Національний аерокосмічний університет
“Харківський авіаційний інститут”
ім.М.Є.Жуковського
Міністерство освіти і науки України

Кваліфікаційна наукова
праця на правах рукопису

ПАДАЛКО ГАЛИНА АНАТОЛІЇВНА

УДК 004.8:004.912:316.77

ДИСЕРТАЦІЯ

МОДЕЛІ, МЕТОДИ ТА ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ВИЯВЛЕННЯ І АНАЛІЗУ ТЕКСТОВОЇ ДЕЗІНФОРМАЦІЇ ТА ПРОПАГАНДИ В СОЦІАЛЬНИХ МЕРЕЖАХ

122 Комп’ютерні науки
Інформаційні технології

Подається на здобуття наукового ступеня доктора філософії

Дисертація містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело.
_____Падалко Галина Анатоліївна

Науковий керівник
Яковлев Сергій Всеволодович,
доктор фіз.-мат. наук, професор

Харків – 2025

АНОТАЦІЯ

Падалко Галина Анатоліївна. Моделі, методи та інформаційна технологія виявлення і аналізу текстової дезінформації та пропаганди в соціальних мережах. – Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття наукового ступеня доктора філософії за спеціальністю 122 – Комп'ютерні науки. – Національний аерокосмічний університет "Харківський авіаційний інститут", Харків, 2025.

Дисертаційна робота присвячена розробці методів і моделей для виявлення та аналізу текстової дезінформації та пропаганди в соціальних мережах. У роботі досліджено сучасні підходи до виявлення інформаційних маніпуляцій, застосування технологій машинного навчання та глибокого навчання для автоматичного аналізу контенту, а також запропоновано нові методи для підвищення точності класифікації та інтерпретації текстових даних

У роботі детально проаналізовано теоретичні засади інформаційних маніпуляцій у контексті інформаційних воєн, основні типи інформаційних розладів та неконвенційних воєн, застосування технологій штучного інтелекту для протидії дезінформації, методи машинного та глибокого навчання для її аналізу, а також інформаційні технології, що використовуються для виявлення та протидії пропаганді.

З урахуванням проведеного аналізу в роботі поставлено та вирішено наукове завдання розробки технології дослідження текстів для аналізу маніпуляцій інформацією.

Обґрунтована методика проведення досліджень, що використовується в дослідженні. Робота базується на теоріях обчислювального штучного інтелекту, а саме – машинного навчання та глибокого навчання що дозволяють класифікувати та кластеризувати великі набори даних.

Вперше запропоновано фреймворк комплексного аналізу текстової дезінформації та пропаганди, заснований на комбінуванні технологій двоспрямованого кодувального представлення з трансформерів (BERT) та

двоспрямованої довгої короткочасної пам’яті (Bi-LSTM) для створення інтерпретованих векторних представлень тексту, який на відміну від існуючих технік, об’єднує виявлення та тематичний аналіз дезінформації, що дозволяє забезпечити ефективне виявлення та аналіз текстової дезінформації та пропаганди в соціальних мережах.

Вперше розроблено модель класифікації текстової дезінформації та пропаганди, заснований на поєднанні трансформерів і архітектури Bi-LSTM для інтеграції лінійних та нелінійних залежностей текстових даних, яка на відміну від існуючих використовує глибокі контекстуальні ембедінги і механізми уваги, що дозволяє підвищити точність моделі.

Удосконалено моделі глибокого навчання для класифікації текстової дезінформації та пропаганди, засновані на архітектурах XLNet, Bi-LSTM та Attention-Based Bi-LSTM, які на відміну від існуючих застосовують адаптивне оцінювання важливих фрагментів тексту, що дозволяє забезпечувати високий рівень продуктивності класифікації.

Удосконалено метод тематичного моделювання, заснований на застосуванні механізмів багатоголової уваги, що на відміну від існуючих використовує глибокі ембедінги для контекстуалізації даних, що дозволяє забезпечити чітке розмежування між кластерами.

Дістали подальшого розвитку моделі для класифікації текстової дезінформації та пропаганди, засновані на ансамблевих методах машинного навчання, які на відміну від існуючих враховують балансування класів і адаптуються до змін вхідних даних, що дозволяє підвищити продуктивність моделей.

Дістали подальшого розвитку моделі для класифікації інформації, засновані на статистичних методах машинного навчання, які на відміну від існуючих адаптовані до класифікації текстової дезінформації та пропаганди, що дозволяє створення систем для виявлення місінформації, дезінформації та малінформації.

Розроблені моделі, методи та програмне забезпечення використані у Balsillie School of International Affairs (акт впровадження від 18 березня 2025), Center for International Governance Innovation (акт впровадження від 18 березня 2025), Державній установі "Харківський обласний центр контролю та профілактики хвороб при Міністерстві охорони здоров'я" (акт впровадження від 2 грудня 2024), а також у освітньому процесі (акт впровадження від 30 січня 2025) та науковій діяльності (акт впровадження від 4 березня 2025) Національного аерокосмічного університету "Харківський авіаційний інститут".

Ключові слова: дезінформація, пропаганда, фейкові новини, аналіз соціальних мереж, машинне навчання, глибоке навчання, нейронні мережі, класифікація, кластеризація, моделювання, дані, загрози, інформаційна інтелектуальна система, штучний інтелект, система штучного інтелекту.

Список публікацій здобувача за темою дисертації

1. H. Padalko, V. Chomko, S. Yakovlev, and D. Chumachenko, "A Novel Comprehensive Framework for Detecting and Understanding Health-Related Misinformation," Information, vol. 16, no. 3, p. 175, Mar. 2025, doi: https://doi.org/10.3390/info16030175. (Scopus, Q2)

2. H. Padalko, V. Chomko, D. Chumachenko, "A novel approach to fake news classification using LSTM-based deep learning models," Frontiers in big data, vol. 6, 2024, https://doi.org/10.3389/fdata.2023.1320800. (Scopus, Q2)

3. H. Padalko, V. Chomko, S. Yakovlev, D. Chumachenko, "Ensemble machine learning approaches for fake news classification," Radioelectronic and computer systems, no. 4, pp. 5–19, 2023, https://doi.org/10.32620/reks.2023.4.01. (Scopus, Q3, Категорія А)

4. H. Padalko, V. Chomko, S. Yakovlev, P. P. Morita, "Classification of disinformation in hybrid warfare: an application of XLNet during the Russia's war against Ukraine," Radioelectronic and computer systems, vol. 2024, no. 4, pp. 46–58, 2024, https://doi.org/10.32620/reks.2024.4.04. (Scopus, Q3, Категорія А)

5. M. Lotto, Th. Hanjahanja-Phiri, H. Padalko, A. Oetomo, Z.A. Butt, J. Boger, J. Millar, Th. Cruvinel, P.P. Morita “Ethical principles for infodemiology and infoveillance studies concerning infodemic management on social media,” Frontiers in Public Health, vol. 11, 1130079, 2023. https://doi.org/10.3389/fpubh.2023.1130079 (Scopus, Q1)

6. A. Fitz-Gerald, D. Chumachenko, H. Padalko, V. Ganesh, “Artificial Intelligence and New Threat Vectors: Using Scenario Planning for Trend Forecasting” Balsillie Papers, vol. 5, no. 6, 2023, https://doi.org/10.51644/bap56.

7. H. Padalko, V. Chomko, D. Chumachenko, “Misinformation Detection in Political News using BERT Model,” CEUR Workshop Proceedings, vol 3641, Proceedings of the 3rd International Workshop of IT-professionals on Artificial Intelligence (ProfIT AI 2023) 2023 (Ватерлу, Канада, 20-22 листопада 2023), 2023, pp. 117–127. https://doi.org/10.5281/zenodo.15014455 (Scopus)

8. H. Padalko, V. Chomko, D. Chumachenko, “The Impact of Stopwords Removal on Disinformation Detection in Ukrainian language during Russian-Ukrainian war,” CEUR Workshop Proceedings, vol. 3777, Proceedings of the 4th International Workshop of IT-professionals on Artificial Intelligence (ProfIT AI 2024) 2024 (Кембрідж, США, 25-27 вересня 2024), 2024, pp. 87–101. . https://doi.org/10.5281/zenodo.15014498 (Scopus)

9. H. Padalko, V. Chomko, D. Chumachenko, “COVID-19 Misinformation Detection on Weibo by Support Vector Classifier,” IEEE International Conference on Dependable Systems, Services and Technologies (DESSERT) (Афіни, Греція, 13-15 жовтня 2023), 197136, 2023, https://doi.org/10.1109/dessert61349.2023.10416460 (Scopus)

10. H. Padalko, P. Morita, “Covid-19 ‘infodemic’ management: russia’s key tool for influencing geopolitics,” Population Medicine, vol. 5 (Supplement):A1809, 17th World Congress on Public Health (Рим, Італія, 2-6 травня 2023), 2023, p. 512. doi: https://doi.org/10.18332/popmed/164626.

11. H. Padalko “Machine learning approach for classification of russian propaganda”, Матеріали V Міжнародній науково-практичній конференції IT-

професіоналів та аналітиків комп'ютерних систем "ProfIT Conference" (Харків, 28-30 червня 2023), 2023, С. 96-97.

12. Г.А. Падалко, С.В. Яковлев "Модель розповсюдження COVID-19 на основі аналізу соціальних мереж", Матеріали IV Міжнародній науково-практичній конференції IT-професіоналів та аналітиків комп'ютерних систем "ProfIT Conference" (Харків, 13-15 грудня 2021), 2021, С. 106.

13. Г.А. Падалко "Мультиагентний підхід як інструмент моделювання соціальних мереж", Матеріали III Міжнародній науково-практичній конференції IT-професіоналів та аналітиків комп'ютерних систем "ProfIT Conference" (Харків, 8-10 грудня 2020), 2020, С. 112.

14. R. Vysochanskyy, H. Padalko, A. Ollila, "Information Warfare in Canada: Countering AI-backed Misinformation, Disinformation and Malinformation (MDM) Surrounding the Russian Invasion of Ukraine," Defence and Security Foresight Group. AI, Ethics, and Warfare Series, 2023, 10 p. [Online]. Available: https://uwaterloo.ca/defence-security-foresight-group/sites/default/files/uploads/documents/vysochanskyy-et-al_information-warfare-in-canada.pdf

15. A. Fitz-Gerald, H. Padalko, "The Need for a Strategic Approach to Disinformation and AI-Driven Threats," The Royal United Services Institute (RUSI), 2024. [Online]. https://www.rusi.org/explore-our-research/publications/commentary/need-strategic-approach-disinformation-and-ai-driven-threats

16. H. Padalko, "Russian Disinformation about the US Election: AI Analysis of Narratives," Center for International Governance Innovation, 2025. [Online]. Available: https://www.cigionline.org/static/documents/DPH-paper-Padalko.pdf

ANNOTATION

Halyna Anatoliivna Padalko. Models, Methods, and Information Technology for Detecting and Analyzing Textual Disinformation and Propaganda in Social Networks – A Qualification Research Paper in Manuscript Form.

Dissertation for the degree of Doctor of Philosophy in the specialty 122 – Computer Science. – National Aerospace University "Kharkiv Aviation Institute", Kharkiv, 2025.

The dissertation is devoted to the development of methods and models for detecting and analyzing textual disinformation and propaganda in social networks. The study examines modern approaches to identifying information manipulation, applying machine learning and deep learning technologies for automated content analysis, and proposing new methods to enhance classification accuracy and text data interpretation.

The dissertation provides a detailed analysis of the theoretical foundations of information manipulation in the context of information wars, the main types of information disorders and unconventional warfare, the application of artificial intelligence technologies to counter disinformation, machine learning and deep learning methods for its analysis, as well as information technologies used for detecting and combating propaganda.

Based on the conducted analysis, the scientific task of developing a technology for studying texts to analyze information manipulation has been set and solved.

The methodology of the research has been substantiated and explained. The work is based on computational artificial intelligence theories, particularly machine learning and deep learning, which enable the classification and clustering of large datasets.

For the first time, a comprehensive framework for analyzing textual disinformation and propaganda has been proposed, combining Bidirectional Encoder Representations from Transformers (BERT) and Bidirectional Long Short-Term Memory (Bi-LSTM) technologies to create interpretable vector representations of text. Unlike existing techniques, this framework integrates disinformation detection and thematic analysis, ensuring effective identification and analysis of textual disinformation and propaganda in social networks.

A novel classification model for textual disinformation and propaganda has been developed, based on the integration of transformer architectures and Bi-LSTM to incorporate both linear and non-linear dependencies in textual data. Unlike existing

models, it utilizes deep contextual embeddings and attention mechanisms, significantly improving classification accuracy.

Deep learning models for classifying textual disinformation and propaganda have been enhanced, specifically those based on XLNet, Bi-LSTM, and Attention-Based Bi-LSTM architectures. Unlike existing methods, these models incorporate adaptive evaluation of important text fragments, ensuring high classification performance.

Topic modeling methods have been improved using multi-head attention mechanisms. Unlike existing approaches, these methods leverage deep embeddings for data contextualization, ensuring clear separation between clusters.

Further development has been made in models for classifying textual disinformation and propaganda using ensemble machine learning methods. Unlike existing techniques, these models account for class balancing and adapt to changes in input data, enhancing model performance.

Statistical machine learning-based classification models have also been further developed. Unlike previous methods, these models have been adapted for classifying textual disinformation and propaganda, enabling the creation of systems for detecting misinformation, disinformation, and malinformation.

The developed models, methods, and software have been implemented at the “Balsillie School of International Affairs” (implementation act dated March 18, 2025), the “Center for International Governance Innovation” (implementation act dated March 18, 2025), the “State Institution “Kharkiv Regional Center for Disease Control and Prevention under the Ministry of Health of Ukraine”” (implementation act dated December 2, 2024), as well as in the educational process (implementation act dated 30 January 2025) and scientific activities (implementation act dated 4 March 2025) at the National Aerospace University “Kharkiv Aviation Institute”.

Keywords: disinformation, propaganda, fake news, social media analysis, machine learning, deep learning, neural networks, classification, clustering, modeling, data, threats, information intelligent system, artificial intelligence, AI system.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ

ЄСЗС – Європейська служба зовнішніх зв’язків

НАТО – Організація Північноатлантичного договору

США – Сполучені Штати Америки

ШІ – Штучний інтелект

COVID-19 – Коронавірусна хвороба 2019 року

AUC – площа під кривою (Area Under the Curve)

Attention-based BiLSTM – модель на основі двоспрямованої довгої короткострокової пам’яті з механізмом уваги

BERT – двоспрямоване кодувальне представлення з трансформерів (Bidirectional Encoder Representations from Transformers)

Bi-GRU – двоспрямована рекурентна нейронна мережа з ґейтованими рекурентними блоками (Bidirectional Gated Recurrent Unit)

Bi-LSTM – двоспрямована довга короткочасна пам’ять (Bidirectional Long Short-Term Memory)

BiRNN – двоспрямована рекурентна нейронна мережа (Bidirectional Recurrent Neural Network)

BRF – збалансований випадковий ліс (Balanced Random Forest)

CNN – згорткова нейронна мережа (Convolutional Neural Network)

DT – дерево рішень (Decision Tree)

EUvsDisinfo – платформа ЄС для боротьби з дезінформацією

FIMI – іноземне маніпулювання інформацією та втручання (Foreign Information Manipulation and Interference)

GRU – ґейтована рекурентна одиниця (Gated Recurrent Unit)

HDBSCAN – ієрархічний метод кластеризації на основі щільності з шумом (Hierarchical Density-Based Spatial Clustering of Applications with Noise)

HTML-теги – теги мови розмітки гіпертексту

IRA – Агентство Інтернет-досліджень (Internet Research Agency)

KNN – модель к-найближчих сусідів (k-Nearest Neighbors)

LDA – прихований розподіл Діріхле (Latent Dirichlet Allocation)

LightGBM – легка градієнтна бустингова машина (Light Gradient Boosting Machine)

LSTM – довга короткочасна пам’ять (Long Short-Term Memory)

MCC – коефіцієнт кореляції Метьюза (Matthews Correlation Coefficient)

MDM – (треба уточнити значення)

NB – модель наївного Баєса (Naïve Bayes)

RF – випадковий ліс (Random Forest)

RBF – радіальна базисна функція (Radial Basis Function)

RNN – рекурентна нейронна мережа (Recurrent Neural Network)

ROC – характеристична крива роботи приймача (Receiver Operating Characteristic)

SVM – метод опорних векторів (Support Vector Machine)

TF-IDF – частота терміну зворотної частоти документа (Term Frequency-Inverse Document Frequency)

U-MAS – система аналізу дезінформації UbiLab (UbiLab Misinformation Analysis System)

UMAP – рівномірне наближення та проєкція многовиду (Uniform Manifold Approximation and Projection)

URL-адреси – Уніфіковані локатори ресурсів (Uniform Resource Locator)

XGBoost – екстремальний градієнтний бустинг (Extreme Gradient Boosting)

ВСТУП

Обґрунтування вибору теми дослідження. В умовах сучасної гібридної війни, яку росія веде проти України та західних країн, інформаційна складова відіграє ключову роль [1]. Дезінформація та пропаганда стали основними інструментами ведення цієї війни, які дозволяють агресору формувати суспільну думку, маніпулювати масовою свідомістю та підривати демократичні інститути інститутів [2], [3], [4]. Соціальні мережі стали головним середовищем для поширення інформаційних атак, оскільки вони забезпечують швидке розповсюдження контенту, передбачають можливість таргетованого впливу та підвищують складність ідентифікації джерел дезінформації [5].

Росія активно використовує інформаційні операції для дестабілізації ситуації як в Україні, так і в країнах Заходу [6]. Серед ключових стратегій – дискредитація державних інституцій, поширення фейкових наративів щодо війни, посилення внутрішньої політичної поляризації та підрив довіри до офіційних джерел інформації. Це підтверджується численними дослідженнями та звітами міжнародних організацій, таких як NATO StratCom [7], EUvsDisinfo [8], [9], Atlantic Council [10]та інших, які фіксують масове застосування Кремлем дезінформаційних кампаній.

Соціальні мережі стали одним із головних майданчиків для поширення таких інформаційних атак [11]. Алгоритми платформ, орієнтовані на максимізацію залученості користувачів, сприяють вірусному розповсюдженню емоційно забарвленого та маніпулятивного контенту. Ворожі держави використовують як автоматизовані ботоферми, так і реальних агентів впливу для генерування та поширення дезінформації [12]. Наприклад, під час повномасштабного вторгнення росії в Україну було зафіксовано масовані атаки російських інформаційних кампаній, спрямовані на виправдання агресії, дискредитацію українського керівництва та створення розколу в міжнародній підтримці України [13].

У зв’язку з цим існує нагальна потреба у розробці ефективних моделей та методів для виявлення та аналізу текстової дезінформації та пропаганди в соціальних мережах. Традиційні методи фактчекінгу та ручного аналізу не встигають за темпами поширення фейкових новин, тому необхідні автоматизовані системи на основі машинного навчання та штучного інтелекту. Використання таких підходів дозволяє аналізувати великі обсяги даних, ідентифікувати структуру дезінформаційних наративів, визначати мережі поширення фейків та виявляти основні джерела інформаційних атак.

Таким чином, тема дослідження є актуальною. Її розробка сприяє вдосконаленню механізмів інформаційної безпеки, протидії гібридним загрозам та зміцненню стійкості українського суспільства та міжнародної спільноти перед російськими інформаційними атаками. Ефективне виявлення та аналіз дезінформації допоможе не лише у локалізації її впливу, але й у стратегічному плануванні інформаційного захисту держави та суспільства загалом.

Таким чином, **актуальним науковим** завданням є розробка технології дослідження текстів для аналізу маніпуляцій інформацією.

Мета і завдання дослідження формуються відповідно до предмета та об’єкта дослідження.

Об’єкт дослідження – процес комплексного аналізу текстової дезінформації для аналізу маніпуляцій інформацією

Предмет дослідження – моделі та методи аналізу текстової дезінформації та пропаганди в соціальних мережах.

Метою дисертаційної роботи є ефективне виявлення та аналіз текстової дезінформації та пропаганди в соціальних мережах.

Для досягнення поставленої мети розв’язуються наступні **задачі**:

1) аналіз теоретичної бази понять інформаційної війни, дезінформації та пропаганди, розгляд головних принципів та підходів російської дезінформації.

2) аналіз наявних моделей та методів машинного навчання для аналізу текстової дезінформації та пропаганди.

3) розробка моделей для аналізу дезінформації та пропаганди на основі методів машинного навчання

4) розробка моделей на основі ансамблевих методів машинного навчання

5) розробка моделей на основі методів глибокого машинного навчання

6) розробка фреймворку для комплексного аналізу текстової дезінформації та пропаганди у соціальних мережах.

7) вирішення практичних завчань пов’язаних виявленням та аналізом текстової дезінформації та пропаганди за допомогою розроблених моделей.

Методи дослідження базуються на теоріях обчислювального штучного інтелекту, а саме – машинного навчання та глибокого машинного навчання що дозволяють класифікувати та кластеризувати великі набори даних.

Наукова новизна одержаних результатів:

1. Вперше запропоновано фреймворк комплексного аналізу текстової дезінформації та пропаганди, заснований на комбінуванні технологій двоспрямованого кодувального представлення з трансформерів (BERT) та двоспрямованої довгої короткочасної пам’яті (Bi-LSTM) для створення інтерпретованих векторних представлень тексту, який на відміну від існуючих технік, об’єднує виявлення та тематичний аналіз дезінформації, що дозволяє забезпечити ефективне виявлення та аналіз текстової дезінформації та пропаганди в соціальних мережах.

2. Вперше розроблено модель класифікації текстової дезінформації та пропаганди, заснований на поєднанні трансформерів і архітектури Bi-LSTM для інтеграції лінійних та нелінійних залежностей текстових даних, яка на відміну від існуючих використовує глибокі контекстуальні ембедінги і механізми уваги, що дозволяє підвищити точність моделі.

3. Удосконалено моделі глибокого навчання для класифікації текстової дезінформації та пропаганди, засновані на архітектурах XLNet, Bi-LSTM та Attention-Based Bi-LSTM, які на відміну від існуючих застосовують адаптивне оцінювання важливих фрагментів тексту, що дозволяє забезпечувати високий рівень продуктивності класифікації.

4. Удосконалено метод тематичного моделювання, заснований на застосуванні механізмів багатоголової уваги, що на відміну від існуючих використовує глибокі ембедінги для контекстуалізації даних, що дозволяє забезпечити чітке розмежування між кластерами.

5. Дістали подальшого розвитку моделі для класифікації текстової дезінформації та пропаганди, засновані на ансамблевих методах машинного навчання, які на відміну від існуючих враховують балансування класів і адаптуються до змін вхідних даних, що дозволяє підвищити продуктивність моделей.

6. Дістали подальшого розвитку моделі для класифікації інформації, засновані на статистичних методах машинного навчання, які на відміну від існуючих адаптовані до класифікації текстової дезінформації та пропаганди, що дозволяє створення систем для виявлення місінформації, дезінформації та малінформації.

Особистий внесок здобувача. Всі основні наукові положення, результати, висновки та рекомендації дисертаційної роботи отримані автором самостійно та опубліковано у роботах [14-29]. У публікаціях, які написано у співаторстві, здобувачеві належать такі результати: [14] – розробка фреймворку для аналізу місінформації, [15] – моделі глибокого навчання для класифікації текстової дезінформації, [16] – ансамблеві моделі машинного навчання для класифікації текстової дезінформації, [17] – модель XLNet для класифікації дезінформації в гібридних війнах під час російської війни проти України, [18] – формулювання етичних принципів інфодеміки та досліджень інформаційного стеження у соціальних медіа, [19] – розробка сценаріїв для передбачень трендів у дезінформації підсиленої штучним інтелектом, [20] –

модель глибокого навчання для виявлення місінформації в політичних новинах, [21] – аналіз впливу вилучення стоп слів для попередньої обробки даних для моделей глибокого навчання для аналізу україномовної дезінформації, [22] – моделі виявлення місінформації пов'язаної з COVID-19 на китайськомовному наборі даних, [23] – збір та аналіз кейсів російських інформаційних кампаній, які використовували тему Covid-19 для впливу на геополітичні процеси, [24] – аналіз моделей машинного навчання для класифікації російської пропаганди, [25] – порівняння моделей розповсюдження COVID-19 на основі аналізу соціальних мереж, [26]– аналіз мультиагентних підходів для моделювання соціальних мереж, [27] – розробка рекомендацій з питань політики для протидії російській дезінформації у Канаді, [28] – розробка рекомендацій з питань політики для протидій дезінформації, підсиленої штучним інтелектом, [29] – аналіз наративів російської дезінформації у контексті виборів до США, використовуючи розроблений фреймворк аналізу місінформації.

