РОЗДІЛ 8
СТРУКТУРНА, ПРИКЛАДНА ТА МАТЕМАТИЧНА ЛІНГВІСТИКА

УДК 81`32:004.9(510)
DOI https://doi.org/10.32782/tps2663-4880/2024.38.42

РОЗРОБКА ІНСТРУМЕНТІВ АВТОМАТИЗОВАНОГО АНАЛІЗУ ДАНИХ ДЛЯ ІДЕНТИФІКАЦІЇ ВЕРБАЛЬНИХ ЗАСОБІВ РЕАЛІЗАЦІЇ МОВИ ВОРОЖНЕЧІ У КИТАЙСЬКОМОВНОМУ МЕДІАДИСКУРСІ: КІЛЬКІСНИЙ І ЯКІСНИЙ ПІДХІД

DEVELOPMENT OF TOOLS FOR AUTOMATED DATA ANALYSIS TO IDENTIFY VERBAL MEANS OF HATE SPEECH IMPLEMENTATION IN CHINESE-LANGUAGE MEDIA DISCOURSE: A QUANTITATIVE AND QUALITATIVE APPROACH

Калінін В.С.,
orcid.org/0009-0003-9692-0978
аспірант кафедри східної і слов'янської філології
Київського національного лінгвістичного університету

У статті досліджено сучасні інструменти автоматизованого аналізу текстів для ідентифікації вербальних засобів реалізації мови ворожнечі в китайськомовному медіадискурсі. Розглянуто основні теоретичні засади, зокрема визначення мовленнєвої агресії, її класифікацію за різними критеріями та особливості вираження в онлайн-просторі. Проаналізовано правила користування популярними соціальними платформами на предмет регулювання та модерації контенту із ознаками ненависті. Обґрунтовано необхідність створення ефективних механізмів для виявлення негативно забарвленої лексики, враховуючи зростання обсягів інформації та складність контролю її поширення у цифровому середовищі.

У праці було представлено програмне забезпечення OSKAL (Opensource Social Knowledge of Aggressive Language), що дозволяє систематизувати збір, очищення, обробку та подальший аналіз текстових одиниць. Архітектуру системи розроблено з використанням модульного підходу; вона складається із серверної частини для інтерпретації даних і клієнтського інтерфейсу для візуалізації результатів. Для збору інформації застосовано метод веб-скрапінгу, який реалізовано через фреймворк Scrapy. Очищені дані класифіковано за допомогою мовної моделі thu-coai/roberta-base-cold, яка натренована на виявлення образливої лексики у китайськомовних матеріалах.

Для апробації методики використано китайськомовну онлайн-платформу Weibo, з якої було зібрано 18632 дописи та 234944 коментарі. На основі цих даних протестовано ефективність мовної моделі thu-coai/roberta-base-cold, яка виявила ознаки агресії у 1868 дописах та 38933 коментарях із точністю класифікації 82,75%. Розкрито можливості моделі в аспектах ідентифікації дискримінаційного контенту, її обмеження при аналізі емоційно забарвлених висловлювань, а також перспективи для подальшого вдосконалення.

Розроблена методика дозволила якісно проаналізувати великий обсяг текстових даних та забезпечити високу точність ідентифікації мови ворожнечі. Отримані результати сприятимуть створенню ефективних механізмів моніторингу та протидії ненависницькій риториці в китайськомовному сегменті інтернету.

Ключові слова: мова ворожнечі, китайськомовна соціальна мережа, автоматизований аналіз даних, мовна модель, веб-скрапінг.

The article explored modern tools for automated text analysis to identify verbal means of hate speech implementation in Chinese-language media discourse. Key theoretical foundations, including the definition of speech aggression, its classification based on various criteria, and its specific features in the online space, were examined. The content moderation and regulation policies of popular social platforms concerning hateful content were analyzed. Considering the increasing volume of information and the challenges in controlling its dissemination in the digital environment, the necessity of creating effective mechanisms for detecting negatively charged language was substantiated.

The study introduced the OSKAL software (Opensource Social Knowledge of Aggressive Language), which enables the systematic collection, cleaning, processing, and further analysis of textual units. The system's architecture was developed using a modular approach. It consists of a server-side component for data processing and a client interface for visualizing results. The method of web scraping, implemented via the Scrapy framework, was employed for data collection. The cleaned data were classified using the thu-coai/roberta-base-cold language model. This model was trained to detect offensive language in Chinese-language materials.

