https://doi.org/10.31891/2307-5732-2025-351-34
УДК 004.8-9

ЛОЗИНСЬКА ОЛЬГА
Національний університет «Львівська політехніка»
https://orcid.org/0000-0002-5079-0544
e-mail: olha.v.lozynska@lpnu.ua
ВИСОЦЬКА ВІКТОРІЯ
Національний університет «Львівська політехніка»
https://orcid.org/0000-0001-6417-3689
e-mail: victoria.a.vysotska@lpnu.ua
МАРКІВ ОКСАНА
Національний університет «Львівська політехніка»
https://orcid.org/0000-0002-1691-1357
e-mail: oksana.o.markiv@lpnu.ua
ДАНИЛИК ВІТАЛІЙ
Національний університет «Львівська політехніка»
https://orcid.org/0000-0001-5928-7235
e-mail: vitalii.m.danylyk@lpnu.ua
КУЛІКОВ ЮРІЙ
Національний університет «Львівська політехніка»
https://orcid.org/0009-0008-6680-7313
e-mail: yurii.kulikov.sa.2020@lpnu.ua

СИСТЕМА АВТОМАТИЧНОГО ВИЯВЛЕННЯ УКРАЇНОМОВНОЇ ДЕЗІНФОРМАЦІЇ НА ОСНОВІ МАШИННОГО НАВЧАННЯ

Розроблення інструментів для ідентифікації та аналізу інформаційних загроз є актуальним завданням, що має важливе значення для забезпечення інформаційної безпеки України, особливо у теперішній час. Дослідження та аналіз методів та комплексних інструментів для ідентифікації фейків і дезінформації, не лише відповідає критичній потребі українського суспільства в надійних засобах верифікації інформації, але й підвищує загальну культуру інформаційного споживання, зміцнюючи інформаційну стійкість нації.

У статті проведено аналіз існуючих підходів та інструментів для ідентифікації, оцінки та протидії дезінформації, фейковим новинам і пропаганді в українському інформаційному просторі. Розроблено прототип системи текстового аналізу, здатної потенційно виявляти дезінформацію з використанням векторизації на основі TF-IDF та мультиноміального наївного класифікатора Байєса. Наукова новизна полягає у застосуванні цих методів машинного навчання до україномовного контенту, а також використання FAISS для швидкого пошуку найближчих сусідів та кластеризації у векторному просторі. Наведено результати порівняння оцінок точності моделі для правдивого та неправдивого контенту. Це дає змогу визначити для якого типу текстів модель працює краще.

Подальше дослідження буде спрямоване на навчання і тренування моделі на нових даних, тестування та оцінювання запропонованої системи, а також використання даної системи для автоматизованого моніторингу новин та у соціальних медіа для ідентифікації потенційно фальсифікованої інформації.

Ключові слова: джерела дезінформації, інформаційна безпека, аналіз фейків, машинне навчання, обробка природної мови, точність.

LOZYNSKA OLGA, VYSOTSKA VICTORIA, MARKIV OKSANA, DANYLYK VITALII, KULIKOV YURII
Lviv Polytechnic National University

AUTOMATIC DETECTION SYSTEM OF UKRAINIAN-LANGUAGE DISINFORMATION BASED ON MACHINE LEARNING

The development of tools for identifying and analyzing information threats is an urgent task that is of great importance for ensuring the information security of Ukraine, especially at the present time. Research and analysis of methods and complex tools for identifying fakes and disinformation not only meets the critical need of Ukrainian society for reliable means of information verification, but also improves the general culture of information consumption, strengthening the information resilience of the nation.

The article analyzes existing approaches and tools for identifying, assessing and countering disinformation, fake news and propaganda in the Ukrainian information space. A prototype of a text analysis system capable of potentially detecting manipulative fragments has been developed. The capabilities of the applied machine learning technologies for creating an appropriate model of the system, as well as the functional capabilities of the developed system, are described.

The article analyzes existing approaches and tools for identifying, assessing and countering disinformation, fake news and propaganda in the Ukrainian information space. A prototype of a text analysis system capable of potentially detecting disinformation using vectorization based on TF-IDF and a multinomial naive Bayes classifier has been developed. The scientific novelty lies in the application of these machine learning methods to Ukrainian-language content, as well as the use of FAISS for fast search of nearest neighbors and clustering in vector space. The results of comparing the model accuracy estimates for true and false content are presented. This allows us to determine for which type of texts the model works better.

Further research will be aimed at learning and training the model on new data, testing and evaluating the proposed system, as well as using this system for automated monitoring of news and in social media to identify potentially falsified information.

Keywords: sources of disinformation, information security, fake news analysis, machine learning, natural language processing, accuracy.

Стаття надійшла до редакції / Received 11.04.2025
Прийнята до друку / Accepted 26.04.2025

Постановка проблеми у загальному вигляді
та її зв'язок із важливими науковими чи практичними завданнями

У цифрову епоху інформаційна безпека є однією з ключових викликів для багатьох суспільств, особливо для країн, які перебувають у стані політичних змін чи конфліктів. Україна, як країна, що зазнає значного впливу інформаційних операцій, стикається з потребою в боротьбі з дезінформацією, фейковими новинами та пропагандою. Відповідно до цього, розроблення інструментів для ідентифікації та аналізу таких інформаційних загроз є актуальним завданням, що має важливе значення для забезпечення інформаційної безпеки країни.

Актуальність даного дослідження не може бути переоцінена в умовах сучасного інформаційного простору, де боротьба за правду стає майже синонімом збереження національної безпеки. Інформаційні війни, в яких істина стає першою жертвою, нещадно бомбардують громадську свідомість мільйонів людей, спотворюючи реальність та формуючи штучну реальність, що служить інтересам зовнішніх і внутрішніх антагоністів.

