УДК 004.8 DOI https://doi.org/10.35546/kntu2078-4481.2024.4.42

М. О. МОЛЧАНОВА
викладач кафедри комп'ютерних наук
Хмельницький національний університет
ORCID: 0000-0001-9810-936X

ВИЯВЛЕННЯ ТА КЛАСИФІКАЦІЯ ТЕХНІК І ОБ'ЄКТІВ ПРОПАГАНДИ В ТЕКСТОВИХ ПОВІДОМЛЕННЯХ ЗАСОБАМИ МАШИННОГО НАВЧАННЯ

У статті запропоновано підхід до виявлення та класифікації технік і об'єктів пропаганди в текстових повідомленнях засобами машинного навчання, що полягає у виконанні послідовних кроків та дозволяє як виявляти застосування технік в цілому, так і виконувати класифікацію застосованих технік, а також виявляти об'єкти, на які спрямовані класифіковані техніки пропаганди.

Для виявлення застосованих технік пропаганди було використано гібридну модель машинного навчання на основі об'єднання архітектур BiLSTM та шарів архітектури трансформер. Застосоване поєднання забезпечило поглиблене розуміння текстового контенту та сприяло підвищенню виявлення пропаганди на 0.037 в порівнянні з відомими аналогами. Для класифікації технік пропаганди запропоновано як класифікація застосованих технік, так і візуалізація отриманих результатів з використанням алгоритму LIME. Використовується множина моделей машинного навчання, де за класифікацію кожної техніки відповідає окрема навчена модель машинного навчання на базі архітектури трансформерів, а саме BERT-подібні моделі. Таке використання дозволяє класифікувати застосовані техніки з мінімальною оцінкою Accuracy 0.82. Реалізоване виявлення пропагандистських об'єктів дозволяє знаходити не лише на кого спрямована пропаганда, а і на що здійснюється її спрямування в розрізі використаних класифікованих технік. Для інтерпретованості та наочності також забезпечено візуалізацію результатів.

Запропонований у роботі підхід корелює із Цілями сталого розвитку ПРООН та дозволяє автоматизувати процес виявлення та класифікації пропаганди та зробити результати подання повними, інтерпретованими та зрозумілими. Зокрема, виявлення та класифікація технік і об'єктів пропаганди за допомогою методів машинного навчання сприяють досягненню Цілі сталого розвитку ООН № 16 шляхом підвищення прозорості інформаційного простору та зміцнення інституційної довіри, а також Цілі сталого розвитку ООН № 4 через розвиток медіаграмотності та критичного мислення серед населення, що дозволяє ефективно протидіяти дезінформації.

Ключові слова: *NLP, техніки пропаганди, об'єкти пропаганди, візуальна аналітика.*

M. O. MOLCHANOVA
Lecturer of Computer Science Department
Khmelnytskyi National University
ORCID: 0000-0001-9810-936X

DETECTION AND CLASSIFICATION OF PROPAGANDA TECHNIQUES AND OBJECTS IN TEXT MESSAGES USING MACHINE LEARNING

Article proposes approach to detecting and classifying propaganda techniques and objects in text messages using machine learning models, which consists of performing sequential steps and allows both to detect the use of techniques in general and to classify the applied techniques, as well as to detect objects to which the classified propaganda techniques are directed.

To detect the applied propaganda techniques, hybrid machine learning model was used based on the combination of BiLSTM architectures and transformer architecture layers. The applied combination provided in-depth understanding of the text content and contributed to increase in propaganda detection by 0.037 compared to known analogues. To classify propaganda techniques, both the classification of the applied techniques and visualization of obtained results using the LIME algorithm were proposed. Set of machine learning models is used, where separate trained machine learning model based on the transformer architecture is responsible for classification of each technique, namely BERT-like models. This use allows to classify applied techniques with minimum Accuracy score of 0.82. The implemented detection of propaganda objects allows to find not only who the propaganda is directed at, but also what it is directed at in terms of classified techniques used. For interpretability and clarity, visualization of the results is also provided.

Approach proposed in the work correlates with the UNDP Sustainable Development Goals and allows to automate the process of detecting and classifying propaganda and make the results of the presentation complete, interpretable and understandable. In particular, the detection and classification of propaganda techniques and objects using machine learning methods contribute to the achievement of the UN Sustainable Development Goal No. 16 by increasing the transparency of the information space and strengthening institutional trust, as well as the UN Sustainable Development Goal No. 4 by developing media literacy and critical thinking among the population, which allows to effectively counteract disinformation.

