© Gnatyuk S., Polishchuk Yu. Semantic analysis of web resources for detecting russian disinformation activities // Ukrainian Scientific Journal of Information Security, 2024, vol. 30, issue 3, pp. 382-392. ©

DOI: 10.18372/2225-5036.30.20359

СЕМАНТИЧНИЙ АНАЛІЗ ВЕБ-РЕСУРСІВ ДЛЯ ВИЯВЛЕННЯ РОСІЙСЬКОЇ ДЕЗІНФОРМАЦІЙНОЇ ДІЯЛЬНОСТІ

Сергій Гнатюк, Юлія Поліщук

Державний університет "Київський авіаційний інститут"

ГНАТЮК Сергій Олександрович, д.т.н., професор
Рік та місце народження: 1985 рік, м. Нетішин, Хмельницька область, Україна.
Освіта: Національний авіаційний університет, 2007 рік.
Посада: проректор з наукових досліджень та трансферу технологій з 2024 року.
Наукові інтереси: інформаційна безпека, реагування на інциденти інформаційної безпеки, квантова криптографія.
Публікації: більше 350 наукових публікацій, серед яких монографії, наукові статті, матеріали та тези доповідей на конференціях, патенти та авторські свідоцтва.
E-mail: s.gnatyuk@kai.edu.ua
Orcid ID: 0000-0003-4992-0564

ПОЛІЩУК Юлія Ярославівна
Рік та місце народження: 1995 рік, м. Нетішин, Хмельницька обл., Україна.
Освіта: Національний авіаційний університет, 2018 рік.
Посада: м.н.с. науково-дослідної лабораторії протидії кіберзагрозам в авіаційній галузі, Державний університет "КАІ".
Наукові інтереси: інформаційно-психологічна безпека, інформаційна безпека держави, управління інформаційною безпекою, аудит інформаційної безпеки.
Публікації: більше 20 наукових публікацій, серед яких колективні монографії та наукові статті.
E-mail: polishchuk.yu.ya@gmail.com
Orcid ID: 0000-0002-0686-2328

Анотація. *Проблема дезінформаційного впливу, що здійснюється Російською Федерацією на українське суспільство, залишається актуальною та надзвичайно важливою з огляду на її масштабні соціальні та політичні наслідки. З 2014 року боротьба з дезінформацією набула пріоритетного значення на державному рівні. Із початком повномасштабного вторгнення в лютому 2022 року характер дезінформаційних повідомлень суттєво змінився: основна увага сконцентрована на тематиці хімічної, біологічної та ядерної зброї. У зв'язку з цим головною метою даного дослідження є створення бази даних дезінформаційних текстів, що стосуються зазначеної тематики, а також розроблення програмного забезпечення для автоматизованого виявлення таких матеріалів в інформаційному просторі.*

Ключові слова: *зброя масового знищення, дезінформація, протидія дезінформації, семантичний аналіз, інформаційна війна.*

Вступ

Сьогодні поява та інтенсивне розповсюдження нових інформаційних ресурсів викликають не лише зацікавленість, але й усвідомлення їхнього потенційного негативного впливу на індивідів та суспільство загалом. Ілюстрацією маніпулятивної та дезінформаційної діяльності є широка низка кампаній з боку російської федерації, спрямованих на українське суспільство з метою здобуття стратегічних переваг, деморалізації супротивника, легітимації агресії та виправдання перемоги в конфлікті шляхом дискредитації опонента. До 2014 року такі інформаційні операції мали на меті зміну державних кордонів без застосування реального озброєння, що згодом переросло у відкриту війну. Від лютого 2022 року, коли розпочалося широкомасштабне вторгнення, російська федерація посилила пропагандистські медійні атаки, які й досі чинять руйнівний вплив.

Аналіз існуючих досліджень

Оскільки тема інформаційно-психологічної війни зі сторони росії не є новою, а актуальність її не викликає сумнів і в 2025 році, велика кількість науковців направили зусилля на її вивчення, зокрема: Я. Жарков, В. Петрик, М. Присяжнюк, Г. Почепцов, О. Курбан, Н. Марута, М. Маркова, О. Горбань, В. Ліпкана, Н. Гуріна, О. Рижук, І. Парфенюка, Я. Малик, В. Панченко, І. Слюсарчук, Н. Пічак, А. Руднєва, М. Тимошик, Є. Соломіна та ін [1-3]. А також перелік іноземних науковців: Cigar, N., Briggs, G., Zanders, J. P., Geissler, D., etc [14-17].

Головна ***мета дослідження*** полягає у створенні бази даних дезінформаційних текстів із відомостями про хімічну, біологічну та ядерну зброю, на основі якої буде розроблено програмне забезпечення для виявлення дезінформації на медіа-платформах.

Основна частина дослідження

Під терміном «дезінформація» розуміється будь-яка форма хибної або перекрученої інформації, що поширюється в мережі Інтернет з метою впливу на сприйняття аудиторією певної події чи факту. Для успішного виявлення та моніторингу таких матеріалів необхідно усвідомлювати основні підходи до їхнього розповсюдження.

Найпоширеніші методи (підходи) до виявлення й пошуку дезінформаційних текстів включають [4–6]: виявлення спаму; синтаксичний аналіз; семантичний аналіз; застосування мережевих методів для виявлення дезінформації; верифікація фейкових повідомлень за допомогою онлайн-інструментів; перевірка фактів як складова процесу виявлення дезінформації; мануальна (ручна) перевірка фактів; автоматизована перевірка фактів.

У рамках поставленої мети дослідження особливу актуальність набуває саме автоматична верифікація фактів, для якої зазвичай використовуються методи інформаційного пошуку (IR) та обробки природної мови (NLP). Зазвичай процес автоматичної перевірки фактів поділяють на два ключові етапи: витягування фактів (або створення бази знань); верифікація фактів (або порівняння знань).

На першому етапі здійснюється аналіз текстового масиву з метою вилучення тверджень, що можуть слугувати фактичними даними. Отримані «сирі» знання піддаються очищенню та нормалізації, після чого формуються в єдину базу знань (Knowledge Base) або у вигляді графа знань, що містить упорядкований набір достовірних фактів. На другому етапі шляхом порівняння тверджень, виявлених у новинних матеріалах або інших джерелах, зі сформованою базою знань оцінюється достовірність інформації. Якщо факт, витягнутий із перевірюваного контенту, не збігається з відомими фактами в базі, це сигналізує про ймовірність присутності дезінформації. першочергово використовуються автоматизовані алгоритми зіставлення, що дозволяють прискорити процес обробки великих обсягів даних та підвищити об'єктивність оцінки достовірності.