For methodology testing, data were collected from the Chinese-language online platform Weibo, comprising 18,632 posts and 234,944 comments. Based on this dataset, the thu-coai/roberta-base-cold language model was evaluated, detecting signs of aggression in 1,868 posts and 38,933 comments with a classification accuracy of 82.75%. The

model's capabilities in identifying discriminatory content, its limitations in analyzing emotionally charged statements, and prospects for further improvement were identified.

The developed methodology enabled the qualitative analysis of a large volume of textual data and ensured high accuracy in the identification of hate speech. The obtained results will contribute to the creation of effective mechanisms for monitoring and countering hate rhetoric in the Chinese-language segment of the internet.

Key words: hate speech, Chinese social network, automated data analysis, language model, web scraping.

Постановка проблеми. На сучасному етапі дослідження засобів реалізації мови ворожнечі прослідковується зміщення фокусу від традиційних методів контент-аналізу до автоматизованих систем обробки даних. Стрімке зростання обсягів інформації в медіапросторі та збільшення випадків використання агресивних висловлювань в соціальних мережах створили потребу в ефективних механізмах для виявлення й ідентифікації відповідних дискримінаційних маркерів. Особливо гостро ця проблема постає при вивченні китайськомовного медіадискурсу, де інструменти, що існують, часто не враховують специфіку китайської мови та знижують точність отриманих результатів. Відсутність систематизованого опису наявної методології створює необхідність у визначенні найбільш ефективних засобів автоматизованого аналізу і розробці рекомендацій щодо їх використання.

Аналіз останніх досліджень і публікацій. Проблема автоматизованого аналізу мови ворожнечі в медіадискурсі привертає увагу багатьох дослідників. Оцінка наявних інструментів для збору даних із соціальних мереж для їхнього подальшого застосування як емпіричної бази соціологічних досліджень проводилася у роботах вітчизняних науковців Кислової О., Кузіної І., Дирди І. [1], Міщенко Л. [2], Молчанової М. [3], Рудніченко М., Шибаєвої Н., Отрадської Т., Шведова Д., Шпінаревої І., Петрова І. [4], Рабчуна Д., Тищенка В., Голобородька С. [5], Собко О. [6], Браіловського М., Хорошка В. [7] та ін. Особливості використання систем машинного навчання разом з нейронними мережами для ідентифікації та подальшої класифікації ворожих висловлювань вивчали такі зарубіжні спеціалісти, як MacAvaney S., Yao H. R., Yang E., Russell K., Goharian N., Frieder O. [8], Vrysis L., Vryzas N., Kotsakis R., Saridou T., Matsiola M., Veglis A., Arcila-Calderón C., Dimoulas C. [9], Mollas I., Chrysopoulou Z., Karlos S., Tsoumakas G. [10], Röttger P., Vidgen B., Nguyen D., Waseem Z., Margetts H., Pierrehumbert J. B. [11], Raju E., Saketh R., Krishna O. K. V., & Kushanu P. G. [12] та ін. Інтеграцію генеративних методів штучного інтелекту, насамперед великих мовних моделей, для вияву і запобігання поширення образливої лексики на базі китайської мови в інтернеті розглядали Chao A. F., Wang C. S., Li B. Y., Chen H. Y. [13], Zhang Y., Zhong T., Yi T., Li H. [14], Wang C. C., Day M. Y., Wu C. L. [15], Kim J. Y., Kesari A. [16], Li J., Ning Y. [17] та ін.

Метою статті є розробка та впровадження інструментів автоматизованого аналізу даних для виявлення мови ворожнечі в китайськомовному медіадискурсі.

Для досягнення поставленої мети визначено такі **завдання**:

1) проаналізувати теоретичні підходи до визначення мови ворожнечі та особливості її реалізації в медіапросторі;

2) розробити програмне забезпечення для автоматизованого збору й аналізу даних на прикладі китайськомовної платформи Weibo;

3) впровадити використання мовної моделі для класифікації текстів на предмет наявності ворожої лексики;

4) оцінити ефективність розробленого інструментарію на основі аналізу зібраних даних.

