УДК 004.8
DOI https://doi.org/10.26661/2786-6254-2024-2-08

МЕТОД НЕЙРОМЕРЕЖЕВОГО ВИЯВЛЕННЯ ПРИЙОМІВ ПРОПАГАНДИ ЗА МАРКЕРАМИ З ВІЗУАЛЬНОЮ ІНТЕРПРЕТАЦІЄЮ ПРИЙНЯТИХ РІШЕНЬ

Молчанова М. О.
викладач кафедри комп'ютерних наук
Хмельницький національний університет
вул. Інститутська, 11, Хмельницький, Україна
orcid.org/0000-0001-9810-936X
m.o.molchanova@gmail.com

Бармак О. В.
доктор технічних наук, професор,
завідувач кафедри комп'ютерних наук
Хмельницький національний університет
вул. Інститутська, 11, Хмельницький, Україна
orcid.org/0000-0003-0739-9678
alexander.barmak@gmail.com

Ключові слова: *BERT, візуальна інтерпретація прийнятих рішень, прийоми пропаганди, маркери прийомів пропаганди.*

У статті запропоновано метод нейромережевого виявлення прийомів пропаганди за маркерами з візуальною інтерпретацією прийнятих рішень, який відрізняється від існуючих тим, що враховує при навчанні нейромережевих класифікаторів додаткову множину маркерів та дозволяє здійснювати візуальну інтерпретацію отриманих результатів. Під додатковою множиною маркерів мається на увазі використання різноманітних текстових ознак, які притаманні визначеним прийомам пропаганди. Кроками методу є попередня обробка усіх навчальних і тестових текстових даних, навчання нейромережевих моделей для ідентифікації кожного маркера пропаганди, навчання нейромережевих моделей для кожного прийому пропаганди, створення моделі для поясненності та інтерпретації отриманих прогнозів для кожної моделі виявлення прийомів пропаганди, нейромережева оцінка сили прояву прийомів пропаганди у тестовому тексті та інтерпретація значень моделлю LIME.

Для дослідження ефективності методу нейромережевого виявлення прийомів пропаганди було створено програмну реалізацію у вигляді набору ноутбуків, реалізованих у хмарному сервісі «Google Colab», що призначені для навчання нейромережевих моделей BERT із подальшим збереженням їх для використання у вебзастосунку для виявлення прийомів пропаганди, а також набору ноутбуків для збереження нейромережевих моделей для виявлення сили прояву маркерів пропаганди. Створений вебзастосунок не лише дозволяє визначити інтенсивність проявів прийомів пропаганди, а і дає можливість здійснювати візуальну аналітику отриманих результатів. Для навчання нейромережевих моделей виявлення прийомів пропаганди використано набір даних, що налічує 550 статей і представляє собою корпус новинних статей, анотованих вручну на рівні фрагментів за допомогою вісімнадцяти пропагандистських прийомів. Дослідження ефективності встановило, що розроблений метод дозволяє шляхом використання набору з 17 навчених BERT-моделей виявляти 17 відповідних прийомів пропаганди з точністю не нижче 81.87%.

METHOD OF NEURAL NETWORK DETECTING OF PROPAGANDA TECHNIQUES BY MARKERS WITH VISUAL INTERPRETATION OF DECISIONS MADE

Molchanova M. O.
Lecturer at the Department of Computer Sciences
Khmelnytskyi National University
Institutska str., 11, Khmelnytskyi, Ukraine
orcid.org/0000-0001-9810-936X
m.o.molchanova@gmail.com

Barmak O. V.
Doctor of Engineering Sciences, Professor,
Head of the Department of Computer Sciences
Khmelnytskyi National University
Institutska str., 11, Khmelnytskyi, Ukraine
orcid.org/0000-0003-0739-9678
alexander.barmak@gmail.com

Key words: *BERT, visual interpretation of decisions made, propaganda techniques, markers of propaganda techniques.*

The article proposes method of neural network detecting of propaganda techniques by markers with visual interpretation of decisions made, which differs from the existing ones in that it takes into account an additional set of markers when training neural network classifiers, and allows visual interpretation of obtained results. The additional set of markers refers to use of various text features inherent in certain propaganda techniques. The steps of the method are pre-processing of all training and test text data, training of neural network models to identify each propaganda marker, training of neural network models for each propaganda technique, creation of model for explanation and interpretation of obtained predictions for each model of detection of propaganda techniques, neural network evaluation of the manifestation strength of propaganda techniques in the test text and interpretation of the values by LIME model.

To effectiveness research of method for detecting propaganda techniques, software implementation was created in the form of the set of notebooks implemented in the Google Colab cloud service, designed for training BERT neural network models and then saving them for use in a web application for detecting propaganda techniques, as well as a set of notebooks for preservation of neural network models to detect the strength of manifestation of propaganda markers. The created web application allows not only determining the intensity of manifestations of propaganda techniques, but also provides an opportunity to perform visual analytics of the obtained results. For training of neural network models for detection of propaganda techniques, the dataset consisting of 550 articles was used, which is the corpus of news articles manually annotated at fragment level using eighteen propaganda techniques. The effectiveness study established that the developed method allows, by using set of 17 trained BERT models, to detect 17 relevant propaganda techniques with an accuracy of no less than 81.87%.

Вступ. Пропаганда, замаскована під звичайні новини, поширюється протягом багатьох десятиліть, а сучасна цифрова епоха створює додаткові умови для її швидшого, масового та ефективного розповсюдження [1]. Розробляються нові сучасні методи генерації текстів, які дедалі частіше важко відрізнити від створених людиною [2], 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.