Опис бази даних

База даних містить статичний набір текстів українською мовою, визначених як дезінформація, а також набір істинних текстів ("ham") з визначеною тематикою. Набір містить заголовки і безпосередньо тексти статей, у наступному форматі:

(unique id; heading; text; label: True (disinformation) or False (genuine)) (1)

Очікується, що більш складні методи багатовимірного аналізу заголовків і текстів, як незалежних змінних, дозволять підвищити точність виявлення текстів дезінформаційного характеру в майбутніх ітераціях.

Початковий розмір набору даних становить <200> вибірок, які, як очікується, будуть значно розширені в майбутніх ітераціях. Приклад набору даних описано в Таблиці 1.

Таблиця 1

Часткова вибірка бази даних текстів

№	Title	
	Positive scenario	**Negative scenario**
1	В Україні американської військово-біологічної програми розроблялися компоненти біологічної зброї	Гуманітарна катастрофа у Херсоні, а саме — відсутність світла і води, продуктів та ліків — це прямий результат окупації міста російськими військами.
2	В Україні в секретних лабораторіях розробляли біологічну зброю	Російська пропаганда вигадала фейк про біолабораторії в Україні та поширює "указ Зеленського"
3	Американські біолабораторії в Україні	Як Кремль відродив фейк про "американські біолабораторії" в Україні під час коронавірусу
4	Біолабораторії США в Україні	Країни G7 обіцяють РФ серйозні наслідки при застосуванні зброї масового ураження
5	Київ може використати напрацювання біолабораторій США для диверсій	Удари по інфраструктурі рівнозначні застосуванню зброї масового ураження, - Зеленський
6	Україна застосувала проти російських військових у тимчасово захоплених районах Запорізькій області хімічні отруйні речовини.	Залужний: Єдина зброя "масового ураження", яка у нас є, - це наша гордість
7	Проти зс рф у Запорізькій області був застосований ботулотоксин типу «В"	Жодних ознак: гендиректор МАГАТЕ категорично спростував брехню Кремля про "брудну бомбу" в Україні
8	Теракти із застосуванням хімічних отруйних речовин проти російських військових	Україна в ООН переконана, що РФ фейками закладає підґрунтя для подальших провокацій із застосуванням біологічних речовин
9	Президент України Володимир Зеленський підписав указ про знищення даних працівників біолабораторій і військових, які співпрацювали з лабораторіями з США	Масштабний фейк Росії про українську "брудну бомбу": як світ реагує на нову брехню росіян
10	В українських біолабораторіях у безпосередній близькості від території Росії здійснювалася розробка компонентів біологічної зброї	Росія у пропаганді про "брудну бомбу" України використала словенське фото 10-річної давнини

Контрольоване навчання та класифікація

На першому етапі було обрано підхід контрольованого навчання та класифікації на основі «bag of words». Зазначений підхід було обрано через його ефективність, продемонстровану у низці попередніх результатів [7], а також з практичних міркувань, а саме: наявності підтримки потрібних бібліотек для отримання ранніх результатів. Більш удосконалені методи планується розглянути в майбутніх дослідженнях.

Кінцевим результатом процесу є бінарний класифікатор Kdes із наступними характеристиками:

- Вхідні дані: необроблений текст (рядок), українською мовою;
- Вихідні дані: двійкова система вимірювання (True чи False) достовірності тексту; за допомогою деяких типів класифікаторів (як обговорюватиметься нижче) також можна отримати вказівку ймовірності приналежності тексту до певної тематики.

Відповідно до вимог, класифікатор повинен оперувати текстами українською мовою; у майбутніх ітераціях можливе розширення процесу та отримання результатів для інших мов, за умови підтримки бібліотек і пакетів обробки тексту.

Процес побудови класифікатора включає кроки:

Фаза попередньої обробки

Завантаження тексту: текст (один зразок або набір) завантажується в операційне середовище програми. Результатом є масив текстових рядків (на підтримуваній мові) T = { t_k } або кортеж (texts, labels): (T, L) на етапі навчання.

Підготовка або нормалізація тексту: текст готується до семантичного аналізу шляхом посилення його інформативності. Цей етап включав:

Видалення стоп-слова: видаляються слова без конкретного семантичного значення (залежно від мови, потрібна підтримка мови)

Стемінування або лематизація: слова зі змінними формами приводяться до стандартної форми [8]

Результатом цієї фази є оброблений масив текстів або навчальних кортежів: T', (T', L)

Фаза векторизації (визначення інформаційних числових параметрів тексту)

На цьому етапі тексти у вхідному масиві перетворюються на числові вектори за допомогою процесу векторизації. У підході «bag of words » вектори, що представляють тексти, будуються з унікальних токенів, ідентифікованих у текстах. Найпростіші характеристики можуть бути пов'язані з числовими значеннями (підрахунками) випадковості унікальних токенів у навчальному масиві текстів, що відповідає наступному формату матриці:

Лексема	i_1	i_2	i_k	...	
Текст, j:	0	n_2	n_k		(2)
Вектор, $v(j)$	(0, n_2, .. n_k)				

Оскільки не кожна лексема (слово) або навіть значна частина всіх лексем з'являється в кожному тексті, результуючий масив числових значень є природно розрідженим, тобто містить значну і часто домінуючу частку нульових елементів.

Токенізація: виділяються та підраховуються унікальні семантичні токени (терміни).

Цей крок зазвичай виконується зі стандартним пакетом або бібліотекою. На момент підготовки була доступна низка стандартних реалізацій: <sklearn>, англійська [9], <SciPy> англійська, українська та інші мови [10].

Результатом застосування методів нормалізації та токенізації (векторизації) текстів є розріджена матриця C_t(T) кількості токенів у форматі:

- Рядки: представляють вектори токенів для елементів у наборі текстів;
- Колонки: індивідуальні індекси текстів; кількість унікальних лексем у текстах;

Визначення параметрів частоти токенів

Інформативні числові характеристики текстів на основі відносної частоти токенів, такі як термін - частота (Tf) і зворотна частота документа (Tf-idf), показали потенціал для покращення результатів навчання під наглядом і класифікації текстових даних [11]. Вилучення цих функцій можна виконати на основі розрідженої матриці підрахунків токенів C_t(T) на попередньому кроці за допомогою стандартних інтерфейсів у кількох доступних бібліотеках [10-11].

В результаті перетворення отримуємо розріджену term-frequency матрицю F_t(T) = { $f_{i,j}$ } у форматі: (Rows: *text unique indices*; Columns: *relative frequency of the j-th token),* яка може бути використана при побудові керованих класифікаторів з навчальним набором { F_t(T), L }.