В Україні, на передовій протистояння з гібридними загрозами, відсутність надійних інструментів для розпізнавання дезінформації може призвести до системних збоїв у суспільній довірі, ерозії фундаментальних демократичних цінностей і стійкості державних інститутів. Це не просто питання медійної грамотності – це питання стратегічної оборони національної безпеки.

Дослідження та аналіз методів та комплексних інструментів для ідентифікації фейків і дезінформації, не лише відповідає критичній потребі українського суспільства в надійних засобах верифікації інформації, але й підвищує загальну культуру інформаційного споживання, зміцнюючи інформаційну стійкість нації.

Аналіз досліджень та публікацій

Для створення системи автоматичної ідентифікації та аналізу фейків, пропаганди та дезінформації, існує кілька існуючих платформ і проектів.

NewsGuard є інноваційним проектом, який зосереджений на підвищенні прозорості та надійності інформації в інтернеті, зокрема на новинних сайтах [1]. Сервіс працює шляхом надання рейтингів надійності різним медіа-ресурсам, що дозволяє користувачам зрозуміти, наскільки вірогідними та надійними є джерела їхнього новинного контенту.

NewsGuard використовує команду досвідчених журналістів, які аналізують та оцінюють новинні сайти згідно з визначеним набором журналістських критеріїв. На основі цих критеріїв сайту присвоюється рейтинг надійності, який стає доступним для користувачів через браузерні плагіни або через інші партнерські платформи. У науковій роботі [2] досліджено та проаналізовано базу даних NewsGuard, яка містить джерела інформації з таких країн як США, Франція, Італія, Німеччина та Канада. Понад 50% джерел, включених до бази даних, мають оцінку 60 балів або вище відповідно до номенклатури NewsGuard. Ці джерела зазвичай відповідають двом зваженим критеріям: «уникає оманливих заголовків» і «не публікує неправдивий або надзвичайно оманливий вміст».

Hoaxy – це платформа, розроблена Осередком дослідження мережі Індіана університету, яка спеціалізується на візуалізації динаміки поширення інформації та фактчекінгу в соціальних мережах [3]. Ця платформа допомагає дослідникам, журналістам та загалом публіці бачити, як поширюються певні твердження та новини онлайн, а також відстежувати зусилля щодо перевірки фактів.

Авторами статті представлено платформу Hoaxy для збору, виявлення та аналізу дезінформації в Інтернеті та пов'язані з нею заходи перевірки фактів [4], проведено попередній аналіз зразка загальнодоступних твітів, які містять як фейкові новини, так і перевірку фактів. Було досліджено, що поширення вмісту перевірки фактів зазвичай відстає від поширення дезінформації на 10–20 годин.

Hoaxy аналізує та візуалізує ланцюги поширення статей та заяв у соціальних мережах. Користувачі можуть ввести конкретні запити чи ключові слова для відстеження поширення пов'язаної інформації. Платформа використовує алгоритми для виявлення і зображення шляхів, якими інформація поширюється між користувачами та медійними вузлами. Це дає змогу зрозуміти, як інформація та потенційна дезінформація розповсюджується через мережі.

Factmata – це інноваційна платформа, що використовує штучний інтелект для виявлення та аналізу шкідливого контенту та дезінформації в інтернеті [5]. Ця система здатна визначати фейкові новини, пропаганду, упереджені висловлювання та інші види маніпулятивного контенту.

Factmata використовує складні алгоритми машинного навчання та обробки природної мови для аналізу текстів. Платформа аналізує великі обсяги інформації з різноманітних джерел, включаючи новини, соціальні мережі, та блоги. Це дозволяє виявляти потенційно шкідливий контент на ранніх етапах його поширення. Factmata також забезпечує зворотний зв'язок користувачів та експертів, що допомагає постійно удосконалювати та налаштовувати алгоритми платформи.

Загалом, Factmata становить суттєвий крок у розвитку інструментів для боротьби з дезінформацією, пропонуючи потужну платформу, яка може допомогти різним організаціям, від новинних агентств до корпоративних клієнтів, контролювати якість та надійність інформації.

Snopes є однією з перших і найбільш впізнаваних платформ для фактчекінгу в інтернеті [6]. Заснована в 1994 році, Snopes спеціалізується на перевірці міських легенд, популярних міфів, розповсюджених історій та дезінформації, які ширяться через Інтернет та соціальні медіа.

Платформа Snopes використовує дослідницький підхід, залучаючи команду досвідчених фактчекерів та журналістів, які аналізують та перевіряють достовірність інформації. Вони здійснюють глибоке розслідування кожного запиту, опираючись на джерела первинних даних, наукові дослідження, офіційні заяви та інші перевірені факти. Snopes також відомий своїми детальними поясненнями і висновками, які допомагають користувачам зрозуміти контекст та природу інформації [7].

Snopes продовжує бути цінним ресурсом у боротьбі з дезінформацією, надаючи користувачам інструмент для перевірки достовірності контенту, який вони споживають. Однак, зростаюча потреба в більш швидкому та автоматизованому фактчекінгу може вимагати від Snopes та подібних платформ адаптації до нових технологічних рішень.

Bellingcat – це незалежна міжнародна ініціатива, що спеціалізується на розслідуванні за допомогою відкритих даних (open-source intelligence, OSINT) [8]. Заснована у 2014 році, Bellingcat використовує дані з відкритих джерел, такі як соціальні медіа, супутникові знімки, відеоматеріали, фотографії, і навіть рекламні бази даних, для розслідування і верифікації інформації, часто пов'язаної з міжнародними конфліктами, геополітичними кризами, і порушенням прав людини.

