований) сексизм, що унеможливлює ефективне ручне виявлення.

Ручна модерація є трудомісткою та не масштабується в умовах постійного зростання обсягів контенту. Підвищення точності виявлення гендерної дискримінації із застосуванням сучасних нейромережевих засобів є актуальним науково-практичним завданням у сфері ІТ.

Мета і задачі кваліфікаційної роботи

Об'єкт дослідження – процес нейромережевої ідентифікації гендерної дискримінації у мультимодальному контенті соціальних медіа.

Предмет дослідження – нейромережеві засоби, зокрема рекурентні й трансформерні нейромережеві архітектури, для аналізу та класифікації текстових представлень мультимодальних даних.

Мета кваліфікаційної роботи бакалавра – підвищення точності виявлення проявів гендерної дискримінації у мультимодальному контенті соціальних медіа шляхом розроблення методу нейромережевої ідентифікації сексизму на основі архітектур глибокого навчання та програмної реалізації відповідної інтелектуальної системи.

Для досягнення мети необхідно було виконати такі задачі:

- провести аналіз предметної області виявлення проявів гендерної дискримінації у мультимодальному контенті соціальних медіа;
- розробити метод нейромережевої ідентифікації гендерної дискримінації;
- здійснити програмну реалізацію інтелектуальної системи для аналізу тексту, аудіо та відео;
- виконати дослідження розробленого методу за допомогою метрик якості.

Схема методу нейромережевої ідентифікації гендерної дискримінації в мультимодальному контенті

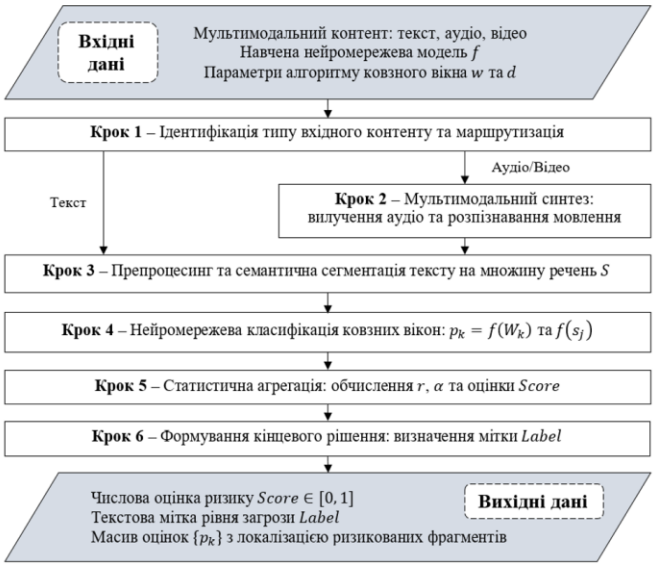


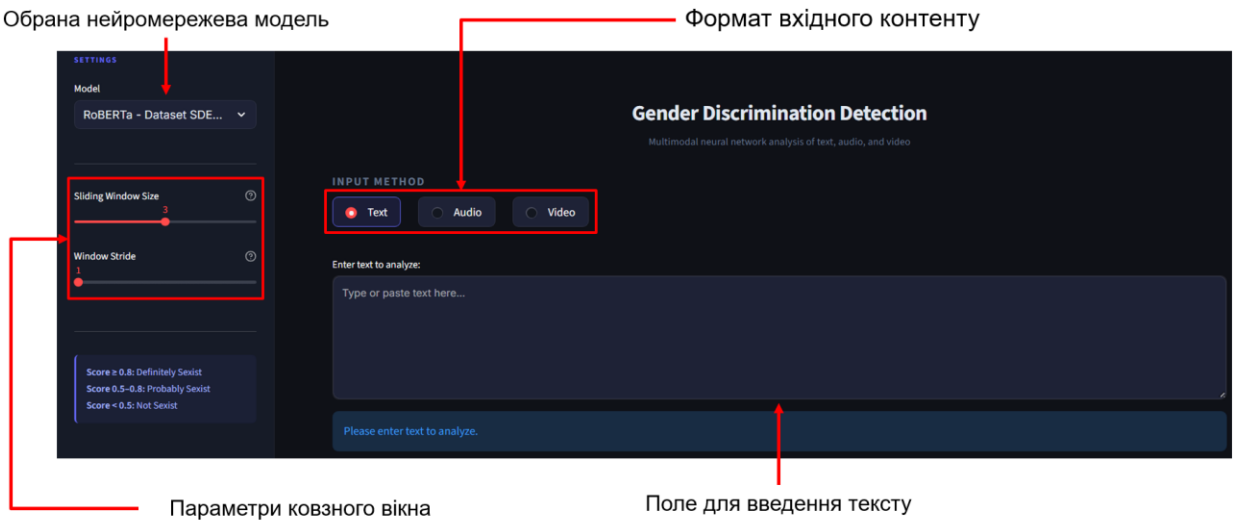
Схема взаємозв'язків програмних компонентів інтелектуальної системи

