УДК 811.581’322
DOI https://doi.org/10.32782/2710-4656/2025.5.1/69

Калінін В. С.
Київський національний лінгвістичний університет

МЕТОДОЛОГІЯ КОРПУСНОГО ДОСЛІДЖЕННЯ ВЕРБАЛЬНИХ ЗАСОБІВ РЕАЛІЗАЦІЇ МОВИ ВОРОЖНЕЧІ В КИТАЙСЬКОМОВНОМУ МЕДІАДИСКУРСІ

У статті витлумачено методологію корпусного дослідження вербальних засобів реалізації мови ворожнечі у китайськомовному медіадискурсі. Визначено, що в умовах жорсткої державної цензури в Китаї реалізація мовної агресії набуває специфічних особливостей, що ускладнює її ідентифікацію традиційними методами. Обґрунтовано необхідність розробки моделі багатокласової типології, здатної категоризувати контент за ознаками ворожості, що є критично важливим для вдосконалення систем модерації.

Інтерпретовано протилежні підходи до регулювання мови ворожнечі: європейський, який зосереджено на правах людини, та китайський, що ставить у пріоритет соціальну гармонію і державний контроль. Проаналізовано наукові погляди дослідників і встановлено, що українські науковці кваліфікують явище під кутом зору лінгвокультурних, соціальних та медійних аспектів, тоді як китайські пропонують більш інструментальні та методологічно розмежовані методики, як-от: формальний, обчислювальний і комунікативно-прагматичний напрями. Обґрунтовано необхідність розробки комплексної методології, що поєднує кількісний корпусний та якісний критичний дискурс-аналіз. Представлено авторське робоче визначення мови ворожнечі.

Для апробації використано адаптовану обчислювальну систему OSKAL та доналаштовану мовну модель на основі архітектури RoBERTa для багатокласової типології за шістьма категоріями: ворожнеча щодо ЛГБТ, за регіональною ознакою, расова ворожнеча, ґендерна ідентичність, загальні образливі висловлювання та нейтральний клас. За допомогою застосунка виконано масовий збір та класифікацію 251 378 дописів та коментарів з китайськомовної соціальної мережі Weibo. У новоствореному репрезентативному корпусі частка контенту з ознаками ненависті сягнула 24,78% (62 283 одиниць). Найпоширенішими категоріями виявилися загальні образи та расова ворожнеча.

Розроблена модель досягла високої загальної точності на тестовій вибірці – 97.86%. Однак, виявлено її системні обмеження при застосуванні до реального дискурсу, пов’язані зі «стерильністю» навчального датасету, труднощами у розпізнаванні імпліцитних форм агресії та вразливістю до стратегій «маскування», які користувачі використовують для обходу цензури. Це підкреслює нагальну потребу у подальшому вдосконаленні обчислювальної методології.

Ключові слова: *мова ворожнечі, корпусний аналіз, китайськомовний медіадискурс, мовна модель, методологія.*

Постановка проблеми. У сучасному інформаційному суспільстві мова ворожнечі (англ. *hate speech*) постає одним із найбільш небезпечних соціальних феноменів і набуває особливо загрозливих форм у глобальному медіадискурсі. Проблема її ефективного виявлення та протидії є особливо гострою у китайськомовному медіапросторі, де поєднання державної цензури та антицензурних стратегій користувачів зумовлює поширення імпліцитних, замаскованих форм мовної агресії. Сучасні наукові підходи, які зосереджено переважно на бінарній класифікації або поверхневій лексико-орієнтованій інтерпретації, нерідко мають обмеження щодо ідентифікації прихованих патернів ненависті та глибшого розуміння способів конструювання агресії. Отже, розробка комплексної, відтворюваної методології корпусного дослідження, яка поєднує кількісний інструментарій та якісний дискурсивний аналіз для виявлення вербальних засобів реалізації мови ворожнечі є критично важливим для вдосконалення систем модерації контенту.

Аналіз останніх досліджень і публікацій. Проблема мови ворожнечі є одним із найбільш актуальних викликів у сучасному медіадискурсі. Науковці витлумачують цей феномен

як комплексне міждисциплінарне явище, яке поєднує лінгвокультурний, соціокультурний та комунікативно-прагматичний виміри, і потребує дослідження на концептуальному й методологічному рівнях. Концепцію вербальних засобів, прагматичних і лінгвістичних механізмів та типології агресивної риторики висвітлено у працях Л. Ажнюк [1], І. Богданової та О. Лептуги [2], які визначили структурні особливості мови ворожнечі та її співвідношення з політкоректністю. Застосування корпусів, колокаційної та лексикографічної інтерпретації для вивчення дискримінаційних номінацій висвітлено у доробках Н. Кондратенко [5] та А. Рудченко [6]. Методологія використання автоматизованих інструментів, нейронних мереж, комп'ютерного та семантичного аналізу є центральною у роботах В. Калініна [3], О. Кислової, І. Кузіної та І. Дирди [4].