Виклад основного матеріалу. Протягом останнього десятиліття феномен мови ворожнечі перебуває у центрі уваги як у суспільстві, так і серед дослідників. Це явище охоплює різноманітні форми вираження, які містять нетерпимість, агресію, образливі або принизливі порівняння щодо окремих осіб або груп [1, с. 254]. Основна проблема полягає в тому, що універсального визначення мови ворожнечі досі не існує, а її межі часто залежать від культурного контексту, суспільних норм, індивідуального сприйняття. Деякі висловлювання можуть бути сприйняті як дискримінаційні одними людьми, але залишатися нейтральними або навіть прийнятними для інших. Така невизначеність ускладнює аналіз та розробку систем для ідентифікації і протидії риториці ненависті [8, с. 4].

Дослідниці Богданова І. та Лептуга О. зазначають, що словосполучення *мова ворожнечі* є аналогом англійського терміна *hate speech*, яке вживається паралельно з подібними, але не тотожними поняттями, такими як словесний екстремізм, мовленнєва агресія, мовна демагогія, мовний конфлікт, насилля чи маніпуляція [18, с. 7]. Колесник Г. пропонує відрізняти мову ворожнечі від нецензурної лайки, образливої лексики чи кривдження. Це, скоріше, агресивна лексика, спрямована на

належність людини до певної соціальної групи за такими ознаками, як політичні, релігійні, етнічні, сексуальні тощо [19, с. 279]. Олейник А. підкреслює, що мова ворожнечі виражає сексизм, расизм, шовінізм, ейджизм, гомофобію, ксенофобію, газлайтінг і класифікує цей феномен як девіантну поведінку, що прирівнюється до таких явищ, як розпалювання ненависті, підбурювання до тероризму та геноциду [20, с. 119].

Погоджуємося зі спостереженнями українських науковців і зазначаємо, що така багатозначність і складність концепції мови ворожнечі створюють значні труднощі для розробки механізмів автоматизованого аналізу цього явища. Визначення чітких критеріїв для ідентифікації негативно забарвленої лексики ускладнюється через варіативність її виявів, що може охоплювати як відверті, так і приховані форми агресії, іронію чи маніпулятивні висловлювання. Така невизначеність ускладнює побудову ефективних систем для виявлення подібного контенту, оскільки алгоритми потребують однозначних і структурованих даних для навчання.

Онлайн середовище значно посилило вплив мови ворожнечі, надавши їй нові платформи для поширення. Соціальні мережі, месенджери, форуми, відеохостинги та інші цифрові ресурси дають змогу швидко і масштабно транслювати контент, разом із агресивними висловлюваннями. Це створює надзвичайну ситуацію, коли цифровий простір стає ареною для розповсюдження стереотипів чи дискримінації, що має значні соціальні, політичні та психологічні наслідки [1, с. 254]. Незважаючи на зусилля платформ щодо регулювання контенту, мова ненависті знаходить способи обходити ці обмеження, використовуючи приховані форми вираження або граючи на тонких нюансах спілкування, особливо, коли це стосується китайськомовного медіадискурсу.

У стандартах Facebook, Instagram, Threads *мова ворожнечі* визначається як прямі напади на людей – а не на концепції чи інституції – на основі раси, етнічного або національного походження, інвалідності, релігійної приналежності, касти, сексуальної орієнтації, статі або гендерної ідентичності. Ці платформи забороняють використання образливих стереотипів, зневажливих висловлювань, принижень, а також закликів до залякування чи дискримінації, спрямованих на окремих осіб або групи. Спільноти активно протидіють подібній агресії, адже вона створює середовище залякування та відчуження, а в деяких випадках може сприяти насильству поза мережею. Проте модератори враховують контекст, щоб дозволити використання висловлювань у сатиричному, інформативному або самовиражальному сенсі, якщо наміри автора є зрозумілими [21].

На противагу чітко визначеним і доступним правилам англійськомовних платформ, китайськомовні можуть мати менш деталізовані норми, які регулюють протидію онлайн агресії. Наприклад, у соціальній мережі Weibo відсутні прямі згадування мови ворожнечі чи суміжних тем. У правилах користування зазначено, що користувачам забороняється використовувати платформу для незаконних або неприязних дій, публікувати нерелевантний чи шкідливий контент, а також порушувати права інших вдаючись до копіювання або використання чужого матеріалу [22]. Однак угода не містить конкретних положень, що стосуються мови ненависті, її заборони або відповідальності за такі порушення.