Апробація матеріалів дисертації. Основні положення, ідеї, висновки дисертаційної роботи доповідалися та обговорювалися на 4th International Workshop of IT-Professionals on Artificial Intelligence, ProfIT AI 2024 (Кембрідж, США, 2024), V Міжнародній науково-практичній конференції IT-професіоналів та аналітиків комп'ютерних систем "ProfIT Conference" (Харків, 2023), 2023 13th International Conference on Dependable Systems, Services and Technologies, DESSERT 2023 (Афіни, Греція, 2023), 3rd International Workshop of IT-Professionals on Artificial Intelligence, ProfIT AI 2023 (Ватерлу, Канада, 2023), 17th World Congress on Public Health (Рим, Італія, 2023), IV Міжнародній науково-практичній конференції IT-професіоналів та аналітиків комп'ютерних систем "ProfIT Conference" (Харків, 2021), III Міжнародній науково-практичній конференції IT-професіоналів та аналітиків комп'ютерних систем "ProfIT Conference" (Харків, 2020)

Структура й обсяг дисертації. Дисертація складається з анотації, змісту, переліку умовних скорочень, вступу, чотирьох розділів, висновку,

списку використаних джерел та додатків. Повний обсяг роботи становить 207 сторінок друкованого тексту, з них анотація – на 7 стор., зміст – на 4 стор., перелік умовних скорочень – на 2 стор., основний текст – на 124 стор., список із 151 використаних джерел – на 17 стор., додатки – на 22 стор. Дисертація містить 54 рисунки, та 16 таблиць.

Зв'язок роботи з науковими програмами, планами, темами. Дослідження, результати яких викладено в дисертації, виконано на кафедрі математичного моделювання та штучного інтелекту Національного аерокосмічного університету "Харківський авіаційний інститут" в рамках виконання науково-дослідних робіт за проєктами "Розробка методологічного та інструментального забезпечення Agile трансформації процесів відбудови медичних закладів України для подолання розладів здоров'я населення у воєнний та повоєнний періоди" (Національний фонд досліджень України, № 2022.01/0017), "Uncovering Information Warfare: Detecting Russian Propaganda on Social Media" (MITACS, №IT37637), "Understanding and Mitigating the Risks of Generative AI in Propaganda and Informational Warfare" (MITACS and Center for International Governance Innovation, №IT36431).

Практичне значення. Використання розроблених моделей та методів дозволило збільшити точність класифікації виявлення текстової дезінформації та пропаганди у соціальних мережах. Використання розробленого фреймворку забезпечує аналіз наративів, який може бути використаний для побудови стратегій контрзаходів та стратегічних комунікацій. Наприклад, застосування фреймворку для аналізу дезінформації про вибори, COVID-19 та російську війну проти України дозволить отримати цінне розуміння наративів у великих корпусах даних.

Результати дисертаційної роботи впроваджено у практичну діяльність низки установ, що підтверджено відповідними актами: Balsillie School of International Affairs (акт впровадження від 18-го березня 2025), Center for International Governance Innovation (акт впровадження від 18-го березня 2025), Державну установу «Харківський обласний центр контролю та профілактики

хвороб при Міністерстві охорони здоров’я» (акт впровадження від 2 грудня 2024), а також у навчальний процес кафедри математичного моделювання та штучного інтелекту (акт впровадження від 30 січня 2025) та наукову роботу (акт впровадження від 4 березня 2025) Національного аерокосмічного університету.

Подяка. Авторка висловлює глибоку вдячність науковому керівнику, члену-кореспонденту Національної академії наук України, Заслуженому працівнику науки та техніки України, лауреату державної премії України в галузі науки та техніки, д.ф.-м.н., професору Сергію Всеволодовичу Яковлеву за цінні професійні поради, конструктивні зауваження та неоціненну підтримку на всіх етапах підготовки і захисту цієї дисертації.

Щира дяка доктору Плініо Моріті – директору Ubiquitous Health Technology Lab та доктору Хелен Чен зі Школи громадського здоров’я University of Waterloo (Канада) за наставництво та можливість навчатись підходам та передовим технологіям штучного інтелекту для аналізу місінформації у соціальних мережах.

Велике спасибі менторам доктору Ен Фітз-Джеральд, – директорці школи міжнародних відносин Балсіллі та професорці Університету Вілфріда Лоріе, доктору Девіду Велчу – директору програми Глобального Урядування Університету Ватерлу, доктору Александру Ланошці – експерту у сфері безпеки та керівнику магістерської роботи в Університеті Ватерлу, та Джіл Сінклер – менторці та канадській експертці з безпеки та оборони та радниці міністра оборони України за те, що були щедрими у своїх порадах та настановах, будуючи надійну фундацію знань про міжнародну безпеку, інформаційні війни та підходи росії до гібридних воєн. Під вашим керівництвом було написано багато статей та проведено подій, які показали справжнє обличчя росії, використовуючи науку, доказові інструменти та формуючи переконливі аргументи, аби відстоювати інтереси світових демократій та України. Ви – справжні друзі України та борці за демократію, внесок яких у розвиток науки та громадянського суспільства неоціненний.

Також важливо подякувати Центру Міжнародних Інновацій Урядування та Хабу Цифрових політик, де авторка проводила дослідження російської дезінформації у контексті президентських виборів США під час написання цієї дисертації, матеріали яких включені у цю роботу. Щира дяка центру Стратегічних Комунікацій НАТО, де авторка провела наукове стажування, опановуючи методи аналізу текстової дезінформації та пропаганди, працюючи з великими наборами даних.

Авторка також щиро вдячна Гарольду Гудвіну, колишньому директору Інституту Штучного Інтелекту Університету Ватерлу та доктору Джефу Каселло за підтримку наукового шляху у Канаді та безмежну відданість науці та служінню людям та демократичним і ліберальним цінностям. Ваша любов та відданість своїй справі робить цей світ кращим.

Щира дяка усім співавторам публікацій за їх безцінний внесок у нашу спільну роботу та результати, які можна було досягнути лише у спільній праці, зокрема Василю Чомку – студенту університету Ватерлу, другу та колезі та Дмитру Чумаченку – колезі, другу та партнеру.

РОЗДІЛ 1. ТЕОРЕТИЧНИЙ АНАЛІЗ ІНФОРМАЦІЙНИХ МАНІПУЛЯЦІЙ ТА АНАЛІЗ ІСНУЮЧИХ МЕТОДІВ ТА ФРЕЙМВОРКІВ ДЛЯ ЇЇ АНАЛІЗУ

1.1. Теорія інформаційних маніпуляцій у контексті інформаційних воєн

Маніпуляція інформацією, як один з найбільших викликів сьогодення, набуває ще більшої значущості у сучасному поляризованому світі протистоянь демократій та автократій. Одну з найголовніших ролей у цьому процесі набувають технології штучного інтелекту (ШІ), які стають підсилювачами та активаторами нових тактик та підходів до створення та поширення дезінформації та маніпуляції суспільною думкою [19]. Сучасні технології змінюють правила гри, перебудовуючи підходи розповсюдження та споживання інформації кінцевими користувачами, інструменталізуючи алгоритми та дані. “Global Risk Report” від Міжнародного економічного форуму у Давосі вже другий рік поспіль (2024 та 2025) визначає дезінформацію як одну з найбільших загроз сучасного світу [30, 31]. Нещодавня доповідь OECD [32] свідчить, що поширення неправдивої інформації посилює поляризацію, що призводить до руйнування соціальної структури відкритих суспільств у довгостроковій перспективі.

1.1.1.Типи інформаційних розладів

З метою маніпуляції зловмисні суб’єкти спотворюють наявну інформацію або створюють неправдивий контент, відомий як фейки. Такі маніпуляції зі спотвореною та неправдивою інформацією називаються інформаційними розладами [33]. Вони охоплюють різні форми контенту, що

можуть підірвати демократичні інститути, соціальну єдність і міжнародні відносини.

Інформаційні розлади характеризуються різними способами створення, поширення та споживання неправдивої інформації. Ці розлади впливають на те, як люди та суспільства сприймають і розуміють реальність. Ця концепція охоплює три основні форми інформаційних розладів: місінформацію, дезінформацію та малінформацію. Графічно ці поняття можна представити, використовуючи виміри правдоподібності та наміру нашкодити, як осі координат (рисунок 1.1). Клер Вардл та Хосейн Деречшан, основоположники цієї класифікації, визначають їх так [33]:

- Місінформація (англ. misinformation) – неправдива інформація, яка поширюється без наміру завдати шкоди.
- Дезінформація (англ. disinformation) – неправдива інформація, яка свідомо поширюється з метою завдати шкоди.
- Малінформація (англ. malinformation) – правдива або частково правдива інформація, яка використовується для завдання шкоди.

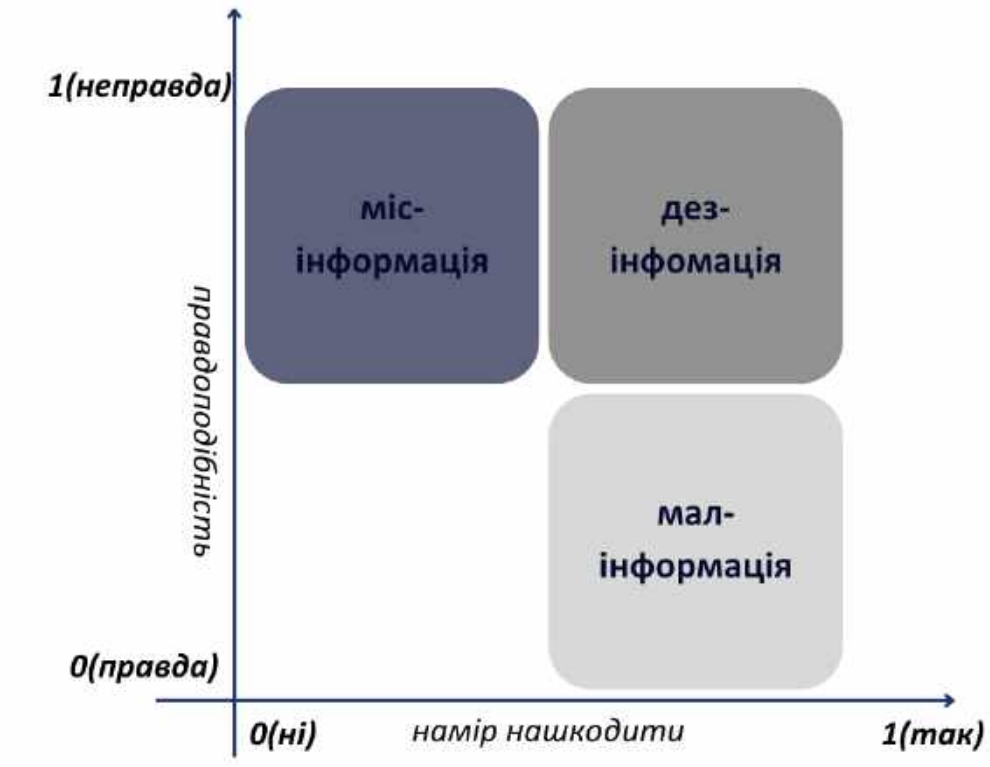

Рисунок 1.1 – Типи інформаційних розладів

Дезінформація мотивована багатьма факторами. Одним з найпоширеніших у сучасному світі є спроби авторитарних країн отримати політичний вплив шляхом сіяння розбрату, поляризації, недовіри і розчарування у демократії та демократичних інституціях. Вона також може бути частиною гібридної війни та військових стратегій, коли дезінформація спрямована на введення в оману військових та цивільного населення країни-противника для забезпечення військової переваги. До прикладу, під час анексії Криму поєднання “зелених чоловічків” (військових без розпізнавальних знаків), дезінформаційних кампаній, кібератак та політичної маніпуляції дозволило росії окупувати півострів [34]. Прикладом дезінформації є сфабриковані росією фейкові відео, на яких, буцімто, видно, як тіла загиблих у Бучі “рухаються” [8]. Ці фейкові новини, аби переконати міжнародну спільноту та населення України й росії, що трагедія у Бучі – несправжня, російські війська не вчиняють військових злочинів, цим самим зменшити єдність українців та підтримку міжнародної спільноти.

У процесі поширення каналами інформації, дезінформація часто перетворюється на місінформацію [35]. Місінформація також описує неправдивий контент, але людина, яка його поширює, не усвідомлює, що ця інформація є хибною або фейковою новиною. Часто дезінформацію підхоплює хтось, хто не розуміє її неправдивої суті, і ця людина поширює її у своїх мережах, вважаючи, що допомагає. Прикладом місінформації є неправдива інформація про вірус, шляхи його поширення та лікування, яку користувачі соцмереж активно розповсюджували під час пандемії коронавірусу [36]. У намаганнях допомогти своїй аудиторії запобігти захворюванню, люди ненавмисно поширювали неперевірену та науково необґрунтовану інформацію, яка у цьому випадку була місінформацією.

Третя категорія інформаційних розладів – це малінформація. Цей термін описує правдиву або частково правдиву інформацію, яка поширюється з метою завдати шкоди. Прикладом цього є випадок, коли російські агенти зламали електронну пошту Демократичного національного комітету та

кампанії Гілларі Клінтон і оприлюднили певні деталі, щоб завдати шкоди репутації [37]. Також прикладом цього типу маніпуляції є створення росіянами сюжетів та історій про окремих право-радикальних українських громадян, аби зобразити всіх українців як нацистів в очах міжнародної спільноти [38].

Деякі країни використовують модель MDM (Misinformation – Disinformation – Malinformation) для обрамлення та написання політик протидії маніпуляціям інформацією [29], проте останнім часом усе частіше використовується FIMI (Foreign Information Manipulation and Interference), введене Європейською службою зовнішніх справ (ЄСЗС). У 2023 ЄСЗС опублікувала свій перший звіт про загрози FIMI, в якому окреслюється поняття FIMI та виклики, які воно створює [40]. У цьому звіті наголошується на необхідності колективного розуміння та реагування на виклики FIMI, які ЄСЗС визначає як “переважно непротиправну модель поведінки, яка загрожує або має потенціал негативно впливати на цінності, процедури та політичні процеси”. Ця діяльність характеризується маніпулятивністю, умисністю, координацією та залученням державних або недержавних суб’єктів, що діють через кордони.

Тактика FIMI є різноманітною та постійно розвивається. Вона включає поширення дезінформації, створення фейкових новинних вебсайтів, імітацію авторитетних медіа ресурсів та використання ШІ для створення оманливого контенту. Наприклад, російська дезінформаційна кампанія “Doppelganger”, передбачала клонування популярних новинних вебсайтів для поширення прокремлівських наративів, що підривало довіру громадськості до авторитетних джерел інформації напередодні виборів у Європі та Америці у 2024 [12, 13]. Проте концепт FIMI включає лише закордонні загрози, лишаючи поза увагою зусилля внутрішніх суб’єктів вплинути на суспільну думку. Авторитарні країни, в свою чергу, є винахідливими у пошуку шляхів дестабілізації демократії зсередини, залучаючи громадян західних країн до поширення пропаганди та вигідних для них наративів.

1.1.2. Типи неконвенційних воєн та місце інформаційної війни серед типів гібридної війни

Маніпуляції інформацією є однією зі зброї гібридної війни, яка у свою чергу містить у собі багато різних варіацій інтервенцій, зокрема, когнітивну, інформаційну та кібервійну. Тзу-Чег Гунг та Тзу-Вей Гунг описують їх взаємозалежності у статті вигляді [42]. На рисунку 1.2 представлене їх графічне відображення.

Рисунок 1.2 – Концептуальний взаємозв’язок між різними видами неконвенційних воєн.

Кібервійна передбачає масштабну скоординовану цифрову атаку для націлювання на інформаційні мережі іншої держави чи суб’єктів протистояння з метою порушення, пошкодження або знищення критичної інфраструктури та систем [43]. Такі атаки можуть завдати значної шкоди, включаючи порушення роботи важливих комп’ютерних систем та сервісів. Когнітивна війна – це діяльність, спрямована на маніпуляцію ментальними станами та поведінкою через контроль когнітивних процесів, використовуючи нейронауку, як зброю [42]. Вона відрізняється від інформаційної війни, яка спрямована на маніпуляцію особами, які приймають рішення, через медіа, соціальні мережі

або міжособистісні комунікації, спотворюючи інформацію та емоційно чи інтелектуально впливаючи на аудиторію [44]. Інформаційна війна часто використовує дезінформацію, формування фейкових наративів та інші методи для маніпуляції емоціями або переконаннями, але не переходить до прямого контролю когнітивних процесів.

Концепція інформаційної війни була формалізована в доктрині збройних сил США наприкінці 1980-х – на початку 1990-х років [45]. Термін став широко відомим завдяки публікації Міністерства оборони США про кібер- та інформаційні стратегії. Російські військові теоретики протягом тривалого часу обговорювали стратегічне використання інформації у війні, часто називаючи це частиною “рефлексивного контролю” або “інформаційно-психологічними операціями” [46]. Їхні роботи в цій сфері беруть початок із часів Холодної війни і залишаються важливими для аналізу та розуміння сучасних російських підходів до інформаційної війни.

Зокрема, концепція “рефлексивного контролю” передбачає процес, за якого одна сторона передає певну інформацію противнику, формуючи його сприйняття та процеси прийняття рішень, щоб змусити його ухвалити рішення, які є вигідними для ініціатора [47]. По суті, це метод впливу на противника шляхом зміни його уявлення про ситуацію, що змушує його діяти відповідно до стратегічних цілей ініціатора. Ця концепція часто використовується під час ядерного шантажу росії, як стратегічна маніпуляція, коли росія поширює інформацію або погрози про свою готовність застосувати ядерну зброю, формуючи страхи у західних лідерів. Формуючи такі повідомлення, росія прагне вплинути на їхнє прийняття рішень, наприклад, щодо обмежень у застосуванні західної зброї на території росії, або ж обмеження збройної підтримки України взагалі.

Стратегія росії у сфері інформаційної війни являє собою безперервні, багатовимірні активності, що пріоритетно спрямовані на маніпуляцію громадською думкою як ключовий елемент геополітичного впливу [1]. Росія інтегрує свої тактики у довгострокові операції, спрямовані на дестабілізацію

супротивників і зміцнення власного впливу. Дослідники зазначають, що російська військова доктрина приділяє значну увагу саме “інформаційно-психологічній” війні, мета якої виходить за межі лише технічних інтервенцій і включає формування наративів, сіяння розбрату та підрив довіри до демократичних інститутів [2], [3], [4]. Цей термін набув особливої популярності під час повномасштабної війни росії проти України, та передбачає використання інформації для впливу на емоції, мотиви, мислення та поведінку, з метою підриву процесів прийняття рішень супротивника та його волі до дій [48]. Ці стратегії відповідають принципам, викладеним у фундаментальних документах, таких як “Доктрина Герасимова”, яка акцентує увагу на нелінійних методах конфлікту, що поєднують військові та невійськові інструменти для досягнення стратегічних цілей [49].

У російській інформаційній війні пропаганда є стратегічним інструментом, який використовується для впливу на сприйняття, маніпулювання наративами та досягнення геополітичних цілей [50]. Вона включає цілеспрямоване поширення інформації, часто через державні медіа та цифрові платформи, для формування громадської думки як всередині країни, так і за її межами. Цей підхід охоплює різні тактики, зокрема дезінформаційні кампанії, психологічні операції та використання соціальних мереж для посилення окремих повідомлень. Зазвичай вона включає вибіркове подання фактів, емоційні заклики, повторюваність та придушення протилежних точок зору, щоб переконати або маніпулювати цільовою аудиторією [51]. Російські дезінформаційні зусилля часто націлені на використання існуючих суспільних конфліктів у цільових країнах. Посилюючи суперечливі питання, ці кампанії прагнуть послабити соціальну згуртованість і дестабілізувати політичні системи. Наприклад, російські суб’єкти відомі поширенням наративів, що поглиблюють расові конфлікти та політичну поляризацію у західних суспільствах, підриваючи довіру до демократичних інституцій [52].

Соціальні мережі стають полем бою у інформаційних війнах, а суб’єкти пов’язані з Кремлем такі як “Агентство соціального проєктування”, успішно

вивчають особливості різних платформ, аби адаптовувати повідомлення та наративи відповідно до форматів та аудиторій різних платформ [45]. Нещодавно Міністерство юстиції США викрило та ліквідувало мережу з майже 1 000 російських пропагандистських акаунтів на платформі X (колишній Twitter) [54]. Ці акаунти, які видавали себе за американців, насправді контролювалися російськими операторами, які отримували фінансування та керівництво з Кремля. Тим часом TikTok також потрапив під пильну увагу через розслідування Європейського Союзу, пов'язане з втручанням росії у президентські вибори в Румунії [55]. Аналогічно, на Facebook російська кампанія "Допельгангер " була виявлена через просування проросійських наративів і купівлю реклами через фейкові акаунти перед важливими виборами в Європі [56]. Незважаючи на санкції ЄС у 2022 році, французькі та німецькі органи влади підтверджують, що російські мережі дезінформації продовжують впливати на європейську аудиторію [57].

Зміни у політиці модерації контенту на основних платформах соціальних мереж додатково дозволяють росії обходити попередні обмеження та розширювати свої кампанії впливу. Під керівництвом Ілона Маска платформа X перейшла до підходу "свободи слова", який, на думку вчених, перетворив її на резонатор для правих поглядів, включаючи посилення наративів на підтримку президента США Дональда Трампа [58]. Дослідники підкреслюють, що інформаційне середовище платформи стає дедалі олігархічним, де кілька впливових акаунтів домінують у потоці новин. Аналогічно, рішення Meta припинити програму фактчекінгу в США після перемоги Трампа на виборах, зробило Facebook та Instagram більш вразливими до дезінформації [59]. Замінивши професійний фактчекінг на систему "заміток спільноти", компанія зменшила свою здатність протистояти координованим кампаніям пропаганди.

Зміни у політиках доступу до контенту для дослідників стають все більш актуальними. Хоча офіційна політика X заявляє про можливість доступу до API для дослідницьких цілей, багато вчених стикаються з тривалими

затримками в отриманні доступу, інколи чекаючи місяцями, і в результаті не отримують потрібні дані [52]. Ця ситуація стає проблемною, оскільки актуальність аналізу з часом знижується. Можливість збирати та аналізувати набори даних є критично важливою для підтримки демократичного дискурсу. Обмеження доступу позбавляє академічну спільноту важливої ролі у функціонуванні соціальних мереж як демократичного суспільного наглядача.

1.2. Застосування технологій ШІ для протидії дезінформації, інформаційним маніпуляціям, пропаганді та фейкам

Потенційне та реальне застосування ШІ для дезінформації є актуальною темою в академічному середовищі, органах влади та корпоративному секторі. Серед загроз, які ШІ може підсилити у створенні та поширенні дезінформації є: ерозія довіри до доказів, зниження когнітивної автономії, експлуатація особистих брендів, посилення мови ненависті, ерозія критичного мислення та інформаційної грамотності, зниження відслідковуваності іноземних операцій, експоненційне погіршення якості контенту, порушення конфіденційності [61], [62].

Росія активно використовує ШІ для того, аби зменшити витрати на створення дезінформації як в Україні, так і закордоном [63]. Наприклад, на початку повномасштабного вторгнення, використовуючи ШІ для дипфейків, росія створила фейкове відео, де президент Зеленський закликає громадян скласти зброю [64]. Також помітною була російська дезінформаційна кампанія, під час якої за допомогою ШІ було створено персонажа Воху – українського військового з синдромом Дауна, і знято ряд відео, де інші військові б’ють, ображають чи знущаються з нього [65]. Ці відеоролики поширювали не лише в російсько- чи українськомовному сегменті, але й активно розповсюджували в англомовному середовищі. Вони були емоційно насичені і активно створювали портрет українських військових, як

безжалісних катів, які здатні до військових злочинів. Також частими були випадки використання ШІ для впливу на міжнародні демократії, особливо під час виборів 2024 року, зокрема у операціях “Допельгангер” та ряді інших [41].

Проте технології також можуть використовуватися для допомоги в боротьбі з інформаційними маніпуляціями. Зокрема, для аналізу контенту, поведінкових шаблонів, наративів, великих наборів даних та для розповсюдження контрнаративів, які полегшують наслідки дезінформації, що у свою чергу, допомагає урядам будувати комунікаційні контрстратегії, базуючись на реальних даних та кількісно-якісній аналітиці, а також у кооперації з громадським сектором і платформами виявляти агентів впливу, їх скоординовану неавтентичну поведінку в цифрових середовищах та наративи, які вони проштовхують.

Також перспективною областю для для аналізу поведінки користувачів у соціальних мережах та дезінформації з якою вони взаємодіють є мультиагентне моделювання. За допомогою мультиагентного підходу можна описати індивідуальні взаємодії, з яких виникають соціальні патерни, та проводити динамічний аналіз таких поведінкових патернів [26]. Мультиагентні моделі соціальних мереж зазвичай аналізуються на основі набору гіпотез. Один із способів перевірки гіпотез – це спостереження за графіками часових рядів для набору показників. При аналізі агентно-орієнтованих соціальних мереж важливим питанням є розуміння ролі соціальних процесів в обмеженні динаміки створюваних мереж. Мета агентних соціальних мереж – вивчити траєкторії даних, що моделюються, і зрозуміти явища, що моделюються. Це дозволяє уникнути недоліків стохастичних моделей для динамічних соціальних мереж, де існуючі дані використовуються для підгонки моделей і оцінки параметрів [25].

1.2.1. Застосування ансамблевих методів машинного навчання для протидії дезінформації, інформаційним маніпуляціям, пропаганді та фейкам

Сучасні дослідження надають всебічний аналіз різноманітних стратегій та інструментів, які використовують для протидії дезінформації та пропаганді. Традиційні методи виявлення фейків, такі як ручна перевірка фактів, та підхід, заснований на ключових словах, виявилися недостатніми для обробки великого обсягу та складності дезінформації, яка циркулює в Інтернеті [66]. У зв'язку з цим машинне навчання стало одним із найефективніших підходів у боротьбі з фейковими новинами, дозволяючи автоматизовано аналізувати великі текстові корпуси, виділяти ключові характеристики та розпізнавати маніпулятивний контент.

