УДК 004.8

Молчанова М.О.*(Хмельницький національний університет, м. Хмельницький, Україна)*

ВИЯВЛЕННЯ ОБ'ЄКТІВ ПРОПАГАНДИ У ТЕКСТОВИХ ПОВІДОМЛЕННЯХ ЗАСОБАМИ ОБРОБКИ ПРИРОДНОЇ МОВИ ІЗ ВІЗУАЛЬНОЮ ІНТЕРПРЕТАЦІЄЮ РЕЗУЛЬТАТІВ

Анотація: *Запропоновано метод виявлення об'єктів пропаганди в текстових повідомленнях нейромережевими засобами обробки природної мови із візуальною інтерпретацією результатів. Відмінністю методу є розширення множини об'єктів пропаганди за рахунок додавання варіантів їх словесних подань та використання контекстних вікон для виявлення зв'язків між прийомами та об'єктами пропаганди. Це дозволяє покращити результати виявлення і забезпечити візуальне представлення об'єктів пропаганди, їх словесних подань і важливих зв'язків між ними. Експериментально доведено ефективність підходу, який забезпечує результати, що корелюють з експертними оцінками, і дозволяє візуально спостерігати об'єкти впливу та їх зв'язки в рамках пропагандистських прийомів.*

Ключові слова: *об'єкти пропаганди, прийоми пропаганди, виявлення пропаганди, обробка природної мови*

Abstract: *Proposes method for detecting propaganda objects in text messages using neural network tools for natural language processing with visual interpretation of the results. The difference of the method is the expansion of the set of propaganda objects by adding variants of their verbal representations and using context windows to detect connections between techniques and propaganda objects. This allows to improve the detection results and provide a visual representation of propaganda objects, their verbal representations and important connections between them. The effectiveness of the approach, which provides results that correlate with expert assessments and allows visually observing objects of influence and their connections within the framework of propaganda techniques, has been experimentally proven.*

Keywords: *propaganda objects, propaganda techniques, propaganda detection, natural language processing*

Постановка проблеми

Пропаганда, спрямована на маніпуляцію різними об'єктами для досягнення політичних, соціальних, економічних або культурних цілей, є одним із найбільших викликів сучасності [1]. Об'єктами пропаганди є особи, групи, організації, соціальні верстви, а також явища або інституції, на які спрямовані пропагандистські зусилля з метою впливу на їхню свідомість, емоції, поведінку та суспільну думку. У сучасних умовах важливим завданням є не лише автоматизоване виявлення пропагандистських прийомів, але й визначення об'єктів, на які спрямовані ці прийоми, з візуальною інтерпретацією результатів.

У даній статті представлено метод виявлення об'єктів пропаганди в текстових повідомленнях засобами обробки природної мови. Відмінністю цього методу є розширення множини об'єктів пропаганди завдяки додаванню варіантів їх словесних подань та використанню контекстних вікон для виявлення зв'язків між прийомами та об'єктами пропаганди. Це дозволяє не тільки покращити результати виявлення, але й забезпечити візуальне представлення об'єктів пропаганди, їх словесних подань та важливих зв'язків між ними. Експериментально доведено ефективність підходу, який забезпечує результати, що корелюють з експертними оцінками, і дозволяє візуально спостерігати об'єкти впливу та їх зв'язки в рамках пропагандистських прийомів.

Запропонований підхід корелює із Цілями сталого розвитку ПРООН та сприяє автоматизації процесу виявлення та класифікації пропаганди, забезпечуючи повні, інтерпретовані та зрозумілі результати. Зокрема, застосування методів обробки природної мови для виявлення та класифікації технік і об'єктів пропаганди сприяє досягненню Цілі сталого розвитку ООН №16 шляхом підвищення прозорості інформаційного простору та зміцнення інституційної довіри. Також це підтримує Ціль сталого розвитку ООН №4, розвиваючи медіаграмотність і критичне мислення серед населення, що допомагає ефективно протидіяти дезінформації.