Платформа WeChat є більш послідовною і характеризує мову ворожнечі як нападки на людей через вік, расу, національність, релігію, стать, сексуальну орієнтацію та інші ознаки. Забороняється публікація будь-якого контенту, який можна трактувати як образливий, расистський або етнічний, наклепницький, принижуючий або такий, який містить або закликає до дискримінації певної групи осіб чи суспільства в цілому. Це розповсюджується не лише на образливу мову, але й зображення, що можуть бути дегуманізуючими, зокрема щодо жертв злочинів на ґрунті ненависті [23]. Отже, соціальна мережа вживає заходів для обмеження поширення цього негативного феномена, хоча застосування таких правил залишається в компетенції адміністрації платформи, що дозволяє їй вносити винятки для випадків, коли контент не порушує основні норми поведінки.

Автоматизація процесів ідентифікації агресивної лексики є складним завданням. Основна проблема полягає в тому, що наявні алгоритми часто базуються на різних наборах даних, які використовують відмінні визначення мови ворожнечі. Відсутність уніфікованих стандартів та обмежена деталізація правил на платформах, як-от Weibo чи WeChat, ускладнюють порівняння результатів і оцінку ефективності таких систем. Крім того, мовна багатозначність створює додаткові виклики, особливо коли ненависть маскується під гумор чи іронію [8, с. 3].

Зважаючи на різноманіття підходів до розуміння мови ворожнечі, пропонуємо власне визначення її як форми комунікації, що виражає ненависть або дискримінацію до осіб чи груп на основі їхніх реальних або припущених характе-

ристик, таких як етнічна належність, стать, гендер, релігійні переконання чи соціальний статус. Особливість цього феномену в сучасному цифровому середовищі полягає у здатності адаптуватися до алгоритмів модерації та маскувати агресію за допомогою іронії, сарказму, візуального контенту чи кодових слів. Вона також може виявлятися через меми, відео або тексти, які формують негативні стереотипи чи поширюють упередження у вигляді, який дозволяє уникнути очевидного порушення правил платформ. Таким чином, подібне вербальне насильство можна вважати динамічним явищем, що не лише підриває принципи рівності і поваги, але й активно модифікується під впливом культурного контексту, соціальних трендів та технологічних можливостей.

Для досягнення мети дослідження, було розроблено програмне забезпечення під назвою OSKAL, яке дозволяє масовий збір і аналіз публічно доступних даних, на прикладі китайськомовної соціальної мережі Weibo. Таке найменування не лише є абревіатурою, що розшифровується як *Opensource Social Knowledge of Aggressive Language*, а й має метафоричне значення. Слово *оскал* асоціюється з виразом обличчя, який символізує приховану загрозу.

Запропоноване технологічне рішення реалізоване мовою програмування Python і складається з двох основних компонентів: серверної частини та клієнтського інтерфейсу (Рис. 1).

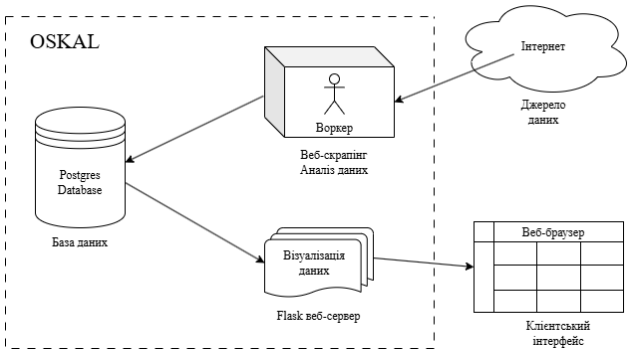

Рис. 1. Архітектура системи

До серверної частини належать база даних, веб-сервер і програма *воркер* (англ. worker), яка виконує обробку та збір інформації. Для взаємодії зі сховищем було обрано бібліотеку SQLAlchemy – ORM (Object-Relational Mapper), що забезпечує зручний і простий доступ до контенту.

Для веб-скрапінгу (сканування та збору інформації з веб-сайтів) використано фреймворк Scrapy, який дозволяє створювати ботів (так звані *спайдери*). Розробивши спеціалізований код, було забезпечено роботу спайдера, який отримував, обробляв і зберігав відомості з онлайн-сторінок в автоматизованому режимі.