Методи машинного навчання показали високу ефективність у виявленні фейкових новин та текстової дезінформації. У статті [67] дослідники використали модель XGBoost для оцінки значущості різних ознак новин, та побудували кілька класифікаційних моделей. Зокрема, заснованих на методі опорних векторів (SVM), випадкового лісу (RF), логістичної регресії (LR), дерева рішень (DT), нейронній мережі NNET. Модель RF продемонструвала найкращий результат, досягнувши точності 94%. Подібним чином у статті [68] була запропонована модель, яка поєднує машинне навчання з методом витягнення біомедичної інформації для виявлення фейкових новин, пов'язаних з COVID-19. Найефективнішою виявилася модель RF, що досягла AUC 0,882, суттєво перевершивши інші традиційні підходи.

SVM є однією з найчастіше застосовуваних моделей у класифікації текстових даних. У статті [69] було проведено оцінку продуктивності кількох класифікаторів, включаючи RF, DT, метод к-найближчих сусідів (KNN), LR та SVM. Найвищу точність (93,61%) продемонструвала саме SVM. Дослідження [70] також показало ефективність SVM у класифікації фейкових новин бенгальською мовою, де ця модель досягла точності 73,20%. У роботі [71] застосовано TF-IDF для вилучення ознак у поєднанні з двома

класифікаторами: SVM та поліноміальною моделлю наївного Баєса (NB). Результати показали, що SVM досяг точності 94,83%, а NB – 91,38%.

Дослідження [72] пропонує новий підхід, заснований на методі KNN, з інтеграцією генетичного та еволюційного відбору ознак. Це дозволило досягти точності 91,3%. Також автори досліджували квантове машинне навчання, використовуючи квантовий KNN, який забезпечив точність 84,4%. У статті [73] було розглянуто генетичні алгоритми у поєднанні з SVM, NB, RF та LR. Кращі результати продемонстрували SVM та RF, які досягли 97% точності при класифікації фейкових вакансій.

Аналіз досліджень свідчить про ефективність методів машинного навчання, зокрема RF та XGBoost, для виявлення фейкових новин, проте серед них SVM залишається одним із найкращих варіантів для задач класифікації текстової дезінформації та пропаганди.

1.2.2. Застосування глибокого навчання для протидії дезінформації, інформаційним маніпуляціям, пропаганді та фейкам

Розвиток технологій, зокрема, глибокого навчання, як підгалузі машинного навчання, спричинив їх широке застосування для задачі аналізу різних даних. Машинне навчання пропонує перспективний підхід для класифікації та боротьби з пропагандою, з потенціалом для постійного навчання та адаптації [24]. Моделі машинного навчання демонструють значний успіх у задачах обробки природної мови, зокрема аналізу емоцій [74], підсумовування текстів [75] та перекладу мов [76].

У контексті класифікації дезінформації, фейкових новин та пропаганди, моделі глибокого навчання застосовуються для аналізу текстового змісту новинних статей та визначення їх достовірності [77]. Ці моделі здатні обробляти великі обсяги даних, виділяти складні закономірності та вловлювати нюанси мови, що є важливим для точного класифікування

інформаційних маніпуляцій [78]. Різні архітектури моделей глибокого навчання, такі як згорткові нейронні мережі (CNN), рекурентні нейронні мережі (RNN) та мережі довгої короткострокової пам’яті (LSTM), досліджувались для класифікації дезінформації. Останнім часом до моделей глибокого навчання додано механізми уваги, які дозволяють моделі зосереджуватись на найбільш релевантних частинах вхідного тексту, що підвищує їх продуктивність [79]. Ці досягнення у глибокому навчанні значно сприяли боротьбі з дезінформацією та проклали шлях до розробки більш точних і ефективних моделей класифікації.

У статті [80] розглядається проблема виявлення фейкових новин на платформах соціальних медіа з використанням гібридного підходу, який включає слабоконтрольоване навчання, Bi-GRU та BiLSTM. Запропонована модель досягла точності 90%, підтвердивши ефективність поєднання глибокого навчання та SVM. Подібний підхід використовувався в дослідженні [81], де для аналізу набору даних “COVID-19 Fake News Dataset” було запропоновано гібридну модель, яка перевершила традиційні LSTM і BiLSTM, досягнувши точності 75.34%.

Дослідження [82] пропонує нову модель для виявлення фейкових новин арабською мовою, використовуючи LSTM-RNN з оптимізацією мисливець-жертва. Модель досягла 96.57% точності для фейкових новин про COVID-19 та 93.53% для сатиричних новин. Аналогічно, у статті [83] досліджено виявлення фейкових новин в’єтнамською мовою за допомогою CNN і RNN, що забезпечило точність 85%.

CNN продемонстрували ефективність у виявленні фейкових новин, особливо у випадку малоресурсних мов. У статті [84] CNN досягла міри F1 95% та точності понад 91% у виявленні фейкових новин курдською мовою. У дослідженні [85] CNN використовувалась для класифікації фейкових новин у соціальних мережах, де точність варіювалася від 81% до 100% залежно від класифікаторів. У статті [86] модель CNN досягла точності 92,38% у

словацькомовному корпусі, а у статті [87] гібридна модель CNN-BiLSTM з Word2Vec показала точність 97%.

Гібридні моделі глибокого навчання також демонструють високу ефективність у виявленні дезінформації. У дослідженні [88] було запропоновано модель, що використовує глибокі мережі вірувань з оптимізацією методу STODL-FNDC, яка досягла точності 95,50%. У статті [89] запропоновано використання BiRNN з різними рівнями виключення (dropout), де найкраща точність (90,1%) була досягнута при значенні виключення 0,3.

Таким чином, RNN, LSTM та Bi-LSTM демонструють високу точність, особливо при застосуванні гібридних підходів. CNN є ефективною у класифікації текстових даних, зокрема у малоресурсних мовах. Гібридні підходи, що поєднують різні глибокі архітектури та оптимізаційні методи, дозволяють покращити продуктивність моделей, що робить їх перспективним напрямом у боротьбі з дезінформацією та фейковими новинами.

1.2.3. Застосування методів машинного навчання для аналізу інформаційного середовища під час надзвичайних ситуацій

Інформаційні розлади залишаються постійними викликами демократій, особливо під час надзвичайних ситуацій. Незалежно від характеру катастрофи, чи це війна, чи природній катаклізм, чи світова пандемія – механізми маніпуляції, а також стратегії аналізу та протидії цим інформаційним загрозам часто мають універсальний характер. Саме ця універсальність робить вивчення дезінформації критично важливим. Під час криз дезінформація поширюється з надзвичайною швидкістю, адже екстремальні ситуації дезорієнтують людей, роблять їх вразливими та спонукають до активного пошуку інформації, що часто призводить до знайдення відповідей на свої запитання у ненадійних джерелах, які покликані маніпулювати користувачами.

Пандемія коронавірусу створила сприятливий ґрунт для поширення дезінформації. Як перша глобальна криза охорони здоров'я, що відбулася у цифрову епоху, вона висвітлила силу соціальних мереж та інших цифрових платформ у поширенні неправдивих відомостей про походження, поширення та лікування вірусу. Це явище, яке отримало назву "інфодемія", та значно ускладнило завдання глобальних систем охорони здоров'я, пояснивши небезпечний зв'язок між дезінформацією та її наслідками для системи громадського здоров'я [90].

Надзвичайні ситуації, які мають характер явищ без кордону, такі як пандемія COVID-19 стали плодовитим ґрунтом для зловмисних суб'єктів у поширенні неправдивої інформації. Кампанії дезінформації в цей період охоплювали широкий спектр тем – від теорій змови про штучне походження вірусу до фальшивих методів лікування та антивакцинальної риторики. Одним із найвідоміших суб'єктів, які використовували ці вразливості, стали російські пропагандистські мережі, зокрема такі організації, як "Агентство Інтернет-досліджень" (IRA) власником якого був Євгеній Прігожин – російський воєнний злочинець, олігарх, та власник ПВК "Вагнер".

IRA активно використовувало соціальну поляризацію під час пандемії, особливо у таких країнах, як Сполучені Штати, де воно сіяло розбрат і недовіру серед громадян [12]. Ці кампанії працювали у двох напрямах: поширення теорій змови щодо походження вірусу та ганьблення невакцинованих людей, що ще більше поглиблювало соціальний розкол. Ця багатогранна стратегія продемонструвала, як маніпуляція дискурсом про охорону здоров'я може бути використана для руйнування суспільної згуртованості [23] .

Інфодемія поглиблює кризу громадського здоров'я, спричиняючи хаос й недовіру. Катастрофічна ціна інфодемії, яка вимірювалась тисячами життів, що були втрачені через споживання неправдивої інформації про вірус дала поштовх до широких дискусій щодо етики місінформації, пов'язаної зі здоров'ям. Благочинність, ненанесення шкоди, автономія, справедливість та

ефективність були визначені дослідниками, як основоположні етичні принципи, підкреслюючи їхню відповідну значущість для досліджень інфодемій [18]. Пандемія COVID-19 продемонструвала, що демократичні уряди особливо вразливі до кампаній дезінформації через їхню відданість свободі слова та відкритим інформаційним середовищам. Натомість авторитарні режими, такі як російський, вправно використовують цифрові інструменти для тоталітарних цілей, застосовуючи іноземну пропаганду для ослаблення систем охорони здоров'я та політичної стабільності інших країн.

Дослідження дезінформації про COVID-19 у Twitter у Південній Африці [91] показало, що трансформерні моделі, зокрема ELECTRA та LAMBERT, досягли високих результатів, причому найкраща міра F1 склала 0,899. Однак дослідження також висвітлило труднощі адаптації цих моделей до різноманітних мовних і культурних контекстів. Аналогічно, аналіз дезінформації щодо постковідного синдрому та повторних заражень виявив, що DeBERTa та XLNet досягли найвищої точності класифікації, досягнувши міри F1 94,96% [92], що підтверджує ефективність механізмів уваги у класифікації текстів. Інше дослідження [92] застосувало BerTURK, модель для виявлення дезінформації, адаптовану до турецької мови, отримавши точність 98,5%, що підкреслює ефективність трансформерних архітектур у морфологічно складних мовах. Налаштовані моделі BERT також були застосовані до конкретних категорій дезінформації, зокрема пов'язаної з COVID-19 та часником, де BERTweet-large перевершив інші моделі з точністю 91,1% і мірою F1 0,894 [94], що підкреслює значущість спеціалізованого навчання моделей.

Архітектури глибокого навчання також широко використовуються для аналізу структурних та часових характеристик дезінформації. Такі техніки, як CNN, Bi-LSTM та CNN+Bi-GRU, виявилися особливо ефективними в задачах класифікації тексту. Дослідження [95] , що використовувало ансамбль CNN+Bi-GRU та Bi-LSTM на наборі даних CovidMis20, досягло високих показників точності у 92,23% і 90,56% відповідно, демонструючи надійність

моделей глибокого навчання у виявленні медичної дезінформації. Інше дослідження [96] протестувало кілька підходів глибокого навчання, зокрема LSTM та MC-CNN, на наборі даних CoAID, причому MC-CNN продемонстрував найвищу точність 99,57%. Гібридний підхід [97], що поєднує CNN та BiLSTM із механізмом самонавчання, покращив виявлення дезінформації, досягнувши 98,15% точності при навчанні з 80% розмічених даних, що підтверджує переваги гібридного та напівконтрольованого навчання. Крім того, оцінка різних наборів даних із використанням ембедингів, таких як Sentence-BERT, показала, що моделі глибокого навчання покращують класифікаційну ефективність і знижують упередженість [98], що підкреслює важливість контекстуальних текстових представлень у аналізі дезінформації.

Низка досліджень запропонували класичні методи, такі як LR, SVM, NB та RF для виявлення дезінформації про COVID-19. У проєкті з аналізу дезінформації у Twitter мовою суахілі метод SVM досяг найвищої точності 83,67%, а поліноміальна модель NB продемонструвала найвищу точність 92,00% [99], що свідчить про конкурентоспроможність класичних методів у певних контекстах. Інше дослідження [100] об’єднало машинне навчання з аналізом соціальних мереж у системі COS2, яка мапувала вузли дезінформації та шляхи її поширення, підкреслюючи потенціал інтеграції обчислювальних методів із мережею аналізу для відстеження та боротьби з дезінформацією.

Гібридні та напівконтрольовані підходи також досліджувалися для покращення виявлення дезінформації. Гібридна модель, що поєднує CNN та Bi-LSTM, працювала як з розміченими, так і з нерозміченими даними, перевершуючи традиційні методи [97]. Аналогічно, моделі ембедингу, такі як Sentence-BERT, покращили продуктивність класифікації [98], демонструючи їхню здатність уточнювати текстові представлення та підвищувати точність моделей. Ці підходи підкреслюють важливість комбінування різних технік для розробки більш надійних систем виявлення дезінформації.

Таким чином, машинне навчання стало критично важливим інструментом у боротьбі з дезінформацією про COVID-19. Трансформерні моделі демонструють високу точність у завданнях класифікації текстів, архітектури глибокого навчання успішно аналізують текстові та часові структури, а традиційні класифікатори залишаються корисними для певних мовних контекстів. Гібридні та напівконтрольовані підходи забезпечують додаткову адаптивність, роблячи їх перспективними рішеннями для виявлення дезінформації. Інтеграція цих різноманітних методологій створює комплексну стратегію для протидії поширенню неправдивої інформації та захисту громадського здоров'я у цифровому середовищі.

Повномасштабна війна росії проти України також спровокувала ряд досліджень та аналізу дезінформації та пропаганди у контексті інформаційно-психологічних операцій росії, оскільки інтенсивність інформаційних маніпуляцій дозволяла збирати набори даних та аналізувати техніки та стратегії пропаганди, використовуючи ШІ. Автори [101] опублікували набір даних VoynaSlov, що складається з понад 38 мільйонів публікацій у соціальних мережах від російських медіа, та використали різні моделі обробки природної мови, зокрема структуровані тематичні моделі, контекстуалізовані нейронні тематичні моделі та попередньо навчені великі мовні моделі, для аналізу стратегій пропаганди по різних типах медіа та періодах часу. Їхні результати виявили значні відмінності в тактиках маніпулювання в залежності від типу медіа, платформи та контексту війни.

Стаття [102] досліджує теми та настрої, що виражаються користувачами україномовного Telegram під час перших шести місяців війни Росії проти України, використовуючи техніки машинного навчання для аналізу даних з соціальних мереж. Дослідження реалізує тематичне моделювання за допомогою факторізації невід'ємної матриці з розходженням Кульбака-Лейблера, та аналіз настроїв з використанням попередньо навчених моделей для категоризації тем і емоційних тонів повідомлень. Стаття аналізує пропаганду, керовану ботами росії, та контрнаративи України в соціальних

мережах під час ключових етапів війни росії проти України 2022 року. Використовуючи TweetBERT для тематичного моделювання та інтегруючи фреймворк BEND з теорією моральних основ, дослідження вивчає, як боти маніпулювали наративами, щоб виправдати дії росії та протистояти НАТО, в той час як Україна використовувала подібні тактики для просування солідарності та стійкості.

Стаття [103] представляє метод OLTW-TEC, передовий підхід машинного навчання для виявлення дезінформації в україномовному контенті за допомогою ансамблю текстових класифікаторів та ковзного вікна для динамічного онлайн-навчання. Цей метод поєднує кілька класифікаторів, щоб адаптуватися до змін у даних, забезпечуючи високу точність та актуальність в реальному часі. Зазначеним обмеженням є висока вимога до обчислювальних ресурсів, що може обмежувати масштабованість, особливо для великих або ресурсозатратних застосувань.

1.3. Інформаційні технології аналізу дезінформації та пропаганди

Перспективним у контексті відслідковування та аналізу дезінформації та пропаганди є розробка комплексних технологічних фреймворків, які комбінуючи різні технології ШІ, зокрема машинного навчання допомагають з таким задачами.

Фреймворк представлений в статті [104], зосереджений на оцінці людиноорієнтованих пояснень ШІ в контексті спільної роботи людини та ШІ для перевірки фактів, зокрема для виявлення дезінформації. Моделі ШІ, залучені до цього завдання, включають текстові моделі обробки природної мови для перевірки фактів. Вони призначені для виявлення та пояснення потенційно оманливого контенту, такого як емоційна мова або маніпулятивне формулювання. Інформаційна технологія також містить “Інформаційну панель перевірки новин”, який демонструє новинні статті та пояснення до них,

створені ШІ. Ефективність системи тестується краудсорсингом для оцінки впливу на здатність користувачів розрізняти фальшиву інформацію.

Фреймворк для боротьби з COVID-19 інфодемією [105] передбачає багатофазний процес, починаючи з етапу збору даних із джерел, таких як соціальні медіа, новинні портали та офіційні урядові канали. Дослідження мало п’ять фаз: етап збору даних, побудови моделі та вибору методів, вдосконалення моделі, перевірки моделі та впровадження моделі. Запропонований фреймворк має потенціал для інтеграції в національну систему зберігання медичних даних Малайзійське сховище медичних даних.

Концептуальний фреймворк для виявлення фейкових новин, представлений у дослідженні [106], використовує трирівневий підхід. Перший рівень включає в себе ідентифікацію джерел фейкових новин, які класифікуються на традиційні медіа, соціальні медіа та пошукові системи. Другий рівень зосереджений на класифікації контенту на категорії, такі як новинний контент, соціальний контекст і політичний контекст. Цей етап включає використання методів машинного навчання для класифікації ознак даних і виявлення шаблонів, пов’язаних з фейковими новинами. Останній рівень пропонує стратегії для боротьби з виявленими фейковими новинами, які можуть бути вбудованими в веб-браузери, соціальні медіа та платформи пошукових систем для автоматичного фільтрування та реагування на дезінформацію.

Представлений у статті [107] фреймворк зосереджений на мультимодальному підході до виявлення фейкових новин, інтегруючи як текстові, так і візуальні характеристики для підвищення продуктивності систем виявлення. Фреймворк наголошує на інженерії ознак, де релевантні характеристики тексту (такі як лінгвістичний тон, емоції та заперечення) і зображень (включаючи маніпуляції з візуальним контентом і емоції, виражені на зображеннях) витягуються та оптимізуються за допомогою алгоритму M-JAYA. Цей метод допомагає вибрати найбільш релевантні ознаки, підвищуючи точність моделей машинного навчання, що використовуються

для виявлення фейкових новин. CNN та LSTM використовуються для обробки та аналізу мультимодальних даних, що дозволяє здійснювати більш надійне виявлення дезінформації порівняно з традиційними одномодальними методами. Фреймворк інтегрує принципи пояснюваного ШІ, щоб забезпечити прозорість і зрозумілість у процесі прийняття рішень, роблячи модель більш надійною.

Фреймворк, описаний у статті [108], пропонує модель на основі трансформерів, AraBERT, для виявлення технік пропаганди в арабських текстах у соціальних мережах, зокрема в твітах. Модель орієнтована на різні методи пропаганди, включаючи “чорно-біле мислення”, маніпулятивну мову, образи, штучні аргументи та інші. Підхід починається з обробки даних, після чого відбувається розробка та оптимізація моделі AraBERT за допомогою оптимізатора ADAM. Модель налаштовується та оцінюється за її результатами, використовуючи міру мікро-F1. Модель AraBERT використовує архітектуру трансформера, натреновану на великому корпусі арабських текстів. Ця архітектура дозволяє розуміти та класифікувати складний нюансований пропагандистський контент. Використання передавального навчання з попередньо навченими ембедингами покращує здатність моделі обробляти різноманітні вхідні дані та узагальнювати нові, невидимі дані.

Фреймворк, описаний у статті [109], пропонує методологію для виявлення фейкових облич за допомогою методів машинного навчання. Процес складається з кількох етапів, починаючи з збору реальних і фейкових наборів зображень, після чого застосовуються техніки попередньої обробки даних, такі як зміна розміру, обрізка та доповнення даних для стандартизації зображень. Наступний етап включає розподіл даних на тренувальний та тестовий набори для полегшення оцінки моделі. Виконується витягнення ознак для ізоляції лицьових областей на зображеннях, що зменшує обчислювальне навантаження, зберігаючи точність. Різні моделі машинного навчання, включаючи LR, SVM, DT, RF та NB, використовуються для класифікації зображень як реальних або фейкових. Для

оцінки продуктивності моделей використовуються такі метрики, як точність, AUC, прецизійність і повнота.

Представлений у статті [110] фреймворк пропонує підхід до класифікації фейкових новин арабською мовою, використовуючи ансамбль CNN і LSTM, поєднаний з ембедингами, такими як ELMo, для контекстуального представлення. Гібридна модель використовує техніки глибокого навчання для покращення виявлення фейкових новин у текстах арабською мовою. Вона інтегрує кілька передових архітектур глибокого навчання, таких як EfficientNetB4, Inception, Xception, ResNet та ConvLSTM, для ефективної обробки та класифікації текстів. Модель також включає техніку LIME, що забезпечує пояснюваність передбачень, допомагаючи користувачам зрозуміти процес прийняття рішень моделі. Експериментальні результати показують, що ансамблева модель CNN-LSTM, вдосконалена з ембедингами ELMo, перевершує попередні методи, досягаючи точності 98,42% з високою прецизійністю, повнотою і мірою F1, що робить її надійним рішенням для виявлення фейкових новин арабською мовою.

Фреймворк, розроблений UbiLab Університету Ватерлу [111], зосереджений на проектуванні та розробці системи реального часу для аналізу дезінформації, UbiLab Misinformation Analysis System (U-MAS), метою якої є виявлення та аналіз дезінформації щодо здоров’я в соціальних мережах, зокрема в Twitter. Система складається з п’яти основних компонентів: фреймворка для збору даних, тематичної моделі прихованого розподілу Діріхле (LDA), аналізатора сентименту, моделі класифікації дезінформації та розгортання в Elastic Cloud для індексації та візуалізації. U-MAS використовує інструменти на базі Python для збору та попередньої обробки даних з Twitter, а потім застосовує моделі машинного навчання для класифікації та аналізу дезінформації, пов’язаної з темами здоров’я. Аналізатор сентименту використовує моделі, такі як VADER і BERT, для оцінки емоційного тону контенту, а модель тем LDA категоризує дані на відповідні теми. Цей фреймворк має значний потенціал для застосувань у громадському здоров’ї, допомагаючи моніторити та зменшувати дезінформацію, особливо під час криз, таких як пандемії.

Фреймворк розроблений лабораторією Лінкольна Массачусетського інституту технологій [112], зосереджений на автоматизації виявлення наративів дезінформації, впливових акаунтів та їхнього причинно-наслідкового впливу в соціальних мережах. Він поєднує методи обробки природної мови, машинного навчання, графової аналітики та техніки причинно-наслідкового підходу для виявлення і кількісної оцінки впливу кампаній з дезінформації. Система спочатку збирає дані з соціальних медіа платформ, таких як Twitter, а потім визначає відповідні наративи за допомогою тематичного моделювання. Акаунти, що беруть участь у цих наративах, класифікуються за допомогою напівконтрольованого ансамблевого класифікатора DT, який аналізує як поведінкові, так і контентні характеристики акаунтів. Графова аналітика використовується для побудови мережі наративів на основі ретвітів і взаємодій, а вплив кожного акаунта на поширення дезінформації оцінюється за допомогою причинно-наслідкового висновку. Це дозволяє виявити акаунти з великим впливом, які можуть не бути відзначені традиційними ознаками, такими як кількість активностей чи центральність.

1.4. Висновки до першого розділу

1. Проаналізовано теоретичні засади інформаційних маніпуляцій у контексті інформаційних воєн.

2. Розглянуто основні типи інформаційних розладів та неконвенційних воєн.

3. Розглянуто застосування технологій ШІ для протидії дезінформації, інформаційним маніпуляціям, пропаганді та фейкам.

4. Проаналізовано методи машинного та глибокого навчання для протидії дезінформації, інформаційним маніпуляціям, пропаганді та фейкам.

5. Розглянуто інформаційні технології аналізу дезінформації та пропаганди.

Основні результати, представлені в цьому розділі, опубліковані в [18], [19], [23], [24], [25], [26], [27], [28], [29].

РОЗДІЛ 2. ДОСЛІДЖЕННЯ МЕТОДІВ МАШИННОГО НАВЧАННЯ ДЛЯ АНАЛІЗУ ТЕКСТОВОЇ ДЕЗІНФОРМАЦІЇ ТА ПРОПАГАНДИ

2.1. Аналіз продуктивності методів машинного навчання для аналізу текстової дезінформації та пропаганди

2.1.1. Огляд класичних методів машинного навчання

У цьому підрозділі здійснено порівняльний аналіз продуктивності найбільш вживаних методів класифікації російської пропаганди в соціальній мережі X (Twitter). Зібрані дані були вручну розмічені відповідно до приналежності до двох класів: російська пропаганда та проукраїнські\нейтральні повідомлення. На цьому наборі даних проведено тестування продуктивності моделей. Методи, обрані для аналізу цього набору даних, включали LR, NB, KNN та SVM.

