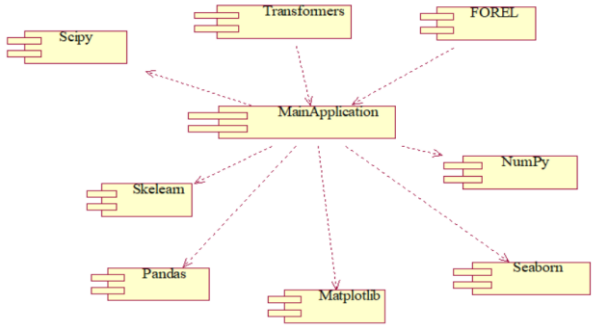

Рисунок 1. *Діаграма компонентів програмного застосування*

За допомогою браузера Google Chrome під час авторизації через обліковий запис gmail здійснюється вхід у хмарний веб-сервіс Google Colab для розгортання, тестування та дослідження системи у вигляді файлу *.ipynb. Для зручності дані завантажуються безпосередньо з хмарного сховища безпосередньо в систему за запитом

Висновки. Результати створеної концепції можуть бути використані при подальшій програмній імлементації модуля на базі застосування різних метрик оцінки якості кластеризації.

Література

[1]. Байраченко О.В. Аналіз призначення та можливостей кластерного аналізу даних для завдань сегментації / О.В. Байраченко, М.Д. Рудніченко // XVI міжнародна науково-практична конференція «Інформаційні технології і автоматизація - 2023» 19-20 жовтня 2023 р., м.Одеса. – 2023. - С.323-324.

[2]. Jannes K. Machine Learning for Finance: Principles and practice for financial insiders. – Packt Publishing, 2019. – 456 p.

УДК 004.8

МЕТОД ВИЯВЛЕННЯ ТА КЛАСИФІКАЦІЇ ПРИЙОМІВ ПРОПАГАНДИ У ТЕКСТОВОМУ КОНТЕНТІ ЗАСОБАМИ ШТУЧНОГО ІНТЕЛЕКТУ

М. Молчанова[0000-0001-9810-936X]
Хмельницький національний університет, Україна
EMAIL: m.o.molchanova@gmail.com

Materials of the XII International Scientific Conference
«Information-Management Systems and Technologies»
23th – 25th September, 2024, Odesa

METHOD FOR DETECTION AND CLASSIFYING OF PROPAGANDA TECHNIQUES IN TEXT CONTENT USING ARTIFICIAL INTELLIGENCE

M. Molchanova
Khmelnytskyi National University, Ukraine

Анотація. *Запропоновано метод виявлення та класифікації прийомів пропаганди за маркерами у текстовому контенті з візуальною інтерпретацією прийнятих рішень, що ґрунтується на використанні набору моделей машинного навчання окремих для кожного прийому пропаганди, що навчаються на модифікованих розмічених даних з доповненою множиною маркерів. Наведено приклад аналізу ефективності, що показує точність запропонованого підходу від 79% до 96% для виявлення окремих прийомів пропаганди.*
Ключові слова: *BERT, прийоми пропаганди, маркери прийомів пропаганди, візуальна інтерпретація отриманих рішень.*

Abstract*. The method for detecting and classifying of propaganda techniques by markers in text content with visual interpretation of the decisions is proposed, based on the use of a set of machine learning models separate for each propaganda technique, which are trained on modified labeled data with a supplemented set of markers. An example of efficiency analysis is given, showing the accuracy of the proposed approach from 79% to 96% for identifying individual propaganda techniques.*
Keywords*: BERT, propaganda techniques, markers of propaganda techniques, visual interpretation of obtained solutions.*

Пропаганда, замаскована під звичайні новини, поширюється протягом багатьох десятиліть, а сучасна цифрова епоха створює додаткові умови для її швидшого, масового та ефективного розповсюдження. Розробляються нові сучасні методи генерації текстів, які дедалі частіше важко відрізнити від створених людиною, що призводить до стрімкого зростання кількості контенту. В свою чергу це підкреслює важливість розробки автоматизованих методів виявлення пропагандистських прийомів, які допоможуть користувачам отримувати інформацію більш усвідомлено [1].