У статті пропонується метод виявлення пропагандистських прийомів за допомогою викори-

стання моделі комбінацій семантичних маркерів, який базується на використанні набору моделей машинного навчання. Пропонується використовувати окремі створені для кожного конкретного пропагандистського прийому моделі машинного навчання, що навчені на модифікованих розмічених даних із доповненою множиною маркерів.

Огляд літератури. На сучасному етапі науковці працюють над виявленням нових маркерів та нових прийомів пропаганди, а також над покращенням існуючих підходів для її виявлення.

Так, у [1] досліджено основні методи аналізу газетних текстів для виявлення маніпулятивних технологій, що допомагає застерегти від дезінформації та пропаганди. Представлено новий набір еталонних даних чеською мовою для навчання та оцінки сучасних і майбутніх методів розпізнавання 18 маніпулятивних прийомів, таких як нагнітання страху, релятивізація та навішування ярликів. Показано, що поєднання контент-аналізу з запропонованим стильовим аналізом підвищує точність виявлення 15 з 17 оцінених маніпулятивних прийомів від 0.05% до 1.46%.

У [2] наведено багатомовний набір даних про пропаганду та проведено експеримент для дослідження маркерів, за якими людські анотатори та алгоритми класифікації відрізняють пропагандистські статті від непропагандистських на певну тему. Показано, що перебільшення, зменшення описовості та відсутність адекватних джерел часто зустрічаються у пропагандистській пресі. Аналізатор VAGO підтвердив, що використання невизначених маркерів значно корелює з цими особливостями. Виявлено, що моделі машинного навчання ефективні для виявлення пропаганди на певну тему, але потребують покращення щодо пояснюваності та узагальнення на інші теми.

У [3] розглядається застосування моделі MVPROP, що використовує багатовимірні контекстні вбудовування, дозволяє покращити точність виявлення пропаганди. Основним обмеженням є слабкі анотації через великий масштаб даних.

У [4] розглядалося виявлення 17 відомих прийомів пропаганди, за які відповідають певні маркери, які притаманні використовуваним прийомам. В [5] аналізувались можливості використання великих мовних моделей (LLMs), зокрема моделі GPT-3.5-Turbo від OpenAI, для виявлення ознак пропаганди в новинних статтях. Дослідження показало, що технологія LLM може давати розумні висновки про пропаганду, хоча точність виявлення складає всього 25.12% за датасетом SemEval-2022.

Метою роботи є створення методу нейромережевого виявлення прийомів пропаганди за маркерами з візуальною інтерпретацією прийнятих рішень.

Метод нейромережевого виявлення прийомів пропаганди за маркерами. Метод нейромережевого виявлення прийомів пропаганди за маркерами призначений для оцінки текстового контенту на предмет наявності прийомів пропаганди та визначення сили їх проявів. Метод відрізняється від існуючих тим, що враховує при навчанні нейромережевих класифікаторів додаткову множину маркерів та дозволяє здійснювати візуальну інтерпретацію отриманих результатів. Під додатковою множиною маркерів мається на увазі використання різноманітних текстових ознак [6], які притаманні визначеним прийомам пропаганди. У таблиці 1 наведено приклад сили проявів додаткових маркерів «Емоційність тексту», «Булінг», «Страх», «Мова ворожнечі» у прийомах пропаганди.

Схема кроків методу виявлення прийомів пропаганди за маркерами з візуальною інтерпретацією прийнятих рішень наведена на рис. 1. Метод призначений для перетворення вхідних даних у вигляді множини навчальних текстів для ідентифікації кожного маркера пропаганди, множини навчальних текстів для кожного прийому пропаганди та тестового тексту для виявлення прийомів пропаганди у вихідні дані у вигляді множини навчених нейромережевих моделей для ідентифікації кожного з прийомів пропаганди, множини навчених нейромережевих моделей для ідентифікації

Таблиця 1

Сила проявів маркерів для прийомів пропаганди

Прийом пропаганди	Емоційність тексту	Булінг	Страх	Мова ворожнечі
«Appeal to Fear-Prejudice»	Висока	Висока	Висока	Нейтрально
«Causal Oversimplification»	Нейтрально	Нейтрально	Нейтрально	Низька
«Doubt»	Нейтрально	Нейтрально	Висока	Нейтрально
«Exaggeration»	Висока	Нейтрально	Нейтрально	Нейтрально
«Labeling»	Нейтрально	Нейтрально	Нейтрально	Висока
«Loaded Language»	Висока	Нейтрально	Нейтрально	Висока
«Minimization»	Нейтрально	Нейтрально	Нейтрально	Низька
«Name Calling»	Нейтрально	Висока	Нейтрально	Висока
«Reductio ad Hitlerum»	Висока	Нейтрально	Висока	Висока
«Whataboutism»	Нейтрально	Нейтрально	Нейтрально	Висока

кожного маркера з доповненої множини маркерів, та відсоток сили прояву кожного прийому пропаганди в тексті із візуальною інтерпретацією прийнятих нейромережами рішень.

Першим кроком 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 основні прийоми пропаганди, такі 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; 7].

На четвертому кроці відбувається створення моделі LIME для поясненності 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.

Відповідно, вихідними даними є множина навчених нейромережевих моделей для ідентифікації кожного з прийомів пропаганди, множина навчених нейромережевих моделей для ідентифікації кожного маркера з доповненої множини маркерів та оцінений текст щодо сил проявів кожного прийому пропаганди, з візуальною інтерпретацією прийнятих нейромережами рішень.

Підхід до формування множини нейромережевих моделей для виявлення прийомів пропаганди та моделей LIME для інтерпретованості прогнозів навчених моделей наведена на рис. 2. Вхідними даними є множина навчальних текстів для кожного прийому пропаганди та доповнена множина маркерів до кожного з навчальних текстів.