Bellingcat використовує сучасні методи аналізу даних, включаючи геолокаційний аналіз, аналіз зображень та відео, а також дослідження соціальних мереж, для того, щоб виявляти і документувати факти, які часто залишаються поза увагою традиційних ЗМІ. Це дозволяє Bellingcat проводити розслідування з високим ступенем незалежності та об'єктивності.

У роботі [9] проаналізовано діяльність Bellingcat у соціальних мережах у Twitter. Досліджено, що на зростання Bellingcat у Twitter значною мірою впливають зовнішні події, такі як російське вторгнення в Україну. Авторами виявлено 13-кратне збільшення кількості ботів, які взаємодіють з обліковим записом Bellingcat.

Bellingcat залишається однією з лідируючих організацій у сфері розслідувань з використанням відкритих даних, що вносить значний вклад у розвиток методів розслідувальної журналістики та аналізу даних.

Порівняння відомих програмних продуктів для виявлення джерел дезінформації, їх переваги та недоліки подано у таблиці 1.

Таблиця 1

Порівняння відомих програмних продуктів для виявлення джерел дезінформації

Назва програмного продукту	Переваги	Недоліки
NewsGuard	1. Пряме залучення експертів. Використання професійних журналістів для оцінки новинних джерел додає вагомість та надійність їхнім рейтингам. Експерти з медіа мають необхідні знання та досвід для адекватної оцінки якості контенту. 2. Широке покриття. NewsGuard аналізує тисячі сайтів, що охоплює широкий спектр медіа-ландшафту, забезпечуючи користувачам інформацію про надійність великої кількості джерел.	1. Висока вартість. Процес аналізу, що залучає кваліфікованих журналістів, є ресурсомістким і може вимагати значних фінансових витрат. Це може обмежувати можливість швидкого розширення покриття нових сайтів або частоти оновлення рейтингів. 2. Потенційна суб'єктивність оцінок. Незважаючи на стандартизований підхід, особисті уподобання та судження журналістів можуть впливати на оцінку. Хоча NewsGuard і намагається мінімізувати суб'єктивність, повністю виключити її неможливо, що може викликати питання щодо об'єктивності та незалежності рейтингів.
Hoaxy	1. Аналіз реального часу. Hoaxy дає змогу користувачам бачити, як інформація поширюється в реальному часі, що є критично важливим для розуміння динаміки інформаційних кампаній. 2. Візуалізація поширення дезінформації. Чіткі, наочні графіки дозволяють легко ідентифікувати основні джерела і вектори поширення дезінформації, а також взаємодію між різними акторами у мережі.	1. Обмежена здатність виявляти дезінформацію. Хоча Hoaxy ефективний у візуалізації та аналізі шляхів поширення інформації, платформа не завжди може точно ідентифікувати, чи є контент дезінформацією. Це вимагає додаткової перевірки та критичного аналізу з боку користувачів. 2. Не враховує лінгвістичні особливості окремих мов. Hoaxy в основному орієнтований на англомовний контент і може не враховувати особливості інших мов, що обмежує його застосування в неангломовних регіонах або для мовних спільнот, що потребують специфічного підходу до аналізу інформації.

Factmata	1. Автоматизований підхід. Використання штучного інтелекту дозволяє автоматизувати процес виявлення шкідливого контенту, знижуючи потребу в ручній перевірці та аналізі великих обсягів даних. 2. Постійне навчання. Factmata постійно навчається на нових даних, що покращує її здатність виявляти різноманітні форми дезінформації та адаптуватися до змінюваних методів маніпуляції інформацією.	1. Високі ресурси для навчання. Розвиток та підтримка алгоритмів, які використовує Factmata, вимагає значних обчислювальних та фінансових ресурсів. Тренування штучного інтелекту на великих наборах даних може бути витратним та складним процесом. 2. Точність та помилки. Як і будь-яка система, що базується на штучному інтелекті, Factmata може мати проблеми з точністю, особливо коли йдеться про контент, що містить суб'єктивні судження, іронію або жаргон, що може ввести систему в оману.
Snopes	1. Велика довіра та авторитет серед користувачів. Завдяки своїй довгій історії та консистентності у перевірці фактів, Snopes здобув репутацію надійного джерела для розрізнення правдивої інформації від неправдивої. Їх робота забезпечила їм довіру та визнання серед користувачів по всьому світу.	1. Залежність від ручної перевірки інформації. Це може сповільнювати процес перевірки, особливо в умовах сучасного новинного циклу, де інформація змінюється дуже швидко. Такий підхід також може виявитись недостатньо ефективним у ситуаціях, коли потрібно оперативно реагувати на велику кількість фальшивих новин, що розповсюджуються в соціальних медіа.
Bellingcat	1. Висока точність та глибина аналізу. Завдяки міждисциплінарному підходу та використанню різноманітних джерел інформації, Bellingcat здатний проводити глибокі аналітичні розслідування. Вони відомі своєю здатністю розкривати складні мережі взаємозв'язків та події, що відбуваються за закритими дверима. 2. Передові методи збору та аналізу даних. Використання передових технологій та методів аналізу дозволяє Bellingcat ефективно використовувати великі обсяги відкритих даних для розкриття інформації, яка часто недоступна через традиційні джерела.	1. Фокус переважно на міжнародних подіях. Оскільки Bellingcat часто зосереджений на глобальних подіях і міжнародних конфліктах, їхня діяльність може бути менш релевантною для локальних або регіональних питань, які вимагають специфічних знань та ресурсів. 2. Потреба у спеціалізованих знаннях для інтерпретації даних. Робота Bellingcat вимагає високого рівня експертизи в областях, таких як аналіз даних, супутникова картографія, верифікація відео та зображень. Це створює бар'єри для залучення нових членів команди та потребує постійного професійного розвитку.