Аналіз останніх публікацій

Є два основних підходи до ідентифікації пропаганди: через розпізнавання іменованих сутностей (NER) та класифікацію повідомлень [2]. Розглядаючи пропаганду як задачу NER, виникає складність через те, що текстові фрагменти з пропагандистськими елементами

зазвичай довші, ніж типові об’єкти NER (наприклад, імена чи назви), і можуть складатися з кількох десятків слів. У дослідженні [3] аналізується вплив довжини текстових сегментів на точність виявлення пропаганди, що підтверджує зростання складності із збільшенням довжини діапазонів. Було випробувано кілька популярних методів для цієї задачі, виміряно, наскільки добре вони відображають розподіл довжини текстових фрагментів, а також запропоновано підхід із адаптивним рівнем згортки, який покращує обмін інформацією між віддаленими словами. Це рішення сприяє більш точному відновленню довжини тексту без втрати загальної ефективності.

У рамках досліджень, орієнтованих на виявлення пропаганди на рівні документів, акцент зроблено на оцінці тексту як цілісного елемента і його окремих речень [4]. Для побудови ознак використовуються різні методи: статистичні індикатори, векторизація тексту [5], лінгвістичне маркування, а також розпізнавання тригерів, таких як абсолютні займенники або підсилювальні слова.

Експериментальні результати продемонстрували, що модель, застосована до аналізу на рівні документа, досягла точності 0,943. Вона змогла правильно класифікувати 6097 непропагандистських статей і 694 пропагандистські статті. Підхід, орієнтований на аналіз окремих речень, показав нижчі результати: точність склала 0,744. Він успішно ідентифікував 205 пропагандистських речень і 1917 непропагандистських, проте 731 статтю було класифіковано невірно.

Аналіз пов’язаних робіт у сфері виявлення прийомів та об'єктів пропаганди виявив низку проблем. По-перше, існує відсутність комплексного аналізу взаємозв’язків між техніками та об'єктами пропаганди в текстах. По-друге, бракує узагальнень для об'єктів пропаганди та їх альтернативних згадувань. Пропаганда, яка виявляється тільки через пошук іменованих сутностей, не демонструє спрямованості технік. Також, техніки пропаганди, що виявляються на рівні документу, не відображають об'єктів впливу. При виявленні пропаганди, як завдання пошуку іменованих сутностей, об'єкти часто подаються власними назвами, що охоплює питання «на кого?», однак не охоплює питання «На що?» мають спрямування використані прийоми

Мета роботи та постановка завдань

Мета роботи полягає в створенні методу виявлення об’єктів пропаганди засобами обробки природної мови з візуальною інтерпретацією результатів, який дозволить у пропагандистських повідомленнях знаходити на кого, і на що, спрямовані конкретні використані в повідомленні прийоми пропаганди, а також бачити візуальну інтерпретацію результату.

Виклад основного матеріалу

В рамках підходу до виявлення об’єктів пропаганди у текстових повідомленнях засобами обробки природної мови із візуальною інтерпретацією результатів буде використано множину нейромережевих моделей для ідентифікації пропагандистських прийомів для подальшого співвіднесення їх з знайденими об’єктами. Кожна з 17 моделей була попередньо навчена для виявлення кожного з 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», «Whataboutism» [6] відповідно.

Метод виявлення об’єктів пропаганди у текстових повідомленнях із застосуванням засобів обробки природної мови та візуальної інтерпретації результатів базується на нейромережевих моделях глибокого навчання та складається з кількох етапів. Спершу здійснюється ідентифікація об’єктів пропаганди шляхом розпізнавання іменованих сутностей (NER). На цьому етапі проводиться попередня обробка тексту, що включає видалення повторів серед іменованих сутностей на рівні їх лем.

Наступним кроком є розширення множини об’єктів пропаганди за рахунок визначення альтернативних варіантів словесного подання іменованих сутностей. Після цього формуються контекстні вікна для кожного об’єкта пропаганди, з урахуванням заданого порогового

значення мінімального розміру вікна. В межах цих контекстних вікон оцінюється інтенсивність використання прийомів пропаганди за допомогою нейромережевих моделей.