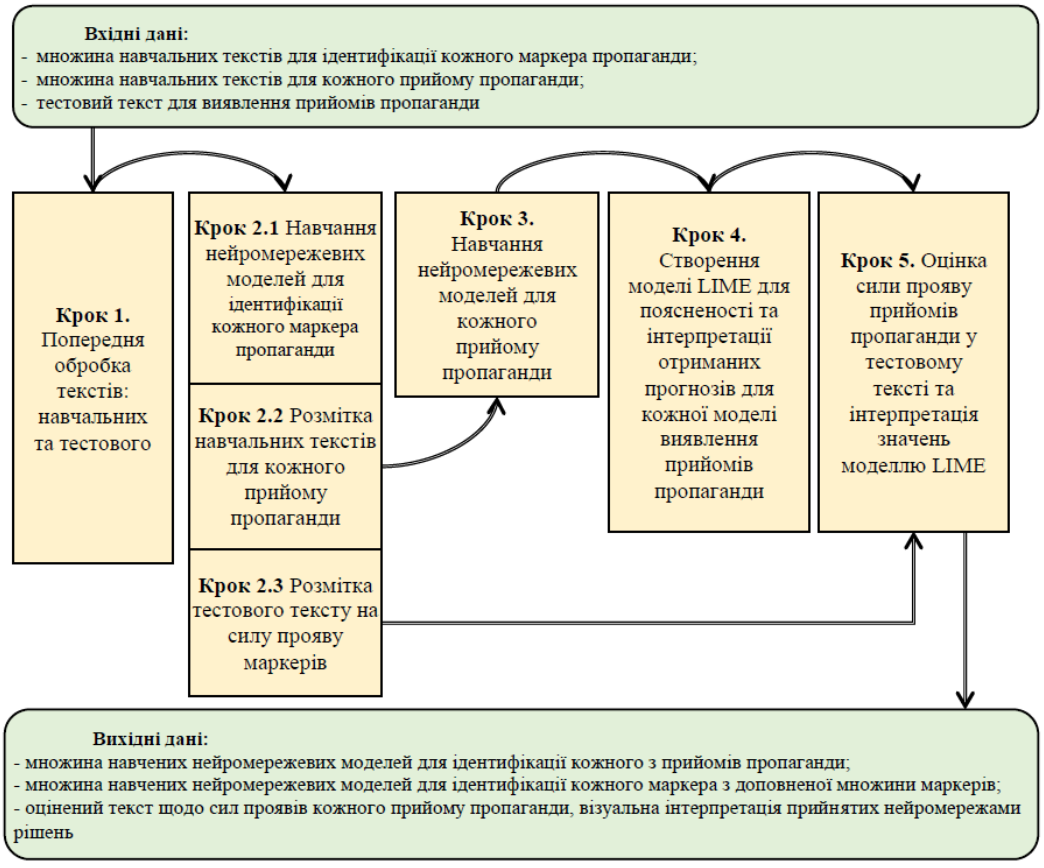

Рис. 1. Схема кроків методу нейромережевого виявлення прийомів пропаганди за маркерами

Першим кроком є попередня обробка навчальних текстів, що включає в себе видалення знаків пунктуації, зайвих пробілів та стоп-слів.

Другим кроком є навчання нейромережевих моделей для кожного прийому пропаганди. У якості нейромережевих моделей було використано BERT-подібні моделі, оскільки даний вид моделей машинного навчання дозволяє розуміти контекст, що є важливим фактором при виявленні прийомів пропаганди.

Було проведено окреме дослідження щодо порівняння BERT-подібних моделей RoBERTa, BERT та ELECTRA. Для цього були використані попередньо натреновані моделі з ресурсу Hugging Face [8], які було донавчені протягом 3-х епох навчання. У табл. 2 наведено одержану ефективність BERT-подібних моделей з виявлення прийомів пропаганди за метрикою Accuracy, згідно з якою обиралися моделі для виявлення окремих прийомів пропаганди.

Після процесу навчання на третьому кроці кожна модель оцінюється за метриками точності, влучності та повноти. Для моделей, що показали високу ефективність за метриками, виконується четвертий крок – створення моделі LIME для пояснення та інтерпретації отриманих прогнозів для кожної моделі виявлення прийомів пропаганди.

Модель LIME слугує як для пояснення текстових даних, так і для числових, дозволяють інтерпретувати, як модель приймає рішення на основі вхідних даних.

Таблиця 2

Ефективність виявлення пропаганди з а метрикою Accuracy

Прийоми пропаганди	**Моделі для виявлення прийомів пропаганди**		
	«bert-base-multilingual-cased»	**«roberta-base»**	**«ukr-electra-base»**
Appeal to fear-prejudice	0.81	0.80	**0.88**
Causal Oversimplification	0.78	0.79	**0.82**
Doubt	**0.93**	0.90	0.87
Exaggeration	0.80	**0.82**	0.80
Flag-Waving	**0.92**	0.90	0.89
Labeling	**0.96**	0.94	**0.96**
Loaded Language	0.93	**0.97**	0.94
Minimisation	0.89	0.86	**0.91**
Name Calling	0.92	0.92	0.91
Repetition	0.93	**0.94**	**0.94**
Appeal to Authority	0.87	**0.89**	0.88
Black and White Fallacy	0.89	**0.91**	0.88
Reductio ad hitlerum	0.85	**0.87**	0.86
Red Herring	0.67	**0.89**	0.78
Slogans	0.84	**0.86**	0.83
Thought terminating Cliches	**0.83**	0.73	0.79
Whataboutism	**0.83**	0.78	0.78