Key words: *NLP, propaganda techniques, propaganda objects, visual analytics.*

Постановка проблеми

На сучасному етапі інтернет-технології сприяють швидкому, масовому та ефективному поширенню різного роду пропаганди та маніпулятивних впливів [1, С. 1-10]. Нові методи генерації текстів забезпечують значне зростання обсягу контенту, що зумовлює потребу в постійному моніторингу нових способів створення пропагандистського вмісту та вдосконаленні методів виявлення та класифікації пропагандистських технік. Це є важливим завданням у контексті протидії дезінформації та забезпечення інформаційної безпеки. Виявлення та класифікація пропаганди у медіапросторі часто здійснюється вручну, без аналізу взаємозв'язку між використаними техніками та відповідними їм об'єктами пропаганди, що значно знижує оперативність та якість результатів.

У зв'язку з цим, автоматизація процесу виявлення технік і відповідних їм об'єктів пропаганди з використанням моделей машинного навчання, їх класифікація та візуальне подання отриманих результатів є надзвичайно актуальною задачею сьогодення. У роботі запропоновано підхід до виявлення та класифікації технік і об'єктів пропаганди у текстових повідомленнях засобами машинного навчання, що складається із трьох послідовних кроків та сприяє досягненню Цілей сталого розвитку № 4 та № 16 відповідно до Програми розвитку ООН.

Аналіз останніх досліджень і публікацій

Проблема виявлення пропаганди [2] є актуальним завданням області обробки природної мови, що не поступається значимістю із сучасними викликами семантичного аналізу [3, С. 591-607], такими як виявлення аб'юзивного змісту [4, С. 16-28] та аналіз емоційної тональності тексту [5, С. 344-356; 6, С. 113-116]. Негативна емоційна тональність та наявність аб'юзивного контенту виступають маркерами пропаганди, що вказують на додаткову ймовірність застосування маніпулятивних впливів. Тому всі ці питання є ключовими аспектами розуміння механізмів маніпуляції, що привертає значну увагу сучасних дослідників.

Сучасні науковці виділяють такі основні підходи до виявлення вмісту пропагандистських технік у тексті: на основі пошуку NER та на основі загальної класифікації текстів. Серед основних проблем для якісного навчання моделей машинного навчання, які б могли виявляти та класифікувати пропагандистські техніки, вчені виділяють проблему відсутності маркованих датасетів [7, С. 2668].

При виявленні та класифікації пропаганди як задачі пошуку іменованих сутностей виникає проблема, пов'язана з тим, що пропагандистські текстові фрагменти зазвичай значно довші за стандартні NER і можуть включати десятки слів. В продовження дослідження впливу довжини пропагандистських фрагментів на ефективність виявлення технік пропаганди, у [8] було зазначено, що зі збільшенням фрагментів завдання ускладнюється.

У [9] визнання наявності пропаганди відбуватиметься на двох рівнях: на загальному рівні, тобто на рівні документа, і на рівні окремих речень. Використовуються такі методи побудови ознак, як статистичний індикатор «TF-IDF», модель векторизації «Bag of Words», маркування частин мови, модель «word2vec» для отримання векторних зображень слів, а також розпізнавання тригера-слова. В якості основного алгоритму моделювання використано логістичну регресію, якою розпізнавання пропаганди на рівні документів показала себе майже в 1,2 рази краще (на 20%). Аналіз необроблених даних показав, що модель розпізнавання пропаганди на рівні документа змогла правильно класифікувати 6097 непропагандистських статей і 694 пропагандистські статті. Отримана оцінка моделі: 0,9433. Модель розпізнавання пропаганди на рівні речення успішно класифікувала 205 пропагандистських статей і 1917 непропагандистських статей. Оцінка моделі: 0,7438 (але 731 статтю було класифіковано неправильно).

З проведеного аналізу сучасних наукових надбань в області виявлення та класифікації технік пропаганди у текстових повідомленнях спостерігається відсутність комплексного підходу, який би забезпечив виявлення пропаганди та класифікацію та її технік цілісно з її об'єктами. Отже, виявлення пропаганди через NER не дозволяє розуміти спрямованість відповідно до використаних технік, а класифікація технік на рівні документу незалежно від об'єктів пропаганди не дозволяє зрозуміти на кого і на що ці маніпулятивні впливи націлені.

