Хмельницький національний університет
Факультет інформаційних технологій
Кафедра комп'ютерних наук

КВАЛІФІКАЦІЙНА РОБОТА

на тему *Метод виявлення мови ворожнечі у зашумлених соціальних текстових даних нейромережевими засобами*

Рівень вищої освіти *другий (магістерський)*

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 2 курсу, група КНм-24-1* (Курс, група виконавця) _____ (Підпис) *Ілля БОЯРЧУК* (Ім'я, ПРІЗВИЩЕ)

Керівник: *Ph.D., ст. викл. кафедри КН* (Науковий ступінь, посада) _____ (Підпис) *Марина МОЛЧАНОВА* (Ім'я, ПРІЗВИЩЕ)

Нормоконтроль: *к.т.н., доцент кафедри 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)

16 грудня 2025 р.

Хмельницький 2025

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

ЗАТВЕРДЖУЮ
Завідувач кафедри *комп'ютерних наук*

_____
(підпис)
д.т.н., професор Олександр БАРМАК
«*28*» *серпня* *2025* року

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

1. Тема кваліфікаційної роботи магістра: *«Метод виявлення мови ворожнечі у зашумлених соціальних текстових даних нейромережевими засобами».*

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

3. Керівник роботи *старший викладач кафедри КН Марина МОЛЧАНОВА*
(Ім'я, ПРІЗВИЩЕ)

4. Затверджені наказом університету від «*25*» *серпня* *2025* р. № *65.*

5. Дата видачі завдання студенту: «*28*» *серпня* *2025* р.

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

Метою кваліфікаційної роботи є підвищення точності виявлення мови ворожнечі у зашумлених соціальних текстових даних. Для досягнення мети слід вирішити такі задачі: провести аналіз природи мови ворожнечі та її класифікаційних ознак; виконати огляд існуючих підходів до виявлення мови ворожнечі, виконати аналіз наукових досліджень; охарактеризувати етичні аспекти автоматизованого виявлення мови ворожнечі; розробити метод виявлення мови ворожнечі у зашумлених соціальних текстових даних нейромережевими засобами; виконати підготовку датасету для файнтюнінгу нейромережі для виявлення мови ворожнечі; виконати програмну реалізацію розробленого методу; провести дослідження методу виявлення мови ворожнечі у зашумлених соціальних текстових даних нейромережевими засобами.

7. Календарний план виконання кваліфікаційної роботи:

№	Назва етапів (розділів) кваліфікаційної роботи бакалавра	Термін виконання	Примітка
1	*Вибір напряму дослідження та узгодження теми кваліфікаційної роботи з керівником, складання календарного графіка виконання роботи*	*вересень 2025*	*Виконано*
2	*Ознайомлення з предметною областю, аналіз існуючих методів і моделей, формулювання мети та завдань дослідження, визначення об'єкта й предмета дослідження*	*вересень 2025*	*Виконано*
3	*Розробка методу чи моделі для вирішення обраного завдання, опис архітектури рішення*	*жовтень 2025*	*Виконано*
4	*Програмна реалізація методу чи моделі*	*жовтень 2025*	*Виконано*
5	*Дослідження ефективності та експериментальна перевірка результатів, порівняння з відомими підходами*	*листопад 2025*	*Виконано*
6	*Написання пояснювальної записки, оформлення відповідно до вимог, врахування зауважень керівника*	*листопад 2025*	*Виконано*
7	*Підготовка презентаційних матеріалів та попередній захист*	*листопад 2025*	*Виконано*
8	*Перевірка пояснювальної записки на відповідність вимогам оформлення (нормоконтроль) та перевірка на академічну доброчесність. Отримання відгуку керівника та рецензії.*	*грудень 2025*	*Виконано*
9	*Публічний захист кваліфікаційної роботи*	*грудень 2025*	*Виконано*

Виконавець: *студент групи КНм-24-1* (Група виконавця) _____ (Підпис) *Ілля БОЯРЧУК* (Ім'я, ПРІЗВИЩЕ)

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

Реферат

Кваліфікаційна робота присвячена вирішенню науково-технічної задачі автоматизованого виявлення мови ворожнечі у зашумлених соціальних текстових даних шляхом розроблення нейромережевого методу, здатного інтерпретувати спотворені, змішаномовні та навмисно масковані мовні конструкції. Досягнення цієї мети забезпечується поєднанням механізмів автоматичного моделювання шумових викривлень у навчальних корпусах із адаптивним налаштуванням архітектури глибинного класифікатора, орієнтованого на роботу з контекстно нестабільними текстовими вхідними даними.

Актуальність теми визначається стрімким зростанням обсягу цифрових комунікацій, у межах яких мова ворожнечі поширюється не лише відкритими формулюваннями, а й через навмисно спотворені, суржикові, масковані та контекстно приховані мовні конструкції. У соціальних мережах, месенджерах і коментарних платформах агресивні висловлювання часто набувають форм, що унеможливлюють їх виявлення традиційними алгоритмами, орієнтованими на нормативно структуровані дані. Наявні методи автоматичного аналізу тексту не враховують системних проявів шуму, мовного змішування, орфографічної девіації та свідомого уникнення прямої лексичної агресії, що знижує їхню ефективність у практичних застосуваннях.

Зростання обсягу інформаційних потоків супроводжується появою нових стратегій мовного маскування, спрямованих на обходження автоматизованої модерації, що актуалізує потребу в методах, здатних інтерпретувати семантично агресивні висловлювання попри їхню формальну деформацію. Цифрові платформи стикаються з ризиками радикалізації, розпалювання ворожнечі та координації деструктивних комунікацій, що посилює суспільну та безпекову значущість створення стійких до шуму нейромережевих рішень. У цих умовах виникає наукова і практична необхідність побудови моделей, здатних виявляти приховані прояви агресії в динамічних, мовно нестабільних і свідомо викривлених текстах, що формують сучасне комунікаційне середовище.

Мета і задачі роботи. Метою кваліфікаційної роботи магістра є підвищення точності виявлення мови ворожнечі у зашумлених соціальних текстових даних шляхом розроблення нейромережевого методу, здатного інтерпретувати спотворені, змішаномовні та навмисно масковані мовні конструкції. На відміну від існуючих підходів, що орієнтовані переважно на попередньо очищені або стандартизовані текстові вибірки, запропонований метод ґрунтується на відтворенні контрольованих шумових спотворень і цілеспрямованому тренуванні нейромережевої моделі в умовах мовної нестабільності, маскування агресивної лексики та змішаності мовних кодів. Такий підхід дозволяє не лише зберігати релевантність класифікації за наявності орфографічних, графічних і семантичних викривлень, а й підвищувати стійкість моделі до свідомого приховування мовленнєвої агресії, що не забезпечується більшістю відомих рішень.

Для досягнення мети слід вирішити такі задачі:

– провести аналіз природи мови ворожнечі та її класифікаційних ознак;

– виконати огляд існуючих підходів до виявлення мови ворожнечі, виконати аналіз наукових досліджень;

– охарактеризувати етичні аспекти автоматизованого виявлення мови ворожнечі;

– розробити метод виявлення мови ворожнечі у зашумлених соціальних текстових даних нейромережевими засобами;

– виконати підготовку датасету для файнтюнінгу нейромережі для виявлення мови ворожнечі;

– виконати програмну реалізацію розробленого методу;

– провести дослідження методу виявлення мови ворожнечі у зашумлених соціальних текстових даних нейромережевими засобами.

Об’єкт дослідження. Процес автоматизованого виявлення мови ворожнечі у соціальних текстових даних із нестабільною мовною структурою та наявністю шумових викривлень.

Предмет дослідження. Моделі, методи та засоби обробки природної мови для автоматизованого виявлення мови ворожнечі у соціальних текстових даних із нестабільною мовною структурою та наявністю шумових викривлень

Методи дослідження, що використані для вирішення поставлених завдань є наступними: методи нейромережевого аналізу тексту для виявлення мови ворожнечі у шумових та змішаномовних даних, методи автоматизованого моделювання лінгвістичних спотворень, методи математичної статистики для оцінювання ефективності запропонованого підходу.

Наукова новизна одержаних результатів полягає у розробленні нейромережевого методу виявлення мови ворожнечі, адаптованого до спотворених, змішаномовних і навмисно маскованих текстових даних, а також у впровадженні механізму автоматизованого формування контрольовано зашумлених корпусів, що забезпечує підвищення стійкості класифікації до мовних викривлень і прихованих агресивних формулювань.

Апробація результатів кваліфікаційної роботи та публікації. За темою кваліфікаційної роботи автором виконано дві наукові публікації, в тому числі підготовлено до публікації статтю у фаховому виданні категорії Б. Основні наукові й практичні результати роботи доповідались у доповіді «Підхід до нейромережевого виявлення мови ворожнечі у зашумлених тексових повідомленнях» на XVII Всеукраїнській науково-практичній конференції «Актуальні проблеми комп’ютерних наук АПКН-2025» (м. Хмельницький) 14-15 листопада 2025 року.

Структура та обсяг роботи. Кваліфікаційна робота складається з реферату, завдання, змісту, переліку скорочень, вступу, 4-х розділів, висновків, переліку посилань із 69 найменувань та 8 додатків. Обсяг основного тексту кваліфікаційної роботи магістра становить 89 сторінок. У роботі наведено 24 рисунки і 3 таблиці.

Ключові слова: мова ворожнечі, зашумлені тексти, обробка природної мови, трансформерні моделі.

Зміст

Перелік скорочень

Скорочення, термін, позначення	Пояснення
ООН	Організація Об’єднаних Націй
РЄ	Рада Європи
ЄС	Європейський Союз
ECRI	European Commission against Racism and Intolerance
FRA	European Union Agency for Fundamental Rights
UNESCO	Організація Об’єднаних Націй з питань освіти науки і культури
OSCE	Organization for Security and Co operation in Europe
NLP	Natural Language Processing
TF IDF	Term Frequency Inverse Document Frequency
CNN	Convolutional Neural Network
LSTM	Long Short Term Memory
GRU	Gated Recurrent Unit
BERT	Bidirectional Encoder Representations from Transformers
RoBERTa	Robustly Optimized BERT Approach
XLM R	Cross lingual Language Model RoBERTa
API	Application Programming Interface

Вступ

Актуальність теми визначається стрімким зростанням обсягу цифрових комунікацій, у межах яких мова ворожнечі поширюється не лише відкритими формулюваннями, а й через навмисно спотворені, суржикові, масковані та контекстно приховані мовні конструкції. У соціальних мережах, месенджерах і коментарних платформах агресивні висловлювання часто набувають форм, що унеможливлюють їх виявлення традиційними алгоритмами, орієнтованими на нормативно структуровані дані. Наявні методи автоматичного аналізу тексту не враховують системних проявів шуму, мовного змішування, орфографічної девіації та свідомого уникнення прямої лексичної агресії, що знижує їхню ефективність у практичних застосуваннях.

Зростання обсягу інформаційних потоків супроводжується появою нових стратегій мовного маскування, спрямованих на обходження автоматизованої модерації, що актуалізує потребу в методах, здатних інтерпретувати семантично агресивні висловлювання попри їхню формальну деформацію. Цифрові платформи стикаються з ризиками радикалізації, розпалювання ворожнечі та координації деструктивних комунікацій, що посилює суспільну та безпекову значущість створення стійких до шуму нейромережевих рішень. У цих умовах виникає наукова і практична необхідність побудови моделей, здатних виявляти приховані прояви агресії в динамічних, мовно нестабільних і свідомо викривлених текстах, що формують сучасне комунікаційне середовище.

Мета і задачі роботи. Метою кваліфікаційної роботи є підвищення точності виявлення мови ворожнечі у зашумлених соціальних текстових даних шляхом розроблення нейромережевого методу, здатного інтерпретувати спотворені, змішаномовні та навмисно масковані мовні конструкції. На відміну від існуючих підходів, що орієнтовані переважно на попередньо очищені або стандартизовані текстові вибірки, запропонований метод ґрунтується на відтворенні контрольованих шумових спотворень і цілеспрямованому тренуванні нейромережевої моделі в умовах мовної нестабільності, маскування агресивної лексики та змішаності мовних

кодів. Такий підхід дозволяє не лише зберігати релевантність класифікації за наявності орфографічних, графічних і семантичних викривлень, а й підвищувати стійкість моделі до свідомого приховування мовленнєвої агресії, що не забезпечується більшістю відомих рішень.

Для досягнення мети слід вирішити такі задачі:

– провести аналіз природи мови ворожнечі та її класифікаційних ознак;

– виконати огляд існуючих підходів до виявлення мови ворожнечі, виконати аналіз наукових досліджень;

– охарактеризувати етичні аспекти автоматизованого виявлення мови ворожнечі;

– розробити метод виявлення мови ворожнечі у зашумлених соціальних текстових даних нейромережевими засобами;

– виконати підготовку датасету для файнтюнінгу нейромережі для виявлення мови ворожнечі;

– виконати програмну реалізацію розробленого методу;

– провести дослідження методу виявлення мови ворожнечі у зашумлених соціальних текстових даних нейромережевими засобами.

Об’єкт дослідження. Процес автоматизованого виявлення мови ворожнечі у соціальних текстових даних із нестабільною мовною структурою та наявністю шумових викривлень.

Предмет дослідження. Моделі, методи та засоби обробки природної мови для автоматизованого виявлення мови ворожнечі у соціальних текстових даних із нестабільною мовною структурою та наявністю шумових викривлень.

Методи дослідження, що використані для вирішення поставлених завдань є наступними: методи нейромережевого аналізу тексту для виявлення мови ворожнечі у шумових та змішаномовних даних, методи автоматизованого моделювання лінгвістичних спотворень, методи математичної статистики для оцінювання ефективності запропонованого підходу.

Наукова новизна одержаних результатів полягає у розробленні нейромережевого методу виявлення мови ворожнечі, адаптованого до спотворених,

змішаномовних і навмисно маскованих текстових даних, а також у впровадженні механізму автоматизованого формування контрольовано зашумлених корпусів, що забезпечує підвищення стійкості класифікації до мовних викривлень і прихованих агресивних формулювань.

Апробація результатів кваліфікаційної роботи та публікації. За темою кваліфікаційної роботи автором виконано дві наукові публікації, в тому числі підготовлено до публікації статтю у фаховому виданні категорії Б. Основні наукові й практичні результати роботи доповідались у доповіді «Підхід до нейромережевого виявлення мови ворожнечі у зашумлених тексових повідомленнях» на XVII Всеукраїнській науково-практичній конференції «Актуальні проблеми комп'ютерних наук АПКН-2025» (м.Хмельницький) 14-15 листопада 2025 року.

Структура та обсяг роботи. Кваліфікаційна робота магістра складається з реферату, завдання, змісту, переліку скорочень, вступу, 4-х розділів, висновків, переліку посилань із 69 найменувань та 8 додатків. Обсяг основного тексту кваліфікаційної роботи магістра становить 89 сторінок. У роботі наведено 24 рисунки і 3 таблиці.

РОЗДІЛ 1 Дослідження сучасного стану щодо виявлення мови ворожнечі у текстових даних

1.1 Основні поняття, аналіз природи мови ворожнечі та її класифікаційні ознаки

У міжнародному правовому полі мова ворожнечі трактується як форма публічної або міжособистісної комунікації, що поширює, підбурює, заохочує чи виправдовує ненависть, дискримінацію або нетерпимість щодо окремих осіб чи соціальних груп за так званими захищеними ознаками, до яких належать, зокрема, раса, етнічне походження, національність, релігія, мова, стать чи інші невід'ємні характеристики [1, 2]. Таке розуміння підкреслює не лише зміст висловлювань, а й їхній потенційний соціальний ефект, пов'язаний із нормалізацією агресії та виключення. Класичним нормативним джерелом у цій сфері вважається Рекомендація Комітету міністрів Ради Європи № R(97)20, яка закріплює базові принципи протидії hate speech і акцентує увагу на особливій відповідальності медіа та інших каналів масової комунікації за недопущення поширення дискримінаційних наративів [3]. Документ також наголошує на необхідності балансу між свободою вираження поглядів і захистом прав та гідності людини.

В архітектурі нормативних і політичних документів Організації Об'єднаних Націй поняття мови ворожнечі подається як образлива, принизлива або підбурлива комунікація, спрямована проти групи чи індивіда на підставі їхніх невід'ємних характеристик. Для уніфікації підходів до тлумачення цього явища ООН підтримує спеціалізований інформаційно-освітній портал «What is hate speech?», а також супровідні матеріали Глобальної стратегії та Плану дій (2019), які окреслюють спільні критерії ідентифікації, профілактики та реагування на прояви мови ворожнечі в офлайн- і онлайн-середовищі [4]. Ці документи підкреслюють міждисциплінарний характер проблеми та необхідність поєднання правових, освітніх і технологічних інструментів.

В українському контексті термін «мова ворожнечі» розглядається переважно в площині інформаційної безпеки, медіарегулювання та цифрових комунікацій, що зумовлено гібридними загрозами й активною роллю онлайн-платформ у формуванні суспільної думки [5]. У висновках Комісії з журналістської етики зазначається, що мова ворожнечі охоплює висловлювання, спрямовані на приниження честі й гідності людини, стигматизацію, дегуманізацію або колективне узагальнення груп за етнічними, мовними, соціальними, релігійними чи іншими ознаками [6]. Таке трактування фокусується на професійних стандартах журналістики та відповідальності медіа за наслідки публічної риторики. На законодавчому рівні Закон України «Про медіа» встановлює пряму заборону на поширення матеріалів, що містять заклики до ворожнечі, насильства чи дискримінації за національною, расовою, мовною, релігійною або іншою ознакою, тим самим інтегруючи міжнародні підходи у національну правову систему [7].

Природа мови ворожнечі є багаторівневою та принципово контекстозалежною, оскільки агресивний або дискримінаційний намір може реалізовуватися не лише через прямі інвективи й відверті образи, а й у значно більш опосередкованих формах. До них належать іронічні висловлювання, евфемістичні заміни, використання кодових позначень і внутрішньогрупових символів, навмисні графічні та орфографічні спотворення, змішування мовних кодів, транслітерація та інші прийоми маскування змісту, що є особливо характерними для цифрових середовищ і соціальних мереж [8, 9]. Такі стратегії дозволяють авторам зберігати агресивний посил, одночасно уникаючи прямої відповідальності. Європейські інституції, зокрема Європейська комісія проти расизму та нетерпимості (ECRI) та Агенція ЄС з фундаментальних прав (FRA), у своїх аналітичних і моніторингових матеріалах системно фіксують зсув від відкритих форм ворожого мовлення до гібридних, латентних і контекстуально зашифрованих проявів, які істотно ускладнюють як правозастосування, так і автоматизовану модерацію цифрового контенту [12].

Мова ворожнечі має не лише лінгвістичну, а й виразну соціокультурну, психологічну та прагматичну природу, що безпосередньо зумовлює складність її формалізації та автоматизованого розпізнавання методами комп'ютерної обробки мови [13]. У цифрових комунікаціях агресія рідко обмежується прямою образою або лайкою, натомість вона виявляється через узуальні й метафоричні конструкції, принизливі порівняння, дегуманізуючі образи, вторинні та імпліцитні значення, а також алюзії на історичні події, політичні наративи чи колективні травми [14]. Окрему проблему становлять випадки, коли ворожість не артикулюється відкрито, а маскується під іронію, псевдогумор, сарказм або ігрові форми мовлення, а також реалізується через гібридну лексику й напівшифровані мовні коди, що функціонують усередині певних онлайн-спільнот. Європейська комісія проти расизму та нетерпимості (ECRI) у щорічних звітах наголошує на прогресуючій трансформації мови ворожнечі в бік латентних та так званих «імовірнісних» форм, які навмисно знижують ризик правових наслідків для автора, але водночас зберігають принижувальний, дискримінаційний або насильницький характер [15].

Аналітичні дослідження Агенції ЄС з фундаментальних прав демонструють, що цифрове середовище сприяє поширенню феномену «нормалізованої агресії», за якого образливі висловлювання поступово втрачають очевидну вербальну форму та замінюються на соціокультурні коди, символи, меми й стилізовані мовні варіанти, зрозумілі насамперед усередині конкретних ідеологічних або субкультурних груп [16]. У такому контексті межа між прийнятною критикою та мовою ворожнечі стає розмитою, що ускладнює як суспільне реагування, так і технічну фільтрацію контенту.

З огляду на завдання автоматизованої детекції у фокусі дослідників опиняються класифікаційні ознаки, придатні до обчислюваної інтерпретації. До них відносять спрямованість висловлювання на приниження, стигматизацію або дегуманізацію захищеної групи; наявність індикаторів підбурювання, виправдання чи нормалізації насильства; маркери маскування лексики через символічні заміни, емодзі, неорфографічні написання, навмисні помилки та код-міксинг; а також ширші дискурсивні патерни, що підвищують ризик ескалації онлайн-агресії в

офлайнові правопорушення [17]. Міжнародні організації акцентують на необхідності врахування цих ознак у комплексі з принципами прав людини та свободою вираження поглядів, пропонуючи збалансовані освітні, етичні й регуляторні рамки для цифрових платформ та національних політик у сфері протидії мові ворожнечі [18].

Для автоматизованого аналізу важливо враховувати не лише очевидну лексичну агресію, а й ті маркери, які виконують маскувальну або контекстуально-підбурливу функцію. Спрямованість висловлювання на групову приналежність часто проявляється через позірно нейтральні лексеми, які у певному контексті набувають дегуманізуючого або дискримінаційного змісту [19]. Міжнародні ініціативи ООН, зокрема в межах впровадження «Strategy and Plan of Action on Hate Speech» [20], підкреслюють, що класифікація не може обмежуватися набором заборонених слів, адже системність мови ворожнечі виявляється через комбінації смислів, алюзії, заміни символів, графічні варіації, емодзі, а також навмисні порушення орфографічних і граматичних норм. Додаткової уваги потребують дискурсивні стратегії, що супроводжують агресивну семантику: узагальнення, дегуманізація, заперечення правової суб'єктності, знеособлення, заклики до обмеження громадянських прав або виправдання насильства. ЮНЕСКО у своїх дослідженнях цифрової комунікації [21] звертає увагу на те, що ненависть дедалі частіше проявляється через візуально-графічні обхідні форми, які вимагають від інтелектуальних систем контекстної обробки, наприклад, емодзі-кодування, де зображення або символ створює смислове навантаження, еквівалентне агресивному висловлюванню.

1.2 Огляд існуючих підходів до виявлення мови ворожнечі

Проблематика автоматизованого виявлення мови ворожнечі належить до кола завдань обробки природної мови (Natural Language Processing), яка формується на перетині комп'ютерної лінгвістики, штучного інтелекту та когнітивного моделювання [22]. У ранніх дослідженнях переважали словникові підходи [23],

засновані на визначенні наборів заборонених слів або стійких словосполучень, що дозволяло здійснювати базову фільтрацію токсичних висловлювань через пошук точних входжень у текст. Такі підходи і нині застосовуються на окремих платформах (наприклад, у політиках модерації коментарів на [24] або у внутрішніх фільтрах Facebook [25]), проте вони не враховують контекст і морфологічні модифікації та не здатні фіксувати непряму, саркастичну або зашифровану агресію. Словникові стратегії втрачають результативність у випадках евфемізації, графічних спотворень, транслітерації та навмисного маскування, що спричинило перехід до статистичних та машинно-навчальних методів [26].

Поступове впровадження методів машинного навчання дозволило перейти від прямого пошуку ключових слів до класифікації текстів за заданими ознаками. Застосування моделей на основі Bag-of-Words [27], TF-IDF [28], n-грам [29], логістичної регресії [30] або опорних векторів [31] дало змогу аналізувати не лише лексичне наповнення, а й структурні та частотні характеристики мови ворожнечі. Одним із перших корпусів, що системно використовувався для таких завдань, був Hate Speech Dataset for Twitter, де вручну анотовані твіт-повідомлення стали базою для тренування класифікаторів. Цей корпус та подібні відкриті набори використовувалися в роботах Davidson, Warmsley, Macy та Weber [32] і Waseem та Hovy [33] для порівняння моделей і визначення релевантних ознак. Проте статистичні методи продемонстрували обмеження в умовах семантичної неоднозначності, багатомовності, контекстно залежних алюзій та нестандартної орфографії. Їх недостатня гнучкість при інтеграції латентних чи непрямих форм агресії спричинила перехід до глибинного навчання.

Поширення нейромережевих методів привело до зміни парадигми розпізнавання мови ворожнечі, оскільки з’явилася можливість працювати з послідовностями слів, контекстуальними залежностями та варіативними мовними конструкціями без створення жорстко фіксованих правил [34]. Архітектури на основі згорткових нейромереж і рекурентних моделей (CNN [35], LSTM [36], GRU [37]) почали застосовуватися для класифікації агресивного контенту на рівні коментарів, постів або коротких повідомлень, зокрема в роботах Pavlopoulos et al.

[38] та Zhang et al. [39]. Ранні експерименти зі згортковими та рекурентними моделями засвідчили вищу адаптивність до орфографічних спотворень і сленгу, однак залишалася проблема узагальнення на різні домени, мови та способи маскування. Вплив інформаційного середовища, де агресія часто проявляється у межах мемів, неологізмів, інтернет-жаргону, культурно маркованих іменувань, вимагав ще гнучкіших представлень, здатних інтерпретувати смислові патерни без прямої залежності від поверхневої форми [40].

Розвиток трансформерних моделей на основі механізму самоуваги (self-attention) сформував новий етап у створенні систем виявлення мови ворожнечі. Впровадження BERT [41], RoBERTa [42], XLM-R [43] та їхніх похідних дало змогу поєднати семантичну варіативність із контекстною інтерпретацією на рівні речення й дискурсу. Аналіз порівняльної ефективності таких архітектур здійснюється, зокрема, в дослідженні Mozafari, Farahbakhsh і Crespi [44], де показано, що моделі з попереднім тренуванням перевершують словникові та статистичні методи і класичні нейромережеві рішення, особливо у випадках маскування агресії через жарт, алюзію чи стилістичне викривлення. Водночас навіть трансформерні моделі стикаються зі складнощами, коли агресивний зміст реалізується поза класичною текстовою формою – через емодзі, цифробуквені заміни, латинізовані написання, графічне фрагментування, суржик чи транслітерацію [46]. У середовищах з активною мовною варіативністю, як-от український сегмент соцмереж у період гібридних конфліктів, ці фактори додатково знижують результативність автоматизованої детекції та вимагають адаптації моделей до нестабільних мовних патернів. Окремим напрямом досліджень стало виявлення прихованої та гібридної мови ворожнечі, яка поєднує елементи подвійного значення, сарказму та політично маркованого контексту. Дослідження Silva, Mondal, Correa, Benevenuto та Weber продемонстрували, що значна частина токсичних висловлювань в онлайн-середовищі не містить явних образливих слів, але має спрямованість на дискримінацію груп через приховану семантику. Подібні спостереження фіксуються у звітах Ради Європи та ECRI, де наголошується, що традиційні системи на основі ключових слів втрачають ефективність через постійне оновлення мовних стратегій, використання сленгу,

локальних кодів та візуально-графічних символів. Актуальність підтримки багатомовних нейромережевих моделей засвідчують також ініціативи ЄС, зокрема у межах досліджень щодо соціальних медіа та радикалізації [11].

Український та східноєвропейський сегменти онлайн-комунікації створюють окремі виклики для автоматизованого виявлення агресії. По-перше, війна, політична поляризація і пропагандистські кампанії посилюють агресивне мовлення з обох боків конфлікту [46]. По-друге, характерною є наявність суржику, морфологічно нестандартних форм, контамінованих конструкцій української та російської мов, латиниці замість кирилиці, навмисного спотворення орфографії та використання меметичних скорочень. По-третє, ворожість дедалі частіше реалізується через калькування з російськомовної або англомовної пропаганди, перенесення псевдокодів та емоційно поляризованих ярликів [47]. Ці чинники ускладнюють створення універсальних словників і потребують формування локалізованих корпусів для тренування моделей. Спроби залучення трансферного навчання (transfer learning) демонструють перспективність мультимовних моделей на кшталт XLM-R у поєднанні з доменною адаптацією, проте повноцінно локалізованих наборів для української мови досі обмаль.

У межах сучасних досліджень нейромережеві системи все частіше поєднуються з попереднім моделюванням шумів, коли тексти штучно трансформуються для тренування стійких класифікаторів. Маскування агресії через цифробуквені комбінації, емодзі, стилістично-символьні деформації або фрагментування слів вимагає від моделей здатності інтерпретувати не лише лінгвістичну форму, а й приховані смислові патерни. У роботах, присвячених adversarial training [48], пропонується адаптація класифікаторів до умов, у яких користувач намагається уникнути автоматичного виявлення. З огляду на реалії цифрової комунікації, де контент часто циркулює між кількома мовами і стилістичними регістрами, нейромережеві методи виявляються більш чутливими до контексту, проте все ж залежать від репрезентативності навчальних даних.

Загальною тенденцією розвитку систем розпізнавання мови ворожнечі є перехід від статичних словників до багаторівневого контекстного аналізу, де

поєднуються семантичні, морфологічні, прагматичні та дискурсивні індикатори [49]. У провідних академічних і практичних розробках нейромережеві архітектури доповнюються спеціалізованими препроцесинговими модулями, в тому числі нормалізацією шумових форм, транслітерацією, виправленням орфографії, а також автоматичним визначенням цільової групи чи об’єкта агресії. Платформи на кшталт Perspective API [50] демонструють спроби інтеграції глибинних моделей у системи модерації, хоча їх ефективність залежить від мовної специфіки та здатності адаптуватися до нових патернів. Сучасні дослідження підтверджують, що повна автоматизація можлива лише за умови інтеграції контекстно чутливих моделей із механізмами адаптивного навчання, які враховують динаміку мови, соціокультурні впливи та навмисне спотворення змісту.

Синтез наведених підходів свідчить, що словникові та статистичні методи поступово втрачають релевантність через зростання обсягу зашумленого контенту, однак виконують допоміжну роль у верифікації цільових груп або задаванні первинних ознак. Нейромережеві моделі, зокрема трансформерні архітектури, забезпечують вищу гнучкість і точність за умов варіативності мовних форм, але їхня результативність напряму залежить від наявності корпусів, що відображають реальне мовне середовище [51]. У ситуації, коли мова ворожнечі поєднує прямі, завуальовані та гібридні прояви, а тексти містять шумові спотворення, змішаність кодів, символічні заміни та порушення орфографічних норм, постає потреба в адаптації підходів глибинного навчання до нестабільних умов. Саме ця залежність між якістю моделей і характером вхідних даних визначає актуальність створення методів, здатних інтерпретувати агресивні висловлювання попри мовні викривлення, що особливо важливо для україномовного та мультикодових середовищ, де алгоритми мають враховувати не лише мовну форму, а й соціально-політичний контекст.

1.3 Етичні аспекти автоматизованого виявлення мови ворожнечі

Автоматизоване виявлення мови ворожнечі безпосередньо пов’язане з комплексом етичних обмежень і нормативних застережень, що стосуються забезпечення свободи вираження поглядів, ризиків надмірного цензурування публічного простору та хибної або спрощеної інтерпретації змісту висловлювань алгоритмічними системами. Використання методів машинного навчання для аналізу мовлення неминуче передбачає формалізацію соціально чутливих понять, що створює загрозу редукції контексту та ігнорування прагматичних і культурних чинників. Міжнародні організації, зокрема Рада Європи та Організація Об’єднаних Націй, у своїх стратегіях протидії hate speech послідовно наголошують, що будь-які заходи з обмеження деструктивної комунікації мають здійснюватися з дотриманням прав людини, передусім права на свободу висловлювання, закріпленого у Конвенції про захист прав людини і основоположних свобод [52]. У цьому контексті регулювання автоматизованого виявлення агресивних висловлювань повинно спиратися на принципи пропорційності, недискримінаційності та мінімального втручання в комунікативні процеси.

Етична напруга суттєво посилюється в ситуаціях, коли системи автоматичного контролю застосовуються для попередньої, масової або прихованої модерації контенту без належного інформування користувачів. Помилкове маркування нейтральних або критичних висловлювань як мови ворожнечі може призводити до блокування або видалення суспільно значущих матеріалів, зокрема політичної критики, журналістських розслідувань, сатири чи фіксації та документування воєнних злочинів і порушень прав людини [53]. Це є особливо проблематичним в умовах збройних конфліктів і кризових ситуацій, де агресивна або емоційно насичена лексика може виконувати не образливу, а оціночно-захисну, мобілізаційну чи свідчильну функцію. З огляду на це автоматизовані технологічні рішення доцільно розглядати виключно як допоміжні інструменти підтримки прийняття рішень, а не як остаточні або самодостатні механізми оцінювання допустимості мовлення.

У документах UNESCO та Організації з безпеки і співробітництва в Європі (OSCE) підкреслюється, що автоматизовані системи виявлення мови ворожнечі мають впроваджуватися з обов’язковим дотриманням принципів прозорості, пояснюваності алгоритмів і наявності процедур оскарження ухвалених рішень [54]. Користувачі та незалежні інституції повинні мати можливість зрозуміти логіку роботи таких систем і перевірити обґрунтованість їхніх висновків. Хибні спрацьовування, контекстні помилки або неповне розуміння дискурсу не можуть слугувати підставою для автоматичного застосування санкцій без участі людини. Відповідальність за остаточну інтерпретацію результатів, вибір модераційних заходів і пов’язані з ними правові наслідки покладається на кінцевого користувача або організацію, що впроваджує й експлуатує відповідну систему, а не на алгоритм як такий.

Окремим етичним виміром є робота з персональними даними та ризики алгоритмічної упередженості. Оскільки повідомлення соціальних мереж можуть містити ідентифікаційні відомості, система має застосовувати принцип мінімізації даних, обмеження доступу та безпечне зберігання корпусу, а за можливості – деперсоналізацію прикладів до рівня, достатнього для навчання й оцінювання. Водночас моделі виявлення мови ворожнечі можуть відтворювати або підсилювати перекоси розмітки та корпусу, наприклад частіше помилково маркувати висловлювання окремих соціальних груп, діалектів або змішаних мовних форм як агресивні. Тому етично обґрунтоване застосування передбачає регулярний аудит якості на різних підмножинах даних, документування правил розмітки, а також використання результатів моделі як сигналу для подальшої перевірки, а не як автоматичної підстави для санкцій.

З огляду на це, автоматизоване виявлення мови ворожнечі має розглядатися як інструмент підтримки, а не заміни аналітичного чи правового оцінювання. Кінцевий контроль, верифікація класифікаційних рішень і прийняття відповідальності за можливі помилки покладаються не на технологію, а на суб’єкта,

який її використовує. Це забезпечує баланс між захистом від шкідливих форм комунікації та збереженням свободи вираження в межах правових і етичних норм.

1.4 Аналіз наукових публікацій у контурі автоматизованого виявлення мови ворожнечі

Розвиток соціальних платформ суттєво посилив поширення мови ворожнечі, що створює ризики для суспільної стабільності та психологічної безпеки користувачів. Традиційні методи виявлення, зокрема словникова фільтрація, правила або класичне машинне навчання, виявляються малоефективними для контекстно залежних, завуальованих та непрямих проявів агресії. У відповідь на це дослідники аналізують потенціал великих мовних моделей (GPT-3, BERT та їх наступників) у задачах автоматизованого виявлення мови ворожнечі. Огляд фокусується на еволюції LLM у сфері NLP, виявленні їхніх переваг і обмежень, а також впливі таких моделей на точність, справедливість і стійкість класифікаційних систем. Узагальнення сучасних досліджень окреслює стан технологій і визначає подальші напрями розвитку, спрямовані на підвищення ефективності та етичної надійності систем виявлення мови ворожнечі на основі LLM [2].

Дослідження [55] зосереджується на виявленні мови ворожнечі у малоресурсних мовах, зокрема арабській та урду, де проблема ускладнюється мовною варіативністю, неявною агресією та відсутністю корпусів. Автори створили власний багато-мовний анотований датасет UAHSD-2025 на основі платформи X, який охоплює як бінарну, так і багатокласову класифікацію за п’ятьма категоріями ворожнечі. Анотація виконана вручну за розробленими інструкціями, що забезпечило якість корпусу. Для подолання міжмовних відмінностей застосовано дві стратегії: переклад усіх текстів в одну цільову мову перед класифікацією та спільне мультимовне навчання без перекладу, де одна модель працює з різними мовами паралельно. Автори провели 54 експерименти, порівнюючи TF-IDF з класичними алгоритмами, глибинні моделі на FastText і GloVe та контекстуальні мовні представлення. Найвищі результати отримано моделлю XLM-R, яка

продемонструвала точність 0.99 у бінарній класифікації для арабської, урду та мультимовного набору, а також 0.95, 0.94 і 0.94 відповідно для багатокласової класифікації. Це підтверджує перевагу мультимовних трансформерів над традиційними підходами при роботі з низькоресурсними мовами.

Дослідження [56] зосереджене на виявленні мови ворожнечі в мовах із деванагарським письмом, зокрема в гінді та непальській, де брак корпусів і моделей ускладнює автоматичний аналіз. Автори застосували гібридну архітектуру Attention BiLSTM-XLM-RoBERTa, яка поєднує послідовну обробку тексту з контекстуальними мультимовними репрезентаціями. Такий підхід забезпечив ефективну інтерпретацію мовних варіантів і захоплення латентних структурних ознак. У межах задачі виявлення мови ворожнечі модель досягла Macro F1 = 0.7481, що засвідчує її здатність працювати в умовах мовної варіативності й нестандартної орфографії.

У дослідженні [57] розглянуто чотири мовні корпуси англійську, каннада, малаялам та тамільську і побудовано ансамблеву модель для класифікації hope speech. До складу ансамблю було включено LSTM, mBERT і XLM-RoBERTa, що дозволило поєднати переваги послідовних нейромереж і багатомовних трансформерів. Результати показали перевагу ансамблевого підходу над окремими моделями за F1-метрикою: 0.93 для англійської, 0.74 для каннада, 0.82 для малаялам і 0.60 для тамільської. Робота демонструє, що позитивна комунікація в онлайн-середовищі також вимагає автоматизованого аналізу, а для малоресурсних мов доцільним є використання мультимовних архітектур та комбінованих моделей.

1.5 Постановка задачі

Метою кваліфікаційної роботи є підвищення точності виявлення мови ворожнечі у зашумлених соціальних текстових даних шляхом розроблення нейромережевого методу, здатного інтерпретувати спотворені, змішаномовні та навмисно масковані мовні конструкції. На відміну від існуючих підходів, що орієнтовані переважно на попередньо очищені або стандартизовані текстові вибірки,

запропонований метод ґрунтується на відтворенні контрольованих шумових спотворень і цілеспрямованому тренуванні нейромережевої моделі в умовах мовної нестабільності, маскування агресивної лексики та змішаності мовних кодів. Такий підхід дозволяє не лише зберігати релевантність класифікації за наявності орфографічних, графічних і семантичних викривлень, а й підвищувати стійкість моделі до свідомого приховування мовленнєвої агресії, що не забезпечується більшістю відомих рішень.

Для досягнення мети слід вирішити такі задачі:

– провести аналіз природи мови ворожнечі та її класифікаційних ознак;

– виконати огляд існуючих підходів до виявлення мови ворожнечі, виконати аналіз наукових досліджень;

– охарактеризувати етичні аспекти автоматизованого виявлення мови ворожнечі;

– розробити метод виявлення мови ворожнечі у зашумлених соціальних текстових даних нейромережевими засобами;

– виконати підготовку датасету для файнтюнінгу нейромережі для виявлення мови ворожнечі;

– виконати програмну реалізацію розробленого методу;

– провести дослідження методу виявлення мови ворожнечі у зашумлених соціальних текстових даних нейромережевими засобами.