На фінальному етапі будується множина важливих зв'язків між об'єктами та прийомами пропаганди, враховуючи порогові значення мінімального рівня прояву пропагандистських прийомів. Отримані результати дозволяють не лише виявити об'єкти та спрямованість пропаганди, але й забезпечити візуалізацію зв'язків між прийомами та їх цільовими об'єктами.

Схематичне представлення запропонованого методу наведено на рисунку 1. Вхідними даними для реалізації методу є текст для аналізу, множина ідентифікованих прийомів пропаганди у тексті та набір попередньо навчених нейромережевих моделей, адаптованих для аналізу кожного прийому. Результатом першого етапу є множина об'єктів пропаганди, визначених за допомогою NER, без повторів.

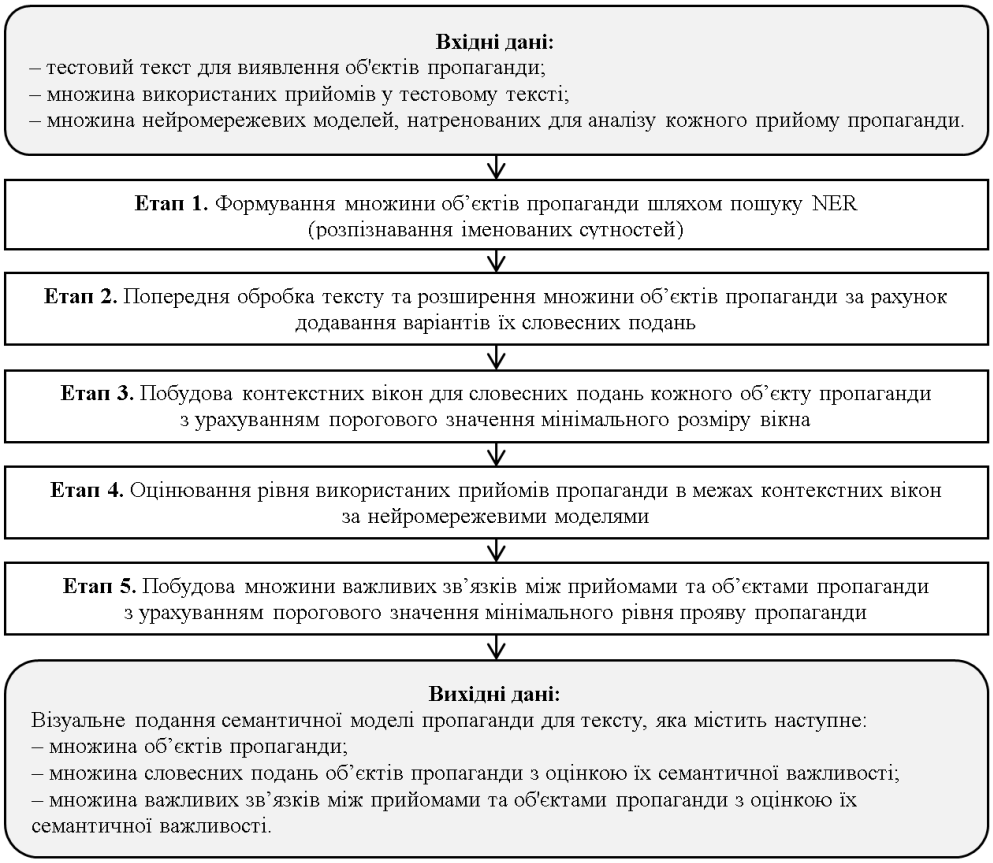

Рисунок 1. Схема методу виявлення об'єктів пропаганди засобами обробки природної мови з візуальною інтерпретацією результатів

На другому етапі методики до кожної ідентифікованої іменованої сутності здійснюється пошук схожих за значенням слів-об'єктів. Це обумовлено тим, що поняття об'єктів пропаганди є ширшим за NER і включає не лише іменовані сутності, але й культурні аспекти, соціальні групи чи узагальнені категорії, об'єднані за спільними характеристиками. Для цього використовується попередньо навчена модель FastText, розроблена Facebook AI Research, яка базується на архітектурах «CBOW» і «Skip-gram». Ця модель дозволяє аналізувати контекст слів, встановлювати семантичні зв'язки та виявляти схожі об'єкти, що дає змогу розширити спектр виявлених об'єктів пропаганди.