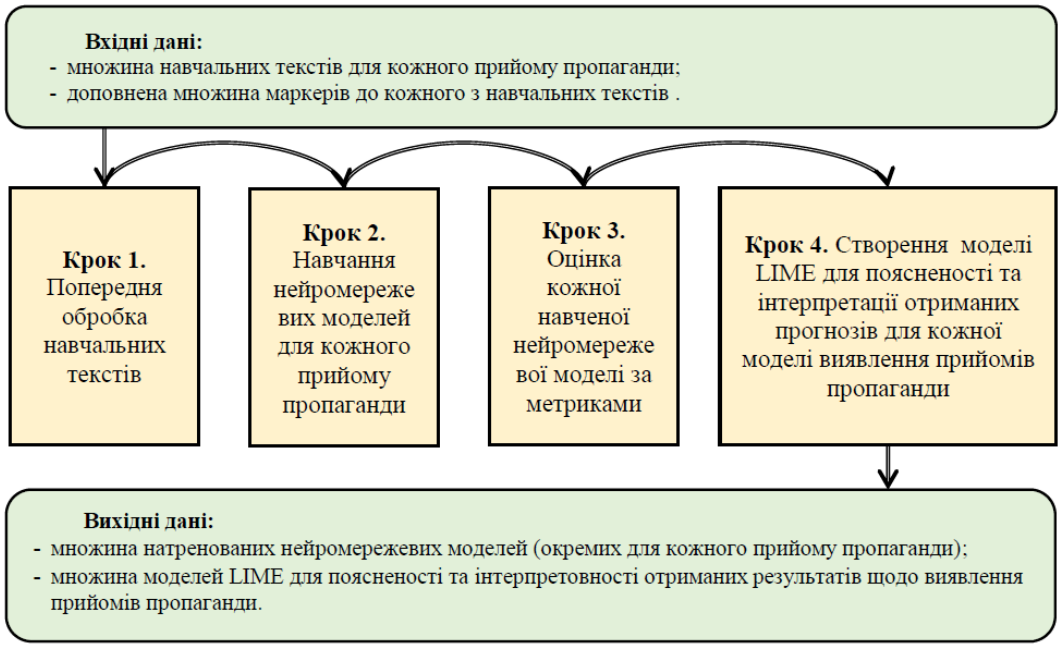

Рис. 2. Схема формування нейромережевих моделей для виявлення прийомів пропаганди та моделей LIME для інтерпретованості прогнозів навчених моделей

Вихідними даними є множина натренованих нейромережевих моделей (окремих для кожного прийому пропаганди) та множина моделей LIME для пояснюваності та інтерпретовності отриманих результатів щодо виявлення прийомів пропаганди нейромережевими моделями.

Формування навчального набору даних для нейромереж виявлення прийомів пропаганди наведено на рис. 3 та відбувається шляхом перетворення вхідних даних у вигляді множини навчених нейромереж до оцінки сили прояву кожного маркеру з доповненої множини маркерів та множини навчальних текстів для кожного прийому пропаганди у вихідні дані у вигляді розміченої множини навчальних текстів для кожного прийому пропаганди за силами прояву кожного з маркерів із доповненої множини маркерів.

Відповідно, вихідними даними є розмічені множина навчальних текстів для кожного із 17 прийомів пропаганди за силами прояву кожного з маркерів із доповненої множини маркерів.

Таким чином, розроблений метод нейромережевого виявлення прийомів пропаганди дозволяє не лише оцінити наявність кожного прийому пропаганди у тексті, а також отримати візуальну пояснюваність отриманих результатів.

Дослідження ефективності методу. Для експериментального дослідження для оцінки ефективності розробленого методу нейромережевого виявлення прийомів пропаганди було створено програмну реалізацію, яка складається із наборів ноутбуків реалізованих у хмарному сервісі «Google Colab», що призначені для навчання нейромережевих моделей BERT із подальшим збереженням їх на жорсткому диску для використання у вебзастосунку для виявлення прийомів пропаганди, а також набору ноутбуків для збереження нейромережевих моделей для виявлення сили прояву маркерів пропаганди. Вебзастосунок реалізовано засобами мови Python використовуючи середовище розробки PyCharm.

Для навчання моделей машинного навчання, що виконують функції виявлення прийомів пропаганди, було використано набір даних «emnlp_trans_uk_dataset», взятий з Kaggle-змагань «Disinformation Detection Challenge» [9]. Набір даних сформований командою «Analysis Project» [10], яка провела аналіз текстів і виявила всі текстові фрагменти, що містять пропагандистські прийоми, а також їх тип. «Analysis Project» створено корпус з 550 новинних статей на основі [11], анотованих вручну на рівні фрагментів за допомогою вісімнадцяти пропагандистських прийомів. Розподіл статей за довжиною в символах наведено на рис. 4.

Як видно з графіку на рисунку 4, для більшості прийомів пропаганди довжина текстів де вони представлені особливої ролі не грають. Однак, «Flag Waving», «Red Herring», «Reductio ad hitlerum» та «Whataboutism» все ж мають меншу максимальну довжину в текстах, де вони представлені.

Для тренування моделей машинного навчання даний датасет було модифіковано таким чином, щоб текст що містить кожен прийом пропаганди був розміщений в окремому каталозі. Після такого перерозподілу було виведено статистику наявних текстів, що репрезентують прийоми пропаганди. Статистика наведена на рис. 5.

