УДК 004.8

Молчанова М.О.

Хмельницький національний університет

КЛАСИФІКАЦІЯ ТЕКСТІВ ЗА ВМІСТОМ ПРОПАГАНДИ НЕЙРОМЕРЕЖЕВИМИ МОДЕЛЯМИ ГЛИБОКОГО НАВЧАННЯ

Розроблено метод класифікації текстів за вмістом пропаганди нейромережевими моделями глибокого навчання, що дозволяє виявляти як явні, так і приховані пропагандистські меседжі, який ґрунтується на об'єднанні можливостей традиційних рекурентних нейромереж з довгостроковою пам'яттю і нейромереж-трансформерів. Отримані результати дослідження ефективності розробленого методу свідчать про спроможність запропонованого методу ефективно класифікувати тексти за вмістом пропаганди нейромережевими моделями глибокого навчання й розроблений метод може використовуватись для оцінки потенційних загроз, пов'язаних з поширенням пропаганди. У ході дослідження вдалось досягнути точності у 97.8 %.

Method for classifying texts by the content of propaganda using neural network models of deep learning has been developed, which allows detecting both explicit and hidden propaganda messages, which is based on combining the capabilities of traditional recurrent neural networks with long-term memory and neural networks-transformers. The obtained results of the study of the effectiveness of the developed method testify to the ability of the proposed method to effectively classify texts by the content of propaganda by neural network models of deep learning, and the developed method can be used to assess potential threats associated with the spread of propaganda. During the research, it was possible to achieve an accuracy of 97.8%.

Інформаційні маніпуляції здійснюються через різноманітні форми, методи й засоби впливу на людей із метою зміни їхніх психологічних характеристик у бажаному напрямку, тому своєчасне виявлення пропаганди є актуальним напрямком наукових досліджень [1]. Такого роду маніпуляції часто використовується для зміни психологічних настроїв в суспільстві, мобілізації підтримки або ж з метою дискредитації опонентів [2, 3].

Загроза маніпуляцій ЗМІ на громадську думку спонукає до наукових досліджень пропаганди та мовних впливів, а також до аналізу комунікаційних факторів в контексті інформаційної безпеки [4]. Зростання споживання онлайн-контенту посилює ризики пропаганди, що загрожує національній безпеці несвоєчасне вирішення якої може призвести до руйнівних наслідків [5, 6]. У соціологічному енциклопедичному словнику термін «пропагандистський допис» розглядається у декількох варіантах: поширення в масах ідеології та політики певних класів, партій, держав; засіб маніпуляції масовою свідомістю [7].

Метою роботи є створення методу класифікації текстів за вмістом пропаганди нейромережевими моделями глибокого навчання, що дозволяє виявляти як явні, так і приховані пропагандистські меседжі, який ґрунтується на об'єднанні можливостей традиційних рекурентних нейромереж з довгостроковою пам'яттю і нейромереж-трансформерів, а також на використанні механізму аугментації навчальних текстових даних, що дозволяє розширити кількість навчальних зразків.

Метод класифікації текстів за вмістом пропаганди нейромережевими моделями глибокого навчання призначений для автоматизованої ідентифікації текстів, які містять пропагандистські елементи. На рисунку 1 представлено структуру методу. Запропонований підхід відрізняється тим, що дає змогу ідентифікувати як відкриті, так і приховані пропагандистські повідомлення. Це забезпечується за рахунок інтеграції рекурентних нейронних мереж із довготривалою пам'яттю з архітектурою трансформерів, а також завдяки застосуванню технік аугментації текстових даних, що сприяє збільшенню кількості навчальних прикладів.

Метод працює шляхом перетворення вхідних даних у вигляді навченої нейромережевої моделі глибокого навчання та тексту для класифікації у вихідні дані у вигляді відсоткової оцінки наявності пропаганди у тексті та присвоєння класу: «текст без пропаганди», «пропагандистський текст» або «підозрілий текст».

