УДК 004.8

Молчанова М.О., Мурава В.В.

Хмельницький національний університет

ВИЯВЛЕННЯ ШАБЛОНІВ ВЕБ-ПРОПАГАНДИ НЕЙРОМЕРЕЖЕВИМИ МЕТОДАМИ

У роботі розглянуто нейромережевий підхід до виявлення шаблонів веб-пропаганди в україномовних текстах на основі трансформерних моделей BERT та RoBERTa з акцентом на оптимізації балансування навчального датасета між пропагандистськими та нейтральними фрагментами. Запропоновано інтегрований метод аналізу, який працює на рівні речень, оцінює ймовірність наявності окремих пропагандистських технік і формує узагальнені показники їх вираженості на рівні всього тексту. Експериментальні результати на корпусі з розміткою за шаблонами UNLP 2025 засвідчують, що оптимальна частка нейтральних текстів (близько 30%) забезпечує досягнення F_1 = 0,725 для україномовної RoBERTa, що перевищує показники відомих міжнародних підходів.

The paper presents a neural-network-based approach to detecting web propaganda patterns in Ukrainian texts using transformer models BERT and RoBERTa, with a focus on optimizing dataset balancing between propagandistic and neutral fragments. An integrated sentence-level analysis method is proposed, which estimates the likelihood of specific propaganda techniques and aggregates their intensity at the document level. Experiments on a corpus annotated with UNLP 2025 propaganda patterns show that an optimal proportion of neutral texts (around 30%) enables the Ukrainian RoBERTa model to reach an F_1 score of 0.725, outperforming a range of existing international baselines.

Стрімка медіатизація суспільних процесів і перенесення політичної комунікації у веб-простір призвели до якісно нової ролі пропагандистського контенту. Веб-платформи, соціальні мережі та новинні агрегатори стали не лише каналами поширення інформації, а й повноцінними інфраструктурами для масштабованого маніпулятивного впливу на громадську думку, електоральну поведінку та уявлення про ключові події [1]. Алгоритмічні стрічки, персоналізовані рекомендації та таргетована реклама підсилюють ефект «інформаційних бульбашок» і сприяють тому, що користувач стикається з пропагандистськими повідомленнями навіть без усвідомлення їх маніпулятивної природи [2, 3].

За таких умов завдання автоматизованого виявлення веб-пропаганди переходить із площини суто теоретичної проблеми медіазнавства в зону критично важливої інформаційної безпеки [4]. Традиційні методи аналізу тексту [5, 6], що спираються на поверхневі статистичні ознаки [7] або фіксовані словники [8], виявляються недостатніми: сучасні пропагандистські повідомлення оперують тонкими семантичними натяками [9], контекстуальними рамками [10, 11], міжтекстовими відсиланнями, стилістичними прийомами [12, 13], які важко

зафіксувати жорстко заданими правилами. У цих умовах саме трансформерні мовні моделі [14] стали базовим інструментом, здатним враховувати широкий контекст, латентні семантичні зв'язки та стилістичні патерни.

Водночас ефективність нейромережевих моделей напряму залежить від структури та збалансованості навчальних вибірок [15]. Для завдання виявлення пропаганди проблема класового дисбалансу є особливо гострою: тексти з маніпулятивними патернами [16], як правило, займають меншу частку інформаційного потоку [17], тоді як нейтральні повідомлення масово переважають. Крім того, окремий текст може містити кілька різних типів пропагандистських технік одночасно [18, 19], що зумовлює багатоміткову природу завдання й робить класичну багатокласову класифікацію непридатною [20].

Запропонований у дослідженні підхід спрямований на підвищення ефективності виявлення шаблонів веб-пропаганди в україномовних текстах на основі трансформерних моделей BERT та RoBERTa через цілеспрямовану оптимізацію балансування датасета, зокрема шляхом варіювання частки нейтральних текстів у навчальній вибірці.