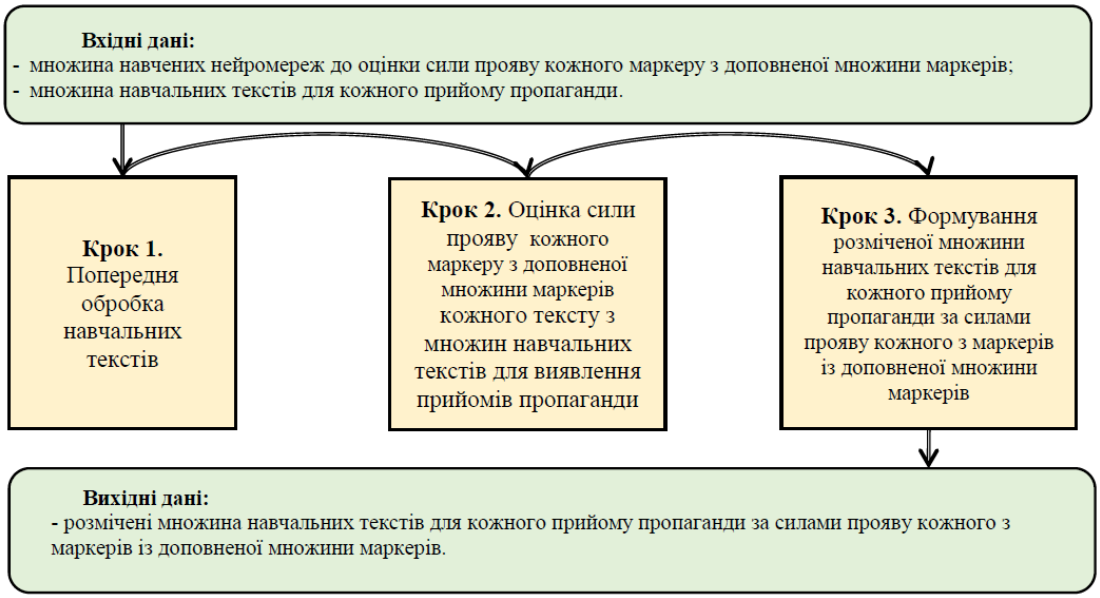

Рис. 3. Кроки для формування навчального набору даних для нейромережевого виявлення прийомів пропаганди