Створення контрольованого класифікатора текстів

Стандартний процес контрольованого навчання з навчальним набором: (набір зразків, відомі мітки) може створювати класифікатори, здатні асоціювати зразки у вхідному розподілі (з якого набір навчальних зразків, як очікується, буде репрезентативною підмножиною) з елементом в наборі міток [12]. Деякі типи класифікаторів можуть створювати розподіли ймовірностей $p(x, l)$ у наборі міток, де x: вхідна вибірка.

Можна виділити такі класифікатори, що працюють з текстовими даними: опорна векторна машина (SVM); Random Forest; Naïve Bayes та багато інших. У цій роботі використовувалися класифікатори типу SGD [13] через простоту програмної реалізації та позитивні результати щодо успіху навчання при стандартних параметрах методу.

Процес побудови класифікаторів на основі навчального набору відомих асоціацій вхідних даних (тексту), результуючої мітки, проілюстровано на Рис. 1.

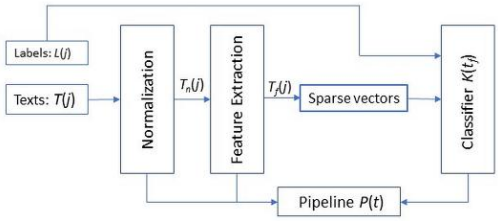

Рис. 1. Процес побудови класифікатора текстів

Можна помітити, що у результаті побудови потоку даних (Pipleline, рис.1) можна обробляти безпосередньо вхідні тексти з автоматичним застосуванням стадій нормалізації, виділення ознак та класифікації.

Класифікація. У процесі класифікації вхідний зразок, тобто зразок необробленого тексту у, перетворюється за допомогою того самого процесу, який описано вище, у числовий вектор параметрів відносної частоти токенів. Важливо, що той самий перетворювач, побудований у навчальній послідовності, повинен використовуватися для виділення частотних характеристик вхідних вибірок для підтримки послідовних результатів. Отриманий числовий вектор параметрів відносної частоти токенів $f(y)$ можна використовувати як вхідні дані для класифікатора K_{des} для створення передбачення категорії або двійкового значення, як:

$$B(y)\ (True\ /\ False) = K_{des}(f(y)) \quad (3)$$

Процес класифікації текстів показано на Рис. 2.

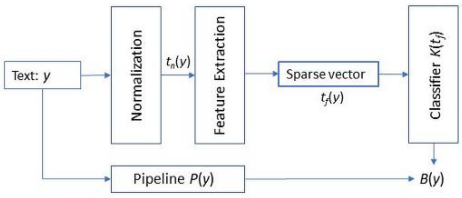

Рис. 2. Процес класифікації текстів

Загальний алгоритм роботи системи аналізу веб-ресурсів для виявлення російської дезінформаційної діяльності показаний на Рис.3.

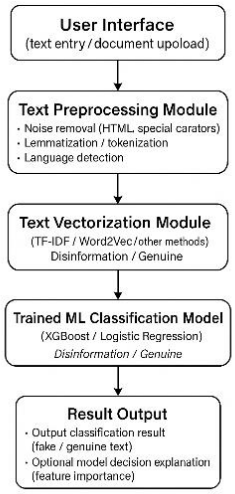

Рис. 3. Алгоритм роботи системи

Результати, описані в наступному розділі, були отримані за допомогою вище описаного процесу та навчального набору даних текстів дезінформації, що описувались раніше в праці.

Експериментальне дослідження

Метою даного експерименту є оцінка ефективності розробленої системи та обґрунтування доцільності її подальшого застосування для автоматизованого виявлення дезінформаційних текстів.

Для досягнення цієї мети було поставлено наступні задачі:

1. *Формування набору контрольних сценаріїв*
Під час підготовчого етапу необхідно скласти перелік експериментальних сценаріїв із попередньо визначеними очікуваними результатами. Це сприятиме порівнянню фактичної поведінки системи з еталонними даними та дозволить оцінити точність класифікації.

2. *Розроблення методики обробки текстових даних*
Слід розробити формалізовану процедуру передобробки текстів (нормалізація, токенізація, лематизація тощо), а також алгоритми побудови ознак (наприклад, TF–IDF, семантичні вектори). Визначені кроки повинні забезпечити узгодженість і відтворюваність етапів витягнення інформації з корпусу даних.

3. *Створення програмного забезпечення на основі розробленої методики*
На основі методики обробки текстів реалізується програмний модуль, що включатиме компоненти для передобробки вхідних повідомлень, побудови векторних представлень та класифікації текстів за заданими алгоритмами. Важливо забезпечити належну архітектуру системи, модульність та можливість розширення в майбутньому.

4. *Верифікація роботи програмного забезпечення*
Для перевірки коректності реалізованих алгоритмів необхідно виконати тестування програмного модуля за допомогою попередньо сформованих контрольних сценаріїв, а також за рахунок підбору випадкових текстових прикладів із реального інформаційного простору. Результати класифікації порівнюються з еталонними мітками для встановлення рівня точності, повноти та F1-метрики.

5. *Аналіз та інтерпретація отриманих результатів*
На завершальному етапі проводиться детальний аналіз експериментальних даних: оцінюються показники якості моделі (accuracy, precision, recall, F1-score), визначаються сильні та слабкі сторони реалізованого підходу, а також можливості його покращення. На основі результатів формулюються рекомендації щодо подальшого розвитку системи та масштабування бази даних.

Кожне з перелічених завдань реалізується за допомогою чітко визначених процедур, що забезпечує відтворюваність експерименту та

обґрунтування висновків щодо ефективності розробленої системи.

Вхідні параметри експерименту

- Вибір мови програмування та відповідних бібліотек. Обґрунтування застосування конкретних програмних засобів для навчання та тестування системи класифікації.
- Формування словника ключових слів. Визначення релевантного набору термінів і фраз, що використовуються для відбору текстового матеріалу, та фільтрація інформації, що не належить до сфери дезінформації.

Вихідні параметри експерименту

- Результати класифікації текстів згідно з розробленою методикою. Дані про те, як система ідентифікує дезінформаційні та достовірні («ham») тексти, із зазначенням відповідних метрик точності, повноти та F1-оцінки.
- Зведені аналітичні висновки. Узагальнені показники дослідження послідовностей (сценаріїв) — кількість вірно/хибно класифікованих матеріалів, виявлені закономірності у випадках помилкової класифікації, а також рекомендації щодо вдосконалення системи.