Автоматизоване виявлення пропаганди в текстах розвивалося від класичних методів обробки природної мови до сучасних глибоких нейромережевих архітектур. На ранніх етапах дослідження фокусувалися на словникових підходах, аналізі полярності, індикаторах суб'єктивності, модальності й емоційної забарвленості висловлювань. Такі методи дозволяли фіксувати загальну настанову тексту, але були малочутливими до конкретних пропагандистських технік, як-от «апеляція до страху», «фальшива дилема», «підміна причинно-наслідкових зв'язків» тощо.

Перехід до корпусних досліджень і появи спеціалізованих датасетів, зокрема новинних корпусів із розміткою пропагандистських прийомів, відкрив можливість більш тонкої класифікації патернів. У низці робіт було показано, що використання класичних векторних представлень (TF-IDF, word2vec, doc2vec) у поєднанні з ансамблевими алгоритмами машинного навчання (SVM, Random Forest, Stacking Classifier) дозволяє досягати прийнятних значень Accuracy і F_1-міри, однак залишає обмеження щодо контекстного розуміння й переносимості моделей між доменами.

Суттєвий прогрес забезпечили трансформерні моделі BERT-подібної архітектури. Їх застосування дало змогу враховувати контекст кожного токена щодо всього тексту, моделювати семантичні та синтаксичні зв'язки й ефективно розрізняти як загальний тон повідомлення, так і специфічні маніпулятивні структури. Розроблені рішення на базі BERT, RoBERTa, XLM-RoBERTa, GPT-подібних моделей демонстрували зростання F1-міри у порівнянні з попередніми підходами, зокрема в задачах класифікації пропагандистських технік за еталонними наборами SemEval та низкою мульти- і мультимодальних корпусів.

Разом із тим у більшості робіт ключова увага приділялася або мультимодальності (поєднанню тексту з візуальними даними, зокрема мемами та зображеннями), або багатомовності (розширенню моделей на кілька мов за допомогою спільних просторових репрезентацій), або мета-навчанню й доменно-

орієнтованому донавчанню. Натомість питання організації навчальної вибірки, зокрема співвідношення між пропагандистськими та нейтральними текстами, розглядалося переважно в контексті аугментації (EDA, back-translation, синтетичні приклади) й компенсації дисбалансу через ваги класів.

Практично відсутні дослідження, де цілеспрямовано аналізується вплив частки текстів без проявів пропаганди у навчальному наборі на здатність моделей розрізняти:

– тексти з конкретним патерном;

– тексти з іншими патернами;

– нейтральні тексти без пропагандистських технік.

Ця прогалина особливо критична для україномовного інформаційного простору, де наявні корпуси маніпулятивних матеріалів лише формуються, а питання збалансованості датасетів у розрізі конкретних патернів і нейтральних повідомлень набуває не лише наукового, а й практичного значення для протидії дезінформації.

Метою дослідження є підвищення ефективності виявлення шаблонів веб-пропаганди в україномовних веб-текстах за допомогою трансформерних нейромережевих моделей шляхом оптимізації балансування датасета, зокрема встановлення оптимальної частки нейтральних текстів у навчальній вибірці.

Початковий набір характеризується нерівномірним розподілом прикладів за патернами: окремі техніки («Loaded Language», «Cherry Picking») представлені значно більше, тоді як інші («Straw Man», «Thought-Terminating Cliché») мають суттєво менші обсяги анотованих фрагментів. Окрім того, існує окрема підмножина фрагментів, де не зафіксовано жодного пропагандистського шаблону.

Балансування здійснюється за двох принципів. По-перше, для кожного патерна p_i формується навчальна вибірка, де позитивні приклади репрезентують реальний розподіл цього патерна, а негативні приклади відбираються так, щоб забезпечити задану частку m нейтральних фрагментів серед усіх негативів. При цьому до негативів входять як нейтральні тексти, так і фрагменти, де присутні інші патерни, але відсутній p_i.