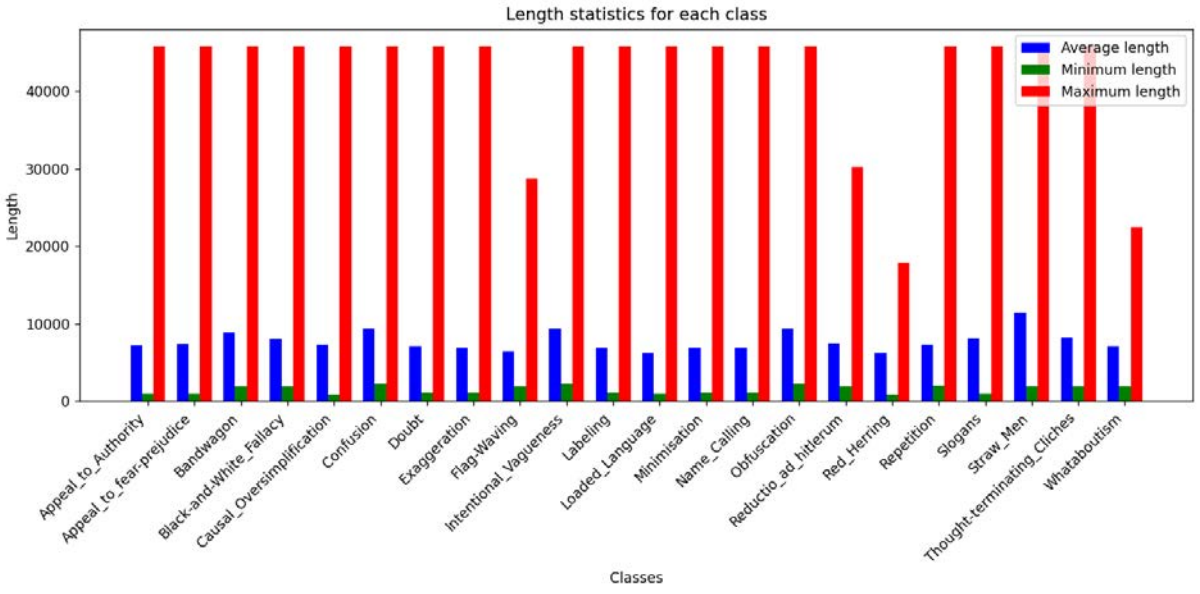

Рис. 4. Статистика за довжиною у символах по прийомам пропаганди

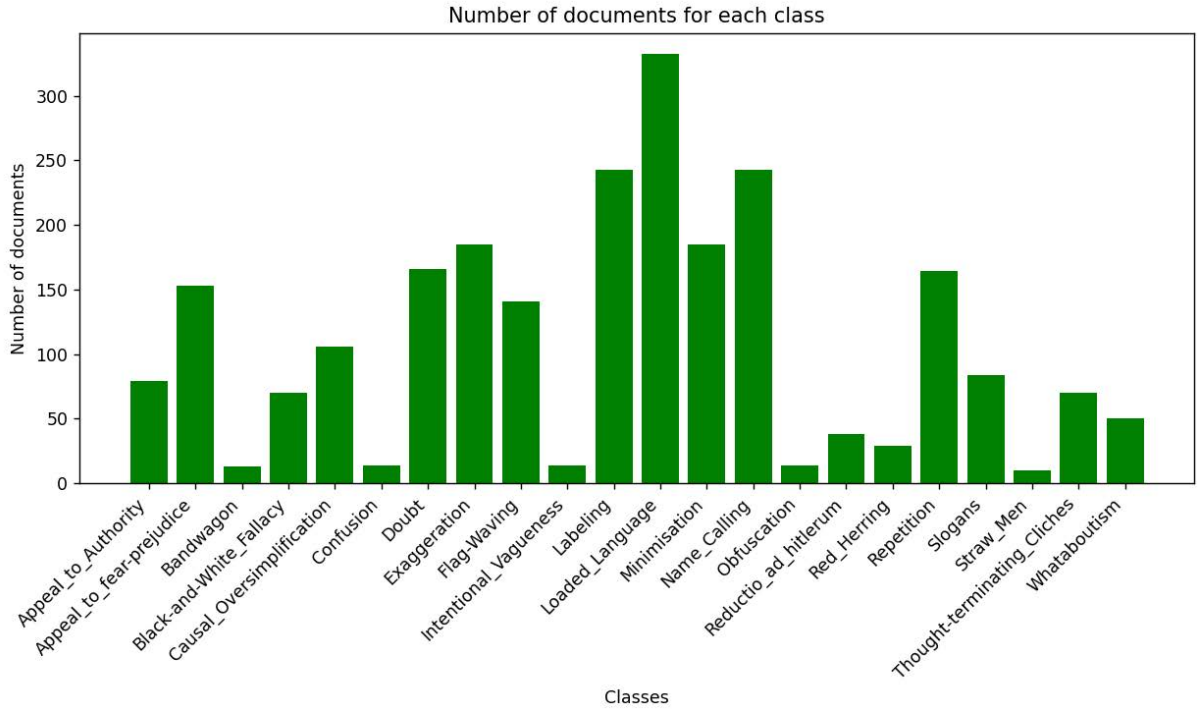

Рис. 5. Статистика по кількості текстів за прийомами пропаганди, шт

Як видно з рис. 5, деякі прийоми пропаганди, такі як «Bandwagon», «Confusion», «Intentional Vagueness», «Obfuscation» та «Straw Men», представлені у критично низькій кількості (менше 20 тестів), тому для них не були створені окремі класифікатори, ці дані було об'єднано у категорію «Інші прийоми пропаганди», однак таким чином, щоб у наявному наборі не були присутні інші прийоми, відмінні від п'яти перерахованих. До прийомів пропаганди, що представлені менш ніж у 100 документах, однак більше ніж 20 було застосовано SMOTE-балансування [12] під час навчання класифікаторів. До таких категорій належать: «Appeal to Authority», «Black and White Fallacy», «Reductio ad hitlerum», «Red Herring», «Slogans», «Thought terminating Cliches» та «Whataboutism».

Із розглянутого набору даних для кожної з 17 типових моделей машинного навчання було сфор-

мовано власний дочірній набір текстів, що має тексти з визначеним прийомом пропаганди та у противагу використовує набір «Інші прийоми пропаганди», доповнений текстами без пропаганди та текстами з іншими прийомами пропаганди. Наприклад, при формуванні набору даних для виявлення прийому «Appeal to fear-prejudice» використовувались 153 документи цільової категорії з цим прийомом і 155 документів нецільової категорії, до якої було віднесено тексти, що містять: інші прийоми пропаганди (15%), «Appeal to Authority» (5%), «Black and White Fallacy» (5%), «Causal Oversimplification» (5%), «Doubt» (5%), «Exaggeration» (5%), «Flag-Waving» (5%), «Labeling» (5%), «Loaded Language» (5%), «Minimisation» (5%), «Name Calling» (5%), «Reductio ad hitlerum» (5%), «Red Herring» (5%), «Repetition» (5%), «Slogans» (5%), «Whataboutism» (5%), «Thought terminating Cliches» (5%), тексти без пропаганди (5%).

Отож у дослідженні буде використано 18 класів: 17 цільових, що є репрезентативними по кількості та відповідають 17 визначеним прийомам пропаганди та 5 об'єднаних в категорію «Інші прийоми пропаганди».

Результати та обговорення. За проведеним експериментом по навченим нейромережевим моделям машинного навчання вдалось досягнути точності виявлення проявів прийомів пропаганди від 0.82 до 0.97 (рис. 6).

Отримані результати забезпечили виявлення різних пропагандистських прийомів з мінімальною точністю 81,87% (мінімальні значення точності отримані для прийому «Causal Oversimplification»), що краще за відомі аналоги [4] щодо виявлення пропаганди незалежно від використовуваних прийомів.

Порівняно з відомими аналогами [5] підвищилась точність виявлення різних пропагандистських прийомів: для прийому «Appeal to Authority» точність виявлення зросла на 10.76% (існуючий метод 77.27%, розроблений метод 88.03%); для прийому «Causal Oversimplification» точність виявлення зросла на 11.99% (існуючий метод 70.1%, розроблений метод 82.09%); для прийому «Doubt» точність виявлення зросла на 75.32% (існуючий метод 17.78%, розроблений метод 93.1%); для прийому «Exaggeration» точність виявлення зросла на 28.04% (існуючий метод 54.17%, розроблений метод 82.21%); для прийому «Flag-Waving» точність виявлення зросла на 27.65% (існуючий метод 64.52%, розроблений метод 92.17%); для прийому «Labeling» точність виявлення зросла на 48.57% (існуючий метод 47.43%, розроблений метод 96.0%); для прийому «Loaded Language» точність виявлення зросла на 42.9% (існуючий метод 54.17%, розроблений метод 97.07%); для прийому «Name Calling» точність виявлення зросла на 44.6% (існуючий метод 47.43%, розроблений метод 92.03%); для прийому «Repetition» точність виявлення зросла на 58.11% (існуючий метод 35.98%, розроблений метод 94.09%); для прийому «Appeal to Authority» точність виявлення зросла на 11.84% (існуючий метод 77.18%, розроблений метод 89.02%); для прийому «Black and White Fallacy» точність виявлення зросла на 36.68% (існуючий метод 54.55%, розроблений метод 91.23%); для прийому «Reductio ad hitlerum» точність виявлення зросла на 62.31% (існуючий метод 25.0%, розробле-