Формулювання мети дослідження

Метою роботи є розробка підходу до виявлення та класифікації технік і об'єктів пропаганди у текстових повідомленнях засобами машинного навчання. Запропонований підхід відрізняється від існуючих тим, що передбачає як загальну оцінку рівня пропаганди у тексті, так і класифікацію використаних технік спільно із множиною об'єктів пропаганди, на які спрямовані використані техніки.

Викладення основного матеріалу дослідження

Підхід до виявлення та класифікація технік і об'єктів пропаганди в текстових повідомленнях засобами машинного навчання призначений для автоматизації процесу модерації текстових повідомлень на предмет наявності маніпулятивних впливів. Схема та кроки методу представлені на рис. 1. Як видно з рисунку, підхід є послідовним до виконання.

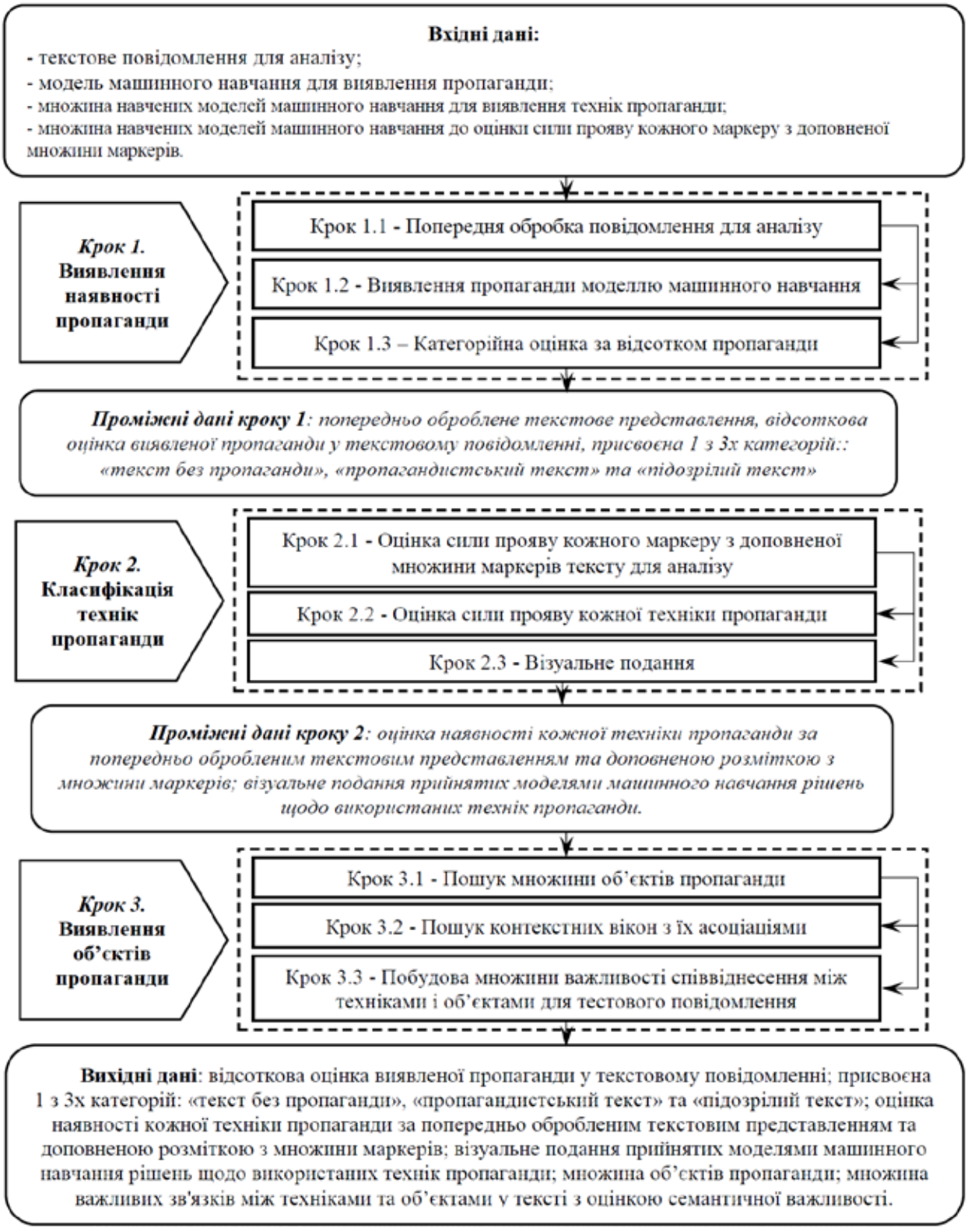

Рис. 1. Схема та кроки методу до виявлення та класифікації технік і об'єктів пропаганди у текстових повідомленнях