На противагу здебільшого гуманітарно-орієнтованим підходам українських досліджень, у китайськомовному науковому полі пріоритетним є створення ресурсної бази та технологічне вдосконалення чинних методів. Розробку спеціалізованих корпусів та лексиконів, які посприяли подоланню лінгвістичних та культурних викликів при автоматизованій детекції мови ворожнечі, запропоновано у працях Deng et al. (COLDATASET) [11], Bai et al. (STATE ToxiCN) [7], Jiang at el. (SWSR) [15], Bennie et al. (PANDA) [8]. Створення мультимодальних моделей та архітектур для підвищення точності бінарної класифікації та протидії технікам маскування виконано у роботах Rao et al. (ROBERTa-CHSD) [17], Xue et al. (MMBERT) [21] та Wang et al. (MultiHateClip) [19]. Застосування просунутих методів машинного навчання та обробки природної мови (NLP) для аналізу соціальної динаміки та прогнозування тенденцій мови ворожнечі презентовано у статтях Wu et al. [20], та Fan et al. [14].

Постановка завдання. Метою статті є розробка та емпіричне обґрунтування комплексної методології для багатокласової типології мови ворожнечі в китайськомовному медіадискурсі.

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

1) сформувати репрезентативний масив даних шляхом застосування системи OSKAL для масового збору контенту з китайськомовної мережі Weibo;

2) розробити та доналаштувати мовну модель для автоматизованої багатокласової типології вербальних засобів агресії за шістьма категоріями;

3) визначити кількісний розподіл ворожих повідомлень в корпусі та встановити системні обмеження обчислювальної методології.

Виклад основного матеріалу. У сучасному інформаційному суспільстві мова ворожнечі є одним із визначальних деструктивних соціальних чинників. Особливої інтенсивності ця риторика набуває в медіадискурсі, де користувачі швидко поширюють агресивні повідомлення не лише вербальними, а й невербальними засобами. Попри те, що протягом останніх десятиліть цей феномен систематично вивчають науковці, єдиного усталеного визначення досі не сформовано. Його трактування варіюється залежно від теоретичної парадигми, дослідницького контексту та соціально-культурних умов.

Міжнародні офіційні інституції по-різному інтерпретують мову ненависті. Наприклад, у документах Європейського суду з прав людини підкреслено, що висловлювання, які принижують чи зневажають групи осіб за ознаками ідентичності, порушують право на гідність і рівність перед законом [13]. Європейський парламент трактує агресивну риторику як публічне підбурювання до насильства або ненависті проти індивіда чи групи осіб, що відрізняються за расою, кольором шкіри, релігією, національним чи етнічним походженням [12].

На противагу західним підходам, у КНР немає єдиного чіткого офіційного визначення поняття hate speech. Проте низка нормативних актів частково охоплює суміжні явища. Зокрема, «Положення про управління екологією мережевого інформаційного контенту» забороняє поширення матеріалу, що «принижує, дискримінує або завдає шкоди іншим» [10]. Зазначаємо, що ці норми є надто загальними й не конкретизують упередженого ставлення за расовою, етнічною чи релігійною ознакою, що ускладнює їхнє застосування як чіткої правової основи для визначення мови ворожнечі. Стаття 249 Кримінального кодексу КНР формально криміналізує «підбурювання до етнічної ненависті або дискримінації» [18] однак фактичне застосування цієї норми визначається не так захистом етнічних меншин, як міркуваннями політичної доцільності – переслідуванню підлягають висловлювання, які влада розцінює як загрозу державі або соціальній гармонії.

Висновуємо, що європейські підходи до регулювання мови ворожнечі вирізняються концептуальною чіткістю та багаторівневістю – вони інтегрують правовий, етичний і соціальний вимір й фіксують основні ознаки феномену: адресність,

вмотивованість ненавистю, наявність підбурювання й шкоди. Натомість у китайському контексті переважає прагматичне регулювання публічного мовлення через обмеження екстремізму, образи національних почуттів та мережеве насильство, без розробки чіткої дефініції. Саме інституційні трактування, що поєднують правові, лінгвістичні та етичні ознаки, стають орієнтиром у подальшому дослідженні.

Науковці також пропонують різні підходи до визначення мови ворожнечі й акцентують увагу на певних аспектах її функціонування. Українські дослідники кваліфікують цей феномен крізь призму кількох взаємодоповнювальних вимірів – лінгвістичного, соціокультурного та комунікативно-прагматичного. І. Богданова та О. Лептуга зосереджуються на мовних механізмах формування ненависті в медіапросторі, трактують явище як систему мовних засобів, що конструюють опозицію «свій – чужий» і ґрунтуються на соціальних стереотипах та дискримінаційних установках [2, с. 7, 10]. Науковиці інтерпретують її у лінгвокультурному та комунікативному вимірах і вказують на приховані форми агресії, маніпуляції та символічне насильство в сучасних медіа, зокрема у форматі мемів і візуально-текстових гібридів. О. Кислова, І. Кузіна та І. Дирда натомість, наголошують на соціально-політичному аспекті явища: мова ворожнечі, на їхню думку, відображає глибші суспільні поділи, закріплює опозицію «ми – вони» та сприяє формуванню «чорно-білого» світогляду, що загрожує соціальній стабільності й толерантності, особливо в онлайновому середовищі [4, с. 253]. У сукупності ці підходи окреслюють тлумачення агресивної риторики як комплексного міждисциплінарного явища, що поєднує вербальне упередження, соціальне протиставлення й медійну маніпуляцію.