Засоби розробки інтелектуальної системи

Для програмної реалізації інтелектуальної системи використовувались такі засоби розробки, як:

- Мова програмування Python;
- Фреймворк глибокого навчання PyTorch;
- Бібліотека трансформерних моделей Transformers;
- Система розпізнавання мовлення OpenAI Whisper;
- Бібліотека для обробки відео MoviePy;
- Фреймворк розробки вебінтерфейсу Streamlit;
- Редактор програмного коду Visual Studio Code.

Інтерфейс інтелектуальної системи

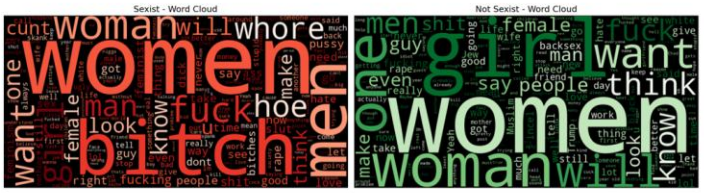


Навчальні набори даних

Характеристика	SDET	SSMB	SDET *(аугментований)*
Повна назва	Sexism Detection in English Texts	Sexism Social Media Balanced	SDET (Augmented)
Структура вибірки	**3 файли** (Train, Dev, Test)	**1 файл**	**3 файли** (Train_aug, Dev, Test)
Розподіл класів	**76%** нейтрально **24%** сексизм	**50%** нейтрально **50%** сексизм	**50%** нейтрально **50%** сексизм
Призначення	Набір для перевірки в реальних умовах	Донор цільових зразків для усунення дисбалансу	Аугментований набір для фінального навчання

Навчальні набори даних

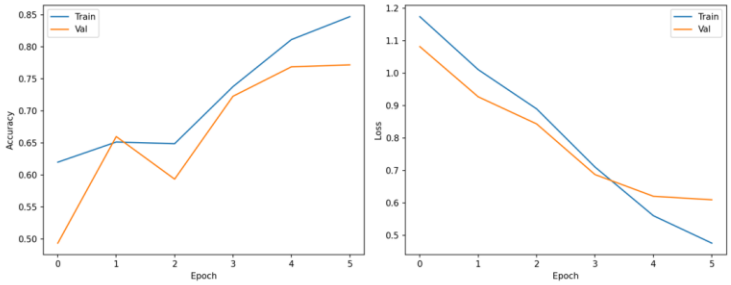
Хмара слів датасету SDET



Хмара слів датасету SSMB

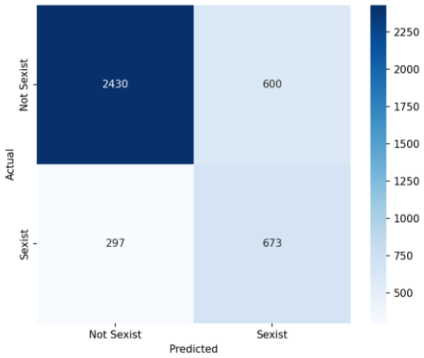
Хмара слів датасету SDET (Augmented)

Результат навчання BiLSTM (датасет SDET)