1. Логістична регресія (LR)

LR – це статистичний метод машинного навчання, який використовується для моделювання ймовірності бінарного результату, що робить його особливо релевантним для завдань класифікації [113]. У контексті дисертаційного дослідження LR використовується для прогнозування ймовірності того, що твіт буде класифікований як проросійська пропаганда або як нейтральний/проукраїнський на основі його текстових характеристик. У рамках дисертаційного дослідження розроблено модель на основі методу LR.

2. Наївний Баєс

Наївний класифікатор Баєса (NB) – це ймовірнісна модель машинного навчання, заснована на теоремі Баєса, яка особливо сприятлива для завдань класифікації тексту завдяки своїй простоті [114]. У контексті дисертаційного дослідження NB використовується для розрізнення твітів, які

характеризуються як проросійська пропаганда, від тих, які є нейтральними або проукраїнськими на основі їхніх текстових атрибутів.

Це дослідження пропонує обчислювально прийнятне та масштабоване рішення на основі моделі NB. Його ймовірнісна основа забезпечує певну міру впевненості щодо класифікацій, а простота забезпечує легкість впровадження та інтерпретації. Незважаючи на припущення про незалежність функцій, модель часто демонструє високу продуктивність у завданнях класифікації тексту, що робить її цінним інструментом для розпізнавання наративів пропаганди в твітах.

3. K-найближчих сусідів (KNN)

KNN – це непараметрична техніка навчання на основі екземплярів, яка класифікує об’єкти на основі більшості класів їхніх K найближчих сусідів у просторі ознак [115]. У контексті дисертаційного дослідження модель KNN використовується для класифікації твітів як проросійська пропаганда або нейтральних/проукраїнських на основі їх текстових репрезентацій шляхом розгляду класифікації найбільш схожих твітів у навчальному наборі даних.

Для поставленого завдання мктод KNN пропонує інтуїтивно зрозумілий підхід до класифікації, використовуючи притаманну структуру та шаблони в даних. Його природа на основі екземплярів гарантує, що модель залишається адаптивною та може включати нові дані без значної підготовки. Однак вибір відповідної метрики відстані та найкращого значення K є ключовими для його успіху. У цьому дослідженні модель KNN є цінним інструментом для розпізнавання пропагандистських механізмів у твітах шляхом розгляду класифікації їхніх найбільш схожих аналогів у навчальних даних.

Модель KNN було оптимізовано за гіперпараметрами, під час якої визначено, що параметр n_neighbors має найкраще значення “19” для досягнення найкращої продуктивності.

4. Метод опорних векторів (Support Vector Machine (SVM))

Застосування моделі SVM визначило найкращу гіперплощину, яка найкраще розділяє дані на окремі класи, забезпечуючи максимальний запас між ними [116].

У контексті цього дослідження математична надійність і здатність SVM обробляти дані великого розміру робить його вдалим вибором для розпізнавання складних шаблонів у текстових даних, допомагаючи таким чином точно класифікувати твіти як проросійські або нейтральні/проукраїнські повідомлення.

Модель SVM було створено за допомогою ядра радіальної базисної функції (RBF), позначеного як SVC (kernel='rbf'). Цей вибір був мотивований здатністю SVM, особливо з ядром RBF, вміло обробляти масиви даних великого розміру, що свідчить про його важливість у завданнях класифікації тексту.

2.1.2. Попередня обробка даних

Для оцінки продуктивності моделей машинного навчання використано набір даних, опис якого наведено в Додатку Г. Щоб підвищити точність моделей та зменшити обчислювальні витрати, необроблені дані Twitter були перетворені у формат, який можна обробляти за допомогою методів машинного навчання. Попередня обробка тексту включає:

- Видалення стоп-слів. Для видалення цих слів із тексту було використано список стоп-слів, що охоплює повну колекцію поширених англійських слів, таких як артиклі, прийменники та сполучники, оскільки вони мають мінімальну значущість для аналізу.
- Заміна згадок. Згадки в Twitter, позначені ідентифікаторами користувачів, що починаються з “@”, були видалені, а хештеги замінено терміном “HASH”. Цей крок важливий для анонімізації інформації користувача та стандартизації тексту.
- Нормалізація регістру. Весь вміст твіту було перетворено на малі літери, щоб гарантувати аналіз без урахування регістру та гарантувати, що слова в різних регістрах інтерпретуються як ідентичні.
- Видалення HTML-тегів і URL-адрес. HTML-теги, які часто зустрічаються в даних, зібраних із веб-сайтів, було видалено. Це включало

ідентифікацію пар тегів (“<” і “>”) і викорінення вмісту, укладеного в них. Крім того, URL-адреси (http/https) також видалено.

- Видалення квадратних дужок. Квадратні дужки, які зазвичай вказують на метадані або цитати, були видалені з тексту.
- Видалення спеціальних символів і знаків пунктуації. Для видалення цих елементів із тексту твіту використовувався попередньо визначений список спеціальних символів і знаків пунктуації, таким чином зберігаючи лише абетко-цифрові символи та пробіли.
- Збереження хештегів. Незважаючи на підготовчі етапи попередньої обробки, хештеги були збережені, оскільки вони поширені в тексті соціальних медіа та інкапсулюють цінну інформацію.
- Визначення кореня. Метод Snowball Stemmer [117], налаштований для англійської мови, використовувався для спрощення відмінюваних або похідних слів до їх основи або форми кореня. Цей процес не тільки забезпечив послідовний аналіз, але й мінімізував розмір словникового запасу.
- Видалення зайвих пробілів. Будь-які зайві пробіли, випадково додані під час етапів попередньої обробки, було видалено, забезпечуючи наявність лише одного пробілу між словами.
- Токенізація. Уточнений текст було сегментовано на окремі слова, позначені як токени, що полегшує більш детальний аналіз, дозволяючи моделі обробляти кожне слово незалежно.
- Усунення дисбалансу класів за допомогою передискретизації. Дисбаланс класів було виправлено шляхом збільшення кількості екземплярів у меншому класі шляхом створення нових екземплярів, або дублікатів, або синтетичних прикладів, доки представлення меншого класу не було збалансованим або менш викривленим відносно класу більшості. Зокрема, було розроблено стратегію передискретизації з використанням класу RandomOverSampler із параметром sampling_strategy, позначеним як “меншість”. Це вказує на бажання вирівняти клас меншості, щоб відповідати кількості

екземплярів класу більшості, тим самим пом’якшуючи упередженість і посилюючи можливості узагальнення навченої моделі.

2.1.3. Результати аналізу продуктивності класифікаційних моделей для аналізу текстової дезінформації та пропаганди

Оцінку продуктивності моделей було проведено на розміченому наборі даних. Ця оцінка була критичною для визначення адаптивності та застосовуваності моделей у різних сценаріях, враховуючи мінливість розміру набору даних. Для цього дослідження було відібрано п’ять вибірок даних від 1000 до 5000 твітів. Ці діапазони було обрано, щоб надати розуміння того, як продуктивність моделей змінюється з різними обсягами даних, тим самим перевіряючи їх надійність і масштабованість. Набори даних були розділені на дві вибірки для забезпечення комплексної оцінки: тренувальна вибірка, яка складала 80% відібраних даних, і тестова вибірка, яка складала решту 20%. Ці дії мали на меті створити надійну модель, навчивши її на більшості даних, зарезервувавши менший, окремий набір для тестування, таким чином запобігаючи перенавчанню.

Продуктивність моделі LR наведено в таблиці 2.1. Продуктивність моделі NB наведено в таблиці 2.2. Продуктивність моделі KNN наведено в таблиці 2.3. Продуктивність моделі SVM наведено в таблиці 2.4. Продуктивність усіх моделей на найбільшій вибірці тренувальних даних з 5000 тисяч твітів наведено в таблиці 2.5.

Щоб уможливити порівняння кожної моделі машинного навчання, застосованої до набору даних, було графічно зображено показники їх точності. На рисунку 2.1 показано динаміку точності моделей для кожного розміру набору даних. При розмірі набору даних у 5000 твітів модель SVM перевершила інші моделі машинного навчання. На рисунку 2.2 показано точність різних моделей для найбільшої вибірки даних.

Таблиця 2.1 – Продуктивність моделі LR

Кількість твітів	Клас	Прецизійність	Повнота	Міра F1	Підтримка	Точність
1000	0	0.72	0.74	0.73	114	0.72
	1	0.72	0.74	0.73	121	
2000	0	0.77	0.78	0.78	240	0.78
	1	0.77	0.78	0.78	234	
3000	0	0.78	0.81	0.79	359	0.79
	1	0.78	0.81	0.79	345	
4000	0	0.79	0.77	0.78	461	0.78
	1	0.79	0.77	0.78	448	
5000	0	0.82	0.78	0.80	543	0.79
	1	0.82	0.78	0.80	593	

Таблиця 2.2 – Продуктивність моделі NB.

Кількість твітів	Клас	Прецизійність	Повнота	Міра F1	Підтримка	Точність
1000	0	0.80	0.78	0.79	106	0.77
	1	0.80	0.78	0.79	129	
2000	0	0.76	0.80	0.78	236	0.77
	1	0.76	0.80	0.78	238	
3000	0	0.75	0.77	0.76	342	0.75
	1	0.75	0.77	0.76	362	
4000	0	0.74	0.82	0.77	462	0.77
	1	0.74	0.82	0.77	447	
5000	0	0.74	0.83	0.78	566	0.77
	1	0.74	0.83	0.78	570	

Таблиця 2.3 – Продуктивність моделі KNN.

Кількість твітів	Клас	Прецизійність	Повнота	Міра F1	Підтримка	Точність
1000	0	0.69	0.79	0.74	106	0.69
	1	0.69	0.79	0.74	129	
2000	0	0.72	0.66	0.69	236	0.70
	1	0.72	0.66	0.69	238	
3000	0	0.73	0.71	0.72	342	0.72
	1	0.73	0.71	0.72	362	
4000	0	0.71	0.75	0.73	462	0.72
	1	0.71	0.75	0.73	447	
5000	0	0.71	0.75	0.73	566	0.72
	1	0.71	0.75	0.73	570	

Таблиця 2.4 – Продуктивність моделі SVM.

Кількість твітів	Клас	Прецизійність	Повнота	Міра F1	Підтримка	Точність
1000	0	0.72	0.92	0.81	122	0.79
	1	0.72	0.92	0.81	113	
2000	0	0.76	0.84	0.80	234	0.78
	1	0.76	0.84	0.80	240	
3000	0	0.76	0.87	0.81	357	0.80
	1	0.76	0.87	0.81	347	
4000	0	0.79	0.81	0.80	457	0.80
	1	0.79	0.81	0.80	452	
5000	0	0.80	0.84	0.82	569	0.82
	1	0.80	0.84	0.82	567	

Таблиця 2.5 – Продуктивність моделей для набору даних у 5000 твітів.

Кількість твітів	Клас	Прецизійність	Повнота	Міра F1	Підтримка	Точність
LR	0	0.82	0.78	0.80	543	0.79
	1	0.82	0.78	0.80	593	
NB	0	0.74	0.83	0.78	566	0.77
	1	0.74	0.83	0.78	570	
KNN	0	0.71	0.75	0.73	566	0.72
	1	0.71	0.75	0.73	570	
SVM	0	0.80	0.84	0.82	569	0.82
	1	0.80	0.84	0.82	567	

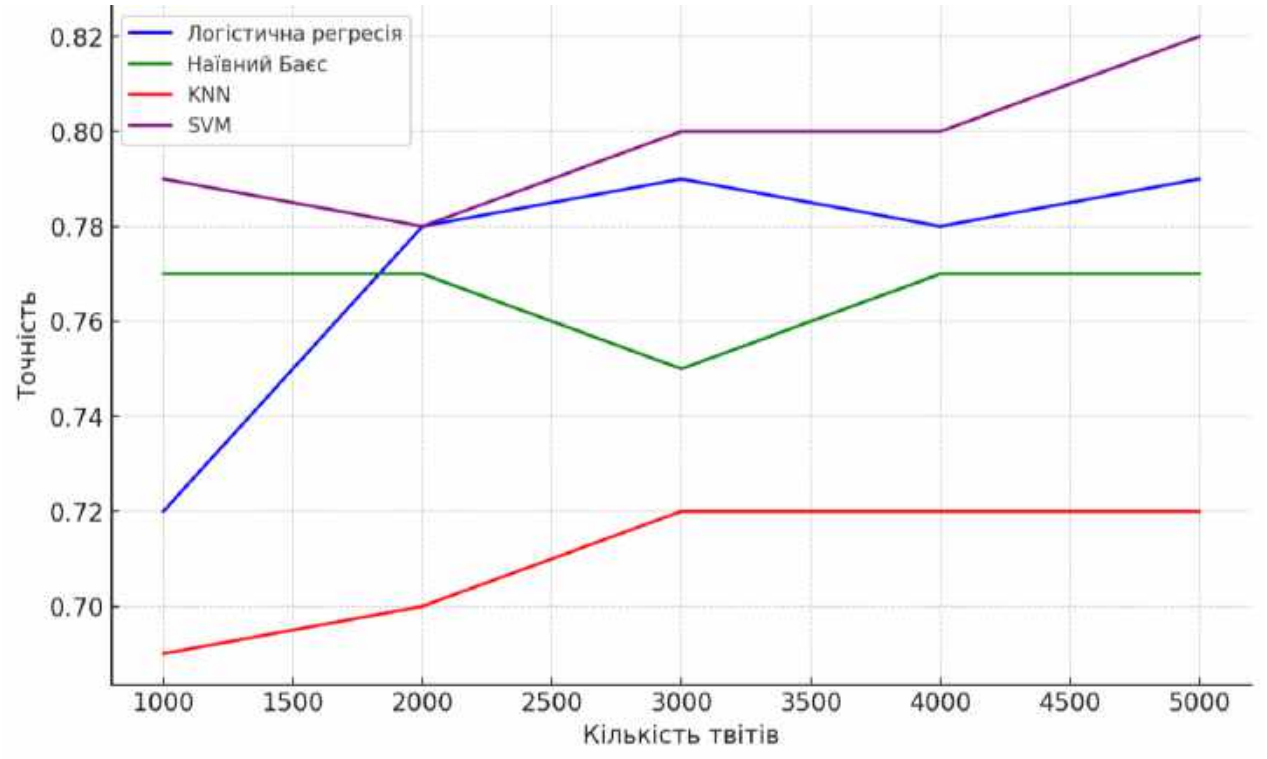

Рисунок 2.1 – Динаміка точності моделей для кожного розміру набору даних

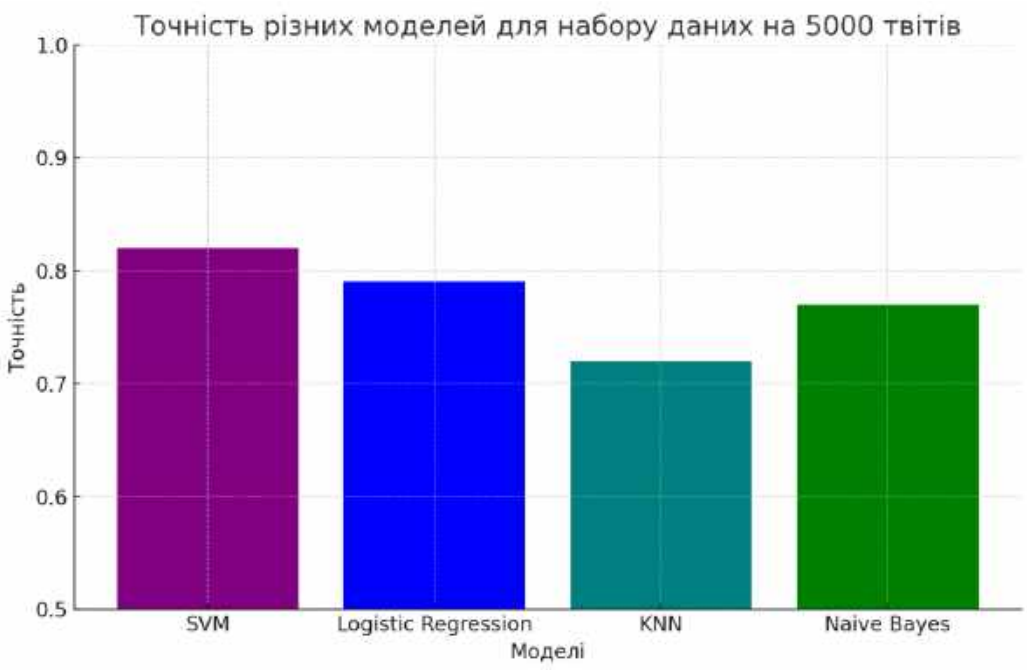

Рисунок 2.2 – Точність різних моделей для найбільшої вибірки даних

Оцінка різних моделей машинного навчання надає важливі уявлення про їх продуктивність і адаптивність. Модель LR з точністю 0.79 продемонструвала свою робастність у виконанні завдань бінарної класифікації текстів. Її збалансовані показники прецизійності та повноти свідчать про її здатність відрізняти проросійську пропаганду від проукраїнських наративів, що робить її життєздатним вибором для завдань, які потребують тонкого розрізнення нюансів у текстових даних.

Модель NB, відома своїми ймовірнісними засадами, досягла точності 0.77. Враховуючи її продуктивність, можна зробити висновок, що витягнуті з твітів ознаки були досить незалежними, що дозволило моделі ефективно оцінити ймовірності, що лежать в основі даних. Її збалансована повнота також підкреслює здатність правильно ідентифікувати твіти з пропагандистським змістом.

Модель KNN, хоч і досягла точності 0.72, могла бути уражена високою розмірністю, властивою текстовим даним. Внутрішні труднощі з обчисленням відстаней у високовимірних просторах могли вплинути на її продуктивність.

Найкраща продуктивність моделі SVM з точністю 0.82 підкреслює її здатність до вирішення завдань класифікації дезінформаційних текстів та

пропаганди. Вибір ядра RBF, ймовірно, дозволив моделі чітко розрізняти складні межі у просторі ознак, ефективно відокремлюючи проросійську пропаганду від проукраїнських наративів.

У проведеному комплексному аналізі виявлено помітну тенденцію, коли показники продуктивності, включаючи прецизійність, повноту, міру F1 та підтримку, демонстрували поліпшення відповідно до збільшення розміру набору даних. Це спостереження підкреслює ключову роль великих обсягів даних у підвищенні продуктивності моделей машинного навчання. Такі великі набори даних можуть охоплювати ширший спектр сценаріїв, варіацій та потенційних крайніх випадків, що дозволяє моделі отримати багатше різноманіття шаблонів. Це, у свою чергу, дозволяє моделі краще розпізнавати складніші та загальніші шаблони, підвищуючи її прогностичну точність.

Великі набори даних діють як захисний механізм проти перенавчання, сприяючи створенню моделей, що краще узагальнюють на нові, незнайомі дані. Серед усіх оцінених методів було встановлено, що SVM постійно показувала кращі результати. Її продуктивність була кращою на всіх розмірах вибірок, що підкреслює її універсальність і надійність. На противагу, хоч вони й мали свої унікальні переваги, методи, як-от NB і KNN, показали відносно стриману продуктивність, особливо за показником точності.

Модель SVM виявилася найбільш продуктивною, демонструючи виражені покращення у всіх показниках продуктивності в міру зростання розміру набору даних. Це спостереження акцентує здатність SVM розпізнавати й адаптуватися до складних шаблонів даних, забезпечуючи кращу узагальнюваність у міру розширення обсягу даних. Навпаки, модель KNN, заснована на принципі обчислення відстаней між точками даних, демонструвала більш стриману відповідь на збільшення обсягу даних. Її покращення продуктивності, особливо за показниками точності та міри F1, були менш вираженими порівняно з іншими методами, що натякає на потенційні проблеми масштабування, властиві методам, заснованим на обчисленні відстаней, при роботі з великими обсягами даних.

2.2. Застосування класифікатору SVM для аналізу текстової дезінформації та пропаганди на інших мовних структурах

2.2.1. Аналіз продуктивності методу SVM на мовах з іншими структурами

У попередньому підрозділі модель SVM показала свою найвищу продуктивність для завдань класифікації тексту. У цьому підрозділі проведено аналіз продуктивності моделі на китайськомовному наборі даних для перевірки її на мові яка має інші властивості та структуру.

Модель SVM адаптовано для класифікації дезінформації на платформі Weibo, популярному китайському мікроблозі. Для застосування моделі класифікації дезінформації було використано набір даних CHECKED [118]. Опис набору даних наведено в Додатку Г. Ця всебічна колекція мікроблогів із китайської соціальної платформи Weibo була створена спеціально для виявлення дезінформації про COVID-19. Це перший у своєму роді набір даних, який фокусується на контенті китайською мовою, пов’язаному з пандемією COVID-19.

Пости Weibo часто містять текст, зображення та відео, тому процес вилучення ознак необхідно було налаштувати для обробки такого типу контенту. У контексті Weibo простір ознак може бути набором ознак, вилучених із тексту, зображень і відео постів. Для текстових даних ознаки можуть бути вилучені за допомогою значень частоти терміну зворотної частоти документа (TF-IDF), n-грам, індексів настроїв тощо.

Класи для прогнозування є бінарними, тобто “фейкові новини” або “реальні новини”. Модель була навчена на розміченому наборі даних постів Weibo для пошуку найкращої гіперплощини, яка розділяє пости з фейковими новинами від реальних у просторі ознак. Після навчання модель здатна класифікувати нові пости в категорії “фейкові” або “реальні”.

2.2.2. Процес налаштування моделі SVM

Процес налаштування моделі включав кілька ключових кроків для забезпечення найкращої продуктивності. Спочатку були протестовані різні моделі шляхом зміни лише функції ядра, що є важливим параметром у моделях SVM. Серед різних функцій ядра, що тестувалися, лінійне ядро виявилося найбільш ефективним, демонструючи найкращі результати у попередніх випробуваннях.

Потім було проведено всебічний пошук найкращих значень решти параметрів, зокрема параметра регуляризації C і коефіцієнта ядра γ, використовуючи метод сіткового пошуку з десятикратною перехресною валідацією. Пошук сітки проводився на діапазоні значень, зі збільшенням порядку величини (*10) для кожного параметра. Цей підхід забезпечив ретельне дослідження простору параметрів і, зрештою, визначив найбільш підходящі значення параметрів для моделі.

Перед застосуванням моделі було важливо вибрати відповідну техніку векторизації для перетворення текстових даних у числовий формат, придатний для методів машинного навчання. Після ретельного розгляду було обрано TF-IDF векторизатор із двох основних причин. По-перше, TF-IDF векторизатор легко інтегрується в бібліотеку Scikit-learn, що усуває потребу в додатковій постобробці. По-друге, серед різних векторизаторів, які не базуються на глибокому навчанні, TF-IDF векторизатор продемонстрував найкращі результати на загальній моделі SVM щодо міри F1, навіть без налаштування параметрів. Ця метрика є важливою, оскільки забезпечує збалансовану оцінку точності та повноти моделі.

Процес налаштування моделі включав систематичний підхід до вибору найбільш підходящої функції ядра, проведення сіткового пошуку для визначення найкращих значень параметрів регуляризації та γ, а також вибір найефективнішої техніки векторизації. Вибір лінійного ядра, всебічний сітковий пошук і TF-IDF векторизатор спільно сприяли розробці надійної та ефективної моделі для вирішення поставленого завдання.

2.2.3. Результати продуктивності моделі SVM для класифікації текстової дезінформації

Продуктивність моделі оцінювалася за допомогою кількох метрик та візуалізацій, включаючи таблицю продуктивності моделі, матрицю невідповідностей, нормалізовану матрицю невідповідностей, графік кривої навчання та графік частоти хибнопозитивних результатів щодо порогів.

Продуктивність моделі була підсумована в таблиці з різними метриками, такими як точність, прецизійність, повнота, міра F1, специфічність, коефіцієнт кореляції Метьюза (MCC) та збалансована точність. Продуктивність моделі представлена в таблиці 2.6. Ці метрики всебічно оцінюють здатність моделі правильно класифікувати фейкові та реальні новини.

Матриця невідповідностей – це таблиця, яка показує кількість істинно позитивних, істинно негативних, хибно позитивних та хибно негативних прогнозів, зроблених моделлю, надає детальний огляд продуктивності моделі, демонструючи, скільки екземплярів кожного класу було правильно або неправильно передбачено. Матриця невідповідностей представлена на рисунку 2.3.

Нормалізована матриця невідповідностей подібна до звичайної матриці невідповідностей, але її значення нормалізовані за кількістю екземплярів у кожному класі. Вона забезпечує більш чітке уявлення про продуктивність моделі, показуючи частку екземплярів кожного класу, які були правильно або неправильно передбачені. Нормалізована матриця невідповідностей представлена на рисунку 2.4.

Графік кривої навчання показує продуктивність моделі на тренувальних та валідаційних даних як функцію кількості навчальної вибірки. Він допомагає визначити, чи уражена модель від перенавчання або недонавчання. Перенавчання відбувається, коли модель добре працює на навчальних даних, але показує погані результати на валідаційних даних. Недонавчання виникає, коли модель показує погані результати як на навчальних, так і на валідаційних даних. Крива навчання представлена на рисунку 2.5.