Наукові роботи, пов’язані з виявленням джерел дезінформації часто фокусуються на використанні технологій машинного навчання, зокрема, методів глибокого навчання для аналізу текстових даних [10]. Наприклад, алгоритми, які базуються на моделях обробки природної мови (NLP) [11], дають змогу виявляти зразки та невідповідності в тексті, що можуть вказувати на дезінформацію. Ці методи включають використання трансформерів, таких як BERT [12], які можуть генерувати та розуміти тексти на дуже високому рівні.

Однак, крім технічних аспектів, багато досліджень підкреслюють важливість контекстуального аналізу. Виявлення дезінформації часто вимагає не лише аналізу тексту, але й залучення контекстуальної інформації, як-от визначення джерела інформації та його надійності. Роботи показують, що інтеграція даних з різних джерел та їх перехресний аналіз може значно підвищити точність виявлення фейкових новин.

Інший аспект, який часто обговорюється в наукових роботах, це розроблення алгоритмів для автоматизованого виявлення змін у поведінці користувачів соціальних мереж, що може бути індикатором маніпулятивних інформаційних кампаній. Вивчення патернів поширення інформації дає змогу ідентифікувати потенційно шкідливий контент до того, як він досягне великої аудиторії.

Враховуючи динаміку сучасних медіа, постійно зростає потреба в розробленні ефективних інструментів для моніторингу та аналізу даних в реальному часі. Багато робіт пропонують

використання комплексних систем, які інтегрують штучний інтелект, машинне навчання та автоматизовані інструменти моніторингу для посилення здатності швидко реагувати на інформаційні загрози.

Варто зазначити, що жоден інструмент або метод не може бути абсолютно ефективним без врахування етичних аспектів та можливих наслідків впровадження таких технологій. Критичний огляд сучасних досліджень показує, що постійний розвиток в області алгоритмічної прозорості, забезпечення приватності та захисту даних є ключовим для підтримки довіри серед користувачів.

Формулювання цілей статті

Метою роботи є: дослідження відомих методів та засобів ідентифікації, аналізу та реагування на дезінформацію, фейкові новини та пропаганду в українському інформаційному просторі, розроблення ефективної системи аналізу тексту з метою виявлення фрагментів, які можуть містити дезінформацію, опис функцій даної системи та використаних технологій для машинного навчання моделі.

Виклад основного матеріалу

Проблема неефективності відомих інструментів для ідентифікації, аналізу та реагування на дезінформацію, фейкові новини та пропаганду в українському інформаційному просторі складається з кількох складових, кожна з яких потребує особливої уваги та індивідуального підходу, а саме:

1. Багатомовність та локалізація. Багато сучасних інструментів і рішень для виявлення дезінформації оптимізовані для англійської мови, що залишає значний пробіл для інших мов, зокрема української. Це створює бар'єр, оскільки алгоритми, які ефективно працюють з англійською мовою, можуть не враховувати лінгвістичні особливості та контекст, які є критичними для правильного аналізу україномовного контенту. Відсутність локалізації та адаптації технологій може призводити до неправильної інтерпретації інформації або недооцінки загроз.

2. Складність виявлення дезінформації. Дезінформація часто використовує витончені методи маскування, включаючи змішування правдивої і неправдивої інформації, що ускладнює її виявлення традиційними методами фактчекінгу. Дезінформація може бути вбудована в логічно правдоподібні наративи, що вимагає глибшого контекстуального аналізу і залучення додаткових джерел для перевірки фактів.

3. Динаміка інформаційних кампаній. Інформаційні кампанії, особливо ті, що стосуються дезінформації, швидко змінюють свої форми та засоби розповсюдження, що вимагає від інструментів виявлення дезінформації гнучкості та здатності адаптуватися до нових викликів.

4. Обмежені ресурси для моніторингу. Великий обсяг інформації, який постійно генерується в інтернеті, робить неможливим ручний моніторинг усіх потенційних джерел дезінформації. Існуючі інструменти часто не в змозі або не ефективні у виявленні дезінформації в реальному часі через обмеження в обчислювальних та людських ресурсах.

Для вирішення викладених вище проблем, потрібно вирішити наступні задачі:

1. Розробка алгоритмів глибокого навчання для української мови. Необхідно створити спеціалізовані алгоритми, які використовують технології обробки природної мови для розпізнавання українського контенту. Ці алгоритми повинні вміти аналізувати не лише словесний контент, але й розуміти контекст, ідіоми, сленг та стилістичні особливості, характерні для української мови. Важливим аспектом буде використання навчальних даних, що включають достатньо прикладів фейків, для тренування моделей розпізнавати недостовірну інформацію.

2. Створення модульної платформи для моніторингу медіа. Ця платформа повинна мати здатність швидко відслідковувати розповсюдження потенційно небезпечних наративів, використовуючи алгоритми для визначення вірусності контенту та його впливу на аудиторію. Важливою функцією буде автоматичний аналіз та звітність, яка допоможе оперативно реагувати на критичні загрози.

3. Інтеграція з медіа-платформами та соціальними мережами. Необхідно розробити API або плагіни для інтеграції системи з популярними медіа-платформами та соціальними мережами. Це забезпечить можливість раннього виявлення та блокування дезінформації на початкових стадіях її поширення. Інтеграція повинна дозволяти аналізувати великі потоки даних в реальному часі, виявляючи патерни поширення дезінформації.

4. Розробка освітніх інструментів для користувачів. Щоб зміцнити інформаційну грамотність суспільства, важливо розробити інструменти, які допоможуть людям критично оцінювати інформацію. Це може включати розробку навчальних курсів, інтерактивних воркшопів, та інших ресурсів, що навчають користувачів розпізнавати ознаки фейків та маніпулятивних технік.

За допомогою цих заходів можна створити ефективну систему, здатну адекватно реагувати на сучасні виклики інформаційного простору, забезпечуючи надійний захист від інформаційних загроз в українському контексті.