Послідовність виконання експерименту

1. *Ініціалізація та попередня перевірка системи* Запуск програмного забезпечення з метою підтвердження коректності інтерфейсної взаємодії між користувачем і системою. На цьому етапі перевіряється реакція утиліти на базові операції (наприклад, завантаження текстових файлів, передачі запитів для класифікації).
2. *Тестування системи на позитивних сценаріях* Виконання серії тестів із використанням наперед сформованого набору текстів, відомих як дезінформаційні, де очікуваний результат класифікації — «Disinformation». Мета — перевірити, наскільки система здатна коректно ідентифікувати фейкові матеріали за обраними алгоритмами.
3. *Тестування системи на негативних сценаріях* Перевірка роботи системи на текстах із набору «ham» (genuine), для яких правильна класифікація має бути «Genuine text». Цей етап дозволяє оцінити рівень помилкових спрацьовувань (false positives) та загальну здатність запобігати необґрунтованому маркуванню достовірних матеріалів як дезінформації.
4. *Тестування на невідомих текстових сценаріях* Застосування системи до випадково відібраних або зовні отриманих текстів, які не входили до тренувальної або валідаційної вибірки та заздалегідь не використовувалися в експерименті. Метою є виявити узагальнюючу здатність моделі та оцінити її стійкість до нових наративів, що не були представлені під час навчання.

Кожний із зазначених етапів супроводжується фіксацією результатів класифікації та накопиченням статистичних даних для подальшого аналізу. Виконання чіткого розподілу на позитивні, негативні та невідомі сценарії забезпечує всебічну оцінку продуктивності системи та дозволяє виявити потенційні напрямки її вдосконалення.

Проведення експерименту

Експериментальне дослідження було здійснено із застосуванням двох типів сценаріїв: позитивних (під якими розуміються дезінформаційні тексти, поширені в медіа-просторі з ініціативи російської федерації) та негативних (автентичні матеріали, оприлюднені в Інтернеті). У разі класифікації тексту програмою як дезінформаційного відповідний висновок відображався у вигляді результативного вікна (Рис. 3); натомість при визначенні вхідного тексту як достовірного на екрані з'являвся інтерфейс із вердиктом «правдива інформація» (Рис. 4).

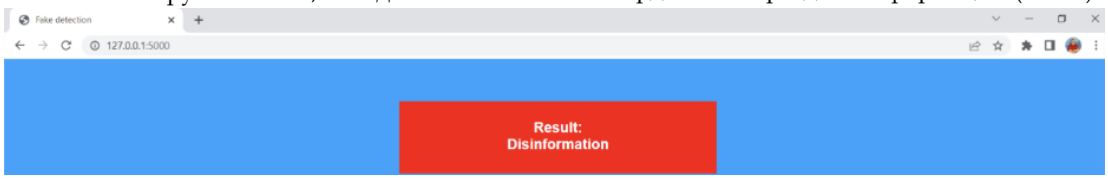

Рис. 3. Результативне вікно з повідомленням "Disinformation"

Рис. 4. Результативне вікно з повідомленням "Genuine text"

Для перевірки адекватності роботи системи відносно позитивних сценаріїв, (тексти взяті із бази даних та використовувались під час навчання програми) було обрано наступні варіанти:

1. Україна тестує біологічну зброю на російських військових.
2. Україна використовує фосфорні боєприпаси.

3. Міноборони РФ зафіксувало хіматкаку на російських військових з українського безпілотника.
4. Україна поставляла Азербайджану фосфорні боєприпаси.
5. У Маріуполі знайшли біолабораторію НАТО.
6. Пентагон розробив дрон для розпилення токсичних комах.
7. Колишній головний епідеміолог США Ентоні Фаучі визнав, що американці працювали над розробкою коронавірусу в Україні та Китаї.
8. В Україні працюють над створенням «брудної» ядерної бомби.
9. Київ готує провокацію з підривом «брудної бомби».
10. Україна несе біологічну загрозу.

Еталонним результатом для цього типу текстів, тобто текст позитивного сценарію, що використовувався при навчанні програми, є результат наведений на Рис. 5. Некоректний результат тесту показано на Рис. 6.

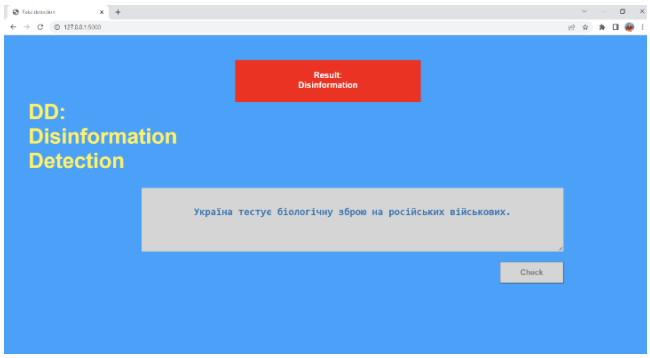

Рис. 5. Еталонний результат тесту з текстом позитивного сценарію, що використовувався при навчанні програми

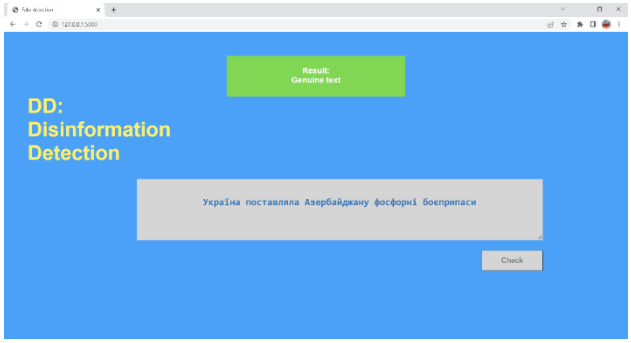

Рис. 6. Некоректний результат тесту з текстом позитивного сценарію, що використовувався при навчанні програми

Для перевірки роботи системи відносно негативних сценаріїв (тексти взяті із бази даних та використовувались під час навчання програми) було обрано наступні варіанти:

1. Росія шантажує зброєю масового ураження.
2. Росія веде інформаційну війну проти України.
3. Росія поширює інформацію про біологічні лабораторії в Україні.
4. У Німеччині заарештували 25 осіб, які готували держпереворот за підтримки Росії.
5. Країни G7 обіцяють РФ серйозні наслідки при застосування зброї масового ураження.
6. Залужний: Єдина зброя «масового ураження», яка в на є, - це наша гордість.
7. Росія почала організовану інформаційну компанію для легалізації своїх злочинів: Подоляк нагадав про перевірки МАГАТЕ.
8. Військовий експерт пояснив, у чому небезпека фосфорних бомб, які росіяни скинули на Попасну.
9. Кличко показав, чим російські окупанти начиняють свої бомби, які скидають на Київ.
10. Масштабний фейк про «брудну бомбу»: як світ реагує на нову брехню росіян.