РОЗДІЛ 2 Метод виявлення мови ворожнечі у зашумлених соціальних текстових даних нейромережевими засобами

2.1 Особливості прояву агресивної лексики в соціальних мережах і комунікаційних платформах

Агресивна мовна поведінка у цифрових середовищах формується як складний багаторівневий феномен, що поєднує лінгвістичні, прагматичні та техно-комунікаційні чинники [58]. Відмінність соціальних платформ від традиційних текстових корпусів зумовлює високий рівень зашумленості даних, варіативність комунікативних стратегій і нестабільність слововживання, що безпосередньо впливає на побудову та навчання моделей виявлення мови ворожнечі.

Цифрове середовище продукує агресивні висловлювання не лише у формі відкритої образи, а й через непрямі мовні конструкції, контекстуальне кодування, емоційно марковані алюзії, візуально-текстові гібриди та ситуативні мовні ігри. Наявність меметичних структур, емодзі, заміни символів, суржику, транслітерації, орфографічного варіювання та свідомих деформацій висловлювання ускладнює автоматизований аналіз і потребує врахування позасловникових патернів. Ці форми не мають сталої семантики й часто залежать від мікрокультур онлайн-спільнот, що робить класичні словникові або регламентні підходи низькоефективними [59].

На відміну від формалізованих текстових корпусів, соціальні мережі характеризуються дискурсивною фрагментованістю й контекстною залежністю висловлювань. Експресивність і агресивність можуть розпізнаватися лише з урахуванням адресата, комунікативної історії або реактивних відповідей у гілці повідомлень. Мовні конструкції часто виконують латентну дискримінаційну чи дегуманізуючу функцію, не містячи прямих інвективних маркерів. Це обумовлює необхідність інтеграції прагматичного аналізу, семантичних векторних моделей і контекстно-чутливих механізмів уваги при побудові нейромережевих методів.

Алгоритми модерації та політики платформ також стимулюють появу прихованої агресії, зміщуючи її у сферу кодових виразів, алюзій, взаємних посилань і візуальних сурогатів [60]. Мовна ворожість нерідко реалізується через групові

практики, коли образливе навантаження розподіляється між авторами, репліками, коментарями чи символічними позначеннями. Ці явища актуалізують застосування багаторівневого аналізу, що охоплює лексичний, морфологічний, синтаксичний та дискурсивний рівні, а також моделювання міжкористувацьких взаємодій.

Оскільки агресивна лексика в онлайні функціонує як динамічна, змінна й контекстно адаптивна структура [61], систему її виявлення неможливо обмежити інвентарем ключових слів або статичною ознаковою моделлю. Ефективність нейромережевого підходу залежить від здатності моделі обробляти неформальні варіанти мови, виявляти приховану семантику, враховувати багатомовність та адаптуватися до нових форм агресії. Особливості даних соціальних платформ формують передумови для побудови спеціалізованих архітектур, комбінованих ембедингів, гібридних класифікаційних схем і методів донавчання на актуальних корпусах.

Таким чином, специфіка прояву агресивної лексики в онлайн-комунікації визначає як вимоги до формування навчальних вибірок, так і архітектуру нейромережевих моделей, які мають працювати з семантично неповними, стилістично викривленими й контекстно опосередкованими даними. Це зумовлює перехід від лексикографічних рішень до контекстно-орієнтованих трансформерних підходів, здатних моделювати приховані агресивні патерни в умовах високої інформаційної шумності.

2.2 Форми маскування, суржику, орфографічних спотворень, транслітерації та контекстного приховування агресії

Мова ворожнечі у цифровому дискурсі дедалі частіше проявляється не у прямій формі, а через приховані, контекстуально зумовлені або навмисно викривлені мовні конструкції, що ускладнює як правове реагування, так і автоматизоване виявлення. Зміщення від відкритих форм агресії до непрямих стратегій комунікації фіксується міжнародними інституціями, зокрема Європейською комісією проти расизму та нетерпимості (ECRI), яка у своїх

щорічних звітах наголошує на зростанні прихованої та квазішифрованої мови ненависті в онлайновому середовищі [10]. На відміну від класичних вербальних образ, що ґрунтуються на відкритому приниженні або дегуманізації, сучасні прояви вербальної агресії маскуються через іронію, квазіжаргон, змішування мовних кодів і використання контекстуальних алюзій. Такі висловлювання можуть виглядати нейтральними поза конкретною соціальною або політичною ситуацією, але в межах визначеної групи мовців набувають точного агресивного змісту.

Суржик як форма змішаної комунікації створює додатковий рівень семантичної невизначеності, оскільки агресивні висловлювання часто поєднують елементи української, російської, англійської, жаргонізованої лексики та транслітерованих запозичень. У дослідженнях Агенції ЄС з фундаментальних прав (FRA) зазначається, що змішаномовне середовище сприяє появі нестандартних словоформ, що виходять за межі словникових баз, використовуваних традиційними алгоритмами модерації [11]. У соціальних мережах суржик та код-міксинг часто застосовуються як інструмент обходу фільтрів і як спосіб приховати агресивну спрямованість висловлювання без втрати його смислової впізнаваності для цільової аудиторії.

Орфографічні спотворення є однією з найпоширеніших стратегій уникнення автоматизованої детекції. Йдеться про навмисне руйнування графічної цілісності слова: випадання або додавання символів, заміну букв цифрами, поєднання різних алфавітів, використання регістру, вставок дефісів або знаків пунктуації в середині слова. Такі конструкції як «х*хли», «ж1д0банди», «укр0націки» чи «russня» дозволяють уникати простого лексичного збігу, зберігаючи образливу семантику. У звітах Ради Європи щодо онлайн-нетерпимості підкреслюється, що диджиталізація агресії супроводжується постійним оновленням так званих тактичних стратегій уникнення санкцій, де орфографічна модифікація перетворюється на форму «мовного шифрування».

Транслітерація, як заміна кириличної графіки на латиницю або її гібридні варіанти, є однією з ключових форм маскування мови ворожнечі. Комбінації на кшталт «mosk@li», «zhydo*banda» чи «ukr_natsi» поєднують латинські форми,

емодзі, знаки підкреслення та символи, створюючи текст, зрозумілий для людини, але малопридатний для словникових фільтрів. Дослідження ОБСЄ з протидії онлайн-ненависті підкреслюють, що проблема транслітерації є критичною у багатомовних регіонах, де користувачі можуть зміщувати алфавіти залежно від комунікативної мети або прагнення приховати зміст повідомлення [62].

З позиції автоматизованого аналізу перелічені прийоми фактично виконують роль цілеспрямованих пертурбацій тексту, тобто спотворень, які порушують лексичні збіги та збільшують частку нестандартних токенів. Унаслідок цього різко знижується ефективність словникових і шаблонних фільтрів, а також моделей, що спираються на поверхневі ознаки написання. Навіть за використання субсловних токенізаторів трансформерних архітектур частина спотворень призводить до фрагментації ключових інвектив у рідкісні підпослідовності, що ускладнює коректне відтворення семантики та зменшує впевненість класифікатора. Водночас агресивний намір у таких випадках часто зберігається для людини, тому задача детекції потребує орієнтації не лише на «чистий» текст, а й на стійкість до варіативного написання, змішаних абеток і навмисного маскування, що підсилює актуальність підходів із керованим моделюванням шумів і тренуванням у наближених до реального середовища умовах.

Контекстне приховування агресії виявляється у формах, де ворожий зміст реалізується не через прямі інвективи, а через натяк, алюзію або вторинне значення. Сюди належать випадки, коли вираз набуває агресивного змісту лише у зв’язку з національними, релігійними чи політичними конотаціями, відомими певній групі. У Стратегії ООН із протидії мові ненависті наголошується, що цифрова культура сприяє появі непрямих форм дегуманізації, де агресивний намір зчитується лише у контексті мемів, політичних гасел, символів і прихованих кодів. Додатково роль контексту простежується в комбінуванні агресивних висловлювань із емодзі, GIF-зображеннями, тональними маркерами, що можуть перетворювати нейтральний текст на вияв приниження або погрози.

Особливо виразною проблемою є комбінація кількох стратегій маскування в одному висловлюванні – змішаномовність, спотворення орфографії, транслітерація,

меметичність, графічні символи, жаргон і сленг. Такі форми не піддаються лінійному аналізу й вимагають інтерпретації на рівні наміру, соціокультурного фону та семантичного мапування. ЮНЕСКО у своїх напрацюваннях щодо цифрової грамотності зазначає, що сучасна агресивна комунікація дедалі частіше набуває фрагментного характеру, коли слово, символ або зображення без контексту не становить проблеми, але у поєднанні створює мову ненависті [63]. Саме тому автоматизовані підходи до виявлення таких проявів стикаються з обмеженнями, що вимагають моделювання контексту, нормалізації нестандартних мовних форм і здатності систем розпізнавати латентні смисли.

Сукупність цих факторів свідчить, що традиційні лінгвістичні методи виявлення мови ворожнечі не можуть бути ефективними без урахування мовних спотворень, навмисної зміни графіки та контекстуальної адаптації. Це підтверджується у звітах Ради Європи, ООН, ЮНЕСКО та FRA, які підкреслюють, що алгоритмічна модерація контенту повинна враховувати не лише текстову форму, а й кодифіковані та приховані прояви комунікативної агресії. Така трансформація у способах вираження ворожості визначає потребу у створенні адаптивних моделей, здатних працювати із зашумленими, нестабільними та навмисно модифікованими мовними структурами.

2.3 Підхід до автоматизованого моделювання шумових викривлень у текстових даних.

Оскільки більшість відкритих корпусів мови ворожнечі, зокрема набір даних з платформи Kaggle, містять переважно стандартизовані тексти, виникає необхідність штучного відтворення типових викривлень, характерних для реальних соціальних платформ. Моделювання шуму виконується не як побічна процедура очищення даних, а як цілеспрямований етап розширення навчальної вибірки з метою підвищення стійкості нейромережної моделі до варіативних, нестандартних та умисно деформованих мовних форм.

Текстовий корпус позначено, як

$$\mathrm{D} = \{(\mathrm{S_i}, \mathrm{y_i})\}_{i=1}^{n}, \tag{2.1}$$

де S_i – текстове повідомлення, а $y_i \in \{0,1\}$ – клас (мова ворожнечі / відсутність проявів мови ворожнечі). Розширений корпус D' формується застосуванням стохастичної функції перетворення до кожного S_i. У загальному вигляді речення представляється як:

$$S = \{w_1, \dots, w_n\} \tag{2.2}$$

і для нього визначається функція трансформації:

$$T: S \rightarrow S' \tag{2.3}$$

що утворює нове речення:

$$S' = \{w'_1, \dots, w'_n\} \tag{2.4}$$

Для моделювання шумів використовуються кілька типів перетворень T_k із імовірностями p_k. Сумарна імовірність вибору певного перетворення на кроці нормується до одиниці:

$$\sum_{k=1}^{K} p_k = 1 \tag{2.5}$$

В межах роботи будуть реалізовані лише ті перетворення, які можна відтворити програмно без ручної переанотації та без руйнування семантики: орфографічні спотворення (перестановка, видалення, дублювання символів), символьне маскування, часткова/повна транслітерація, обмежені лексичні підміни (суржик/жаргон) за словником. Нехай w – слово (послідовність символів). Кожне слово w_i у текстовій послідовності розглядається як кандидат на трансформацію, і з імовірністю p_k до нього застосовується одне з визначених перетворень T_k, тоді як з імовірністю $1-p_k$ воно зберігає початкову форму. Стохастичне правило модифікації токенів задається виразом:

$$w'_i = \begin{cases} T_k(w_i), \text{з імовірністю } p_k \\ w_i, \text{з імовірністю } 1 - p_k. \end{cases} \tag{2.6}$$

На рівні речення така аплікація виконується покроково, причому загальна частка модифікованих токенів обмежується в межах 20-40% для збереження цілісності синтаксичної структури та коректності мітки y_i.

Розширений корпус формується шляхом об’єднання вихідних і трансформованих вибірок, після чого здійснюється донавчання моделі на змішаних

даних із домішкою α ($0 \leq \alpha \leq 1$), що визначає співвідношення «чистих» і зашумлених прикладів у кожному батчі:

$$D* = D \cup D' \tag{2.7}$$

$$D\alpha = \alpha \cdot D + (1 - \alpha) \cdot D' \tag{2.8}$$

Оптимізація нейромережевої моделі виконується за критерієм мінімізації крос-ентропійної втрати на розширеному корпусі. Якщо класифікатор позначити як F_θ, тоді задача навчання має вигляд:

$$\theta^* = arg \min_\theta \mathcal{L}(F_\theta(D_\alpha), Y) \tag{2.9}$$

Значення параметрів p_k, α і частка модифікованих токенів визначаються експериментально на валідаційній підвибірці з метою досягнення рівноваги між семантично стабільним відтворенням класів і необхідною варіативністю синтетичного шуму.

2.4 Етапи та кроки методу виявлення мови ворожнечі у зашумлених соціальних текстових даних

Наведена на рисунку 2.1 схема ілюструє етапи та кроки методу виявлення мови ворожнечі у зашумлених соціальних текстових даних.

Метод призначений для автоматизованого виявлення мови ворожнечі в соціальних текстових даних, що містять орфографічні, графічні, лексичні або транслітераційні викривлення, характерні для природного цифрового спілкування. Його побудова ґрунтується на поєднанні контрольованого моделювання шумових спотворень у навчальному корпусі з подальшим донавчанням нейромережевої архітектури, здатної інтерпретувати нестандартні мовні форми без втрати класифікаційної здатності.

Перший етап охоплює підготовку нейромережі, а другий охоплює застосування навченої моделі для детекції мови ворожнечі в нових повідомленнях. Вхідними даними першого етапу є анотований датасет із двома класами (мова ворожнечі / відсутність мови ворожнечі) та базова архітектура нейромережевого класифікатора, обрана для подальшого донавчання. Результатом стають розширений

корпус із контрольованими викривленнями та параметри навченої моделі, придатної до використання на етапі інференсу.

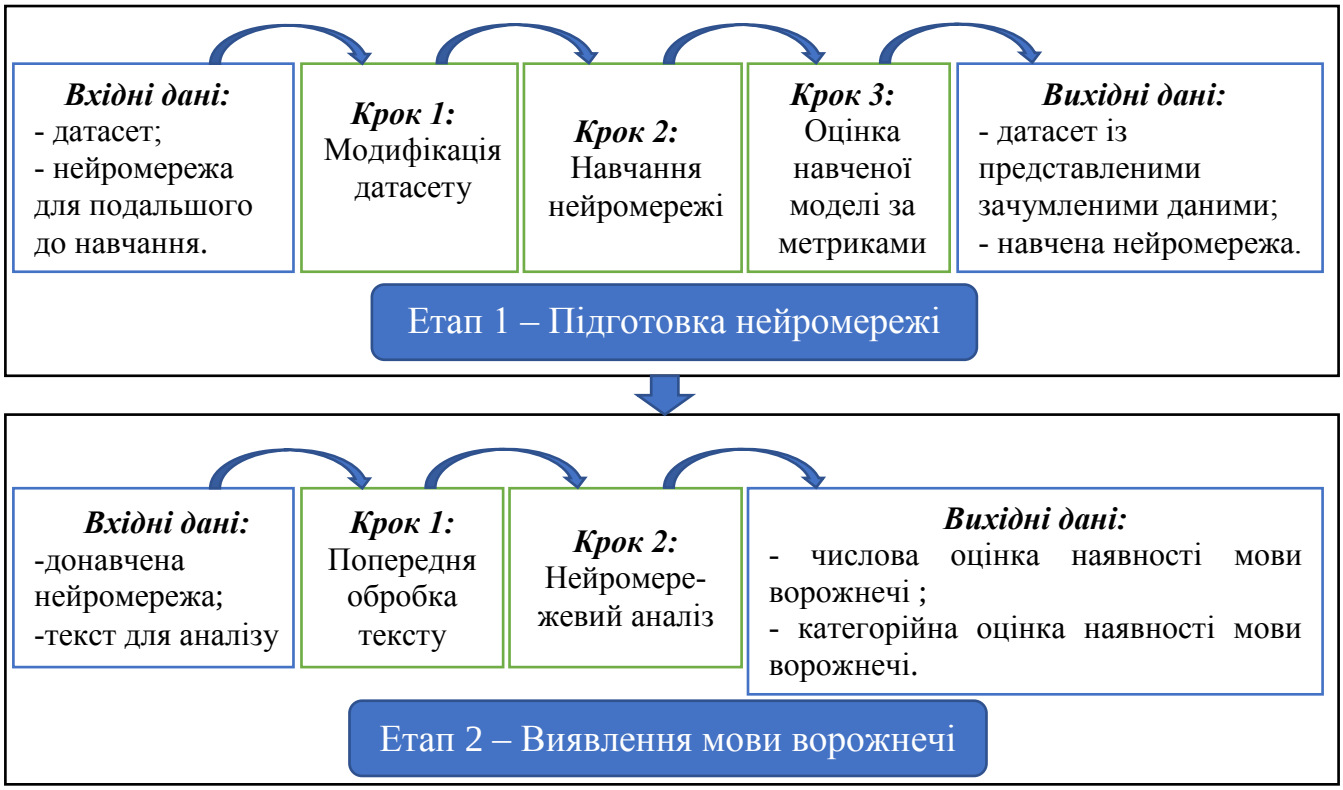


Рисунок 2.1 – Етапи та кроки методу виявлення мови ворожнечі у зашумлених соціальних текстових даних

На першому етапі спочатку виконується модифікація вихідного корпусу. Для кожного тексту формується стохастично трансформований варіант шляхом застосування заздалегідь визначених операторів шуму. До перетворень віднесено орфографічні деформації на рівні символів, символьне маскування з використанням альтернативних графем, часткову або повну транслітерацію, а також лексичні підміни в межах словника суржику та жаргону. Задається вектор імовірностей застосування перетворень, після чого генерується додаткова вибірка, яка разом з оригінальними прикладами утворює розширений датасет. Співвідношення «чистих» і зашумлених прикладів регулюється параметром домішування, що дозволяє керувати мірою варіативності даних без зміни первинної розмітки. Перед навчанням здійснюється уніфікований препроцесинг: токенізація під вибрану модель,

узгодження довжин послідовностей, формування батчів і, за потреби, часткова нормалізація, яка не знищує індикатори маскування, важливі для задачі.

Навчання нейромережі відбувається на змішаній вибірці з мінімізацією крос-ентропійної втрати. Оптимізація параметрів виконується методом стохастичного градієнтного спуску з фіксацією випадкового зерна та журналюванням конфігурації, що забезпечує відтворюваність. Для запобігання перенавчанню застосовуються валідаційні зупинки, регуляризація та підбір гіперпараметрів у межах вузької сітки. Після завершення навчання проводиться оцінка за метриками точності, повноти, F_1-міри та, за наявності дисбалансу, додатково обчислюється ROC-AUC або PR-AUC. За підсумками оцінювання відбирається найкращий контрольний знімок моделі; за потреби виконується калібрування ймовірностей (наприклад, температурним масштабуванням) і налаштування робочого порога прийняття рішення відповідно до цільового компромісу між хибними спрацьовуваннями та пропусками. Вихідними артефактами першого етапу є збережений розширений корпус, навчена й відкалібрована модель та зафіксований набір параметрів.

Другий етап призначений для виявлення мови ворожнечі в нових даних. На вхід подається текстове повідомлення та параметри донавченої моделі. Спочатку виконується попередня обробка, яка відтворює умови, прийняті на етапі навчання: та сама токенізація, ті самі обмеження довжини та правила поводження з нестандартними формами, щоб зберегти чутливість моделі до маскування, транслітерації та орфографічних відхилень. Після перетворення в токенізоване представлення повідомлення здійснюється прямий прохід крізь класифікатор, у результаті чого формується ймовірнісна оцінка належності до класу «мова ворожнечі». Калібрована ймовірність порівнюється з робочим порогом; рішення про клас супроводжується числовою оцінкою, яка відображає впевненість моделі. За потреби додатково повертається категорійна інтерпретація на основі порогової логіки або заздалегідь визначеної схеми ризиків, що дає змогу відокремлювати прикордонні випадки для ручної перевірки.

Змістовно перший етап забезпечує підвищення стійкості до зашумлення за рахунок навчання на контрольованих викривленнях, тоді як другий забезпечує

відтворюване застосування цієї стійкості в експлуатації. Взаємозв’язок між етапами підтримується узгодженістю препроцесингу, ідентичними токенізаторами та фіксацією параметрів шумогенерації, що виключає розрив між умовами навчання й використання. У такій конфігурації метод забезпечує стабільне виявлення мови ворожнечі навіть за наявності символічного маскування, суржику, транслітерації та орфографічних спотворень, водночас зберігаючи контроль за рівнем хибних класифікацій через чітко визначені пороги та процедури калібрування.

2.5 Формування і підготовка набору навчальних даних

Набір даних «Hate Speech Detection curated Dataset» [64] представляє собою анотовану колекцію текстових повідомлень англійською мовою, призначену для виявлення мови ворожнечі у соціальних мережах. Його особливістю є включення сучасних елементів онлайн-комунікації – емодзі, емоційних символів, скорочень та сленгових виразів, що забезпечує відповідність реальним мовним практикам у цифровому середовищі. Корпус поділений на два класи: повідомлення, що містять мову ворожнечі, та нейтральні або неворожі тексти. Попередня фільтрація і лінгвістичний аналіз гарантують якість даних та відсутність контенту, який міг би завдати шкоди користувачам.

Набір є цінним ресурсом для досліджень у галузі обробки природної мови та глибинного навчання, оскільки може використовуватись для побудови й навчання моделей автоматичного розпізнавання ворожих висловлювань у текстах соціальних мереж. Крім того, він може слугувати репрезентативною базою для оцінювання ефективності алгоритмів класифікації, спрямованих на виявлення агресивної чи дискримінаційної комунікації в онлайн-середовищі. Даний датасет (рисунок 2.2) буде використано для перевірки роботи методу, як валідаційний.

Набір даних «Hate Speech and Offensive Language Detection» [65] є систематизованою колекцією твітів англійською мовою, призначеною для дослідження та автоматизованого виявлення мови ворожнечі й образливих висловлювань у соціальних мережах.

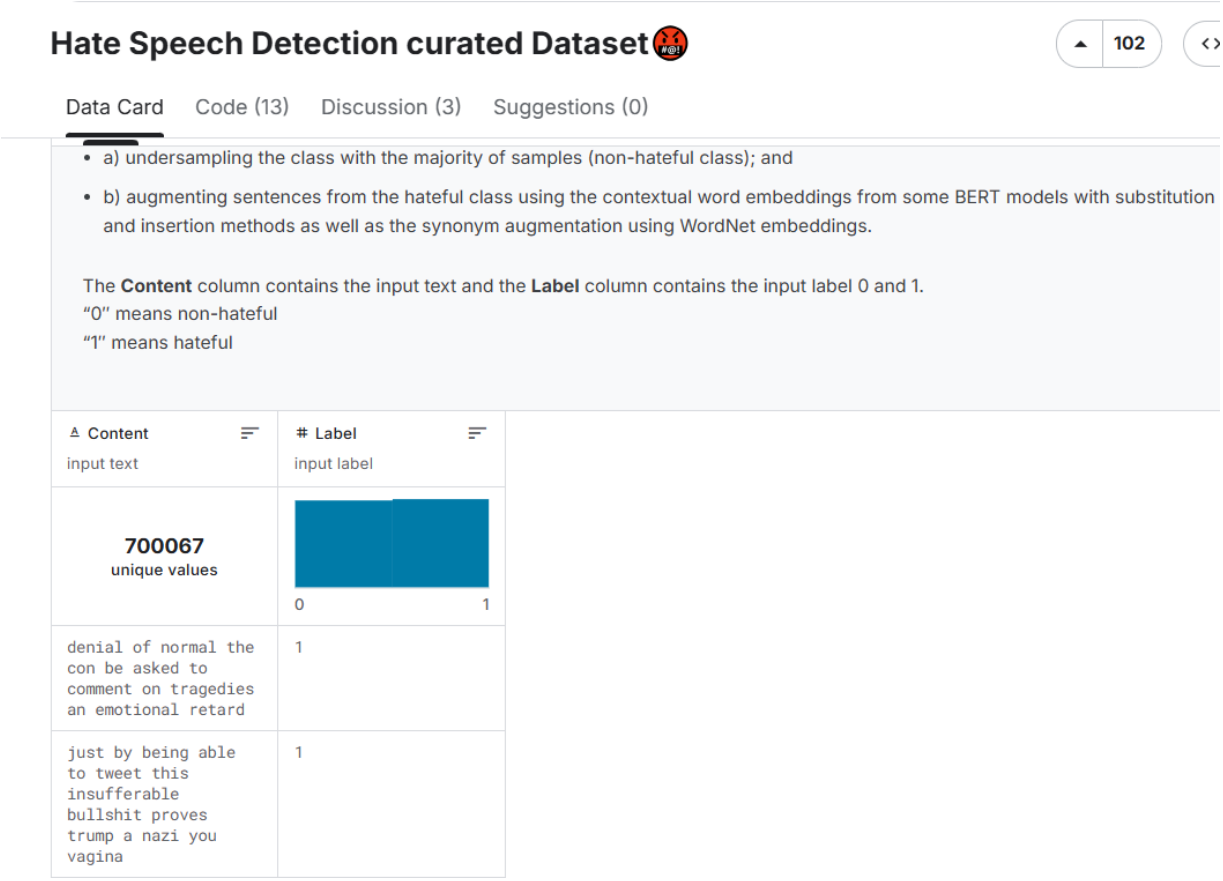


Рисунок 2.2 – Приклад даних з датасету «Hate Speech Detection curated Dataset»

Він створений шляхом збору твітів через публічний API Twitter із використанням цільових пошукових запитів, пов'язаних із проявами ворожості чи образливого контенту. Кожен запис (рисунок 2.3) містить текст повідомлення та результати його багаторазової ручної анотації, що дозволяє фіксувати різні рівні суб'єктивності в оцінках. Для кожного твіту наведено кількість загальних оцінок, а також кількість випадків, коли його класифікували як мову ворожнечі, як образливий або як нейтральний. Така структура забезпечує можливість статистичного аналізу достовірності класифікації й дозволяє будувати моделі з урахуванням ступеня узгодженості між анотаторами.

Набір даних є цінним ресурсом для досліджень у галузі обробки природної мови, машинного та глибинного навчання, оскільки надає репрезентативний корпус сучасних текстів із соціальних мереж, що містять як пряму, так і латентну агресію.

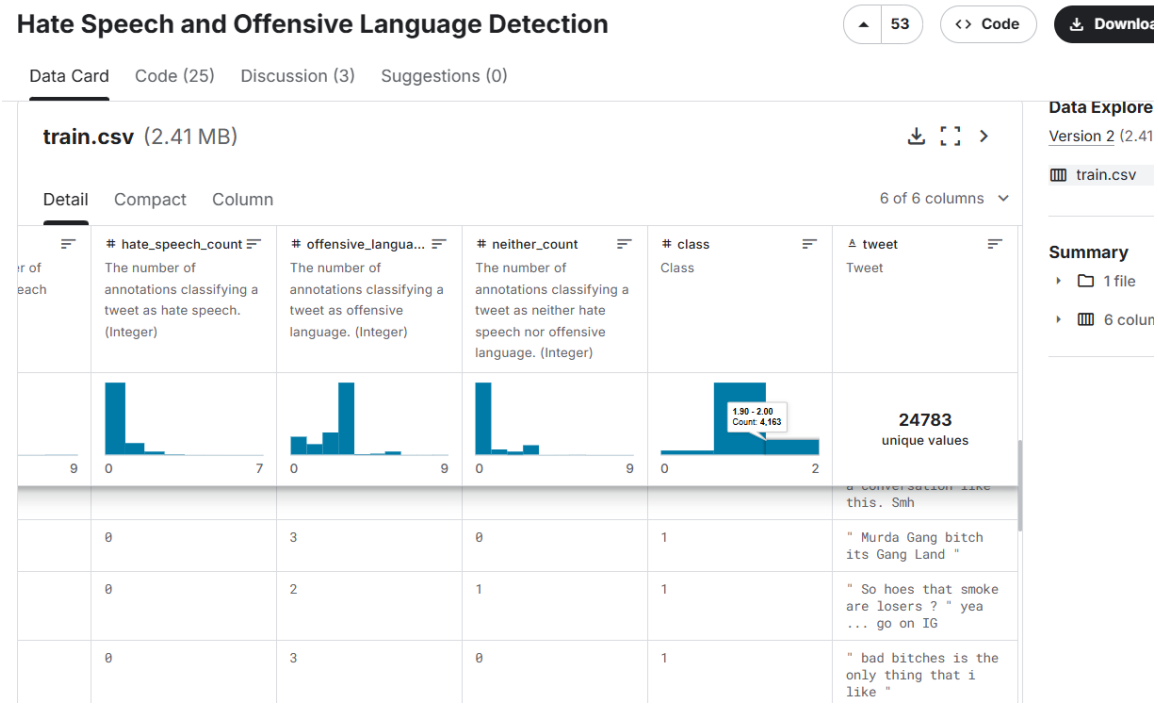


Рисунок 2.3 – Вигляд датасету «Hate Speech and Offensive Language Detection»

Його можна використовувати для розроблення моделей автоматичного модераційного контролю, аналізу емоційного тону повідомлень або створення систем виявлення токсичності контенту. Завдяки детальній структурі й відкритій ліцензії набір даних може служити як навчальною базою, так і еталонним матеріалом для порівняння ефективності різних алгоритмів класифікації у сфері боротьби з онлайн-дискримінацією та мовою ненависті.

Висновки до розділу 2

У розділі обґрунтовано метод виявлення мови ворожнечі, орієнтований на роботу з контекстно нестабільними та навмисно викривленими соціальними текстами. Показано, що агресивна комунікація у цифрових середовищах формується як багаторівневе явище, де поєднуються лінгвістичні, прагматичні й техно-комунікаційні чинники, а прояви ворожості часто маскуються через суржик, орфографічні спотворення, транслітерацію, символічні заміни. Така природа даних унеможливлює ефективне застосування словникових і статичних ознакових підходів та визначає потребу в контекстно-орієнтованих нейромережевих рішеннях.

Запропоновано підхід до автоматизованого моделювання шумових викривлень, у межах якого вихідний корпус доповнюється синтетично згенерованими прикладами за допомогою стохастичних перетворень на рівні токенів і слів. Формалізовано механізм застосування перетворень із керованими імовірностями та домішуванням за параметром α, що дозволяє підтримувати баланс між «чистими» і зашумленими прикладами без повторної анотації. Навчання моделі здійснюється на змішаній вибірці з мінімізацією крос-ентропійної втрати та подальшим калібруванням імовірнісних виходів, що забезпечує узгодження між обчислюваними оцінками та практичними порогами прийняття рішень.

Структуру методу подано як послідовність. На етапі підготовки формується розширений корпус, виконується уніфікований препроцесинг і донавчання класифікатора; на етапі застосування відтворюються умови тренування для нових повідомлень, здійснюється інференс і повертається як бінарне рішення, так і числова оцінка впевненості. Узгодженість токенізації, правил обробки та параметрів шумогенерації між етапами виключає розрив між навчанням та експлуатацією і забезпечує стабільність роботи моделі в реальному середовищі.

Окреслено експериментальну базу у вигляді двох відкритих наборів даних з різним ступенем варіативності й наявністю сучасних елементів онлайн-комунікації. Таке поєднання дає можливість оцінити стійкість методу до орфографічних, графічних та лексичних викривлень і водночас перевірити переносимість рішень на матеріалі реальних соціальних платформ. Передбачене оцінювання якості з урахуванням дисбалансу класів базується на узгодженні точності, повноти та F_1-міри, доповнених пороговими та калібрувальними процедурами.

Сукупний результат розділу полягає у формуванні цілісної методології: від теоретичного обґрунтування природи шумів і їх формального моделювання до визначення архітектури навчального та інференсного конвеєра і вибору даних для перевірки гіпотези. Така методологія створює підґрунтя для програмної реалізації, експериментальної верифікації та аналізу ефективності запропонованого підходу, що буде розгорнуто в наступних розділах.

РОЗДІЛ 3 Проєктування інтелектуальної системи виявлення мови ворожнечі

3.1 Вибір засобів розробки інтелектуальної системи

Розроблення інтелектуальної системи виявлення мови ворожнечі у зашумлених текстових даних вимагало використання такого інструментарію, який забезпечує поєднання обчислювальної ефективності, підтримки сучасних нейромережевих архітектур і гнучкості при роботі з неструктурованими мовними даними. Основою реалізації обрано мову програмування Python, що на сьогодні є стандартом у сфері обробки природної мови та глибинного навчання. Її застосування зумовлене можливістю використовувати попередньо навчені трансформерні моделі, виконувати моделювання шумових трансформацій без втрати семантики та поєднувати текстову препроцесуалізацію з подальшим донавчанням класифікатора.

Обчислювальне середовище побудовано на основі Google Colab [66], що дає доступ до апаратного прискорення та уніфікованої конфігурації з можливістю відтворення експериментів. Такий підхід дозволяє виконувати генерацію трансформованих корпусів, змішування даних, токенізацію, векторизацію та навчання моделі в межах єдиного середовища, не вдаючись до розгортання окремої інфраструктури. Інтеграція попередньо налаштованих мовних представлень забезпечує адаптацію до контекстно нестабільних даних, тоді як використання автоматизованих середовищ виконання усуває обмеження, пов’язані з локальними обчислювальними ресурсами.

Побудова всього циклу – від формування розширеного корпусу до оцінювання класифікаційної стійкості здійснюється в межах єдиного технологічного стеку, що унеможливлює розрив між етапами моделювання шуму, тренуванням і подальшим інференсом. Це забезпечує контрольоване перенесення результатів до застосунку та гарантовану сумісність між процесами підготовки даних і виконанням класифікації.

3.2 Проєктування складових інтелектуальної системи

Діаграма компонентів інтелектуальної системи виявлення мови ворожнечі наведена на рисунку 3.1.

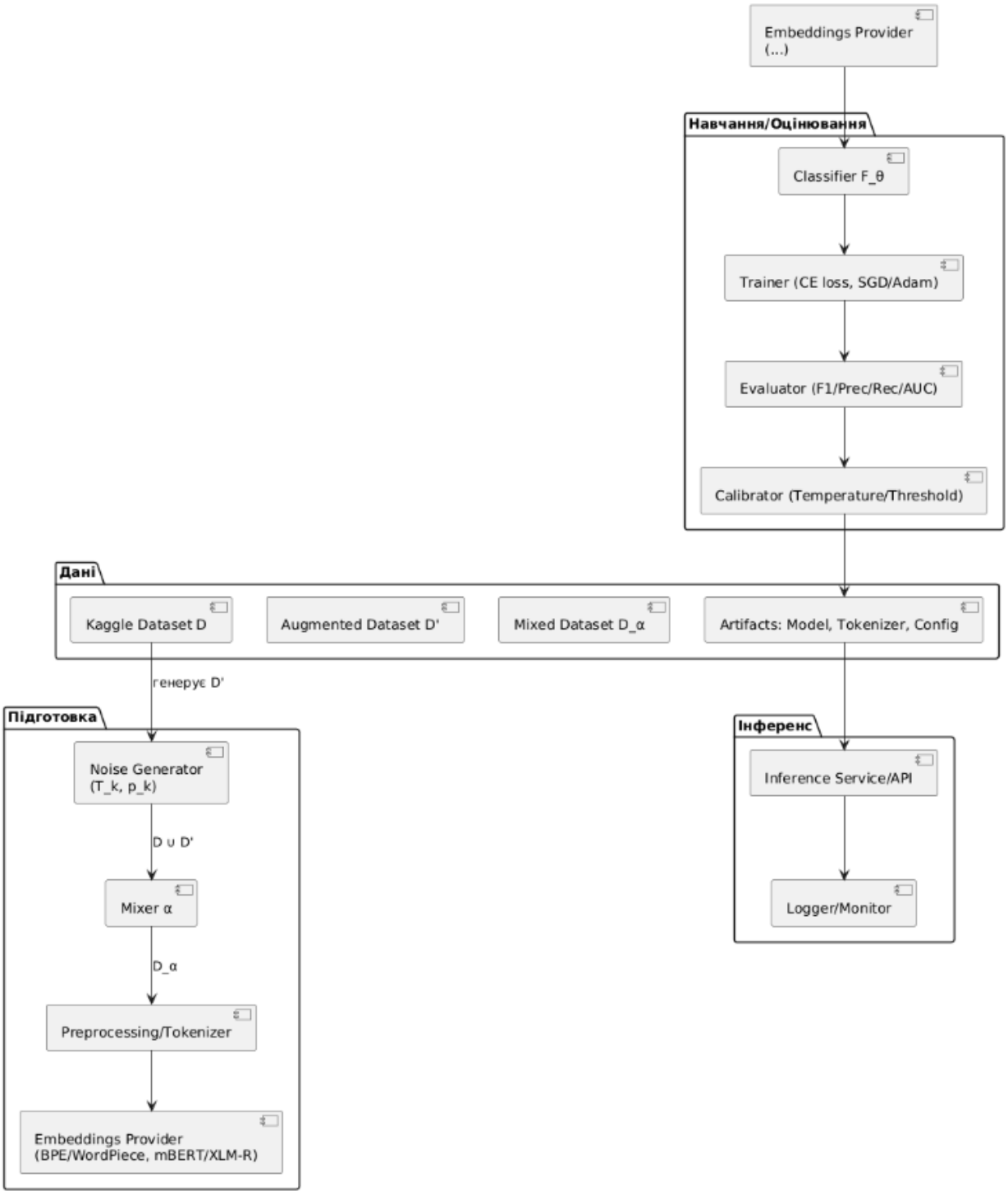


Рисунок 3.1 – Діаграма компонентів

Діаграма відтворює структурно-процесуальну організацію методу виявлення мови ворожнечі у зашумлених текстових даних як послідовність пов’язаних підсистем, що забезпечують безперервний перехід від формування даних до застосування моделі. Логіка побудови передбачає поділ на три взаємозалежні змістові блоки: підготовку, навчання та інференс, кожен з яких оперує власними об'єктами та артефактами, але узгоджується з попереднім і наступним рівнями.

Початковою компонентною основою слугує корпус даних, отриманий з відкритого джерела, який розглядається як базовий набір для генерації похідних вибірок. На основі цього корпусу формується зашумлена підмножина шляхом застосування стохастичних трансформацій, після чого модифіковані приклади поєднуються з вихідними в єдину структуру. Далі здійснюється первинна обробка тексту із збереженням елементів, значущих для інтерпретації нестандартних мовних форм, та підготовка представлень, придатних для подальшого кодування нейромережею. На цьому етапі вбудовано трансформаційний механізм, що забезпечує перехід від токенізованих послідовностей до контекстуальних векторних репрезентацій.

Отримані дані передаються до блоку навчання й оцінювання, де локалізується архітектура класифікатора та модуль оновлення його параметрів. Навчання здійснюється за критерієм мінімізації похибки на змішаній вибірці, що дозволяє адаптувати модель до шумових викривлень. Після досягнення прийнятної збіжності виконується оцінювання якості на незалежних даних із фіксацією ключових метрик. Калібрування вихідних ймовірнісних оцінок забезпечує узгодженість між результатом класифікації і прийняттям рішення в умовах нечіткої або маскованої агресивної лексики. Завершення навчального циклу супроводжується формуванням артефактів, серед яких параметризована модель, словник токенізації та конфігураційні налаштування, необхідні для подальшого використання.