Для вирішення цих задач було розроблено прототип системи, яка змогу виявляти дезінформацію у текстових документах. Дезінформація може бути виявлена у різних формах, таких як неправдиві новини, маніпулятивні статті або навіть у вигляді спотвореної інформації, яка вводить в оману читачів.

Функції системи:

1. Виявлення дезінформації в тексті. Основна функція системи полягає в аналізі тексту з метою виявлення фрагментів, які можуть містити дезінформацію. Це може включати в себе тексти, що поширюють неправдиву або спотворену інформацію, маніпулюють фактами або створюють оманливе враження.

2. Вибір моделі та параметрів. Система розроблена з урахуванням можливості аналізу різних форматів текстових документів. Вона підтримує такі формати, як CSV, що дозволяє зручно завантажувати та аналізувати текстові дані. Формат CSV є одним з найпоширеніших форматів для зберігання табличних даних, що робить його універсальним для обробки програмним забезпеченням.

3. Навчання моделі. Ця функція системи дає змогу створювати модель машинного навчання, яка може "вчитися" на основі попередньо зібраних даних. Процес навчання моделі полягає у тому, щоб система автоматично аналізувала набір даних, виявляла в ньому закономірності та встановлювала зв'язки між різними характеристиками тексту, які можуть вказувати на наявність дезінформації. Для цього було створено власний датасет, орієнтований на українську мову, що включає різноманітні приклади дезінформації та правдивих текстів [13]. Цей датасет містить тексти з різних джерел, включаючи новини, соціальні мережі, блоги та інші публікації. Збір даних проводився з метою створення максимально точного та репрезентативного набору для навчання моделі.

Серед методів машинного навчання використовувалися векторизація на основі TF-IDF та навчання на основі мультиноміального наївного класифікатора Байєса (Multinomial NB) [14]. Цей класифікатор бере до уваги середнє значення кожної ознаки для кожного класу, а також включає коефіцієнт згладжування alpha.

Алгоритм роботи системи наведено на рис. 1.

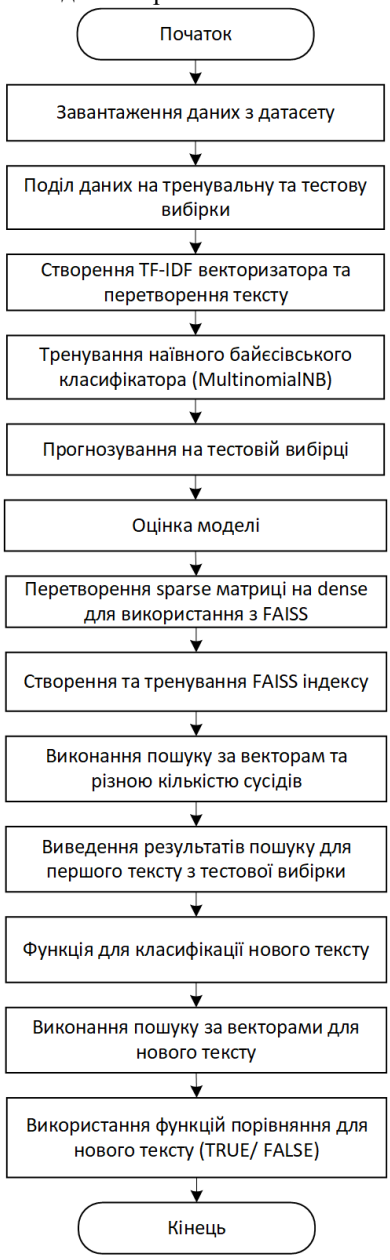

Рис. 1. Алгоритм роботи системи з використанням TF-IDF та мультиноміального наївного класифікатора Байєса

На одному з етапів відбувається використання FAISS (Facebook AI Research Similarity Search), що дає змогу отримати кращі результати класифікації для україномовного контенту.

4. Відображення результатів. Ця функція системи забезпечує зручний інтерфейс для відображення результатів аналізу користувачеві через командний рядок. Програма виводить результати, включаючи прогнозовану мітку (правдива інформація або дезінформація) та ймовірності для кожного класу, що спрощує процес інтерпретації і використання інформації. Крім того, програма може генерувати звіти у форматі CSV, що дозволяє користувачам зберігати та аналізувати результати в подальшому.

Аналіз тексту. Після того, як користувач ввів текст або завантажив файл, він може запустити аналіз, натиснувши відповідну кнопку в командному інтерфейсі. Це ініціює процес обробки, під час якого система починає аналізувати текст. Спочатку текст перетворюється у вектори, які слугують вхідними даними для моделей машинного навчання. Векторизація допомагає представити текст у числовому форматі, що дозволяє програмі ефективно його обробляти і виявляти характерні шаблони [15].

Під час векторизації система може враховувати додаткові опції, якщо їх обрав користувач. Наприклад, якщо було вибрано опції "References" або "Sources", ці параметри додаються до векторів, і їх ураховують під час аналізу. Таким чином, система здатна обробляти різні види інформації і адаптуватися до додаткових вимог.

Після того як система перетворює текст у вектори за допомогою процесу векторизації, ці вектори передаються в модель машинного навчання для подальшого аналізу. Модель використовує ці вектори, щоб зробити прогноз щодо ймовірності того, що текст містить дезінформацію.

Модель машинного навчання, яка використовується в системі, була попередньо навчена на великому наборі даних, що включає тексти, які містять як правдиву інформацію, так і дезінформацію. Завдяки цьому вона здатна розпізнавати характерні шаблони, що зустрічаються у текстах, які містять дезінформацію. Якщо система виявляє такі шаблони або особливості в аналізованому тексті, це впливає на результати аналізу, збільшуючи ймовірність того, що текст містить дезінформацію.