На противагу вітчизняним науковцям, китайські дослідники інтерпретують мову ворожнечі через призму більш вузькоспеціалізованих методологічних підходів, що зумовлено конкретними завданнями: створенням мовних корпусів, експертизою великих даних або глибоким прагматичним аналізом. Замість єдиного комплексного визначення, їхні праці пропонують кілька операційних підходів. По-перше, для навчання мовних моделей використовується формальний, класифікаційний підхід. У його межах агресивна риторика визначається як будь-яка комунікація, що принижує групу на основі захищених ознак, як-от: раса, стать чи релігія, і часто містить принизливі вислови, стереотипи та дегуманізацію [9, с. 93, 98]. По-друге, для обробки великих масивів інформації, зокрема із соціальних мереж, застосовується обчислювальний, лексико-орієнтований метод, де мовою ненависті вважається будь-який текст, що містить слова зі спеціалізованих словників дискримінаційної лексики, наприклад, Hatebase [14, с. 2]. По-третє, є комунікативно-прагматичний підхід, який фокусується на імпліцитних, прихованих формах агресії. З цієї перспективи мова ворожнечі витлумачується не через лексичне наповнення, а як мовленнєва дія – непряма «атака» на об'єкт, де прагматична функція висловлювання (сарказм чи зневага) суперечить його буквальному значенню [22, с. 458, 459].

З огляду на різноманіття проаналізованих підходів, у межах цього дослідження запропоновано власне робоче визначення. Мову ворожнечі трактовано як динамічну систему вербальних і невербальних одиниць, що виражає, конструює та поширює ненависть, упередження чи дискримінацію щодо осіб чи груп на основі їхніх реальних або припущених характеристик, а саме: етнічна належність, раса, стать, ґендер, регіональна ознака, сексуальна орієнтація, релігійні переконання чи соціальний статус. Особливістю такої риторики у сучасному цифровому середовищі є здатність підлаштовуватися під алгоритми модерації та маскуватися через використання сарказму, іронії, кодованих слів чи візуального контенту [3, с. 230]. Саме тому корпусний аналіз дослідження є ключовим для емпіричної верифікації мовних патернів у великих масивах даних, що уможливило автоматизовану ідентифікацію вербальних засобів агресії.

Попри глобальне поширення мови ворожнечі, її системні корпусні дослідження в китайськомовному медіадискурсі активізувалися лише в останнє десятиліття. Пояснюємо це частковим браком надійних та загальнодоступних наборів даних. Однією з перших спроб заповнити цю прогалину стало створення COLD (Chinese Offensive Language Dataset) – першого відкритого бенчмарку для виявлення образливої лексики в китайській мові, що охоплює теми раси, ґендеру та регіональної приналежності [11, с. 11580, 11582]. Поява таких ресурсів прискорила розробку автоматизованих методів аналізу, хоча наявні напрацювання здебільшого зорієнтовані на вирішення технічних завдань класифікації та прогнозування.

Іншим важливим напрацюванням є двомовний корпус MultiHateClip, що містить відео з YouTube та китайської платформи Bilibili. Ця робота виходить за межі суто вербального аналізу і доводить,

що в китайському медіапросторі мова ворожнечі є переважно мультимодальною: у понад 80% випадків ненависницький зміст передається через комбінацію тексту, візуальних образів та аудіо [19, с. 7494]. Погоджуємося з обґрунтуванням авторів щодо обмеженості тільки лише текстоцентричних методів під час дослідження сучасних форм мовної агресії в медіа. Водночас у межах пропонованої розвідки увагу сфокусовано саме на вербальній складовій, як фундаментальній основі риторики ненависті.

Отже, попри те, що згадані праці мають переважно технічну спрямованість, вони надають ключові інсайти: по-перше, про соціальну динаміку поширення ворожнечі, і, по-друге, про її мультимодальний характер. Водночас комплексний лінгвістичний аналіз лексико-граматичних, стилістичних та прагматичних характеристик китайськомовної риторики ненависті досі залишається недостатньо розробленим, що й визначає актуальність дослідження.

Особливий інтерес становить саме китайськомовний простір, адже тут поєднано кілька суперечливих тенденцій. З одного боку, він характеризується наявністю однієї з найскладніших та технологічно розвинених систем інтернет-цензури у світі, спрямованої на видалення контенту, що може спровокувати колективні дії [16, с. 2, 8]. З іншого боку, таке середовище жорсткого контролю не запобігає поширенню ворожих висловлювань, а лише модифікує їх і робить більш прихованими.