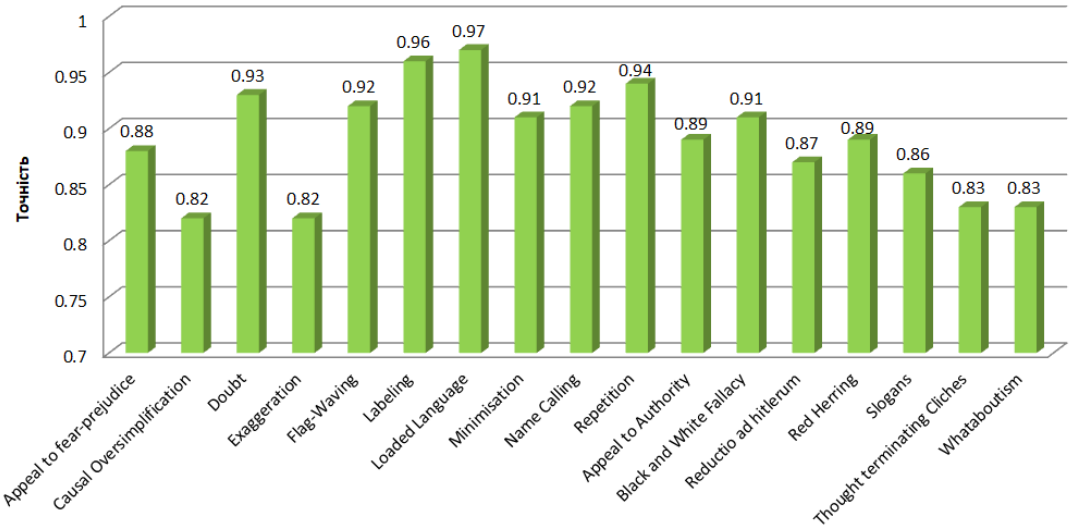

Рис. 6. Точність виявлення прийомів пропаганди BERT-моделями, навченими запропонованим методом

ний метод 87.31%); для прийому «Red Herring» точність виявлення зросла на 49.8% (існуючий метод 39.22%, розроблений метод 89.02%); для прийому «Slogans» точність виявлення зросла на 10.54% (існуючий метод 75.5%, розроблений метод 86.04%); для прийому «Thought terminating Cliches» точність виявлення зросла на 29.5% (існуючий метод 53.57%, розроблений метод 83.07%); для прийому «Whataboutism» точність виявлення зросла на 43.91% (існуючий метод 39.22%, розроблений метод 83.13%).

Отримані результати свідчать про спроможність запропонованого методу ефективно виявляти прийоми пропаганди, а також дозволяє візуально оцінити, які саме дані вплинули на рішення моделей щодо наявних прийомів пропаганди у користувацькому тексті.

Приклад поясненності з використанням моделі LIME щодо наявності у тексті прийому «Appeal to Fear-Prejudice» наведено на рис. 7.

Суть прийому «Appeal to Fear-Prejudice» полягає у створенні або підсиленні відчуття загрози та страху, з метою змусити людей прийняти певні ідеї або дії, які вважаються захисними або необхідними для уникнення небезпеки. Як видно з рис. 7, вагомими словами, що вплинули на рішення моделі, що тут присутній прийом «Appeal to Fear-Prejudice», є такі слова: «захопити», «знищити», «майбутнє», «варварів» тощо, що цілком відповідає визначенню даного прийому.

Отже, запропонований метод виявлення прийомів пропаганди за маркерами 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 дозволяє виявляти прийоми пропаганди з точністю понад 81.87 %, що зважаючи на здатність пропаганди маскуватись у контекстах повідомлень та новин є високим показником.

Подальші дослідження будуть спрямовані на розширення датасету, а також на розширення множини маркерів та навчання відповідних моделей машинного навчання.

Висновки. У статті розглянуто поточний стан наукового напряму виявлення прийомів пропаганди. З огляду на невирішені задачі предметної області запропоновано метод нейромережевого виявлення прийомів пропаганди за маркерами з візуальною інтерпретацією прийнятих рішень, який дозволяє шляхом використання набору з 17 навчених BERT-моделей виявляти 17 відповідних прийомів пропаганди. Метод відрізняється від існуючих тим, що враховує при навчанні нейромережевих класифікаторів додаткову множину маркерів та дозволяє здійснювати візуальну інтерпретацію отриманих результатів. Під додатковою множиною маркерів мається на увазі використання різноманітних текстових ознак, які притаманні визначеним прийомам пропаганди. Кроками методу є попередня обробка усіх навчальних і тестових текстових даних, навчання нейромережевих моделей для ідентифікації кожного маркера пропаганди, навчання нейромережевих моделей для кожного прийому пропаганди, створення моделі для поясненності та інтерпретації отриманих прогнозів для кожної моделі виявлення прийомів пропаганди, нейромережева оцінка сили прояву прийомів пропаганди у тестовому тексті та інтерпретація значень моделлю LIME. У методі використано 17 нейромережевих моделей для виявлення прийомів пропаганди: «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» та «Whataboutism».