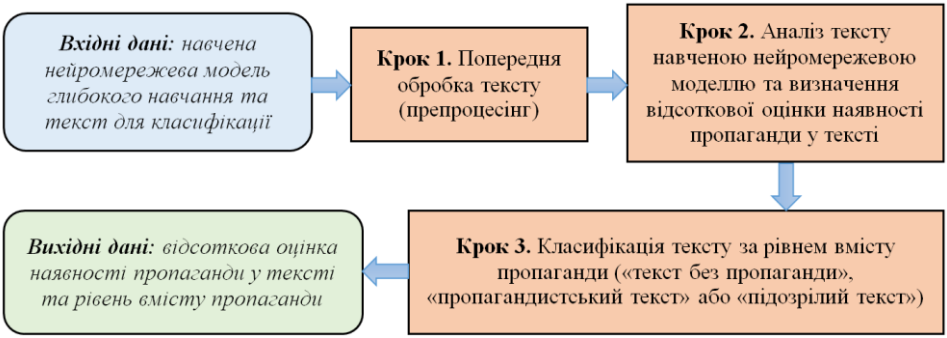

Рисунок 1 – Кроки методу класифікації текстів за вмістом пропаганди нейромережевими моделями глибокого навчання

Першим етапом процесу є попередня обробка тексту для класифікації. Цей етап включає кілька кроків препроцесингу, таких як приведення тексту до нижнього регістру, видалення стоп-слів, пунктуації та інших зайвих елементів. Після цього оброблений текст перетворюється в числові послідовності, які надалі передаються у нейронну мережу глибокого навчання з гібридною архітектурою для оцінки рівня пропаганди.

Другий етап полягає у застосуванні навченої моделі нейронної мережі для аналізу тексту. В результаті цього аналізу модель виконує числову оцінку рівня пропагандистського контенту. Використовується нейронна мережа, що поєднує рекурентну архітектуру з довготривалою пам'яттю (LSTM) та трансформери. Така комбінація забезпечує глибоке розуміння послідовностей і контексту в текстових даних. Схематичне зображення архітектури моделі представлено на рисунку 2.

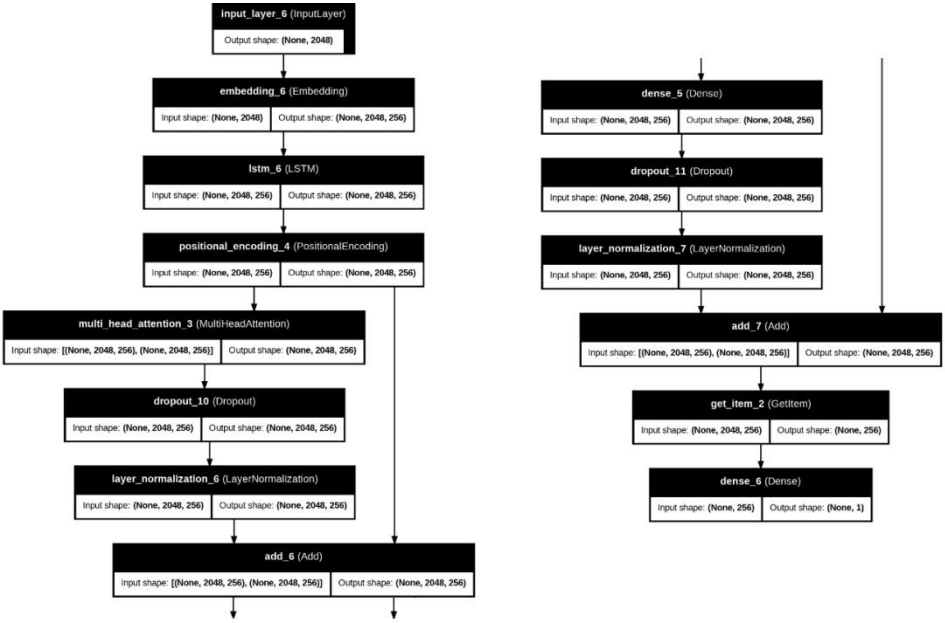

Рисунок 2 – Архітектура нейронної мережі для класифікації пропаганди

Вибір зазначеної архітектури обумовлений перевагами та обмеженнями її складових. LSTM ефективно обробляє послідовні дані, але має труднощі з утриманням довготривалих залежностей, що особливо відчутно при роботі з великими текстами. З іншого боку, шар Multi-Head Attention в трансформерах дозволяє аналізувати весь текст одразу, що сприяє більш точному збереженню контексту. Однак цей підхід вимагає значних обчислювальних ресурсів, особливо при роботі з довгими текстовими послідовностями.