Фаза інференсу реалізує застосування навченої системи до нових текстових прикладів. Вона використовує параметри, збережені після навчання, і відтворює структуру обробки, яка використовувалася під час тренування, що унеможливлює

розрив між підготовкою та застосуванням. Результати аналізу реєструються та можуть бути піддані моніторингу, що відкриває можливість виявлення деградації продуктивності або накопичення нових типів шуму. Уся система функціонує як замкнений цикл, у якому потоки даних, модульні компоненти та результати навчання зберігають послідовність переходів і забезпечують відтворюваність процедур.

На рисунку 3.2 наведено діаграму варіантів використання. UML-діаграма відображає розподіл ролей та відповідальностей у процесі побудови й застосування методу виявлення мови ворожнечі у зашумлених текстових даних. Кожен актор пов’язаний з окремими функціональними сценаріями, що відповідають його зоні компетенції та етапу життєвого циклу системи.

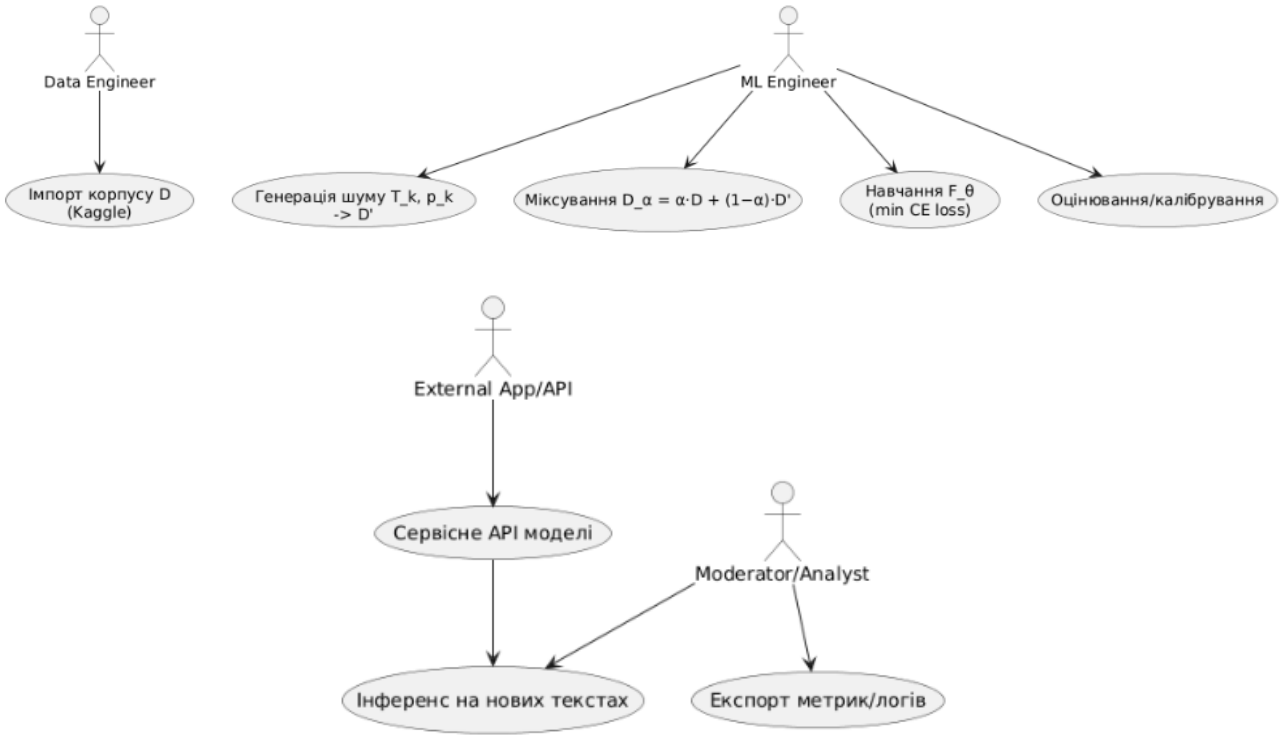


Рисунок 3.2 – Діаграма варіантів використання

На першому фрагменті зображено взаємодію двох учасників: фахівця з підготовки даних та інженера з машинного навчання. Підготовка датасету покладається на Data Engineer, який відповідає за пошук, імпорт і базову структуризацію анотованих вибірок. Наступний етап передбачає моделювання

шумових викривлень, під час якого наявний корпус доповнюється штучно згенерованими прикладами. Далі відбувається поєднання вихідних та трансформованих даних у пропорції, що дозволяє контролювати баланс між «чистими» та зміненими зразками. Після формування такої вибірки інженер з машинного навчання здійснює навчання моделі на основі крос-ентропійного критерію та забезпечує калібрування вихідних оцінок. Йому також належить аналіз отриманих результатів і прийняття рішень щодо повторного тренування чи корекції гіперпараметрів.

Нижній фрагмент показує застосування системи після завершення етапу навчання. Взаємодія з інтелектуальною системою може здійснюватися через зовнішній застосунок або програмний інтерфейс, який ініціює інференс, тобто класифікацію нових текстів із використанням навченої моделі. У свою чергу, роль модератора або аналітика полягає у відстеженні результатів, збереженні метрик продуктивності та аналізі журнальних записів, що дозволяє здійснювати оцінку ефективності моделі, виявляти аномалії або потребу в оновленні даних. Такий розподіл відповідальності забезпечує послідовність між етапами побудови системи та її практичним використанням, уникаючи змішування технічних і експлуатаційних функцій.

Послідовність взаємодій у процесі навчання моделі відображає поетапний рух даних і поступове оновлення параметрів класифікатора (рисунок 3.3). Вихідним об’єктом є корпус, отриманий із зовнішнього джерела, який містить анотовані текстові приклади. Цей датасет передається до модуля шумогенерації, де для кожного речення застосовується стохастична функція перетворення. У результаті формується додаткова вибірка, у якій наявні орфографічні, символічні, транслітераційні або лексичні спотворення, що імітують природний шум соціальних комунікацій.

Після цього вихідний і трансформований набори поєднуються в одне змішане середовище відповідно до заздалегідь визначеного співвідношення, що дозволяє регулювати частку модифікованих прикладів у навчальній множині. Отримані дані передаються до блока попередньої обробки, де виконується

токенізація та перетворення тексту у формат, придатний до векторизації. На наступному кроці формується тензорне представлення, яке подається безпосередньо на вхід нейромережевої моделі.

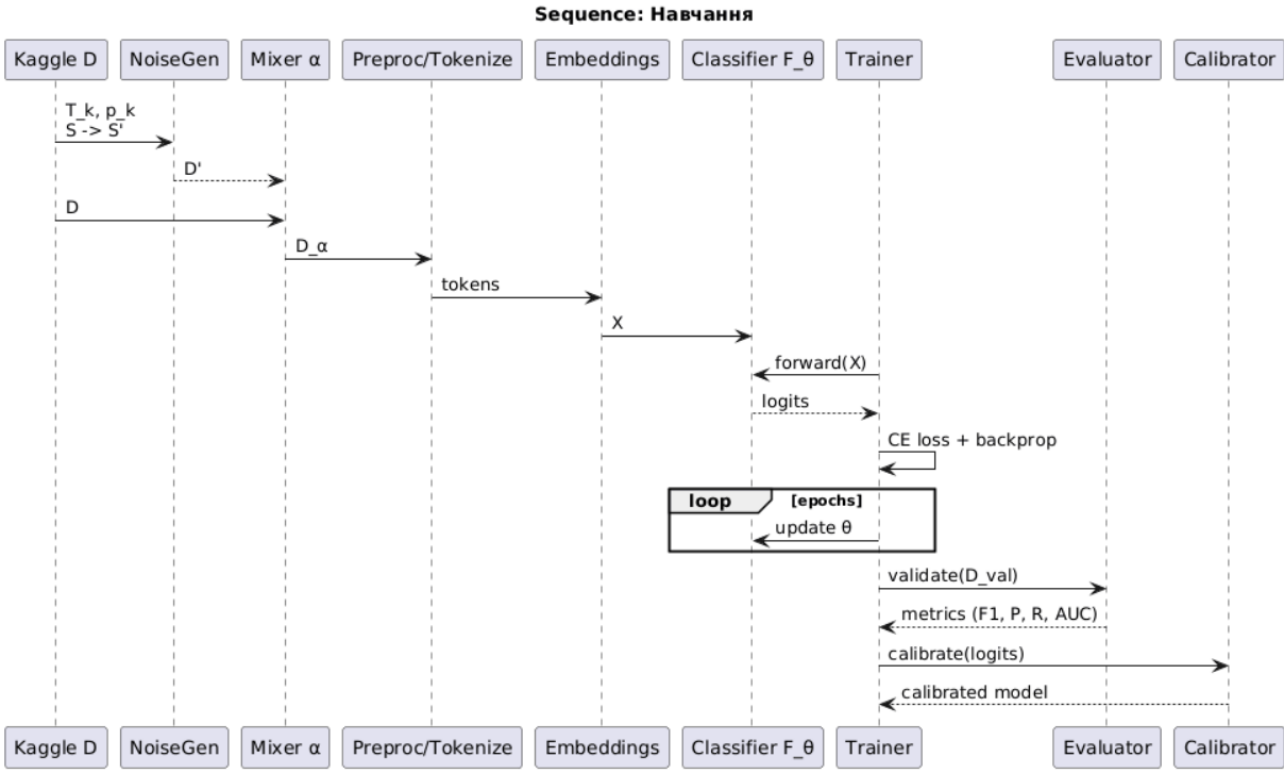


Рисунок 3.3 – Діаграма послідовності процесу навчання

Під час прямого проходу обчислюються логіти, які слугують основою для подальшого обчислення функції втрат. Градієнтне оновлення параметрів виконується ітеративно протягом кількох циклів навчання, що забезпечує поступову адаптацію моделі до змішаної вибірки. Після завершення кожної навчальної ітерації або епохи модель проходить валідаційне оцінювання на окремій підмножині даних, де обчислюються показники точності, повноти, F1-міри, а також інші узагальнені метрики.

Кінцевим етапом є калібрування вихідних імовірнісних оцінок, що дозволяє узгодити статистичну достовірність прогнозів із практичними порогами прийняття рішень. На виході формується стабілізована версія моделі, готова до подальшого використання на етапі інференсу. Уся послідовність забезпечує єдність між фазами моделювання шуму, оптимізацією параметрів і контролем якості.

Послідовність інференсу (рисунок 3.4) відображає рух даних від моменту взаємодії користувача з сервісом до отримання класифікаційного результату. Процес розпочинається з надсилання текстового запиту через прикладний інтерфейс, який забезпечує маршрутизацію даних до підсистеми попередньої обробки. На цьому етапі здійснюється нормалізація, токенізація та приведення вхідного повідомлення до уніфікованого формату, що відповідає параметрам моделі.

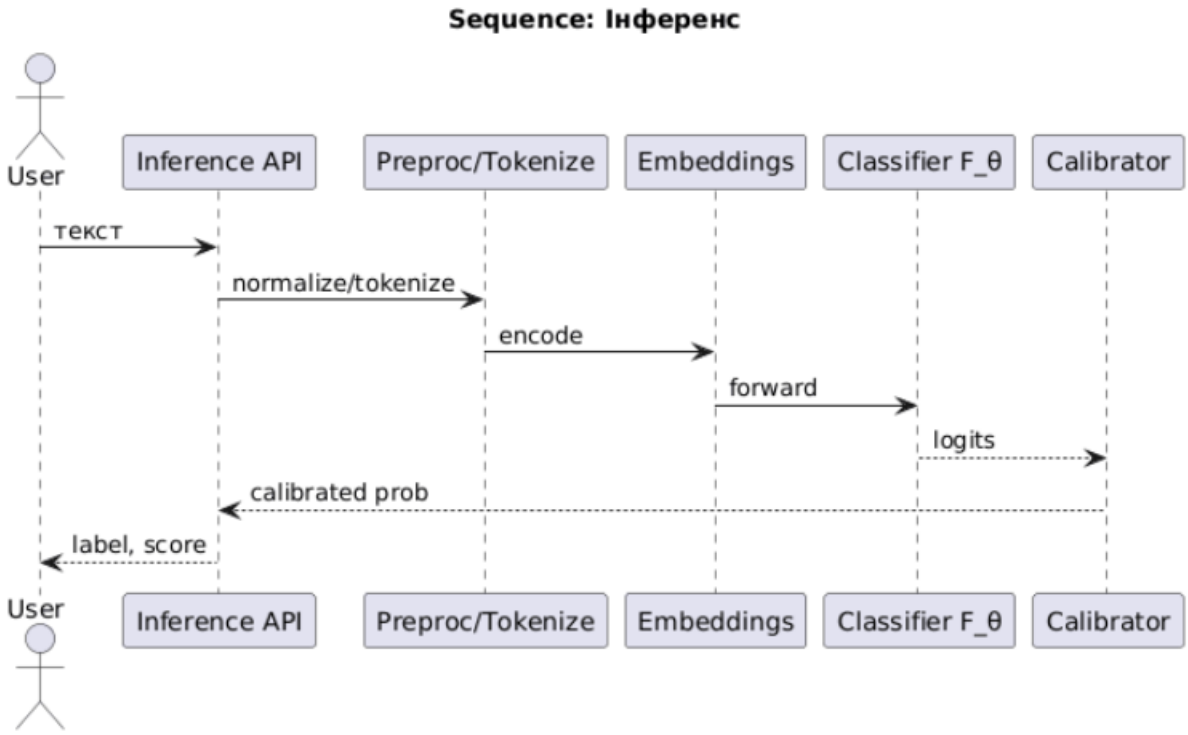


Рисунок 3.4 – Діаграма послідовності процесу інференсу

Після перетворення в текстові токени дані кодуються у векторне представлення в модулі ембеддингів, яке відтворює семантичні зв’язки між одиницями тексту. Отримані тензори подаються на вхід класифікаційної моделі, де відбувається пряме обчислення логітів. Результати обробки передаються до етапу калібрування ймовірнісних оцінок, що узгоджує величини вихідного розподілу з емпіричною статистикою, отриманою при навчанні.

Скориговані значення повертаються до сервісного інтерфейсу у вигляді цільової мітки та відповідного балу впевненості. Користувач отримує результат у тому самому сеансі взаємодії, що забезпечує безперервність і прозорість

інтерпретації. Уся послідовність демонструє цілісність обчислювального конвеєра, де жоден компонент не функціонує автономно, а узгоджене проходження даних гарантує стабільність роботи системи в реальному часі.

3.3 Функціональні можливості інтелектуальної системи

У межах кваліфікаційної роботи проєктується інтелектуальна система для автоматизованого виявлення повідомлень із ознаками мови ворожнечі у соціальних текстах за умов наявності шумів і навмисних спотворень (опечатки, маскування символами, транслітерація, лексичні підміни). Інтелектуальна система проєктується як інтерактивний вебінструмент для дослідження автоматизованого виявлення мови ворожнечі у коротких соціальних повідомленнях за умов навмисної або природної обфускації тексту. У межах підходу передбачається поєднання: (1) керованого моделювання текстових спотворень (шумів) для імітації реальних комунікаційних викривлень, (2) змішаного навчання на чистих і зашумлених прикладах у строго заданій пропорції на рівні міні-пакетів, (3) ручного циклу тонкого налаштування трансформерної моделі з моніторингом історії навчання, (4) калібрування вихідних оцінок і вибору порогу рішення за критерієм F1, (5) оцінювання узагальнювальної здатності на тестовій вибірці та (6) керування реєстром навчальних артефактів для відтворюваності експериментів.

Нижче наведено проєктні підсистеми, їх призначення, інформаційні потоки та функціональні обмеження.

Підсистема керування набором даних

Призначення. Підсистема проєктується для завантаження, первинної валідації та уніфікації навчальних даних у форматі таблиці повідомлень. Її роль полягає у стандартизації структури вхідного набору, усуненні некоректних записів і підготовці узгодженого внутрішнього подання даних для подальших етапів.

Вхідні дані. Файл набору даних у форматі CSV; назва колонки з текстом повідомлення; назва колонки з міткою класу (бінарна мітка 0/1); параметри формування вибірок (частки для валідаційної і тестової підвибірок, початкове

значення генератора випадковості); параметр обмеження кількості прикладів на клас.

Вихідні дані. Внутрішній масив повідомлень і відповідних міток класу; метадані про кількість рядків до/після очищення, розподіл класів, факт застосування обмеження; попередній перегляд кількох прикладів.

Основні функції:

– завантаження набору даних із CSV і зчитування колонок «текст повідомлення» та «мітка класу»;

– валідація коректності значень міток класу та відсів порожніх/некоректних записів;

– підрахунок базових статистик (обсяг даних, розподіл класів);

– формування короткого попереднього перегляду прикладів для контролю коректності завантаження.

Обмеження. Підсистема передбачає бінарну постановку задачі та вимагає узгоджених міток класу (0/1). Якість подальших етапів істотно залежить від репрезентативності і чистоти початкового набору даних, а також від достатнього обсягу прикладів обох класів.

Підсистема балансування та формування підвибірок

Призначення. Підсистема проєктується для керування співвідношенням класів у використаній частині набору даних та для формування стратифікованих підвибірок навчання, валідації і тестування, що забезпечує коректність оцінювання та відтворюваність експериментів.

Вхідні дані. Масив повідомлень і міток класу після первинної валідації; максимальна кількість прикладів на клас; частки валідаційної і тестової підвибірок; початкове значення генератора випадковості.

Вихідні дані. Підвибірки навчання/валідації/тестування у вигляді множин індексів; метадані застосованого обмеження (скільки прикладів відібрано для кожного класу) та розміри підвибірок.

Основні функції:

– відбір не більш ніж заданої кількості прикладів для кожного класу із випадковою перестановкою;

– стратифіковане розбиття на навчальну, валідаційну та тестову підвибірки із збереженням пропорцій класів;

– фіксація параметрів розбиття для відтворюваності.

Обмеження. За надто малого обсягу даних або за сильного дисбалансу класів стратифіковане розбиття може втрачати стабільність, що знижує надійність оцінювання. Додатково, агресивне обмеження кількості прикладів на клас може погіршувати узагальнювальну здатність моделі.

Підсистема керування лексиконом підмін

Призначення. Підсистема проєктується для підключення зовнішнього лексикону підмін як керованого джерела лексичних замін, що використовується під час моделювання шумів. Це дозволяє імітувати заміну лексем на евфемізми, жаргонні або обхідні форми написання.

Вхідні дані. Файл лексикону у форматі JSON, що містить відповідності «слово / заміна»; вимоги до нормалізації ключів (нижній регістр, обрізання пробілів).

Вихідні дані. Внутрішнє відображення лексикону (словник підмін); статистика розміру лексикону для контролю завантаження.

Основні функції:

– завантаження та синтаксична перевірка файлу лексикону;

– нормалізація ключів і значень та відсів некоректних пар;

– надання лексикону як параметра для підсистеми моделювання шумів.

Обмеження. Лексикон підмін є доменно-залежним: некоректні або надто агресивні підміни можуть руйнувати семантику висловлювання і створювати шум, який не відповідає реальним сценаріям обфускації.

Підсистема моделювання шумів

Призначення. Підсистема проєктується для керованого внесення штучних спотворень у текстові повідомлення з метою імітації типових варіантів обфускації у

вебкомунікації та підвищення стійкості моделі до варіативного написання, часткової втрати символів і навмисних графічних підмін.

Вхідні дані. Текст повідомлення; параметри конфігурації шумів, що визначають імовірності застосування окремих перетворень, інтервал частки слів для модифікації, а також символ маскування; за потреби – лексикон підмін; початкове значення генератора випадковості.

Вихідні дані. Зашумлений текст; набір демонстраційних прикладів зашумлення для попереднього перегляду налаштувань.

Основні функції:

– сегментація тексту із збереженням пробілів та розділових знаків;

– вибір частки слів, що підлягають модифікації, відповідно до заданого інтервалу;

– внесення опечаток шляхом перестановки, видалення або дублювання символів;

– маскування частини символів із заміною їх на символ маскування;

– графемні підміни між кирилицею та латиницею для моделювання змішаного написання;

– лексичні підміни на основі підключеного лексикону;

– формування кількох зашумлених варіантів одного повідомлення для візуальної перевірки.

Обмеження. Інтенсивність шумів повинна бути узгоджена з природою даних: надмірні спотворення можуть зменшувати інформаційну цінність повідомлення та ускладнювати навчання. Крім того, окремі типи шумів можуть мати різну релевантність для різних мов і платформ комунікації.

Підсистема змішаного навчання на рівні міні-пакетів

Призначення. Підсистема проєктується для реалізації керованого змішування чистих і зашумлених прикладів у кожному міні-пакеті навчання. Такий механізм забезпечує точне дотримання заданої частки чистих повідомлень у міні-пакеті та стабілізує експериментальні умови при зміні загальної інтенсивності шумів.

Вхідні дані. Міні-пакет повідомлень і міток класу; частка чистих повідомлень у міні-пакеті; конфігурація шумів; лексикон підмін (за наявності); параметри токенізації (максимальна довжина, стратегія доповнення).

Вихідні дані. Пакет тензорів для моделі (ідентифікатори токенів, маски уваги тощо) та відповідні мітки класу, сформовані з урахуванням точного співвідношення чистих і зашумлених прикладів.

Основні функції:

– випадковий вибір множини «чистих» прикладів у межах міні-пакета відповідно до заданої частки;

– застосування шумів до решти прикладів у міні-пакеті;

– токенізація текстів і формування вхідних тензорів для моделі;

– синхронізація генератора випадковості для забезпечення відтворюваності.

Обмеження. Точне змішування на рівні міні-пакетів залежить від розміру міні-пакета: при малих значеннях розміру міні-пакета дискретизація може спричиняти відхилення від бажаної пропорції у конкретних пакетах, навіть якщо в середньому вона зберігається.

Підсистема навчання трансформерної моделі

Призначення. Підсистема проєктується для тонкого налаштування попередньо навченої трансформерної моделі у бінарній задачі виявлення мови ворожнечі з урахуванням зашумлених даних та керованого змішування прикладів.

Вхідні дані. Навчальна та валідаційна підвибірки; вибрана базова трансформерна архітектура; гіперпараметри навчання (швидкість навчання, кількість епох, розмір міні-пакета, коефіцієнт регуляризації); параметри токенізації; конфігурація шумів і параметри змішування.

Вихідні дані. Набір параметрів моделі після навчання; історія значень функції втрат на кроках навчання та метрики на валідації по епохах; артефакти моделі і токенізатора для подальшого завантаження.

Основні функції:

– ініціалізація токенізатора і моделі за обраною базовою архітектурою;

– формування потоків даних для навчання і валідації;

– виконання циклу навчання з обмеженням градієнтів та оптимізацією AdamW;

– обчислення функції втрат на навчанні та її логування по кроках;

– обчислення функції втрат і метрик якості на валідації після кожної епохи;

– вибір найкращого стану моделі за критерієм валідаційної F_1-міри.

Обмеження. Навчання є обчислювально затратним і залежить від доступності графічного прискорення. Додатково, результати можуть бути чутливими до параметрів шуму, розміру міні-пакета та вибраної базової архітектури, що вимагає систематичного налаштування.

Підсистема моніторингу та візуалізації процесу навчання

Призначення. Підсистема проєктується для інтерактивного контролю стану навчання та аналізу динаміки оптимізації. Передбачається подання двох ключових графіків: залежність навчальної втрати від номера кроку та узагальнювальна динаміка на валідації.

Вхідні дані. Поточний стан навчального процесу; історія навчальної втрати по кроках; історія валідаційної втрати та F_1-міри по епохах; параметри нормалізації валідаційної втрати для візуалізації.

Вихідні дані. Текстовий статус навчання; графік «навчальна втрата – крок»; графік на одній шкалі 0..1 для порівняння «F_1-міра» та «інвертована нормалізована валідаційна втрата».

Основні функції:

– відображення поточного статусу навчання та останніх оцінених метрик;

– побудова графіка навчальної втрати у функції номера кроку;

– побудова валідаційного графіка на одній шкалі 0..1: F_1 та інвертоване нормалізоване значення валідаційної втрати;

– періодичне оновлення графіків під час виконання навчання.

Обмеження. Нормалізація валідаційної втрати для відображення на шкалі 0..1 є візуалізаційним перетворенням і не повинна інтерпретуватися як пряме порівняння чисельних значень втрати та F1. Такий графік слугує для якісного аналізу узгодженості тенденцій, а не для заміни чисельних метрик.

Підсистема оцінювання якості на тестовій підвибірці

Призначення. Підсистема проєктується для підсумкового оцінювання узагальнювальної здатності моделі на відкладеній тестовій підвибірці із застосуванням зафіксованих параметрів калібрування та порогу рішення.

Вхідні дані. Тестова підвибірка (тексти і мітки класу); навчений артефакт моделі; параметри токенізації; температура калібрування і поріг рішення.

Вихідні дані. Набір тестових метрик (точність, повнота, точність класифікації, F_1-міра) для прийнятого порогу.

Основні функції:

– виконання інференсу на тестовій підвибірці пакетами;

– застосування калібрування і порогу рішення;

– обчислення стандартних метрик якості бінарної класифікації.

Обмеження. Результати тестування є валідними лише за умови коректного формування тестової підвибірки та відсутності перетину із навчальними даними. Для нестабільних або малих тестових підвибірок метрики можуть мати високу дисперсію.

Підсистема інференсу окремих повідомлень

Призначення. Підсистема проєктується для прикладного використання навченої моделі у режимі класифікації одиничних повідомлень із поверненням оцінки належності до класу мови ворожнечі з урахуванням калібрування і порогу.

Вхідні дані. Текст повідомлення; навчений артефакт моделі; параметри токенізації; температура калібрування та поріг рішення.

Вихідні дані. Прогнозований клас; назва класу у зрозумілому вигляді; ймовірнісна оцінка; зафіксовані параметри порогу і калібрування, що використовувалися під час прийняття рішення.

Основні функції:

– нормалізація тексту повідомлення (уніфікація пробілів та службових символів);

– токенізація і пропускання через модель;

– застосування калібрування і порогу;

– формування структурованого результату для вебінтерфейсу.

Обмеження. Інференс передбачає відповідність даних домену навчання; при суттєвому доменному зсуві або появі нових способів обфускації зростає ризик деградації якості. Бінарна інтерпретація також не відображає усієї палітри токсичних проявів, що може вимагати розширення класів у подальших дослідженнях.

Підсистема реєстру моделей та відтворюваності експериментів

Призначення. Підсистема проєктується для накопичення навчальних артефактів, їх опису метаданими та подальшого завантаження для повторного використання без необхідності повторного навчання. Це забезпечує відтворюваність експериментів і порівнянність результатів між різними конфігураціями шумів, змішування та гіперпараметрів.

Вхідні дані. Параметри навчання, шумів, змішування, токенізації; навчена модель; токенізатор; параметри калібрування і порогу; часові мітки та ідентифікатори запуску.

Вихідні дані. Запис у реєстрі моделей із унікальною назвою; файли артефакту моделі й токенізатора; файл метаданих експерименту; перелік доступних записів реєстру; завантажений стан моделі за вибраним записом.

Основні функції:

– збереження артефакту моделі та токенізатора з унікальним ідентифікатором;

– збереження метаданих експерименту (параметри, найкращі валідаційні показники, калібрування, поріг);

– формування переліку доступних записів реєстру для перегляду;

– завантаження вибраного запису реєстру із відновленням моделі, токенізатора та параметрів рішення.

Обмеження. Реєстр потребує дисципліни експериментування: при великій кількості запусків необхідні правила іменування, архівації та контролю версій даних. Також переносимість артефактів може залежати від сумісності версій бібліотек і базових моделей.

3.4 Метрики оцінювання нейромережі для виявлення мови ворожнечі

Оцінювання якості нейромережевої моделі для виявлення мови ворожнечі ґрунтується на зіставленні передбачених нею міток із фактичними розподілами у валідаційній підмножині корпусу. Для коректного вимірювання здатності системи розрізняти токсичні висловлювання та нейтральні повідомлення застосовують метрики, що відображають не лише загальну точність, а й баланс між повнотою виявлення та коректністю класифікації. В умовах асиметричного розподілу класів, коли прикладів ворожих висловлювань суттєво менше, ніж нейтральних, традиційна accuracy не відображає реальної якості роботи моделі, оскільки може приховувати значну частку пропущених випадків.

Тому ключового значення набувають показники precision та recall, які характеризують відповідно частку правильно ідентифікованих випадків серед позитивних передбачень і здатність моделі виявляти всі релевантні зразки [67]. Їхнє узгоджене співвідношення репрезентується через F_1-міру, що зменшує вплив дисбалансу даних і дозволяє оцінити модель як цілісну систему . Для моделей із ймовірнісним виходом додатково аналізується площа під ROC-кривою, що демонструє, як змінюється співвідношення істинно позитивних і хибно позитивних рішень за різних порогів. Каліброваність моделі також відіграє значну роль, адже адекватність прогнозованих імовірностей визначає коректність подальшого використання системи в автоматизованих рішеннях або при інтеграції з модераційними інтерфейсами [68].

Загалом система оцінювання не зводиться до одноразового підрахунку метрик, а дає змогу простежити динаміку навчання, порівнювати різні конфігурації гіперпараметрів і виявляти типові помилки на прикордонних випадках. Таке трактування забезпечує не лише кількісне підтвердження ефективності моделі, а й створює основу для її подальшого вдосконалення та адаптації до змінних мовних і соціокультурних контекстів.

Висновки до розділу 3

У розділі сформовано архітектуру інтелектуальної системи виявлення мови ворожнечі та обґрунтовано вибір технологічного середовища, здатного забезпечити відтворюваність експериментів і безперервність повного циклу обробки даних. Використання Python у поєднанні з середовищем Google Colab створює умови для інтеграції попередньо навчених трансформерних моделей, формування розширених корпусів із контрольованими шумовими викривленнями, а також для проведення тренування та інференсу в єдиній конфігурації без розривів між етапами.

Запропонована компонентна структура відображає послідовний перехід від підготовки даних до застосування моделі, де кожен модуль виконує окреслену функцію й взаємодіє з іншими через чітко визначені артефакти. Механізми шумогенерації та змішування даних інтегровані безпосередньо у конвеєр попередньої обробки, що дозволяє навчати класифікатор у режимі, максимально наближеному до реальних умов зашумлення. Діаграми варіантів використання та послідовностей фіксують рольовий розподіл і порядок викликів, завдяки чому забезпечується узгодженість між інженерними та експлуатаційними практиками.

Процедури валідації та калібрування закладають підґрунтя для керованої роботи системи в операційному середовищі: порогові рішення приймаються на основі узгоджених імовірнісних оцінок, а оцінювання якості враховує дисбаланс класів і прикордонні випадки.

Отримані результати проєктування створюють методологічну й технічну основу для наступного етапу програмної реалізації, проведення експериментів і емпіричної перевірки гіпотези щодо підвищення точності виявлення мови ворожнечі в умовах зашумлених соціальних текстових даних.

РОЗДІЛ 4 Експериментальна установка та дослідження методу виявлення мови ворожнечі у зашумлених соціальних текстових даних

4.1 Програмна структура компонентів інтелектуальної системи

Діаграма пакетів і класів, яка відображає модульну організацію програмної реалізації системи наведена на рисунку 4.1. Вона показує, як програмні компоненти згруповані за функціональними блоками й у якій послідовності між ними відбувається передавання даних та виклик функцій.

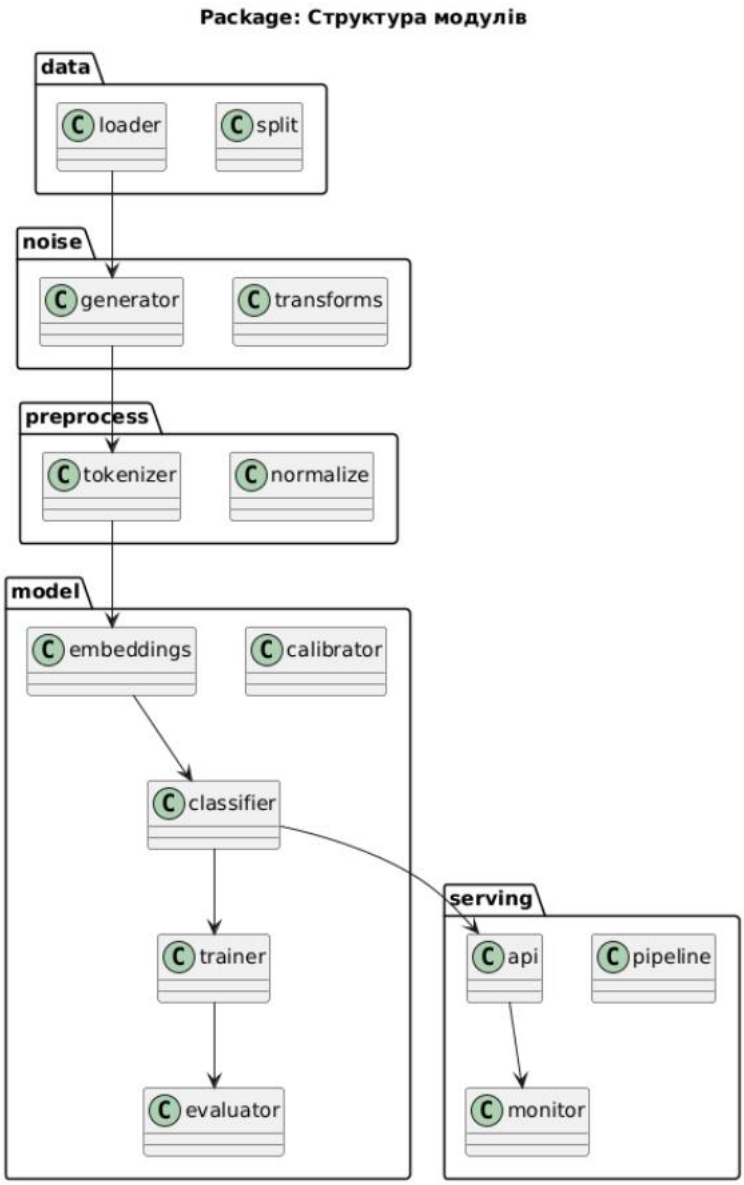


Рисунок 4.1 – Діаграма пакетів

Пакет data містить базову інфраструктуру роботи з корпусом. Класи завантаження забезпечують отримання текстів з джерела, тоді як модуль розподілу формує навчальні, валідаційні й тестові підвибірки. Саме на цьому рівні задається структура даних, що передається далі у конвеєр.

Пакет noise відповідає за формування штучно зашумлених прикладів. Генератор застосовує стохастичні перетворення до тексту, а окремі трансформаційні модулі реалізують конкретні типи викривлень, сумісні з програмною реалізацією. Тут формується проміжний корпус, який поєднується з початковими даними.

Пакет preprocess відтворює кроки підготовки тексту перед передачею в модель. Токенізатор виконує сегментацію та формує послідовності, придатні до векторизації, а модуль нормалізації узгоджує їх із форматом обраної архітектури без руйнування релевантних шумових маркерів.

Пакет model репрезентує ядро нейромережевої частини. Підсистема ембеддингів формує векторні подання токенів, класифікатор виконує розпізнавання, тренувальний модуль відповідає за оновлення параметрів, а компонент оцінювання здійснює обчислення метрик. Окремо виділений калібратор коригує ймовірнісні виходи для узгодженості з реальним розподілом даних.

Пакет serving моделює етап використання моделі після навчання. Сервісний інтерфейс забезпечує обробку зовнішніх запитів і виклик інференсу, а конвеєр формує послідовність застосування токенізації, ембеддингів і класифікації. Компонент моніторингу акумулює результати роботи, фіксує якість і виявляє аномалії або деградацію точності.

Уся структура побудована так, що потоки даних узгоджено переходять від завантаження корпусу до застосування моделі. Шумогенерація інтегрована до підготовчого етапу, препроцесинг зберігає зв’язок із навчанням, а сервісний шар працює на тих самих артефактах, що й тренувальне середовище. Така організація мінімізує розрив між експериментальною частиною та застосуванням і забезпечує розширюваність системи без модифікації цілісної архітектури.

Діаграма, наведена на рисунку 4.2, демонструє логіку розгортання системи виявлення мови ворожнечі як сервісу, доступного зовнішнім користувачам або

модераторам через програмний інтерфейс. Вона відображає ієрархію компонентів, послідовність обміну даними та розподіл відповідальності між обчислювальними, сервісними й інфраструктурними вузлами.

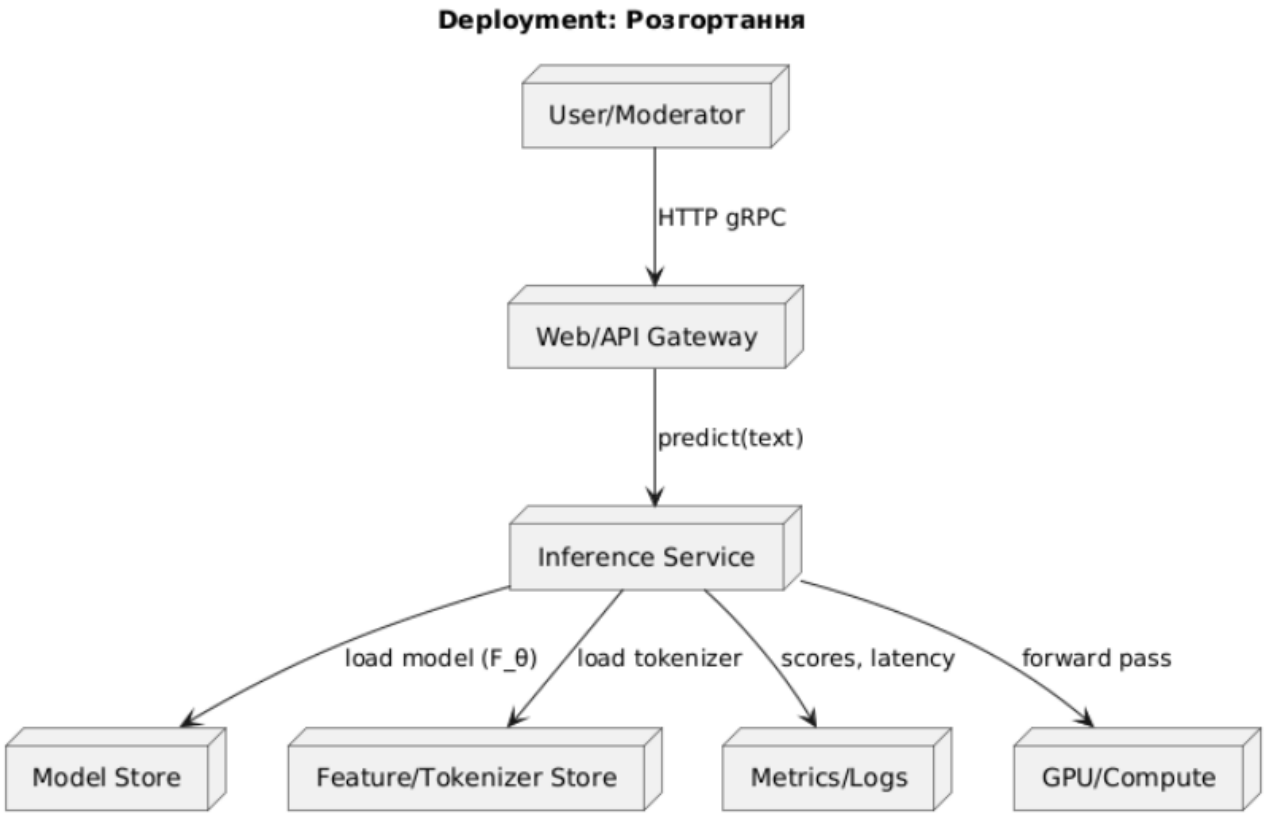


Рисунок 4.2 – Схема розгортання інтелектуальної системи

На верхньому рівні джерелом запиту є користувач або модератор, який ініціює аналіз тексту через клієнтський інтерфейс. Запит передається в систему з використанням HTTP-з’єднання, що забезпечує платформонезалежний виклик сервісу. Далі він надходить до веб-шлюзу або API-шару, який виконує маршрутизацію й валідацію вхідних повідомлень, а також ініціює процедуру передбачення через метод, аналогічний виклику predict(text).