Зібрані ботом вихідні дані обробляються модулем нормалізації результатів, який відповідає за очищення та видалення наявного HTML-коду, скриптів, емодзі, інших метатегів.

Очищені матеріали передаються до великої мовної моделі, яка натренована на виявлення ознак ворожості в текстах. У роботі було використано thu-coai/roberta-base-cold, яка доступна на платформі Hugging Face [24]. Вона побудована на основі трансформерної архітектури BERT (Bidirectional Encoder Representations from Transformers) – нейромережевого методу обробки природної мови, що дозволяє аналізувати значення слів з урахуванням всього контексту речення.

Модель навчена на китайськомовному корпусі COLD (Chinese Offensive Language Dataset) – першому загальнодоступному наборі даних для виявлення образливої лексики китайською мовою, що містить 37480 розмічених речень з різними типами упереджень щодо раси, статі, регіону. Корпус було сформовано шляхом збору дописів із різних китайських соціальних платформ, зокрема Zhihu і Weibo, з подальшою фільтрацією та розміткою. Навчальний набір (32157 прикладів) створювався напівавтоматично: спочатку базовий алгоритм маркував тексти, а потім розмітка перевірялась вручну анотаторами. Модель здійснює бінарну класифікацію, розподіляючи висловлювання на два класи – ворожі або нейтральні, і досягає 82,75% точності. Макро-середня оцінка F1 на тестовому наборі є гармонійним середнім між влучністю прогнозів та часткою справжніх позитивних результатів і складає 82,39%. Завдяки можливостям цього лінгвістичного інструмента, було автоматизовано ідентифікацію агресивної мови у зібраній статистиці.

Описані вище етапи обробки контенту виконуються програмою воркер, яка запрограмована на щоденний запуск і здійснює повний цикл збору та аналізу нових дописів з соціальної мережі Weibo.

Для дослідження було обрано саме цю спільноту, оскільки вона надає доступ через API (інтерфейс програмування додатків), що дозволяє автоматично звертатися до її репозиторію та отримувати публічно доступну інформацію. Воркер здійснював запити до певних ендпойнтів (англ. endpoints), що є точками доступу для отримання даних у вигляді JSON-об'єктів, і проводив збір й обробку матеріалів без необхідності вручну взаємодіяти з платформою. Зокрема, він опрацьовував дописи користувачів, їхні коментарі до них разом з іншими метаданими.

Отже, один цикл воркера проходить у декілька етапів.

На першому етапі здійснюється збір інформації:

1) завантаження списку категорій мережі;

2) завантаження дописів з кожної з категорій разом з їхніми унікальними ідентифікаторами і часом публікації;

3) завантаження коментарів до допису разом з їхніми унікальними ідентифікаторами і часом публікації.

На другому етапі відбувається очищення даних від HTML-коду для видалення візуальних елементів допису, оскільки мета дослідження зосереджена на вивченні саме вербальної складової.

Третій етап полягає в обробці текстових даних за допомогою тренованої мовної моделі thu-coai/roberta-base-cold, яка класифікує дописи на нейтральні (0) і негативні (1) з вказівкою відповідної відсоткової ймовірності.

На четвертому етапі зібрана інформація, що містить вихідні дописи і коментарі з їхніми ідентифікаторами та категоріями Weibo, очищений текст, час публікації, результати класифікації, додається до бази даних.

Підсумкові дані дослідження відображаються у клієнтському інтерфейсі, що є HTML-представленням результатів на веб-сторінці з використанням мікрофреймворка Flask.

Варто зазначити, що у процесі дослідження не збиралися персональні дані користувачів. Для отримання й аналізу контенту було використано методи OCINT (Open Source Intelligence), що надають можливість отримувати публічно доступні деталі з відкритих джерел без порушення приватності.

У період з 14 жовтня 2024 року по 14 січня 2025 року запрограмований воркер зібрав 18632 дописи та 234944 коментарі до них із 11 категорій платформи Weibo, а саме: «Популярне», «Чарти», «Місто», «Суспільство», «Наука і техніка», «Зірки», «Фільми», «Музика», «Цифровий контент», «Транспорт», «Ігри». За допомогою мовної моделі, було виявлено ознаки ворожої лексики у 1868 дописах та 38933 коментарях, а також визначено відсоткову ймовірність наявності агресії у цих текстах.