Дослідження Chen et al., засноване на масштабному опитуванні користувачів Weibo, емпірично підтверджує цей парадокс. Через цензуру понад 90% користувачів вдаються до самоцензури, особливо при обговоренні гострих соціально-політичних тем [9, с. 99]. Водночас, щоб обійти обмеження, користувачі активно застосовують численні антицензурні стратегії: використання омофонів, піньїню, акронімів, сатири, а також перетворення тексту на зображення [9, с. 101]. Саме ці практики надають змогу ворожому контенту поширюватися, маскуючись під нейтральні чи неоднозначні повідомлення. Це створює унікальні умови для функціонування мовної агресії, яка стає більш імпліцитною, кодованою та контекстуально залежною, що передусім потребує значно складніших методів для її ґрунтовного вивчення.

Кількісні та обчислювальні підходи, що набули поширення з появою перших китайських корпусів мови ворожнечі, а саме: CHSD [17] та COLD [11], хоч і уможливили роботу з великими обсягами даних, також мають критичні обмеження. По-перше, базові кількісні методи, що покладаються на зіставлення слів зі списками токсичної лексики, демонструють високу неточність. У дослідженнях зазначено, що чутливі слова трапляються не лише у ворожих, а й в абсолютно нейтральних чи навіть антидискримінаційних контекстах, що зумовлює появу значної кількості хибних спрацьовувань [11, с. 11585, 11596]. По-друге, навіть досконаліші корпусні дослідження, що застосовують моделі глибокого навчання на кшталт ROBERTa-CHSD, здебільшого націлені на вирішення технічного завдання бінарної класифікації («ворожий» / «неворожий» текст) [17, с. 503, 508]. Такий підхід корисний для автоматичної модерації, проте не дає відповіді на запитання, як саме конструюється мова ненависті: які лексичні, граматичні чи прагматичні стратегії використовуються для створення упередженого висловлювання.

Одним із суттєвих недоліків текстоцентричних методів є їхня вразливість до стратегій «маскування», які широко використовуються в китайському інтернет-просторі для обходу цензури [21, с. 1, 2]. Користувачі навмисно спотворюють текст за допомогою омофонічної заміни (заміни ієрогліфів на схожі за звучанням), деформації ієрогліфів (розкладання на графічні складники) та змішування коду (вкраплення піньїню, англійських літер чи емодзі). Наприклад, здійснюють заміну ієрогліфа 满» (*mǎn*, повний) на «蛮» (*mán*, варвар) через їхню співзвучність [21, с. 2]. Такі маніпуляції, що апелюють до фонетичних та візуальних знань носія мови, роблять ворожий зміст невидимим для моделей, що аналізують лише стандартну вербальну форму.

Отже, наявні методології окреслюють подвійний виклик: з одного боку, якісні підходи обмежені в масштабованості, а з іншого – кількісні методи є або поверхневими, або менш ефективними у протистоянні креативним практикам вираження ненависті в умовах жорсткої цензури китайського цифрового середовища. У цьому контексті констатуємо нагальну потребу у комплексній методології, яка б поєднувала інструменти корпусної лінгвістики з якісним дискурсивним аналізом. Застосування такого підходу уможливить визначення закономірностей функціонування ворожих висловлювань, визначити їхні лексико-граматичні та стилістичні характеристики, а також окреслити прагматичні стратегії, що формують специфіку агресивної комунікації в медіапросторі Китаю.

У праці застосовано комплексну обчислювальну методологію, що поєднує корпусний аналіз та просунуті моделі машинного навчання. Реалізація цього підходу передбачала насамперед створення репрезентативного масиву текстів, що послугував емпіричною базою дослідження. Зібрані дописи та коментарі було класифіковано доналаштованою мовною моделлю за ознаками агресії. Такий інструментальний підхід уможливив, з одного боку, виявлення та кількісну оцінку лексико-семантичних патернів, притаманних ворожим висловлюванням, а з іншого – закладення основи для їхньої подальшої критико-дискурсивної інтерпретації у ширшому соціокультурному контексті.

Для збору та формування корпусу текстів було використано попередньо розроблене В. Калініним програмне забезпечення OSKAL (Opensource Social Knowledge of Aggressive Language) [3]. Це рішення, реалізоване мовою програмування Python, уможливило масовий збір та аналіз публічно доступних матеріалів із китайськомовної соціальної мережі Weibo. Архітектура системи складається з двох ключових компонентів: серверної частини та клієнтського інтерфейсу. Перша містить базу даних, вебсервер та програму-воркер, яка в автоматизованому режимі виконує обробку інформації. Для взаємодії з базою даних використано бібліотеку SQLAlchemy – ORM (Object-Relational Mapper), що забезпечило зручний доступ до контенту. Для веб-скрапінгу застосовано фреймворк Scrapy, який надав змогу створювати спеціалізованих ботів («спайдерів»). Розроблений спайдер автоматично отримував дописи та коментарі користувачів, обробляв і зберігав їх. Після збору початкові відомості проходили етап очищення, під час якого модуль нормалізації видаляв HTML-код, скрипти, емодзі та інші метатеги [3, с. 229–231].

