УДК 004.91:004.62

Oleksandr Terentiev, Doctor of Technical Sciences, Associate Professor, Principal researcher
ORCID ID: https://orcid.org/0000-0002-4288-1753 ***e-mail:*** o.terentiev@gmail.com

Tetyana Prosyankina-Zharova, Doctor of Technical Sciences, Associate Professor, senior researcher
ORCID ID: https://orcid.org/0000-0002-9623-8771 ***e-mail:*** t.pruman@gmail.com

Yurii Abroskin, graduate student
ORCID ID: https://orcid.org/0009-0009-9828-5596 ***e-mail:*** abroskin21@gmail.com

Volodymyr Duda, graduate student
ORCID ID: https://orcid.org/0009-0002-4278-4635 ***e-mail:*** dudavolodimir@gmail.com

Institute of Telecommunications and Global Information Space of the National Academy of Sciences of Ukraine, Kyiv, Ukraine

ANALYSIS METHODOLOGY OF PRO-KREMLIN DESINFORMATION IN INTERNET NEWS ARTICLES

Abstract. *The article proposes and implements a methodology for analyzing pro-Kremlin disinformation in Internet sources based on the integration of automated data collection, natural language processing methods, topic modeling, and statistical analysis. The study utilized an open multilingual dataset containing 18,249 links to web articles in 42 languages, developed within the framework of the European anti-disinformation initiatives VERA.AI and EUvsDisinfo. The proposed methodology includes the stages of automated extraction of texts from web resources, text preprocessing, language filtering, thematic clustering, and the development of classification models using the SAS Text Miner system. For automated collection of textual content, a specialized Python-based software application was developed using the PLAYWRIGHT and ASYNCIO libraries, optimized for high-performance processing of large-scale web article corpora.*

The results of the study revealed a significant relationship between the type of content, the language of the source, and the necessity of VPN access for retrieving texts. The Pearson chi-square statistic was 8847 with 10 degrees of freedom and a p-value < 0.000001, indicating a high statistical significance of the obtained results. It was found that Russian-language disinformation resources in most cases require the use of VPN access due to sanctions and geographical access restrictions, whereas trustworthy English-language and Ukrainian-language sources demonstrate substantially higher openness and accessibility stability. Thematic analysis showed that pro-Kremlin disinformation is concentrated around anti-Ukrainian, anti-NATO, and conspiracy-oriented narratives, demonstrating high thematic repetition and characteristics of coordinated FIMI campaigns. The proposed methodology can be applied in the fields of information security, OSINT analytics, information space monitoring, and the development of automated disinformation detection systems.

Keywords: *FIMI, Foreign Information Manipulation and Interference, OSINT, Open Source Intelligence, fake news detection, SAS, chi-square statistic.*

© О.М. Терентьєв, Т.І. Просянкіна-Жарова, Ю.Ю. Аброскін, В.О. Дуда, 2026

О.М. Терентьєв, Т.І. Просянкіна-Жарова, Ю.Ю. Аброскін, В.О. Дуда

Інститут телекомунікацій і глобального інформаційного простору НАН України, м. Київ, Україна

МЕТОДИКА АНАЛІЗУ ПРОКРЕМЛІВСЬКОЇ ДЕЗІНФОРМАЦІЇ В ДЖЕРЕЛАХ МЕРЕЖІ ІНТЕРНЕТ

Анотація. *У статті запропоновано та реалізовано методику аналізу прокремлівської дезінформації в джерелах мережі Інтернет, що базується на поєднанні автоматизованого збору даних, методів обробки природної мови, тематичного моделювання та статистичного аналізу. Для проведення дослідження використано відкритий багатомовний набір даних, який містить 18 249 посилань на веб-статті 42 мовами та сформований у межах європейських ініціатив протидії дезінформації VERA.AI та EUvsDisinfo. Запропонована методика включає етапи автоматизованого вивантаження текстів із веб-ресурсів, попередньої обробки текстових даних, мовної фільтрації, тематичної кластеризації та побудови моделей класифікації із використанням системи SAS Text Miner. Для автоматизованого збору текстового контенту було розроблено спеціалізовану програму мовою Python із використанням бібліотек PLAYWRIGHT та ASYNCIO, оптимізовану для високопродуктивної обробки великих масивів веб-статей.*