Через обмежену кількість даних для навчання нейромережевих моделей у рамках методу заплановано використання техніки аугментації тексту, що сприятиме підвищенню різноманітності навчальної вибірки. Для цього буде застосована модель трансформерної нейромережевої архітектури «Text-to-Text Transfer Transformer», яка вирішує всі задачі шляхом перетворення тексту в текст. Це

охоплює такі задачі, як переклад, узагальнення, відповіді на запитання та перефразування.

Третій етап методу полягає в класифікації аналізованого тексту до одного з трьох класів: «текст без пропаганди», «пропагандистський текст» або «підозрілий текст». Для цього емпіричним шляхом було визначено межі для кожної категорії. Зокрема, тексти без пропаганди мають оцінку від 0 до 0.45, підозрілі тексти — від 0.45 до 0.55, а пропагандистські — від 0.55 до 1. Ці порогові значення можуть бути змінені та адаптовані залежно від специфіки даних і типів пропаганди, що аналізуються.

У результаті метод генерує два типи вихідних даних: відсоткову оцінку ймовірності наявності пропаганди в тексті та належність до одного з трьох класів: «текст без пропаганди», «пропагандистський текст» або «підозрілий текст».

Для навчання нейромережі було сформовано набір даних з 25 000 записів, що належать категоріям «Пропаганда» та «Не пропаганда». Переліки пропагандистських та верифікованих джерел було сформовано згідно офіційних каналів Президента й Верховної Ради України, а також за даними аналітичних міжнародних авторитетних досліджень та аналітичних зведень. Для нормалізації вхідних даних, було відкинуто записи довжиною менше 1000 і більше 8000 символів. В результаті фільтрації даних, отримано набір, що складається із 10 000 записів, де 5 000 належать категорії «Пропаганда» та 5 000 категорії «Не пропаганда». Описаний набір даних було використано для навчання моделей нейромереж у межах розробленого методу класифікації текстів за вмістом пропаганди.

Для дослідження ефективності розробленого методу класифікації текстів за вмістом пропаганди нейромережевими моделями глибокого навчання було створено програмну реалізацію для навчання нейромережевої моделі гібридної архітектури та застосунок з графічним інтерфейсом користувача. За допомогою розробленої програмної системи було виконано експерименти для дослідження ефективності розробленого методу класифікації текстів за вмістом пропаганди.

У ході дослідження вдалось досягнути точності у 97.8 %. При використанні лише двох категорій – «пропаганда» та «не пропаганда», найбільша кількість помилок зосереджена в проміжку 0.4 – 0.6 (рис. 3). Тому введення категорії «підозрілий текст» із можливістю плаваючих меж адаптивним предметній області є доцільним.

Завдяки уведенню категорії «підозрілий текст», показник Accuracy незначно зменшився, однак показник Precision та Recall зросли, що свідчить про можливість ефективної автоматизованої модерації текстів на предмет пропаганди. Із застосуванням аугментації кращі показники досягаються при більшій кількості епох. Це пояснюється розширенням навчальної вибірки, що призводить до потреби більшої кількості епох. В той же час, при застосуванні аугментації вдалося

досягнути точності 97.83 %, при тому що без аугментації цей показник максимально досяг рівня 96.94% [8, 9].