Критерії відбору дописів були розроблені з урахуванням специфіки дослідження та технічних можливостей платформи Weibo. Вибірка охоплювала:

1) публічні дописи з відкритих акаунтів опублікованих у визначений період дослідження;

2) дописи з текстовим наповненням (без урахування таких, що містили лише зображення чи відео);

3) дописи з не менш ніж 5 коментарями для забезпечення контекстуального аналізу.

Для класифікації речень як ворожих було встановлено поріг ймовірності на рівні 90%. Цей показник було обрано емпірично після попереднього тестування моделі на пілотній вибірці з 1000 дописів. При нижчому порозі (75–85%) вона демонструвала значну кількість хибнопозитивних результатів, маркуючи як насильницькі звичайні емоційні висловлювання. Натомість поріг вище 95% призводив до пропуску текстів з ознаками прихованої агресії.

Наведемо приклад аналізу одного з ворожих дописів, що було розміщено на платформі Weibo.

Оригінальний текст:

好烦，自己微博日常吐槽一下工页大模底好多编造的段子，怎么被转到男博主那边了？有些男的满身赛博爹味来指指点点，跟有病似的。哦对了，他们的价值真的蛮低，因为只转发不点赞。拉黑拉黑，全部拉黑。

Переклад:

«*Так набридло! Я, зазвичай, висловлюю своє невдоволення з приводу того, чому на платформі Weibo так багато вигаданих жартів. Але чому вони всі розміщуються на сторінках блогерів чоловіків? Деякі чоловіки притримуються кібер-батьківського стилю і хизуються цим, ніби у них проблеми з головою. О, до речі, їхня цінність й справді низька, оскільки вони тільки роблять перепост контенту і навіть не ставлять вподобайку! Блокую всіх, всіх у чорний список*».

У тексті автор демонструє мову ворожнечі, яка виявляється у приниженні, дискримінації та негативному ставленні до конкретної групи людей – блогерів чоловічої статі, які певним чином поводяться в інтернеті. Вираз *кібер-батьківський стиль* підкреслює стереотипне уявлення про чоловіків із домінуючою або авторитарною поведінкою в онлайн-просторі. Фраза *ніби у них проблеми з головою* є прямим вираженням вербальної агресії, спрямованої на приниження блогерів, а твердження про їхню *низьку цінність* додає зневажливий тон.

Результати дослідження демонструють, що використана мовна модель thu-coai/roberta-base-cold, попри високі показники точності на тестовому наборі (82,75%), має певні обмеження при роботі з текстами соціальних мереж. Аналіз зібраних даних показав, що модель часто класифікує як ворожі висловлювання, які містять емоційне забарвлення (обурення, роздратування, засудження тощо), але не мають ознак дискримінації чи агресії щодо певних соціальних груп. Це може бути пов'язано з особливостями навчаль-

ного корпусу COLD, який, хоча і містить різноманітні приклади упереджень, може не повністю відображати специфіку сучасного китайськомовного медіадискурсу.

Порівнявши отримані результати з іншими дослідженнями у сфері автоматизованого аналізу мови ворожнечі, можемо зазначити певні спільні тенденції і відмінності. Так, Chao et al [13, с. 11] у роботі також використовували модель BERT для аналізу китайськомовних текстів і досягли точності 94,42% у першій фазі та 81,48% у другій фазі виявлення мови ненависті, що вище за показники нашої моделі (82,75%). Їхній підхід фокусувався специфічно на виявлення вербальної дискримінації в контексті пандемії COVID-19, що використовував генеративний штучний інтелект для перефразування негативних висловлювань, тоді як пропонована розвідка охоплює ширший спектр виявів агресивної лексики у різних контекстах китайськомовного медіадискурсу. Zhang et al [14, с. 13774] також розробили власний спеціалізований доменно-орієнтований двоетапний підхід для мультидоменної класифікації на наборі масиву з 20000 текстів, що демонструє ефективність використання великих мовних моделей для виявлення словесного насильства в китайськомовному середовищі.

Розроблене програмне забезпечення OSKAL продемонструвало ефективність у автоматизованому зборі й обробці даних з платформи Weibo. Проте детальний аналіз результатів вказує на необхідність вдосконалення механізмів класифікації і розширення критеріїв оцінки текстового контенту.