У результаті дослідження встановлено суттєву залежність між типом контенту, мовою джерела та необхідністю використання VPN-доступу для отримання текстів. Значення статистики хі-квадрат Пірсона становило 8847 при 10 ступенях свободи та p-value < 0,000001, що свідчить про високу статистичну значущість отриманих результатів. Виявлено, що російськомовні дезінформаційні ресурси у більшості випадків потребують використання VPN через санкційні та географічні обмеження доступу, тоді як правдиві англомовні та україномовні джерела характеризуються значно вищою відкритістю та стабільністю доступу. Тематичний аналіз показав, що прокремлівська дезінформація концентрується навколо антиукраїнських, антинатовських та конспірологічних наративів, демонструючи високу повторюваність та ознаки координованих FIMI-кампаній. Запропонована методика може бути використана у задачах інформаційної безпеки, OSINT-аналітики, моніторингу інформаційного простору та побудови систем автоматичного виявлення дезінформації.

Ключові слова: *FIMI, Іноземне інформаційне маніпулювання та втручання, OSINT, аналіз відкритих джерел інформації, виявлення дезінформацій, SAS, статистика хі-квадрат Пірсона.*

https://doi.org/10.32347/2411-4049.2026.2.154-160

Вступ

Сучасний розвиток цифрових технологій та глобалізація інформаційного простору призвели до появи нових актуальних загроз масового поширення дезінформації, маніпулятивного контенту та координованих інформаційних операцій. Все це становить особливу небезпеку та загрозу на тлі військової агресії, в рамках якої прокремлівська дезінформація використовується як інструмент інформаційно-психологічного впливу, який призводить до формування небезпечних суспільних настроїв, дискредитує органи державної влади та призводить до послаблення міжнародної підтримки України.

У зв'язку з цим виникає необхідність створення ефективних підходів для аналізу дезінформаційних матеріалів, здатних працювати з великими масивами неструктурованих текстових даних. В даній роботі запропоновано методику аналізу прокремлівської дезінформації в джерелах мережі Інтернет, що включає автоматизоване вивантаження текстових даних, їх попередню обробку та тематичний аналіз із використанням системи SAS Text Miner. Методика базується на використанні відкритого багатомовного набору даних, сформованого в межах європейських дослідницьких ініціатив у сфері протидії дезінформації, та орієнтована на виявлення інформаційних наративів, оцінку доступності джерел, аналіз мовних особливостей контенту й статистичне підтвердження виявлених закономірностей. Отримані результати можуть бути використані у системах моніторингу інформаційного простору, OSINT-дослідженнях, задачах інформаційної безпеки та побудові моделей автоматичного виявлення дезінформації.

Опис методики аналізу прокремлівської дезінформації в джерелах мережі Інтернет

В рамках дослідження було запропоновано та реалізовано методику аналізу прокремлівської дезінформації в джерелах мережі Інтернет, що складається з наступних етапів.

1. Формування списку джерел, що підлягають дослідженню, або підписка на спеціалізовані агрегатори новин в мережі Інтернет [1, 2].

2. Вивантаження даних з мережі Інтернет, в автоматизованому режимі з метою отримання текстового вмісту веб-статей за URL-посиланнями з набору даних та збереження результатів у структурованому форматі [3, 4, 5].

3. Попередня обробка текстів із використанням системи SAS Text Miner [6, 7], після вивантаження, що включає наступні кроки.

3.1. Очистка текстів від HTML-тегів, рекламних блоків, службових символів, дублікатів, навігаційних елементів, стоп-слів [6].

3.2. Нормалізація тексту, що включає токенізацію, лематизацію, стемінг, уніфікація кодування [6].

3.3. Мовна фільтрація, після якої із загального корпусу залишаються англійська, українська та російська мови. Це дозволяє зменшити обчислювальні витрати, звузити предмет дослідження та підвищити якість кластеризації [6, 7].

4. Аналіз текстів із використанням спеціалізованих модулів та компонентів системи SAS Text Miner [6]. Для виявлення інформаційних наративів використовуються інструменти тематичного моделювання, кластеризації та аналіз латентних семантичних структур [6, 7].