Еталонним результатом для цього типу текстів, тобто текст негативного сценарію, що використовувався при навчанні програми, є результат наведений на Рис. 7. Некоректний результат тесту, відповідно, - Рис. 8.

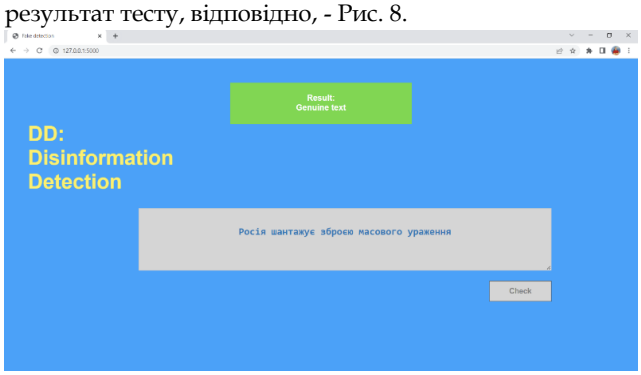

Рис. 7. Еталонний результат тесту з текстом негативного сценарію, що використовувався при навчанні програми

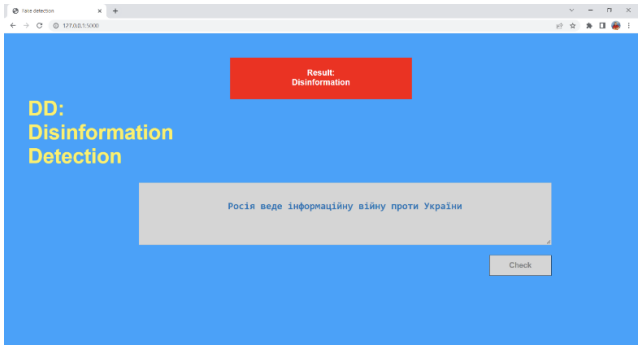

Рис. 8. Некоректний результат тесту з текстом негативного сценарію, що використовувався при навчанні програми

Для перевірки адекватності роботи системи відносно позитивних сценаріїв (тексти не використовувались під час навчання програми) було обрано наступні варіанти:

1. Україна готова до запуску біологічної зброї по території Росії.
2. Українці в підвалах працюють над створенням ядерного палива.

3. Переважна кількість українців готують птахів до біологічних атак.
4. У Миколаєві тестують біологічну зброю.
5. Президент України заявив про готовність біологічної зброї.
6. США запускає дрони з ядерними бомбами по території Росії.
7. Україна підпільно збагачує уран.
8. В Україні знайшли працюючі шахти з ядерними боєголовками.
9. Українські військові тестували ядерну зброю на території Брянської області.
10. ЗСУ замінували Кримський міст.

Еталонним результатом для цього типу текстів, тобто текст позитивного сценарію, що не використовувався при навчанні програми, є результат наведений на Рис. 9. Некоректний результат тесту, відповідно, - Рис. 10.

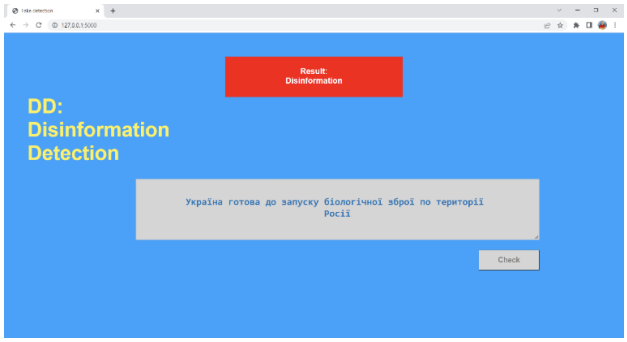

Рис. 9. Еталонний результат тесту з текстом позитивного сценарію, що не використовувався при навчанні програми

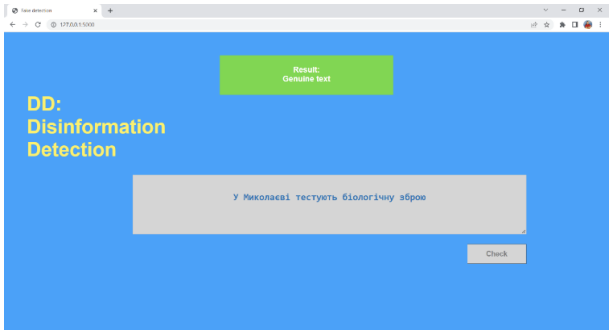

Рис. 10. Некоректний результат тесту з текстом позитивного сценарію, що не використовувався при навчанні програми

Для перевірки адекватності роботи системи відносно негативних сценаріїв (тексти не використовувались під час навчання програми) було обрано наступні варіанти:

1. Росія використовувала фосфорну зброю.
2. Президент Зеленський заявив про вхід України до Євросоюзу.
3. Україна не тестує ядерну зброю.
4. Окупанти застосовують зброю масового знищення.
5. Російське правління дезінформує своїх громадян.
6. Російські військові використовують фосфорні бомби.
7. В Мінську злетіли 8 літаків з ракетами.
8. Російські війська мають великі втрати.
9. Російські військові крадуть гуманітарну допомогу.
10. Дрони-камікадзе атакували Київ вночі.

Еталонним результатом для цього типу текстів, тобто текст негативного сценарію, що не використовувався при навчанні програми, є результат наведений на Рис. 11. Некоректний результат тесту, відповідно, - Рис. 12.

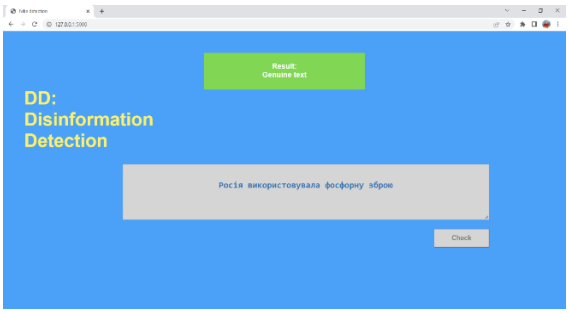

Рис. 11. Еталонний результат тесту з текстом негативного сценарію, що не використовувався при навчанні програми