Метою роботи є розробка методу виявлення та класифікації прийомів пропаганди у текстовому контенті, який базується на використанні набору моделей машинного навчання. Пропонується

використовувати окремі створені для кожної пропагандистської техніки моделі машинного навчання [2, 3], що навчені на модифікованих розмічених даних із доповненою множиною маркерів.

Метод виявлення та класифікації прийомів пропаганди призначений для оцінки текстового контенту на предмет наявності прийомів пропаганди та визначення сили їх проявів. Під додатковою множиною маркерів мається на увазі використання різноманітних текстових ознак, які притаманні визначеним прийомам пропаганди. Схема кроків методу наведена на рис. 1.

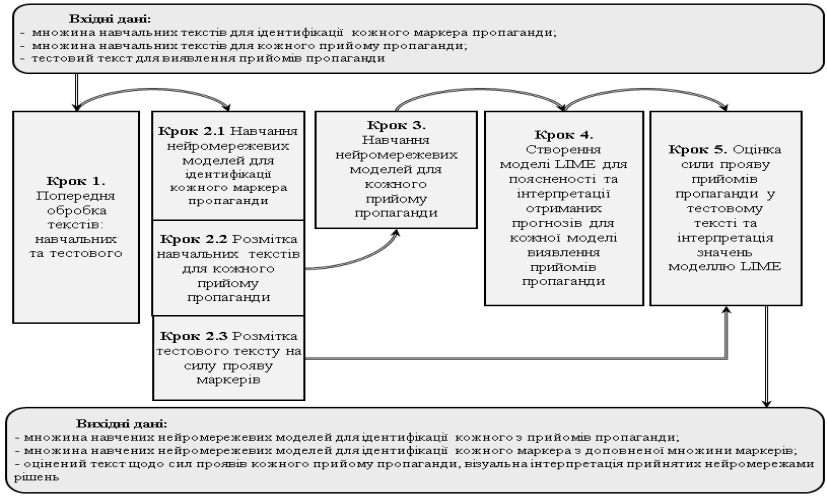

Рисунок 1. *Схема методу виявлення та класифікації прийомів пропаганди*

Першим кроком виконання методу є попередня обробка усіх текстових даних, як навчальних так і текстових. Вона включає видалення знаків пунктуації та видалення стоп-слів. На другому кроці здійснюється навчання нейромережевих моделей для ідентифікації кожного маркера пропаганди, які використовуються для розмітки навчальних текстів для кожного прийому пропаганди, а також для розмітки щодо наявності маркерів тестових текстів. Третім кроком є навчання нейромережевих моделей для кожного прийому пропаганди. Кількість нейромережевих моделей у даному дослідженні складає 17 і покриває такі прийоми пропаганди, як: «Appeal to fear-prejudice»,

«Causal Oversimplification», «Doubt», «Exaggeration», «Flag-Waving», «Labeling», «Loaded Language», «Minimisation», «Name Calling», «Repetition», «Appeal to Authority», «Black and White Fallacy», «Reductio ad hitlerum», «Red Herring», «Slogans», «Thought terminating Cliches», «Whataboutism» [4]. На четвертому кроці відбувається створення моделі LIME для поясненості та інтерпретації отриманих прогнозів для кожної моделі виявлення прийомів пропаганди, які разом із навченими нейромережевими моделями на кроці 3 будуть оцінювати користувацький текст. На п'ятому кроці відбувається нейромережева оцінка сили прояву прийомів пропаганди у тестовому тексті та інтерпретація значень моделлю LIME [5].

Отже, було запропоновано метод виявлення та класифікації прийомів пропаганди, який дозволяє шляхом використання набору з 17 навчених BERT-моделей виявляти 17 прийомів пропаганди з точністю від 79% до 96%. Для навчання BERT-моделей, що виконують функції виявлення прийомів пропаганди, використано набір даних, що представляє собою корпус з 788 новинних статей, анотованих вручну на рівні фрагментів за допомогою вісімнадцяти пропагандистських прийомів.

Література

[1] G. Martino, S. Yu, A. Barron-Cedeno, R. Petrov, P. Nakov, Fine-Grained Analysis of Propaganda in News Article, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019, pp. 5640-5650.

[2] 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, pp.16-28.

[3] 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, pp. 344–356.

[4] Propaganda Analysis Project, 2024. URL: https://propaganda.qcri.org/index.html.

[5] GitHub, Lime, 2024. URL: https://github.com/marcotcr/lime.