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 ґрунтується 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 НМТіЗ-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.