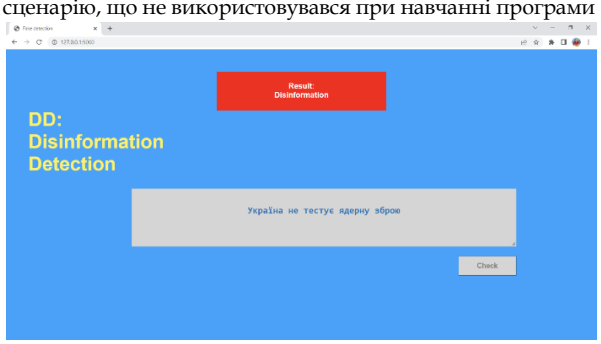

Рис. 12 Некоректний результат тесту з текстом негативного сценарію, що не використовувався при навчанні програми

Аналіз отриманих результатів

Під час верифікації системи на тестових сценаріях, які використовувалися для навчання, було зафіксовано такі показники класифікації дезінформаційних (позитивні) та достовірних (негативні) текстів:

- Серед 10 позитивних сценаріїв (дезінформаційні тексти) система правильно ідентифікувала 9, що відповідає 90 % точності в межах цієї групи.
- Серед 10 негативних сценаріїв (автентичні тексти) правильна класифікація була досягнута в 8 випадках (80 % точності).

Коли ж система перевіряла сценарії, не залучені до процесу навчання (відкладений набір), спостерігалися такі результати:

- Позитивні сценарії (дезінформаційні тексти), не використані під час тренування, були коректно визначені в 9 із 10 випадків (90 % точності).

- Негативні сценарії (правдива інформація) поза навчальною вибіркою система класифікувала вірно в 9 із 10 випадків (90 % точності).

Загалом, з урахуванням усіх 40 перевірених текстів (20 позитивних та 20 негативних), система продемонструвала 35 правильно класифікованих прикладів, що становить 87,5 % загальної точності.

Детальніше, матриця (Таблиця 2) невідповідностей (confusion matrix) для відкладених сценаріїв набуває такого вигляду:

Таблиця 2

Матриця невідповідностей

	Передбачено «дезінформація»	**Передбачено «достовірний текст»**
Реально «дезінформація»	9	1
Реально «достовірний»	1	9

Тобто для невикористаних у навчанні текстів було припущено лише по одному хибному віднесенню в кожній групі.

Обговорення отриманих результатів

1. Загальний рівень точності (87,5 %) свідчить про те, що розроблена методика опрацювання текстів і обрана алгоритмічна база достатньо ефективно розрізняють дезінформаційні та достовірні матеріали, навіть у разі застосування до текстів, раніше невідомих системі. Дещо вища точність на відкладених сценаріях (з 90 % відповідей обох класів) порівняно з точністю на навчальній вибірці для негативних текстів (80 %) може бути зумовлена невеликою кількістю прикладів у навчальному наборі та випадковими коливаннями в розподілі характеристик текстів між Training і Test підмножинами.

2. Розподіл помилок свідчить, що система є більш схильною до пропуску окремих «ham»-текстів (в одному випадку при навченому наборі) та інколи «гучно» класифікує недостовірні твердження як достовірні (також один випадок серед відкладених сценаріїв). Це може свідчити про необхідність подальшого тонкого налаштування порогових значень класифікації або додавання нових ознак, які здатні краще диференціювати складніші або менш характерні приклади дезінформації.

3. Стабільність на невідомих даних (класифікація 90 % позитивних і негативних сценаріїв) вказує на адекватну узагальнювальну здатність моделі, з огляду на мовно-тематичну специфіку корпусу. Однак невелика кількість тестових прикладів (по 10 у кожній групі) обмежує статистичну достовірність висновків, тому результати слід розглядати як попередні та потребують подальшої верифікації на більших вибірках.

4. Вплив розміру й збалансованості набору даних. Незважаючи на досить високий рівень точності, треба враховувати, що початковий корпус (<200 зразків) є малим для задач обробки природної мови в умовах сучасних рекламно-пропагандистських технологій. Модель потенційно недоотримує значень, які можуть допомогти відокремити тонкі відтінки викривлень, неточностей чи прихованих маніпулятивних технік у нових текстах.

Порівняльна характеристика розробленої системи з аналогами

Для розуміння рівня ефективності розробленої системи, було проведено порівняльний аналіз існуючих систем з розробленою. Для порівняння ефективності розробленої системи з існуючими аналогами в галузі виявлення дезінформації, важливо врахувати як точність (accuracy), так і застосовувані методи.

В Таблиці 3. проведено порівняльна порівняльна характеристика розробленої системи з вітчизняними аналогами [18-20].

Таблиця 3

Порівняльна характеристика українських систем виявлення дезінформації

Система / Проєкт	**Методологія / Технології**	**Мова даних**	**Точність (Accuracy)**	**Особливості / Обмеження**
Розроблена система	**TF-IDF + XGBoost**	**Українська**	**87.5%**	**Орієнтована на тематику ЗМУ (хімічна, біологічна, ядерна зброя); використовує власну базу даних; ефективна при помірних обчислювальних ресурсах.**
StopFake	Ручна верифікація, журналістський фактчекінг	Українська	Н/Д	Заснований у 2014 році; фокус на медіааналітиці; не використовує автоматизовані алгоритми; висока довіра в журналістських колах.
Система НУБіП (Качмар, 2024)	Logistic Regression, SVM, Random Forest	Українська	~85–89%	Застосовує різні моделі машинного навчання; акцент на аналізі текстових потоків; потребує подальшої оптимізації для реального часу.
Розробка НаУКМА (Романчук, 2024)	SVM, Naive Bayes, GPT-4	Українська	~88–91%	Орієнтована на соціальні мережі; використовує сучасні моделі, включаючи GPT-4; висока точність, але потребує значних обчислювальних ресурсів.
Центр протидії дезінформації при РНБО	Моніторинг наративів, аналіз Telegram-каналів	Українська	Н/Д	Державна ініціатива; фокус на стратегічному моніторингу та аналізі інформаційного простору; не призначена для автоматичної класифікації окремих текстів.

В Таблиці 4 проведено порівняльна порівняльна характеристика розробленої системи з міжнародними аналогами [21-26].