Система здійснює аналіз цих текстів, і видає результати, які показують, чи містить кожен текст дезінформацію. Таким чином, процес запуску аналізу тексту є центральним етапом, який включає перетворення тексту в зручний для аналізу формат і запуск машинного навчання для визначення його ймовірного походження.

Аналіз результатів. Коли аналіз тексту завершено, програма відображає результати у спеціальному полі для виводу, де користувач може легко їх переглянути й інтерпретувати. Це поле зазвичай знаходиться в командному інтерфейсі або відображається у вигляді текстового виводу після завершення аналізу.

У полі для виводу користувач отримує інформацію про відсоток тексту, який, ймовірно, містить дезінформацію. Цей відсоток розраховується на основі аналізу, проведеного моделлю машинного навчання. Модель виявляє характерні ознаки, що можуть свідчити про наявність дезінформації. Відсоток може змінюватися залежно від складності та особливостей тексту, а також від використовуваної моделі.

На рис. 2 та рис. 3 подано результати класифікації для текстів 1-4 з використанням моделей на основі TF-IDF та мультиноміального наївного класифікатора Байєса. Текст 1 та текст 2 система класифікувала за міткою “0”, що означає правдиву інформацію. Ймовірності прогнозу склали 69.09% на користь правдивої інформації та 30.91% на користь дезінформації для тексту 1 і 74.56% (правда) та 25.44% (неправда) для тексту 2 відповідно. Найближчі сусіди вектору цих текстів вказали на те, що вони також були класифіковані як правдива інформація.

```
true_text1 = "В Україні прийняли новий закон, який дозволяє громадянам зберігати зброю вдома для самозахисту."
classify_new_text(true_text1)
true_text2 = "Міністерство охорони здоров'я України повідомило про зростання кількості вакцинацій проти коронавірусу."
classify_new_text(true_text2)
false_text3 = "В Україні введено комендантську годину для всіх громадян з 20:00 до 06:00."
classify_new_text(false_text3)
false_text4= "Україна стала членом Європейського Союзу з правом голосу у всіх рішеннях."
classify_new_text(false_text4)
```

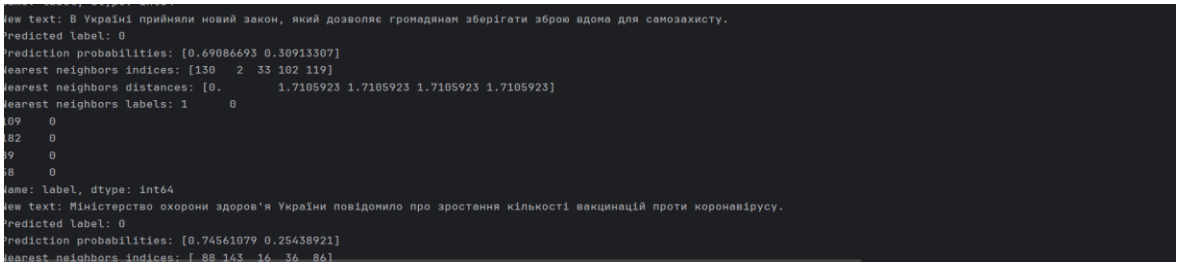

Рис. 2. Результати моделі на основі TF-IDF та мультиноміального наївного класифікатора Байєса для україномовних новин (для тексту 1 та 2)

Текст 3 та текст 4 отримали мітку “1”, що означає дані тексти містять дезінформацію. Ймовірності прогнозу для тексту 3 склали 47.11% на користь правдивої інформації та 52.89% на користь дезінформації та 43.83% (правда) і 56.17% (неправда) для тексту 4 відповідно. Найближчі сусіди вектору цього тексту підтвердили класифікацію як дезінформація.

Цей функціонал допомагає користувачеві розпізнати та класифікувати тексти як правдиві або дезінформацію, що є важливим для подальшого аналізу та прийняття рішень. Система надає детальну інформацію про ймовірності прогнозу та найближчі сусіди вектору, що допомагає користувачеві краще зрозуміти основи класифікації.

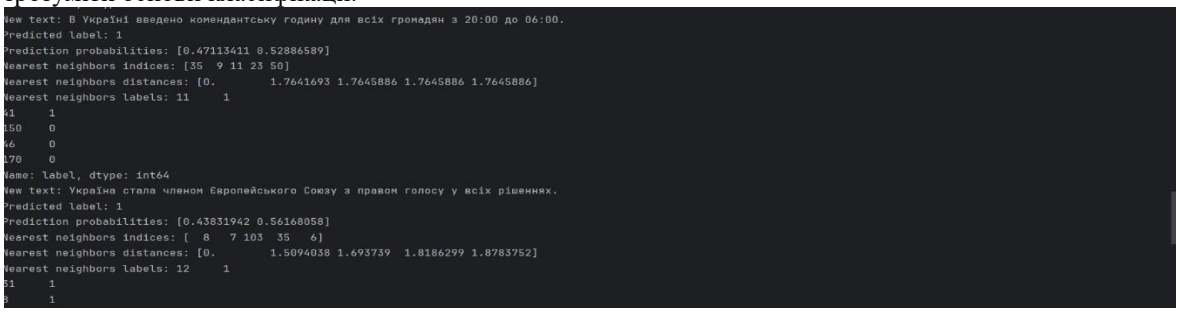

Рис. 3. Результати моделі на основі TF-IDF та мультиноміального наївного класифікатора Байєса для україномовних новин (для тексту 3 та 4)

Щоб полегшити візуальне розпізнавання, система може підкреслювати червоним кольором текстові фрагменти, які визначені як дезінформація. Це надає користувачеві наочний спосіб побачити, які частини тексту, найімовірніше, містять неправдиву або маніпулятивну інформацію. Таке підкреслення допомагає швидко ідентифікувати підозрілі ділянки тексту.