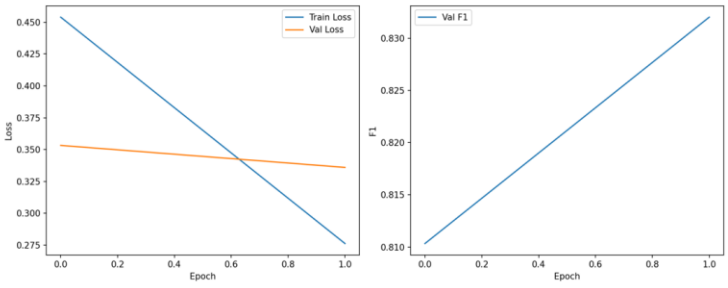


Графік навчання нейромережевої моделі

Матриця плутанини

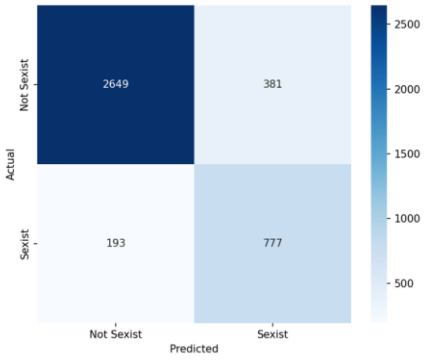


Результат навчання RoBERTa (аугментований SDET)



Графік навчання нейромережевої моделі

Матриця плутанини



Порівняння нейромережевих моделей

Архітектура	Датасет	Accuracy	Precision	Recall	F1-score	Macro F1
BiLSTM	SDET	78%	0.53	0.69	0.60	0.72
BiLSTM	SSMB	82%	0.83	0.79	0.81	0.82
RoBERTa	SDET	85%	0.67	0.80	0.73	0.81
RoBERTa	SSMB	87%	0.87	0.86	0.86	0.87
RoBERTa	**SDET (аугментований)**	**86%**	**0.67**	**0.80**	**0.73**	**0.82**

RoBERTa з крос-датасетною аугментацією є найбільш практично придатною конфігурацією: вона досягає Macro F1 = 0.82 на незбалансованому розподілі даних, що відповідає реальним умовам експлуатації. Конфігурація SSMB демонструє високі результати лише у штучно збалансованому сценарії, який не відображає реального розподілу класів у даних соціальних мереж.

Порівняльне тестування з існуючими програмними рішеннями

Тип текстового фрагмента	Cardiff NLP	Kimola	BiLSTM (SDET)	RoBERTa (аугментований SDET)
Явний сексизм	+ (99.5%)	+ (95.0%)	+ (90.0%)	+ (99.2%)
Прихований сексизм	–	–	–	+ (92.5%)
Нейтральний текст	+ (99.6%)	+ (95.0%)	+ (55.0%)	+ (84.0%)
Довгий змішаний текст	–	+ (95.0%)	+ (81.6%)	+ (97.7%)
Загальний результат	**2 з 4 (50%)**	**3 з 4 (75%)**	**3 з 4 (75%)**	**4 з 4 (100%)**

Умовні позначення:

(+): коректна класифікація; (–): невірна класифікація.

Висновки

Досягнуто мети кваліфікаційної роботи бакалавра – підвищення точності виявлення проявів гендерної дискримінації у мультимодальному контенті соціальних медіа.

Розроблено метод нейромережевої ідентифікації гендерної дискримінації та реалізовано інтелектуальну систему, що підтримує аналіз мультимодального контенту (текст, аудіо, відео), обробку великих фрагментів алгоритмом ковзного вікна та візуалізацію оцінок ризику.

Застосування трансформерної архітектури RoBERTa з крос-датасетною аугментацією вирішило проблему дисбалансу класів, забезпечивши стабільну якість класифікації (Macro F1 = 0.82) на реалістичних даних.

Порівняльне тестування довело перевагу розробленого рішення: правильно розпізнано усі текстові фрагменти, включно з прихованим (доброзичливим) сексизмом та довгими текстами, на яких аналоги зазнали невдачі.

Визначено перспективи розвитку розробленого рішення: розширення мовної підтримки, мультикласова класифікація типів сексизму та оптимізація для модерації в режимі реального часу.