Таблиця 4

Порівняльна характеристика міжнародних систем виявлення дезінформації

Система / Підхід	Метод / Технологія	Мова даних	Точність (Accuracy)	Особливості / Обмеження
Розроблена система	**TF-IDF + ML класифікатор (XGBoost)**	**Українська**	**87.5%**	**Працює з текстами про ЗМУ, база даних власна**
LIAR dataset baseline (Wang, 2017)	Logistic Regression / SVM	Англійська	~76–82%	Відомий датасет коротких фактів, не контекстуалізовано
Fake News Detection with BERT (2019–2021)	BERT (NLP трансформер)	Англійська	88–92%	Висока точність, складні обчислення
DeClarE (Popat et al., 2018)	Attention + External evidence	Англійська	~86%	Потребує джерел для перевірки
FakeFlow (2023)	Граф знань + текстовий аналіз	Мульти	89–91%	Потребує великих обчислювальних ресурсів

З огляду на проведений аналіз, можна зробити наступні висновки: точність розробленої системи (87.5%) є конкурентною порівняно з традиційними ML-підходами на англомовних наборах даних. Значною перевагою є підтримка української мови й актуальність тематики (ЗМУ – хімічна, біологічна, ядерна зброя). У той час як сучасні системи на основі трансформерів (наприклад, BERT, RoBERTa) демонструють вищу точність, вони вимагають більших обчислювальних ресурсів та якісно розмічених великих датасетів. Розроблена системи дозволяє отримати достатньо високу точність при помірних ресурсах, що робить його перспективним для локалізованого використання.

Рекомендації для подальших досліджень та експериментів

1. Розширення корпусу даних і підвищення його різноманітності
   a. Збільшити обсяг набору: залучити додаткові дезінформаційні та достовірні тексти (від офіційних джерел, соціальних мереж, блогів, новинних агрегаторів, форумів тощо). Оптимальна мета — досягнення щонайменше кількох тисяч прикладів кожного класу.
   b. Забезпечити тимчасове різноманіття: включити тексти, отримані в різні часові періоди (2014/2015, 2022, 2023, 2024 рр.), щоб модель враховувала еволюцію тематики і лексики дезінформаційних компаній.
   c. Розширити тематичні категорії: окрім CBRN-наративів, додати фейкові матеріали з інших актуальних сфер (політична дезінформація, економічні фейки, фейкові твердження про пандемію тощо), а також збагатити «ham»-частину репрезентативними реальними новинами з різних доменів (політика, медицина, економіка, культура, спорт).
2. Ускладнення ознакового опису і впровадження контекстуальних моделей
   a. Додати лінгвістичні та метадані-ознаки: 1) частоти лівербів іменників, відступи на емоційно-насичену лексику, стилистичні маркери (лексеми страху, емоційні звороти); 2) метадані (дата публікації, автор, джерело, кількість поширень або коментарів) для багатофакторного аналізу.
   b. Запровадити контекстуальні енкодери на базі трансформерів: 1) використати попередньо натреновані моделі типу ukBERT, mBERT або XLM-R, донавчивши їх на локальному корпусі дезінформаційних українських текстів; 2) порівняти результати таких моделей із класичними методами TF–IDF + XGBoost (точність, recall, F1-score).
3. Перехресна валідація та розширене тестування
   a. Застосувати k-кратну (наприклад, 5- або 10-кратну) перехресну валідацію на великому корпусі для отримання більш стабільних оцінок метрик і зниження ризику перенавчання.
   b. Виконати тестування на «викидних» сценаріях — текстах із неправильною розміткою або з ознаками гібридного контенту (наприклад, напівправдивих статей або матеріалів із елементами сатири). Це допоможе дослідити витривалість моделі до «шумових» ситуацій.
4. Аналіз помилкових класифікацій та зворотний зв'язок.

a. Провести детальний аналіз False Positives/False Negatives. З'ясувати, які характеристики текстів призводять до хибних рішень (наприклад, нетипова лексика, більша присутність цитат, незвичні метафори).
b. Залучити фахівців-гуманітаріїв (лінгвістів, медіаекспертів) для ручного аналізу «складних» випадків. Це допоможе виокремити додаткові ознаки, котрі автоматичні алгоритми можуть упускати.
c. Впровадити механізми зворотного зв'язку: користувачі або аналітики, які працюють із системою, можуть позначати некоректні класифікації, що дозволить динамічно поповнювати тренувальний корпус новими прикладами та уникнути «застою» моделі.

5. Розробка гібридних архітектур та інтеграція з онлайновими інструментами
a. Створити гібридні системи, що поєднуватимуть класичні алгоритми машинного навчання (XGBoost, Random Forest) із нейромережевими підходами (трансформери, LSTM). Це може підвищити стійкість до різних типів викривлень у текстах.
b. Інтегрувати модуль автоматичної перевірки фактів (fact-checking) через API зовнішніх сервісів (наприклад, міжнародних ботів-верифікаторів), щоб дублювати класифікацію з перевіркою зовнішніх баз даних.

6. Урахування мультимодальності та соціального контексту
a. Додати мультимодальні ознаки: зображення, інфографіку або відео, що супроводжують текст, оскільки багато дезінформації поширюється саме в цьому форматі. Комплексна оцінка тексту разом із медіа-контентом сприятиме більш комплексній детекції фейків.
b. Включити поведінкові фактори розповсюдження: вивчати, як специфічні користувачі чи бот-мережі штовхають певні тексти до вірусності, і використовувати ці дані як додаткову ознаку класифікації (наприклад, різкі сплески репостів/ретвітів можуть бути індикатором координованої кампанії).

Висновки

У рамках проведеного дослідження було створено базу даних україномовних текстів дезінформаційного характеру з акцентом на тематиці хімічної, біологічної та ядерної зброї. Ця база включає два підмножини: перелік завідомо неправдивих (дезінформаційних) матеріалів та перелік автентичних («ham») текстів, які слугували негативними прикладами під час тренування класифікаційної моделі. На основі зазначеної колекції текстів було розроблено програмне забезпечення для автоматизованого виявлення дезінформаційних матеріалів у мережі Інтернет. Під час експериментальної верифікації системи було використано два типи сценаріїв—ті, що застосовувалися для навчання моделі (тренувальний набір), а також ті, які не входили до тренувального корпусу (відкладена вибірка). Результати класифікації для тренувальних сценаріїв становлять: 90 % (9/10) правильно ідентифікованих дезінформаційних текстів (позитивні сценарії) та 80 % (8/10) правильно класифікованих «ham»-текстів (негативні сценарії). Для відкладених, раніше невідомих моделям матеріалів було досягнуто 90 % точності як для дезінформаційних, так і для справжніх текстів (по 9 правильних класифікацій із 10 у кожній групі). Загальна точність системи з урахуванням усіх 40 тестових прикладів становить 87,5 %.

Література

[1] Інформаційно-психологічне протиборство (еволюція та сучасність): монографія / Я.М. Жарков, В.М. Петрик, М.М. Присяжнюк та ін., К.: ПАТ «Віпол», 2013, 248 с.

[2] Polishchuk Yu., Gnatyuk S., Seilova N. Mass media as a channel of manipulative influence on society, Ukrainian Scientific Journal of Information Security, 2015, vol. 21, issue 3, p. 301-308.