Таблиця 2.6 – Продуктивність моделі SVM

Метрика	Значення
Точніть	0.8907
Міра F1	0.8733
Повнота	0.8907
Прецизійність	0.9038
Специфічність	0.4458
MCC	0.6264
Збалансована точність	0.7229

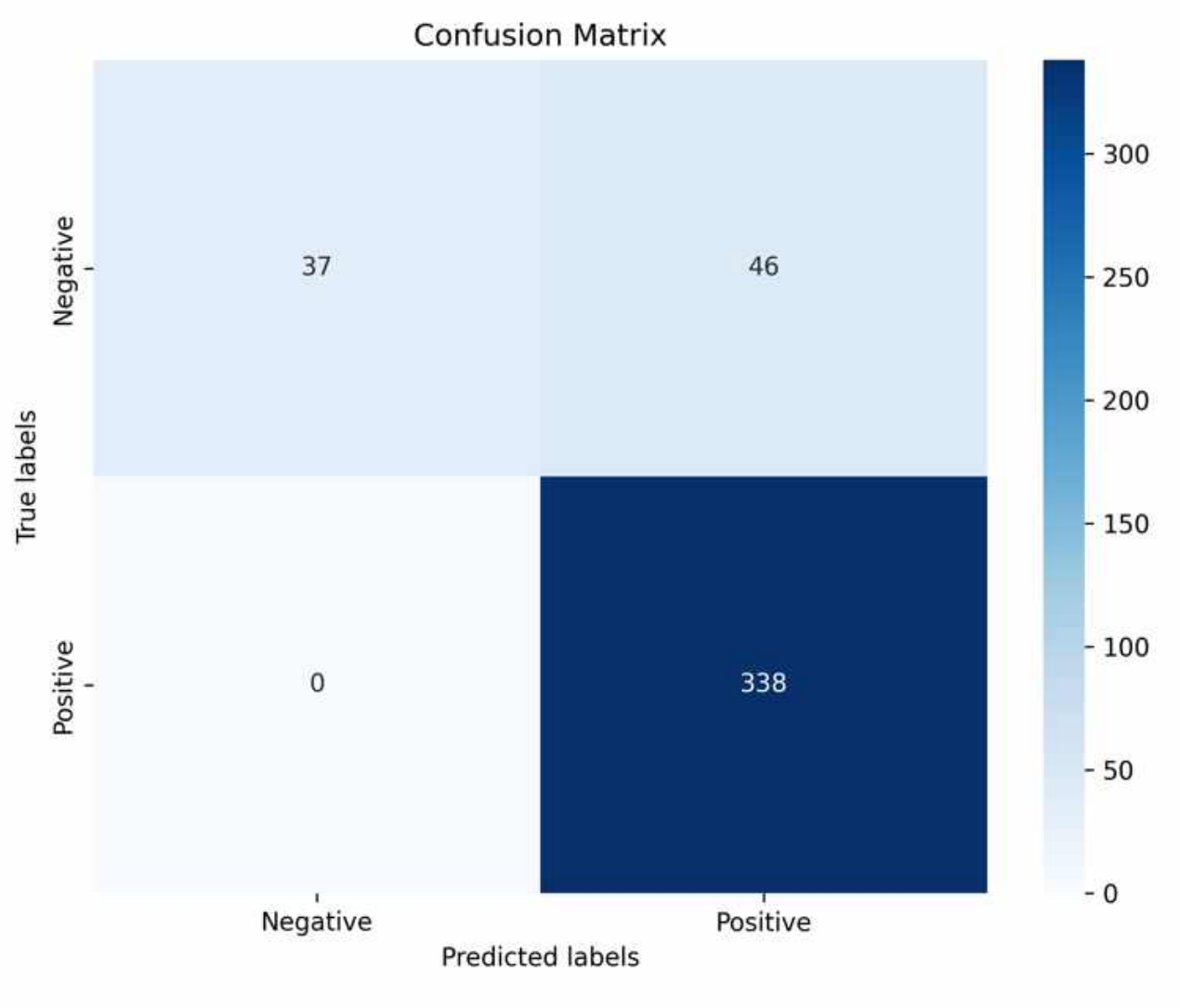

Рисунок 2.3 – Матриця невідповідностей

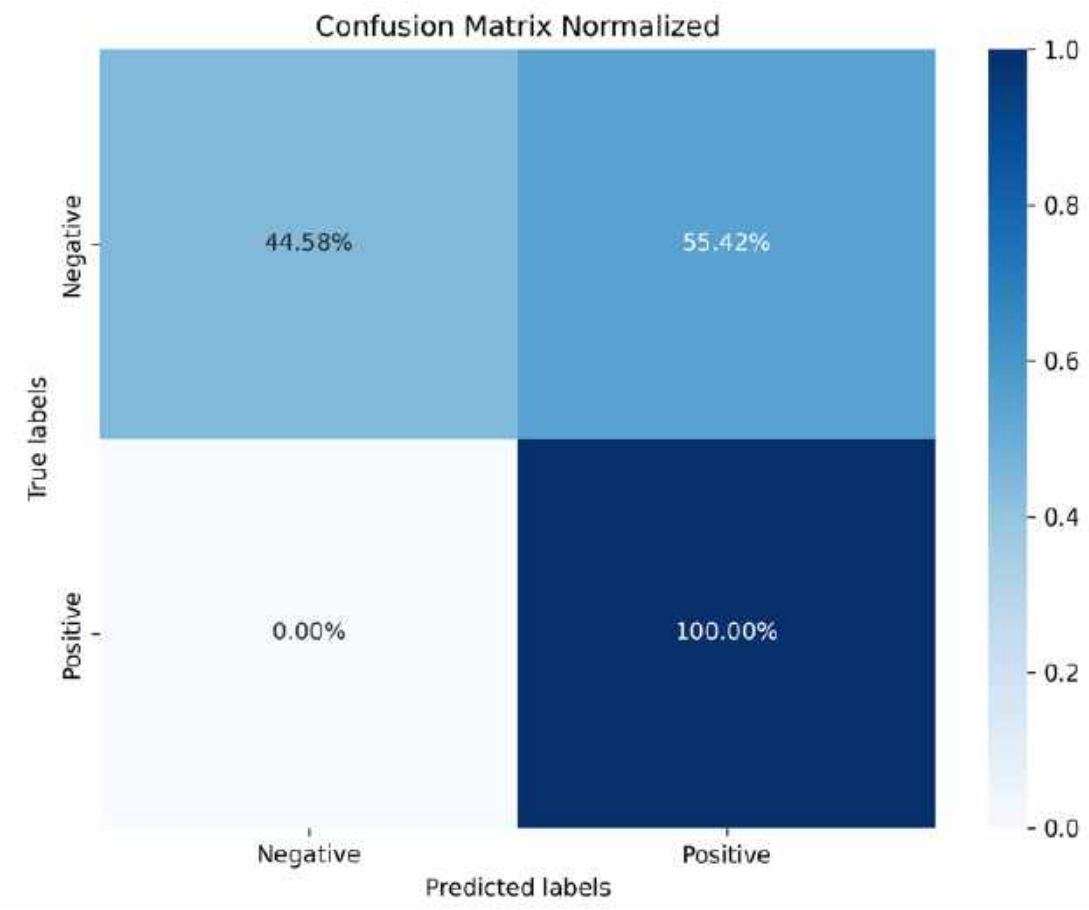

Рисунок 2.4 – Нормалізована матриця невідповідностей

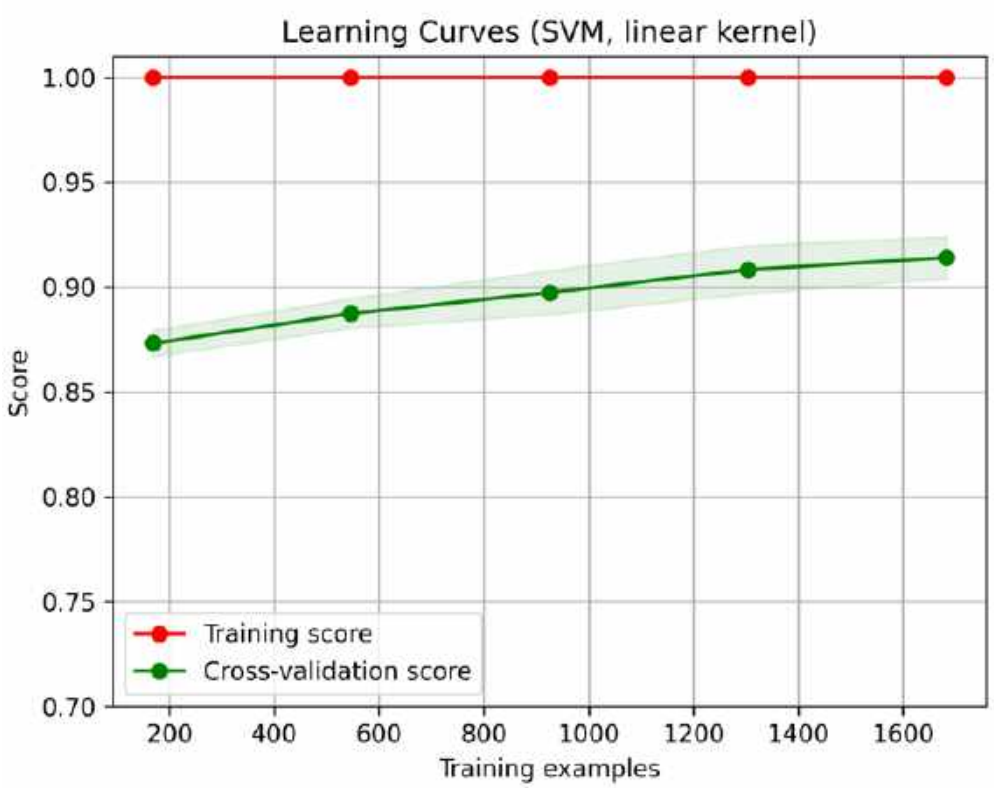

Рисунок 2.5 – Крива навчання

Графік показника хибно позитивних результатів залежно від порогових значень демонструє зміну показника хибно позитивних результатів залежно від порогового значення, яке використовується моделлю. Це допомагає визначити порогове значення, яке мінімізує показник хибно позитивних результатів, водночас максимізуючи показник істинно позитивних результатів. Графік показника хибно позитивних результатів залежно від порогових значень представлено на Рисунку 2.6.

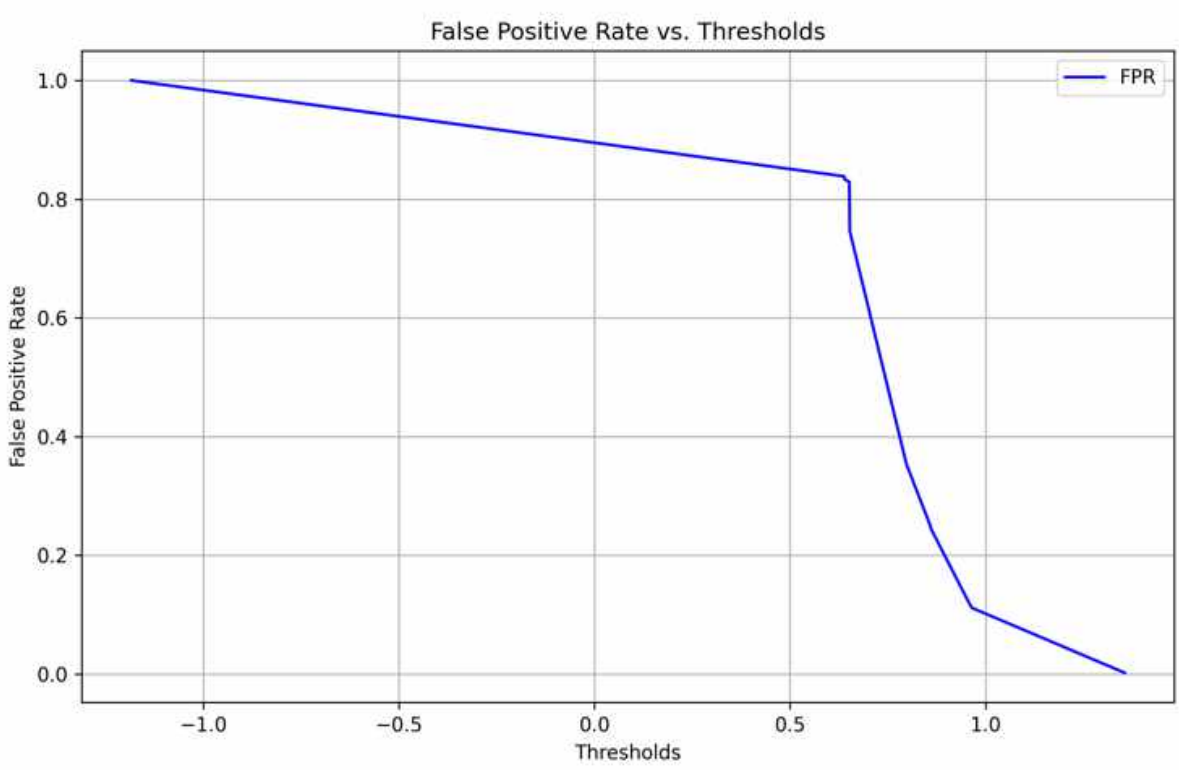

Рис. 2.6 – Графік показника хибнопозитивних результатів залежно від порогових значень

Результати свідчать про те, що модель SVM добре впоралася з класифікацією фейкових новин про COVID-19 на платформі Weibo. Модель досягла точності 89.07% і міри F1 87.33%. Матриця невідповідностей та нормалізована матриця невідповідностей надають детальний огляд продуктивності моделі, показуючи, що модель правильно класифікувала високу частку екземплярів кожного класу. Графік кривої навчання вказує на те, що модель не була перенавчена або недонавчена.

2.2.4. Обговорення результатів аналізу методу SVM для класифікації текстової дезінформації

Результати свідчать про відносно високу продуктивність моделі у класифікації фейкових новин про COVID-19. Точність моделі становить 89.07%, що означає, що модель правильно класифікувала 89.07% загальних випадків. Міра F1 становить 87.33%, що вказує на збалансовану продуктивність моделі за цими показниками. Повнота моделі становить 89.07%, що означає, що модель правильно виявила 89.07% усіх випадків фейкових новин. Прецизійність моделі становить 90.38%, що означає, що 90.38% випадків, які модель передбачила як фейкові новини, дійсно були дезінформацією.

Однак підтримка моделі складає лише 44.58%, що є відносно низьким показником. Підтримка – це здатність моделі правильно визначати реальні новини. Підтримка на рівні 44.58% означає, що модель правильно визначила лише 44.58% усіх реальних новин. Це свідчить про високу частоту хибно позитивних результатів, тобто модель неправильно класифікувала багато реальних новин як фейкові.

MCC становить 0.6264, що вважається помірною кореляцією. MCC вимірює якість бінарних класифікацій і варіюється від -1 до +1. Коефіцієнт +1 означає ідеальне передбачення, 0 вказує на випадкове передбачення, а -1 – на протилежне передбачення. MCC 0.6264 вказує на те, що модель має помірну кореляцію між спостережуваними та передбаченими бінарними класифікаціями.

Збалансована точність моделі становить 72.29%, що є середнім значенням повноти та підтримки. Збалансована точність свідчить про те, що модель має збалансовану продуктивність за повнотою та підтримкою, але все ж є простір для покращення.

Класифікація фейкових новин про COVID-19 на платформі Weibo має кілька суттєвих викликів. Один із головних викликів – це мовні та культурні нюанси, пов’язані з китайською мовою, які створюють унікальні труднощі для

обробки та аналізу тексту. Культурні відмінності та контекстуальні посилання можуть ускладнювати точну класифікацію контенту моделлю. Крім того, набір даних часто є незбалансованим із більшою кількістю реальних новин, що може призвести до потенційного перекосу в бік класифікації новин як реальних, що впливає на специфічність моделі.

Ще одним викликом є виявлення сарказму та гумору, які модель часто неправильно класифікує як реальні або фейкові новини. Сарказм і гумор часто включають гру слів або подвійний сенс, що може бути неочевидним для моделі. Якість міток у навчальних даних також має вирішальне значення для продуктивності моделі. Неправильно позначені екземпляри в навчальних даних можуть призвести до того, що модель вивчить неправильні шаблони.

Фейкові новини можуть суттєво відрізнятися за змістом, стилем і подачею, що є значним викликом. Деякі статті можуть бути повністю вигаданими (дезінформація), тоді як інші можуть містити суміш правдивої та неправдивої інформації (малінформація). Ця варіативність ускладнює точну класифікацію фейкових новин. Крім того, природа новин і інформації швидко змінюється, особливо під час криз охорони здоров’я, як-от пандемії COVID-19. Модель має бути стійкою до цих часових динамік і точно класифікувати новини, навіть коли ситуація стрімко еволюціонує.

Вибір ознак є ще одним складним аспектом цього завдання. Модель повинна враховувати різноманітні ознаки, такі як текст, зображення та взаємодії користувачів (наприклад, коментарі, лайки, репости) для точної класифікації новин. Однак використання занадто великої кількості ознак може призвести до перенавчання, а занадто малої – до недонавчання.

Класифікація фейкових новин про COVID-19 на платформі Weibo має кілька суттєвих викликів, включаючи мовні та культурні нюанси, незбалансовані дані, виявлення сарказму та гумору, якість маркування, змінність фейкових новин, вибір ознак, тощо. Для подолання цих викликів потрібен комплексний підхід, який враховує різні аспекти даних і моделі, щоб досягти високої продуктивності класифікації та забезпечити інтерпретованість і стійкість моделі до часових змін.

2.3. Ансамблеві методи машинного навчання для класифікації текстової дезінформації та пропаганди

2.3.1. Аналіз ансамблевих моделей для задач класифікації

1. Збалансований випадковий ліс

Збалансований випадковий ліс (BRF) – це метод ансамблевого машинного навчання, призначений для вирішення проблеми незбалансованості класів у наборах даних [119]. Незбалансованість класів часто зустрічається в багатьох реальних наборах даних, де один клас значно переважає інші, що призводить до зміщених прогнозів. Метод BRF модифікує стандартний підхід RF, використовуючи методи для балансування розподілу класів перед навчанням кожного дерева. Метод RF будує ансамбль дерев рішень, випадково вибираючи вибірки з набору даних і навчаючи кожне дерево на різних вибірках. Остаточний прогноз отримують шляхом агрегування прогнозів з усіх дерев, зазвичай за принципом більшості голосів для класифікації або усереднення для регресії. Перед навчанням кожного дерева у лісі метод BRF виконує вибірку більшості класу, щоб збалансувати розподіл класів. Це гарантує, що кожне дерево навчається на наборі даних, де класи приблизно рівномірно представлені.

BRF є надійним ансамблевим методом, призначеним для роботи з незбалансованими наборами даних. Забезпечуючи, що кожне дерево в ансамблі навчається на збалансованій підмножині даних, BRF надає більш рівноправне представлення всіх класів, що призводить до точніших і менш упереджених прогнозів.

2. XGBoost

XGBoost або eXtreme Gradient Boosting – це вдосконалена реалізація методу градієнтного бустингу, розроблена для швидкості та продуктивності [120]. Цей метод здобув значну увагу в спільноті машинного навчання завдяки своїй ефективності та здатності вирішувати різноманітні завдання

прогнозування. XGBoost базується на концепції бустингу, де слабкі моделі (зазвичай дерева рішень) навчаються послідовно, щоб виправляти помилки своїх попередників [121]. Остаточний прогноз є ансамблем цих слабких моделей. XGBoost використовує градієнтний бустинг, в якому нові моделі навчаються для прогнозування залишків або помилок попередніх моделей

XGBoost – це надійний метод машинного навчання, який використовує принципи градієнтного бустингу разом з регуляризацією, що робить його потужним, ефективним та точним інструментом прогнозування.

3. Модель LightGBM

LightGBM, що означає Light Gradient Boosting Machine, – це фреймворк градієнтного бустингу, спеціально розроблений для швидкості та підвищення продуктивності [122]. Розроблений Microsoft, він впроваджує декілька інновацій та оптимізацій, що роблять його швидшим та більш ощадливим у використанні пам’яті порівняно з іншими фреймворками градієнтного бустингу, без шкоди для точності. Подібно до інших методів градієнтного бустингу, LightGBM будує ансамбль рішень дерев послідовно, де кожне дерево виправляє помилки свого попередника. Остаточний прогноз є сумою прогнозів від окремих дерев. На відміну від традиційних методів градієнтного бустингу, які будують дерева рівнево, LightGBM використовує стратегію росту за листками. Він вибирає листок із максимальним приростом втрат для розширення, що дозволяє зменшити більше втрат у порівнянні з рівневими методами, але за рахунок збільшення складності моделі.

LightGBM – це просунутий фреймворк градієнтного бустингу, який завдяки своїм інноваційним підходам забезпечує баланс між обчислювальною продуктивністю та точністю прогнозування, що робить його особливо придатним для великих наборів даних або ситуацій, коли обчислювальні ресурси обмежені.

2.3.2. Оптимізація гіперпараметрів та валідація моделей

З огляду на велику кількість гіперпараметрів у фреймворках градієнтного бустингу, таких як XGBoost або LightGBM, необхідна ефективна техніка оптимізації. Для цієї задачі було використано Optuna [123], сучасну бібліотеку для оптимізації гіперпараметрів:

- Баєсівська оптимізація: Замість традиційних методів перебору сітки або випадкового пошуку, Optuna використовує баєсівську оптимізацію. Моделюючи функцію мети (для поставленої задачі, помилку валідації) за допомогою гауссівських процесів, Optuna оновлює свої припущення після кожного випробування, спрямовуючи пошук до гіперпараметрів, які можуть мінімізувати помилку;
- Обрізання: Для підвищення ефективності механізм обрізання Optuna достроково припиняє випробування, якщо набір гіперпараметрів, ймовірно, не перевершить поточні найкращі результати.

Було налаштовано наступні гіперпараметри XGBoost:

- n_estimators: У межах від 100 до 1000, що визначає кількість раундів бустингу;
- learning_rate: Логарифмічно варіювався між 1E-3 та 1E-1, що визначає розмір кроку на кожній ітерації;
- subsample та colsample_bytree: Обидва параметри вибиралися в межах від 0.5 до 1, визначаючи частку даних і ознак, використаних на кожному раунді бустингу відповідно;
- max_depth: Ціле число між 3 і 9, що задає максимальну глибину дерева;
- параметри регуляризації: Включають gamma (від 0 до 1), lambda (логарифмічна шкала між 1E-3 до 10) та alpha (логарифмічна шкала між 1E-3 до 10), що колективно працюють для запобігання перенавчанню.

Для LightGBM було налаштовано такі гіперпараметри:

- num_leaves: Визначає кількість листків у дереві, було проведено експерименти зі значеннями від 2 до 256, розуміючи, що більша кількість може підвищити точність, але також призвести до перенавчання;
- learning_rate: Цей параметр визначає розмір кроку на кожній ітерації бустингу. Його було варійовано логарифмічно між 1E-3 та 1E-1, забезпечуючи баланс між швидкістю збіжності моделі та точністю; (а) bagging_fraction та feature_fraction: (б) bagging_fraction: Визначає частку даних, використаних на кожному раунді бустингу. Було обрано значення між 0.4 і 1.0, забезпечуючи як різноманітність, так і достатню кількість даних для кожного раунду;
- feature_fraction: Представляє частку ознак, які враховуються на кожному раунді бустингу. Цей параметр було налаштовано у межах від 0.4 до 1.0, дозволяючи моделі зосереджуватись на різних підмножинах ознак на різних раундах;
- max_depth: Важливий параметр, що задає максимальну глибину дерев у LightGBM, та був визначений від 3 до 9. Глибше дерево може уловлювати більш складні патерни, але також ризикує перенавчитися;
- параметри регуляризації: (а) lambda_l1: Цей параметр L1 регуляризації був налаштований на логарифмічній шкалі від 1E-8 до 10. Він допомагає у виборі ознак та запобіганні перенавчанню; (б) lambda_l2: Виконує роль параметра L2 регуляризації, який також було налаштовано на логарифмічній шкалі від 1E-8 до 10. Він зменшує складність моделі та запобігає перенавчанню.

З найкращими гіперпараметрами, остаточний етап навчання включав:

- Повторне навчання на повному наборі даних: Найкращі гіперпараметри з етапу оптимізації були використані для навчання моделей на повному тренувальному наборі даних. Такий підхід гарантує, що модель отримує користь з усіх доступних даних, використовуючи патерни та взаємозв'язки, які могли бути втрачені в підмножині;
- Валідація: Після навчання продуктивність моделі була оцінена на окремому тестовому наборі. Цей систематичний підхід дозволяє створити модель, здатну ефективно розрізняти справжні новини від фейкових.

2.3.4. Результати аналізу ансамблевих моделей для задачі класифікації текстової дезінформації та пропаганди

Для класифікації фейкових новин було проведено навчання трьох ансамблевих моделей машинного навчання: XGBoost, LightGBM та BRF. Використовуючи набір даних WELFake [124], було виділено 80% для тренування, що дозволило моделям виявити шаблони та тонкощі в новинних статтях. Решта 20% була зарезервована для тестування, щоб надати неупереджену оцінку здатності кожної моделі класифікувати новини. Опис набору даних наведено в Додатку Г.

Хоча модель BRF була створена для роботи з незбалансованими класами, вона показала особливу ефективність для вирішення поставленої задачі, навіть на збалансованому наборі даних. Використовуючи стратегію вибірки під час навчання кожного дерева, модель BRF забезпечила рівне ставлення до справжніх і фейкових новин. Незважаючи на диспропорцію класів, цей підхід посилив стійкість моделі до потенційного перенавчання. Це гарантувало, що сукупний результат ансамблю був тонким і надійним у класифікації.