Після збору та нормалізації текстів було здійснено підготовку власної мовної моделі для автоматизованого виявлення ознак агресивної мови. Навчання охоплювало три основних етапи.

На першому етапі здійснено формування та підготовку спеціалізованого датасету. Основою для нього послугувала таблиця, що містила два стовпці: text (оригінальний текст висловлювання) та sentiment (категорія його тональності). Було визначено шість категорій: ворожнеча щодо ЛГБТ, за регіональною ознакою, расою, ґендерною ідентичністю, а також загальна категорія образливих висловлювань та нейтральний клас. Для забезпечення різноманітності та повноти даних використано декілька джерел: частина речень була згенерована за допомогою штучного інтелекту, інша – запозичена з наявних корпусів (наприклад, COLD [11], SWSR [15], State ToxiCN [7]), а решта – відібрана з масиву, зібраного на попередньому етапі дослідження [3, с. 231]. У результаті створено збалансований набір, який містив по 3000 прикладів для кожної категорії, що у сумі становило 18 000 речень.

З метою забезпечення високої якості даних та надійності майбутньої моделі, весь масив текстів пройшов ретельну перевірку. Було проведено валідацію, очищення від вербального «шуму», HTML-тегів, непотрібних символів, емодзі та дублікатів, що забезпечило фокус саме на вербальному змісті. Цей крок був критично важливим, оскільки унеможливив тренування на некоректних або повторюваних прикладах, що зумовлює перенавчання та зниження її узагальнювальної здатності. Ці процедури гарантували високу якість навчального матеріалу, зменшили ризик хибнопозитивних або хибнонегативних класифікацій і підвищили репрезентативність корпусу.

На завершення першого етапу підготовлений датасет розподілено на три незалежні вибірки: навчальну, валідаційну та тестову. Для збереження однакових пропорцій класів у кожному блоці застосовано метод стратифікованого поділу. Це гарантувало, що всі підгрупи даних зберегли те ж саме відсоткове співвідношення категорій, яке було у початковому наборі. Такий підхід забезпечив навчання й перевірку моделі на репрезентативних матеріалах, що є ключовою умовою для отримання об'єктивної оцінки ефективності.

На другому етапі було проведено доналаштування попередньо навченої мовної моделі для завдання класифікації текстів. Для цього обрано **hfl/chinese-roberta-wwm-ext** – потужну версію архітектури RoBERTa, спеціально навчену на великому корпусі китайської мови. Такий вибір обґрунтований тим, що подібні моделі вже мають глибоке «розуміння» мовних патернів, що робить їх чудовою основою для спеціалізованих завдань.

Підготовка розпочалася з токенізації, де кожен текст перетворено на послідовність числових кодів. Для стандартизації вхідних даних усі послідовності були уніфіковані до однакової довжини у 200 токенів шляхом доповнення або обрізання. Після цього підготовлений контент було завантажено у спеціальні контейнери, які забезпечили їхню ефективну порційну подачу для навчання та валідації невеликими партіями (розміром у 16 прикладів).

Для тренування ініціалізовано класифікаційну модель, яку було сконфігуровано для завдання розподілу текстів на шість визначених раніше категорій. Процес доналаштування відбувався протягом трьох епох з використанням оптимізаційного алгоритму та низької швидкості навчання (2e-5), що є стандартною практикою для тонкого налаштування трансформерних моделей. Для стабілізації навчання застосовано планувальник швидкості навчання з лінійним зменшенням та коротким проміжком «розігріву».

Навчальний цикл складався з ітеративного процесу: після кожної епохи тренування на навчальній вибірці модель оцінювалася на валідаційній. Для збереження було використано стратегію відбору найкращого варіанта: зберігалася лише та версія, яка демонструвала найвищу точність даних. Це надало змогу уникнути перенавчання та гарантувати, що фінальний класифікатор ефективно узагальнює патерни, а не просто запам'ятовує навчальні приклади.

Третій, завершальний етап зосереджено на всебічній оцінці та тестуванні найкращого класифікатора. Насамперед, фінальна перевірка на контрольній вибірці зафіксувала високу загальну точність моделі (97.86%). Для глибшого аналізу продуктивності розраховано низку деталізованих метрик. Побудовано матрицю плутанини, а також обчислено показники точності, повноти та F1-міри для кожної з шести категорій, що уможливило детальне розуміння сильних та слабких сторін системи. Зокрема, найвищий показник F1-міри (близько 0.988) вона продемонструвала для категорії регіональної ворожнечі. Інтерпретація матриці плутанини засвідчила, що найпоширенішою помилкою була класифікація висловлювань расової ворожнечі як ґендерної, що свідчить про потенційну контекстуальну схожість мовних тенденцій, спрямованих на ці дві групи. Важливою частиною цього етапу став якісний розбір помилок, де випадки незбігу прогнозів із реальними мітками зібрано та проаналізовано для виявлення патернів – наприклад, чи має модель труднощі з коротшими текстами. Насамкінець, ефективність розробленого рішення порівняно з базовим методом, який завжди прогнозує найпоширеніший клас, що наочно продемонструвало його перевагу у багатокатегорійній ідентифікації контенту.