По-друге, формується спеціальна валідаційна вибірка, у якій усі патерни представлені рівномірно, а кількість нейтральних фрагментів співмірна з сумарною кількістю пропагандистських прикладів. Така конструкція дозволяє об'єктивно оцінити, чи не «плутає» модель окремі патерни між собою та чи зберігається її здатність розрізняти маніпуляційні й нейтральні тексти при зміні m.

Критично важливим є те, що розмітка виконується на рівні речень, а не повних текстів. Це дає змогу підвищити щільність корисних прикладів у вибірці та сконцентрувати навчання на безпосередніх носіях патернів. Водночас така глибина розмітки породжує окремі обмеження, які будуть розглянуті в обговоренні.

Побудований метод (рисунок 1) на основі бінарних моделей використовується як єдиний інтегрований інструмент аналізу веб-контенту [21]. Вхідними даними слугує довільний текст – публікація в соціальній мережі, стаття онлайн-видання або фрагмент новинної стрічки.

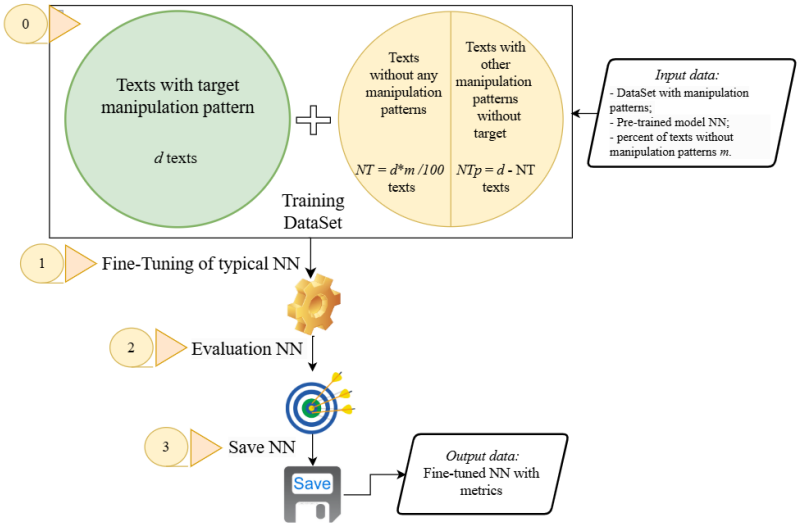

Рисунок 1 – Метод виявлення шаблонів веб-пропаганди

На першому етапі виконується попередня обробка: текст розбивається на окремі речення, здійснюється їх токенізація та перетворення у формат, придатний для подальшого опрацювання трансформерними моделями.

На другому етапі кожне речення незалежно подається на вхід усім бінарним моделям, де кожна з них відповідає окремому пропагандистському шаблону. Для кожної моделі обчислюється ймовірність того, що у відповідному реченні присутня конкретна техніка пропаганди. Якщо ця ймовірність перевищує заздалегідь обраний пороговий рівень, вважається, що речення містить відповідний шаблон. У результаті для кожного речення формується набір пропагандистських технік, які в ньому виявлено.

На третьому етапі здійснюється узагальнення результатів на рівні всього тексту. Для кожного шаблону обчислюються інтегральні показники його вираженості з урахуванням усіх речень, у яких було зафіксовано його прояв. Паралельно виділяються речення, у яких модель оцінює наявність тієї чи іншої техніки як найбільш ймовірну; саме вони можуть демонструватися користувачеві як найхарактерніші приклади. Такий підхід дає змогу не лише встановити факт наявності чи відсутності пропаганди в тексті, а й деталізувати, які саме техніки використовуються та в яких фрагментах вони проявляються найінтенсивніше.