Модель XGBoost відіграла ключову роль, використовуючи механізм градієнтного бустингу для ітераційного фокусування на неправильно класифікованих прикладах. Така методична точність забезпечила здатність моделі розрізняти справжній та сфабрикований контент новин.

Модель LightGBM, використовуючи свою унікальну стратегію росту за листками, продемонструвала високу продуктивність як за швидкістю обчислень, так і за точністю класифікації. З огляду на складність та можливі диспропорції класів у наборах даних з фейковими новинами, акцент моделі на максимізацію зменшення втрат був особливо корисним.

У Таблиці 2.7 наведено оцінку продуктивність моделей. Для глибшого розуміння продуктивності моделей класифікації та визначення можливих зон для вдосконалення, в якості метрики було розглянуто матрицю невідповідностей.

На рисунку 2.6 зображено матрицю невідповідностей для моделі BRF, на рисунку 2.7 – для моделі XGBoost, на рисунку 2.8 – для моделі LightGBM.

Таблиця 2.7 – Оцінка продуктивності моделей

Метрика	BRF	XGBoost	LightGBM
Міра F1 (тест)	0.944557408	0.968226763	0.97741164
Точність (тест)	0.943300756	0.966588105	0.976431443
Повнота (тест)	0.949577542	0.984186545	0.985794693
Прецизійність (тест)	0.939590076	0.952776336	0.96916996
Втрата (тест)	2.043647915	0.095977571	0.123440667
Міра F1 (тренування)	1	0.988855286	1
Точність (тренування)	1	0.98844161	1
Повнота (тренування)	1	0.996497727	1
Прецизійність (тренування)	1	0.981329177	1
Втрата (тренування)	2.22E-16	0.055085406	1.68E-06

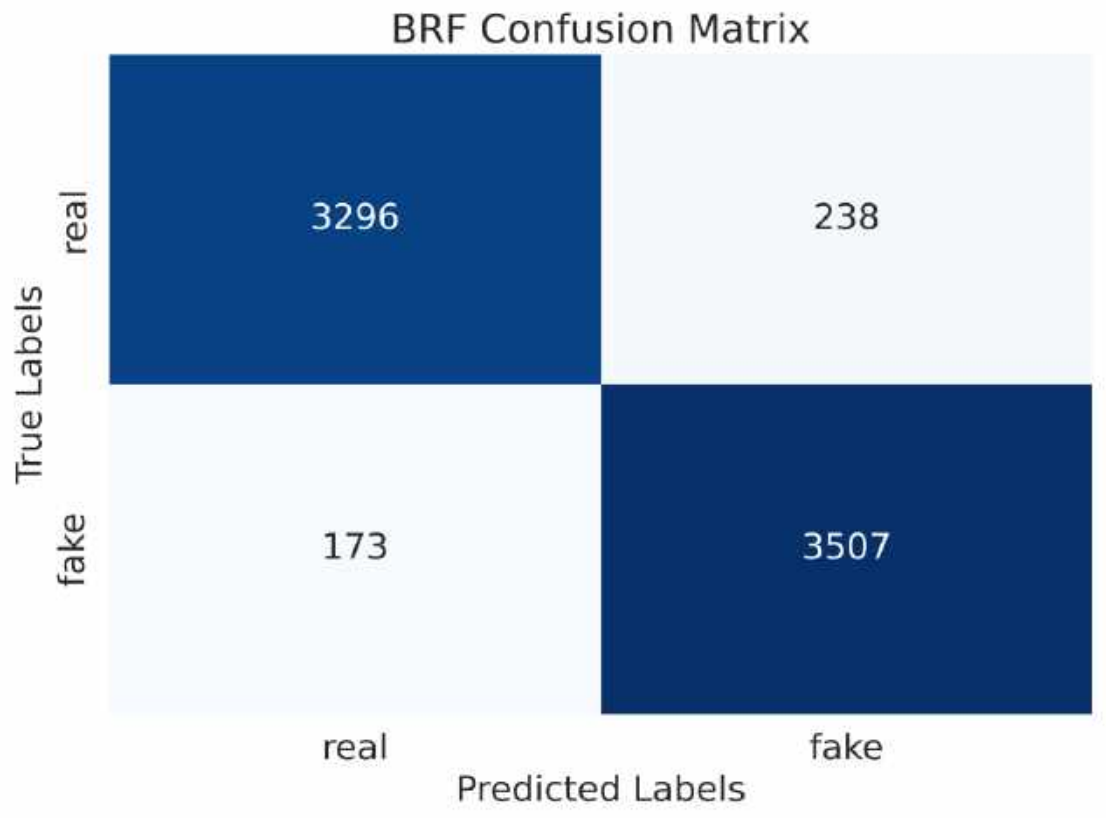

Рисунок 2.6 – Матриця невідповідностей для моделі BRF.

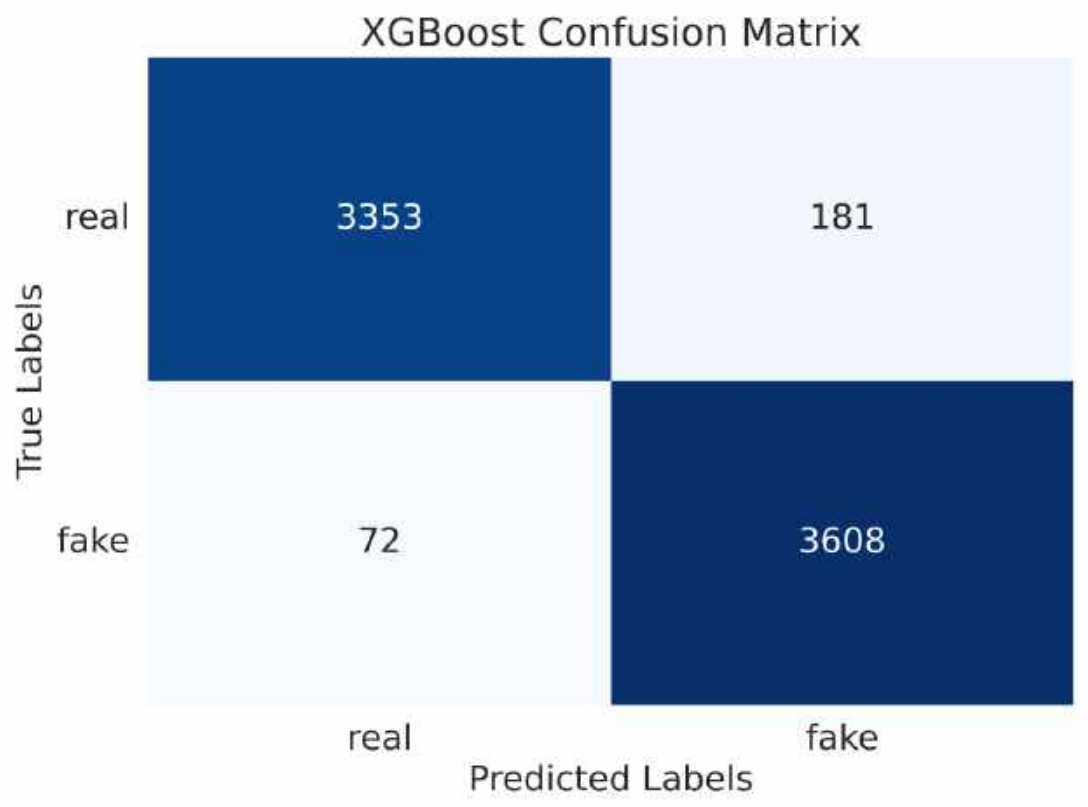

Рисунок 2.7 – Матриця невідповідностей для моделі XGBoost

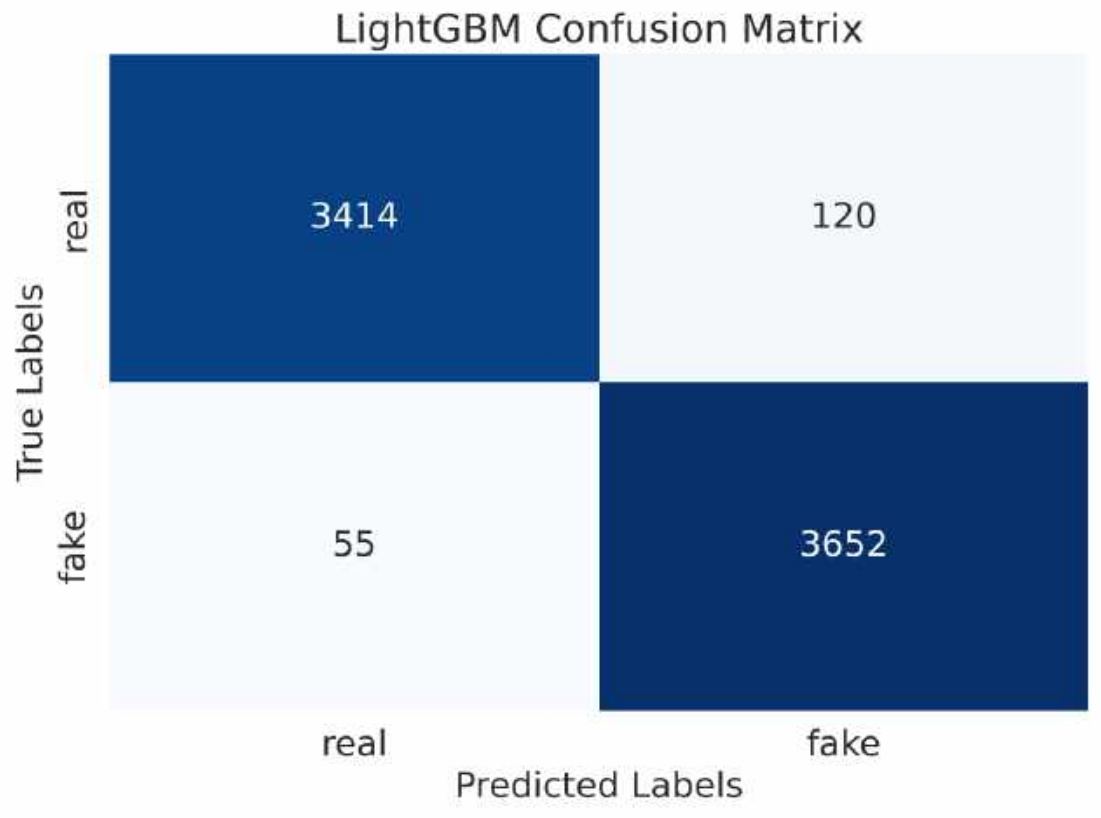

Рис. 2.8 – Матриця невідповідностей для LightGBM моделі

Існує помітна кількість хибнопозитивних і хибнонегативних результатів у моделі BRF незважаючи на те, що вона досягла високої продуктивності в більшості випадків. Це свідчить про те, що, хоча модель загалом продуктивна, певні випадки або шаблони можуть бути неправильно інтерпретовані, що призводить до цих помилок.

Модель XGBoost, з іншого боку, демонструє більш точну класифікацію з меншою кількістю хибнопозитивних і хибнонегативних результатів, ніж модель BRF. Це свідчить про вищу точність і повноту, що означає, що модель ефективно виявляє фейкові новини і не класифікує справжні новини як фейкові. Було включено криву ROC (Receiver Operating Characteristic) в систему оцінювання, щоб пояснити компроміси між чутливістю та специфічністю за різних порогових значень. Крива ROC – це графічне зображення, що відображає істиннопозитивний показник (чутливість) у порівнянні з хибнопозитивним показником (специфічність) для системи бінарної класифікації при зміні дискримінаційного порогу. Крива надає уявлення про продуктивність класифікатора на всіх рівнях порогу, при цьому площа під кривою (AUC) служить як єдиний показник для кількісної оцінки здатності моделі розрізняти позитивні та негативні класи. Модель з ідеальною розрізнювальною здатністю матиме AUC, що дорівнює 1, тоді як модель без розрізнювальної здатності матиме AUC 0.5. Включивши криву ROC, можливо оцінити надійність моделі та її здатність збалансувати виявлення істинно позитивних результатів і уникати хибно позитивних.

Модель LightGBM продемонструвала найкращі результати, з найменшою кількістю хибнопозитивних і хибнонегативних результатів. Це свідчить про найкращий баланс між точністю та повнотою, що забезпечує точність прогнозів моделі та її чутливість до нюансів даних.

Хоча всі три моделі демонструють надійні можливості класифікації, модель LightGBM виділяється своєю точністю та повнотою, що робить її найпродуктивнішої для розрізнення справжніх і фейкових новин у цьому наборі даних. На рисунку 2.9 зображено криву ROC для моделі BRF, на рисунку 2.10 – для моделі XGBoost, на рисунку 2.11 – для моделі LightGBM.

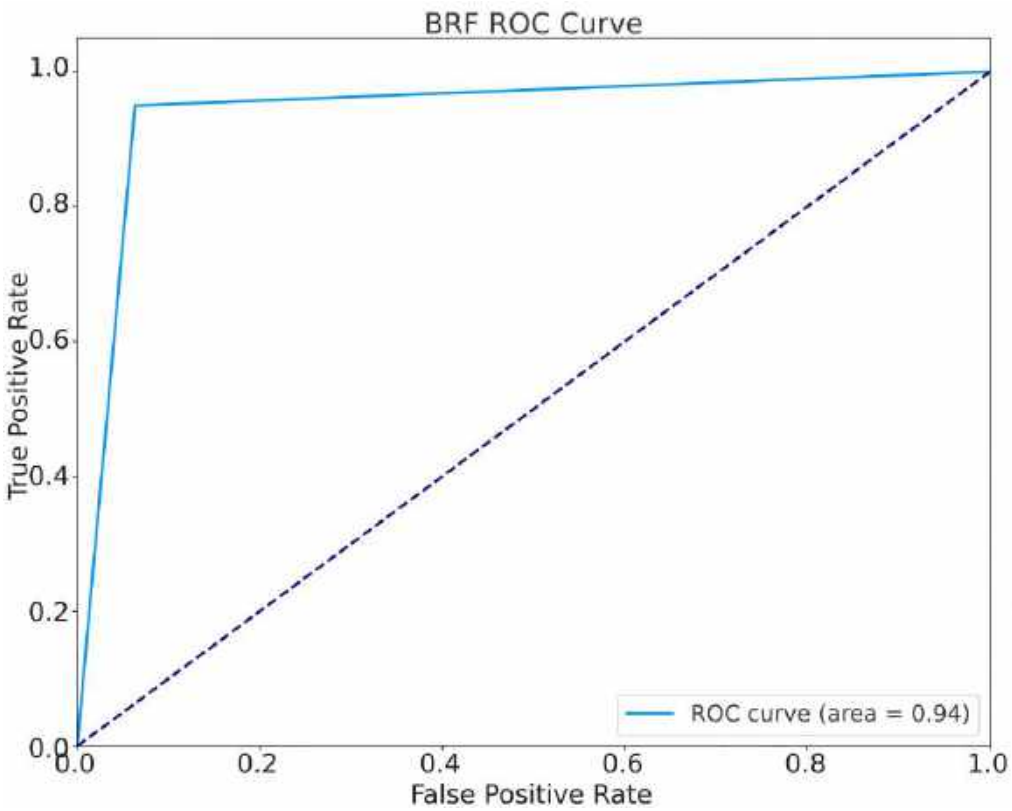

Рис 2.9 – Крива ROC для моделі BRF

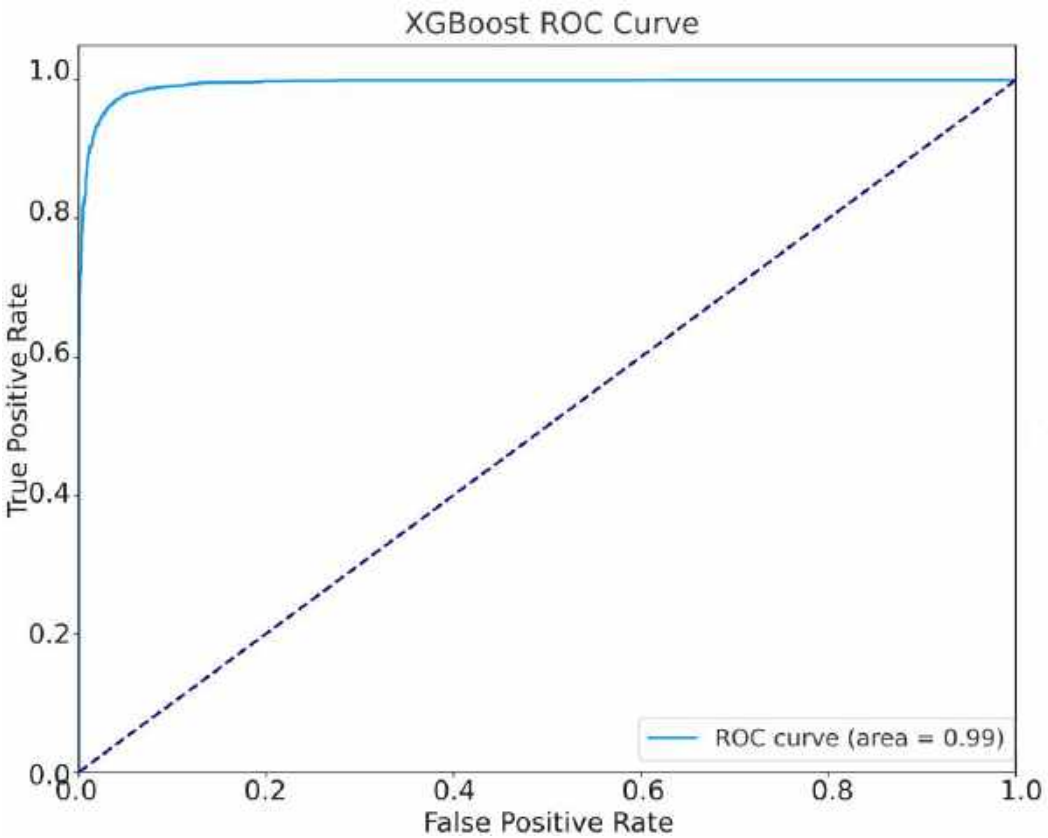

Рис 2.10 – Крива ROC для моделі XGBoost

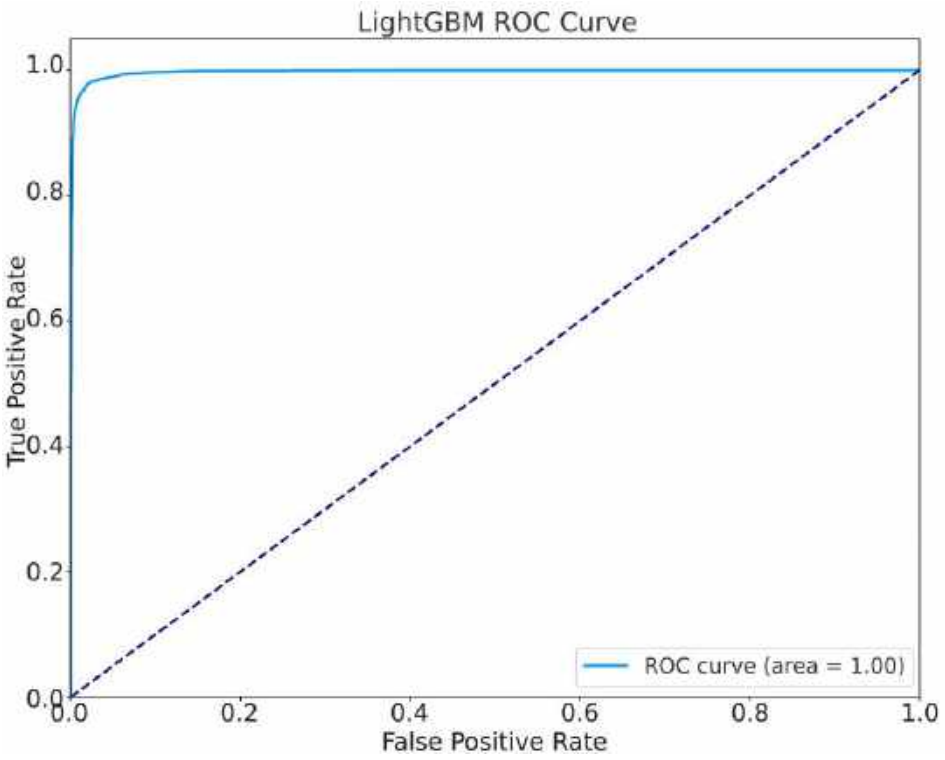

Рис 2.11 – Крива ROC для моделі LightGBM

Було проаналізовано різні метрики продуктивності для кожної моделі. Міри F1 на тестовому наборі були помітно високими для всіх моделей, причому модель LightGBM досягла найвищого значення приблизно 0.977, за нею слідує XGBoost з показником близько 0.968 та BRF з 0.945. Це свідчить про гармонійний баланс між прецизійністю та повнотою для кожної моделі, особливо для LightGBM. Метрики точності також підтверджують це, із значеннями, що наближаються або перевищують 94% для всіх моделей на тестовому наборі. Повнота для XGBoost та LightGBM була високою, що свідчить про те, що ці моделі добре визначають більшість випадків фейкових новин. Однак за прецизійністю, модель LightGBM трохи перевершила XGBoost, що означає меншу кількість хибнопозитивних результатів. Значення втрат на тестовому наборі надають уявлення про відхилення моделей від фактичних результатів, при цьому BRF мала значно вищі втрати порівняно з іншими моделями. Під час аналізу навчальних метрик як LightGBM, так і BRF досягли ідеальних показників за мірою F1, прецизійністю, повнотою та точністю, що свідчить про можливе перенавчання на тренувальних даних. На

противагу, XGBoost, хоча і показує високі метрики, продемонстрував незначне відхилення від ідеальних показників, що може вказувати на більш узагальнену модель. Загалом, ці результати підкреслюють ефективність моделей у класифікації фейкових новин, причому кожна модель має свої унікальні сильні сторони та сфери для подальшого дослідження.

Модель XGBoost показала високу міру F1 0.968 на тестовому наборі, що свідчить про баланс між прецизійністю та повнотою. Точність 0.966 підтверджує її здатність правильно класифікувати новинні статті. Особливо примітно, що показник повноти 0.984 вказує на сильну здатність ідентифікувати більшість фейкових новин. Однак прецизійність 0.953 свідчить про дещо вищий рівень хибнопозитивних результатів порівняно з іншими моделями.

Модель LightGBM продемонструвала найвищу міру F1 та точність на тестовому наборі, що перевищує 0.976. Це свідчить про те, що LightGBM правильно класифікує більшість випадків і підтримує баланс між прецизійністю та повнотою. Показник повноти дорівнює 0.986, що означає, що модель майже не пропускає фейкових новин. Точність 0.969 ще більше підкреслює її здатність мінімізувати хибно позитивні результати.

Нарешті, модель BRF, розроблена для вирішення проблеми незбалансованих класів, має міру F1 0.945 на тестовому наборі. Хоча це дещо нижче, ніж у двох інших моделей, вона все ще свідчить про надійну продуктивність. Точність 0.943 та повнота 0.950 підтверджують її ефективність. Однак прецизійність 0.940, хоча і гідна, вказує на дещо вищу схильність до хибнопозитивних результатів. Значення втрат для тестового набору є помітно вищими, ніж у інших моделей, що може вказувати на області для оптимізації. Хоча всі три моделі продемонстрували потужні можливості у класифікації фейкових новин, LightGBM виявилася найкращою в цій оцінці.

Отримані результати надали важливі інсайти щодо класифікації фейкових новин за допомогою ансамблевих моделей машинного навчання. Метрики продуктивності, отримані з матриці невідповідностей та кривої ROC,

надають комплексне уявлення про сильні сторони кожної моделі та можливі зони для покращення. LightGBM продемонстрував виняткову ефективність, особливо враховуючи баланс між швидкістю та точністю. XGBoost проявив свою силу в ітеративному вдосконаленні, акцентуючи увагу на неправильно класифікованих випадках з попередніх ітерацій. BRF запропонував надійний механізм класифікації, який зменшив вроджені упередження.

2.4. Висновки до другого розділу

1. Дістали подальшого розвитку моделі для класифікації текстової дезінформації та пропаганди, засновані на ансамблевих методах машинного навчання, які на відміну від існуючих враховують балансування класів і адаптуються до змін вхідних даних, що дозволяє підвищити продуктивність моделей.

2. Дістали подальшого розвитку моделі для класифікації інформації, засновані на статистичних методах машинного навчання, які на відміну від існуючих адаптовані до класифікації текстової дезінформації та пропаганди, що дозволяє створення систем для виявлення місінформації, дезінформації та малінформації.

3. Проведено експериментальні дослідження моделей машинного навчання LR, NB, KNN, SVM для класифікації російської дезінформації про війну в Україні у X (Twitter).

4. Проведено експериментальні дослідження ансамблевих моделей машинного навчання BRF, XGBoost, LightGBM для аналізу набору даних фейкових та реальних новин WELFake.

5. Проведено експериментальні дослідження моделей машинного навчання SVM для аналізу текстової дезінформації та пропаганди на інших мовних структурах, зокрема для аналізу китайськомовного набору даних з соціальної мережі Weibo.