Thu Jun 18 18:14:19 EEST 2026, Петровський Сергій Степанович, Хмельницький національний університет, ХНУ

Anti-Plagiarism (http://ap.km.ua) v-16.718

Максимальне співпадіння з одним документом 3.0%

Словники перевірки: UA, US, RU. **Помилок в документах: 17%**

	Документ		Сумарний збіг по Базі Даних	
ID: 275971 Назва: КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА на тему Метод нейромережевої ідентифікації гендерної дискримінації у соціальних медіа за мультимодальними даними Додано в БД: 2026-06-18 Автора: Даніл ШАШОК Керівники: Марина МОЛЧАНОВА Консультанти: Опоненти:	Символи	Лексеми	Символи	Лексеми
	104930	790	4716 (4%)	68 (9%)

Джерело плагіату

ID	Опис	Наявність плагіату в документі	
		Символи	Лексеми

Протокол аналізу звіту подібності науковим керівником

Заявляю, що я ознайомився (-лась) з Повним звітом подібності, який був згенерований Системою виявлення і запобігання плагіату щодо роботи:

Автор: Данііл ШАШОК

Співавтор:

Назва: КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА на тему Метод нейромережевої ідентифікації гендерної дискримінації у соціальних медіа за мультимодальними даними

Науковий керівник: Марина МОЛЧАНОВА, д-р філ., ст.викл. каф. КН

Підрозділ: Кафедра комп'ютерних наук

Коефіцієнт подібності 1:4.2%

Коефіцієнт подібності 2:2.22%

Мікропробіли: 0

Заміна букв: 0

Інтервали: 0

Білі знаки: 7

Дата створення звіту: 2026-06-18 07:17:42.0

Після аналізу Звіту подібності констатую наступне:

☒ **Запозичення, виявлені в роботі є законними і не є плагіатом. Рівень подібності не перевищує допустимої межі. Таким чином робота незалежна і приймається.**

☐ **Запозичення не є плагіатом, але перевищено граничне значення рівня подібностей. Таким чином робота повертається на доопрацювання.**

☐ **Виявлено запозичення і плагіат або навмисні текстові спотворення (маніпуляції), як передбачувані спроби укриття плагіату, які роблять роботу невідповідною вимогам законодавства (Ст. 32. ЗУ Про вищу освіту, пункт 3.1, Ст. 42. ЗУ Про освіту) та вимог НАЗЯВО (Критерій 5), а також кодексу етики і процедурам. Таким чином робота не приймається.**

Обгрунтування:

2026-06-18

Дата

експерт Петровський В. І. ст.

РІШЕННЯ ЕКСПЕРТНОЇ КОМІСІЇ КАФЕДРИ КОМП'ЮТЕРНИХ НАУК

ПРО ДОПУСК КВАЛІФІКАЦІЙНОЇ РОБОТИ ДО ЗАХИСТУ

Назва кваліфікаційної роботи *Метод нейромережевої ідентифікації гендерної дискримінації у соціальних медіа за мультимодальними даними*
Автор *студент групи КН-22-1 Шашок Даниіл Анатолійович*
Освітня програма *Комп'ютерні науки*
Рівень вищої освіти *перший (бакалаврський)*
Спеціальність *122 – Комп'ютерні науки*
Науковий керівник: *ст. викладач каф. КН, д-р філософії з комп. наук Марина МОЛЧАНОВА*

На основі аналізу кваліфікаційної роботи на дотримання вимог академічної доброчесності (у т.ч. відсутності ознак академічного плагіату) з урахуванням результатів перевірки роботи спеціалізованим програмними засобами комісія зробила такий висновок:

№	Висновок	Позначка про відповідність
1	Ознаки академічного плагіату	
1.1	Запозичення, виявлені в роботі, є законними і не є академічним плагіатом (далі – зазначаються підстави віднесення запозичень до правомірних, якщо потрібно). Робота приймається до захисту.	***відповідає***
1.2	Виявлені запозичення не є академічним плагіатом, розміщені в розділах, які не описують безпосередньо авторське дослідження, але кількість цитат перевищує обсяг, виправданий поставленою метою роботи (далі – зазначаються детальні та аргументовані підстави віднесення запозичень до правомірних). Робота приймається до захисту, але має бути відкоригована.	
1.3	Виявлені запозичення не є академічним плагіатом, але частково розміщені в розділах, які описують безпосередньо авторське дослідження, а кількість цитат перевищує обсяг, виправданий поставленою метою роботи. Робота може бути допущена до захисту після того як буде відкоригована та доопрацьована і успішно пройде повторну перевірку на академічний плагіат.	
1.4	Робота містить навмисні текстові спотворення, передбачувані спроби укриття текстових запозичень або інші прояви академічного плагіату. Робота містить фабрикацію або фальсифікацію даних. Робота не допускається до захисту.	
2	Інші види порушень академічної доброчесності	***відсутні***

Підтвердження:

Запозичення, виявлені в роботі Даниіла Шашка, не є плагіатом, оскільки: запозичення розміщені в розділі огляду існуючих підходів, не описують безпосередньо авторську роботу і не стосуються її результатів; усі запозичення фрагментарні; до запозичень входять фрагменти, які не мають авторства і містять поширені конструкції та загальновідомі терміни, скорочення. Рівень подібності не перевищує допустимої межі. Таким чином, робота є законною та приймається до захисту.