Висновки. У результаті проведеного дослідження було досягнуто поставленої мети щодо розробки та тестування інструментів автоматизованого аналізу даних для виявлення мови ворожнечі в китайськомовному медіадискурсі. Для вирішення поставлених завдань розроблено програмне забезпечення OSKAL, яке дозволяє здійснювати автоматизований збір й обробку публічно доступних даних з китайськомовної соціальної мережі Weibo. Впроваджено використання мовної моделі thu-coai/roberta-base-cold для класифікації текстів на предмет наявності принизливої лексики, що дозволило розглянути значний обсяг даних.

За період дослідження було проаналізовано понад 250 тисяч текстових одиниць (18632 дописи та 234944 коментарі), з яких 10,03% дописів та 16,59% коментарів містили потенційні ознаки агресивної комунікації.

У роботі було визначено обмеження наявних підходів до автоматизованого аналізу мови ворожнечі, зокрема складність розрізнення емоційно забарвлених висловлювань та справжніх виявів дискримінації. Створено базу розмічених прикладів риторики ненависті з китайськомовного сегменту соціальних мереж, яка може бути використана для подальших досліджень та розробки більш точних алгоритмів класифікації.

Перспективу подальших досліджень вбачаємо в розробці та тренуванні власної мовної моделі, здатної не лише визначати наявність мови ворожнечі, але й класифікувати її за типами (расова дискримінація, сексизм, ейджизм тощо).

Реалізація цих напрямів дозволить створити більш ефективні інструменти для моніторингу та протидії мові насилля в онлайн-просторі.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ:

1. Кислова О., Кузіна І., Дирда, І. Дослідження онлайнової мови ворожнечі щодо ромської меншини в українському інтернет-просторі. *Ideology and Politics Journal*. 2020. Вип. 2. Ч. 16. С. 252–278. DOI: https://doi.org/10.36169/2227-6068.2020.01.00021.

2. Міщенко Л. Д. Метод розпізнавання фейкових новин у мережі інтернет на основі обробки природної мови : дис. … д-ра філософії : 123. Київ, 2024. 232 с. URL: https://ela.kpi.ua/handle/123456789/68509 (дата звернення: 20.01.2025).

3. Молчанова М. Метод виявлення та класифікації технік пропаганди у текстовому контенті засобами штучного інтелекту. *Розвитки інформаційно-керуючих систем та технологій* : монографія. Львів-Торунь : Liha-Pres, 2024. С. 245–266. URL: https://files.znu.edu.ua/files/Bibliobooks/Inshi80/0060461.pdf#page=245 (дата звернення: 20.01.2025).

4. Рудніченко М., Шибаєва Н., Отрадська Т., Шведов Д., Шпінарева І., Петров І. Інтелектуальна система аналізу та детекції текстового бот-контенту великого обсягу у соціальних мережах. *Розвитки інформаційно-керуючих систем та технологій* : монографія. Львів-Торунь : Liha-Pres, 2024. С. 197–222. URL: https://files.znu.edu.ua/files/Bibliobooks/Inshi80/0060461.pdf#page=197 (дата звернення: 20.01.2025).

5. Рабчун Д. І., Тищенко В. С., Голобородько С. О. Ефективне розпізнавання дезінформації за допомогою нейронних мереж: фокус на виявленні емоційного впливу. *Телекомунікаційні та інформаційні технології*. 2024. Вип. 2, Ч. 83. С. 37–48. DOI: https://doi.org/10.31673/2412-4338.2024.024860.

6. Собко О. Метод класифікації кіберзалякувань в україномовному текстовому контенті засобами штучного інтелекту. *Наука і техніка сьогодні*. 2024. Вип. 13. Ч. 41. С. 1252–1263. DOI: https://doi.org/10.52058/2786-6025-2024-13(41).

7. Браіловський М., Хорошко В. Методи деструктивного впливу та захисту контенту в соціальних мережах. *Безпека інформаційних систем і технологій*. 2023. Вип. 1. Ч. 6. С. 5–12. DOI: https://doi.org/10.17721/ISTS.2023.1.5-12.

8. MacAvaney S., Yao H. R., Yang E., Russell K., Goharian N., Frieder O. Hate speech detection: Challenges and solutions. *PloS one*. 2019. Vol. 14. № 8. P. 1–16. DOI: https://doi.org/10.1371/journal.pone.0221152.