У рамках роботи FastText донавчається на текстах, які містять пропаганду, для забезпечення специфічності до задачі. В результаті цього етапу формується розширена множина об'єктів пропаганди, яка включає альтернативні варіанти їх словесних подань.

Мінімальний рівень семантичної близькості визначається емпірично залежно від специфіки завдання; у даному випадку порогове значення не використовувалося.

На третьому етапі формуються контекстні вікна для кожного об’єкта пропаганди. Контекстним вікном вважається речення, у якому згадується конкретний об’єкт. Якщо одне речення містить кілька об’єктів пропаганди, контекстне вікно створюється один раз і включає всі об’єкти. У разі, якщо об’єкт пропаганди має кілька словесних подань, контекстні вікна дублюються для кожного з них, зберігаючи зв'язок із вихідним об’єктом. Мінімальний розмір контекстного вікна визначається пороговим значенням, яке встановлюється відповідно до вимог аналізу.

На четвертому етапі методу здійснюється аналіз контекстних вікон для визначення рівня використання пропагандистських прийомів. Це реалізується через векторизацію текстового контенту контекстних вікон із застосуванням відповідних векторизаторів, після чого нейромережеві моделі аналізують приналежність кожного контекстного вікна до конкретних прийомів пропаганди. Оцінка здійснюється для всіх виявлених у тексті прийомів, що дозволяє визначити, які саме з них були задіяні у межах кожного контексту.

На фінальному, п’ятому етапі, будується множина важливих зв’язків між пропагандистськими прийомами та об’єктами. Це виконується з урахуванням порогового значення мінімального рівня прояву пропаганди. У випадках, коли сила прояву прийому в межах контекстного вікна не перевищує встановлений поріг, такий прийом не вважається застосованим до відповідної групи об’єктів.

Такий підхід забезпечує наочне подання результатів, сприяючи ефективному аналізу повідомлень і розумінню взаємозв’язків між об’єктами та прийомами пропаганди.

Для оцінки ефективності розробленого методу виявлення об’єктів пропаганди було створено спеціалізоване програмне забезпечення, яке дозволяє ідентифікувати об’єкти пропаганди, зіставляти їх із використаними прийомами та відображати результати у формі візуальної аналітики. Отримані дані порівнювалися з висновками авторитетних ресурсів і експертів у сфері протидії пропаганді, що дало змогу оцінити якість запропонованого підходу.

Для тестування використовувалися розмічені повідомлення із соціальних мереж, підготовлені Центром стратегічних комунікацій [7], які містили експертні висновки. Це забезпечило можливість порівняння результатів роботи методу з незалежними оцінками фахівців. Програмне забезпечення для реалізації методу було розроблено у вигляді вебзастосунку на мові програмування Python. В межах створеного програмного забезпечення застосовано: 17 попередньо навчених нейромережевих моделей, створених на основі попередніх досліджень; нейромережеву бібліотеку Stanza для розпізнавання іменованих сутностей (NER); фреймворк Flask для організації вебінтерфейсу; модель FastText, адаптовану до специфіки аналізу пропагандистських повідомлень шляхом донавчання. Приклад повідомлення та його аналізу авторитетним джерелом наведено на рисунку 2.

У ході дослідження ефективності запропонованого методу виявлення об’єктів і прийомів пропаганди було встановлено, що результати, отримані за допомогою розробленого підходу, демонструють високу кореляцію з експертними оцінками, представленими Центром стратегічних комунікацій [18].