На Кроці 1 відбувається виявлення наявності пропаганди шляхом використання гібридної моделі машинного навчання на основі об'єднання архітектур BiLSTM та шарів трансформера. Крок 1 полягає у послідовному перетворенні інформації з текстового повідомлення для аналізу та моделі машинного навчання для виявлення

пропаганди у попередньо оброблене текстове представлення, відсоткову оцінку виявленої пропаганди у текстовому повідомленні та присвоєнні відповідної категорії: «повідомлення без пропаганди», «пропагандистське повідомлення» та «підозріле повідомлення». Для цього текстове повідомлення проходить підкроки попередньої обробки повідомлення для аналізу (видалення стоп-слів, лишніх пробілів тощо), виявлення пропаганди моделлю машинного навчання з гібридною архітектурою та визначення категорійної оцінки за відсотком пропаганди. Категорійні межі з'ясовуються емпірично, однак за замовчуванням рекомендовано такі межі: «повідомлення без пропаганди» від 0 до 0.45, «підозріле повідомлення» має значення від 0.45 до 0.55, а «пропагандистське повідомлення» від 0.55 до 1. Використання архітектури в поєднанні BiLSTM та шарів трансформера дозволяє досягти глибшого розуміння послідовності та контексту в тексті.

Крок 2 присвячений класифікації технік пропаганди за маркерами із використанням візуального представлення результатів. Виконується завдяки застосуванню 17 альтернативних моделей-трансформерів BERT-архітектури машинного навчання, де кожна окрема модель відповідає за ідентифікацію окремої з 17 технік: «Name Calling», «Doubt», «Causal Oversimplification», «Appeal to fear-prejudice», «Exaggeration», «Flag-Waving», «Black and White Fallacy», «Loaded Language», «Red Herring», «Minimisation», «Repetition», «Thought terminating Cliches», «Appeal to Authority», «Slogans», «Labeling», «Reductio ad hitlerum», «Whataboutism». Візуальне представлення результату здійснюється завдяки використанню методу LIME.

Крок 3 відповідає за виявлення об'єктів пропаганди та складається із пошуку множини об'єктів пропаганди, пошуку відповідних до знайдених об'єктів контекстних вікон та побудови множини важливості співвіднесення між використаними техніками пропаганди і знайденими об'єктами для тестового повідомлення. Під об'єктами пропаганди у поточному контексті розуміється не тільки пошук NER (за допомогою бібліотеки Stanza), а і семантично-близьких до NER слів-асоціацій, які знаходяться шляхом застосування моделі FastText.

Для дослідження ефективності запропонованого підходу до виявлення та класифікація технік і об'єктів пропаганди в текстових повідомленнях засобами машинного навчання розроблено відповідне програмне забезпечення. Результати, отримані у вигляді оцінки виявлення пропаганди, класифікованих технік пропаганди з об'єктами їх спрямування були порівняні з висновками експертів у сфері виявлення та протидії пропаганді з. «Центру стратегічних комунікацій» [10].

На рис. 2 представлено приклад використання створеного програмного забезпечення, що демонструє його ефективність та спроможність у виявленні та класифікації пропагандистських технік та об'єктів пропаганди.