Нарешті, для забезпечення практичного застосування моделі було створено функцію для прогнозування на нових текстових даних. Усі отримані результати – метрики, графіки, звіти про помилки та фінальні прогнози на тестовій вибірці – збережено у структурованому вигляді для подальшого аналізу та забезпечення повної відтворюваності дослідження.

Отже, застосована триетапна методологія забезпечила прозорий та відтворюваний процес створення класифікатора. Вона надала змогу не лише кількісно оцінити загальну продуктивність на тестових даних, але й провести глибокий якісний аналіз її сильних та слабких сторін у розрізі різних категорій ворожнечі. Усе це є необхідною умовою для подальшого практичного застосування моделі у прикладних завданнях.

Збір даних здійснено у проміжок з 20 вересня 2025 року по 20 жовтня 2025 року. Протягом цього часу воркер зібрав 16 931 дописи та 234 447 коментарі з 11 категорій платформи Weibo: «Популярне», «Чарти», «Місто», «Суспільство», «Наука і техніка», «Зірки», «Фільми», «Музика», «Стосунки», «Мода» та «Краса». Після автоматизованої класифікації, проведеної розробленою мовною моделлю, ознаки ворожості було виявлено у 5303 дописах та 56 980 коментарях. Отже, створений корпус репрезентує широкий спектр китайськомовного медіадискурсу, охоплює різні тематики й жанри, та слугує надійною емпіричною базою для подальшого аналізу мови ненависті у соціальних мережах.

Для кількісного представлення виявленого ворожого контенту, його розподіл за шістьма визначеними категоріями подаємо у (табл. 1).

Таблиця 1

Розподіл виявлених ворожих висловлювань за категоріями

Категорія	Кількість дописів	Кількість коментарів
Нейтральний клас	11 628	177 467
Загальні образливі висловлювання	1653	19 918
Ворожнеча за расою	1489	15 215
Ворожнеча за статтю	956	12 463
Ворожнеча за регіональною ознакою	764	6 467
Ворожнеча щодо ЛГБТ	441	2917
Всього	16 931	234 447

На етапі контрольованого тестування доналаштована мовна модель продемонструвала високу ефективність із загальною точністю 97.86%, проте при застосуванні до реального корпусу виявилися системні проблеми неправильної ідентифікації деяких категорій. Ці відмінності між результатами перевірки та практичним застосуванням вказують на суттєві обмеження поточного обчислювального підходу.

Основні системні обмеження класифікатора пов'язані, по-перше, зі «стерильністю» навчального датасету, оскільки навчальний корпус, створений частково зі штучно згенерованих та заздалегідь розмічених прикладів, виявився занадто «чистим» у порівнянні з динамічним реальним дискурсом Weibo, що зумовило зниження здатності моделі до узагальнення. По-друге, констатуємо чутливість до емоційного забарвлення, коли класифікатор помилково визначає нейтральні тексти як ворожі через їхнє імпліцитне значення чи використання сленгу, який не був достатньо представлений у навчальному датасеті, що відображає загальну проблему обчислювальної лінгвістики у розпізнаванні агресії, замаскованої сарказмом чи іронією. Крім того, виявлена контекстуальна невизначеність в класифікації підтверджується значною кількістю помилок у матриці плутанини (зокрема між расовою та ґендерною ворожнечею), що свідчить про труднощі з чітким розмежуванням цих класів, ймовірно, через їхню схожість або спільне використання певних лексичних маркерів, що ускладнює точне визначення об'єкта ненависті.

Отже, розроблена система OSKAL разом із новою натренованою мовною моделлю виявила свою ефективність в автоматизованому зборі та первинній класифікації великих обсягів даних, однак вбачаємо нагальну необхідність у подальшому вдосконаленні обчислювальної методології для підвищення її стійкості до імпліцитних форм агресії та стратегій маскування.

Висновки. У результаті проведеного дослідження досягнуто поставленої мети щодо розробки та емпіричного тестування комплексної, відтворюваної методології корпусного дослідження вербальних засобів реалізації мови ворожнечі в китайськомовному медіадискурсі. Для вирішення поставлених завдань було створено комплексний обчислювальний інструментарій, що забезпечив масштабованість кількісного аналізу та заклав основу для подальшої глибинної інтерпретації феномена. Зокрема, впроваджено адаптовану обчислювальну систему OSKAL та здійснено доналаштування власної мовної моделі на основі архітектури RoBERTa для багатокласової типології.