Основне навантаження з виконання класифікації зосереджено в сервісі інференсу. Цей модуль містить реалізацію прямого проходу моделі, послідовного застосування токенізатора та механізму калібрування ймовірностей. Для коректної роботи він завантажує параметри класифікатора з окремого сховища моделей, де зберігається навчена версія F_θ. Паралельно здійснюється доступ до словника

токенізатора та допоміжних конфігурацій, які містяться в репозиторії мовних представлень або ознак.

Під час кожного запиту сервіс може задіювати обчислювальні ресурси центрального чи графічного процесора, залежно від обсягу даних і вимог до латентності. Обчислення логітів і ймовірностей виконується як послідовний forward pass, що повторює структуру навчальної архітектури. Після отримання результату сервіс реєструє продуктивність, час відповіді та успішність виконання у системі моніторингу або логування. Такі журнали можуть надалі використовуватися для повторного аналізу, виявлення деградації точності або адаптації системи до нових типів шуму.

Таким чином, розгортання організоване як оркестрація взаємопов’язаних компонентів, де кожен вузол відповідає за власну підфункцію: прийом запиту, його маршрутизацію, виконання інференсу, доступ до артефактів моделі, реєстрацію показників і опрацювання результатів. Архітектура дозволяє масштабувати систему, оновлювати модель без зупинки сервісу та інтегрувати її у модераційні панелі або зовнішні програмні рішення.

4.2 Алгоритмічна специфікація прикладних компонентів інтелектуальної системи

Подано узгоджену алгоритмічну специфікацію прикладних компонентів інтелектуальної системи у вигляді псевдокодів та їх опису. Подання у псевдокодовій формі виконує роль проєктної фіксації логіки оброблення даних, структури вхідних і вихідних інформаційних потоків, а також контрольованих параметрів експерименту. Для забезпечення однозначності трактування в цьому пункті використовується єдина схема опису: процедура розглядається як завершена алгоритмічна одиниця, що має визначені вхідні та вихідні дані; функція – як допоміжна операція, що повертає значення та може використовуватися в межах процедур. Надалі усі ключові операції подано саме як процедури, а допоміжні перетворення трактуються як функції.

У проєктованій системі дані імпортуються з табличного файлу, який містить дві суттєві колонки: текст повідомлення та мітка класу. Процедура завантаження та валідації набору повідомлень виконує перевірку наявності відповідних колонок, відсіювання порожніх повідомлень, приведення міток до цілочисельного формату та фільтрацію допустимого набору міток у бінарній постановці (0/1). Окремо передбачено контроль мінімально допустимого обсягу валідних записів, оскільки надто малий обсяг робить подальше розбиття та навчання статистично нестійкими.

```
ПРОЦЕДУРА ЗавантажитиТабличніДані(шлях_до_файлу, назва_колонки_повідомлення, назва_колонки_мітки)
ВХІД: табличний файл; назви колонок для повідомлення і мітки
ВИХІД: повідомлення[], мітки[]

1: відкрити файл як таблицю із заголовком
2: якщо назва_колонки_повідомлення або назва_колонки_мітки відсутні → ПОМИЛКА
3: ініціалізувати повідомлення[] ← порожній, мітки[] ← порожній
4: ДЛЯ кожного рядка таблиці:
5:     t ← обрізати пробіли(значення(назва_колонки_повідомлення))
6:     y ← значення(назва_колонки_мітки)
7:     якщо t порожній → ПРОПУСТИТИ
8:     якщо y не перетворюється у ціле число → ПРОПУСТИТИ
9:     якщо y ∉ {0,1} → ПРОПУСТИТИ
10:    додати t до повідомлення[]
11:    додати y до мітки[]
12: якщо |повідомлення[]| < мінімально_достатньо → ПОМИЛКА
13: ПОВЕРНУТИ повідомлення[], мітки[]
КІНЕЦЬ
```

Далі передбачається процедура обмеження кількості прикладів на клас, яка у проєктованій системі виконує дві задачі: контроль експериментального бюджету обчислень та вирівнювання впливу класів під час швидких ітерацій. Ідея полягає у випадковому відборі не більш ніж M повідомлень для кожного класу з використанням фіксованого зерна генератора випадковостей. Це забезпечує відтворюваність підвибірки та порівнюваність серій експериментів при зміні параметрів шумів або частки чистих даних.

```
ПРОЦЕДУРА ОбмежитиПрикладиЗаКласом(повідомлення[], мітки[], максимум_на_клас, зерно)
ВХІД: масив повідомлень; масив міток; максимум_на_клас; зерно
ВИХІД: повідомлення2[], мітки2[], метадані

1: якщо максимум_на_клас ≤ 0 → ПОВЕРНУТИ повідомлення[], мітки[], {"обмеження": НІ}
2: індекси0 ← всі i, де мітки[i]=0
3: індекси1 ← всі i, де мітки[i]=1
4: перемішати(індекси0, зерно); перемішати(індекси1, зерно)
```

```
        5: вибрані0 ← перші min(максимум_на_клас, |індекси0|) елементів індекси0
        6: вибрані1 ← перші min(максимум_на_клас, |індекси1|) елементів індекси1
        7: вибрані ← об’єднати(вибрані0, вибрані1); перемішати(вибрані, зерно)
        8: сформувати повідомлення2[], мітки2[] за індексами вибрані
        9:  метадані  ←  {"обмеження":  ТАК,  "максимум_на_клас":  максимум_на_клас,  "клас0":
|вибрані0|, "клас1": |вибрані1|}
        10: ПОВЕРНУТИ повідомлення2[], мітки2[], метадані
        КІНЕЦЬ
```

Наступним необхідним кроком у проєктуванні є процедура стратифікованого розбиття даних на підвибірки навчання, валідації та тестування. Принциповою вимогою є збереження пропорцій класів у кожній підвибірці, оскільки в іншому випадку метрики можуть бути зміщені через нерепрезентативність підмножин. У процедурі застосовується двоетапний підхід: спочатку формується відкладена сукупність для «валідація+тест» (рисунок 4.3), після чого вона додатково розбивається на валідаційну та тестову частини у заданих пропорціях.

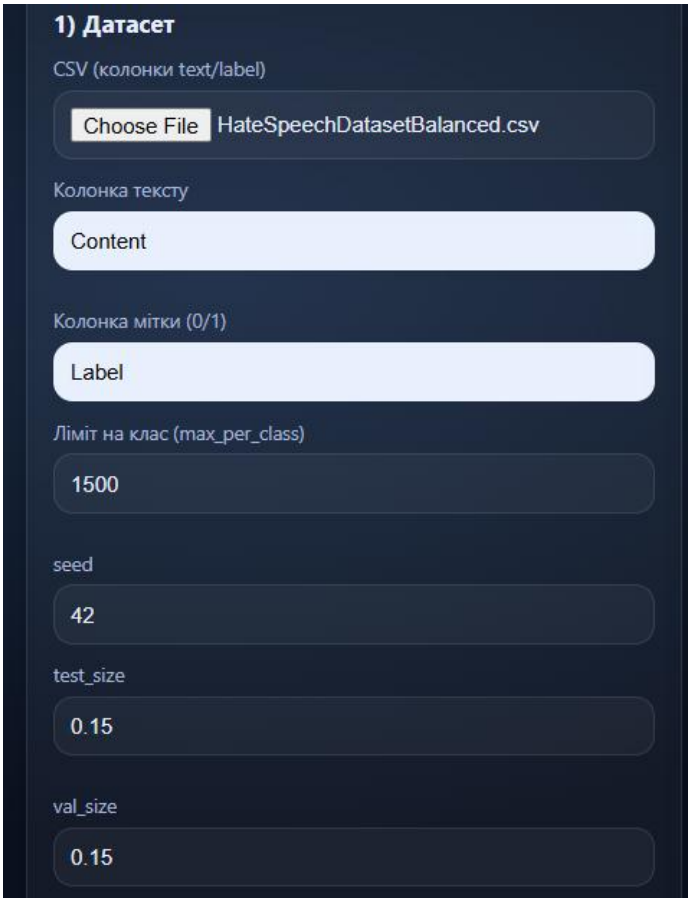


Рисунок 4.3 – Приклад розбиття датасету

```
ПРОЦЕДУРА СтратифікованоРозбитиДані(мітки[], частка_тест, частка_валідації, зерно)
ВХІД: масив міток; частки тест/валідації; зерно
```

```
ВИХІД: індекси_навчання[], індекси_валідації[], індекси_тесту[]

1: індекси0 ← всі i, де мітки[i]=0
2: індекси1 ← всі i, де мітки[i]=1
3: перемішати(індекси0, зерно); перемішати(індекси1, зерно)

4: частка_відкладення ← частка_тест + частка_валідації
5: k0 ← округлити(|індекси0| * частка_відкладення)
6: k1 ← округлити(|індекси1| * частка_відкладення)

7: індекси_відкладені ← (перші k0 з індекси0) ∪ (перші k1 з індекси1)
8: індекси_навчання ← (решта індекси0) ∪ (решта індекси1)

9: частка_тест_у_відкладених ← частка_тест / частка_відкладення
10: стратифіковано поділити індекси_відкладені на індекси_тесту та індекси_валідації
за частка_тест_у_відкладених

11: ПОВЕРНУТИ індекси_навчання, індекси_валідації, індекси_тесту
КІНЕЦЬ
```

Окремим прикладним компонентом у проєктованій системі є процедура завантаження лексикону підмін, який використовується для контрольованого відтворення лексичних замін у межах моделювання шумів. Лексикон задається як словник відповідностей «вхідне слово / заміна» (рисунок 4.4) та нормалізується до нижнього регістру для ключів, що мінімізує залежність від регістру символів у повідомленнях.

Рисунок 4.4 – Приклад лексикону

Такий механізм дозволяє моделювати евфемізми, заміни на нейтральні або обхідні форми, а також типові варіанти «сленгового» написання.

```
ПРОЦЕДУРА ЗавантажитиЛексиконПідмін(шлях_до_JSON)
ВХІД: JSON-файл із парами "слово":"заміна"
ВИХІД: лексикон_підмін

1: прочитати JSON як словник
2: лексикон_підмін ← порожній
3: ДЛЯ кожної пари (ключ, значення):
4:     якщо ключ і значення є рядками:
5:         k ← нижній_регістр(обрізати_пробіли(ключ))
6:         v ← обрізати_пробіли(значення)
7:         якщо k не порожній і v не порожній:
8:             лексикон_підмін[k] ← v
9: ПОВЕРНУТИ лексикон_підмін
КІНЕЦЬ
```

Ключовим елементом проєктування є процедура моделювання шумів D', яка формує зашумлену версію повідомлення за рахунок контрольованих спотворень. У межах цієї процедури проєктується: (1) розбиття тексту зі збереженням пробілів та розділових знаків, (2) вибір частки слів для модифікації у межах заданого інтервалу, (3) стохастичний вибір типу перетворення відповідно до ваг (ймовірностей), (4) застосування конкретного перетворення (опечатка, маскування символів, підміна графем, лексична підміна). Важливо, що параметри шумів задаються як числові величини та використовуються саме як керовані регулятори інтенсивності, що дозволяє відтворювати однакові умови при порівнянні різних конфігурацій навчання.

```
ПРОЦЕДУРА ЗашумитиПовідомлення(повідомлення, параметри_шуму, лексикон_підмін, зерно)
ВХІД: текст повідомлення; параметри шуму; лексикон підмін (може бути порожнім); зерно
ВИХІД: зашумлене_повідомлення

1: частини ← розбити_зі_збереженням_пробілів_і_пунктуації(повідомлення)
2: позиції_слів ← позиції елементів у частини, що є словами
3: якщо позиції_слів порожній → ПОВЕРНУТИ повідомлення

4: частка ← випадкове_значення(мін_частка_слів, макс_частка_слів, зерно)
5: m ← max(1, округлити(частка * |позиції_слів|))
6: обрані ← випадкова_підмножина(позиції_слів, m, зерно)

7: ДЛЯ кожної позиції p з обрані:
8:     тип ← вибрати_тип_перетворення_за_ймовірностями(параметри_шуму)
9:     якщо тип = "опечатка" → застосувати_опечатку(частини[p], зерно)
10:            якщо тип = "маскування" → замінити частину символів на
символ_маскування(частини[p])
```

```
        11:     якщо тип = "підміна_графем" → виконати графемні заміни (латиниця↔кирилиця) з імовірністю
        12:    якщо тип = "лексична_підміна" і лексикон_підмін не порожній:
        13:          замінити слово за лексиконом з урахуванням регістру

        14: зашумлене_повідомлення ← об’єднати(частини)
        15: ПОВЕРНУТИ зашумлене_повідомлення
        КІНЕЦЬ
```

На базі D′ проєктується процедура точного змішування $D\alpha$ на рівні міні-пакетів. Її призначення полягає в тому, щоб у кожному міні-пакеті забезпечити наперед задану частку «чистих» повідомлень, а решту замінити на зашумлені версії. Такий механізм зменшує варіативність режиму навчання та забезпечує строгий контроль над тим, наскільки модель «бачить» зашумлені приклади протягом оптимізації. У процедурі передбачено: обчислення кількості чистих повідомлень у пакеті, випадкове перемішування індексів пакета, поділ на чисті та зашумлені, а також формування вихідних масивів повідомлень та міток для подальшого токенізування і подачі у модель.

```
        ПРОЦЕДУРА      СформуватиМініПакетЗіЗмішуванням(пакет_прикладів,      частка_чистих, параметри_шуму, лексикон_підмін, зерно)
        ВХІД: пакет пар (повідомлення, мітка); частка чистих; параметри шуму; лексикон; зерно
        ВИХІД: пакет_повідомлень[], пакет_міток[]

        1: B ← кількість прикладів у пакеті
        2: n_чистих ← округлити(частка_чистих * B)
        3: n_чистих ← обмежити у межах 0..B
        4: індекси ← [0..B-1]; перемішати(індекси, зерно)
        5: чисті ← перші n_чистих індексів; зашумлені ← решта

        6: ініціалізувати пакет_повідомлень[] і пакет_міток[] як порожні
        7: ДЛЯ j від 0 до B-1:
        8:     (t, y) ← пакет_прикладів[j]
        9:     якщо j ∈ зашумлені:
        10:         t ← ЗашумитиПовідомлення(t, параметри_шуму, лексикон_підмін, зерно)
        11:    додати t до пакет_повідомлень[]
        12:    додати y до пакет_міток[]
        13: ПОВЕРНУТИ пакет_повідомлень[], пакет_міток[]
        КІНЕЦЬ
```

Процедура навчання трансформерної моделі проєктується як ручний цикл оптимізації з журналюванням історії навчання та контролем якості на валідаційній підвибірці. Важливою проєктною вимогою є фіксація двох типів історії: (1) значення навчальної втрати на кожному кроці оптимізації, що дозволяє аналізувати збіжність

і стабільність, (2) значення валідаційної втрати та F_1-міри після кожної епохи, що використовується для вибору найкращого стану моделі. У межах проєктування передбачається підбір порогу рішення за критерієм максимальної F_1-міри на валідації, що є релевантним для задачі з потенційно асиметричними помилками.

```
ПРОЦЕДУРА НавчитиМодельТрансформера(індекси_навчання, індекси_валідації, повідомлення[], мітки[], конфігурація, процедура_Dα)
ВХІД: індекси навчання/валідації; повідомлення; мітки; конфігурація; процедура Dα
ВИХІД: модель_найкраща, історія_навчання, логіти_валідації_найкращі, мітки_валідації

1: ініціалізувати токенізатор і трансформерну модель для бінарної класифікації
2: ініціалізувати оптимізатор і функцію втрат
3: історія ← порожні масиви: кроки[], навч_втрата[], епохи[], вал_втрата[], вал_F1[]
4: найкращий_F1 ← -∞; стан_найкращий ← НІ; логіти_найкращі ← НІ

5: ДЛЯ епоха = 1..E:
6:     встановити модель у режим навчання
7:     ДЛЯ кожного міні-пакета з навчальних індексів:
8:         (пакет_повідомлень, пакет_міток) ← процедура_Dα(міні-пакет)
9:         виконати токенізацію пакета; обчислити логіти; обчислити втрату
10:        виконати зворотний прохід; оновити параметри
11:        додати (номер_кроку, навч_втрата) до історія

12:    встановити модель у режим оцінювання
13:    обчислити логіти на валідації без зашумлення
14:    вал_втрата ← значення функції втрат на валідації
15:    ймовірності ← перетворити логіти у ймовірності позитивного класу
16:    поріг_епохи ← підібрати поріг, що максимізує F1 на валідації
17:    вал_F1 ← обчислити F1 за поріг_епохи
18:    додати (епоха, вал_втрата, вал_F1) до історія

19:    якщо вал_F1 > найкращий_F1:
20:         найкращий_F1 ← вал_F1
21:         зберегти поточний стан моделі як стан_найкращий
22:         зберегти логіти валідації як логіти_найкращі

23: завантажити стан_найкращий у модель
24: ПОВЕРНУТИ модель, історія, логіти_найкращі, мітки_валідації
КІНЕЦЬ
```

Після вибору найкращого стану моделі проєктується процедура калібрування оцінок і фіксації порогу рішення. Калібрування виконується шляхом масштабування логітів температурою, яку підбирають так, щоб мінімізувати втрату на валідації. Після цього поріг рішення повторно підбирається за максимальною F_1-мірою вже на каліброваних оцінках. З позиції проєктування це дозволяє отримувати більш узгоджені ймовірності та стабільні правила прийняття рішення під час тестування і подальшого використання моделі.

```
ПРОЦЕДУРА КалібруватиТаПідібратиПоріг(логіти_валідації, мітки_валідації)
ВХІД: логіти на валідації; істинні мітки
ВИХІД: температура, поріг, метрики_валідації

1: температура ← підібрати параметр температури, що мінімізує втрату на валідації
2: логіти2 ← логіти_валідації / температура
3: ймовірності ← softmax(логіти2) для позитивного класу
4: поріг ← підібрати поріг, що максимізує F1 на валідації
5: метрики_валідації ← обчислити (точність, точність_позитивного, повноту, F1) за поріг
6: ПОВЕРНУТИ температура, поріг, метрики_валідації
КІНЕЦЬ
```

Для контролю узагальнювальної здатності проєктується процедура оцінювання на тестовій підвибірці, яка використовує вже зафіксовані параметри калібрування та поріг рішення. Це усуває «підгонку» під тест і зберігає коректність фінальної оцінки. Процедура формує узагальнювальні метрики та, за потреби, може бути розширена для виведення матриці помилок або аналізу прикладів помилкових класифікацій.

```
ПРОЦЕДУРА ОцінитиНаТесті(індекси_тесту, повідомлення[], мітки[], модель, токенізатор, температура, поріг)
ВХІД: тестові індекси; повідомлення; мітки; модель; токенізатор; температура; поріг
ВИХІД: метрики_тесту

1: встановити модель у режим оцінювання
2: ініціалізувати масив ймовірностей і масив істинних міток
3: ДЛЯ кожного пакета тестових індексів:
4:     обчислити логіти моделі
5:     застосувати калібрування: логіти ← логіти / температура
6:     отримати ймовірності позитивного класу
7:     накопичити ймовірності та істинні мітки
8: метрики_тесту ← обчислити метрики за фіксованим порогом
9: ПОВЕРНУТИ метрики_тесту
КІНЕЦЬ
```

Завершальним компонентом проєктування є процедура керування реєстром артефактів моделей, що дозволяє накопичувати результати експериментів і відтворювати їх без повторного навчання. Реєстр розглядається як структура каталогів, де кожен запис містить параметри конфігурації, збережені ваги моделі та токенізатор, а також паспорт експерименту (час створення, параметри шуму, частка чистих повідомлень у міні-пакеті, температура калібрування, поріг рішення, метрики на валідації). Окремо проєктується процедура завантаження моделі з

реєстру за ідентифікатором запису, що забезпечує повторне використання у сценаріях оцінювання та інференсу.

```
ПРОЦЕДУРА ЗберегтиДоРеєстру(модель, токенізатор, паспорт_експерименту, каталог_реєстру)
ВХІД: модель; токенізатор; паспорт експерименту; каталог реєстру
ВИХІД: ідентифікатор_запису

1: ідентифікатор_запису ← сформувати унікальну назву (час + назва_запуску)
2: створити каталог запису у каталозі реєстру
3: зберегти параметри моделі у каталог запису
4: зберегти токенізатор у каталог запису
5: зберегти паспорт_експерименту у файл метаданих
6: ПОВЕРНУТИ ідентифікатор_запису
КІНЕЦЬ


ПРОЦЕДУРА ЗавантажитиЗРеєстру(каталог_реєстру, ідентифікатор_запису)
ВХІД: каталог реєстру; ідентифікатор запису
ВИХІД: модель, токенізатор, паспорт_експерименту

1: перевірити існування каталогу запису
2: завантажити токенізатор із каталогу запису
3: завантажити модель із каталогу запису
4: прочитати паспорт_експерименту з файлу метаданих
5: ПОВЕРНУТИ модель, токенізатор, паспорт_експерименту
КІНЕЦЬ
```

Узгодженість представлення експериментальних кривих у проєктованій системі забезпечується також введенням допоміжної функції перетворення валідаційної втрати до єдиної шкали з F_1-мірою для візуального аналізу. Оскільки F_1-міра належить інтервалу [0;1], а втрата має іншу шкалу, використовується нормалізація втрати за мінімумом і максимумом в історії та інверсія (чим менша втрата, тим більша «якість» на графіку). Це не змінює числових метрик, а лише уніфікує інтерпретацію кривих при аналізі навчання.

```
ФУНКЦІЯ НормалізуватиТаІнвертуватиВтрату(втрати[])
ВХІД: масив значень валідаційної втрати
ВИХІД: значення_на_шкалі_0_1[]

1: mn ← мінімум(втрати[]); mx ← максимум(втрати[])
2: якщо mx = mn → ПОВЕРНУТИ масив одиниць довжини |втрати[]|
3: ДЛЯ кожного v у втрати[]:
4:     norm ← (v - mn) / (mx - mn)
5:     inv ← 1 - norm
6:     додати inv до вихідного масиву
7: ПОВЕРНУТИ вихідний масив
КІНЕЦЬ
```

Наведена сукупність процедур і функції утворює проєктну основу прикладних компонентів системи та визначає її ключові властивості: кероване моделювання шумів у повідомленнях, строгий контроль частки зашумлених прикладів на рівні міні-пакетів, відтворюваність навчання за рахунок фіксації параметрів випадковості, коректний вибір найкращого стану моделі за валідаційною F_1-мірою, подальше калібрування оцінок і фіксацію порогу рішення, а також накопичення артефактів у реєстрі для повторного використання. Це забезпечує можливість систематичного порівняння конфігурацій шумів і часток змішування без апеляції до конкретної реалізації інтерфейсу, зберігаючи фокус на проєктуванні потоків даних, алгоритмів та правил прийняття рішення.

Подана алгоритмічна специфікація формалізує ключові прикладні компоненти інтелектуальної системи та узгоджує їхню взаємодію на рівні вхідних і вихідних інформаційних потоків. У псевдокодовій формі зафіксовано базові процедури підготовки даних, керованого моделювання спотворень повідомлень, точного змішування чистих і зашумлених прикладів у мініпакетах, навчання з валідаційним контролем, калібрування оцінок і фіксації правила прийняття рішення, а також ведення реєстру експериментальних артефактів. Такий опис забезпечує однозначність трактування логіки системи, відтворюваність експериментів.

4.3 Особливості використання інтелектуальної системи

Розроблена інтелектуальна система функціонує як вебзастосунок, що запускається у середовищі Google Colab і відкривається у вбудованому вікні браузера. Інтерфейс згруповано у три логічні панелі: «Датасет», «Навчання + графіки», «Оцінювання / Інференс / Реєстр». Вигляд головного вікна навадена на рисунку 4.5. Взаємодія користувача із системою реалізується через заповнення параметрів, завантаження файлів та натискання керувальних кнопок; у відповідь система виводить структурований протокол виконання (службові повідомлення,

параметри конфігурації, проміжні та підсумкові метрики), а також графіки перебігу навчання і валідації.

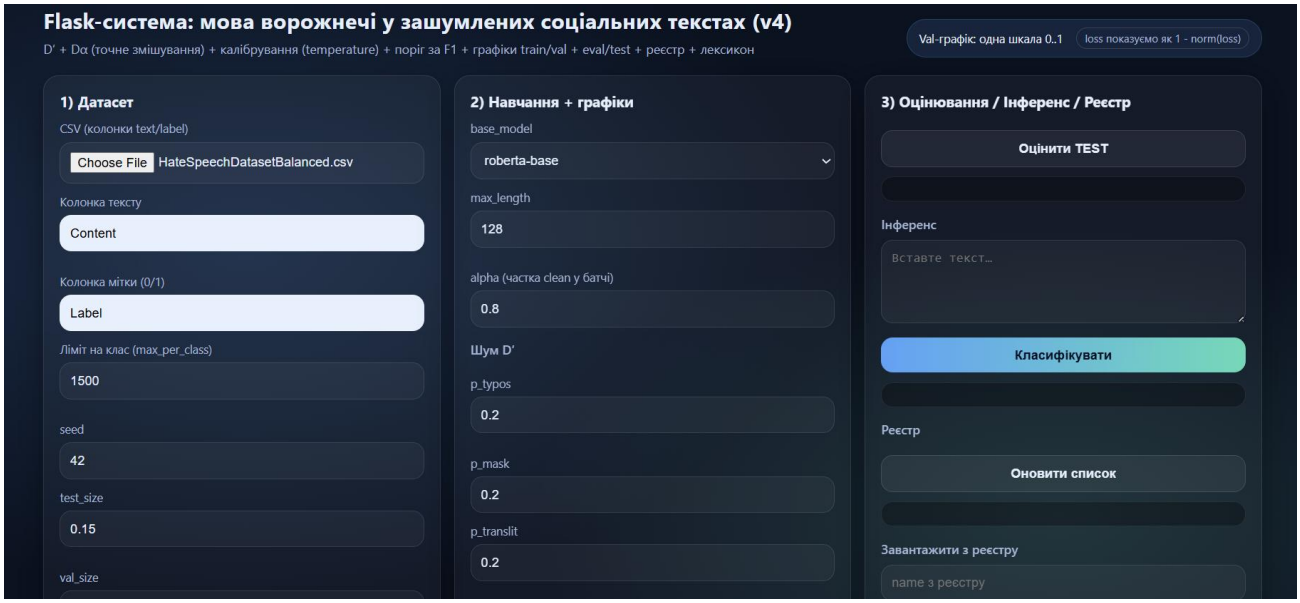


Рисунок 4.5 – Вигляд вікна після завантаження

У типовому сценарії використання робота починається з підготовки табличного набору даних у форматі CSV, який містить щонайменше дві колонки: колонку з текстом повідомлення та колонку з міткою класу у бінарній постановці (0 або 1). Далі користувач переходить до панелі «Датасет» і виконує такі дії: через кнопку вибору файлу завантажує CSV; у полях «Колонка тексту» та «Колонка мітки» задає назви відповідних колонок у файлі; встановлює параметр ліміту на клас (максимальна кількість прикладів на кожний клас, що дозволяє керувати обсягом експерименту) та значення зерна випадковості для відтворюваного відбору; задає частки тестової та валідаційної підвибірок. Після натискання кнопки «Завантажити, обмежити та розбити» система послідовно: зчитує дані, відсіює некоректні записи, за потреби застосовує обмеження кількості прикладів на клас, а далі формує стратифіковане розбиття на навчальну, валідаційну та тестову підвибірки. Результатом виконання є виведення протоколу (рисунок 4.6) із кількістю використаних повідомлень, розподілом за класами та розмірами підвибірок, а також

короткий попередній перегляд кількох записів для контролю коректності завантаження.

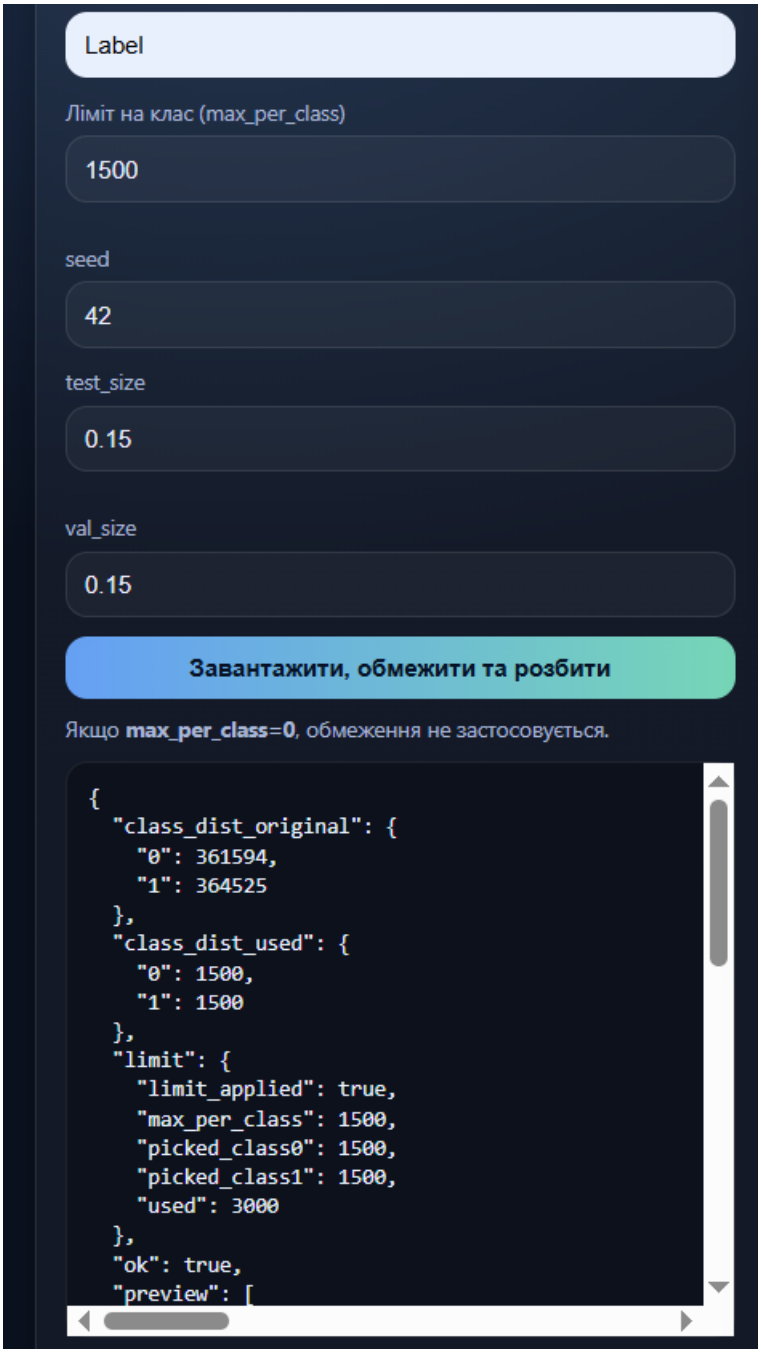


Рисунок 4.6 – Протокол завантаження датасету

За наявності потреби у моделюванні лексичних підмін користувач може додатково активувати завантаження словника підмін у форматі JSON (панель «Лексикон»). Для цього обирають файл та натискають кнопку «Завантажити лексикон». Після успішного завантаження система повідомляє кількість зчитаних

пар підмін, а в подальших кроках ці відповідності враховуються під час формування зашумлених варіантів повідомлень (рисунок 4.7). Якщо лексикон не завантажується, система працює коректно і без нього, обмежуючись іншими типами шумів.

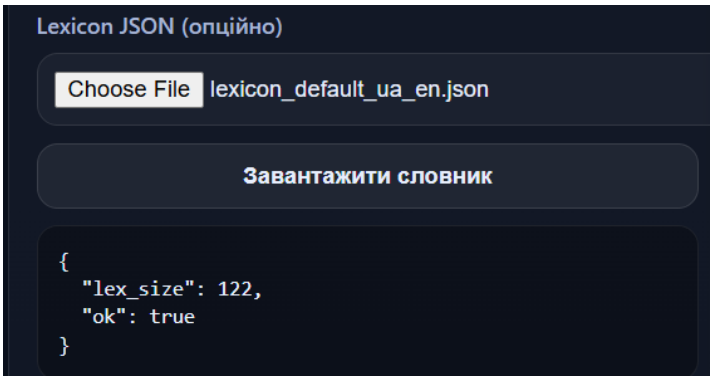


Рисунок 4.7 – Приклад завантаження лексикону

Після підготовки датасету користувач переходить до панелі «Навчання + графіки», де задаються параметри навчання та параметри моделювання шумів (рисунок 4.8).

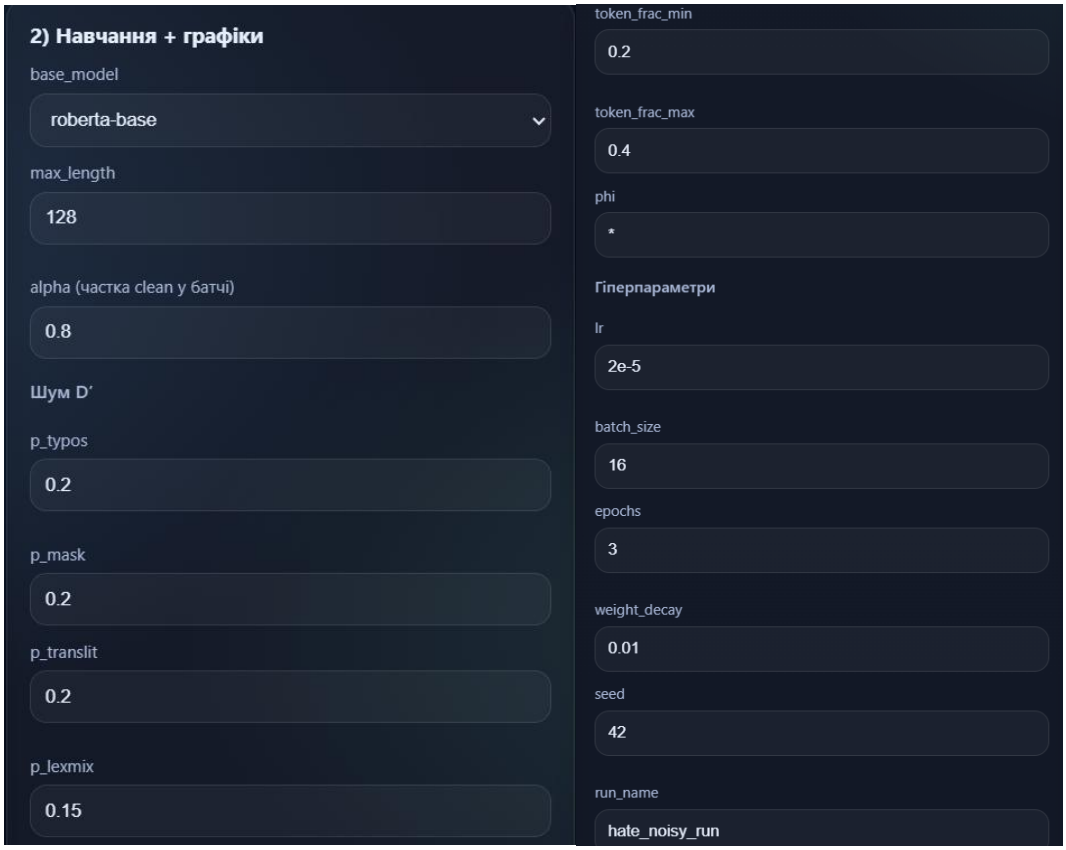


Рисунок 4.8 – Параметри навчання

На цьому етапі спочатку обирається базова трансформерна модель із наданого переліку, а також встановлюється максимальна довжина послідовності (обмеження довжини вхідного тексту для подачі у модель). Далі задається параметр частки чистих даних у мініпакеті, який визначає, яка частина прикладів у кожному мініпакеті залишається без змін, а решта піддається зашумленню. У блоці параметрів шуму користувач задає інтенсивності окремих механізмів спотворення (опечатки, маскування символів, графемні підміни, лексичні підміни), а також інтервал частки слів, які можуть модифікуватися у повідомленні. Окремо задається символ маскування, який використовується під час заміни частини символів у словах. У блоці гіперпараметрів задають швидкість навчання, розмір мініпакета, кількість епох, коефіцієнт регуляризації ваг, зерно випадковості та назву запуску, за якою результат буде ідентифіковано у реєстрі.

Запуск навчання здійснюється натисканням кнопки «Старт навчання». Після цього система ініціалізує модель і токенізатор, запускає цикл оптимізації та починає накопичувати історію навчання. У процесі користувач може контролювати перебіг двома способами. Перший спосіб натиснути кнопку «Оновити статус» (рисунок 4.9), після чого система виведе поточне повідомлення про стан (епоха, крок, поточні значення втрати, валідаційні показники) або повідомлення про помилку, якщо вона виникла.

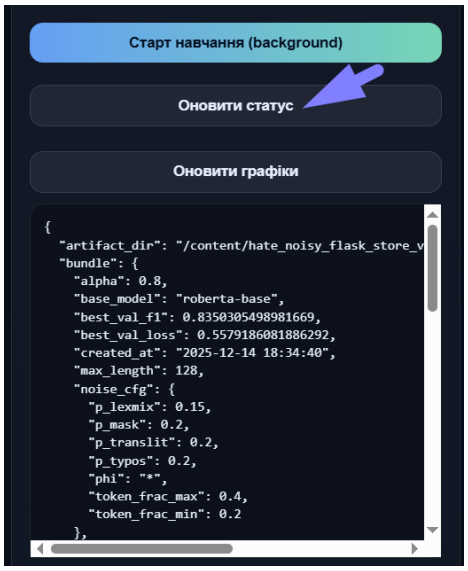


Рисунок 4.9 – Виведення інформації про процес навчання

Другий спосіб натиснути кнопку «Оновити графіки»; тоді система оновить візуалізації (рисунок 4.10).

Рисунок 4.10 – Виведення графіків навчання

Окрім ручного оновлення, передбачено автоматичне оновлення статусу та графіків під час навчання з регулярним інтервалом.

У системі формуються два ключові графіки. Перший графік відображає значення навчальної втрати залежно від номера кроку оптимізації; його призначення контроль збіжності та стабільності навчання (загальна тенденція до зменшення втрати інтерпретується як позитивна динаміка за умови відсутності різких неконтрольованих коливань). Другий графік призначений для аналізу якості на валідаційній підвибірці та побудований в єдиній шкалі від 0 до 1: на ньому одночасно відображається значення F_1-міри та перетворений показник валідаційної втрати, поданий у вигляді інвертованої нормалізованої величини (тобто менша валідаційна втрата відображається як більше значення на шкалі). Така уніфікація шкали забезпечує читабельність і дозволяє інтерпретувати обидві криві в єдиній логіці: «вище / краще». Важливо підкреслити, що перетворення втрати

застосовується лише для візуалізації та не підмінює числові метрики, які використовуються для вибору найкращого стану моделі.