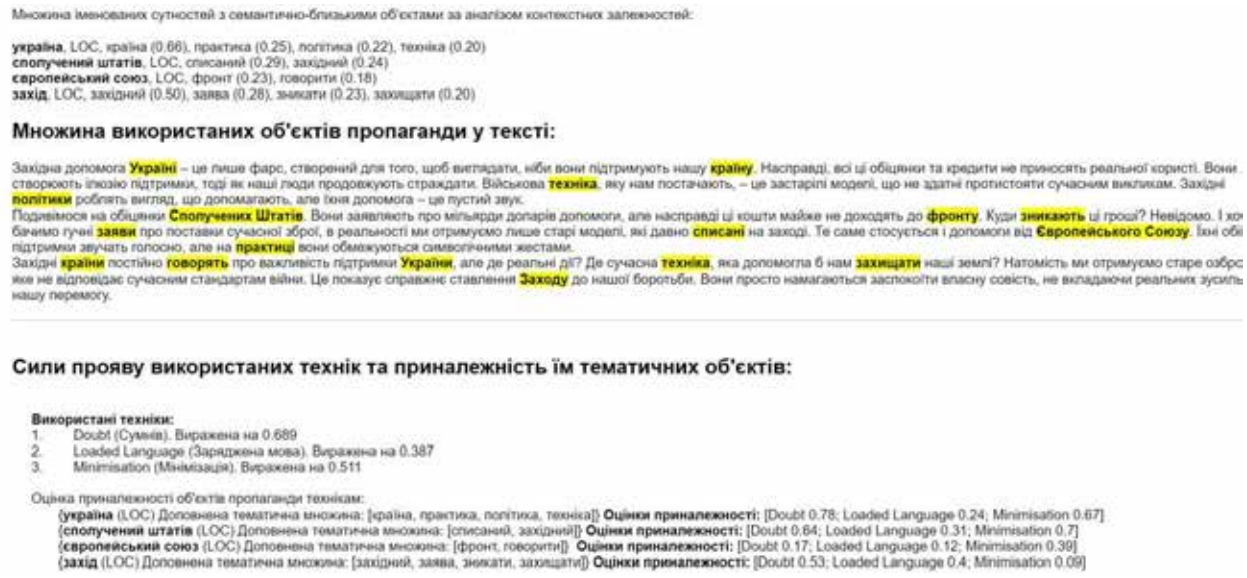

Рис. 2. Приклад використання розробленого застосунку для класифікації використаних технік та об'єктів пропаганди

В ході навчання машинних моделей, що відповідають за класифікацію використаних технік, було досягнуто значень за метрикою Accuracy від 0.82 до 0.97. Більш детально дані дослідження наведені на рис. 3.

Запропонований підхід виявлення і класифікації технік пропаганди та відповідних технікам об'єктів забезпечує виявлення пропаганди з точністю 0.98, що у порівнянні з [9] на 0.037 підвищує ефект виявлення пропаганди у тексті, а техніки пропаганди дозволяє класифікувати з точністю від 0.82 до 0.97, що також переважає відомі аналоги. Ще однією значною перевагою використання такого підходу полягає у поясненності отриманих результатів

та можливості бачити результат комплексно: як загальну оцінку рівня пропаганди у тексті, так і класифікацію використаних технік спільно із множиною об'єктів пропаганди, на які спрямовані використані техніки із оцінками відповідності до використаних технік.