За час дослідження було проаналізовано значний масив даних із соціальної мережі Weibo, який склав 251 378 одиниць (дописи та коментарі). Зазначаємо, що 75,22% текстів є нейтральними (189 095 одиниць), тоді як частка контенту з ознаками ворожнечі становить 24,78% (62 283 одиниць). Серед агресивних висловлювань найбільш поширеною категорією виявилися загальні образи та расова ворожнеча.

У роботі було визначено обмеженість обчислювальної моделі, які полягають у її вразливості до «стерильності» навчального датасету, труднощах у розпізнаванні імпліцитних форм агресії та контекстуальній плутанині між схожими класами. Для подальшого вдосконалення методології та верифікації результатів створено базу розмічених прикладів для шести категорій ворожості, яка може бути використана як еталонна вибірка.

Перспективу подальших досліджень вбачаємо у двох ключових напрямах: по-перше, у розробці обчислювальних механізмів, які протистоятимуть стратегіям «маскування» ворожості, як-от: омофони, графічні заміни тощо; по-друге, в інтеграції мультимодального аналізу для вивчення невербальних засобів агресії, що є критично важливим для комплексного розуміння медіадискурсу.

Реалізація цих напрямів надасть змогу створити більш стійкі, надійні та точні обчислювальні моделі, що матимуть значне практичне значення для підвищення ефективності систем модерації контенту та сприятимуть ґрунтовному соціолінгвістичному розумінню динаміки мови ворожнечі у контрольованому медіапросторі.

Список літератури:

4. Ажнюк Л. Мова ворожнечі й політкоректність у сучасній суспільно-політичній комунікації. *Мовознавство.* 2024. Вип. 2. С. 9–23. DOI: https://doi.org/10.33190/0027-2833-335-2024-2-002.

5. 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.* 2020. Вип. 7. С. 6–11. URL: http://repositsc.nuczu.edu.ua/handle/123456789/12378 (дата звернення: 21.10.2025).

6. Калінін В. Розробка інструментів автоматизованого аналізу даних для ідентифікації вербальних засобів реалізації мови ворожнечі у китайськомовному медіадискурсі: кількісний і якісний підхід. *Закарпатські філологічні студії.* 2024. Вип. 38. С. 227–233. DOI: https://doi.org/10.32782/tps2663-4880/2024.38.42.

7. Кислова О., Кузіна І., Дирда І. Дослідження онлайнової мови ворожнечі щодо ромської меншини в українському інтернет-просторі. *IDEOLOGY AND POLITICS JOURNAL.* 2020. Вип. 2(16). С. 250–275. DOI: https://doi.org/10.36169/2227-6068.2020.01.00021.

8. Кондратенко Н. Методи корпусної лінгвістики в дослідженні етно-і гетеростереотипів. *Іван Огієнко і сучасна наука та освіта*. 2023. Вип. 20. С. 80–89. DOI: https://doi.org/10.32626/2309-7086.2023-20.80-89.

9. Рудченко А. Російський наратив про «українських фашистів» у публікаціях українських медіа. *Вісник Львівського університету. Серія Журналістика*. 2020. Вип. 47. С. 171–178. DOI: http://dx.doi.org/10.30970/vjo.2020.47.10516.

10. Bai Z., Yin S., Lu J., Zeng J., Zhu H., Sun Y., Yang L., Lin H. STATE ToxiCN: A Benchmark for Span-level Target-Aware Toxicity Extraction in Chinese Hate Speech Detection. *arXiv preprint arXiv:2501.15451*. 2025. DOI: https://doi.org/10.48550/arXiv.2501.15451.

11. Bennie M., Zhang D., Xiao B., Cao J., Liu C. X., Meng J., Tripp, A. PANDA – Paired Anti-hate Narratives Dataset from Asia: Using an LLM-as-a-Judge to Create the First Chinese Counterspeech Dataset. *arXiv preprint arXiv:2501.00697*. 2025. DOI: https://doi.org/10.48550/arXiv.2501.00697.

12. Chen X., Xie J., Wang Z., Shen B., Zhou Z. How we express ourselves freely: Censorship, self-censorship, and anti-censorship on a Chinese social media. In: *International Conference on Information*. Cham: Springer Nature Switzerland. 2023. Pp. 93–108. DOI: https://doi.org/10.1007/978-3-031-28032-0_8.

13. Cyberspace Administration of China. *Regulations on the Governance of Network Information Content Ecology (*网络信息内容生态治理规定*)*. 2020. URL: https://www.cac.gov.cn/2019-12/20/c_1578375159509309.htm (дата звернення: 21.10.2025).

14. Deng J., Zhou J., Sun H., Zheng C., Mi F., Meng H., Huang M. COLD: A benchmark for Chinese offensive language detection. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 2022. Pp. 11580–11599. DOI: https://doi.org/10.18653/v1/2022.emnlp-main.796.

15. European Parliament. *Resolution of 18 January 2024 on extending the list of EU crimes to hate speech and hate crime (2023/2068(INI))*. 2024. URL: https://www.europarl.europa.eu/doceo/document/TA-9-2024-0044_EN.html (дата звернення: 21.10.2025).