5. Статистичний аналіз результатів. На цьому етапі проводиться оцінка частоти появи наративів, аналіз мовних відмінностей, аналіз доступності джерел, оцінка залежності між мовою та типом блокування. Для статистичного підтвердження гіпотез можуть використовуватися критерій хі-квадрат Пірсона, кореляційний аналіз, аналіз частот та дисперсійний аналіз.

Опис вхідного набору даних для дослідження

Однією із проблем при проведенні досліджень є наявність вже розмічених даних, тому для аналізу та побудови моделей було взято готовий розмічений набір даних [1], що містить 18 249 посилань на статті в мережі Інтернет, при

цьому це багатомовний файл, що включає 42 різні мови. Цей набір даних [1] можна отримати у відкритому доступі, за посиланням в мережі Інтернет: https://zenodo.org/records/10514307.

Зауважимо, що ці дані збиралися протягом майже 9 років, починаючи з січня 2015 по серпень 2023 року, колективом авторів, в рамках спеціальної європейської ініціативи VERA.AI за підтримки проєкту Horizon Europe.

Цей ресурс [1] призначений для досліджень проблем дезінформації, задач обробки природньої мови та машинного навчання, побудови моделей виявлення дезінформацій, аналізу інформаційних операцій, вивчення прокремлівських наративів на 42 мовах.

Окрім цього, це дослідження [1] тісно пов'язано з іншим дослідницьким проєктом EUvsDisinfo [2], що є флагманським проєктом Європейської служби зовнішніх дій (EEAS – European External Action Service), створений для виявлення, аналізу та протидії дезінформаційним кампаніям, насамперед прокремлівським інформаційним операціям. Зазначений проєкт реалізується командою "Східна оперативна група зі стратегічних комунікацій ЄС" (East StratCom Task Force), яка в свою чергу фактично є одним із перших наддержавних центрів Європи, що системно займається аналізом дезінформації, інформаційною безпекою, цифровими інформаційними операціями, FIMI-аналітикою (Foreign Information Manipulation and Interference) та OSINT-моніторингом.

Програмна реалізація деяких етапів методики аналізу прокремлівської дезінформації в джерелах мережі Інтернет

Спеціально для вирішення задачі по вивантаженню даних з усіх Інтернет-джерел, що наведені у наборі даних [1], колективом авторів було розроблено спеціалізовану програму на мові програмування Python, з якою можна ознайомитися на GitHub [3]. Програма реалізує асинхронний конвеєр обробки тексту для масового автоматизованого збору даних з веб-сайтів (HTML-документів) із використанням бібліотек PLAYWRIGHT [4] та ASYNCIO [5]. Архітектура системи оптимізована для високопродуктивного завантаження великих корпусів веб-статей і придатна для задач обробки природних мов, аналізу відкритих джерел інформації (OSINT, Open Source Intelligence), текстової аналітики та аналізу дезінформаційних кампаній.

Після попереднього розвідувального аналізу текстів було прийнято рішення з 42 мов зазначеного набору даних [1] залишити лише три – англійську, російську та українську, для того щоб звузити дослідницьку задачу та зменшити час на аналіз та обчислення.

На рисунку наведена стандартна схема технологічного процесу, що містить етапи та модулі:

1. Завантаження даних, як з мережі Інтернет по вказаних джерелах, так і з локального диску комп'ютера.

2. Парсинг тексту, перетворення неструктурованого тексту у дані, придатні для обробки програмою.

3. Фільтрація тексту, попередня обробка, під час якої видаляються непотрібні слова, очищуються текстові дані, зменшується шум, залишаються лише інформативні терміни для подальшого аналізу.

4. Кластеризація тексту, автоматичне групування у кластери на основі їх тематичної або статистичної подібності.

5. Модуль побудови текстових правил, призначений для автоматичного створення правил класифікації текстів, на основі виявлення закономірностей у документах, з послідовною побудовою моделей прогнозування на основі тексту.