9. Vrysis L., Vryzas N., Kotsakis R., Saridou T., Matsiola M., Veglis A., Arcila-Calderón C., Dimoulas C. A web interface for analyzing hate speech. *Future Internet*. 2021. Vol. 13. № 80. P. 1–18. DOI: https://doi.org/10.3390/fi13030080.

10. Mollas I., Chrysopoulou Z., Karlos S., Tsoumakas G. ETHOS: a multi-label hate speech detection dataset. *Complex & Intelligent Systems*. 2022. Vol. 8. P. 4663–4678. DOI: https://doi.org/10.1007/s40747-021-00608-2.

11. Röttger P., Vidgen B., Nguyen D., Waseem Z., Margetts H., Pierrehumbert J. B. HateCheck: Functional tests for hate speech detection models. *In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2021. Vol. 1. P. 41–58. DOI: https://doi.org/10.18653/v1/2021.acl-long.4.

12. Raju E., Saketh R., Krishna O. K. V., Kushanu P. G. Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media. *International Journal of Multidisciplinary Research in Science, Engineering and Technology*. 2024. Vol. 7. № 12. P. 17872–17878. URL: https://www.ijmrset.com/upload/53_Advances.pdf (дата звернення: 20.01.2025).

13. Chao A. F., Wang C. S., Li B. Y., Chen, H. Y. From hate to harmony: Leveraging large language models for safer speech in times of COVID-19 crisis. *Heliyon*. 2024. Vol. 10. P. 1–19. DOI: https://doi.org/10.1016/j.heliyon.2024.e35468.

14. Zhang Y., Zhong T., Yi T., Li, H. Domain-enhanced prompt learning for Chinese implicit hate speech detection. *IEEE Access*. 2024. Vol. 12. P. 13773–13782. DOI: https://doi.org/10.1109/ACCESS.2024.3351804.

15. Wang C. C., Day M. Y., Wu C. L. Political hate speech detection and lexicon building: A study in Taiwan. *IEEE Access*. 2022. Vol. 10. P. 44337–44346. DOI: https://doi.org/10.1109/ACCESS.2022.3160712.

16. Kim J. Y., Kesari A. Misinformation and hate speech: The case of anti-Asian hate speech during the COVID-19 pandemic. *Journal of Online Trust and Safety*. 2021. Vol. 1. № 1. P. 1–14. DOI: https://doi.org/10.54501/jots.v1i1.13.

17. Li J., Ning Y. Anti-asian hate speech detection via data augmented semantic relation inference. *Proceedings of the International AAAI Conference on Web and Social Media*. 2022. Vol. 16. P. 607–617. DOI: https://doi.org/10.1609/icwsm.v16i1.19319.

18. Богданова І., Лептуга О. «Мова ворожнечі» в текстах українськомовного медіапростору. *Український світ у наукових парадигмах : збірник наукових праць Харківського національного педагогічного університету імені Г. С. Сковороди.* Харків : ХІФТ, 2020. Вип. 7. С. 6–11. URL: http://repositsc.nuczu.edu.ua/handle/123456789/12378 (дата звернення: 20.01.2025).

19. Колесник Г. О. «Мова ворожнечі» як соціальний та лінгвістичний феномен. *Вчені записки Таврійського національного університету імені В. І. Вернадського. Серія: Філологія. Журналістика*. Київ : Гельветика, 2021. Вип. 71. № 4. Ч. 3. С. 278–283. DOI: https://doi.org/10.32838/2710-4656/2021.4-3/46.

20. Олейник А. Етичні феномени соціальних мереж: мова ворожнечі, кіберненависть та кібербулінг. *Вісник Львівського університету. Серія філос.-політолог. студії*. Вип. 41. С. 118–124. DOI: https://doi.org/10.30970/PPS.2022.42.15.

21. Community Standards: Hateful Conduct. 2025. URL: https://transparency.meta.com/en-gb/policies/community-standards/hateful-conduct/ (дата звернення: 20.01.2025).

22. Weibo Service Agreement. 2025. URL: https://www.weibo.com/signup/v5/protocol (дата звернення: 20.01.2025).

23. WeChat Safety Center: Community Guidelines. 2025. URL: https://safety.wechat.com/en_US/community-guidelines/cover/hateful-spam-or-other-inappropriate-behaviour (дата звернення: 20.01.2025).

24. Hugging Face: Model card. URL: https://huggingface.co/thu-coai/roberta-base-cold (дата звернення: 20.01.2025).