Для об'єктивної оцінки результатів розроблений підхід було зіставлено з низкою відомих рішень, які використовують трансформерні моделі й ансамблеві підходи для виявлення пропаганди в англомовних та мультимовних корпусах. У більшості таких робіт F_1-міра коливається на рівні 0,57–0,60 для складних задач класифікації технік пропаганди за еталонними наборами даних.

На цьому тлі досягнуте значення F_1 = 0,725 для україномовної RoBERTa при m = 30% свідчить про суттєве покращення якості. Важливо, що йдеться не про полегшену задачу «є/немає пропаганди», а саме про виявлення різних шаблонів у межах багатоміткової постановки на складному корпусі з перетином патернів.

Цей результат особливо показовий з огляду на те, що мова йде про українську мову, для якої обсяги маркованих даних традиційно менші, ніж для англійської. Використання попередньо натренованих україномовних моделей у поєднанні з ретельно продуманим балансуванням вибірки компенсує цей недолік і демонструє, що якісні рішення для автоматизованого виявлення пропаганди можуть бути створені й для «малих» мов за умови правильного підходу до структурування даних.

Незважаючи на отримані високі показники, розроблений підхід має низку принципових обмежень.

По-перше, аналіз здійснюється на рівні речень. Це дозволяє чітко локалізувати патерни та збільшити щільність позитивних прикладів, однак водночас ігнорує ширший дискурсивний контекст. Частина маніпулятивних технік працює саме через композицію кількох речень, розміщення інформації у певній послідовності, контрастування з попереднім або наступним контекстом. Речення, яке в ізоляції виглядає нейтральним, у контексті пропагандистського наративу може виконувати ключову роль, але така роль не фіксується при суто реченнєвому аналізі.

По-друге, корпус розмічений вручну, що неминуче містить елемент суб'єктивності. Різні експерти можуть по-різному інтерпретувати межу між, наприклад, «Loaded Language» та емоційно забарвленими, але не пропагандистськими висловлюваннями. Наявність суб'єктивних похибок у розмітці впливає як на процес навчання, так і на оцінювання, формуючи верхню межу досяжної якості.

По-третє, у дослідженні розглядалася дискретна шкала значень m (10%, 30%, 50%, 70%). Теоретично, оптимальне значення може лежати між цими точками, а також залежати від конкретної архітектури, типу патерна чи домену текстів. Подальші дослідження можуть передбачати використання більш дрібного кроку варіювання m або застосування методів байєсівської оптимізації для пошуку m* в неперервному просторі.

Перспективним напрямом є також перехід від суто реченнєвого аналізу до ієрархічних моделей, які поєднують рівень речень і рівень документа. Такі моделі можуть спочатку виявляти локальні патерни, а потім інтегрувати їх у глобальний дискурсивний контекст, оцінюючи, як локальні маніпуляції формують загальну рамку повідомлення.

Окремим вектором розвитку є інтеграція пояснювальних механізмів. Для практичних застосувань (модерація, фактчекінг, медіаграмотність) важливо не лише автоматично детектувати патерн, а й пояснити користувачеві, чому модель вважає той чи інший фрагмент маніпулятивним, які мовні маркери чи структурні особливості на це вказують. Застосування методів пояснювальної ШІ (attention-

візуалізація, градієнтні карти, локальні сурогатні моделі) може суттєво підвищити довіру до систем виявлення пропаганди.

Отримані результати мають як наукове, так і практичне значення. З наукового боку продемонстровано, що зміна частки нейтральних текстів у навчальній вибірці є окремим параметром, який суттєво впливає на узагальнювальну здатність моделей виявлення пропаганди й повинен розглядатися як об'єкт оптимізації нарівні з гіперпараметрами навчання. З практичного боку запропонований підхід може бути покладений в основу програмних систем для автоматизованого моніторингу інформаційного простору, підтримки фактчекінгових ініціатив, а також освітніх інструментів із розвитку медіаграмотності, де користувачі отримуватимуть не лише індикатор наявності пропаганди, а й марковані приклади конкретних маніпулятивних технік.