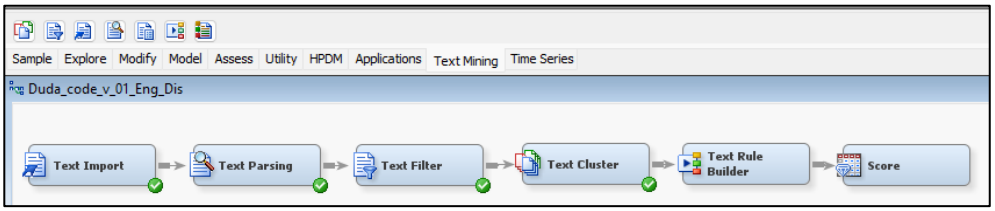

Рисунок 1. Стандартна схема технологічного процесу для аналізу текстової інформації в системі SAS Text Miner

Основною задачею наведеної на рис. 1 технологічної схеми процесу є виявлення повторюваних пропагандистських тез, координації наративів, тематичних груп дезінформації. Для чого модуль кластеризації дозволяє автоматично групувати документи, виявляти схожі інформаційні кампанії, визначати центри тематичних кластерів. Типові кластери можуть включати антиукраїнські наративи, антинатівську риторику, маніпуляції щодо санкцій, дискредитацію західних медіа, дезінформацію щодо військових подій.

Побудова моделей класифікації, із використанням модуля побудови текстових правил, дозволяє будувати правила класифікації, виявляти ключові терміни дезінформації та оцінювати інформативність ознак, для чого можуть використовуватися дерева рішень, регресійні моделі, TF-IDF метод та ентропійного зважування [6, 7].

Аналіз отриманих результатів дослідження

При вивантаженні текстів за посиланнями в мережі Інтернет було з'ясовано, що близько 43% інтернет-джерел вивантажується без будь-яких проблем, 42,29% можна вивантажити, лише використовуючи VPN-з'єднання, а 14,67% джерел взагалі було видалено. Зазначимо, що VPN (Virtual Private Network – віртуальна приватна мережа) – це технологія, яка створює зашифрований тунель між пристроєм і віддаленим сервером.

Найбільша кількість заблокованого та видаленого контенту припадає саме на той, що класифіковано як дезінформація на англійській та російській мовах. В таблиці 1 наведена більш детальна інформація з розбивкою по мовах та категоріях (правдива інформація чи дезінформація).

Значення статистики хі-квадрат Пірсона, для результатів наведених в таблиці 1, дорівнює 8847, при кількості ступенів свободи 10, та значенням p-value меншим за 0,000001. Значення хі-квадрат є надзвичайно великим і статистично значущим, що в сукупності означає, що існує дуже сильна залежність між типом контенту, мовою контенту та необхідністю використання VPN. Найбільший внесок у значення статистики хі-квадрат дають: російська дезінформація, по причині надзвичайно високого відсотку

необхідності VPN-доступу (86,7%); правдивість інформації із англомовних джерел, через дуже високу відкриту доступність без використання VPN (76%); правдивість інформації із україномовних джерел через майже повну відкритість ресурсів (94%).

Таблиця 1. Кількість статей для різних мов, що було вивантажено з мережі Інтернет без використання та з використанням VPN, а також процент видалених статей

Мова джерела	Категорія	Кількість новин у наборі даних	Процент статей, що вивантажені без використання VPN	Процент статей, що вивантажені із використанням VPN	Процент видалених статей
Англійська	правдива інформація	6121	76,44	4,313	19,25
Англійська	дезінформація	425	25,65	62,82	11,53
Російська	правдива інформація	469	53,31	40,94	5,76
Російська	дезінформація	5356	2,427	86,71	10,87
Українська	правдива інформація	315	93,65	0,634	5,71
Українська	дезінформація	56	56	0	0

Більшість правдивих англомовних статей доступні напряму (76%), мають високу стабільність доступу, що свідчить про використання провідних перевірених засобів масової інформації та відсутності геоблокувань.

Окремо необхідно відмітити той факт, що 19% статей, що відносяться до правдивих, протягом останніх років було видалено або недоступні. Для великих англомовних ресурсів це природний процес, пов'язаний із появою платного доступу до архівних матеріалів, видаленням дублюючих посилань, зміною URL внаслідок міграції ресурсів та реструктуризації, зміною у налаштуванні правил безпеки (захист від DDoS-атак та ботів), а також видаленням застарілих новин. Між тим англомовні ресурси, що позначені як дезінформаційні, потребують використання VPN у 62%, по причині геоблокування регіонального обмеження та обмеження доступу за IP.