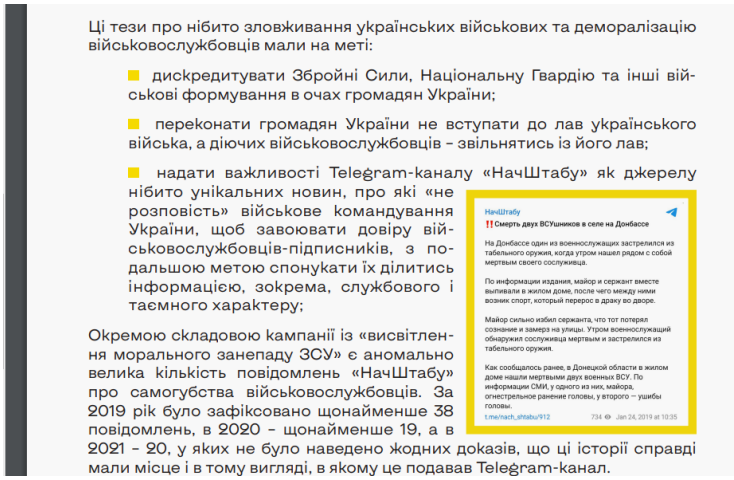

Рисунок 2. Аналіз повідомлення що містить пропаганду від авторитетного джерела [18]

Для підтвердження наведено приклад аналізу допису з пропагандистського каналу (рисунок 2). У цьому прикладі результати автоматичного аналізу, виконаного за допомогою розробленого методу, співпали з висновками експертів. Це свідчить про здатність методу точно ідентифікувати об’єкти пропаганди та визначати, які прийоми було використано для їх маніпулятивного впливу.

Отримані результати підтверджують практичну застосовність розробленого підходу для автоматизованого аналізу пропагандистських повідомлень і можливість його використання в реальних умовах для підтримки роботи аналітиків та дослідників. Результати наведені на рисунку 3.

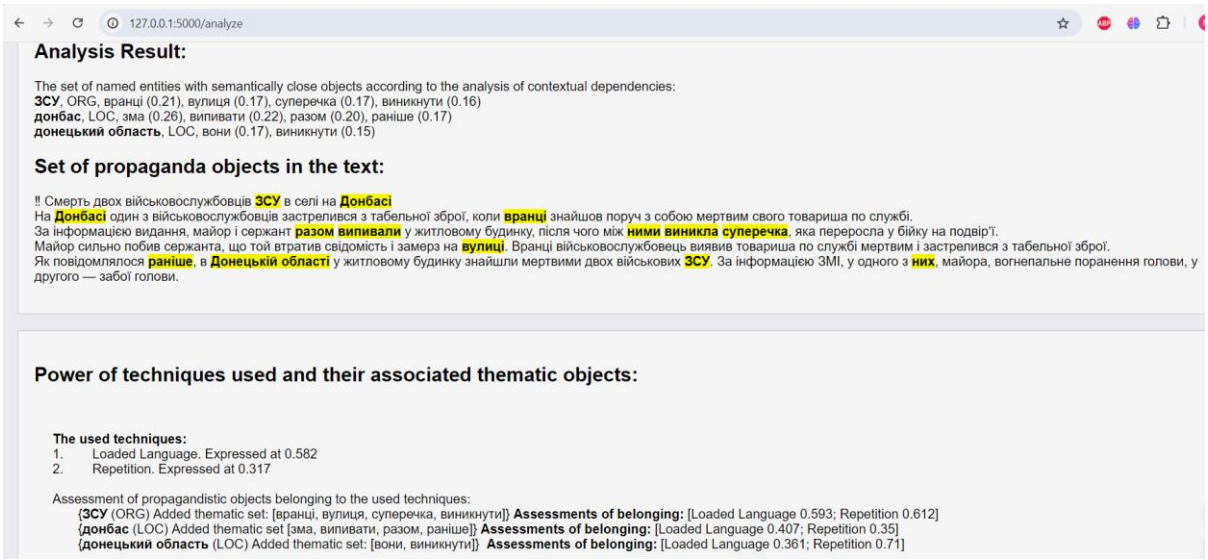

Рисунок 3. Візуальна інтерпретацією прийнятих рішень за методом виявлення об’єктів пропаганди

Аналіз за допомогою розробленого програмного забезпечення показав використання пропагандистських прийомів «Loaded Language» (0.582) і «Repetition» (0.317), ідентифікував об’єкти пропаганди (наприклад, ЗСУ, Донбас, Донецька область) разом із семантично близькими словами та оцінив відповідність об’єктів пропаганди до прийомів: ЗСУ («Loaded Language» – 0.593, «Repetition» – 0.612), Донбас («Loaded Language» – 0.407, «Repetition» – 0.35), Донецька область («Loaded Language» – 0.361, «Repetition» – 0.71), з подальшим візуальним відображенням знайдених об’єктів у тексті.