Обсяг запозичень, визначений системами виявлення збігів/ідентичності/схожості:
- за системою Anti-Plagiarism: 3%;
- за системою StrikePlagiarism КП1: 4.2%.

18.06.2025

Завідувач кафедри _____ *Олександр БАРМАК*

Гарант освітньої програми _____ *Олександр МАЗУРЕЦЬ*

Керівник кваліфікаційної роботи _____ *Марина МОЛЧАНОВА*

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
МОН УКРАЇНИ

Кафедра комп'ютерних наук

ВІДГУК НАУКОВОГО КЕРІВНИКА

на кваліфікаційну роботу бакалавра

студента *гр. КН-22-1 Шашка Даниіла Анатолійовича*

за темою *Метод нейромережевої ідентифікації гендерної дискримінації у соціальних медіа за мультимодальними даними*

1. Актуальність теми

Тема роботи є актуальною у зв'язку з поширенням соціальних медіа та необхідністю автоматизованого виявлення проявів гендерної дискримінації. Особливу значущість має використання мультимодальних даних, оскільки дискримінаційний зміст може передаватися як текстовими, так і візуальними засобами. Робота відповідає сучасним напрямам розвитку етичного аналізу контенту, машинного навчання та інтелектуальної модерації інформаційного середовища.

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

Зміст роботи відповідає предметній області спеціальності 122 «Комп'ютерні науки», оскільки охоплює аналіз даних, математичне моделювання, алгоритмізацію, застосування нейромережевих методів і розроблення програмного забезпечення. У роботі виконано постановку прикладної задачі, підготовку мультимодальних даних, вибір моделей, програмну реалізацію та експериментальне оцінювання результатів.

3. Професійні та особистісні якості бакалавра

Під час виконання кваліфікаційної роботи студент продемонстрував достатній рівень фахової підготовки, відповідальність і наполегливість. Він уміє працювати з науковими джерелами, аналізувати дані та приймати обґрунтовані технічні рішення. Здобувач послідовно виконував поставлені завдання, уважно ставився до рекомендацій керівника та прагнув забезпечити належний рівень обґрунтованості практичної частини.

4. Ступінь самостійності під час виконання кваліфікаційної роботи

Основні етапи дослідження виконано студентом самостійно. Ним проведено аналіз предметної області, підготовлено дані, обрано й реалізовано нейромережеві моделі, виконано експерименти та сформульовано висновки. Консультації керівника мали спрямувальний характер і стосувалися переважно методики дослідження, структури

роботи та інтерпретації отриманих результатів. Рівень самостійності відповідає вимогам до кваліфікаційної роботи бакалавра.

5. Ступінь оволодіння методами дослідження

Студент продемонстрував належне володіння методами наукового дослідження – від аналізу фахової літератури та формалізації задачі до програмної реалізації й експериментальної перевірки. У роботі обґрунтовано застосування нейромережевих засобів для спільного аналізу текстового та візуального контенту. Здобувач не лише використав відповідні програмні бібліотеки та моделі, а й пояснив їх роль у загальній структурі запропонованого методу.

6. Повнота та якість розкриття теми роботи

Тему роботи розкрито послідовно та достатньо повно. Представлено аналіз предметної області, огляд існуючих підходів, обґрунтування розробленого методу, опис програмної реалізації та результати експериментальної перевірки. Теоретичні положення логічно пов'язані з практичною частиною, а конкретизація етапів розроблення дозволяє оцінювати дослідження як цілісне та завершене.

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

Матеріал викладено логічно, послідовно та зрозуміло. Структура роботи дозволяє простежити перехід від постановки проблеми й аналізу існуючих підходів до розроблення методу, його реалізації та оцінювання результатів.