Таким чином, на етапі відображення результатів користувач отримує повну картину щодо аналізованого тексту. Система надає відсоток правдивості чи неправдивості тексту, а також рекомендації щодо подальших дій. Ця інформація дає змогу користувачеві швидко та ефективно визначити, наскільки текст може бути підозрілим, і прийняти відповідні рішення щодо його автентичності або необхідності подальшого аналізу.

Висновки з даного дослідження і перспективи подальших розвідок у даному напрямі

У підсумку, створення системи для надійного виявлення дезінформації у текстах є значним досягненням у галузі обробки природної мови та машинного навчання. Завдяки використанню передових алгоритмів та великих обсягів даних для тренування, система продемонструвала високу точність у розпізнаванні текстів з правдивою інформацією та дезінформацією, зменшуючи ризик хибних висновків.

Модель успішно визначає характерні ознаки, притаманні текстам, що містять дезінформацію, забезпечуючи точні результати аналізу. Це є важливим інструментом для користувачів, які прагнуть відрізнити правдиву інформацію від маніпулятивних або неправдивих текстів. В процесі роботи системи тексти отримують мітки "0" (правдива інформація) або "1" (неправдива інформація). Система дає змогу аналізувати тексти різного походження, виявляти підозрілі фрагменти та надавати рекомендації щодо їхньої достовірності. Крім того, система надає детальну інформацію про ймовірності прогнозу та найближчі сусіди вектору, що допомагає користувачеві краще зрозуміти основи класифікації.

Розроблена система може стати цінним інструментом у багатьох галузях, де важливо відрізнити дезінформацію від правдивої інформації. Вона також підкреслює потенціал машинного навчання у розв'язанні складних проблем виявлення дезінформації та відкриває нові перспективи для подальших досліджень у цій сфері. Система може бути використана в журналістиці, освіті, урядових та неурядових організаціях, що займаються боротьбою з фейковими новинами та маніпуляціями у медіа.

Завдяки своїй функціональності та точності, розроблена система забезпечує користувачам ефективний інструмент для аналізу текстів, сприяючи підвищенню рівня інформаційної безпеки та медіаграмотності у суспільстві.

Подяка

Дана стаття підготована завдяки грантової підтримки Національного Фонду Досліджень України, реєстраційний номер проєкту 33/0012 від 3/03/2025 (2023.04/0012) «Розроблення інформаційної системи автоматичного виявлення джерел дезінформації та неавтентичної поведінки користувачів чатів» за конкурсом «Наука для зміцнення обороноздатності України».

Література

1. NewsGuard: Global Leader in Information Reliability [Електронний ресурс]. – Режим доступу: https://www.newsguardtech.com/ – (Дата звернення 20.02.2025 р.). – Назва з екрана.

2. Lühring J., Metzler H., Lazzaroni R., Shetty A., Lasser J. Best practices for source-based research on misinformation and news trustworthiness using NewsGuard / J. Lühring, H. Metzler, R. Lazzaroni, A. Shetty, J. Lasser // Journal of Quantitative Description: Digital Media. – 2025. – Vol. 5. – P. 1-53. https://doi.org/10.51685/jqd.2025.003

3. Hoaxy [Електронний ресурс]. – Режим доступу: https://hoaxy.osome.iu.edu/ – (Дата звернення 20.02.2025 р.). – Назва з екрана.

4. Shao C., Ciampaglia G.L., Flammini A., Menczer F. Hoaxy: A Platform for Tracking Online Misinformation / C. Shao, G.L. Ciampaglia, A. Flammini, F. Menczer. // WWW '16 Companion: Proceedings of the 25th International Conference Companion on World Wide Web. – 2016. – P. 745 – 750. https://doi.org/10.1145/2872518.2890098.

5. FactMata [Електронний ресурс]. – Режим доступу: https://geniusee.com/portfolio/geniusee/factmatahttps://hoaxy.osome.iu.edu/ – (Дата звернення 20.02.2025 р.). – Назва з екрана.

6. Snopes [Електронний ресурс]. – Режим доступу: https://www.snopes.com/ – (Дата звернення 20.02.2025 р.). – Назва з екрана.

7. Alzurkani R. Fake News Detection Using Snopes and Politifact Data / R. Alzurkani // Researchgate. – 2023. https://www.researchgate.net/publication/370268556_Fake_News_Detection_Using_Snopes_and_Politifact_Data

8. Bellingcat [Електронний ресурс]. – Режим доступу: https://www.bellingcat.com/ – (Дата звернення 20.02.2025 р.). – Назва з екрана.

9. Bär D., Calderon F., Lawlor M., Lickleder S., Totzauer M., Feuerriegel S. Analyzing Social Media Activities at Bellingcat / D. Bär, F. Calderon, M. Lawlor, S. Lickleder, M. Totzauer, S. Feuerriegel // arXiv:2209.10271. – 2022. https://doi.org/10.1145/3578503.3583604.

10. Capuano N., Fenza G., Loia V., Nota F. D. Content-based fake news detection with machine and deep learning: A systematic review / N. Capuano, G. Fenza, V. Loia, F. D. Nota // Neurocomputing. – 2023. – Vol. 530. – P. 91–103.

11. Oshikawa R., Qian J., Wang W. Y. A survey on natural language processing for fake news detection / R. Oshikawa, J. Qian, W. Y. Wang // arXiv:1811.00770. – 2018. https://doi.org/10.48550/arXiv.1811.00770.