У результаті дослідження запропонованого методу виявлено, що він дозволяє отримувати результати, які корелюють із результатами з авторитетних маркованих джерел. За допомогою застосування комплексного підходу до виявлення пропаганди, та з використанням візуальної інтерпретації результатів, вирішується задача взаємозв'язків між використаними прийомами та об'єктами пропаганди.

ВИСНОВКИ.

Розроблено метод виявлення об'єктів пропаганди засобами обробки природної мови з візуальною інтерпретацією прийнятих рішень, що відрізняється від існуючих розширенням множини об'єктів пропаганди завдяки додаванню варіантів їх словесних подань і використанню контекстних вікон для виявлення взаємозв'язків між використаними прийомами та об'єктами пропаганди. Це покращує результати виявлення та їх візуальне представлення. Метод включає розпізнавання іменованих сутностей, попередню обробку тексту, розширення множини об'єктів пропаганди, побудову контекстних вікон, оцінювання рівня використаних прийомів та побудову важливих зв'язків між прийомами та об'єктами пропаганди.

Для підвищення точності та якості виявлення прийомів та об'єктів пропаганди за семантичними маркерами у повідомленнях засобами обробки природної мови з подальшою інтерпретацією результатів, було розроблено підхід, який дозволяє ідентифікувати об'єкти пропаганди у текстах, а також на кого і на що спрямовані пропагандистські прийоми. Метод вирішує проблеми відсутності комплексного аналізу взаємозв'язків прийомів та об'єктів пропаганди в повідомленнях і відсутності узагальнень для об'єктів пропаганди та їх альтернативних згадок. Експериментально доведено ефективність підходу, що дозволяє, окрім пошуку NER за допомогою бібліотеки нейронної мережі «STANZA», розширювати перелік об'єктів пропаганди за допомогою бібліотеки машинного навчання «FastText», а також оцінювати їх зв'язок з використаними прийомами. Результати методу корелюють з експертними оцінками, а візуальна аналітика забезпечує наочне спостереження об'єктів впливу в рамках пропагандистських прийомів.

Список посилань:

1. *M.B. Shevtsiv, K. A. Honcharuk, Propaganda as socio-legal phenomenon: problems of understanding, in: Current problems of historical and legal and of international legal science. South Ukrainian magazine 1, 2019, pp. 119 – 122*
2. *Молчанова М.О. Нейромережеве виявлення і класифікація прийомів та об'єктів пропаганди у текстовому контенті. Міжнародний науково-технічний журнал «Вимірювальна та обчислювальна техніка в технологічних процесах», № 4, 2024. с. 153-161. https://doi.org/10.31891/2219-9365-2024-80-19 URL: https://vottp.khmnu.edu.ua/index.php/vottp/article/view/385/361*
3. *Przybyla P. Long Named Entity Recognition for Propaganda Detection and Beyond / P. Przybyla, K. Kaczynski // Proceedings of the International Conference of the Spanish Society for Natural Language Processing. – 2023*
4. *Vysotska V. Information technology for recognizing propaganda, fakes and disinformation in textual content based on nlp and machine learning methods. Radio Electronics, Computer Science, Control. 2024. No. 2. P. 126. URL: https://doi.org/10.15588/1607-3274-2024-2-13 (date of access: 29.11.2024).*
5. *From Bag-of-Words to Transformers: A Comparative Study for Text Classification in Healthcare Discussions in Social Media / E. De Santis et al. IEEE Transactions on Emerging Topics in Computational Intelligence. 2024. P. 1–15. URL: https://doi.org/10.1109/tetci.2024.3423444 (date of access: 29.11.2024).*
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. *Tsentr stratehichnykh komunikatsii [Elektronnyi resurs]. – Rezhym dostupu: https://spravdi.gov.ua/*