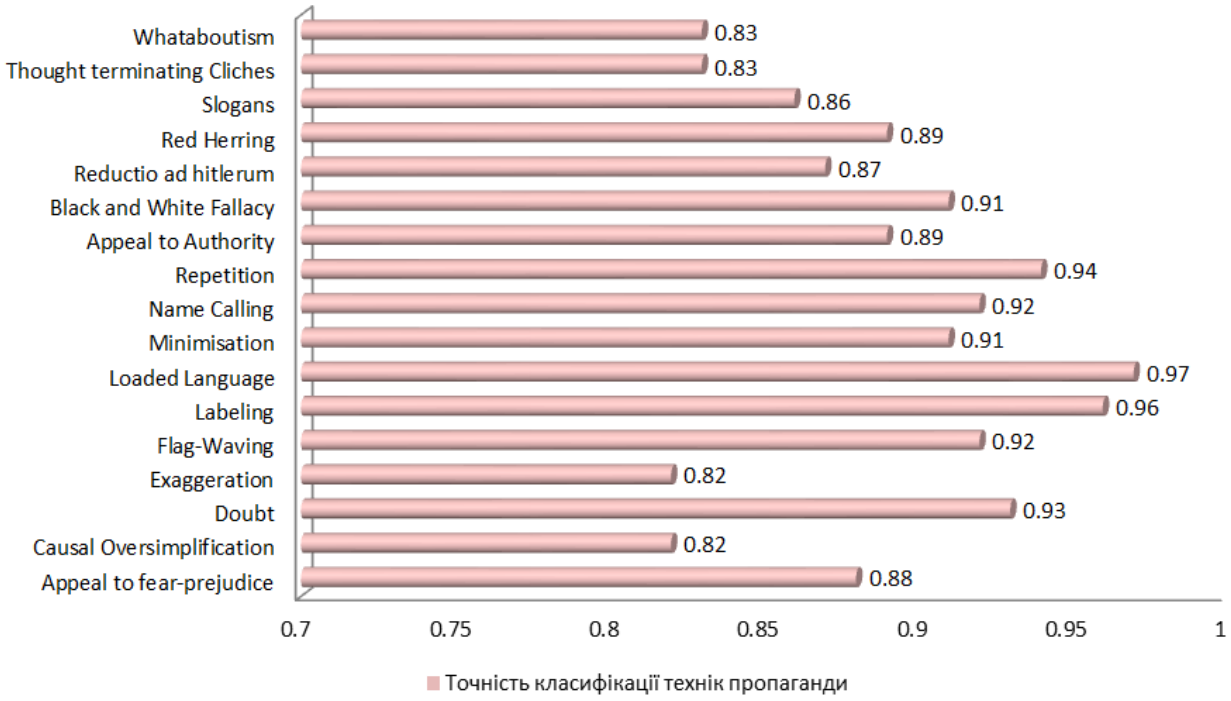

Рис. 3. Точність класифікації використаних технік

Висновки

У статті запропоновано підхід до виявлення та класифікація технік і об'єктів пропаганди в текстових повідомленнях засобами машинного навчання, що полягає у виконанні ряду послідовних кроків, яке забезпечує як виявлення застосування технік в цілому, так і виконання класифікації застосованих технік та виявлення об'єктів, на які спрямовані класифіковані техніки пропаганди.

Для виявлення застосованих технік пропаганди було використано гібридну модель машинного навчання на основі об'єднання архітектур BiLSTM та шарів архітектури трансформер. Застосоване поєднання забезпечило поглиблене розуміння текстового контенту та сприяло підвищенню виявлення пропаганди на 0.037 в порівнянні з відомими аналогами. Для класифікації технік пропаганди запропоновано як класифікація застосованих технік, так і візуалізація отриманих результатів з використанням алгоритму LIME. Використовується множина моделей машинного навчання, де за класифікацію кожної техніки відповідає окрема навчена модель машинного навчання на базі архітектури трансформерів, а саме BERT-подібні моделі. Таке використання дозволяє класифікувати застосовані техніки з мінімальною оцінкою Accuracy 0.82. Реалізоване виявлення пропагандистських об'єктів дозволяє знаходити не лише на кого спрямована пропаганда, а і на що здійснюється її спрямування в розрізі використаних класифікованих технік. Для інтерпретованості та наочності також забезпечено візуалізацію результатів.

Запропонований підхід сприяє виконанню Цілей сталого розвитку ПРООН шляхом автоматизації процесу виявлення та класифікації технік пропаганди з їх об'єктами. Зокрема, виявлення та класифікація технік і об'єктів пропаганди за допомогою методів машинного навчання сприяють досягненню Цілі сталого розвитку ООН № 16 шляхом підвищення прозорості інформаційного простору та зміцнення інституційної довіри, а також Цілі сталого розвитку ООН № 4 через розвиток медіаграмотності та критичного мислення серед населення, що дозволяє ефективно протидіяти дезінформації.

Список використаної літератури

1. Ahmad P.N., Yuanchao L., Aurangzeb K. et al. Semantic web-based propaganda text detection from social media using meta-learning. SOCA. 2024. Vol. 11761. C. 1–10. https://doi.org/10.1007/s11761-024-00422-x.

2. Malik M.S.I., Imran T., Mona Mamdouh J. How to detect propaganda from social media? Exploitation of semantic and fine-tuned language models. *PeerJ Computer Science. 2023. Vol. 9.* e1248. https://doi.org/10.7717/peerj-cs.1248.

3. Kovalchuk O., Slobodzian V., Sobko O. et al. Visual Analytics-Based Method for Sentiment Analysis of COVID-19 Ukrainian Tweets. *Lecture Notes on Data Engineering and Communications Technologies. 2023. Vol. 149*. С. 591–607.

4. Krak I., Zalutska O., Molchanova M., Mazurets O., Bahrii R., Sobko O., Barmak O. Abusive Speech Detection Method for Ukrainian Language Used Recurrent Neural Network. *CEUR Workshop Proceedings. 2024. Vol. 3688.* С. 16–28. https://doi.org/10.31110/COLINS/2024-3/002.