Для якісного контролю того, як саме застосовуються спотворення, передбачено сценарій попереднього перегляду шумів. Користувач вводить довільне повідомлення у поле «Попередній перегляд шуму» та натискає кнопку «Згенерувати приклади» (рисунок 4.11).

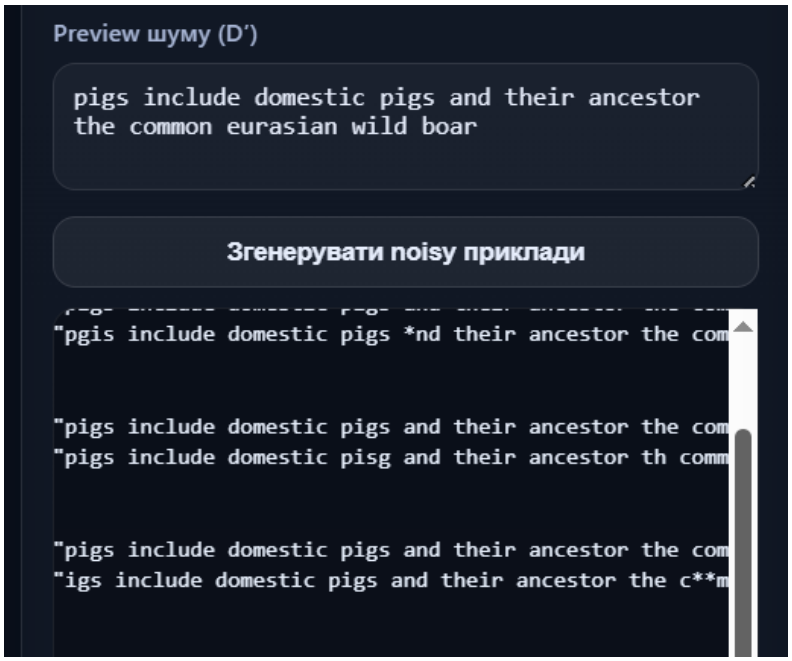


Рисунок 4.11 – Приклад зашумлення даних

У відповідь система формує кілька варіантів зашумленого повідомлення, що дозволяє ще до старту або під час навчання перевірити адекватність інтенсивності спотворень і, за потреби, скоригувати параметри шуму.

Після завершення навчання система зберігає модельні артефакти та паспорт експерименту до реєстру. На цьому етапі користувач переходить до панелі «Оцінювання / Інференс / Реєстр» і може виконати оцінювання на тестовій підвибірці. Для цього натискають кнопку «Оцінити тест»; система застосовує зафіксовані правила прийняття рішення (з урахуванням калібрування та порогу, визначеного за валідацією) і повертає підсумкові метрики (рисунок 4.12).

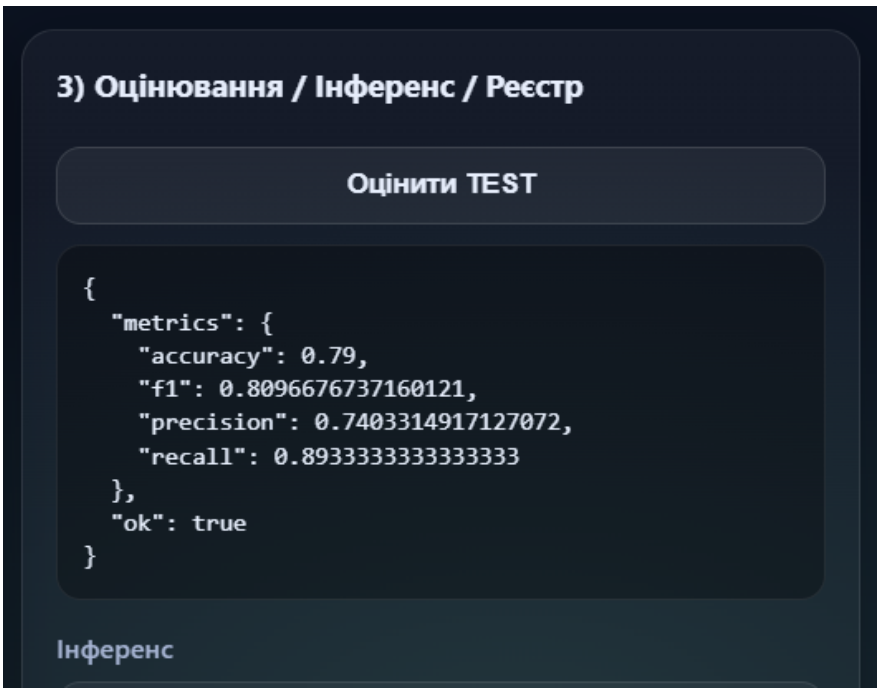


Рисунок 4.12 – Оцінка моделі

Далі доступний сценарій інференсу для одиничного повідомлення: користувач вводить текст у поле «Інференс» і натискає «Класифікувати». Система повертає прогнозований клас, а також числову оцінку впевненості та використаний поріг рішення, що забезпечує прозорість інтерпретації результату (рисунок 4.13).

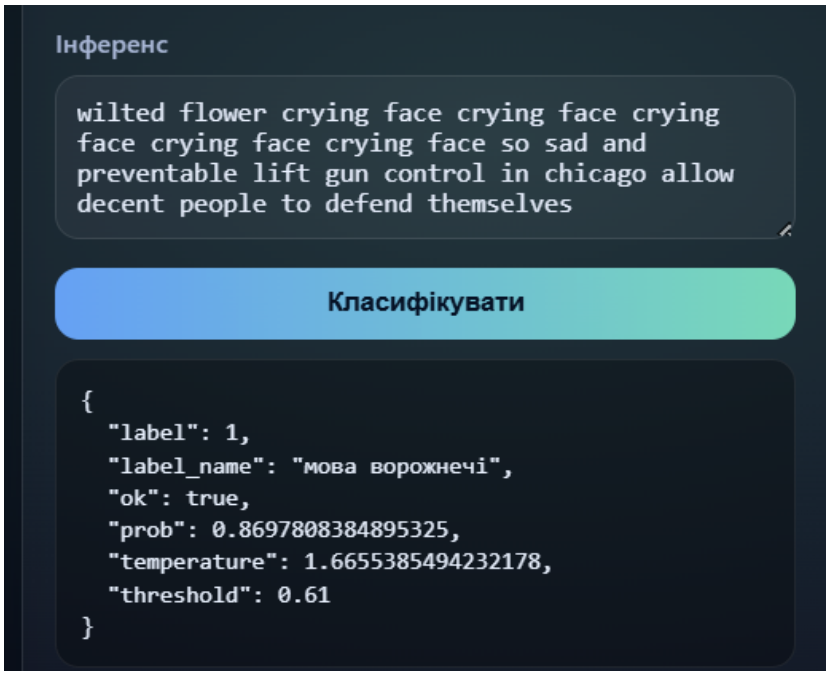


Рисунок 4.13 – Приклад аналізу повідомлення

Реєстр моделей підтримує два основні сценарії. Користувач натискає «Оновити список», після чого система виводить перелік збережених запусків із ключовими характеристиками (час створення, базова модель, частка чистих даних у мініпакеті, параметри калібрування, узагальнена якість на валідації). Приклад наведено на рисунку 4.14.

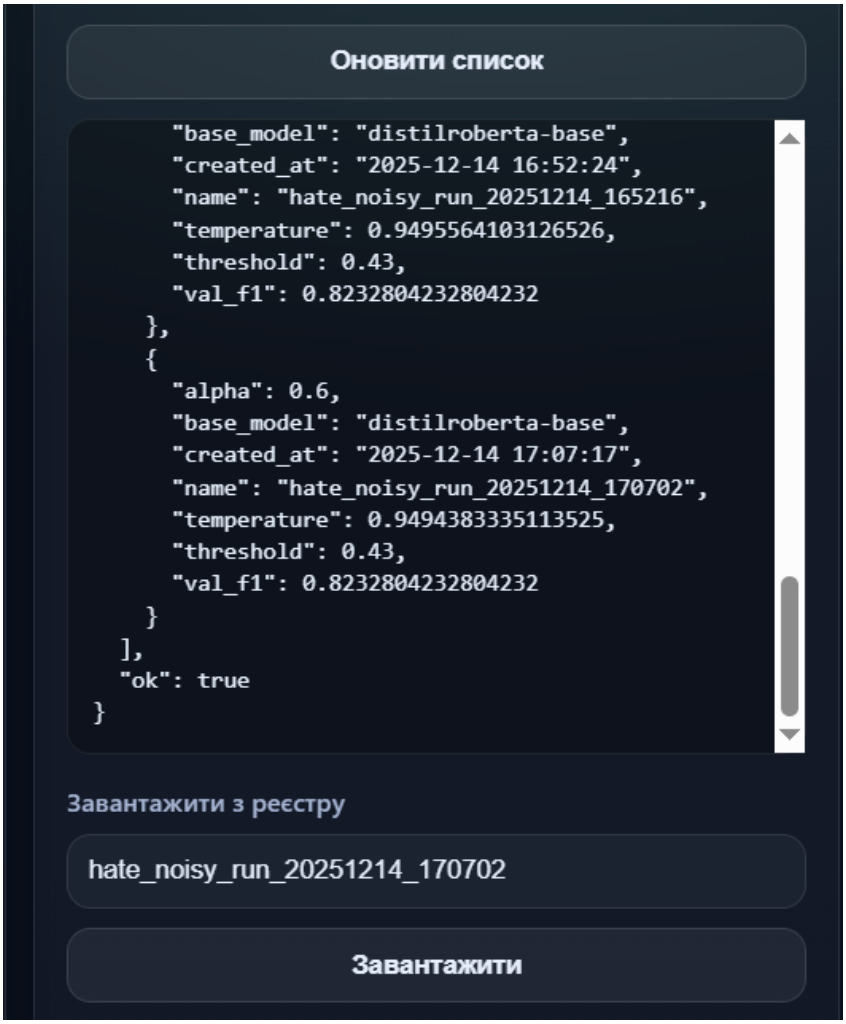


Рисунок 4.14 – Реєстр моделей

Користувач може завантажити раніше збережену модель: у полі «Назва з реєстру» вводиться ідентифікатор запису, після чого натискається кнопка «Завантажити» (рисунок 4.15). У відповідь система відновлює модель, токенізатор та паспорт експерименту, що дозволяє одразу виконувати оцінювання або інференс без повторного навчання.

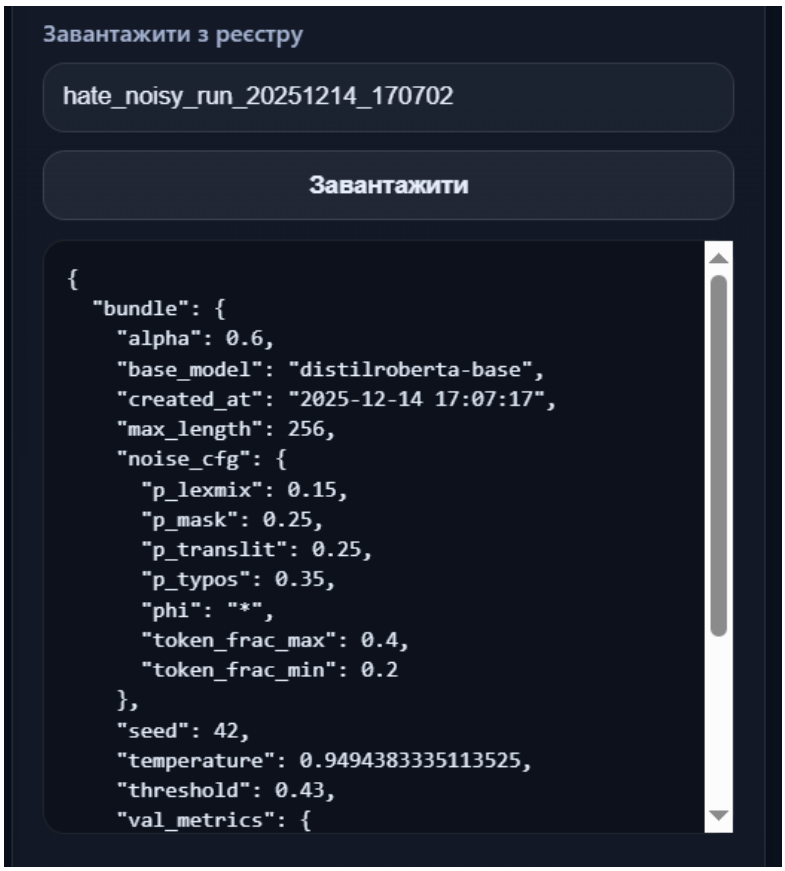


Рисунок 4.15 – Завантаження моделі з реєстру

Отже, наведено типовий порядок роботи з інтелектуальною системою та практичну логіку взаємодії користувача з її інтерфейсом. Показано, що система забезпечує повний цикл експерименту: підготовку й стратифіковане розбиття набору даних, налаштування параметрів моделювання шумів і навчання, моніторинг процесу за протоколом та графіками, попередній перегляд зашумлення, підсумкове оцінювання на тестовій підвибірці, інференс для окремих повідомлень і повторне використання результатів через реєстр моделей. Така організація сценаріїв використання підвищує відтворюваність експериментів і спрощує контроль якості на всіх етапах роботи.

4.4 Дослідження ефективності та інтерпретація отриманих результатів

Ефективність розробленого методу з використанням інтелектуальної системи оцінювалась у бінарній постановці виявлення мови ворожнечі за текстом

повідомлення. Дослід проводився на збалансованому наборі даних із контрольованим обмеженням кількості прикладів на клас та стратифікованим розбиттям на навчальну, валідаційну і тестову підвибірки. У процесі навчання застосовувалося кероване моделювання шумів у повідомленнях і кероване змішування «чистих» та «зашумлених» прикладів у мініпакетах. Після завершення оптимізації для найкращого стану моделі виконувалося калібрування оцінок та підбір порога рішення за критерієм максимальної F_1-міри на валідаційній підвибірці.

Таблиця 4.1 – Характеристики використаних даних та протокол обмеження

Показник	**Значення**
Загальна кількість використаних повідомлень	3000
Розподіл за класами після обмеження	1500 / 1500
Максимум прикладів на клас	1500
Частка тестової підвибірки	0,15
Частка валідаційної підвибірки	0,15
Зерно випадковості	42

Далі виконано серію запусків із різними базовими трансформерними архітектурами та різною часткою чистих прикладів у мініпакеті. Для коректного порівняння у реєстрі зберігаються ключові параметри запуску: архітектура, частка чистих прикладів, температура калібрування, підібраний поріг рішення та валідаційна F_1-міра.

Таблиця 4.2 – Порівняння запусків за валідаційною F1-мірою

Архітектура	**Частка чистих прикладів у мініпакеті**	**Температура калібрування**	**Поріг рішення**	**Валідаційна F_1-міра**	**Ідентифікатор запуску**
roberta-base	0,6	1,636	0,50	0,8339	hate_noisy_run_20251214_181658
distilroberta-base	0,6	1,661	0,37	0,8086	hate_noisy_run_20251214_182101

З таблиці 4.2 видно, що за однакової частки чистих прикладів у мініпакеті (0,6) архітектура roberta-base має вищу узагальнювальну якість на валідації порівняно з distilroberta-base. Це узгоджується з тим, що повнорозмірна модель краще відтворює контекстні залежності у складних і варіативних формулюваннях, притаманних соціальним повідомленням.

Окремо проаналізовано запуск для roberta-base із підвищеною часткою чистих прикладів у мініпакеті (0,8). Навчання проводилося до 4 епох, при цьому найкращий стан моделі відбирався за максимумом F_1-міри на валідації (зафіксовано на 3-й епосі).

Таблиця 4.3 – Параметри та метрики найкращої конфігурації roberta-base

Показник	**Значення**
Архітектура	roberta-base
Максимальна довжина повідомлення	128
Частка чистих прикладів у мініпакеті	0,8
Найкраща епоха за F1 на валідації	3
Середня навчальна втрата (епоха 3)	0,2963
Валідаційна втрата (епоха 3)	0,5579
Валідаційна точність	0,82
Валідаційна точність позитивного класу	0,7707
Валідаційна повнота позитивного класу	0,9111
Валідаційна F1-міра	0,8350
Температура калібрування	2,1447
Підібраний поріг рішення	0,46
Ідентифікатор запуску	hate_noisy_run_20251214_183428

Отриманий профіль метрик (висока повнота за помірної точності позитивного класу) інтерпретується як орієнтація моделі на мінімізацію пропусків повідомлень із мовою ворожнечі, що є практично виправданим для задачі фільтрації ризикового контенту. Калібрування температурою та підбір порога рішення

забезпечують відповідність ймовірнісних оцінок і стабільність правил класифікації під час оцінювання та інференсу.

Для візуального контролю процесу навчання система формує два ключові графіки. Перший графік відображає зміну навчальної втрати за кроками оптимізації; спадна тенденція при наявності стохастичних коливань відповідає типовій динаміці навчання на мініпакетах і підтверджує збіжність оптимізації (рисунок 4.16).

Рисунок 4.16 – Графік навчальної втрати за кроками оптимізації

Другий графік призначений для аналізу узагальнення на валідації та будується у спільній шкалі 0…1, що підвищує читабельність: одночасно відображається значення F_1-міри та інвертована нормалізована валідаційна втрата (чим менша втрата, тим більше значення на шкалі після перетворення) (рисунок 4.17).

Рисунок 4.17 – Графік валідації в одній шкалі: F1-міра та інвертована нормалізована валідаційна втрата

Ще однією перевіркою стало тестування розробленої методології на іншому датасеті. Проведене тестування показало, що всі досліджувані моделі, що навчались

на штучно зачумлених даних мали приріст точності щодо моделей, навчених на чистих даних датасету. Приріст сягав від 1,5 % до 5,5%.

Подані результати підтверджують, що розроблена інтелектуальна система на основі розробленого методу забезпечує відтворюване навчання та оцінювання трансформерних моделей у задачі виявлення мови ворожнечі за умов керованого зашумлення текстів. Порівняння запусків показало перевагу архітектури roberta-base за валідаційною F_1-мірою, а використання калібрування та підбору порога рішення підвищує узгодженість прогнозів і спрощує практичну інтерпретацію результатів. Візуалізація валідаційних показників в одній шкалі забезпечує кращу читабельність і дозволяє оперативно співставляти динаміку F_1-міри зі зміною валідаційної втрати.

Однак, наведена методологія має ряд обмежень. Отримані результати залежать від складу і репрезентативності використаного набору повідомлень: за умови збалансованого датасету та обмеження кількості прикладів на клас модель демонструє стабільні метрики, однак при зміні домену (інша платформа, тематика, мова, часовий період) можливе часткове зниження узагальнюваності. Додатковим чинником є чутливість якості до параметрів зашумлення та частки «чистих» прикладів у мініпакеті: надмірна інтенсивність спотворень або невдало підібрані ймовірності типів шуму можуть змінювати баланс між точністю і повнотою, що потребує підбору конфігурації під конкретний сценарій застосування. Також на відтворюваність і швидкість експериментів впливають обчислювальні ресурси середовища виконання, оскільки навчання трансформерних моделей є ресурсомістким і може вимагати компромісів щодо довжини послідовності, розміру мініпакета та кількості епох.

Подальші дослідження доцільно спрямувати на розширення експериментальної бази за рахунок мультидоменної і мультимовної валідації, а також на систематичний аналіз внеску окремих типів шумів (абляційні дослідження) з метою формування рекомендованих профілів зашумлення для різних умов даних. Перспективним є розвиток модуля інтерпретації рішень (пояснення найбільш впливових токенів/фрагментів), а також дослідження більш стійких схем оптимізації і добору порога, зокрема з урахуванням різної вартості помилок для практичних

сценаріїв модерації. Окремий напрям становить порівняння із сучаснішими архітектурами та підходами до робастного навчання (контрастивні або змагальні режими), а також перенесення навченої моделі на потоки даних у режимі наближеному до реального часу.

Висновки до розділу 4

У розділі виконано експериментальне дослідження методу виявлення мови ворожнечі в умовах керованого зашумлення соціальних текстових даних та обґрунтовано його прикладну реалізацію у складі інтелектуальної системи. Сформовано модульну програмну структуру компонентів із логічним розподілом відповідальності між завантаженням і розбиттям даних, генерацією шумів, підготовкою тексту, навчанням і оцінюванням трансформерної моделі, калібруванням імовірнісних виходів та сервісним інференсом, що мінімізує розрив між експериментальною частиною та практичним використанням системи.

Алгоритмічну основу системи зафіксовано у вигляді узгодженої специфікації ключових процедур, які формалізують оброблення датасету, контрольоване моделювання спотворень, точне змішування чистих і зашумлених прикладів у мініпакетах, навчання з валідаційним контролем, калібрування оцінок і підбір порога рішення, а також ведення реєстру експериментальних артефактів. Така формалізація забезпечує відтворюваність експериментів, керованість параметрів та однозначність інтерпретації результатів незалежно від конкретного інтерфейсу реалізації.

Показано практичну логіку використання системи у форматі вебзастосунку: реалізовано повний цикл роботи користувача від завантаження та стратифікованого розбиття даних до налаштування параметрів шуму і навчання, моніторингу процесу за протоколом і графіками, попереднього перегляду зашумлення, підсумкового оцінювання на тестовій підвибірці, інференсу одиничних повідомлень та повторного використання моделей через реєстр.

За результатами експериментів на збалансованому наборі даних із контрольованим обмеженням кількості прикладів на клас продемонстровано конкурентну якість класифікації для трансформерних архітектур та обґрунтовано перевагу roberta-base порівняно з distilroberta-base за валідаційною F1-мірою в однакових умовах змішування чистих і зашумлених прикладів. Для найкращої конфігурації зафіксовано профіль метрик із підвищеною повнотою позитивного класу за помірної точності, що є практично доцільним для задачі фільтрації ризикового контенту, де критичним є мінімізація пропусків повідомлень із ознаками мови ворожнечі.

Додатково підтверджено корисність калібрування та підбору порога рішення за валідаційними даними: за рахунок узгодження імовірнісних оцінок і фіксації правила прийняття рішення підвищується стабільність результатів під час тестування та інференсу. Використання графічного контролю (динаміка навчальної втрати та спільна шкала для F_1-міри і перетвореної валідаційної втрати) забезпечує інтерпретоване відстеження збіжності оптимізації та узагальнювальної здатності моделі, що підвищує надійність вибору найкращого стану.

Отже, розділ підтверджує, що розроблений метод і його програмна реалізація забезпечують кероване робастне навчання трансформерних моделей у задачі виявлення мови ворожнечі в умовах зашумлення текстів, а також створюють відтворювану експериментальну базу для порівняння архітектур і конфігурацій зашумлення з можливістю практичного використання через інтерфейс інтелектуальної системи.

Загальні висновки

Метою кваліфікаційної роботи було підвищення точності виявлення мови ворожнечі у зашумлених соціальних текстових даних шляхом розроблення нейромережевого методу, здатного інтерпретувати спотворені, змішаномовні та навмисно масковані мовні конструкції. Розроблений метод дозволив підвищити точність на даних іншого датасету щонайменше на 1,5 % в порівнянні з нейромережею, навченою на чистих даних. На відміну від більшості існуючих підходів, зорієнтованих на попередньо очищені або стандартизовані текстові вибірки, запропонований метод передбачає відтворення керованих шумових спотворень і цілеспрямоване навчання нейромережевої моделі в умовах мовної нестабільності, маскування агресивної лексики та змішаності мовних кодів. Це дає змогу зберігати коректність класифікації за наявності орфографічних, графічних і семантичних викривлень, а також підвищує стійкість моделі до навмисного приховування мовленнєвої агресії, що зазвичай недостатньо враховано у відомих рішеннях. Для досягнення мети були сформульовані та вирішити такі *задачі*:

– проведено аналіз природи мови ворожнечі та її класифікаційних ознак;

– виконано огляд існуючих підходів до виявлення мови ворожнечі, виконано аналіз наукових досліджень;

– охарактеризовано етичні аспекти автоматизованого виявлення мови ворожнечі;

– розроблено метод виявлення мови ворожнечі у зашумлених соціальних текстових даних нейромережевими засобами;

– виконано підготовку датасету для файнтюнінгу нейромережі для виявлення мови ворожнечі;

– виконано програмну реалізацію розробленого методу;

– проведено дослідження методу виявлення мови ворожнечі у зашумлених соціальних текстових даних нейромережевими засобами.

Розроблено інтелектуальну систему у форматі вебзастосунку, що забезпечує повний цикл експерименту з виявлення мови ворожнечі у соціальних текстах: завантаження та валідацію табличних даних, стратифіковане розбиття, кероване моделювання шумів і контрольоване змішування чистих та зашумлених прикладів

під час навчання трансформерної моделі, моніторинг навчання за протоколом і графіками, підбір порога рішення та калібрування оцінок, фінальне оцінювання на тестовій підвибірці, інференс для одиничних повідомлень і повторне використання результатів через реєстр моделей. Практично значущим є те, що система дозволяє швидко порівнювати конфігурації шумів і базові архітектури, фіксувати параметри запусків та метрики, а також переносити навчений класифікатор у прикладні сценарії фільтрації ризикового контенту й підтримки модерації, де реальні тексти часто містять викривлення.

Обмеження. Результати залежать від репрезентативності використаного корпусу та відповідності типів штучно змодельованих спотворень тим, що трапляються в реальному середовищі: за появи нових стратегій обфускації або доменно-специфічної лексики якість може змінюватися. Додатково на узагальнювальну здатність впливають налаштування частки чистих прикладів у мініпакеті, інтенсивності шумів і вибір порога рішення, оскільки різні прикладні задачі можуть вимагати іншого балансу між пропусками й хибними спрацюваннями. Обмеженням також є бінарна постановка, яка не розрізняє типи мови ворожнечі, цільові групи та контекст, що може бути суттєвим у практиці модерації.

Перспективи подальших досліджень. Доцільним є розширення постановки до багатокласової або багатоміткової класифікації з виділенням типів агресії та цільових категорій, а також поглиблення моделювання шумів шляхом додавання контекстних та семантичних перетворень і автоматизованого підбору інтенсивностей шумів під конкретний домен. Перспективним напрямом є оцінювання робастності на різнорідних наборах даних і в режимі доменної адаптації, а також інтеграція механізмів інтерпретованості для підвищення прозорості результатів у модераційних сценаріях.

Основні наукові й практичні результати роботи доповідались у доповіді «Підхід до нейромережевого виявлення мови ворожнечі у зашумлених тексових повідомленнях» [69] на XVII Всеукраїнській науково-практичній конференції «Актуальні проблеми комп’ютерних наук АПКН-2025» (м.Хмельницький) 14-15 листопада 2025 року. Також за темою кваліфікаційної роботи підготовлено до публікації статтю у фаховому виданні категорії Б.

Перелік посилань

1. Exploring Textual Hate Speech Detection Methods and Datasets: A Comprehensive Literature Review / R. Bisoi et al. *2024 International Conference on Intelligent Computing and Emerging Communication Technologies (ICEC)*, Guntur, India, 23–25 November 2024. 2024. P. 1–6. URL: https://doi.org/10.1109/icec59683.2024.10837448 (date of access: 17.10.2025).

2. Hate Speech Detection using Large Language Models: A Comprehensive Review / A. Albladi et al. *IEEE Access*. 2025. P. 1. URL: https://doi.org/10.1109/access.2025.3532397 (date of access: 17.10.2025).

3. Recommendation No. R (97) 20 of the Committee of Ministers to Member States on "Hate Speech". *Council of Europe*, 30 October 1997. URL: https://rm.coe.int/1680505d5b (date of access: 17.10.2025).

4. What is hate speech? *United Nations.* URL: [https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech (date of access: 17.10.2025).

5. Що таке мова ворожнечі? – Путівник з прав людини. *Путівник з прав людини*. URL: https://www.rights.in.ua/themes/zlochini-na-runt-njenavist-ta-mova-vorozhnjech/hate-speech/what-is-hate-speech (дата звернення: 17.10.2025).

6. Головна - Комісія з журналістської етики. *Комісія з журналістської етики*. URL: https://cje.org.ua/ (дата звернення: 17.10.2025).

7. Про медіа. *Офіційний вебпортал парламенту України*. URL: https://zakon.rada.gov.ua/laws/show/2849-IX (дата звернення: 17.10.2025).

8. Hietanen M., Eddebo J. Towards a Definition of Hate Speech–With a Focus on Online Contexts. *Journal of Communication Inquiry*. 2022. P. 019685992211243. URL: https://doi.org/10.1177/01968599221124309 (date of access: 17.10.2025).

9. Hate Speech Detection Using Machine Learning / S. Futane et al. *International Journal for Research in Applied Science and Engineering Technology*.

2023. Vol. 11, no. 4. P. 1186–1188. URL: https://doi.org/10.22214/ijraset.2023.50265 (date of access: 17.10.2025).

10. European Commission against Racism and Intolerance (ECRI). *COE*. URL: https://www.coe.int/en/web/european-commission-against-racism-and-intolerance (date of access: 17.10.2025).

11. European Union Agency for Fundamental Rights. *European Union Agency for Fundamental Rights*. URL: https://fra.europa.eu/en (date of access: 17.10.2025).

12. Junod M. Guidelines, All in One Place: Website Review: ECRI Guidelines Trust. *Journal of Electronic Resources in Medical Libraries*. 2024. P. 1–7. URL: https://doi.org/10.1080/15424065.2024.2317882 (date of access: 17.10.2025).

13. Chhabra A., Vishwakarma D. K. A literature survey on multimodal and multilingual automatic hate speech identification. *Multimedia Systems*. 2023. URL: https://doi.org/10.1007/s00530-023-01051-8 (date of access: 17.10.2025).

14. Meske C., Bunde E. Design Principles for User Interfaces in AI-Based Decision Support Systems: The Case of Explainable Hate Speech Detection. *Information Systems Frontiers*. 2022. URL: https://doi.org/10.1007/s10796-021-10234-5 (date of access: 17.10.2025).

15. Cyberbullying-related Hate Speech Detection Using Shallow-to-deep Learning / D. Sultan et al. *Computers, Materials & Continua*. 2023. Vol. 74, no. 1. P. 2115–2131. URL: https://doi.org/10.32604/cmc.2023.032993 (date of access: 17.10.2025).

16. Peršak N. Criminalising Hate Crime and Hate Speech at EU Level: Extending the List of Eurocrimes Under Article 83(1) TFEU. *Criminal Law Forum*. 2022. URL: https://doi.org/10.1007/s10609-022-09440-w (date of access: 17.10.2025).

17. Textual Feature Extraction Using Ant Colony Optimization for Hate Speech Classification / S. Gite et al. *Big Data and Cognitive Computing*. 2023. Vol. 7, no. 1. P. 45. URL: https://doi.org/10.3390/bdcc7010045 (date of access: 17.10.2025).

18. Hate speech and real harm. *United Nations.* URL: https://www.un.org/en/hate-speech/understanding-hate-speech/hate-speech-and-real-harm (date of access: 17.10.2025).

19. Towards safer online communities: Deep learning and explainable AI for hate speech detection and classification / H. Kibriya et al. *Computers and Electrical Engineering*. 2024. Vol. 116. P. 109153. URL: https://doi.org/10.1016/j.compeleceng.2024.109153 (date of access: 17.10.2025).

20. United Nations Strategy and Plan of Action on Hate Speech. *GCED Clearinghouse*. URL: https://www.gcedclearinghouse.org/resources/united-nations-strategy-and-plan-action-hate-speech?language=en (date of access: 17.10.2025).

21. Izquierdo Montero A., Laforgue-Bullido N., Abril-Hervás D. Hate speech: a systematic review of scientific production and educational considerations. *Revista Fuentes*. 2022. Vol. 2, no. 24. P. 222–233. URL: https://doi.org/10.12795/revistafuentes.2022.20240 (date of access: 17.10.2025).

22. Hate Speech Detection in Twitter: Natural Language Processing Exploration / K. Egode et al. *Global Advanced Research Journal of Educational Research and Reviews*. 2023. Vol. 11, no. 8. P. 325–336. URL: https://doi.org/10.5281/zenodo.11178221 (date of access: 17.10.2025).

23. dictNN: A Dictionary-Enhanced CNN Approach for Classifying Hate Speech on Twitter / M. *Kupi, M. Bodnar, N. Schmidt, C. E. Posada*. *arXiv* preprint arXiv:2103.08780, 2021. URL: https://doi.org/10.48550/arXiv.2103.08780 (date of access: 17.10.2025).

24. Hate speech policy. *YouTube Help*. URL: https://support.google.com/youtube/answer/2801939 (date of access: 17.10.2025).

25. Hateful Conduct. Meta. URL: https://transparency.meta.com/en-gb/policies/community-standards/hateful-conduct/ (date of access: 17.10.2025).

26. Magu R., Luo J. Determining Code Words in Euphemistic Hate Speech Using Word Embedding Networks. *Proceedings of the 2nd Workshop on Abusive Language*

Online (ALW2), Brussels, Belgium. Stroudsburg, PA, USA, 2018. URL: https://doi.org/10.18653/v1/w18-5112 (date of access: 17.10.2025).

27. From Bag-of-Words to Transformers: A Comparative Study for Text Classification in Healthcare Discussions in Social Media / E. De Santis et al. *IEEE Transactions on Emerging Topics in Computational Intelligence*. 2024. P. 1–15. URL: https://doi.org/10.1109/tetci.2024.3423444 (date of access: 17.10.2025).

28. Jain S., Jain S. K., Vasal S. An Effective TF-IDF Model to Improve the Text Classification Performance. *2024 IEEE 13th International Conference on Communication Systems and Network Technologies (CSNT)*, Jabalpur, India, 6–7 April 2024. 2024. P. 1–4. URL: https://doi.org/10.1109/csnt60213.2024.10545818 (date of access: 17.10.2025).

29. Xiang R. F. Use of n-grams and K-means clustering to classify data from free text bone marrow reports. *Journal of Pathology Informatics*. 2024. P. 100358. URL: https://doi.org/10.1016/j.jpi.2023.100358 (date of access: 17.10.2025).

30. Classification of User's Review Using Modified Logistic Regression Technique / R. Reddy, U. A. Kumar. *International Journal of System Assurance Engineering and Management*, 15(1), 2024. P. 279–286. URL: https://doi.org/10.1007/s13198-022-01711-4 (date of access: 17.10.2025).

31. Email spam detection: a comparison of svm and naive bayes using bayesian optimization and grid search parameters / D. Budiman et al. *Journal of Student Research Exploration*. 2024. Vol. 2, no. 1. P. 53–64. URL: https://doi.org/10.52465/josre.v2i1.260 (date of access: 17.10.2025).

32. Automated Hate Speech Detection and the Problem of Offensive Language / T. Davidson et al. *Proceedings of the International AAAI Conference on Web and Social Media*. 2017. Vol. 11, no. 1. P. 512–515. URL: https://doi.org/10.1609/icwsm.v11i1.14955 (date of access: 17.10.2025).

33. Waseem Z., Hovy D. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. *Proceedings of the NAACL Student Research*

Workshop, San Diego, California. Stroudsburg, PA, USA, 2016. URL: https://doi.org/10.18653/v1/n16-2013 (date of access: 17.10.2025).

34. Rawat A., Kumar S., Samant S. S. Hate speech detection in social media: Techniques, recent trends, and future challenges. *WIREs Computational Statistics*. 2024. Vol. 16, no. 2. URL: https://doi.org/10.1002/wics.1648 (date of access: 17.10.2025).

35. Li J., Yang Y., Sun J., Wang F., Chen S. DT-GCNN: Dynamic Triplet Network with GRU-CNN for Enhanced Text Classification. *International Journal of Machine Learning and Cybernetics*. 2025. P. 1–13. URL: https://doi.org/10.1007/s13042-025-02769-9 (date of access: 17.10.2025).

36. Effective Text Classification using BERT, MTM LSTM, and DT / S. Jamshidi et al. *Data & Knowledge Engineering*. 2024. P. 102306. URL: https://doi.org/10.1016/j.datak.2024.102306 (date of access: 17.10.2025).

37. An Improved Hybrid GRU and CNN Models for News Text Classification / I. Y. Khudhair et al. *JOIV : International Journal on Informatics Visualization*. 2025. Vol. 9, no. 1. P. 303. URL: https://doi.org/10.62527/joiv.9.1.2658 (date of access: 17.10.2025).

38. Eriguchi A., Tsuruoka Y., Cho K. Learning to Parse and Translate Improves Neural Machine Translation. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vancouver, Canada. Stroudsburg, PA, USA, 2017. URL: https://doi.org/10.18653/v1/p17-2012 (date of access: 17.10.2025).

39. Hate Speech Dataset from a White Supremacy Forum / O. de Gibert et al. *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, Brussels, Belgium. Stroudsburg, PA, USA, 2018. URL: https://doi.org/10.18653/v1/w18-5102 (date of access: 17.10.2025).

40. Schmid U. K., Kümpel A. S., Rieger D. Social Media Users' Motives for (Not) Engaging With Hate Speech: An Explorative Investigation. *Social Media + Society*.

2024. Vol. 10, no. 4. URL: https://doi.org/10.1177/20563051241306322 (date of access: 17.10.2025).

41. Li X., Jia L. English text topic classification using BERT-based model. *Journal of Computational Methods in Sciences and Engineering*. 2025. URL: https://doi.org/10.1177/14727978251321982 (date of access: 17.10.2025).

42. Hazim L. R., Ata O. Textual Authenticity in the AI Era: Evaluating BERT and RoBERTa with Logistic Regression and Neural Networks for Text Classification. *2024 International Symposium on Electronics and Telecommunications (ISETC)*, Timisoara, Romania, 7–8 November 2024. 2024. P. 1–6. URL: https://doi.org/10.1109/isetc63109.2024.10797291 (date of access: 17.10.2025).

43. G B. M., Sampreetha V., Meghana K. U. Sentiment Analysis for Low-Resource Languages Using a Hybrid XLM-Roberta and Bi-LSTM Model. *2025 International Conference on Networks & Advances in Computational Technologies (NetACT)*, Trivandrum, India, 7–9 August 2025. 2025. P. 1–6. URL: https://doi.org/10.1109/netact65906.2025.11188922 (date of access: 17.10.2025).

44. Transfer learning for hate speech detection in social media / L. Yuan et al. *Journal of Computational Social Science*. 2023. URL: https://doi.org/10.1007/s42001-023-00224-9 (date of access: 17.10.2025).

45. Du J., Jiang Y., Liang Y. Transformers in Opinion Mining: Addressing Semantic Complexity and Model Challenges in NLP. *Transactions on Computational and Scientific Methods*. 2024. Vol. 4, no. 10. URL: https://doi.org/10.5281/zenodo.14058500 (date of access: 17.10.2025).

46. Livingston S., Bahador B. Propaganda feedback loops as communication rituals: Hate speech on talk radio. *Media, Culture & Society*. 2025. URL: https://doi.org/10.1177/01634437251331809 (date of access: 17.10.2025).

47. Mazepa S. War, Hate, Propaganda and the Internet: A Dangerous Combination. Cham : *Springer Nature Switzerland*, 2024. URL: https://doi.org/10.1007/978-3-031-69008-2 (date of access: 17.10.2025).

48. Revisiting Single-Step Adversarial Training for Robustness and Generalization / Z. Li et al. *SSRN Electronic Journal*. 2023. URL: https://doi.org/10.2139/ssrn.4377055 (date of access: 17.10.2025).

49. Untangling Hate Speech Definitions: A Semantic Componential Analysis Across Cultures and Domains / K. Korre et al. *Findings of the Association for Computational Linguistics: NAACL 2025*, Albuquerque, New Mexico. Stroudsburg, PA, USA, 2025. P. 3184–3198. URL: https://doi.org/10.18653/v1/2025.findings-naacl.175 (date of access: 17.10.2025).