Правдиві російськомовні джерела доступні лише в 53%, а в 40% потребують вже використання VPN, що пов’язано із санкційними та регіональними обмеженнями та частковим блокуванням незалежних медіа. В той час як дезінформаційні російськомовні джерела доступні лише в 2,4%, а для 86,7% необхідне використання VPN, що пов’язано в першу чергу із санкційним блокуванням на рівні держави доменів та розгортанням інфраструктури обмеження доступу в ЄС та Україні.

Аналіз текстів за тематиками та наративами показали наступні основні результати. Російськомовна дезінформація концентрується навколо геополітичних конфліктів, активно використовує антинатівські, антиукраїнські та конспірологічні наративи, демонструє ознаки

скоординованих FIMI-кампаній, має значно вищий рівень тематичної повторюваності та емоційної поляризації. В той час як правдива інформація характеризується більшою тематичною різноманітністю, орієнтована на політичну та суспільну аналітику, містить менше конфліктно-пропагандистських конструкцій.

СПИСОК ЛІТЕРАТУРИ \ REFERENCES

1. Leite, J., Razuvayevskaya, O., Bontcheva, K., & Scarton, C. (2024). *EUvsDisinfo: A dataset for multilingual detection of pro-Kremlin disinformation in news articles* [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.10492913
2. *EUvsDisinfo: Official website.* (n.d.). https://euvsdisinfo.eu/
3. Duda, V., & Terentiev, O. (n.d.). *Program for download EUvsDisinfo* [GitHub repository]. GitHub. https://github.com/oterentiev/download_EUvsDisinfo
4. *Playwright: Browser automation library website.* (n.d.). https://playwright.dev/
5. Python Software Foundation. (n.d.). *Asyncio – Asynchronous I/O: Python 3 documentation.* https://docs.python.org/3/library/asyncio.html
6. Jade, T., Belamaric-Wilsey, B., & Wallis, M. (2019). *SAS text analytics for business applications* (1st ed.): Concept rules for information extraction models. SAS Press. https://sasinstitute.redshelf.com/book/1878372
7. Terentiev, O. M., Duda, V. O., Abroskin, Yu. Yu., & Prosyankina-Zharova, T. I. (2026). Analysis of text analytics methods for knowledge extraction from Ukrainian-language social media. *Environmental Safety and Natural Resources, 1*(57), 161–170. https://doi.org/10.32347/2411-4049.2026.1.161-170 [in Ukrainian]

Стаття надійшла до редакції 30.01.2026, надійшла після рецензування 23.03.2026, прийнята 06.04.2026

The article was received 30.01.2026, received after revision 23.03.2026, accepted 06.04.2026

Терентьєв Олександр Миколайович
доктор технічних наук, доцент, провідний науковий співробітник, Інститут телекомунікацій і глобального інформаційного простору НАН України
Адреса робоча: бульв. Чоколівський, 13, Київ, Україна, 03186
ORCID ID: https://orcid.org/0000-0002-4288-1753 ***e-mail:*** o.terentiev@gmail.com

Просянкіна-Жарова Тетяна Іванівна
доктор технічних наук, доцент, старший науковий співробітник, Інститут телекомунікацій і глобального інформаційного простору НАН України
Адреса робоча: бульв. Чоколівський, 13, Київ, Україна, 03186
ORCID ID: https://orcid.org/0000-0002-9623-8771 ***e-mail:*** t.pruman@gmail.com

Аброскін Юрій Юрійович
аспірант, Інститут телекомунікацій та глобального інформаційного простору НАНУ
Адреса робоча: бульв. Чоколівський, 13, Київ, Україна, 03186
ORCID ID: https://orcid.org/0009-0009-9828-5596 ***e-mail:*** abroskin21@gmail.com

Дуда Володимир Олександрович
аспірант, Інститут телекомунікацій і глобального інформаційного простору НАН України
Адреса робоча: бульв. Чоколівський, 13, Київ, Україна, 03186
ORCID ID: https://orcid.org/0009-0002-4278-4635 ***e-mail:*** dudavolodimir@gmail.com