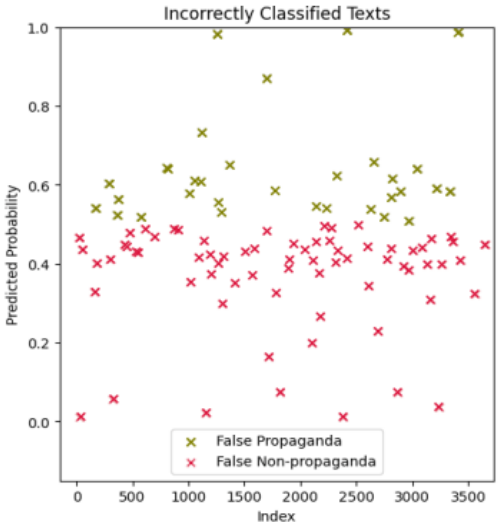

Рисунок 3 – Розподіл некоректно ідентифікованих текстів до категорій «пропаганда» та «не пропаганда»

Отже, було запропоновано метод класифікації текстів за вмістом пропаганди нейромережевими моделями глибокого навчання, який дозволяє виявляти як явні, так і приховані пропагандистські меседжі, ґрунтуючись на об'єднанні рекурентних нейромереж довгострокової пам'яті із трансформерами, а також на використанні механізму аугментації навчальних текстових даних, що дозволяє розширити кількість навчальних зразків. Отримані результати дослідження ефективності розробленого методу свідчать про спроможність запропонованого методу ефективно класифікувати тексти за вмістом пропаганди нейромережевими моделями глибокого навчання й розроблений метод може використовуватись для оцінки потенційних загроз, пов'язаних з поширенням пропаганди. Застосування додаткової категорії «підозрілий текст» дозволило підняти показники Precision та Recall, що у свою чергу дає можливість автоматизованої модерації текстів на предмет пропаганди з помилками не більше 1.83% для хибного виявлення пропаганди.

Перелік посилань

1. Faye G., Icard B., Casanova M., Chanson J., Maine F., Bancilhon F., Gadek G., Gravier G., Egre P. Exposing propaganda: an analysis of stylistic cues comparing human annotations and

machine classification. Proceedings of the Third Workshop on Understanding Implicit and Underspecified Language, 2024. pp. 62–72.
2. Vijayaraghavan P., Vosoughi, S. TWEETSPIN: Fine-grained Propaganda Detection in Social Media Using Multi-View Representations. Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2022, pp. 3433–3448.
3. Martino G., Yu S., Barron-Cedeno A., Petrov R., Nakov, P. Fine-Grained Analysis of Propaganda in News Article. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019, pp. 5640–5650.
4. Молчанова М. Метод виявлення та класифікації прийомів пропаганди у текстовому контенті засобами штучного інтелекту. Матеріали XII Міжнародної науково-практичної конференції «Інформаційні управляючі системи та технології ІУСТ-ОДЕСА-2024». 23-25.09.2024. Одеса. 2024. С.251-254.
5. Молчанова М.О. Дослідження ефективності методу класифікації текстів за вмістом пропаганди нейромережевими моделями глибокого навчання. Інформаційні технології і автоматизація. Матеріали XVII міжнародної науково-практичної конференції. 31 жовтня – 1 листопада 2024 р. Одеса, ОНТУ. 2024. С.665-668.
6. Krak I., Molchanova M., Mazurets O., Sobko O., Zalutska O., Barmak O. Method for Neural Network Detecting Propaganda Techniques by Markers With Visual Analytic. CEUR Workshop Proceedings, 2024, vol. 3790, pp. 158-170.
7. Krak I., Didur V., Molchanova M., Mazurets O., Zalutska O., Manziuk E., Barmak O. Method for Political Propaganda Detection in Internet Content Using Recurrent Neural Network Models Ensemble. CEUR Workshop Proceedings, 2024, vol. 3806, pp. 312-324.
8. Молчанова М. О. Застосування аугментації даних для підвищення точності виявлення пропаганди в інтернет-джерелах нейромережевими моделями глибокого навчання. Матеріали VIII Міжнародної науково-практичної конференції «Перспективи сучасної науки: теорія і практика». 16-18.09.2024. Львів – 2024. с. 199-205.
9. Молчанова М.О. Метод виявлення та класифікації технік пропаганди у текстовому контенті засобами штучного інтелекту. Розвитки інформаційно-керуючих систем та технологій.: монографія. Львів-Торунь : Lina-Pres, 2024. – С.245-266.