50. Perspective API. *Perspective API*. URL: https://perspectiveapi.com (date of access: 17.10.2025).

51. Tunison S. Content Analysis. *Springer Texts in Education*. Cham, 2023. P. 85–90. URL: https://doi.org/10.1007/978-3-031-04394-9_14 (date of access: 17.10.2025).

52. Pukallus S., Arthur C. Combating Hate Speech on Social Media: Applying Targeted Regulation, Developing Civil-Communicative Skills and Utilising Local Evidence-Based Anti-Hate Speech Interventions. *Journalism and Media*. 2024. Vol. 5, no. 2. P. 467–484. URL: https://doi.org/10.3390/journalmedia5020031 (date of access: 17.10.2025).

53. White J. Advancing Ethical and Accurate Hate Speech Detection with Machine Learning Techniques. *International Journal of Scientific Research and Engineering Trends*. 2024. Vol. 10, no. 2. P. 99–104. URL: https://doi.org/10.61137/ijsret.vol.10.issue2.135 (date of access: 17.10.2025).

54. Rodríguez-Peral E. M., Gómez Franco T., Rodríguez-Peral Bustos D. Propagation of Hate Speech on Social Network X: Trends and Approaches. *Social Inclusion*. 2025. Vol. 13. URL: https://doi.org/10.17645/si.9317 (date of access: 17.10.2025).

55. UA-HSD-2025: Multi-Lingual Hate Speech Detection from Tweets Using Pre-Trained Transformers / M. Ahmad et al. *Computers*. 2025. Vol. 14, no. 6. P. 239. URL: https://doi.org/10.3390/computers14060239 (date of access: 17.10.2025).

56. Kodali R. G., Manukonda D. P., Iglesias D. byteSizedLLM@ NLU of Devanagari Script Languages 2025: Hate Speech Detection and Target Identification Using Customized Attention BiLSTM and XLM-RoBERTa Base Embeddings. In: *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*. 2025. P. 242–247. URL: https://aclanthology.org/2025.chipsal-1.25/ (date of access: 17.10.2025).

57. Stop the Hate, Spread the Hope: An Ensemble Model for Hope Speech Detection in English and Dravidian Languages / D. Sharma et al. *ACM Transactions on Asian and Low-Resource Language Information Processing*. 2025. URL: https://doi.org/10.1145/3716383 (date of access: 17.10.2025).

58. Mane S., Kundu S., Sharma R. A Survey on Online Aggression: Content Detection and Behavioural Analysis on Social Media Platforms. *ACM Computing Surveys*. 2025. URL: https://doi.org/10.1145/3711125 (date of access: 17.10.2025).

59. Online and Offline Aggressive Behaviors in Adolescence: The Role of Self-Regulatory Self-Efficacy Beliefs / A. Favini et al. *Behavioral Sciences*. 2024. Vol. 14, no. 9. P. 776. URL: https://doi.org/10.3390/bs14090776 (date of access: 17.10.2025).

60. Trabelsi Z. Mitigating Digital Misconduct: Content Moderation and Antisocial Behavior in Online Communities. 2025. URL: https://corpus.ulaval.ca/entities/publication/6954faf6-3b71-4feb-9d58-da24e7902690 (date of access: 17.10.2025).

61. Abdullina L. R., Ageeva A. V., Sabirova D. R. Investigating Verbal Aggression as a Type of Communication Behavior of English Speakers in the Internet Space. *Journal of Research in Applied Linguistics*. 2023. Vol. 14, no. 3. P. 300–304. URL: https://doi.org/10.22055/rals.2023.19555 (date of access: 17.10.2025).

62. OSCE Representative on Freedom of the Media. *Organization for Security and Co-operation in Europe | OSCE*. URL: https://www.osce.org/representative-on-freedom-of-media (date of access: 17.10.2025).

63. Countering Hate Speech. *UNESCO*. URL: https://www.unesco.org/en/countering-hate-speech (date of access: 17.10.2025).

64. Hate Speech Detection Curated Dataset. Waalbannyantudre. *Kaggle*. URL: https://www.kaggle.com/datasets/waalbannyantudre/hate-speech-detection-curated-dataset (date of access: 17.10.2025).

65. Hate Speech and Offensive Language Detection. TheDevastator. *Kaggle*. URL: https://www.kaggle.com/datasets/thedevastator/hate-speech-and-offensive-language-detection (date of access: 17.10.2025).

66. Google Colaboratory. *Google Colab.* URL: https://colab.research.google.com/ (date of access: 17.10.2025).

67. Evaluation Metrics in Machine Learning. *GeeksforGeeks*. URL: https://www.geeksforgeeks.org/machine-learning/metrics-for-machine-learning-model/ (date of access: 17.10.2025).

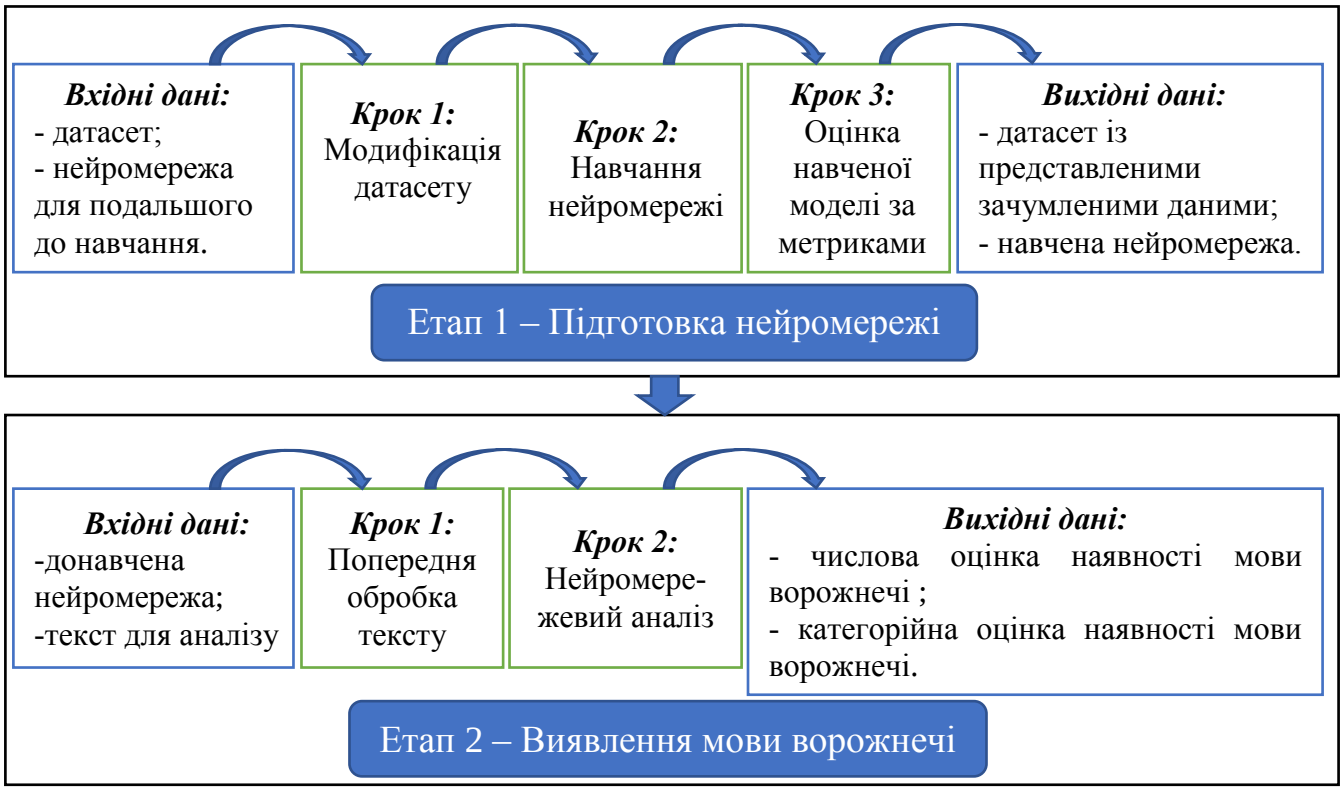
68. Bodner B. Unlocking Success: The 5 Essential Metrics You Must Track in Neural Network Training. *Medium*. URL: https://medium.com/@benjybo7/unlocking-success-the-5-essential-metrics-you-must-track-in-neural-network-training-52dcb8874ff0 (date of access: 17.10.2025).

69. Підхід до нейромережевого виявлення мови ворожнечі у зашумлених тексових повідомленнях / Боярчук І.О., Молчанова М.О. // *Актуальні проблеми комп'ютерних наук : зб. наук. пр. за матеріалами XVII Всеукр. наук.-практ. конф. (АПКН-2025).* – Хмельницький, 14–15 листоп. 2025 р. – Хмельницький, 2025. – С. 46–51.

ДОДАТКИ

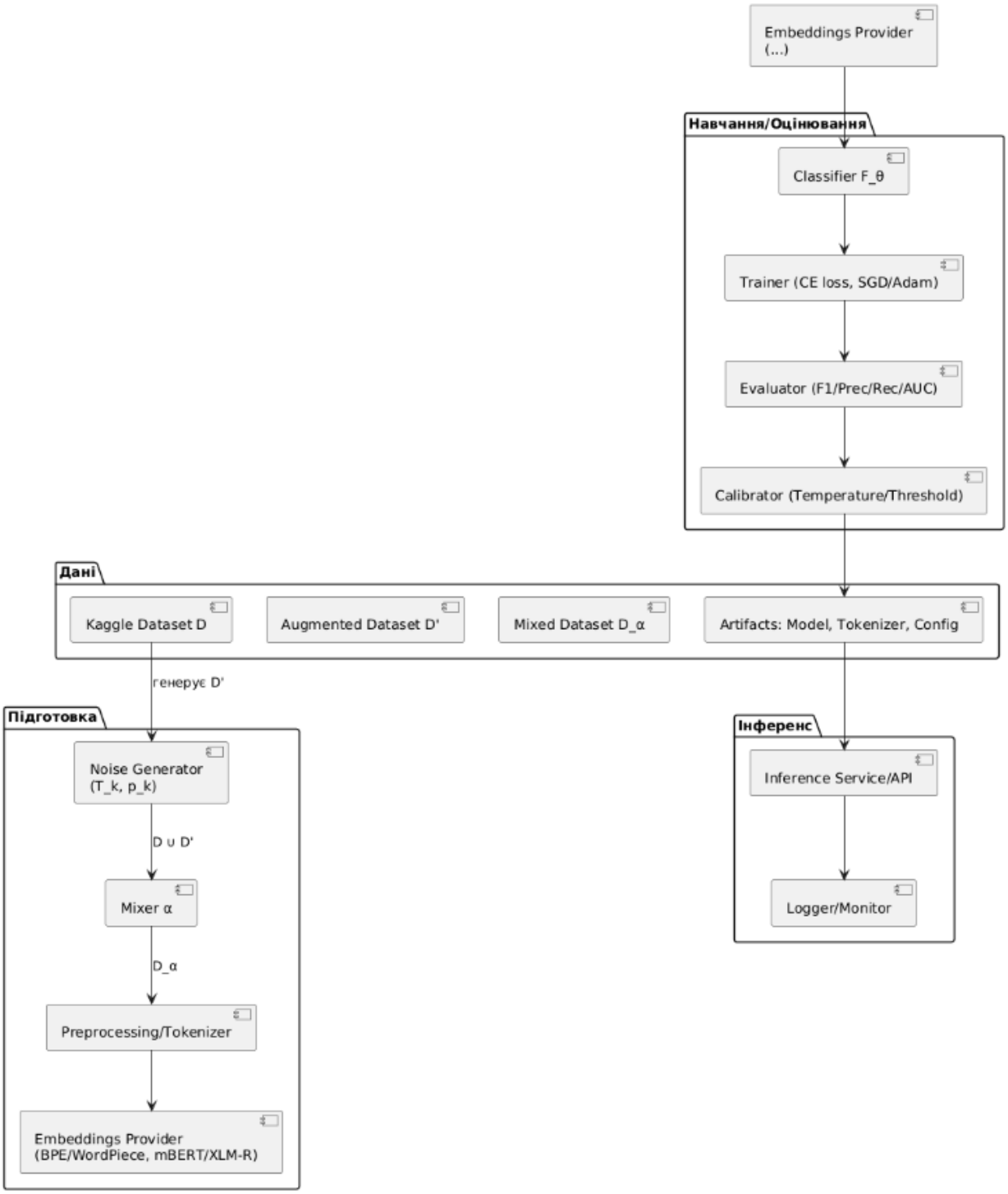
Додаток А

Етапи та кроки методу виявлення мови ворожнечі у зашумлених соціальних текстових даних



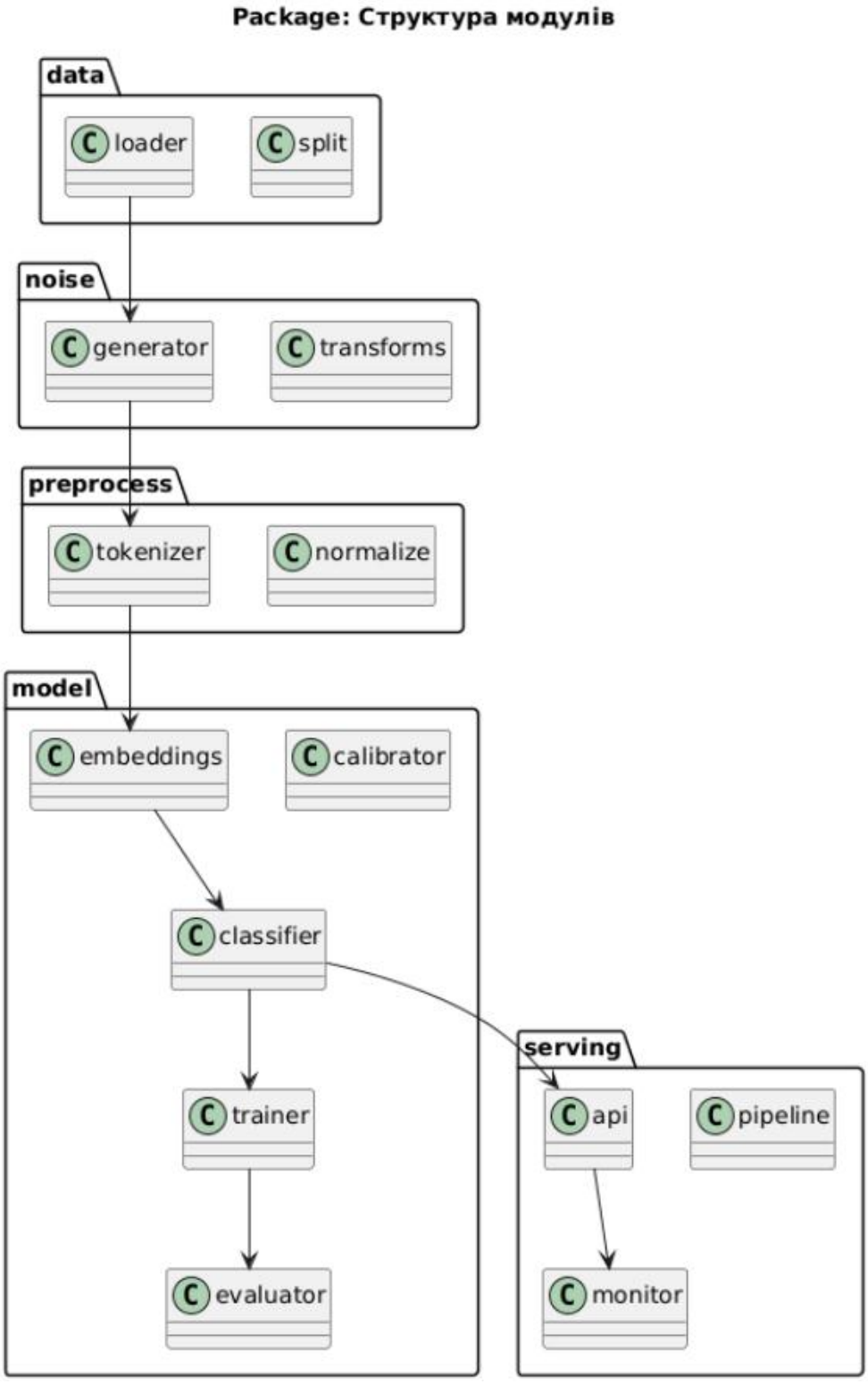
Додаток Б

Діаграма компонентів



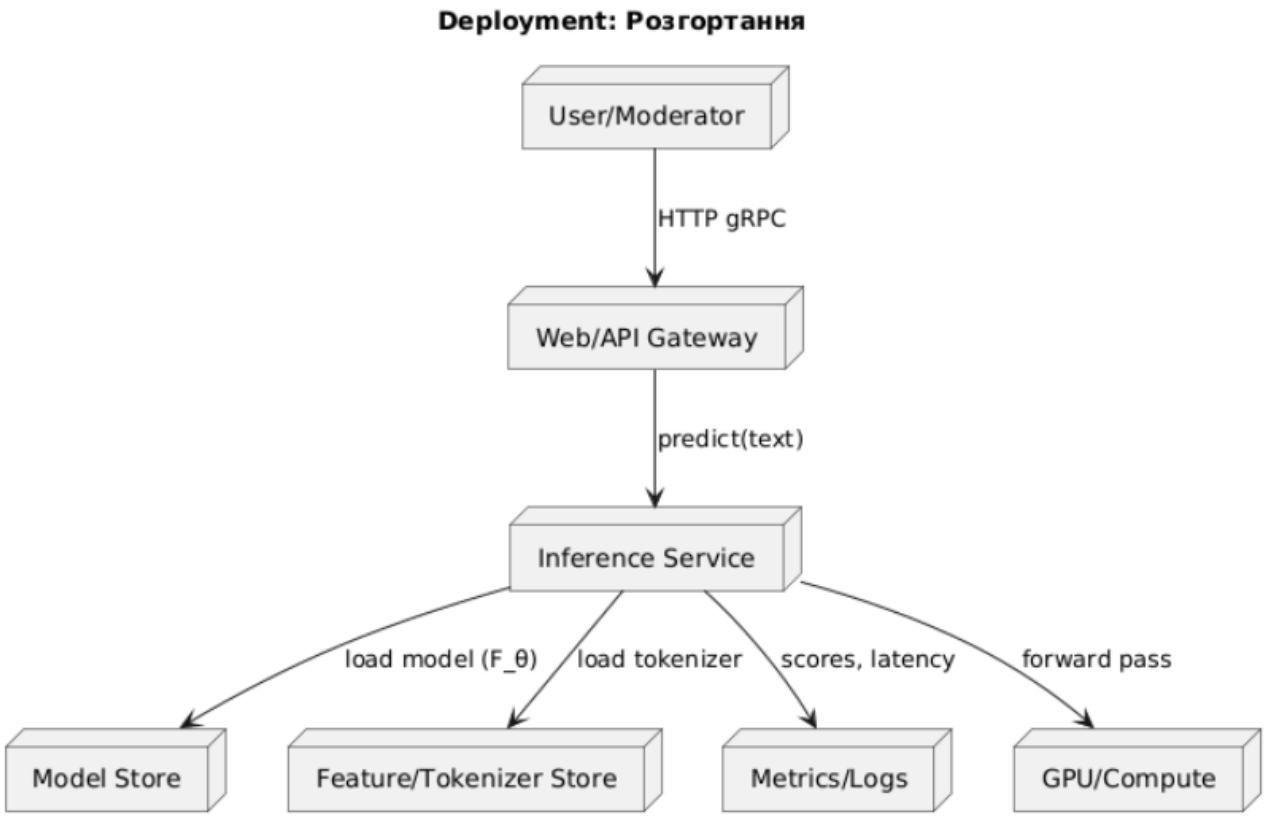
Додаток В

Діаграма пакетів



Додаток Г

Схема розгортання інтелектуальної системи



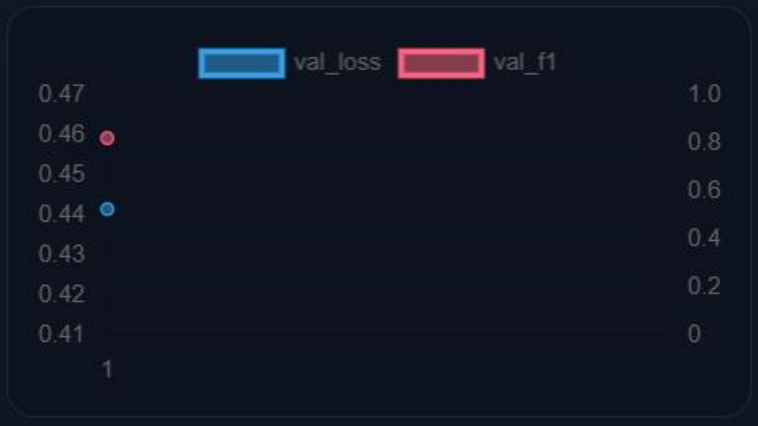
Додаток Д

Світлини екрану інтелектуальної системи

```
  },
  "ok": true,
  "progress": "Epoch 2/3 | step 275 | train_loss=0.3
  "running": true,
  "started_at": 1765734431.6249807
}
```

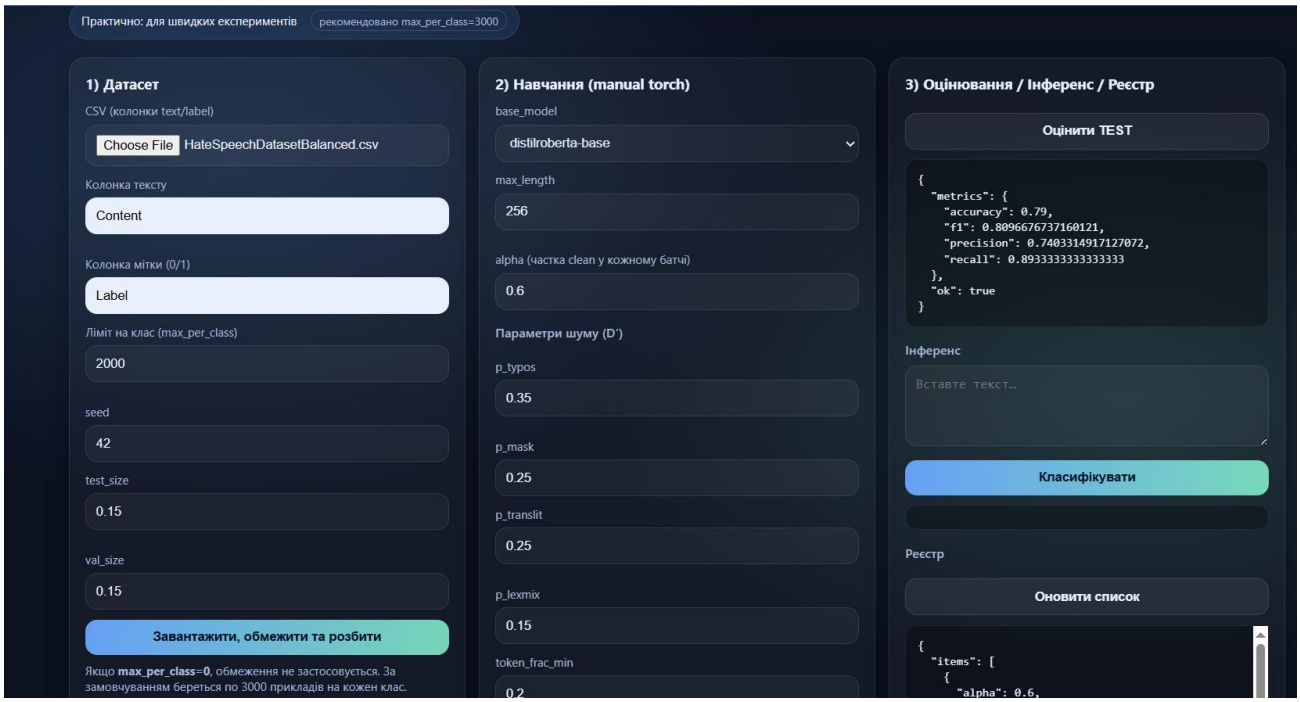

1) Графік на навчальних (train)

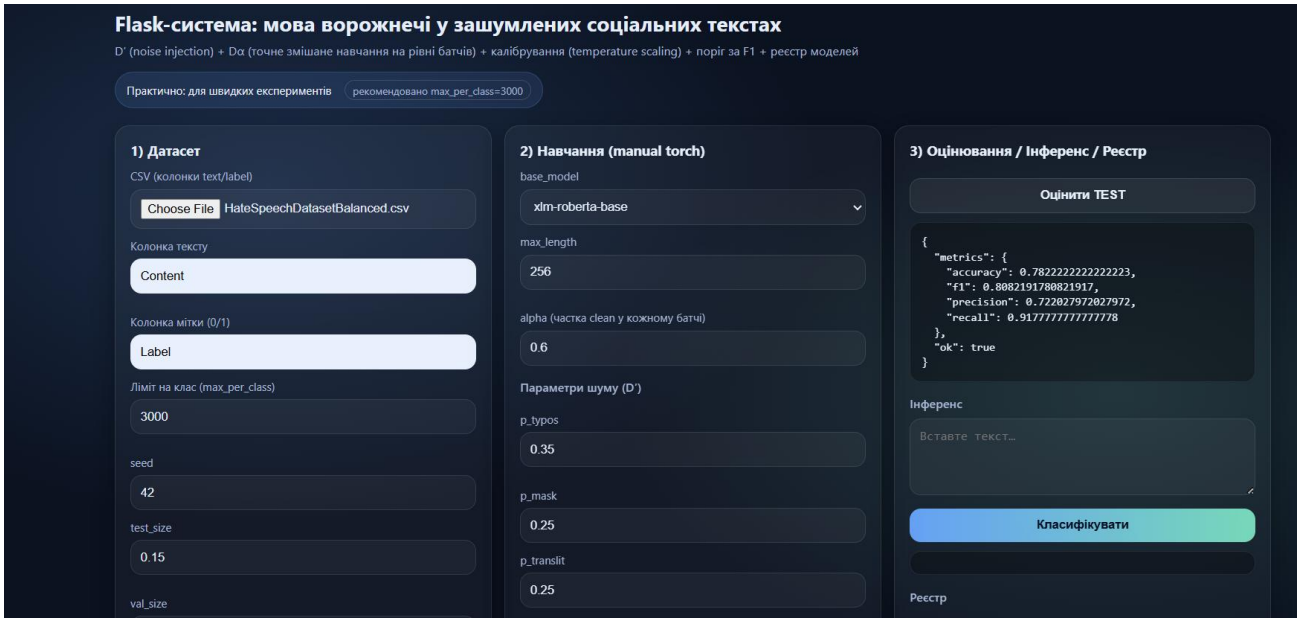
2) Графік на валідаційних (val)



Під час навчання графіки оновлюються автоматично раз на 2 секунди. Після завершення — залишаються на екрані.


1) Датасет
CSV (колонки text/label)
Choose File
No file chosen
Please select a file.
Колонка тексту
Content
Колонка мітки (0/1)
Label
Ліміт на клас (max_per_class)
2000
seed
42
test_size
0.15
val_size
0.15
Завантажити, обмежити та розбити

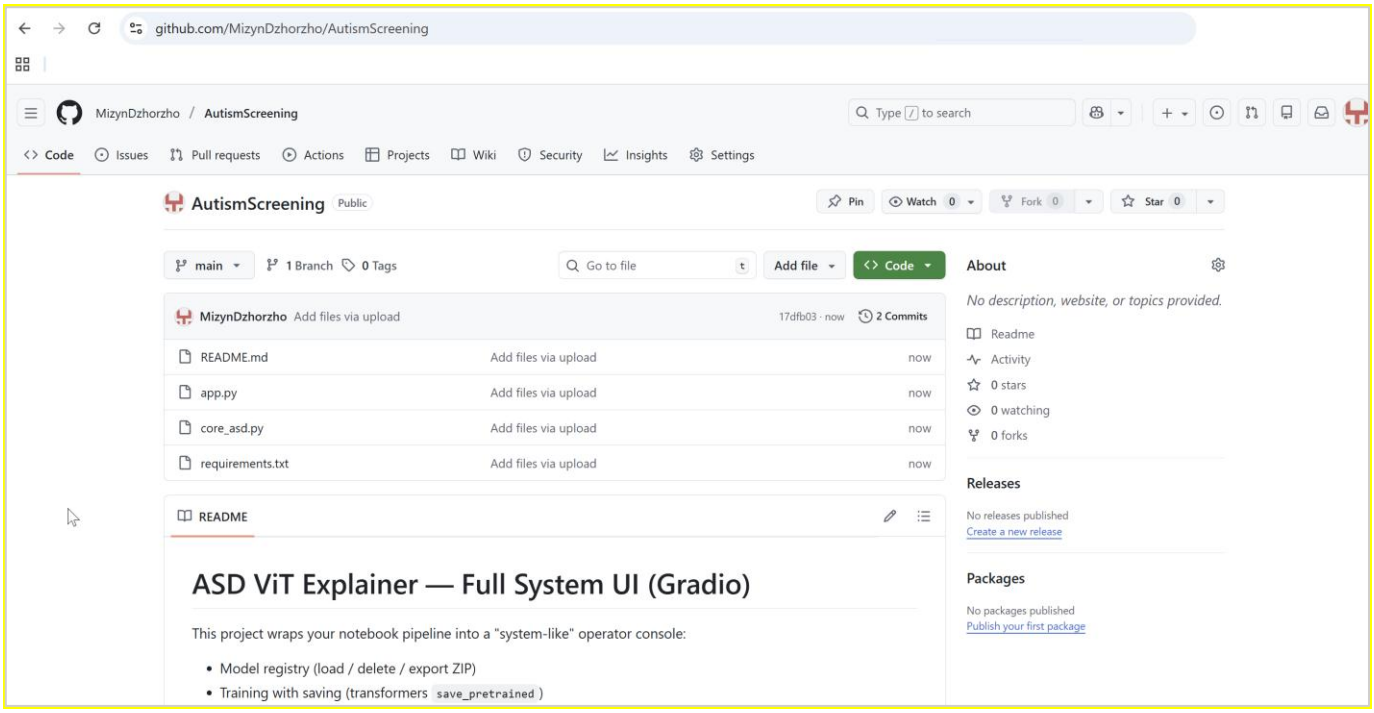

Практично: для швидких експериментів
рекомендовано max_per_class=3000
1) Датасет
CSV (колонки text/label)
Choose File
HateSpeechDatasetBalanced.csv
Колонка тексту
Content
Колонка мітки (0/1)
Label
Ліміт на клас (max_per_class)
2000
seed
42
test_size
0.15
val_size
0.15
Завантажити, обмежити та розбити
Якщо max_per_class=0, обмеження не застосовується. За замовчуванням береться по 3000 прикладів на кожен клас.
2) Навчання (manual torch)
base_model
distilroberta-base
max_length
256
alpha (частка clean у кожному батчі)
0.6
Параметри шуму (D')
p_typos
0.35
p_mask
0.25
p_translit
0.25
p_lexmix
0.15
token_frac_min
0.2
3) Оцінювання / Інференс / Реєстр
Оцінити TEST
{
"metrics": {
"accuracy": 0.79,
"f1": 0.8096676737160121,
"precision": 0.7403314917127072,
"recall": 0.8933333333333333
},
"ok": true
}
Інференс
Вставте текст…
Класифікувати
Реєстр
Оновити список
{
"items": [
{
"alpha": 0.6,


Flask-система: мова ворожнечі у зашумлених соціальних текстах
D' (noise injection) + Dα (точне змішане навчання на рівні батчів) + калібрування (temperature scaling) + поріг за F1 + реєстр моделей
Практично: для швидких експериментів
рекомендовано max_per_class=3000
1) Датасет
CSV (колонки text/label)
Choose File
HateSpeechDatasetBalanced.csv
Колонка тексту
Content
Колонка мітки (0/1)
Label
Ліміт на клас (max_per_class)
3000
seed
42
test_size
0.15
val_size
2) Навчання (manual torch)
base_model
xlm-roberta-base
max_length
256
alpha (частка clean у кожному батчі)
0.6
Параметри шуму (D')
p_typos
0.35
p_mask
0.25
p_translit
0.25
3) Оцінювання / Інференс / Реєстр
Оцінити TEST
{
"metrics": {
"accuracy": 0.7822222222222223,
"f1": 0.8082191780821917,
"precision": 0.722027972027972,
"recall": 0.9177777777777778
},
"ok": true
}
Інференс
Вставте текст…
Класифікувати
Реєстр

Додаток Е

Програмні коди

Вихідний код, використаний у дослідженні, доступний у репозиторії GitHub: https://github.com/MizynDzhorzho/AutismScreening (дата звернення 13.12.2025).



Додаток Ж

Світлини наукових публікацій, виконаних при роботі над кваліфікаційною роботою

1. Підхід до нейромережевого виявлення мови ворожнечі у зашумлених тексових повідомленнях / Боярчук І.О., Молчанова М.О. // *Актуальні проблеми комп'ютерних наук : зб. наук. пр. за матеріалами XVII Всеукр. наук.-практ. конф. (АПКН-2025).* – Хмельницький, 14–15 листоп. 2025 р. – Хмельницький, 2025. – С. 46–51.

2. Молчанова М.О., Мазурець О.В., Боярчук І.О., Залуцька О.О. Об'єктно-орієнтована система для нейромережевого виявлення мови ворожнечі з використанням cloud-технологій / М.О. Молчанова, О.В. Мазурець, І.О. Боярчук, О.О. Залуцька // Вісник Хмельницького національного університету. Серія: Технічні науки. – Хмельницький, 2026. – № 3. (Прийнято до друку).

Міністерство освіти і науки України
Хмельницький національний університет

ЗБІРНИК НАУКОВИХ ПРАЦЬ

за матеріалами XVII Всеукраїнської науково-практичної конференції
«Актуальні проблеми комп'ютерних наук АПКН-2025»

14-15 листопада 2025

Хмельницький 2025

ЗМІСТ

УДК 004.8

Боярчук І.О., Молчанова М.О.

Хмельницький національний університет

ПІДХІД ДО НЕЙРОМЕРЕЖЕВОГО ВИЯВЛЕННЯ МОВИ ВОРОЖНЕЧІ У ЗАШУМЛЕНИХ ТЕКСОВИХ ПОВІДОМЛЕННЯХ

У роботі розглянуто нейромережевий підхід до виявлення мови ворожнечі у зашумлених текстових повідомленнях соціальних мереж і месенджерів, де широко присутні орфографічні відхилення, суржик, емодзі, транслітерація та змішані мовні коди. Метою є підвищення точності та стійкості класифікації за рахунок адаптації моделі до спотворених і навмисно маскованих мовних конструкцій. Запропоновано двоетапний конвеєр: на першому етапі формується контрольовано зашумлений корпус на основі анотованого набору «Hate Speech Detection curated Dataset» з побудовою чистої та зашумленої підвибірок; на другому – здійснюється поетапне донавчання трансформерної моделі типу BERT/RoBERTa з мінімальною нормалізацією тексту. Результати підтверджують доцільність включення параметризованих шумових операторів у навчальний цикл та окреслюють перспективи використання підходу у сервісах модерації контенту й моніторингу інформаційної безпеки.

The paper presents a neural network–based approach to hate speech detection in noisy short texts from social media and messaging platforms, which are characterised by spelling deviations, code-mixing, emojis, transliteration and intentionally distorted tokens. The aim is to improve the accuracy and robustness of hate speech classification by adapting the model to corrupted and masked linguistic patterns. A two-stage pipeline is proposed. First, a controllably noised training corpus is constructed on top of the annotated "Hate Speech Detection curated Dataset" by generating clean and noisy subsets. Second, a transformer model of the BERT/RoBERTa family is progressively fine-tuned on these data under conditions of minimal text normalisation. The proposed approach can serve as a foundation for content moderation services, information security monitoring systems and analytical tools for public online communication.

Підхід до нейромережевого виявлення мови ворожнечі у зашумлених повідомленнях грунтується на усвідомленні того, що сучасні цифрові комунікації формуються не в умовах «лабораторної» мови, а в середовищі постійних викривлень [1]. Соціальні мережі, месенджери та коментарні платформи насичені суржиком [2], орфографічними девіаціями [3], емодзі, навмисними замінами символів, транслітерацією та змішаністю мовних кодів [4]. У такому контексті навіть потужні трансформерні моделі [5, 6], навчені на відносно чистих корпусах [7], втрачають чутливість до завуальованої агресії, оскільки ключові лексичні маркери мови ворожнечі маскуються або систематично спотворюються [8]. Це зумовлює потребу не лише у більш складних архітектурах, а насамперед у методах,

які прямо враховують шумовий характер вхідних даних [9] і відтворюють його на етапі навчання [10].

Актуальність дослідження виявлення мови ворожнечі у зашумлених повідомленнях зумовлена стрімким зростанням обсягів неформального, спонтанного та навмисно модифікованого текстового контенту в соціальних мережах і месенджерах [11]. У середовищах, де користувачі активно застосовують нестандартні правописні форми, жаргон, суржик, коди-міксинг, емодзі та символи, що виконують семантичні функції, традиційні методи автоматичної модерації втрачають ефективність [12]. Мова ворожнечі дедалі частіше маскується шляхом орфографічних викривлень [13], введенням латинізмів, навмисним «каламутінням» токсичних слів або їхнім креативним поділом [14], що ускладнює роботу як класичних алгоритмів, так і сучасних моделей без додаткової адаптації [15]. У контексті посилення інформаційної безпеки та необхідності швидкого реагування на токсичний контент, що може сприяти ескалації конфліктів, радикалізації або порушенню прав користувачів, розроблення методів стійкого до шумів розпізнавання мови ворожнечі постає як критично важливе завдання.

Сучасні підходи NLP відкривають широкі можливості для роботи з такими ускладненими, гетерогенними текстовими потоками, оскільки моделі глибинного навчання здатні оперувати контекстуальними представленнями, враховувати багатозначність та латентні структури, притаманні неформальному онлайн-дискурсу [16]. Трансформерні архітектури [17], зокрема BERT-подібні моделі [18], демонструють здатність інтерпретувати послідовності зі змішаними кодами, відносно стійко працювати з неповними або деформованими токенами [19] та вловлювати семантичну подібність [20] навіть у разі суттєвих орфографічних модифікацій. Додатковий потенціал криється у можливості навчання на спеціально створених зашумлених корпусах [21], де моделі поступово формують інваріантні до шумів представлення й набувають здатності узагальнювати токсичні патерни у значно ширшому діапазоні їхніх проявів, ніж той, який міститься у вихідних «чистих» даних.

Розвиток напрямку також пов'язаний із дедалі активнішим впровадженням методів робастного навчання, спрямованих на забезпечення стабільної роботи моделей у реалістичних умовах. Параметризовані оператори шуму, регуляризаційні техніки та доменно-орієнтоване донавчання дозволяють наближати тренувальні дані до фактичної комунікації користувачів, де помилки, жаргонізми та креативні викривлення є нормою. У поєднанні з методами тонкого донавчання та динамічного машинного розуміння контексту це дає змогу суттєво підвищити якість і надійність класифікації мови ворожнечі, забезпечуючи адаптивність моделей до різноманітних типів спотворених вхідних сигналів [22].

З огляду на те, що токсичний контент часто виникає саме в умовах підвищеної емоційності та неструктурованого письма, NLP-підходи, орієнтовані на

роботу з шумами, стають ключовим інструментом для побудови ефективних систем модерації, моніторингу та аналізу інформаційних загроз. Дослідження у цьому напрямку не лише підвищують точність автоматичного виявлення небезпечних повідомлень, а й сприяють формуванню більш стійких, адаптивних та етично орієнтованих технологій аналізу онлайн-комунікацій.