Подальші дослідження доцільно спрямувати на розширення корпусу за рахунок інших доменів (соціальні мережі, коментарі, блоги), перехід до ієрархічного аналізу на рівні документів, інтеграцію мультимодальних даних (текст плюс зображення) та розроблення пояснювальних модулів, що підвищать прозорість і довіру до нейромережевих систем виявлення веб-пропаганди.

Перелік посилань

1. Scorzato L. Reliability and Interpretability in Science and Deep Learning. Minds and Machines. 2024. Vol. 34, no. 3.
2. Geissler, D., Bär, D., Pröllochs, N., &Feuerriegel, S. "Russian propaganda on social media during the 2022 invasion of Ukraine". EPJ Data Science, vol. 12, no. 1, pp. 35, 2023.
3. Lande D. Formation and analysis of networks of events in the field of parliamentary control based on the application of artificial intelligence systems. INFORMATION AND LAW. 2024. No. 1(48). P. 84–89.
4. Marchenko O., Isoieva M. Automatic Generation of Coherent Natural Language Texts. Flexible Query Answering Systems. Cham, 2023. P. 79–92.
5. Lawrence G. The Power to Lie: Propaganda and Post-truth Politics. Societal Deception. London, 2024. P. 211–270.
6. Shevchuk P., Molchanova M., Mazurets O. Software for Text Messages Reliability Analysis Based on the Machine Learning Models Ensemble. Proceedings of IV International Scientific and Practical Conference «Innovative research and perspectives of the development of science and technology». January 29-31, 2024. Stockholm, Sweden. 2024. Pp. 347-354.
7. 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.
8. 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.
9. 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.

10. 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.
11. 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.
12. Murava V., Zalutska O., Didur V., Mazurets O. Software architecture of information system for exchanging LLM thematic prompts. Global Trends in the Development of Information Technology and Science. Proceedings IV International Scientific and Practical Conference. June 25-27, 2025. Stockholm, Sweden. Pp. 121-127.
13. Молчанова М.О., Мазурець О.В., Собко О.В., Кліменко В.І., Андрощук В.І. Метод нейромережевого виявлення кібербулінгу з використанням хмарних сервісів та об'єктно-орієнтованої моделі. Науковий журнал «Вісник Хмельницького національного університету» серія: Технічні науки. Хмельницький, 2024. №2 (333). С. 200-206.
14. 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.
15. 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.
16. Молчанова М.О., Мазурець О.В., Собко О.В., Віт Р.В., Назаров В.В. Алгоритм виявлення аб'юзивного вмісту в україномовному аудіоконтенті для імплементації в об'єктно-орієтовану інформаційну систему. Науковий журнал «Вісник Хмельницького національного університету» серія: Технічні науки. Хмельницький, 2024. №1 (331). С. 101-106.
17. Денисенко Б.О., Молчанова М.О., Мазурець О.В. Інтелектуальна система виявлення дезінформації з застосуванням штучних нейронних мереж. Збірник наукових праць за матеріалами XVI Всеукраїнської науково-практичної конференції «Актуальні проблеми комп'ютерних наук АПКН-2024». 15-16 листопада 2024. Хмельницький, 2024. с. 167-174.
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. 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.
20. Крак Ю.В., Дідур В.О., Молчанова М.О., Мазурець О.В., Собко О.В., Залуцька О.О., Бармак О.В. Метод виявлення політичної пропаганди в інтернет-контенті нейромережевими засобами обробки природної мови. Науковий журнал «Проблеми програмування». Київ, 2024, №2-3. с. 288-295.
21. Molchanova M., Didur V., Sobko O., Mazurets O. Detection of Web Propaganda Patterns by Transformer Neural Networks: Improving Efficiency via Dataset Balancing, CEUR Workshop Proceedings, 2025, vol. 3988, pp. 112-126.