Для дослідження ефективності методу виявлення прийомів пропаганди було створено програмну реалізацію у вигляді набору ноутбуків реалізованих у хмарному сервісі «Google Colab», що призначені для навчання нейромережевих моделей BERT із подальшим збереженням їх на жорсткому диску для використання у вебзастосунку для виявлення прийомів пропаганди, а також набору ноутбуків для збереження нейромережевих моделей для виявлення сили прояву маркерів пропаганди. Створений вебзастосунок дозв-

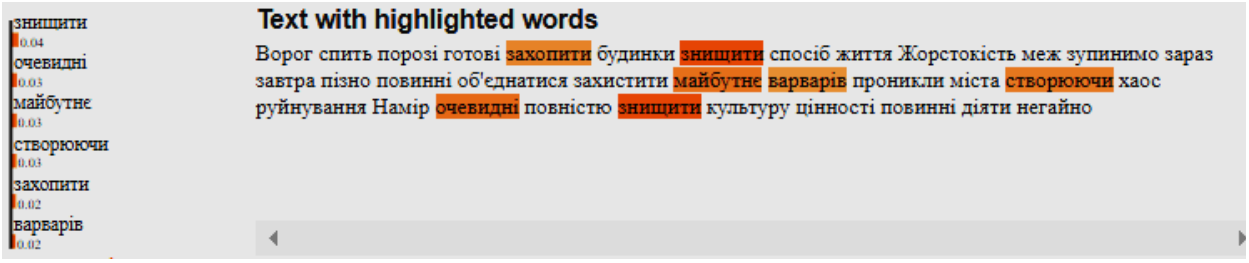

Рис. 7. Приклад пояснення результатів щодо наявності у тексті прийому «Appeal to Fear-Prejudice»

воляє не лише визначити інтенсивність проявів прийомів пропаганди, а і дає можливість здійснювати візуальну аналітику отриманих результатів. Для навчання BERT-моделей, що виконують функції виявлення прийомів пропаганди, використано набір даних «emnlp_trans_uk_dataset» від «Analysis Project». Набір даних є корпусом із 550 новинних статей, анотованих вручну на рівні фрагментів за допомогою вісімнадцяти пропагандистських прийомів. Дослідження ефективності встановило, що розроблений метод дозволяє шляхом використання набору з 17 навчених BERT-моделей виявляти 17 відповідних прийомів пропаганди з точністю не нижче 81.87%.

ЛІТЕРАТУРА

1. Horak A., Sabol R., Herman O., Baisa, V. Recognition of propaganda techniques in newspaper texts: Fusion of content and style analysis. *Expert Systems with Applications*. 2024, Vol. 251. DOI: https://doi.org/10.1016/j.eswa.2024.124085.
2. 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. DOI: https://doi.org/10.48550/arXiv.2402.03780.
3. 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.
4. 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. DOI: https://doi.org/10.18653/v1/D19-1565.
5. Martino G. D. S., Barron-Cedeno A., Wachsmuth H., Petrov R., Nakov P. SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles. *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. 2020, pp. 1377–1414.
6. 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».* 2024. С. 251–254.
7. 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.
8. Hugging Face, The AI community building the future, 2024. URL: https://huggingface.co/ (дата звернення: 06.11.2024).
9. Kaggle Competition – Disinformation Detection Challenge, 2023. URL: https://aihouse.org.ua/en/event/disinformation-detection-challenge/ (дата звернення: 06.11.2024).
10. Propaganda Analysis Project, 2023. URL: https://propaganda.math.unipd.it/index.html (дата звернення: 06.11.2024).
11. Zenodo. Propaganda. Proppy Corpus 1.0, 2019. URL: https://zenodo.org/records/3271522#.XS6qRUUzau4 (дата звернення: 06.11.2024).
12. 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*». 2024, С. 199–205.

REFERENCES

1. Horak A., Sabol R., Herman O., Baisa, V. (2024) Recognition of propaganda techniques in newspaper texts: Fusion of content and style analysis. *Expert Systems with Applications*. Vol. 251. DOI: https://doi.org/10.1016/j.eswa.2024.124085
2. Faye G., Icard B., Casanova M., Chanson J., Maine F., Bancilhon F., Gadek G., Gravier G., Egre P. (2024) Exposing propaganda: an analysis of stylistic cues comparing human annotations and machine classification. *Proceedings of the Third Workshop on Understanding Implicit and Underspecified Language*, pp. 62–72. DOI: https://doi.org/10.48550/arXiv.2402.03780
3. Vijayaraghavan P., Vosoughi, S. (2022) 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. pp. 3433–3448.

4. Martino G., Yu S., Barron-Cedeno A., Petrov R., Nakov, P. (2019) 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)*. pp. 5640–5650. DOI: https://doi.org/10.18653/v1/D19-1565
5. Martino G. D. S., Barron-Cedeno A., Wachsmuth H., Petrov R., Nakov P. (2020) SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles. *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. pp. 1377–1414.
6. Molchanova, M. (2024) Method vyiavlennia ta klasyfikatsii pryiomiv prohandy u tekstovomu kontenti zasobamy shtuchnoho intelektu. *Materialy XII Mizhnarodnoi naukovo-praktychnoi konferentsii "Informatsiini upravliaiuchi systemy ta tekhnolohii IUST-ODESA-2024*", S. 251–254. (in Ukrainian)
7. 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*. Vol. 3688, pp. 16–28.
8. Hugging Face, The AI community building the future, 2024. URL: https://huggingface.co/ (дата звернення: 06.11.2024).
9. Kaggle Competition – Disinformation Detection Challenge, 2023. URL: https://aihouse.org.ua/en/event/disinformation-detection-challenge/ (дата звернення: 06.11.2024).
10. Propaganda Analysis Project, 2023. URL: https://propaganda.math.unipd.it/index.html (дата звернення: 06.11.2024).
11. Zenodo. Propaganda. Proppy Corpus 1.0, 2019. URL: https://zenodo.org/records/3271522#.XS6qRUUzau4 (дата звернення: 06.11.2024).
12. Molchanova, M. O. (2024) Zastosuvannia auhmentatsii danykh dlia pidvyshchennia tochnosti vyiavlennia prohandy v internet-dzherelakh neiromerezhevymy modeliamy hlybokoho navchannia. *Materialy VIII Mizhnarodnoi naukovo-praktychnoi konferentsii "Perspektyvy suchasnoi nauky: teoriia i praktyka"*. S. 199–205. (in Ukrainian)