Метою запропонованого підходу є підвищення точності та стійкості виявлення мови ворожнечі у зашумлених соціальних текстових даних шляхом розроблення нейромережевого методу, адаптованого до спотворених, змішаномовних і навмисно маскованих мовних конструкцій. На відміну від традиційних рішень, що припускають попереднє очищення текстів або їхню максимальну нормалізацію, у роботі реалізовано концепцію контрольованого зашумлення корпусу з подальшим донавчанням моделі у «реалістичних» умовах цифрового мовлення. Об'єктом дослідження виступає процес автоматизованого виявлення мови ворожнечі у соціальних текстах із нестабільною мовною структурою та наявністю шумових викривлень, а предметом – моделі, методи та засоби обробки природної мови, здатні інтерпретувати такі тексти без істотної втрати класифікаційної якості.

Ключова ідея методу полягає у побудові двоетапного конвеєра. На першому етапі формується контрольовано зашумлений навчальний корпус на основі початкового датасету виявлення мови ворожнечі, до якого застосовуються спеціально спроєктовані оператори лінгвістичних викривлень. Вони імітують типові для соціальних мереж практики модифікації тексту: випадкові та систематичні орфографічні помилки, заміна літер схожими символами, часткова або повна транслітерація, домішування елементів іншої мови, вставка емодзі та графічних маркерів, що виконують семантичну або маскувальну функцію. Введення таких спотворень дає змогу сфокусувати навчання моделі не на поверхневій формі токенів, а на стійкіших контекстуальних патернах агресивної комунікації.

Для валідації методу використано анотований набір даних «Hate Speech Detection curated Dataset» з платформи Kaggle, що містить бінарну розмітку повідомлень за класами «мова ворожнечі» / «нейтральний текст» та відображає сучасні практики онлайн-комунікації із включенням сленгу, скорочень і емодзі. На основі цього корпусу побудовано чисту та зашумлену підвибірки, які застосовано для поетапного донавчання трансформерної моделі типу BERT/RoBERTa. Така організація даних дозволяє порівнювати поведінку класифікатора на стандартизованих і спотворених текстах та оцінювати внесок шумового донавчання у підвищення стійкості до реальних цифрових викривлень.

На другому етапі реалізується власне нейромережевий аналіз. Попередньо донавчена модель приймає на вхід текстові фрагменти, які проходять мінімальну попередню обробку (токенізація, приведення до формату, сумісного з архітектурою трансформера) без агресивної нормалізації, що могла б знищити інформативні

шумові патерни. Мережа видає як категоріальну оцінку наявності мови ворожнечі, так і числовий бал впевненості, який може бути використаний у політиках модерації та для побудови людиноорієнтованих інтерфейсів пояснення. Оцінювання якості здійснюється за класичними для задач бінарної класифікації метриками – точністю, повнотою, F_1-мірою, а також за допомогою аналізу помилок на підмножинах із різними типами шуму. Особлива увага приділяється збереженню повноти виявлення ворожих висловлювань, оскільки асиметрія класів робить просту точність ненадійним показником.

Експериментальні дослідження засвідчили, що попереднє формування контрольовано зашумлених корпусів у поєднанні з цілеспрямованим донавчанням трансформерної моделі підвищує стійкість класифікації до орфографічних, графічних і змішаномовних викривлень. У порівнянні з базовою моделлю, навченою на «очищеній» вибірці, запропонований підхід забезпечує вищі значення F1-міри для класу мови ворожнечі саме на повідомленнях зі штучно змодельованим шумом, зменшуючи кількість пропущених агресивних випадків за незначного зростання кількості хибних спрацьовувань. Отримані результати свідчать, що параметризація шумових операторів та їх включення у навчальний цикл є ефективним способом адаптації нейромережевих моделей до реальних умов функціонування в соціальних мережах.

Запропонований підхід може бути використаний як основа для побудови сервісних модулів модерації контенту в соціально-орієнтованих платформах, системах моніторингу інформаційної безпеки та інструментах аналітики публічних комунікацій. Подальші дослідження доцільно спрямувати на розширення номенклатури шумових трансформацій із урахуванням специфіки україномовних і змішаномовних онлайн-спільнот, експерименти з гібридними архітектурами та інтеграцію інтерпретованих методів аналізу (LIME, SHAP) для підвищення пояснюваності рішень нейромережі у чутливих з етичної та правової точки зору сценаріях використання.

Перелік посилань

1. Unnava S., Parasana S. R. A Study of Cyberbullying Detection and Classification Techniques: A Machine Learning Approach. Engineering, Technology & Applied Science Research. 2024. Vol. 14, no. 4. P. 15607–15613.
2. Hladun O., Mazurets O., Molchanova M., Sobko O. Real Time Detection the Person Emotion State Using Neural Network. Scientific Research: Modern Innovations and Future Perspectives. Proceedings of the 2 International scientific and practical conference. November 25-27, 2024. Montreal, Canada. 2024. Pp. 119-123.
3. Mazurets O., Molchanova M., Klimenko V., Prosvitliuk M Practice Implementation of Neural Network Model BART-Large-CNN for Text Annotation. Prospects of Scientific Research in the Conditions of the Modern World. Proceedings of XXVII International scientific and practical conference. June 12-14, 2024. Rotterdam, Netherlands. 2024. Pp. 97-102.

4. Віт Р.В., Мазурець О.В. Тематична класифікація текстової інформації засобами обробки природної мови. Збірник наукових праць XXIII Міжнародної наукової конференції «Нейромережні технології та їх застосування НМТі3-2024». 11-12 грудня 2024. Краматорськ-Тернопіль, ДДМА. 2024. с. 63-66.
5. Овчарук О.М., Мазурець О.В. Нейромережеве діагностування проявів ПТСР у текстовому контенті з використанням помилко-орієнтованого навчального набору даних. Науковий журнал «Вісник Хмельницького національного університету» серія: Технічні науки. Хмельницький, 2024. №6, Т.1 (343). С. 195-200.
6. Civila S. Cyberbullying. Comprehensive Sexuality Education for Gender-Based Violence Prevention. 2024. P. 229–245.
7. Овчарук О.М., Мазурець О.В. Нейромережева архітектура з квантовим шаром для аналізу текстових повідомлень на прояви посттравматичного стресового розладу. Науковий журнал «Наука і техніка сьогодні». Київ, 2024. №13 (41). С. 1192-1204.
8. Мазурець О.В., Тимофієв І.А., Кліменко В.І., Тищенко О.О. Метод виявлення депресивного стану пов'язаного із навчанням у закладах освіти із використанням нейромережі дуальної архітектури. Науковий журнал «Вісник Херсонського національного технічного університету». 2024. №4 (91). С. 311-318.
9. Віт Р.В., Мазурець О.В. Підхід до тематичної класифікації текстової інформації засобами обробки природної мови. Науковий журнал «Наукові праці Донецького національного технічного університету», серія «Проблеми моделювання та автоматизації проектування». 2025. №1 (21). С. 94-99.
10. Овчарук О.М., Мазурець О.В. Нейромережевий метод діагностування психологічних розладів за аналізом повідомлень на основі роздільного підходу до класифікації. Міжнародний науково-технічний журнал «Вимірювальна та обчислювальна техніка в технологічних процесах». № 1, 2025. с. 210-216.
11. Blazhuk V., Mazurets O., Zalutska O. An Approach to Using the mBERT Deep Learning Neural Network Model for Identifying Emotional Components and Communication Intentions. The Impact of Scientific Research on the Development of the Modern World. Proceedings of the XLIV International scientific and practical conference. October 23-25, 2024. Dubrovnik, Croatia. 2024. Pp. 79-84.
12. Tymofiiev I., Mazurets O., Hardysh D., Molchanova M. Neural Network Dual Architecture for Depression Detection Using Cloud Services. Scientific Research in the Era of Digital Technologies: Challenges and Opportunities. Proceedings of the XLVI International scientific and practical conference. November 6-8, 2024. Barcelona, Spain. 2024. Pp. 84-88.
13. Юрченко Д.Ю., Овчарук О.М., Мазурець О.В., Шевчук П.О. Метод використання нейромережі гібридної архітектури для визначення емоційної тональності текстових повідомлень. Міжнародний науково-технічний журнал «Вимірювальна та обчислювальна техніка в технологічних процесах». № 2, 2025. с. 136-141.
14. Mazurets O., Tymofiiev I., Dydo R. Approach for Using Neural Network BERT-GPT2 Dual Transformer Architecture for Detecting Persons Depressive State. Ricerche scientifiche e metodi della loro realizzazione: esperienza mondiale e realtà domestiche. Raccolta di articoli scientifici con gli atti della VI Conferenza scientifica e pratica internazionale. 15 novembre, 2024. Bologna, Repubblica Italiana. 2024. Pp. 147-151.
15. Віт Р.В., Мазурець О.В. Метод виявлення психологічного цифрового перевантаження за аналізом текстових даних нейромережевими моделями глибокого навчання. Науковий

журнал «Вісник Херсонського національного технічного університету». 2025. №2 (93). Т. 2. С. 107-114.

16. Mazurets O., Sobko O., Vit R., Pasternak V. Practical Approach for Detection by Deep Learning of Target Objects of Subject Area Based on Semantic Connectivity Indicators in Audio Database. Proceedings of XXIV International Scientific and Practical Conference «Modern Scientific Challenges are the Driving Force of the Development of Scientific Research». May 22-24, 2024. Bruges, Belgium. International Scientific Unity. 2024. Pp. 91-96.

17. Віт Р.В., Мазурець О.В. Метод виявлення комунікаційних об'єктів як індикаторів цифрової втоми. Інтелектуальний метод виявлення цільових об'єктів предметної області для класифікації текстової інформації. Матеріали XIII Міжнародної науково-практичної конференції «Інформаційні управляючі системи та технології ІУСТ-ОДЕСА-2025». 24-26.09.2025. Одеса. 2025. С.119-121.

18. Yurchenko D., Mazurets O., Didur V., Molchanova M. Approach to Using Cloud Services for Visual Analytics of Neural Network Analysis of Texts Emotional Tonality. The Future of Scientific Discoveries: New Trends and Technologies. Proceedings of the XLVII International scientific and practical conference. November 13-15, 2024. Marseille, France. 2024. Pp. 108-113.

19. Mazurets O.V., Sobko O.V., Molchanova M.O., Zalutska O.O., Yurchak A.V. Practical Implementation of Neural Network Method for Stress Features Detection by Social Internet Networks Posts. Global Science: Prospects and Innovations. Proceedings of the II International Scientific and Theoretical Conference «Scientific Review of the Actual Events, Achievements and Problems». May 31, 2024. Berlin, Federal Republic of Germany: International Center of Scientific Research. 2024. Pp. 160-167.

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

21. Casas F. Age Discrimination. Encyclopedia of Quality of Life and Well-Being Research. Cham, 2023. P. 118–121.

22. Lee H. Lived Religion in Religious Vaccine Exemptions. Perspectives in Biology and Medicine. 2024. Vol. 67, no. 1. P. 96–113.

Довідка: ВХНУ ТН 15-12/2024

Видання: Herald of Khmelnytskyi National University. Technical Sciences (Вісник Хмельницького національного університету. Технічні науки)

Категорія фаховості видання: затверджено як наукове фахове видання України, у якому можуть публікуватися результати дисертаційних робіт на здобуття наукових ступенів доктора наук, кандидата наук та ступеня доктора філософії, категорії «Б» (наказ МОН №1643 від 28.12.2019, наказ МОН №409 від 17.03.2020).

Напрям – технічні науки за спеціальностями – 101, 121, 122, 123, 124, 125, 141, 151, 161, 172, 181, 182 (28.12.2019), спеціальності – 131, 132, 133 (17.03.2020).

Назва статті: ОБ'ЄКТНО-ОРІЄНТОВАНА СИСТЕМА ДЛЯ НЕЙРОМЕРЕЖЕВОГО ВИЯВЛЕННЯ МОВИ ВОРОЖНЕЧІ З ВИКОРИСТАННЯМ CLOUD-ТЕХНОЛОГІЙ

Автори: Молчанова Марина, Мазурець Олександр Боярчук Ілля, Залуцька Ольга.
Хмельницький національний університет

Номер, у який прийнято статтю: №3 за 2026 рік.

15.12.2025

Начальника відділу
інтелектуальної власності та трансферу технологій Ю.В.Кравчик

ВІДДІЛ КАДРІВ
І.С.Мартинюк

УДК 004.8

МОЛЧАНОВА МАРИНА
Хмельницький національний університет
https://orcid.org/0000-0001-9810-936X
e-mail: m.o.molchanova@gmail.com
МАЗУРЕЦЬ ОЛЕКСАНДР
Хмельницький національний університет
https://orcid.org/0000-0002-8900-0650
e-mail: exe.chong@gmail.com
БОЯРЧУК ІЛЛЯ
Хмельницький національний університет
e-mail: romaboy2005@gmail.com
ЗАЛУЦЬКА ОЛЬГА
Хмельницький національний університет
https://orcid.org/0000-0003-1242-3548
e-mail: zalutsk.olha@gmail.com

ОБ'ЄКТНО-ОРІЄНТОВАНА СИСТЕМА ДЛЯ НЕЙРОМЕРЕЖЕВОГО ВИЯВЛЕННЯ МОВИ ВОРОЖНЕЧІ З ВИКОРИСТАННЯМ CLOUD-ТЕХНОЛОГІЙ

У статті представлено результати розроблення та експериментального дослідження об'єктно-орієнтованої системи нейромережевого виявлення мови ворожнечі з використанням cloud-технологій. Запропоновано метод нейромережевого виявлення мови ворожнечі, що передбачає двоетапну обробку: підготовку стійкої нейромережевої моделі шляхом модульного введення шуму у навчальні дані та подальше використання цієї моделі для інференсу у хмарному середовищі. Введення шуму дозволяє імітувати типові спотворення, характерні для соціальних платформ (орфографічні варіації, символічні заміни, часткове маскування), що підвищує стійкість класифікатора до реальних текстових умов. Архітектура системи реалізована на базі модулів TextIndexDataset, BatchNoisyCollator та TemperatureScaler, які відповідають за інкапсуляцію даних, формування батчів зі спотвореннями та калібрування ймовірнісних прогнозів відповідно. Хмарне розгортання забезпечує масштабованість обчислень, централізоване збереження моделей і параметрів, а також повторюваність експериментів.

Експериментальні дослідження проведено на датасетах «Hate Speech Detection curated Dataset» (для навчання) та «Hate Speech and Offensive Language Detection» (для зовнішньої валідації). Отримані результати доводять, що навчання моделей у змішаному режимі (чисті та зашумлені приклади) забезпечує кращу узагальнюваність: на внутрішньому тесті моделі без шуму показують вищу F1-міру, проте на зовнішньому датасеті перевага моделей, навчальних зі спотвореннями, становить 1,5–1,7 %. Це підтверджує ефективність модульного введення шуму для підвищення робастності моделей і зменшення ефекту переадаптації до навчального корпусу.

Запропонований підхід поєднує принципи об'єктно-орієнтованого проєктування, хмарних обчислень і глибинного навчання, що робить його придатним для масштабованих систем моніторингу та модерації контенту. Перспективи подальших досліджень полягають у розширенні набору стратегій зашумлення, удосконаленні калібрування прогнозів і перевірці запропонованого рішення на багатомовних корпусах та реальних потоках повідомлень.

Ключові слова: мова ворожнечі, трансформерні моделі, робастність, модульне введення шуму.

MOLCHANOVA MARYNA, MAZURETS OLEKSANDR, BOIARCHUK ILLIA, ZALUTSKA OLHA
Khmelnytskyi National University

OBJECT-ORIENTED SYSTEM FOR NEURAL NETWORK DETECTION OF HATE SPEECH USING CLOUD TECHNOLOGIES

The article presents the results of the development and experimental study of an object-oriented neural network system for hate speech detection using cloud technologies. A method for neural network detection of hate speech is proposed, which involves two-stage processing: training a stable neural network model by modularly introducing noise into the training data and further using this model for inference in a cloud environment. Introducing noise allows you to simulate typical distortions characteristic of social platforms (spelling variations, symbolic substitutions, partial masking), which increases the stability of the classifier to real text conditions. The system architecture is implemented on the basis of the TextIndexDataset, BatchNoisyCollator and TemperatureScaler modules, which are responsible for data encapsulation, the formation of batches with distortions and the calibration of probabilistic forecasts, respectively. Cloud deployment ensures scalability of calculations, centralized storage of models and parameters, as well as repeatability of experiments.

Experimental studies were conducted on the datasets "Hate Speech Detection curated Dataset" (for training) and "Hate Speech and Offensive Language Detection" (for external validation). The obtained results prove that training models in mixed mode (clean and noisy examples) provides better generalization: on the internal test, models without noise show a higher F1-measure, however, on the external dataset, the advantage of models trained with distortions is 1.5–1.7%. This confirms the effectiveness of modular noise injection to increase the robustness of models and reduce the effect of overfitting to the training corpus.

The proposed approach combines the principles of object-oriented design, cloud computing and deep learning, which makes it suitable for scalable content monitoring and moderation systems. Prospects for further research are to expand the set of noise reduction strategies, improve the calibration of predictions and verify the proposed solution on multilingual corpora and real message streams.

Keywords: hate speech, transformative models, robustness, modular noise injection.

Постановка проблеми у загальному вигляді та її зв'язок із важливими науковими чи практичними завданнями

Стрімке зростання обсягу цифрових комунікацій зумовлює поширення мови ворожнечі у формах, що часто є навмисно спотвореними, змішаномовними та маскованими, через що традиційні підходи аналізу тексту, орієнтовані на «нормативні» дані, демонструють зниження ефективності в реальних умовах. Це формує науково-технічну проблему розроблення нейромережевого методу, здатного інтерпретувати такі нестабільні текстові конструкції шляхом моделювання шумових викривлень у навчальних корпусах та адаптивного налаштування класифікатора [1].

Практичне значення проблеми пов'язане із завданнями автоматизованої модерації контенту та забезпечення інформаційної безпеки [2], де потрібні стійкі моделі й масштабоване розгортання; використання Cloud-технологій є доцільним для організації обчислювально інтенсивного навчання, сервісного інференсу та відтворюваності експериментів у прикладних системах [3].

Актуальність досліджень у сфері нейромережевого виявлення мови ворожнечі зумовлена стрімким зростанням обсягів користувацького контенту в соціальних мережах, месенджерах і цифрових платформах [4], а також підвищеними вимогами до швидкої, точної та масштабованої модерації повідомлень [5]. Традиційні підходи, засновані на словниках заборонених слів або ручній перевірці, виявляються малоефективними в умовах динамічної еволюції мови, контекстної багатозначності та навмисного маскування агресивних висловлювань [6]. Саме тому методи обробки природної мови, зокрема сучасні трансформерні нейромережі, стають ключовим інструментом для автоматизованого аналізу семантики, прагматики та контексту текстових повідомлень [7]. NLP-моделі здатні враховувати не лише лексичні маркери ворожнечі, а й приховані смислові конструкції [8], іронію, сарказм, контекстні залежності між токенами, що є критично важливим для надійного розпізнавання мови ворожнечі в реальних умовах [9].

Використання глибинних NLP-моделей у поєднанні з методами контрольованого зашумлення навчальних даних відкриває нові можливості для підвищення робастності систем аналізу тексту [10]. Соціальні платформи характеризуються високим рівнем лінгвістичної варіативності: користувачі активно застосовують орфографічні помилки, скорочення, транслітерацію, емодзі, символічні заміни та навмисні спотворення, спрямовані на обходження автоматичних фільтрів [11]. Інтеграція шумових операторів у навчальний цикл дозволяє NLP-моделям формувати більш узагальнені семантичні представлення, зменшуючи залежність від поверхневих ознак і підвищуючи стійкість до атак на класифікатор. У цьому контексті запропонований підхід демонструє практичну цінність для побудови систем, здатних працювати з «нечистими» текстами без агресивної нормалізації, що часто призводить до втрати смислових нюансів.

Хмарні технології суттєво розширюють можливості застосування NLP у задачах виявлення мови ворожнечі, забезпечуючи еластичне масштабування обчислювальних ресурсів, централізоване керування моделями та обробку потокових даних у режимі, наближеному до реального часу. Поєднання cloud-інфраструктури з модульною архітектурою NLP-системи створює умови для інтеграції таких рішень у великі платформи моніторингу контенту, де важливими є стабільність сервісу, відтворюваність експериментів і можливість швидкого оновлення моделей [12]. Крім того, калібрування ймовірнісних прогнозів у хмарному середовищі підвищує інтерпретованість результатів і дозволяє використовувати виходи моделі як основу для прийняття управлінських або юридично значущих рішень.

Перспективи розвитку NLP у цьому напрямку пов’язані з переходом до багатомовних і крослінгвальних моделей, здатних виявляти мову ворожнечі незалежно від мови повідомлення, а також із поглибленням аналізу дискурсивних і соціолінгвістичних характеристик тексту [13]. Подальше поєднання трансформерних моделей із контекстною інформацією про користувачів, часову динаміку комунікації та мережеві взаємодії може суттєво підвищити точність і практичну цінність систем. Таким чином, застосування сучасних методів NLP у хмарних нейромережевих системах виявлення мови ворожнечі є перспективним напрямком, що відповідає актуальним викликам цифрової безпеки та створює основу для ефективної, масштабованої та адаптивної модерації текстового контенту.

Аналіз досліджень та публікацій

Розвиток соціальних платформ суттєво інтенсифікував поширення мови ворожнечі, що підвищує ризики для суспільної стабільності та психологічної безпеки користувачів. Класичні підходи детекції типу словникові фільтри, правила та традиційні методи машинного навчання часто не забезпечують належної якості для контекстно залежних, завуальованих і непрямих проявів агресії. У зв’язку з цим у фокусі сучасних досліджень перебуває застосування великих мовних моделей і трансформерних архітектур (GPT-3, BERT та наступників) для автоматизованого виявлення hate speech, із аналізом їхніх переваг, обмежень і впливу на точність, справедливість та стійкість класифікаційних систем; узагальнення робіт окреслює поточний стан технологій і напрями подальшого розвитку в бік підвищення ефективності та етичної надійності [14].

У дослідженні [15] розглянуто виявлення мови ворожнечі в малоресурсних мовах (арабська, урду), де задачу ускладнюють мовна варіативність, неявна агресія та дефіцит корпусів. Автори сформували вручну анотований мультимовний датасет UAHSD-2025 на основі платформи X для бінарної та багатокласової класифікації (п’ять категорій). Для зменшення міжмовних відмінностей застосовано дві стратегії: попередній переклад усіх текстів у спільну цільову мову та спільне мультимовне навчання без перекладу. За результатами експериментів із TF-IDF і класичними алгоритмами, моделями на FastText/GloVe та контекстуальними представленнями найкращі показники отримано на XLM-R: точність 0.99 для бінарної класифікації (арабська, урду та мультимовний набір) і 0.95/0.94/0.94 відповідно для багатокласового режиму, що підтверджує перевагу мультимовних трансформерів у низькоресурсних умовах.

Робота [16] присвячена детекції мови ворожнечі для мов деванагарі (гінді, непальська), де автоматизований аналіз обмежений нестачею корпусів і адаптованих моделей. Запропоновано гібридну архітектуру Attention BiLSTM-XLM-RoBERTa, яка поєднує послідовне моделювання з контекстуальними

мультимовними репрезентаціями, що дає змогу краще враховувати варіативність і нестандартну орфографію. У задачі виявлення мови ворожнечі модель досягла Macro $F_1 = 0.7481$, демонструючи працездатність за умов мовної неоднорідності.

Формулювання цілей статті

Метою роботи є: розроблення та обґрунтування об'єктно-орієнтованої системи для нейромережевого виявлення мови ворожнечі з використанням Cloud-технологій, яка забезпечує стійкість класифікації до зашумлених і навмисно спотворених текстів шляхом модульного введення шуму на етапі підготовки даних та подальшого калібрування прогнозів моделі.

Виклад основного матеріалу

В основі запропонованої об'єктно-орієнтованої системи лежить двоетапний метод (рис. 1), який поєднує підготовку стійкої нейромережевої моделі та її подальше застосування для аналізу текстів у прикладному режимі.

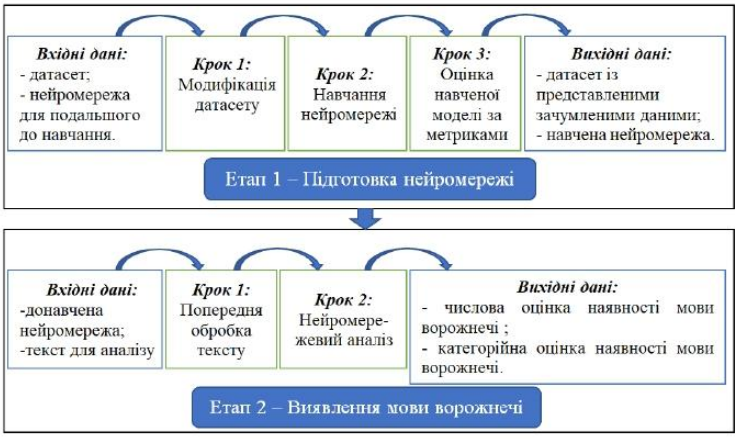


Рис. 1. Схема методу нейромережевого виявлення мови ворожнечі

Етап 1 – підготовка нейромережі. На вхід подається датасет і базова нейромережа, призначена для донавчання. На першому кроці виконується модифікація датасету, зокрема формування спотворених варіантів текстів (імітація реальних умов соціальних платформ: помилки, заміни символів, маскування лексем), що підвищує різноманітність навчальних прикладів і зменшує чутливість моделі до таких викривлень. Далі здійснюється навчання нейромережі на підготовлених даних. На третьому кроці проводиться оцінювання якості за метриками класифікації та відбір конфігурації з найкращими показниками. Виходом етапу є оновлений датасет із зашумленими даними та навчена нейромережа, готова до інференсу.

Етап 2 – виявлення мови ворожнечі. На вхід подаються донавчена нейромережа та текст для аналізу. Спочатку виконується попередня обробка тексту (уніфікація формату, підготовка до подачі в модель). Далі здійснюється нейромережевий аналіз, у межах якого модель формує прогноз щодо наявності ознак мови ворожнечі. Результатом є числова оцінка та категоріальна інтерпретація, що може використовуватися для автоматизованої модерації та підтримки рішень.

Cloud-технології застосовуються як середовище виконання обчислювально інтенсивних операцій (навчання/інференс), зберігання артефактів (модель, конфігурації, метрики) та забезпечення відтворюваності експериментів і масштабованості сервісу.

Об'єктно-орієнтована система наведена на рис. 2 і реалізує модульний підхід до підготовки даних, формування батчів та постобробки прогнозів нейромережі. Основою підсистеми даних є клас TextIndexDataset (наслідує Dataset), який інкапсулює тексти, мітки та індекси прикладів і надає стандартизований доступ через

методи __len__() та __getitem__(idx). Формування батчів виконується засобами DataLoader, для якого колейт-функцією використовується клас BatchNoisyCollator. Він відповідає за кероване внесення шуму в тексти відповідно до заданої конфігурації (параметри інтенсивності та словникових/символьних замін), а також за подальшу токенізацію з використанням сумісного токенізатора (PreTrainedTokenizerBase). Введення шуму винесено в окрему функцію apply_noise_to_text(), що забезпечує заміннність і розширюваність стратегій зашумлення без зміни решти компонентів.

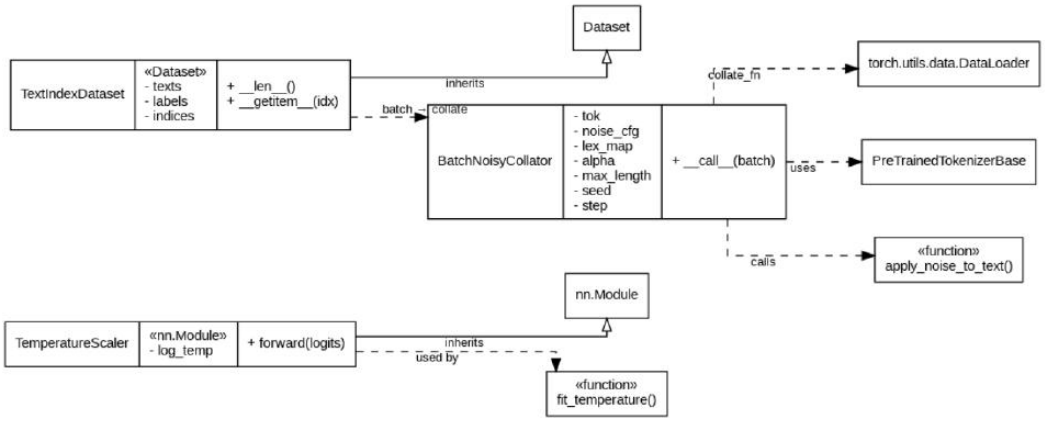


Рис. 2. Діаграма класів об’єктно-орієнтованої системи

Для підвищення надійності ймовірнісних оцінок у системі передбачено модуль калібрування прогнозів – клас TemperatureScaler (наслідує nn.Module), який виконує температурне масштабування логітів у методі forward(logits). Параметр калібрування оцінюється процедурою fit_temperature() на валідаційних даних без модифікації ваг базової нейромережі.

Хмарна архітектура розгортання (рис. 3) відображає сервісну організацію системи, орієнтовану на масштабоване виконання нейромережевого інференсу та відтворюване навчання в керованому хмарному середовищі.

Взаємодія користувача або модератора із системою здійснюється через вебінтерфейс адміністратора, який передає запити на аналіз тексту захищеним каналом до шлюзу прикладного програмного інтерфейсу та балансувальника навантаження. Далі запит типу predict(text) маршрутизується до сервісу інференсу, що розгортається у вигляді контейнеризованого компонента з можливістю автоматичного масштабування відповідно до поточного потоку звернень.

Сервіс інференсу під час виконання аналізу використовує кілька інфраструктурних ресурсів: зі сховища моделей завантажуються артефакти навченої нейромережі, зі сховища токенізаторів і конфігурацій, параметри попередньої обробки та перетворення тексту. Обчислювальна частина інференсу виконується на GPU/обчислювальних вузлах, що забезпечує низьку затримку відповіді при обробці потокових даних. Паралельно результати роботи сервісу (оцінки, часові характеристики, помилки) передаються до підсистеми спостережуваності хмарного середовища (метрики, журнали, трасування), що забезпечує контроль продуктивності, виявлення збоїв та підтримку експлуатації.

У межах сервісу інференсу логіка обробки реалізована як узгоджений набір об’єктно-орієнтованих модулів (підготовка даних, формування батчів, калібрування), що забезпечує відокремлення відповідальностей і розширюваність. Зокрема, модуль калібрування виконує постобробку прогнозів (температурне масштабування), а модуль внесення шуму може застосовуватися як додатковий шлях під час експериментів або навчання для оцінювання стійкості моделі до спотворених текстів.

Окремо на схемі виділено контур експериментів і навчання: ініціювання серій експериментів

здійснюється через інтерфейс керування навчанням, який формує завдання для хмарного навчального процесу на GPU. Навчальний процес використовує датасети зі сховища даних, формує ваги моделі та зберігає їх у реєстрі моделей, а також передає метрики й журнали до системи моніторингу. Така організація забезпечує централізоване керування артефактами (дані/моделі/конфігурації), повторюваність експериментів, можливість швидкого оновлення моделей у сервісі інференсу та масштабування обчислень без прив'язки до локальної інфраструктури.

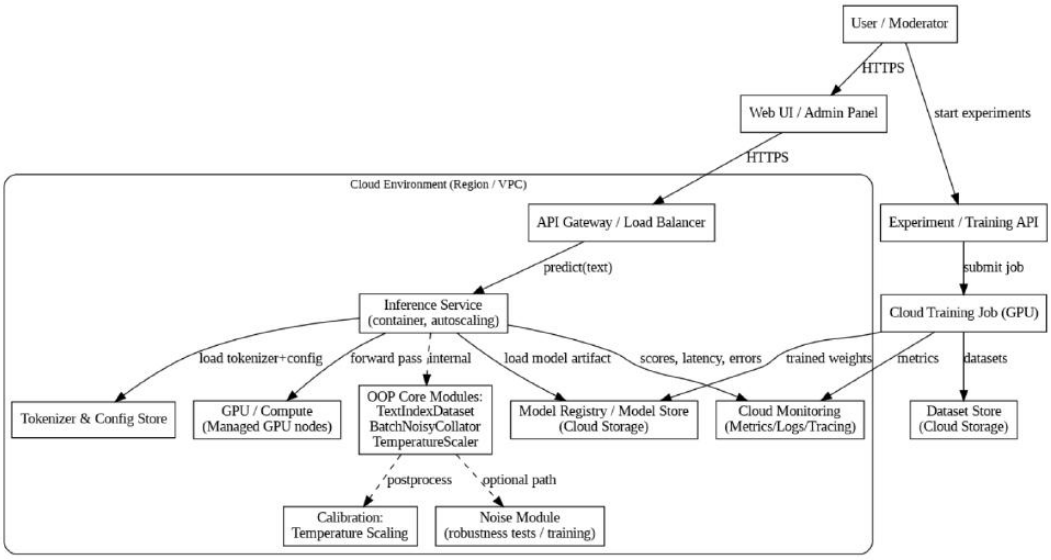


Рис. 3. Хмарна архітектура розгортання об'єктно-орієнтованої системи нейромережевого виявлення мови ворожнечі

У таблиці 1 наведено експериментальні дані, отримані розробленим програмним забезпеченням. Навчання всіх нейромережевих архітектур здійснювалось на датасеті «Hate Speech Detection curated Dataset» [17], а валідація відбувалась на підвибірці цього ж датасету, яка не брала участь у навчанні, та на датасеті «Hate Speech and Offensive Language Detection» [18] для дослідження узагальнювано здатності моделей.

Таблиця 1

Порівняння запусків за валідаційною F1-мірою

Архітектура	Частка чистих прикладів у мініпакеті	Режим навчання	Температура калібрування	F1 (тестова вибірка навчального датасету)	F1 (тестова вибірка валідаційного датасету)
roberta-base	0,6	змішані (чисті+шум)	1,636	0,8339	0,8133
distilroberta-base	0,6	змішані (чисті+шум)	1,661	0,8086	0,781
roberta-base	1,0	лише чисті	1,590	0,8589	0,7580
distilroberta-base	1,0	лише чисті	1,620	0,8329	0,763

За даними таблиці 1 встановлено, що навчання моделей у змішаному режимі (чисті приклади та зашумлені) забезпечує кращу узагальнюваність порівняно з навчанням лише на чистих текстах [19]. Для обох архітектур у межах тестової вибірки того самого датасету, на якому виконувалось навчання, моделі, навчені без шуму, демонструють вищу F1-міру (для roberta-base: 0,8589 проти 0,8339; для distilroberta-base: 0,8329 проти

0,8086), що є очікуваним через більшу відповідність розподілу даних навчальній вибірці.

Водночас під час перевірки на зовнішньому датасеті «Hate Speech and Offensive Language Detection» спостерігається протилежна тенденція: моделі, навчені зі змішаними даними, забезпечують суттєво вищі значення F1 (для roberta-base: 0,8133 проти 0,7580; для distilroberta-base: 0,7810 проти 0,7630). Це підтверджує, що модульне введення шуму під час навчання знижує переадаптацію до специфіки навчального корпусу [20] та підвищує стійкість до варіативності й спотворень, характерних для реальних соціальних текстів і різних датасетів [21].

Висновки з даного дослідження
і перспективи подальших розвідок у даному напрямі

У ході дослідження розроблено та апробовано об'єктно-орієнтовану систему нейромережевого виявлення мови ворожнечі з використанням хмарних технологій, що забезпечує відтворюваність експериментів і придатність до масштабованого розгортання. За результатами порівняльних запусків встановлено, що навчання у змішаному режимі із модульним введенням шуму підвищує узагальнювальну здатність моделей на зовнішньому датасеті, знижуючи чутливість до доменного зсуву та типових спотворень соціальних текстів. Водночас навчання лише на чистих даних забезпечує вищі показники на тестовій вибірці того самого навчального датасету, але гірше переноситься на інші джерела даних.

Перспективи подальших досліджень пов'язані з розширенням набору стратегій зашумлення з урахуванням мовозмішування та навмисного маскування, а також із вивченням впливу цих стратегій на різні трансформерні архітектури. Доцільним є також поглиблення калібрування прогнозів і порогової оптимізації для сценаріїв модерації, а також проведення масштабніших експериментів на багатомовних корпусах і реальних потоках повідомлень у хмарному середовищі.

Література

1. What is hate speech? United Nations. URL: https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech.

2. Мазурець О.В., Тимофієв І.А., Кліменко В.І., Тищенко О.О. Метод виявлення депресивного стану, пов'язаного із навчанням у закладах освіти, із використанням нейромережі дуальної архітектури / О.В. Мазурець, І.А. Тимофієв, В.І. Кліменко, О.О. Тищенко // Вісник Херсонського національного технічного університету. – 2024. – № 4 (91). – С. 311–318.

3. Murava V., Zalutska O., Didur V., Mazurets O. Software Architecture of Information System for Exchanging LLM Thematic Prompts / V. Murava, O. Zalutska, V. Didur, O. Mazurets // Global Trends in the Development of Information Technology and Science. Proceedings of IV International Scientific and Practical Conference. – Stockholm, Sweden, 25–27 June 2025. – P. 121–127.

4. Молчанова М.О., Мазурець О.В., Собко О.В., Кліменко В.І., Андрощук В.І. Метод нейромережевого виявлення кібербулінгу з використанням хмарних сервісів та об'єктно-орієнтованої моделі / М.О. Молчанова, О.В. Мазурець, О.В. Собко, В.І. Кліменко, В.І. Андрощук // Вісник Хмельницького національного університету. Серія: Технічні науки. – Хмельницький, 2024. – № 2 (333). – С. 200–206.

5. Molchanova M., Didur V., Sobko O., Mazurets O. Detection of Web Propaganda Patterns by Transformer Neural Networks: Improving Efficiency via Dataset Balancing / M. Molchanova, V. Didur, O. Sobko, O. Mazurets // CEUR Workshop Proceedings. – 2025. – Vol. 3988. – P. 112–126.

6. Sobko O., Mazurets O., Didur V., Chervonchuk I. Recurrent Neural Network Model Architecture for Detecting a Tendency to Atypical Behavior of Individuals by Text Posts / O. Sobko, O. Mazurets, V. Didur, I. Chervonchuk // Theoretical and Practical Aspects of Modern Research. Proceedings of XXVI International Scientific and Practical Conference. – Ottawa, Canada, 5–7 June 2024. – International Scientific Unity, 2024. – P. 113–117.

7. Віт Р.В., Мазурець О.В. Метод виявлення комунікаційних об'єктів як індикаторів цифрової втоми. Інтелектуальний метод виявлення цільових об'єктів предметної області для класифікації текстової інформації /

Р.В. Віт, О.В. Мазурець // Матеріали XIII Міжнародної науково-практичної конференції «Інформаційні управляючі системи та технології ІУСТ-ОДЕСА-2025». – Одеса, 24–26.09.2025. – 2025. – С. 119–121.