Основні результати, представлені в цьому розділі, опубліковані в [16], [22], [24], [25].

РОЗДІЛ 3. МОДЕЛІ ГЛИБОКОГО НАВЧАННЯ ДЛЯ АНАЛІЗУ ТЕКСТОВОЇ ДЕЗІНФОРМАЦІЇ ТА ПРОПАГАНДИ

У цьому розділі дисертаційного дослідження було удосконалено та розроблено нові моделі глибокого навчання для задачі класифікації текстової дезінформації та пропаганди.

3.1 Модель BERT для класифікації текстової дезінформації та пропаганди

3.1.1. Опис вдосконаленої моделі на основі BERT

Модель BERT (Bidirectional Encoder Representations from Transformers) є значним проривом у сфері обробки природної мови, особливо у виявленні дезінформації, пропаганди та фейкових новин. Розроблена дослідниками з Google, основна інновація BERT полягає в здатності аналізувати контекст слів у реченні двосторонньо, на відміну від традиційного підходу, що базується на односторонній або послідовній обробці [125]. Ця здатність є ключовою для правильного сприйняття тонкощів і складностей мови, що є важливим для точного виявлення дезінформації.

BERT працює на основі трансформеру – архітектури, яка обробляє слова у взаємозв’язку з усіма іншими словами в реченні, замість традиційного покрокового підходу. Це дозволяє отримати більш цілісне розуміння структури і змісту речення. Модель попередньо тренується на великому корпусі текстів, що дозволяє їй враховувати різноманітні мовні закономірності та нюанси. Це попереднє навчання включає задачу прогнозування відсутніх слів у реченнях, що допомагає моделі враховувати закономірності контекстуальних відносин між словами.

Для виявлення фейкових новин, BERT можна донавчити на специфічних для новин наборах даних. Це донавчання включає тренування моделі на наборі даних, що містить як справжні, так і фейкові новини, дозволяючи їй вивчати характеристики та закономірності, які відрізняють справжні новини від

дезінформації. Перевага BERT полягає в здатності моделі розуміти тонкі підказки та мовні варіації, що часто вказують на фейкові новини, такі як перебільшені твердження, непослідовна інформація або сенсаційні, емоційні заяви.

Модель розпізнавання фейкових новин на основі BERT функціонує у кілька етапів. Текстовий вміст новинних статей виступає входом моделі. Модель обробляє цей вміст, враховуючи контекст кожного слова та речення. Модель створює репрезентації тексту, які захоплюють як лінгвістичні властивості, так і вивчені закономірності фейкових та справжніх новин. Нарешті, ці репрезентації використовуються для класифікації новин як фейкових або справжніх.

Однією з ключових переваг моделі у цьому контексті є її здатність до трансферного навчання. Після попереднього навчання на великому масиві текстів вона може ефективно адаптуватися до специфічної мови та стилю новинних статей навіть при відносно невеликій кількості даних для донавчання. Це робить її дуже ефективною і точною у виявленні фейкових новин навіть у випадках з обмеженим обсягом тренувальних даних.

Однак варто зазначити деякі обмеження моделі у цьому застосуванні. Незважаючи на високу ефективність у аналізі мови, BERT потребує значних обчислювальних ресурсів, що може бути обмеженням у певних середовищах. Крім того, продуктивність моделі може залежати від якості та репрезентативності тренувальних даних. Використання упереджених або незбалансованих наборів даних може призвести до менш точних класифікацій.

У процесі навчання вдосконалена модель BERT була натренована протягом декількох епох за допомогою оптимізатора AdamW та планувальника лінійної швидкості навчання. Навчання включає зворотне поширення для налаштування ваг на основі втрат, що обчислюються від порівняння прогнозів моделі з фактичними мітками. Продуктивність моделі періодично оцінюється на валідаційному наборі даних, аналізуючи метрики та створюючи звіти по класифікації. Цей цикл навчання та оцінювання ітеративно покращує здатність моделі класифікувати текст як правдивий або фейковий.

3.1.2. Результати вдосконаленої моделі на основі BERT

Для експериментального дослідження розробленої моделі використано набір даних “Fake-Real News" [126], описаний у Додатку Г. У цьому підрозділі використовується набір показників продуктивності для оцінки ефективності моделі машинного навчання у класифікації новинних статей як “Фейкові” або “Справжні”. До цих показників належать точність, прецизійність, повнота, міра F1 та AUC. Разом ці показники надають всебічну картину класифікаційних можливостей моделі, підкреслюючи як її сильні сторони, так і можливі напрямки для покращення. Результати класифікації фейкових новин наведені в Таблиці 3.1.

Таблиця 3.1 – Результати класифікації методом BERT

Клас	Прецизійність	Повнота	Міра F1	Підтримка	Точність
Фейкові	0.82	0.83	0.83	733	0.80
Справжні	0.77	0.75	0.76	554	

Ці результати вказують на відносно високу продуктивність моделі у розрізненні між двома класами, при цьому вона демонструє трохи кращі результати у виявленні справжніх новин порівняно з фейковими. Загальна точність свідчить про надійність моделі в обробці цього завдання класифікації.

На Рисунку 3.1 зображено втрати моделі за епохами, що ілюструє зменшення втрат під час навчання протягом 15 епох. Як видно з графіка, втрата різко знижується на початкових етапах навчання, що свідчить про те, що модель швидко навчається на основі тренувальних даних. Найстрімкіше зниження відбувається між 1-ю та 4-ю епохами, після чого крива починає згладжуватися, вказуючи на сповільнення темпу покращення. До 5-ї епохи втрата значно зменшилася і продовжує знижуватися лише незначно, поступово наближаючись до плато на 15-й епосі. Це згладжування кривої свідчить про те, що модель досягає точки збіжності, де подальше навчання

приносить мало або жодних покращень у зменшенні втрат. Це також вказує на те, що модель, ймовірно, досягла свого найкращого рівня продуктивності на тренувальному наборі даних.

Загалом графік демонструє успішний процес навчання, у ході якого модель значно покращує свою здатність мінімізувати функцію втрат, причому цей процес стабільний та ефективний протягом усіх епох навчання.

На Рисунку 3.2 зображено криву прецизійності-повноти, яка є графічним інструментом для оцінки продуктивності бінарного класифікатора. Як видно з кривої, зі збільшенням повноти прецизійність знижується, що свідчить про те, що коли модель ідентифікує більше позитивних прикладів (істиннопозитивних), вона також починає включати більше хибнопозитивних, зменшуючи точність. AUC зазначена як 0.82, що вказує на високий рівень загальної продуктивності моделі.

На рисунку 3.3 показано криву ROC, яка є графічним зображенням здатності бінарного класифікатора змінювати поріг розрізнення. ROC-крива демонструє, що запропонована модель добре відрізняє два класи. Площа під кривою ROC дорівнює 0.87, що кількісно відображає загальну здатність класифікатора розрізняти позитивні та негативні класи.

На рисунку 3.4 показано графік, що відображає зміну точності моделі на валідаційних даних протягом 15 епох. Графік показує значну варіабельність точності від епохи до епохи. Точність досягає піків на деяких етапах (близько 1-ї, 5-ї та 11-ї епох), досягаючи максимумів трохи вище 0.805. Водночас на інших етапах вона різко знижується, з найбільш помітним падінням на 9-й епосі, де точність падає нижче 0.795. Загальна тенденція не вказує на постійне покращення або погіршення точності у процесі навчання.

На рисунку 3.5 показано матрицю невідповідностей, яка свідчить про те, що модель досить ефективно виявляє фейкові новини, про що свідчить високий показник істинно позитивних для класу “Фейкові новини”. Однак матриця також демонструє простір для покращення, зокрема у зменшенні кількості хибно негативних, коли фейкові новини не виявляються, та хибно позитивних, коли справжні новини помилково позначаються як фейкові.

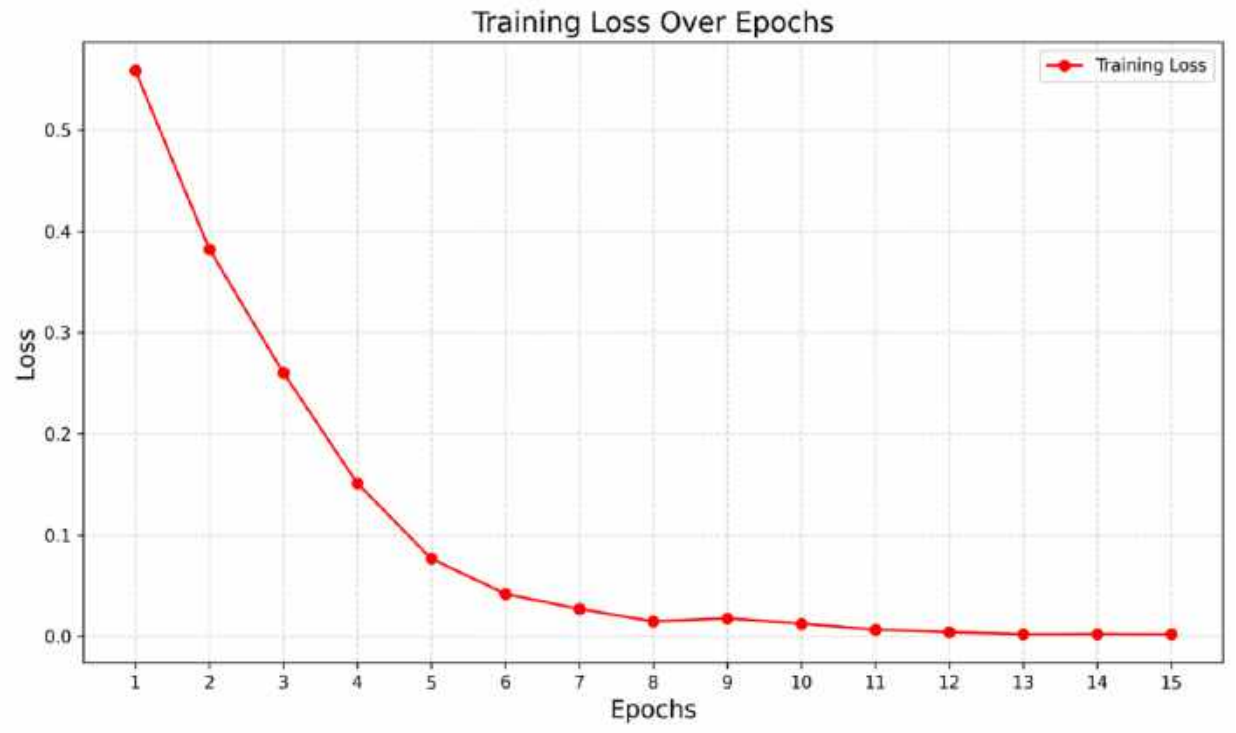

Рисунок 3.1 – Втрати моделі за епохами

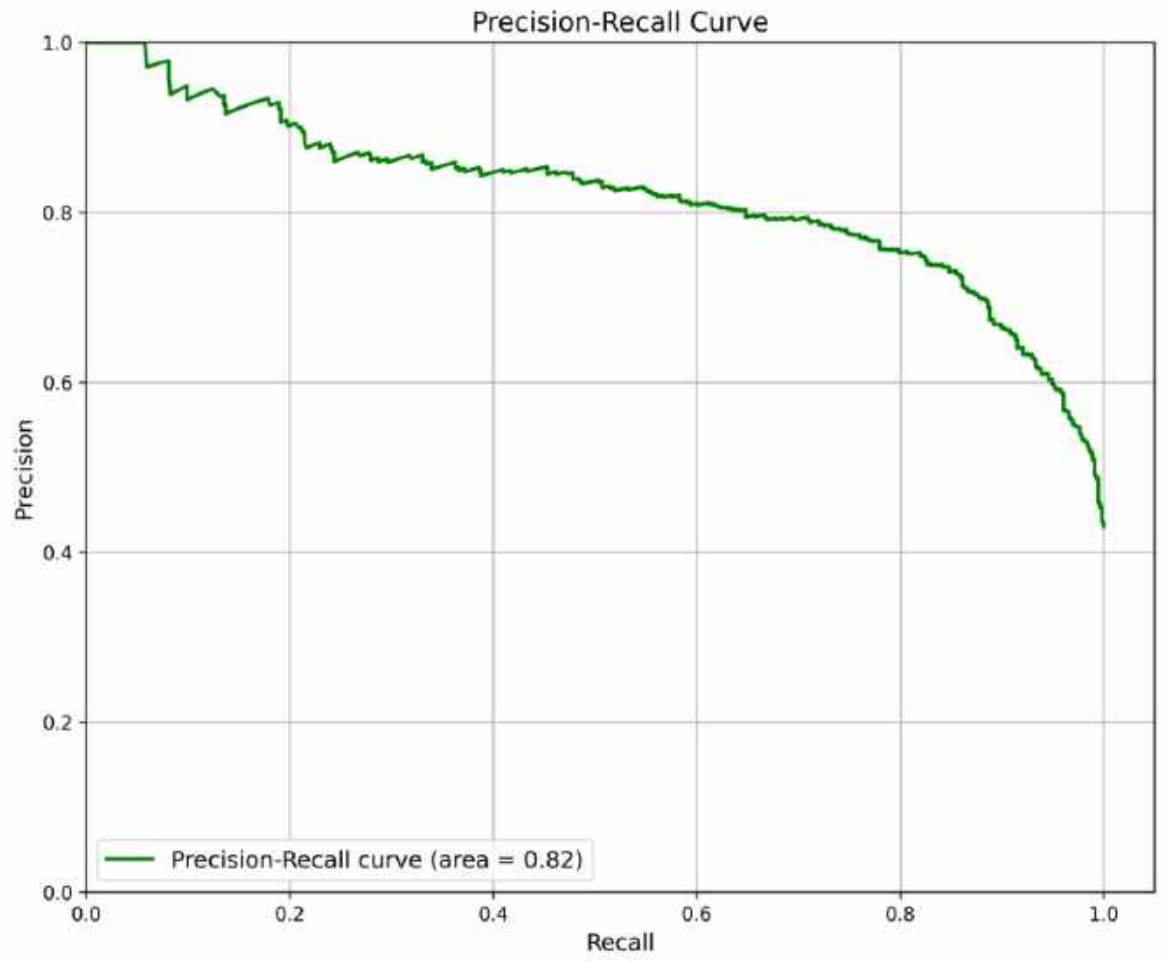

Рисунок 3.2 – Крива прецизійності-повноти

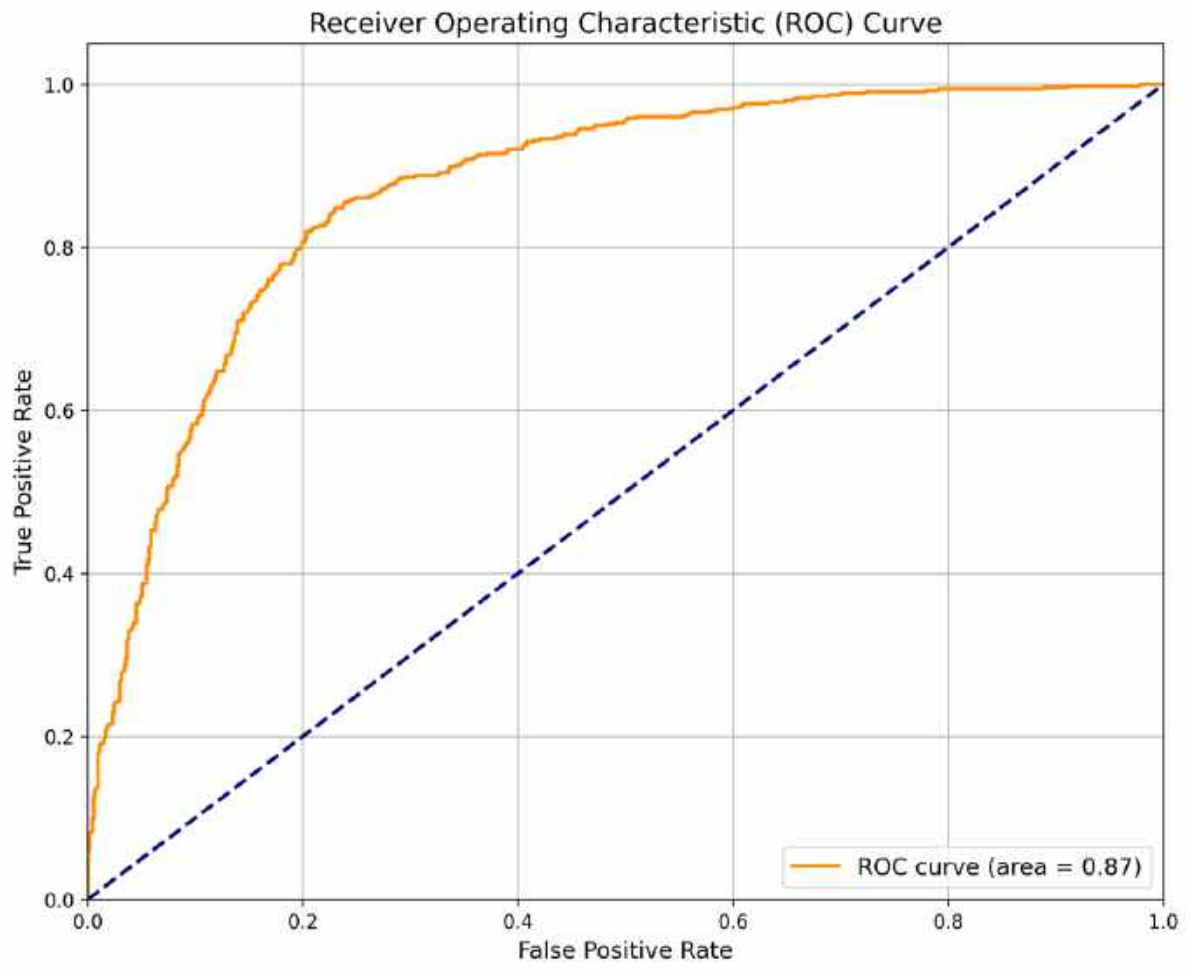

Рисунок 3.3 – Крива приймально-операційної характеристики (ROC)

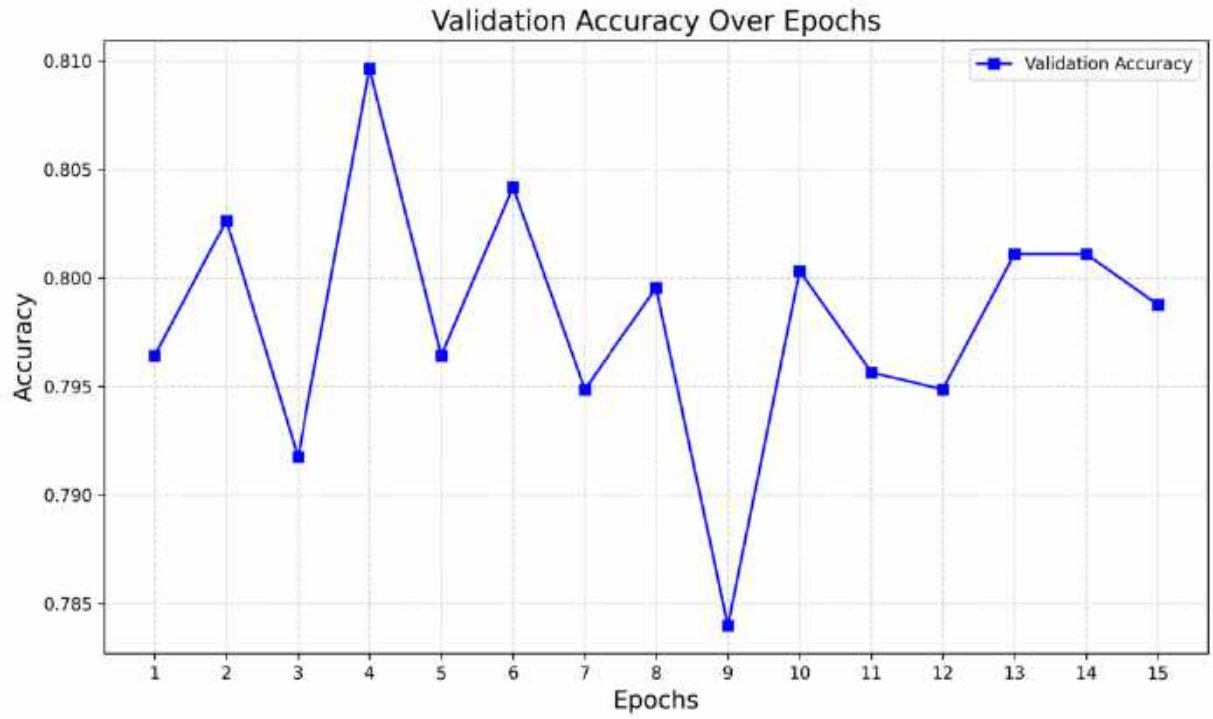

Рисунок 3.4 – Валідація точності даних за епохами

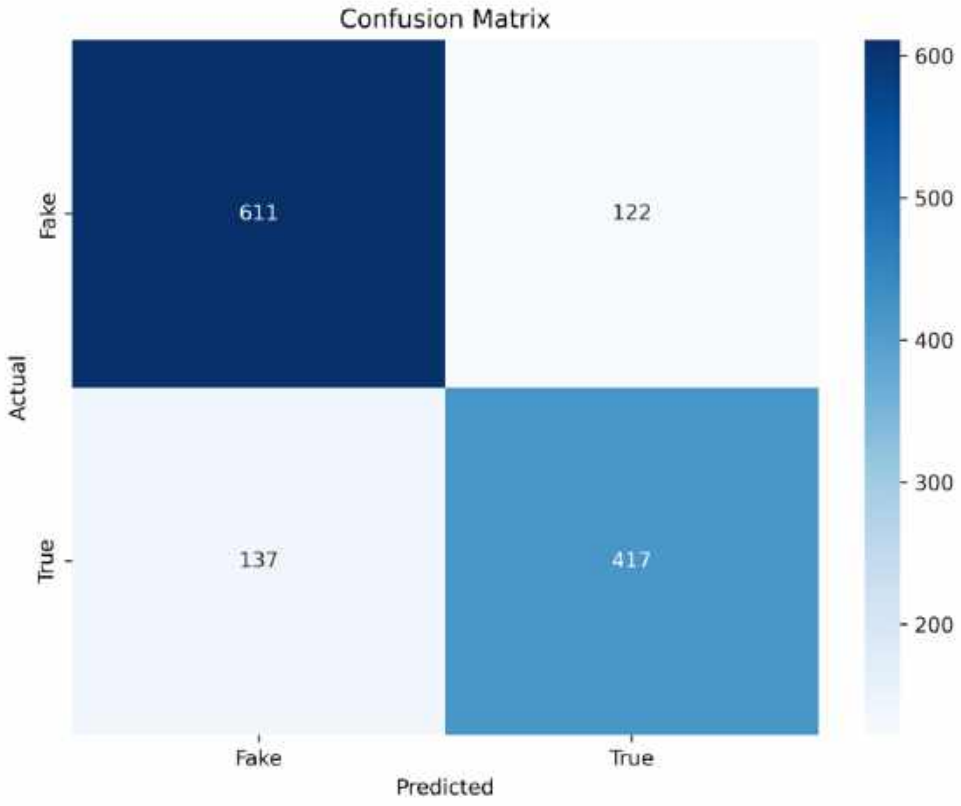

Рисунок 3.5 – Матриця невідповідностей

Точність на валідаційних даних досягла 79.88%, що є високим показником і свідчить про значний потенціал моделі у визначенні правдивості новинних статей. Ця точність надає базову оцінку загальної продуктивності моделі, однак вона не розкриває всіх тонкощів її передбачувальної здатності для різних класів новин.

Подальший аналіз результатів показав, що прецизійність для класу “Фейкові” склала 82%, а для класу “Справжні” – 77%. Ці показники вказують на високий рівень точності, особливо у виявленні фейкових новин. Хоча прецизійність для класу “Справжні” трохи нижча, вона все ж є суттєвою. Показники повноти склали 83% для класу “Фейкові” і 75% для “Справжні”, що свідчить про те, що модель краще виявляє фейкові новини, ніж охоплює всі справжні новини, що підтверджується показниками прецизійності.

Міра F1, яка об’єднує прецизійність і повноту, склала 83% для класу “Фейкові” і 76% для “Справжні”, що вказує на добре збалансовану

продуктивність, особливо для фейкових новин. Цей баланс є важливим у сценаріях, де релевантність результату (прецизійність) та здатність виявити всі релевантні випадки (повнота) мають вирішальне значення.

AUC дорівнює 0.87, що підтверджує здатність моделі розрізняти класи. AUC, наближене до 1, вказує на те, що модель має високий рівень істинно позитивних результатів щодо хибнопозитивних, демонструючи її здатність відрізняти “Фейкові” від “Справжні” новин. Такий високий показник AUC свідчить про надійну модель, яка добре працює при різних порогах, надаючи гнучкість у її використанні в реальних сценаріях залежно від конкретних потреб і компромісів.

Проте матриця невідповідностей надає додаткові відомості про розбіжності в здатності моделі класифікувати “Фейкові” новини порівняно зі “Справжні”. Незважаючи на те, що показник істинно позитивних результатів для фейкових новин є високим (611 із 733), модель також неправильно класифікувала значну кількість справжніх статей як фейкові (137 із 554).