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

Результати роботи можуть використовуватися як основа для систем автоматизованого моніторингу й модерації контенту в соціальних медіа. Розроблений метод має потенціал для інтеграції у сервіси аналізу мультимодальних даних, а також для подальшого розширення переліку виявлюваних проявів дискримінації.

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

Кваліфікаційна робота Шашка Даниїла Анатолійовича має завершений характер, відповідає темі та предметній області спеціальності 122 «Комп'ютерні науки». Робота може бути допущена до захисту. Рекомендована оцінка — «відмінно».

Керівник _____ *ст. викладач каф. КН, д-р філософії Марина МОЛЧАНОВА*

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
МОН УКРАЇНИ

Кафедра комп'ютерних наук

РЕЦЕНЗІЯ

на кваліфікаційну роботу бакалавра

студента *гр. КН-22-1 Шашка Даниіла Анатолійовича*
за темою: *Метод нейромережевої ідентифікації гендерної дискримінації у соціальних медіа за мультимодальними даними*

1. Актуальність обраної теми

Робота присвячена актуальній задачі автоматизованого виявлення проявів гендерної дискримінації у соціальних медіа. Її значущість зумовлена потребою у підвищенні безпеки цифрової комунікації та вдосконаленні засобів моніторингу дискримінаційного контенту. Тема має міждисциплінарний характер, оскільки поєднує методи машинного навчання, аналіз мультимодальних даних і соціально значущу прикладну проблему. Це забезпечує практичну спрямованість роботи та перспективність її подальшого розвитку.

2. Повнота розкриття мети та завдань роботи

Мету й завдання роботи розкрито послідовно та достатньо повно. Автор виконав аналіз предметної області, формалізував задачу, обґрунтував вибір нейромережевих засобів і реалізував відповідне програмне рішення. Дослідження доведено до експериментальної перевірки, а кожний його етап логічно спирається на результати попереднього. Це свідчить про цілісність роботи та розуміння автором методики проведення дослідження.

3. Зміст кожного розділу роботи

У першому розділі розглянуто предметну область, прояви гендерної дискримінації у соціальних медіа та сучасні підходи до аналізу відповідного контенту. Другий розділ присвячено розробленню методу нейромережевої ідентифікації дискримінаційних проявів за текстовими й візуальними даними, а також опису етапів їх підготовки та критеріїв оцінювання. У третьому розділі представлено програмну реалізацію і результати експериментальної перевірки. Структура роботи дозволяє простежити перехід від постановки задачі до оцінювання ефективності створеного рішення.

4. Оцінка розробленої інформаційної системи, її практична цінність

Розроблена система має прикладний потенціал для автоматизованого моніторингу контенту у соціальних медіа та виявлення можливих проявів гендерної дискримінації. Використання мультимодальних даних дозволяє враховувати як текстову, так і візуальну складові повідомлень. Практична цінність роботи полягає у скороченні обсягу ручного аналізу та створенні передумов для більш оперативного й послідовного опрацювання цифрового контенту. Запропоноване рішення може бути основою для подальшої інтеграції у системи модерації та інформаційного моніторингу.

5. Якість оформлення кваліфікаційної роботи бакалавра

Кваліфікаційну роботу оформлено акуратно та відповідно до основних академічних вимог. Матеріал подано логічно й послідовно, а теоретична і практична частини належно узгоджені між собою. Таблиці та ілюстративні матеріали доповнюють текст і сприяють

кращому розумінню отриманих результатів. Загальне оформлення створює позитивне враження від роботи.

6. Недоліки кваліфікаційної роботи бакалавра

До незначних недоліків можна віднести те, що окремі пояснення параметрів експериментального дослідження подано дещо стисло. Їх детальніше висвітлення полегшило б відтворення окремих етапів перевірки методу. Зазначене зауваження має рекомендаційний характер і не впливає на обґрунтованість отриманих результатів та загальну якість роботи.

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

Кваліфікаційна робота Шашка Даниіла Анатолійовича є завершеним самостійним дослідженням, у якому поєднано аналіз предметної області, розроблення нейромережевого методу, програмну реалізацію та експериментальну перевірку. Робота відповідає предметній області спеціальності 122 «Комп'ютерні науки», має практичну спрямованість і демонструє достатній рівень фахової підготовки автора. Рекомендована оцінка — «відмінно».

Рецензент _____ Бедрина І.А.