8. Mazurets O.V., Sobko O.V., Molchanova M.O., Zalutska O.O., Yurchak A.V. Practical Implementation of Neural Network Method for Stress Features Detection by Social Internet Networks Posts / O.V. Mazurets, O.V. Sobko, M.O. Molchanova, O.O. Zalutska, A.V. Yurchak // Global Science: Prospects and Innovations. Proceedings of the II International Scientific and Theoretical Conference «Scientific Review of the Actual Events, Achievements and Problems». – Berlin, Federal Republic of Germany, 31 May 2024. – International Center of Scientific Research, 2024. – P. 160–167.

9. Молчанова М.О., Мазурець О.В., Собко О.В., Віт Р.В., Назаров В.В. Алгоритм виявлення аб'юзивного вмісту в україномовному аудіоконтенті для імплементації в об'єктно-орієнтовану інформаційну систему / М.О. Молчанова, О.В. Мазурець, О.В. Собко, Р.В. Віт, В.В. Назаров // Вісник Хмельницького національного університету. Серія: Технічні науки. – Хмельницький, 2024. – № 1 (331). – С. 101–106.

10. Mazurets O., Molchanova M., Klimenko V., Prosvitliuk M. Practice Implementation of Neural Network Model BART-Large-CNN for Text Annotation / O. Mazurets, M. Molchanova, V. Klimenko, M. Prosvitliuk // Prospects of Scientific Research in the Conditions of the Modern World. Proceedings of XXVII International Scientific and Practical Conference. – Rotterdam, Netherlands, 12–14 June 2024. – 2024. – P. 97–102.

11. Mazurets O., Sobko O., Vit R., Pasternak V. Practical Approach for Detection by Deep Learning of Target Objects of Subject Area Based on Semantic Connectivity Indicators in Audio Database / O. Mazurets, O. Sobko, R. Vit, V. Pasternak // Proceedings of XXIV International Scientific and Practical Conference «Modern Scientific Challenges are the Driving Force of the Development of Scientific Research». – Bruges, Belgium, 22–24 May 2024. – International Scientific Unity, 2024. – P. 91–96.

12. Shevchuk P., Molchanova M., Mazurets O. Software for Text Messages Reliability Analysis Based on the Machine Learning Models Ensemble / P. Shevchuk, M. Molchanova, O. Mazurets // Proceedings of IV International Scientific and Practical Conference «Innovative Research and Perspectives of the Development of Science and Technology». – Stockholm, Sweden, 29–31 January 2024. – 2024. – P. 347–354.

13. Мазурець О.В., Козенко О.В., Собко О.В. Метод автоматизованого підбору відповідей на користувацькі запитання за семантичною подібністю / О.В. Мазурець, О.В. Козенко, О.В. Собко // Матеріали XII Всеукраїнської науково-практичної конференції «Глушковські читання». – Київ, 2023. – С. 106–109.

14. Albladi A., et al. Hate Speech Detection using Large Language Models: A Comprehensive Review / A. Albladi et al. // IEEE Access. – 2025. – P. 1. – URL: https://doi.org/10.1109/access.2025.3532397 (accessed 17.10.2025).

15. Ahmad M., et al. UA-HSD-2025: Multi-Lingual Hate Speech Detection from Tweets Using Pre-Trained Transformers / M. Ahmad et al. // Computers. – 2025. – Vol. 14, No. 6. – P. 239. – URL: https://doi.org/10.3390/computers14060239 (accessed 17.10.2025).

16. Kodali R.G., Manukonda D.P., Iglesias D. byteSizedLLM@ NLU of Devanagari Script Languages 2025: Hate Speech Detection and Target Identification Using Customized Attention BiLSTM and XLM-RoBERTa Base Embeddings / R.G. Kodali, D.P. Manukonda, D. Iglesias // In: Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025). – 2025. – P. 242–247. – URL: https://aclanthology.org/2025.chipsal-1.25/.

17. Hate Speech Detection Curated Dataset. Waalbannyantudre. Kaggle. URL: https://www.kaggle.com/datasets/waalbannyantudre/hate-speech-detection-curated-dataset.

18. Hate Speech and Offensive Language Detection. TheDevastator. Kaggle. URL: https://www.kaggle.com/datasets/thedevastator/hate-speech-and-offensive-language-detection.

19. E. A. Manziuk, O. V. Sobko, I. O. Podhorniuk, M. O. Molchanova, O. V. Mazurets. Multifactorial analysis of mobbing behavioral signs in educational environments posts by NLP means / E.A. Manziuk, O.V. Sobko, I.O. Podhorniuk, M.O. Molchanova, O.V. Mazurets // Journal of Physics: Conference Series. – 2025. – Vol. 3105, No. 1. – P. 012025. – DOI: 10.1088/1742-6596/3105/1/012025. – URL: https://iopscience.iop.org/article/10.1088/1742-

6596/3105/1/012025.

20. Mazurets O., Vit R., Molchanova M., Tymofiiev I., Sobko O. Context-enriched approach to students depression monitoring in education using BERT-GPT hybrid model / O. Mazurets, R. Vit, M. Molchanova, I. Tymofiiev, O. Sobko // CEUR Workshop Proceedings. – 2025. – Vol. 4096. – P. 167–176.

21. Molchanova M., Didur V., Sobko O., Mazurets O. Detection of Web Propaganda Patterns by Transformer Neural Networks: Improving Efficiency via Dataset Balancing / M. Molchanova, V. Didur, O. Sobko, O. Mazurets // CEUR Workshop Proceedings. – 2025. – Vol. 3988. – P. 112–126.

References

1. What is hate speech? United Nations. URL: https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech.

2. Mazurets O.V., Tymofiiev I.A., Klimenko V.I., Tyshchenko O.O. Metod vyiavlennia depresyvnoho stanu, poviazanoho iz navchanniam u zakladakh osvity, iz vykorystanniam neiromerezhi dualnoi arkhitektury / O.V. Mazurets, I.A. Tymofiiev, V.I. Klimenko, O.O. Tyshchenko // Visnyk Khersonskoho natsionalnoho tekhnichnoho universytetu. – 2024. – № 4 (91). – S. 311–318.

3. Murava V., Zalutska O., Didur V., Mazurets O. Software Architecture of Information System for Exchanging LLM Thematic Prompts / V. Murava, O. Zalutska, V. Didur, O. Mazurets // Global Trends in the Development of Information Technology and Science. Proceedings of IV International Scientific and Practical Conference. – Stockholm, Sweden, 25–27 June 2025. – P. 121–127.

4. Molchanova M.O., Mazurets O.V., Sobko O.V., Klimenko V.I., Androshchuk V.I. Metod neiromerezhevoho vyiavlennia kiberbulinhu z vykorystanniam khmarnykh servisiv ta obiektno-oriientovanoi modeli / M.O. Molchanova, O.V. Mazurets, O.V. Sobko, V.I. Klimenko, V.I. Androshchuk // Herald of Khmelnytskyi National University. Technical sciences. – Khmelnytskyi, 2024. – № 2 (333). – S. 200–206.

5. Molchanova M., Didur V., Sobko O., Mazurets O. Detection of Web Propaganda Patterns by Transformer Neural Networks: Improving Efficiency via Dataset Balancing / M. Molchanova, V. Didur, O. Sobko, O. Mazurets // CEUR Workshop Proceedings. – 2025. – Vol. 3988. – P. 112–126.

6. Sobko O., Mazurets O., Didur V., Chervonchuk I. Recurrent Neural Network Model Architecture for Detecting a Tendency to Atypical Behavior of Individuals by Text Posts / O. Sobko, O. Mazurets, V. Didur, I. Chervonchuk // Theoretical and Practical Aspects of Modern Research. Proceedings of XXVI International Scientific and Practical Conference. – Ottawa, Canada, 5–7 June 2024. – International Scientific Unity, 2024. – P. 113–117.

7. Vit R.V., Mazurets O.V. Metod vyiavlennia komunikatsiinykh obiektiv yak indykatoriv tsyfrovoi vtomy. Intelektualnyi metod vyiavlennia tsilovykh obiektiv predmetnoi oblasti dlia klasyfikatsii tekstovoi informatsii / R.V. Vit, O.V. Mazurets // Materialy XIII Mizhnarodnoi naukovo-praktychnoi konferentsii «Informatsiini upravliaiuchi systemy ta tekhnolohii IUST-ODESA-2025». – Odesa, 24–26.09.2025. – 2025. – S. 119–121.

8. Mazurets O.V., Sobko O.V., Molchanova M.O., Zalutska O.O., Yurchak A.V. Practical Implementation of Neural Network Method for Stress Features Detection by Social Internet Networks Posts / O.V. Mazurets, O.V. Sobko, M.O. Molchanova, O.O. Zalutska, A.V. Yurchak // Global Science: Prospects and Innovations. Proceedings of the II International Scientific and Theoretical Conference «Scientific Review of the Actual Events, Achievements and Problems». – Berlin, Federal Republic of Germany, 31 May 2024. – International Center of Scientific Research, 2024. – P. 160–167..

9. Molchanova M.O., Mazurets O.V., Sobko O.V., Vit R.V., Nazarov V.V. Alhorytm vyiavlennia abiuzyvnoho vmistu v ukrainomovnomu audiokontenti dlia implementatsii v obiektno-oriientovanu informatsiinu systemu / M.O. Molchanova, O.V. Mazurets, O.V. Sobko, R.V. Vit, V.V. Nazarov // Herald of Khmelnytskyi National University. Technical sciences. – Khmelnytskyi, 2024. – № 1 (331). – S. 101–106.

10. Mazurets O., Molchanova M., Klimenko V., Prosvitliuk M. Practice Implementation of Neural Network Model BART-Large-CNN for Text Annotation / O. Mazurets, M. Molchanova, V. Klimenko, M. Prosvitliuk // Prospects of Scientific Research in the Conditions of the Modern World. Proceedings of XXVII International Scientific

and Practical Conference. – Rotterdam, Netherlands, 12–14 June 2024. – 2024. – P. 97–102.

11. Mazurets O., Sobko O., Vit R., Pasternak V. Practical Approach for Detection by Deep Learning of Target Objects of Subject Area Based on Semantic Connectivity Indicators in Audio Database / O. Mazurets, O. Sobko, R. Vit, V. Pasternak // Proceedings of XXIV International Scientific and Practical Conference «Modern Scientific Challenges are the Driving Force of the Development of Scientific Research». – Bruges, Belgium, 22–24 May 2024. – International Scientific Unity, 2024. – P. 91–96.

12. Shevchuk P., Molchanova M., Mazurets O. Software for Text Messages Reliability Analysis Based on the Machine Learning Models Ensemble / P. Shevchuk, M. Molchanova, O. Mazurets // Proceedings of IV International Scientific and Practical Conference «Innovative Research and Perspectives of the Development of Science and Technology». – Stockholm, Sweden, 29–31 January 2024. – 2024. – P. 347–354.

13. Mazurets O.V., Kozenko O.V., Sobko O.V. Metod avtomatyzovanoho pidboru vidpovidei na korystuvatski zapytannia za semantychnoiu podibnistiu / O.V. Mazurets, O.V. Kozenko, O.V. Sobko // Materialy XII Vseukrainskoi naukovo-praktychnoi konferentsii «Hlushkovski chytannia». – Kyiv, 2023. – S. 106–109.

14. Albladi A., et al. Hate Speech Detection using Large Language Models: A Comprehensive Review / A. Albladi et al. // IEEE Access. – 2025. – P. 1. – URL: https://doi.org/10.1109/access.2025.3532397 (accessed 17.10.2025).

15. Ahmad M., et al. UA-HSD-2025: Multi-Lingual Hate Speech Detection from Tweets Using Pre-Trained Transformers / M. Ahmad et al. // Computers. – 2025. – Vol. 14, No. 6. – P. 239. – URL: https://doi.org/10.3390/computers14060239.

16. Kodali R.G., Manukonda D.P., Iglesias D. byteSizedLLM@ NLU of Devanagari Script Languages 2025: Hate Speech Detection and Target Identification Using Customized Attention BiLSTM and XLM-RoBERTa Base Embeddings / R.G. Kodali, D.P. Manukonda, D. Iglesias // In: Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025). – 2025. – P. 242–247. – URL: https://aclanthology.org/2025.chipsal-1.25/.

17. Hate Speech Detection Curated Dataset. Waalbannyantudre. Kaggle. URL: https://www.kaggle.com/datasets/waalbannyantudre/hate-speech-detection-curated-dataset.

18. Hate Speech and Offensive Language Detection. TheDevastator. Kaggle. URL: https://www.kaggle.com/datasets/thedevastator/hate-speech-and-offensive-language-detection.

19. E. A. Manziuk, O. V. Sobko, I. O. Podhorniuk, M. O. Molchanova, O. V. Mazurets. Multifactorial analysis of mobbing behavioral signs in educational environments posts by NLP means / E.A. Manziuk, O.V. Sobko, I.O. Podhorniuk, M.O. Molchanova, O.V. Mazurets // Journal of Physics: Conference Series. – 2025. – Vol. 3105, No. 1. – P. 012025. – DOI: 10.1088/1742-6596/3105/1/012025. – URL: https://iopscience.iop.org/article/10.1088/1742-6596/3105/1/012025.

20. Mazurets O., Vit R., Molchanova M., Tymofiiev I., Sobko O. Context-enriched approach to students depression monitoring in education using BERT-GPT hybrid model / O. Mazurets, R. Vit, M. Molchanova, I. Tymofiiev, O. Sobko // CEUR Workshop Proceedings. – 2025. – Vol. 4096. – P. 167–176.

21. Molchanova M., Didur V., Sobko O., Mazurets O. Detection of Web Propaganda Patterns by Transformer Neural Networks: Improving Efficiency via Dataset Balancing / M. Molchanova, V. Didur, O. Sobko, O. Mazurets // CEUR Workshop Proceedings. – 2025. – Vol. 3988. – P. 112–126.

Додаток К

Презентаційний матеріал

КВАЛІФІКАЦІЙНА РОБОТА МАГІСТРА

МЕТОД ВИЯВЛЕННЯ МОВИ ВОРОЖНЕЧІ У ЗАШУМЛЕНИХ СОЦІАЛЬНИХ ТЕКСТОВИХ ДАНИХ НЕЙРОМЕРЕЖЕВИМИ ЗАСОБАМИ

Виконав:
студент групи КНм-24-1
Ілля БОЯРЧУК

Керівник:
Ph.D., ст. викл. кафедри КН
Марина МОЛЧАНОВА

Актуальність

Актуальність теми визначається стрімким зростанням обсягу цифрових комунікацій, у межах яких мова ворожнечі поширюється не лише відкритими формулюваннями, а й через навмисно спотворені, суржикові, масковані та контекстно приховані мовні конструкції. У соціальних мережах, месенджерах і коментарних платформах агресивні висловлювання часто набувають форм, що унеможливлюють їх виявлення традиційними алгоритмами, орієнтованими на нормативно структуровані дані. Наявні методи автоматичного аналізу тексту не враховують системних проявів шуму, мовного змішування, орфографічної девіації та свідомого уникнення прямої лексичної агресії, що знижує їхню ефективність у практичних застосуваннях.

Зростання обсягу інформаційних потоків супроводжується появою нових стратегій мовного маскування, спрямованих на обходження автоматизованої модерації, що актуалізує потребу в методах, здатних інтерпретувати семантично агресивні висловлювання попри їхню формальну деформацію. Цифрові платформи стикаються з ризиками радикалізації, розпалювання ворожнечі та координації деструктивних комунікацій, що посилює суспільну та безпекову значущість створення стійких до шуму нейромережевих рішень. У цих умовах виникає наукова і практична необхідність побудови моделей, здатних виявляти приховані прояви агресії в динамічних, мовно нестабільних і свідомо викривлених текстах, що формують сучасне комунікаційне середовище.

Мета і задачі роботи

Об'єкт дослідження – процес автоматизованого виявлення мови ворожнечі у соціальних текстових даних із нестабільною мовною структурою та наявністю шумових викривлень.

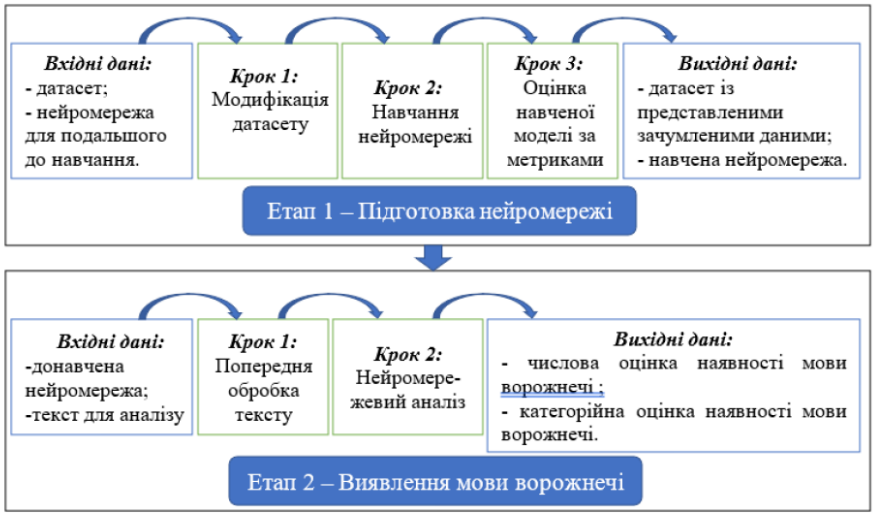
Предмет дослідження – моделі, методи та засоби обробки природної мови для автоматизованого виявлення мови ворожнечі у соціальних текстових даних із нестабільною мовною структурою та наявністю шумових викривлень.

Метою кваліфікаційної роботи магістра є підвищення точності виявлення мови ворожнечі у зашумлених соціальних текстових даних шляхом розроблення нейромережевого методу, здатного інтерпретувати спотворені, змішаномовні та навмисно масковані мовні конструкції.

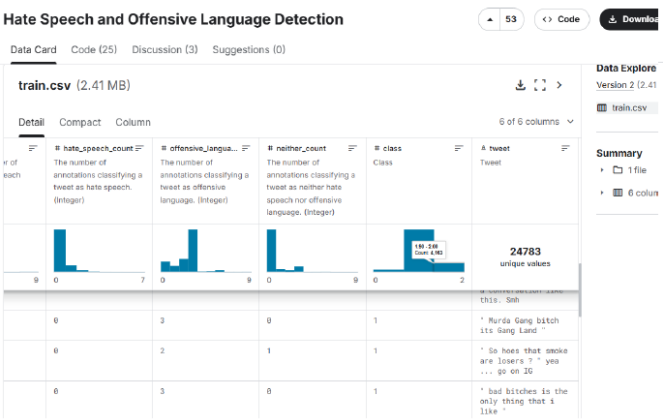
Для досягнення поставленої мети слід вирішити такі **завдання**:

- провести аналіз природи мови ворожнечі та її класифікаційних ознак;
- виконати огляд існуючих підходів до виявлення мови ворожнечі, виконати аналіз наукових досліджень;
- охарактеризувати етичні аспекти автоматизованого виявлення мови ворожнечі;
- розробити метод виявлення мови ворожнечі у зашумлених соціальних текстових даних нейромережевими засобами;
- виконати підготовку датасету для файнтюнінгу нейромережі для виявлення мови ворожнечі;
- виконати програмну реалізацію розробленого методу;
- провести дослідження методу виявлення мови ворожнечі у зашумлених соціальних текстових даних нейромережевими засобами.

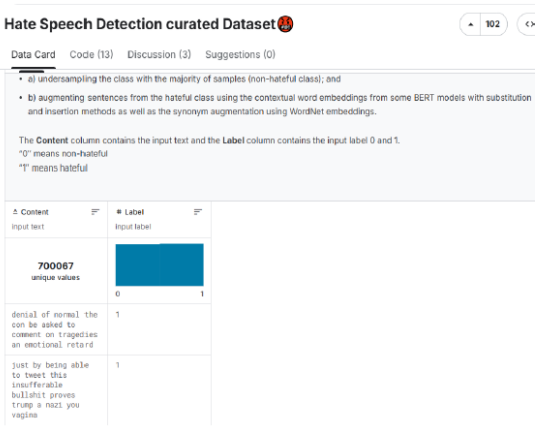
Етапи та кроки методу виявлення мови ворожнечі у зашумлених соціальних текстових даних



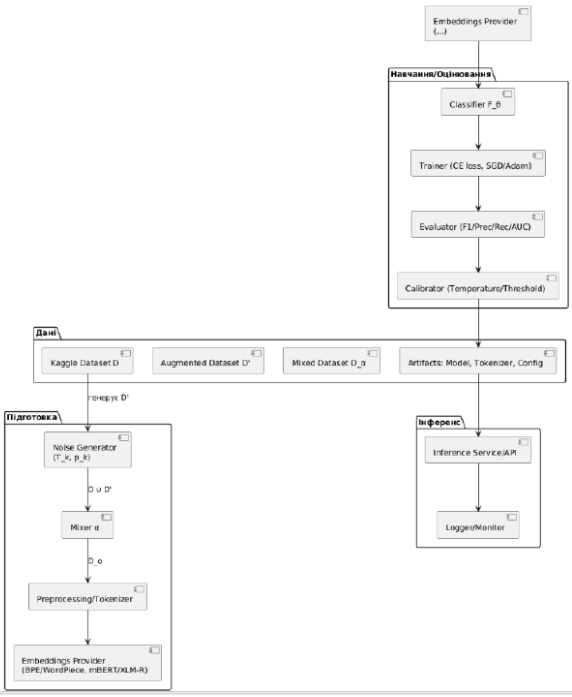
Датасет



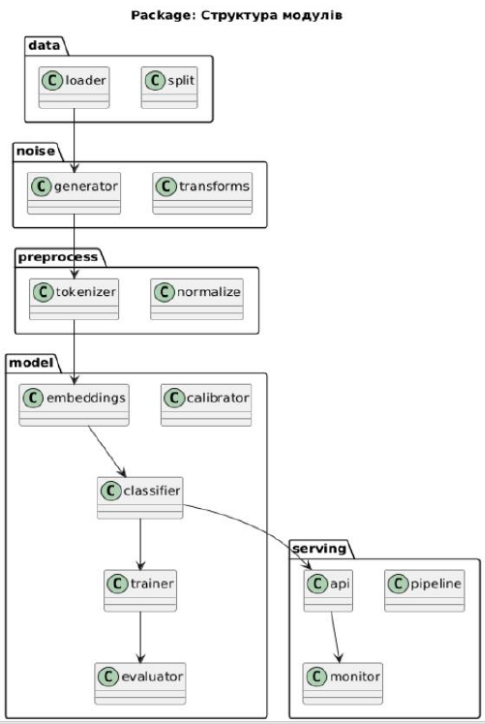
Датасет «Hate Speech and Offensive Language Detection»



Приклад даних з датасету «Hate Speech Detection curated Dataset»



Проєктування складових інтелектуальної системи



Програмна структура компонентів інтелектуальної системи

Діаграма пакетів

Програмна структура компонентів інтелектуальної системи

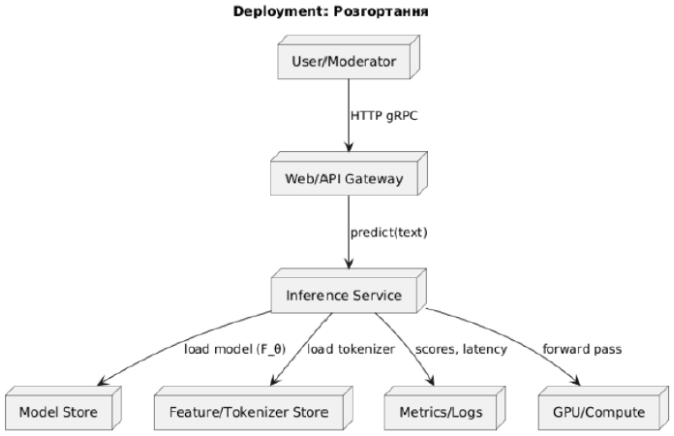
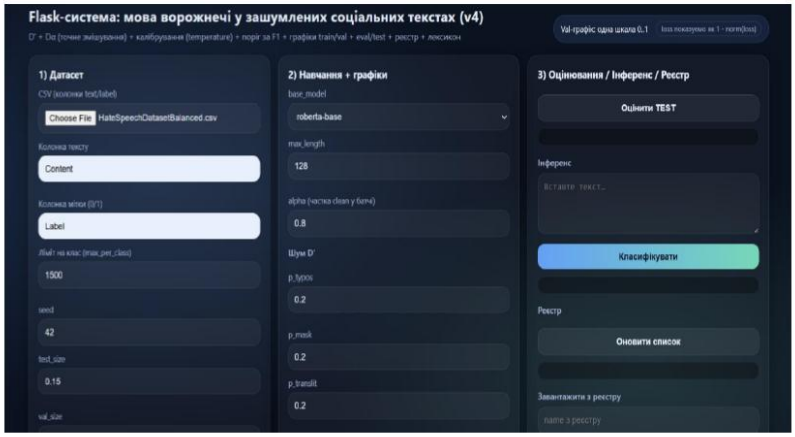


Схема розгортання інтелектуальної системи

Інтелектуальна система



Дослідження методу

Характеристики використаних даних та протокол обмеження

Показник	Значення
Загальна кількість використаних повідомлень	3000
Розподіл за класами після обмеження	1500 / 1500
Максимум прикладів на клас	1500
Частка тестової підвибірки	0,15
Частка валідаційної підвибірки	0,15
Зерно випадковості	42

Порівняння запусків за валідаційною F1-мірою

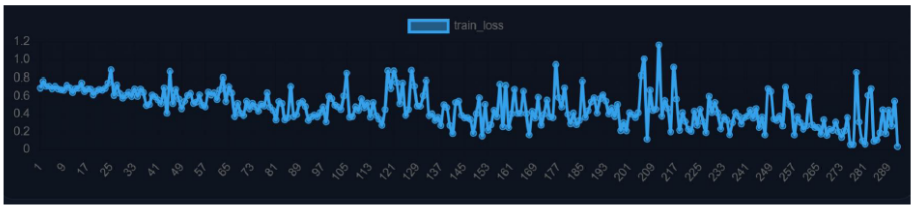
Архітектура	Частка чистих прикладів у мініпакеті	Температура калібрування	Поріг рішення	Валідаційна F_1-міра	Ідентифікатор запуску
roberta-base	0,6	1,636	0,50	0,8339	hate_noisy_run_20251214_181658
distilroberta-base	0,6	1,661	0,37	0,8086	hate_noisy_run_20251214_182101

Дослідження методу

Параметри та метрики найкращої конфігурації roberta-base

Показник	Значення
Архітектура	roberta-base
Максимальна довжина повідомлення	128
Частка чистих прикладів у мініпакеті	0,8
Найкраща епоха за F1 на валідації	3
Середня навчальна втрата (епоха 3)	0,2963
Валідаційна втрата (епоха 3)	0,5579
Валідаційна точність	0,82
Валідаційна точність позитивного класу	0,7707
Валідаційна повнота позитивного класу	0,9111
Валідаційна F1-міра	0,8350
Температура калібрування	2,1447
Підібраний поріг рішення	0,46
Ідентифікатор запуску	hate_noisy_run_20251214_183428

Дослідження методу



Графік навчальної втрати за кроками оптимізації

Графік валідації в одній шкалі: F1-міра та інвертована нормалізована валідаційна втрата

Висновки

Було досягнуто мету кваліфікаційної роботи магістра, а саме було підвищення точності виявлення мови ворожнечі у зашумлених соціальних текстових даних шляхом розроблення нейромережевого методу, здатного інтерпретувати спотворені, змішаномовні та навмисно масковані мовні конструкції.

Для досягнення поставленої мети було поставлено та вирішено такі завдання:

- проведено аналіз природи мови ворожнечі та її класифікаційних ознак;
- виконано огляд існуючих підходів до виявлення мови ворожнечі, виконано аналіз наукових досліджень;
- охарактеризовано етичні аспекти автоматизованого виявлення мови ворожнечі;
- розроблено метод виявлення мови ворожнечі у зашумлених соціальних текстових даних нейромережевими засобами;
- виконано підготовку датасету для файнтюнінгу нейромережі для виявлення мови ворожнечі;
- виконано програмну реалізацію розробленого методу;
- проведено дослідження методу виявлення мови ворожнечі у зашумлених соціальних текстових даних нейромережевими засобами.

ДЯКУЮ ЗА УВАГУ!

StrikePlagiarism.com

Документ прийнятий

Звіт подібності

Метадані

Назва організації	підрозділ
Khmelnytskyi National University	Кафедра комп'ютерних наук

Заголовок

КВАЛІФІКАЦІЙНА РОБОТА на тему Метод виявлення мови ворожнечі у зашумлених соціальних текстових даних нейромережевими засобами

Автор	Науковий керівник / Експерт
Ілля БОЯРЧУК	Марина МОЛЧАНОВА, Ph.D., ст. викл. кафедри КН

Кількість слів	Кількість символів	Дата звіту	Дата редагування	ІД документу
18035	151279	12/16/2025	12/16/2025	332880404

Обсяг знайдених подібностей

Коефіцієнт подібності визначає, який відсоток тексту по відношенню до загального обсягу тексту було знайдено в різних джерелах. Зверніть увагу, що високі значення коефіцієнта не автоматично означають плагіат. Звіт має аналізувати компетентна / уповноважена особа.

КП 1

18035
Кількість слів

КЦ

151279
Кількість символів

Тривога

У цьому розділі ви знайдете інформацію щодо текстових спотворень. Ці спотворення в тексті можуть говорити про МОЖЛИВІ маніпуляції в тексті. Спотворення в тексті можуть мати навмисний характер, але частіше характер технічних помилок при конвертації документа та його збереженні, тому ми рекомендуємо вам підходити до аналізу цього модуля відповідально. У разі виникнення запитань, просимо звертатися до нашої служби підтримки.

Заміна букв	ß	17
Інтервали	A→	0
Мікропробіли	¶	0
Білі знаки	ß	75
Парафрази (SmartMarks)	a	10

Джерела

Нижче наведений список джерел. В цьому списку є джерела із різних баз даних. Колір тексту означає в якому джерелі він був знайдений. Ці джерела і значення Коефіцієнту Подібності не відображають прямого плагіату. Необхідно відкрити кожне джерело і проаналізувати зміст і правильність оформлення джерела.

10 найдовших фраз	Колір тексту

ПОРЯДКОВИЙ НОМЕР	НАЗВА ТА АДРЕСА ДЖЕРЕЛА URL (НАЗВА БАЗИ)	КІЛЬКІСТЬ ІДЕНТИЧНИХ СЛІВ (ФРАГМЕНТІВ)
1	https://elar.khmnu.edu.ua/bitstreams/5738fc92-0211-483b-82ea-e040d41d4576/download	33 0.18 %
2	https://elar.khmnu.edu.ua/bitstreams/5738fc92-0211-483b-82ea-e040d41d4576/download	31 0.17 %
3	Метод нейромережевої ідентифікації переломів кісток нижніх кінцівок за рентгенівськими знімками [redacted] 12/18/2024 Khmelnytskyi National University (Кафедра комп'ютерних наук)	18 0.10 %

Tue Dec 16 09:49:41 EET 2025, Петровський Сергій Степанович, Хмельницький національний університет, ХНУ

Anti-Plagiarism (UA) v-15.284 Educational

The maximum coincidence with one document 1.0%

Dictionaries check: en_US, ru_RU, ua_UA. **Errors in the documents: 14%**

	Document		Sum coincidence on the DB	
ID: 253201 Title: КВАЛІФІКАЦІЙНА РОБОТА на тему Метод виявлення мови ворожнечі у зашумлених соціальних текстових даних нейромережевими засобами Added in a DB: 2025-12-16 Authors: Ілля БОЯРЧУК Heads: Марина МОЛЧАНОВА Consultants: Opponents:	Symbols	Lexemes	Symbols	Lexemes
	135680	939	3159 (2%)	49 (5%)

Plagiarism sources

ID	Description	Plagiarism presence in the document	
		Symbols	Lexemes

РІШЕННЯ ЕКСПЕРТНОЇ КОМІСІЇ КАФЕДРИ КОМП'ЮТЕРНИХ НАУК

ПРО ДОПУСК КВАЛІФІКАЦІЙНОЇ РОБОТИ ДО ЗАХИСТУ

Назва кваліфікаційної роботи *Метод виявлення мови ворожнечі у зашумлених соціальних текстових даних нейромережевими засобами*
Автор *студент групи КНм-24-1 Ілля БОЯРЧУК*
Освітня програма *Комп'ютерні науки*
Рівень вищої освіти *другий (магістерський)*
Спеціальність *122 – Комп'ютерні науки*
Науковий керівник: *Ph.D., ст. викл. каф. КН 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 Іллі Боярчука, 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: 2%,
- за системою StrikePlagiarism КП1: 4,2%, КП2: 1,1%.

16.12.2025

Завідувач кафедри _____ *Олександр БАРМАК*

Гарант освітньої програми _____ *Руслан БАГРІЙ*

Керівник кваліфікаційної роботи _____ *Марина МОЛЧАНОВА*

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
МОН УКРАЇНИ
кафедра комп'ютерних наук

ВІДГУК НАУКОВОГО КЕРІВНИКА

на кваліфікаційну роботу магістра

гр. КНм-24-1 Іллі БОЯРЧУКА за темою: *Метод виявлення мови ворожнечі у зашумлених соціальних текстових даних нейромережевими засобами.*

1. Актуальність обраної теми

З розвитком цифрових комунікацій мова ворожнечі у соціальних мережах та месенджерах стає складним соціальним викликом. Автоматизоване виявлення агресивних висловлювань у текстах, особливо за умов спотворення, транслітерації та маскування, є надзвичайно важливим для модерації контенту та забезпечення безпеки онлайн-середовищ. Розробка нейромережевих методів, здатних працювати з «за шумленими» даними, має значний науковий і практичний інтерес.

2. Відповідність роботи предметній області спеціальності 122 Комп'ютерні науки та загальним вимогам до наукових робіт

Робота повністю відповідає предметній області спеціальності 122 «Комп'ютерні науки», оскільки передбачає застосування нейромережевих моделей для обробки текстових даних, розробку методів їхньої адаптації до шумових спотворень та реалізацію програмної системи для практичного використання.

3. Професійні та особистісні якості магістранта

Магістрант продемонстрував високий рівень організованості та аналітичного мислення. Під час виконання роботи він виявив здатність самостійно опановувати сучасні нейромережеві технології та інтегрувати їх у комплексний метод, демонструючи відповідальність, уважність до деталей та прагнення до високої якості результатів.

4. Ступінь самостійності під час виконання кваліфікаційної роботи

Усі етапи дослідження виконані самостійно. Магістрант приймав ключові рішення щодо архітектури моделей, алгоритмів адаптації до спотворень та оцінки ефективності, що свідчить про високий рівень самостійності.

5. Наукова новизна та оригінальність запропонованих підходів

Наукова новизна полягає у створенні нейромережевого методу, адаптованого до спотворених і маскованих текстових даних, а також у впровадженні механізму автоматичного формування контрольовано зашумлених корпусів, що дозволяє підвищити стійкість моделі до мовних викривлень та прихованих агресивних висловлювань, що розширює можливості практичного застосування.

6. Ступінь оволодіння методами дослідження

Магістрант ефективно застосував сучасні методи обробки природної мови, нейромережеві моделі та алгоритми роботи з «зашумленими» даними. Він продемонстрував уміння інтегрувати різні підходи для підвищення стійкості та точності класифікації мови ворожнечі у соціальних текстах.

7. Повнота та якість розкриття теми роботи

Тема роботи розкрита комплексно: виконано огляд сучасних методів виявлення мови ворожнечі, розроблено нейромережевий метод для роботи з зашумленими текстовими даними, побудовано інтелектуальну систему та проведено експериментальне тестування її ефективності. Додатково проаналізовано етичні аспекти застосування методу, що підкреслює повноту та практичну цінність роботи.

8. Логічність, послідовність, аргументованість, літературна грамотність викладу матеріалу

Робота структурована логічно, послідовність викладення відповідає поставленій меті. Матеріал викладено зрозуміло, наведені експериментальні результати та їх аналіз, рисунки й таблиці оформлені відповідно до академічних стандартів.

9. Можливість практичного застосування кваліфікаційної роботи, окремих її частин

Розроблений метод може бути інтегрований у системи модерації контенту соціальних платформ, платформи для моніторингу соціальних медіа або аналітичні системи для дослідження мови ворожнечі. Адаптація до спотворених даних підвищує надійність класифікації та дозволяє застосовувати систему у реальних умовах онлайн-комунікацій.

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

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

Науковий керівник _____ Ph.D., ст. викл. кафедри КН Марина МОЛЧАНОВА

ВІДГУК ОПОНЕНТА

на кваліфікаційну роботу магістра

гр. КНм-24-1 Іллі БОЯРЧУКА за темою: *Метод виявлення мови ворожнечі у зашумлених соціальних текстових даних нейромережевими засобами.*

1. Актуальність обраної теми

Виявлення мови ворожнечі у соціальних мережах та цифрових платформах є надзвичайно актуальним завданням у сучасних умовах поширення онлайн-агресії та маскованих формулювань. Розробка нейромережевого методу, здатного працювати зі зашумленими та змішаномовними текстами, має практичне значення для підвищення ефективності модерації контенту та забезпечення безпеки комунікацій.

2. Відповідність роботи предметній області спеціальності 122 Комп'ютерні науки та загальним вимогам до наукових робіт

Тема роботи повністю відповідає предметній області спеціальності 122 «Комп'ютерні науки». Робота демонструє застосування трансформерних моделей, методів обробки природної мови, керованого формування навчальних корпусів та оцінки ефективності нейромереж у прикладній задачі виявлення мови ворожнечі.

3. Повнота розкриття мети та завдань дослідження

Мета та завдання дослідження чітко сформульовані та повністю виконані.

4. Наявність наукової новизни

Наукова новизна роботи полягає у поєднанні трансформерних моделей із керованим зашумленням текстових даних та автоматизованим формуванням контрольованих навчальних корпусів, що забезпечує підвищення стійкості моделей до маскованих та спотворених формулювань агресивного контенту та розширює практичні можливості інтелектуальних систем модерації.

5. Зміст кожного розділу роботи

Робота складається з чотирьох розділів. Перший розділ присвячено аналізу сучасного стану досліджень у сфері виявлення мови ворожнечі та етичних аспектів автоматизації. Другий розділ описує розроблений метод, етапи його реалізації та

підготовку датасету. Третій розділ містить проєктування інтелектуальної системи, включаючи компоненти, функціональні можливості та метрики оцінювання нейромережевої моделі. Четвертий розділ присвячено експериментальній перевірці методу та інтерпретації результатів.

6. Ступінь розкриття теми роботи

Тема роботи розкрита всебічно: виконано огляд літератури, розроблено метод виявлення мови ворожнечі у зашумлених соціальних текстових даних, проведено експериментальні дослідження та оцінку ефективності системи, враховано практичні та етичні аспекти використання.

7. Якість оформлення кваліфікаційної роботи

Робота оформлена грамотно, розділи структуровані логічно, таблиці та рисунки наочно ілюструють результати досліджень. Оформлені посилання на джерела відповідають академічним стандартам.

8. Недоліки кваліфікаційної роботи

Рекомендується розширити перевірку моделі на мультидоменних та мультимовних корпусах для підвищення узагальнюваності, а також провести абляційний аналіз впливу окремих типів шумів на якість класифікації.

9. Загальний висновок (допускається чи не допускається до захисту), якої оцінки заслуговує кваліфікаційна робота

Робота виконана на високому науковому та практичному рівні, поставлені завдання реалізовані, наукова новизна підтверджена. Роботу можна допустити до захисту, рекомендована оцінка – відмінно.

Опонент (прізвище, ім'я, по батькові, посада, місце роботи)

Тимчук Олизавета Теннадіївна, д.т.н., професор

«15» 12 2025 р професор каф. КІС

підпис