16. European Court of Human Rights. *Factsheet: Hate speech*. 2023. URL: https://www.echr.coe.int/documents/d/echr/fs_hate_speech_eng (дата звернення: 21.10.2025).

17. Fan L., Yu H., Yin Z. Stigmatization in social media: Documenting and analyzing hate speech for COVID-19 on Twitter. *Proceedings of the Association for Information Science and Technology*. 2020. № 57(1). Pp. 1–11. DOI: https://doi.org/10.1002/pra2.313.

18. Jiang L., Ma Y., Lin H., Xia R., Wang Y. SWSR: A Chinese Dataset and Lexicon for Online Sexism Detection. *arXiv preprint arXiv:2108.03070*. 2021. Pp. 1–44. URL: https://arxiv.org/pdf/2108.03070 (дата звернення: 21.10.2025).

19. Liu J. Internet censorship in China: Looking through the lens of categorisation. *Journal of Current Chinese Affairs*. 2024. Pp. 1–16. DOI: https://doi.org/10.1177/18681026231220948.

20. Rao X., Zhang Y., Jia Q., Liu X., Peng S. 基于 RoBERTa 的中文仇恨言论侦测方法研究 (Chinese Hate Speech detection method Based on RoBERTa-WWM). In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*. 2023. Pp. 501–511. URL: https://aclanthology.org/2023.ccl-1.44/ (дата звернення: 21.10.2025).

21. The Supreme People's Procuratorate of the People's Republic of China. *Criminal Law of the People's Republic of China*. 2020. URL: https://en.spp.gov.cn/2020-12/26/c_948417_12.htm (дата звернення: 21.10.2025).

22. Wang H., Yang T. R., Naseem U., Lee R. K. W. Multihateclip: A multilingual benchmark dataset for hateful video detection on youtube and bilibili. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 2024. Pp. 7493–7502. DOI: https://doi.org/10.1145/3664647.3681521.

23. Wu X. K., Zhao T. F., Lu L., Chen W. N. Predicting the hate: A GSTM model based on COVID-19 hate speech datasets. *Information Processing & Management*. 2022. № 59(4). Pp. 1–15. DOI: https://doi.org/10.1016/j.ipm.2022.102998.

24. Xue Q., Dou Y., Shi R., Li X. L., Gao W. MMBERT: Scaled Mixture-of-Experts Multimodal BERT for Robust Chinese Hate Speech Detection under Cloaking Perturbations. *arXiv preprint arXiv:2508.00760*. 2025. DOI: https://doi.org/10.48550/arXiv.2508.00760.

25. 徐俊扬. 语料库驱动的网络隐性仇恨言论的语用分析——以 “专家建议” 话题评论为例. 人文与社会科学学. 2025. № 1(2). Pp. 456–460. DOI: https://doi.org/10.70693/rwsk.v1i2.539.

Kalinin V. S. THE METHODOLOGY FOR CORPUS ANALYSIS OF VERBAL MEANS USED TO REALIZE HATE SPEECH IN CHINESE-LANGUAGE MEDIA DISCOURSE

The article explores the methodology for corpus analysis of verbal means used to realize hate speech in Chinese-language media discourse. It is determined that, given the context of strict state censorship in China, the realization of verbal aggression acquires specific characteristics, which complicates its identification using traditional methods. The necessity of developing a multi-class typology model is substantiated, which can categorize content based on the features of hostility, a factor critically important for improving moderation systems.

Opposing approaches to hate speech regulation are examined: the European one, focused on human rights, and the Chinese one, which prioritizes social harmony and state control. Scientific views of researchers are analyzed: Ukrainian scholars interpret the phenomenon from the perspective of linguacultural, social, and media aspects, while the Chinese propose more instrumental and methodologically differentiated models, such as formal, computational, and communicative-pragmatic directions. Considering these differences, the necessity of a mixed methodology that combines quantitative corpus analysis and qualitative Critical Discourse Analysis is substantiated. The study introduces the author's working definition of hate speech.

The adapted computational system OSKAL and a fine-tuned language model based on the RoBERTa architecture are used for multi-class typology across six categories: anti-LGBT hostility, regional bias, racial hostility, gender identity, general offensive language, and a neutral class. Mass collection and classification of 251,378 posts and comments from the Chinese social network Weibo are performed using the application. This led to a representative corpus creation, revealing that the share of content with signs of hostility was 24.78% (62,283 units). The most common categories were general offensive language and racial hostility.

The model achieves a high overall accuracy on the test sample – 97.86%. However, the study also highlights its systemic limitations when applied to real discourse, which are related to the "sterility" of the training dataset, difficulties in recognizing implicit forms of aggression, and vulnerability to "masking" strategies used to bypass censorship. The findings contribute to underscoring the urgent need for further improvement of the computational methodology.

Key words: *hate speech, corpus analysis, Chinese-language media discourse, language model, methodology.*

Дата надходження статті: 27.10.2025
Дата прийняття статті: 20.11.2025
Опубліковано: 29.12.2025