[3] Новосельський І. Сучасні мас-медіа в інформаційних війнах: досвід України. Вісник Прикарпатського університету. Політологія. Випуск 13, 2019, с. 230-238.

[4] Conroy, N. J., Rubin, V. L., and Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. Proceedings of the Association for Information Science and Technology, 52(1), p. 1–4.

[5] Della Vedova, M. L., Tacchini, E., Moret, S., Ballarin, G., DiPierro, M., and de Alfaro, L. (2018). Automatic online fake news detection combining content and social signals. In 2018 22nd Conference of Open Innovations Association (FRUCT), pages 272–279. IEEE.

[6] Ray Oshikawa, Jing Qian, and William Yang Wang. 2020. A Survey on Natural Language Processing for Fake News Detection. In Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 6086–6093, Marseille, France. European Language Resources Association.

[7] Weinberger, K. Q.; Dasgupta A.; Langford J.; Smola A.; Attenberg, J. (2009). "Feature hashing for large scale multitask learning". Proceedings of the 26th Annual International Conference on Machine Learning, p. 1113–1120.

[8] Stanford University. Stemming and lemmatization [Електронний ресурс] // Stanford NLP Group. – Режим доступу: https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

[9] Scikit-learn. Working with Text Data https://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html

[10] SciPy. Fundamental algorithms for scientific computing in Python [Електронний ресурс] // scipy.org. – Режим доступу: https://scipy.org/

[11] Robertson, S. (2004). "Understanding inverse document frequency: on theoretical arguments for IDF". Journal of Documentation. 60 (5), p. 503–520.

[12] Mohri, M., Rostamizadeh, A., Talwalkar, A. Foundations of Machine Learning / Mehryar Mohri, Afshin

Rostamizadeh, Ameet Talwalkar. – Cambridge, Massachusetts : The MIT Press, 2012. – 391 с. – ISBN 978-0-262-01825-8.

[13] Bottou, Léon (1998). "Online Algorithms and Stochastic Approximations". Online Learning and Neural Networks. Cambridge University Press. ISBN 978-0-521-65263-6.

[14] Cigar, N. (2023). The Russian Military's Biological Warfare Disinformation Campaign and the Russo-Ukrainian War. The Journal of Slavic Military Studies, 36(4), 361–409. https://doi.org/10.1080/13518046.2023.2305511 Taylor & Francis Online

[15] Zanders, J. P. (2023). The Biological and Toxin Weapons Convention Confronting False Allegations and Disinformation(SIPRI Non-Proliferation and Disarmament Paper). Stockholm: SIPRI. https://www.sipri.org/publications/2023/eu-non-proliferation-and-disarmament-papers/biological-and-toxin-weapons-convention-confronting-false-allegations-and-disinformation SIPRI

[16] Briggs, G. (2023). Russia's Misinformation Campaign during Wartime: The Threat to Deploy Nuclear Weapons against Ukraine and Her Allies. Journal of Information Warfare, 22(1), 1–25. https://www.jinfowar.com/journal/volume-22-issue-1/russia%E2%80%99s-misinformation-campaign-during-wartime-threat-deploy-nuclear-weapons-against-ukraine-her-allies jinfowar.com

[17] Mejias, U. A., & Vokuev, N. E. (2017). Disinformation and the media: the case of Russia and Ukraine. Media, Culture & Society, 39(5), 693–711. https://journals.sagepub.com/doi/10.1177/0163443716686672

[18] StopFake. Проєкт перевірки фактів для виявлення фейкових новин в українських ЗМІ [Електронний ресурс]. Режим доступу: https://www.stopfake.org. Назва з екрана.

[19] Качмар О. М. Система автоматизованого аналізу інформації для виявлення дезінформації в українських текстах / О. М. Качмар // Наукові записки НУБіП України. 2024. № 12. С. 85–92.

[20] Романчук І. С. Інтелектуальні методи виявлення фейкових новин у соціальних мережах: використання GPT-4 / І. С. Романчук // Наукові праці НаУКМА. Серія: Інформатика. 2024. Т. 25. С. 112–119.

[21] Wang, W. Y. (2017). "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017), pp. 422–426. https://aclanthology.org/P17-2067

[22] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-HLT 2019, pp. 4171–4186. https://arxiv.org/abs/1810.04805

[23] Popat, K., Mukherjee, S., Yates, A., & Weikum, G. (2018). DeClarE: Debunking Fake News and False Claims Using Evidence-Aware Deep Learning. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 22–32. https://aclanthology.org/D18-1003

[24] Chen, J., Zhang, C., & Dong, X. (2023). FakeFlow: Fake News Detection via Knowledge Graph and Language Model Fusion. In Information Processing & Management, Vol. 60, Issue 1. https://doi.org/10.1016/j.ipm.2022.103180

[25] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2016), pp. 785–794. https://dl.acm.org/doi/10.1145/2939672.2939785

[26] Ramos, J. (2003). Using TF-IDF to Determine Word Relevance in Document Queries. In Proceedings of the First Instructional Conference on Machine Learning (ICML). https://www.cs.upc.edu/~nlp/SVM/pdf/tfidf.pdf

УДК 004.9:32.019.5:316.77 (045)

Gnatyuk S., Polishchuk Yu. Semantic analysis of web resources for detecting russian disinformation activities

Abstract. *The issue of disinformation influence exerted by the Russian Federation on Ukrainian society remains relevant and critically important due to its large-scale social and political consequences. Since 2014, the fight against disinformation has become a national priority. With the onset of the full-scale invasion in February 2022, the nature of disinformation messages has changed significantly: the main focus has shifted to topics related to chemical, biological, and nuclear weapons. In this regard, the main goal of this study is to create a database of disinformation texts related to these topics, as well as to develop software for the automated detection of such materials in the information space.*

Keywords: *weapons of mass destruction, disinformation, countering disinformation, semantic analysis, information warfare.*

Гнатюк Сергій Олександрович, д.т.н., професор, проректор з наукових досліджень та трансферу технологій Державного університету «Київський авіаційний інститут».

Sergiy Gnatyuk, Doctor of Technical Sciences, Professor, Vice-Rector for Research and Technology Transfer at State University "Kyiv Aviation Institute."

Поліщук Юлія Ярославівна, молодший науковий співробітник науково-дослідної лабораторії протидії кіберзагрозам в авіаційній галузі Державного університету «Київський авіаційний інститут».

Yulia Polishchuk, Junior Researcher at the Research Laboratory for Countering Cyber Threats in the Aviationat State University "Kyiv Aviation Institute."