5. Zalutska O., Molchanova M., Sobko O., Mazurets O., Pasichnyk O., Barmak O., Krak I. Method for Sentiment Analysis of Ukrainian-Language Reviews in E-Commerce Using RoBERTa Neural Network. *CEUR Workshop Proceedings. 2023. Vol. 3387*. С. 344–356.

6. Молчанова М.О., Залуцька О.О., Бармак О.В. Метод інтелектуального аналізу тональності текстів. *Матеріали XII Всеукраїнської науково-практичної конференції «Глушковські читання».* Київ, 2023. С. 113–116.

7. Ahmad P.N. Robust Benchmark for Propagandist Text Detection and Mining High-Quality Data. *Mathematics. 2023. Vol. 11.* С. 2668.

8. Przybyla P. Long Named Entity Recognition for Propaganda Detection and Beyond. *Proceedings of the International Conference of the Spanish Society for Natural Language Processing.* 2023.

9. Інформаційна технологія розпізнавання пропаганди, фейків та дезінформації у текстовому контенті на основі методів NLP та машинного навчання. 2024. URL: https://openurl.ebsco.com/EPDB%3Agcd%3A11%3A14025707/detailv2?sid=ebsco%3Aplink%3Ascholar&id=ebsco%3Agcd%3A178891833&crl=c (дата звернення: 24.11.2024).

10. Центр стратегічних комунікацій 2024. URL: https://spravdi.gov.ua/ (дата звернення: 24.11.2024).

References

1. Ahmad, P. N., Yuanchao, L., Aurangzeb, K., et al. (2024). Semantic web-based propaganda text detection from social media using meta-learning. *SOCA, 11761,* pp. 1–10. https://doi.org/10.1007/s11761-024-00422-x.

2. Malik, M. S. I., Imran, T., & Mona Mamdouh, J. (2023). How to detect propaganda from social media? Exploitation of semantic and fine-tuned language models. *PeerJ Computer Science, 9,* e1248. https://doi.org/10.7717/peerj-cs.1248.

3. Kovalchuk, O., Slobodzian, V., Sobko, O., et al. (2023). Visual analytics-based method for sentiment analysis of COVID-19 Ukrainian tweets. *Lecture Notes on Data Engineering and Communications Technologies, 149*, pp. 591–607.

4. Krak, I., Zalutska, O., Molchanova, M., Mazurets, O., Bahrii, R., Sobko, O., & Barmak, O. (2024). Abusive speech detection method for Ukrainian language used recurrent neural network. *CEUR Workshop Proceedings, 3688*, pp. 16–28. https://doi.org/10.31110/COLINS/2024-3/002.

5. Zalutska, O., Molchanova, M., Sobko, O., Mazurets, O., Pasichnyk, O., Barmak, O., & Krak, I. (2023). Method for sentiment analysis of Ukrainian-language reviews in e-commerce using RoBERTa neural network. *CEUR Workshop Proceedings, 3387*, pp. 344–356.

6. Molchanova, M. O., Zalutska, O. O., & Barmak, O. V. (2023). Metod intelektualnoho analizu tonalnosti tekstiv [Intellectual text sentiment analysis method.]. *Materialy XII Vseukrainskoi naukovo-praktychnoi konferentsii «Hlushkovski chytannia», Kyiv*, pp. 113–116 [in Ukrainian].

7. Ahmad, P. N. (2023). Robust benchmark for propagandist text detection and mining high-quality data. *Mathematics, 11,* 2668.

8. Przybyla, P. (2023). Long named entity recognition for propaganda detection and beyond. *Proceedings of the International Conference of the Spanish Society for Natural Language Processing*.

9. Informatsiina tekhnolohiia rozpiznavannia propahandy, feikiv ta dezinformatsii u tekstovomu kontenti na osnovi metodiv NLP ta mashynnoho navchannia [Information technology for recognizing propaganda, fake news, and disinformation in text content based on NLP methods and machine learning] (2024). URL: https://openurl.ebsco.com/EPDB%3Agcd%3A11%3A14025707/detailv2?sid=ebsco%3Aplink%3Ascholar&id=ebsco%3Agcd%3A178891833&crl=c (Accessed: November 24, 2024) [in Ukrainian].

10. Tsentr stratehichnykh komunikatsii [Strategic Communications Center]. (2024). URL: https://spravdi.gov.ua/ (Accessed: November 24, 2024) [in Ukrainian].