12. Kozik R., Kula S., Choras M., Wozniak M. Technical solution to counter potential crime: Text analysis to detect fake news and disinformation / R. Kozik, S. Kula, M. Choras, M. Wozniak // Journal of Computational Science. – 2022. – Vol. 60. – 8 p. https://doi.org/10.1016/j.jocs.2022.101576.

13. Лозинська О., Марків О., Висоцька В., Романчук Р., Назаркевич М. Інформаційна технологія розроблення та наповнення датасету дезінформації з використанням інтелектуального пошуку дипфейків та клікбейтів / О. Лозинська, О. Марків, В. Висоцька, Р. Романчук, М. Назаркевич // Herald of Khmelnytskyi National University. Technical Sciences – 2024. – 343(6(1)). – С. 158-167. https://doi.org/10.31891/2307-5732-2024-343-6-24.

14. Xu S., Li Y., Zheng W. Bayesian Multinomial Naïve Bayes Classifier to Text Classification / S. Xu, Y. Li, W. Zheng // Lecture Notes in Electrical Engineering. – 2017. https://doi.org/10.1007/978-981-10-5041-1_57.

15. López W., Merlino J., Rodríguez-Bocca P. Learning semantic information from internet domain names using word embeddings / W. López, J. Merlino, P. Rodríguez-Bocca// Engineering Applications of Artificial Intelligence. – 2020. – Vol. 94. – P. 103823. http://dx.doi.org/10.1016/j.engappai.2020.103823.

References

1. NewsGuard: Global Leader in Information Reliability [Електронний ресурс]. – Режим доступу: https://www.newsguardtech.com/ – (Дата звернення 20.02.2025 р.). – Назва з екрана.

2. Lühring J., Metzler H., Lazzaroni R., Shetty A., Lasser J. Best practices for source-based research on misinformation and news trustworthiness using NewsGuard / J. Lühring, H. Metzler, R. Lazzaroni, A. Shetty, J. Lasser // Journal of Quantitative Description: Digital Media. – 2025. – Vol. 5. – P. 1-53. https://doi.org/10.51685/jqd.2025.003

3. Hoaxy [Електронний ресурс]. – Режим доступу: https://hoaxy.osome.iu.edu/ – (Дата звернення 20.02.2025 р.). – Назва з екрана.

4. Shao C., Ciampaglia G.L., Flammini A., Menczer F. Hoaxy: A Platform for Tracking Online Misinformation / C. Shao,

G.L. Ciampaglia, A. Flammini, F. Menczer. // WWW '16 Companion: Proceedings of the 25th International Conference Companion on World Wide Web. – 2016. – P. 745 – 750. https://doi.org/10.1145/2872518.2890098.

5. FactMata [Електронний ресурс]. – Режим доступу: https://geniusee.com/portfolio/geniusee/factmatahttps://hoaxy.osome.iu.edu/ – (Дата звернення 20.02.2025 р.). – Назва з екрана.

6. Snopes [Електронний ресурс]. – Режим доступу: https://www.snopes.com/ – (Дата звернення 20.02.2025 р.). – Назва з екрана.

7. Alzurkani R. Fake News Detection Using Snopes and Politifact Data / R. Alzurkani // Researchgate. – 2023. https://www.researchgate.net/publication/370268556_Fake_News_Detection_Using_Snopes_and_Politifact_Data

8. Bellingcat [Електронний ресурс]. – Режим доступу: https://www.bellingcat.com/ – (Дата звернення 20.02.2025 р.). – Назва з екрана.

9. Bär D., Calderon F., Lawlor M., Lickleder S., Totzauer M., Feuerriegel S. Analyzing Social Media Activities at Bellingcat / D. Bär, F. Calderon, M. Lawlor, S. Lickleder, M. Totzauer, S. Feuerriegel // arXiv:2209.10271. – 2022. https://doi.org/10.1145/3578503.3583604.

10. Capuano N., Fenza G., Loia V., Nota F. D. Content-based fake news detection with machine and deep learning: A systematic review / N. Capuano, G. Fenza, V. Loia, F. D. Nota // Neurocomputing. – 2023. – Vol. 530. – P. 91–103.

11. Oshikawa R., Qian J., Wang W. Y. A survey on natural language processing for fake news detection / R. Oshikawa, J. Qian, W. Y. Wang // arXiv:1811.00770. – 2018. https://doi.org/10.48550/arXiv.1811.00770.

12. Kozik R., Kula S., Choras M., Wozniak M. Technical solution to counter potential crime: Text analysis to detect fake news and disinformation / R. Kozik, S. Kula, M. Choras, M. Wozniak // Journal of Computational Science. – 2022. – Vol. 60. – 8 p. https://doi.org/10.1016/j.jocs.2022.101576.

13. Lozynska O., Markiv O., Vysotska V., Romanchuk R., Nazarkevych M. Informatsiina tekhnolohiia rozroblennia ta napovnennia datasetu dezinformatsii z vykorystanniam intelektualnoho poshuku dypfeikiv ta klikbeitiv / O. Lozynska, O. Markiv, V. Vysotska, R. Romanchuk, M. Nazarkevych // Herald of Khmelnytskyi National University. Technical Sciences. – 2024. – 343(6(1)). – S. 158-167. https://doi.org/10.31891/2307-5732-2024-343-6-24.

14. Xu S., Li Y., Zheng W. Bayesian Multinomial Naïve Bayes Classifier to Text Classification / S. Xu, Y. Li, W. Zheng // Lecture Notes in Electrical Engineering. – 2017. https://doi.org/10.1007/978-981-10-5041-1_57.

15. López W., Merlino J., Rodríguez-Bocca P. Learning semantic information from internet domain names using word embeddings / W. López, J. Merlino, P. Rodríguez-Bocca// Engineering Applications of Artificial Intelligence. – 2020. – Vol. 94. – P. 103823. http://dx.doi.org/10.1016/j.engappai.2020.103823.