Показники продуктивності та матриця невідповідностей свідчать про ефективну модель. Результати вказують на значну здатність виявляти фейкові новини, що є основною метою в контексті поширення дезінформації.

3.2. Модель XLNet для класифікації текстової дезінформації та пропаганди

3.2.1. Опис моделі XLNet

XLNet – це вдосконалена модель обробки природної мови, розроблена для подолання обмежень ранніх моделей. Вона базується на трансформерній архітектурі, але її відрізняє новий підхід до тренування на основі перестановок [127]. На відміну від BERT, яка використовує техніку маскування слів, XLNet використовує авторегресійну ціль тренування на основі перестановок, що дозволяє моделі ефективно враховувати двоспрямований контекст, не

жертвуючи природною мовною структурою. Цей підхід до перестановок забезпечує, що модель може навчатися залежностям між словами у більш узагальненій формі, тим самим покращуючи розуміння мовних відносин у межах послідовності.

Основою моделі XLNet є її авторегресійне формулювання на основі моделювання мови з перестановками. На відміну від BERT, який спирається на масковане мовне моделювання, XLNet генерує передбачення, максимізуючи ймовірність послідовності за всіма можливими перестановками порядку факторизації. Зокрема, для заданої послідовності x = (x1, x2, …, xT), XLNet максимізує таку цільову функцію:

$$L_{XLNet} = E_{\pi \epsilon P(T)} \sum_{t=1}^{T} \log P\left(x_{\pi_t} | x_{\pi_1}, \dots, x_{\pi_{t-1}}\right), \quad (3.1)$$

де π позначає можливу перестановку індексів послідовності x_i, представляє і-й токен у послідовності, P(T) – спільна ймовірність послідовності токенів T, а E – ембединги та навчальний процес передбачає обчислення спільної ймовірності токенів у послідовності з урахуванням множини перестановок. Ця авторегресійна природа дозволяє XLNet зберігати двоспрямований контекст без маскування токенів, забезпечуючи, що модель повністю вловлює взаємозалежності між усіма словами в реченні під час навчання. Натомість масковане мовне моделювання BERT може призводити до розбіжностей між навчанням та інференцією через штучно замасковані вхідні дані.

Механізм уваги в XLNet базується на сегментній рекурентності Transformer-XL, що покращує здатність моделі вловлювати довготривалі залежності між послідовностями. Рекурентний механізм дозволяє передавати контекстну інформацію з одного сегмента в наступний, подолавши обмеження фіксованої довжини, властиве багатьом попереднім трансформерним архітектурам. Формально, нехай $h_t^{(l)}$ позначає прихований стан на шарі l та позиції t для заданого сегмента послідовності. Тоді XLNet розширює

приховані стани між сегментами, таким чином моделюючи рекурентність за наступною формулою:

$$h_t^{(l)} = \text{TransformerXL}\left(h_t^{(l-1)}, h_{t-1}^{(l)}\right), \quad (3.2)$$

де увага для кожного шару може звертатися до попередніх сегментів, щоб зберігати контекст через кілька речень. Ця техніка є ключовою для виявлення дезінформації, оскільки контент, що поширюється у соціальних мережах, часто включає довготривалі залежності між повідомленнями чи постами.

Іншою важливою інновацією в XLNet є механізм уваги з двома потоками, який розрізняє передбачуваний контент (потік контенту) і позиційний запит, який використовується для передбачення (потік запиту). Цей механізм дозволяє XLNet окремо моделювати цільовий токен і контекст під час навчання. Механізм уваги з двома потоками визначається як:

$$q_t = f_q(h_t, x_t), \quad (3.3)$$
$$h_t = f_c(h_t, q_t), \quad (3.4)$$

де q_t позначає потік запитів, h_t – прихований стан потоку вмісту. Завдяки відокремленню цих двох потоків, XLNet може більш надійно працювати із залежностями між контекстом і цільовими токенами, що забезпечує кращу продуктивність у задачах, пов’язаних із тонкими мовними структурами, наприклад, при виявленні багатомовної дезінформації.

XLNet оптимізується за допомогою оптимізатора Adam із адаптивними швидкостями навчання. Цільова функція тренування на основі перестановок вводить значну обчислювальну складність, тому такі техніки, як ефективна у використанні пам’яті увага та обрізання градієнта, стабілізують тренування. Гіперпараметри моделі, включаючи швидкість навчання та кроки прогріву, налаштовані спеціально для забезпечення збіжності, зокрема враховуючи зашумлений характер даних із соціальних мереж, які використовуються для класифікації дезінформації.

3.2.2. Вдосконалення моделі XLNet

Для експериментального дослідження було використано набір даних російської пропаганди про війну в Україні у Твітері, описаний у Додатку Г. Процес вдосконалення моделі XLNet розпочався з комплексної попередньої обробки даних. Це включало очищення текстових даних шляхом видалення небажаних елементів, таких як хештеги, спеціальні символи, URL-адреси та стоп-слова, з метою покращення загальної якості набору даних. Було застосовано стемінг для зведення слів до їхніх кореневих форм, що забезпечило однорідність даних.

Набір даних було розподілено на навчальну (80%), валідаційну (10%) та тестову (10%) вибірки, що забезпечило ефективну основу для оцінювання та налаштування моделі. Навчальна вибірка використовувалася для навчання моделі, валідаційна допомагала в налаштуванні гіперпараметрів, а тестова оцінювала фінальну продуктивність.

Після попередньої обробки було ініціалізовано модель XLNet для виконання задачі бінарної класифікації, орієнтованої на виявлення проросійської пропаганди або нейтральних/проукраїнських повідомлень. Токенізація використовувалася для перетворення текстових даних у числовий формат, який модель могла інтерпретувати.

Зокрема, текст було токенізовано в послідовності з максимальною довжиною 256 токенів, щоб забезпечити узгодженість між вибірками. Така довжина була обрана для забезпечення балансу між захопленням достатнього контексту кожного твіту та підтриманням обчислювальної ефективності. Токенізація також включала заповнення коротших твітів до встановленої довжини та обрізання довших твітів, щоб вони відповідали встановленому ліміту.

Архітектуру моделі було оптимізовано шляхом ретельного вибору ключових гіперпараметрів для покращення стабільності тренування та точності. Навчання проводилось протягом трьох епох, оскільки це було

найкращим для балансу між часом навчання та підвищенням продуктивності моделі. Розмір пакета становив 8 для фаз навчання та оцінки, що дозволило ефективно використовувати пам’ять і забезпечити надійне оновлення моделі. Графік швидкості навчання включав фазу прогріву на 500 кроків, що дозволяло уникнути швидкої розбіжності моделі на початку навчання. Зниження ваг було встановлено на рівні 0.01, щоб зменшити ризик перенавчання, сприяючи кращій здатності моделі до узагальнення, шляхом покарання за занадто складні рішення.

Навчання проводилося з використанням оптимізатора Adam, який динамічно налаштовував швидкість навчання протягом тренування, щоб забезпечити збіжність. Під час фази оцінювання використовувалося раннє завершення навчання, що дозволяло зупиняти навчання, як тільки продуктивність на валідаційній вибірці переставала покращуватися, запобігаючи перенавчанню.

Метрики продуктивності, включаючи точність, прецизійність, повноту та міру F1, були зафіксовані для оцінювання ефективності моделі в різних класах. Остання натренована модель була оцінена на тестовій вибірці, де вона продемонструвала здатність точно класифікувати твіти в відповідні категорії, ефективно балансуючи між обчислювальною ефективністю та передбачуваною потужністю.

3.2.3. Результати продуктивності вдосконаленої моделі на основі архітектури XLNet

В таблиці 3.2 наведено метрики продуктивності (точність, повнота, міра F1, підтримка та прецизійність) вдосконаленої моделі XLNet для різних кількостей твітів та класів (0 та 1).

Таблиця 3.2 – Продуктивність вдосконаленої моделі на основі архітектури XLNet

Кількість твітів	Клас	Прецизійність	Повнота	Міра F1	Підтримка	Точність
1000	0	0.95	0.92	0.94	46	0.95
	1	0.95	0.97	0.96	54	
2000	0	1	0.94	0.97	77	0.98
	1	0.96	1	0.98	123	
3000	0	0.93	0.89	0.91	122	0.93
	1	0.93	0.96	0.94	178	
4000	0	0.93	0.90	0.91	172	0.93
	1	0.93	0.95	0.94	228	
5000	0	0.93	0.89	0.91	214	0.92

Продуктивність моделі XLNet демонструє стабільну здатність класифікувати твіти на проросійську дезінформацію та нейтральні/проукраїнські класи із стабільно високою точністю при різних обсягах даних. Точність у цьому контексті вимірює частку правильних позитивних передбачень серед усіх позитивних передбачень. Для найменшого набору з 1000 твітів точність для обох категорій залишалася стабільно високою на рівні 0.95, що відображає здатність моделі мінімізувати хибно позитивні твіти. Зі збільшенням набору даних точність залишалася здебільшого стабільною з незначними коливаннями. Наприклад, на рівні 2000 твітів клас 0 досяг ідеальної точності, що свідчить про те що всі проросійські твіти були ідентифіковані без жодних хибнопозитивних. Навіть при більших наборах даних з 3000, 4000 та 5000 твітів значення точності залишалися вище 0.91 для обох класів, що вказує на те, що модель ефективно розрізняла проросійську пропаганду та проукраїнські / нейтральні повідомлення.

Повнота є ще однією важливою метрикою, що вимірює здатність моделі ідентифікувати всі релевантні елементи для кожного класу. Для набору даних

з 1000 твітів повнота для проукраїнських твітів становила 0.97, трохи перевершуючи клас проросійську пропаганду, у якого була повнота 0.92. Ця тенденція зберігалася зі збільшенням розміру наборів даних, і клас проукраїнських та нейтральних твітів постійно показував вищі значення повноти, ніж проросійська пропаганда. Наприклад, на рівні 5000 твітів проукраїнські та нейтральні твіти мали повноту 0.95, тоді як проросійська пропаганда – 0.89. Ця різниця вказує на те, що модель була трохи більш ефективною у розпізнаванні всіх нейтральних/проукраїнських твітів порівняно з проросійською пропагандою, можливо, через більш узгоджені мовні шаблони в нейтральному/проукраїнському контенті. Тим не менш, повнота залишалася на рівні вище 0.89 для всіх обсягів даних, що підкріплює надійність моделі в ідентифікації релевантних твітів для обох категорій.

Для набору даних з 1000 твітів міра F1 для класу проросійської пропаганди становила 0.94, тоді як для проукраїнської та нейтральної інформації – 0.96, що свідчить про високу ефективність у класифікації твітів. Зі збільшенням набору даних показники міри F1 залишалися стабільно високими, зі значеннями між 0.91 і 0.98. При 2000 твітів міра F1 для класу проукраїнської та нейтральної інформації досягла 0.98, що показує здатність моделі підтримувати високий стандарт із додаванням нових даних. Навіть при більшому обсязі з 4000 та 5000 твітів міра F1 залишалася вище 0.91 для обох класів, що демонструє здатність моделі підтримувати баланс між точністю та повнотою, тим самим зменшуючи як хибно позитивні, так і хибно негативні результати.

Зі збільшенням розміру набору даних підтримка для кожного класу зростала, забезпечуючи ширшу базу для оцінки ефективності моделі. На рівні 5000 твітів клас російської пропаганди мав підтримку 214, тоді як клас проукраїнської та нейтральної інформації – 286. Стабільна продуктивність моделі при різних рівнях підтримки вказує на те, що вона добре обробляла різні розподіли класів. Високі значення підтримки також допомогли забезпечити, що метрики продуктивності не були викривлені через

недостатню кількість даних в одному з класів, сприяючи більш надійній оцінці результатів.

Точність класифікації залишалася високою для всіх обсягів даних. Для набору даних з 1000 твітів точність складала 95%, що свідчить про надійність навіть з обмеженою кількістю даних. Точність досягла піку у 98% для 2000 твітів, що вказує на відмінну роботу моделі з помірним обсягом даних. Однак із зростанням кількості твітів до 3000, 4000 і 5000 точність трохи знизилася до 93%. Це невелике зниження може бути викликане збільшенням різноманітності та складності даних, що внесло більш складні варіації та крайові випадки. Незважаючи на це зниження, точність вище 90% для всіх розмірів наборів даних підтверджує, що вдосконалена модель XLNet залишається потужним інструментом для класифікації контенту в соціальних мережах, особливо в контекстах, що включають інформаційну війну та аналіз пропаганди.

На рисунку 3.6 показано точність моделі XLNet в залежності від кількості твітів. Це дає важливе уявлення про те, як розвивалася продуктивність моделі зі збільшенням обсягу даних. Спочатку точність моделі значно покращилася з 95% до 98% при збільшенні кількості твітів з 1000 до 2000. Це покращення свідчить про те, що модель отримала користь від додаткових тренувальних даних, які надали більше контексту та прикладів для навчання, що призвело до кращої класифікації твітів. Однак після позначки у 2000 твітів точність почала поступово знижуватися зі збільшенням обсягу даних. Це зниження продуктивності, що впало до близько 93% на рівні 5000 твітів, може свідчити про те, що додана складність у більших наборах даних внесла більш складні варіації та крайові випадки, що, можливо, ускладнило моделі підтримку пікової точності.

Тенденція, що спостерігається на графіку, також говорить про те, що може існувати найкращий обсяг даних для тренування моделі XLNet для цього конкретного завдання. Хоча більший обсяг даних зазвичай допомагає покращити продуктивність моделі, у цьому випадку збільшення набору даних

понад 2000 твітів призвело до зменшення ефективності, можливо, через збільшення шуму або різноманітність мовних шаблонів, що ускладнило ефективне узагальнення моделі. Це підкреслює важливість ретельного балансу між розміром і якістю набору даних під час навчання.

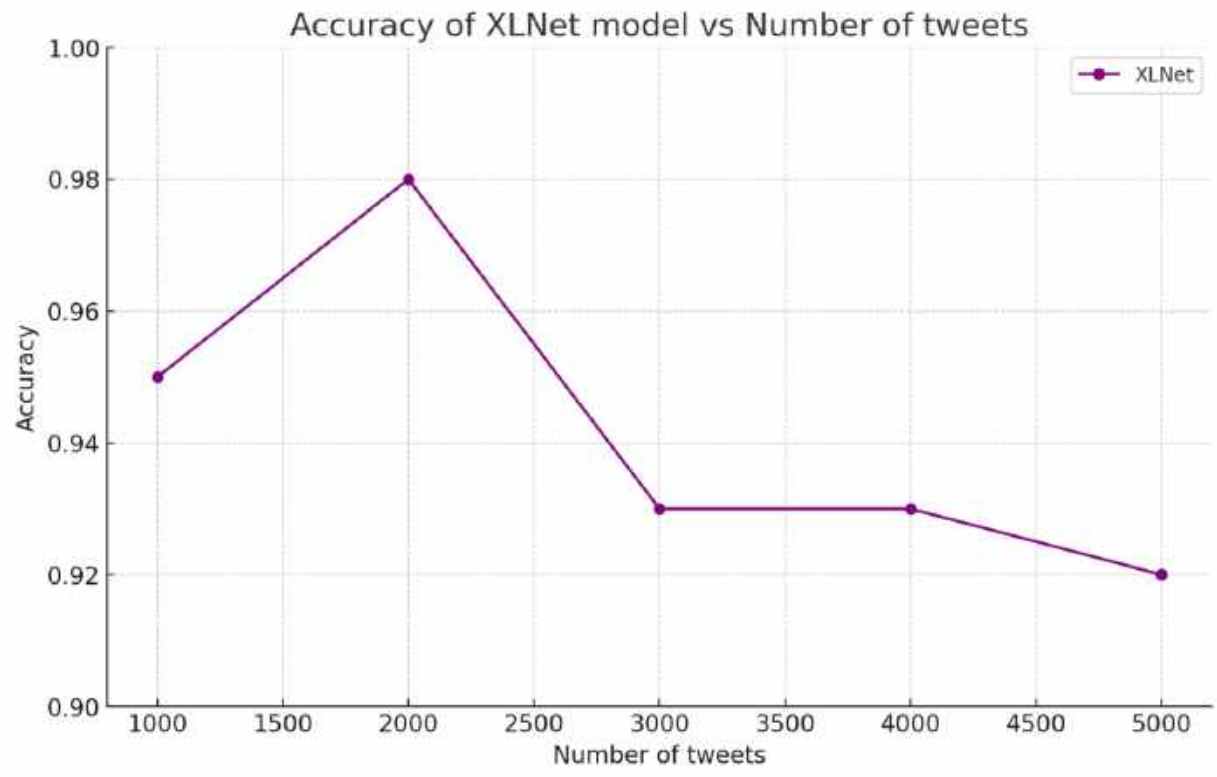

Рисунок 3.6 – Точність вдосконаленої моделі на основі архітектури XLNet

3.3. Вдосконалені моделі класифікації текстової дезінформації на основі архітектур LSTM, BiLSTM та Attention Based BiLSTM

3.3.1. Аналіз архітектур моделей глибокого навчання

1. LSTM – це особливий вид RNN, здатний навчатися на довготривалих залежностях у даних [128]. Традиційні RNN страждають від проблеми затухання або вибуху градієнта, що ускладнює навчання на даних, де для розуміння майбутніх точок даних необхідна інформація з минулого [129]. LSTM було розроблено для подолання цього обмеження, і вони добре підходять для класифікації, обробки та прогнозування на основі часових рядів.

Мережа LSTM складається з пам’яткових комірок, розташованих у рекурентному прихованому шарі, які часто називають одиницями або вузлами [128]. Кожна комірка пам’яті має три основні компоненти: вхідні ворота, ворота забуття та вихідні ворота, а також стан комірки. Ці ворота (гейти) та стан комірки працюють разом, дозволяючи LSTM зберігати або забувати інформацію в довгих послідовностях даних:

- Вхідні ворота. Визначають, скільки нового вхідного сигналу слід додати до стану комірки. Складаються з сигмоїдної (sigmoid) функції активації, яка стискає значення в діапазоні від 0 до 1, та тангенціальної (tanh) функції активації, що стискає значення між -1 і 1. Сигмоїдна функція визначає, які значення пропустити (0 означає “не пропускати нічого”, 1 означає “пропустити все”), а tanh задає вагу переданих значень, які потім додаються до стану комірки.
- Ворота забування Визначають, скільки інформації з поточного стану комірки слід забути або зберегти. Містять сигмоїдну функцію активації, яка стискає значення в діапазоні від 0 до 1. Значення, близькі до 0, означають забування, а близькі до 1 – збереження.
- Вихідні ворота. Визначають, скільки інформації з поточного стану комірки слід передати на наступний шар. Складаються з сигмоїдної функції активації (яка стискає значення між 0 і 1) та функції tanh, яка застосовується до стану комірки (стискаючи значення між -1 і 1). Вихідний сигнал отримується як добуток цих двох результатів.
- Стан комірки. Виступає як “пам’ять” блоку LSTM. Це шлях, що проходить через усі LSTM-елементи з мінімальними лінійними взаємодіями. Оновлюється воротами забування та вхідними воротами.

На кожному кроці часу блок LSTM отримує вхідні дані, попередній прихований стан і попередній стан комірки:

- Ворота забування вирішують, які частини стану комірки слід забути.

- Вхідні ворота визначають, які значення з вхідного сигналу слід використати для оновлення стану комірки. Після цього стан комірки оновлюється за рахунок видалення непотрібних даних і додавання нових.
- Вихідні ворота визначають, які частини стану комірки слід передати як прихований стан для поточного кроку часу.

Рисунок 3.7 ілюструє структуру моделі LSTM.

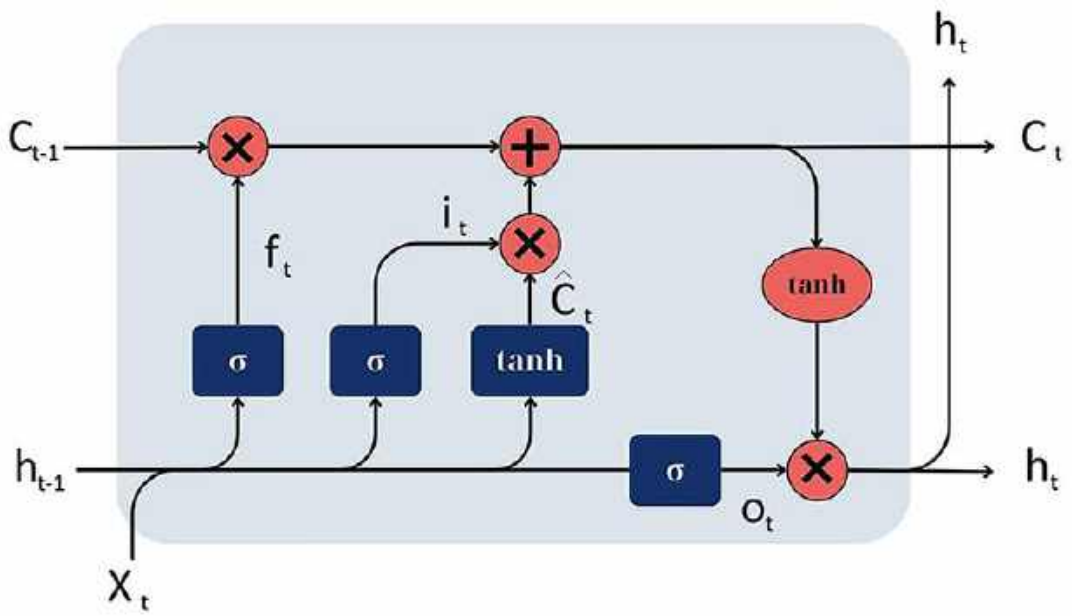

Рисунок 3.7 – структура моделі LSTM

Переваги моделі LSTM:

- Захоплення контекстуальної інформації. Фейкові новини часто містять тонкі підказки й вимагають розуміння контексту протягом послідовності слів або речень. LSTM здатні захоплювати довгострокові залежності та контекстуальну інформацію, що є критично важливим для точної класифікації фейкових новин.
- Обробка послідовностей змінної довжини. Новинні статті можуть значно відрізнятися за довжиною. LSTM можуть обробляти послідовності різної довжини, що робить їх підходящими для класифікації новинних статей різної довжини.

- Зменшення проблеми затухання та вибуху градієнта. LSTM стійкі до цієї проблеми, роблячи їх більш стабільними та ефективними в навчанні складних закономірностей у тексті.

Недоліки моделі LSTM:

- Обчислювальна складність. LSTM мають складну структуру з кількома воротами й станом комірки, що підвищує обчислювальну складність моделі. Це робить їх інтенсивними щодо ресурсів і часу для тренування.
- Ризик перенавчання. LSTM мають багато параметрів, що збільшує ризик перенавчання, особливо коли даних обмаль.
- Інтерпретація. Як і інші моделі глибокого навчання, LSTM мають низьку інтерпретованість, що може бути недоліком у завданнях, де важливо розуміти причини певної класифікації.
- Залежність від даних. Ефективність моделей LSTM значною мірою залежить від якості та кількості навчальних даних. Якщо навчальні дані не є репрезентативними для реальних даних або їх недостатньо, модель може працювати неефективно. Це є суттєвою проблемою для класифікації фейкових новин, оскільки фейкові новини постійно змінюються, а отримання репрезентативного та всебічного набору даних може бути складним завданням.

2. BiLSTM

Двоспрямована довга короткострокова пам’ять (BiLSTM) є важливою архітектурою для завдань класифікації послідовностей, оскільки вона допомагає покращити продуктивність моделі [130].

BiLSTM є важливою архітектурою для завдання класифікації фейкових новин, оскільки допомагає покращити продуктивність моделі на проблемах класифікації послідовностей. У класифікації фейкових новин усі кроки вхідної послідовності доступні; BiLSTM навчає дві LSTM на вхідній послідовності – першу на послідовності в оригінальному вигляді, а другу – на зворотній копії вхідної послідовності. Виходи на одному кроці з обох LSTM потім об’єднуються. Це надає додатковий контекст мережі, що призводить до

швидшого та повнішого навчання на проблемі, що є важливим для точної класифікації фейкових новин.

- Шар прямої LSTM. Цей шар обробляє вхідну послідовність (статтю новин) від початку до кінця. Він захоплює контекстуальну інформацію від минулого до поточного кроку.
- Шар зворотної LSTM. Цей шар обробляє вхідну послідовність (статтю новин) від кінця до початку. Він захоплює контекстуальну інформацію від майбутнього до поточного кроку.
- Об’єднання. Виходи шарів прямої і зворотної LSTM на кожному кроці об’єднуються і передаються на наступний шар. Це забезпечує повний огляд вхідної послідовності, захоплюючи минулу і майбутню контекстуальну інформацію на кожному кроці.

На кожному кроці пряма LSTM обробляє поточний вхід і попередній прихований стан, тоді як зворотна LSTM обробляє поточний вхід і наступний прихований стан. Виходи обох LSTM потім об’єднуються і передаються на наступний шар. Це дозволяє BiLSTM захоплювати контекстуальну інформацію як з минулого, так і з майбутнього на кожному кроці, що є важливим для точної класифікації фейкових новин, оскільки це часто вимагає тонких підказок і розуміння контексту протягом послідовності слів або речень.

На рисунку 3.7 наведено архітектуру розробленої моделі на основі BiLSTM.

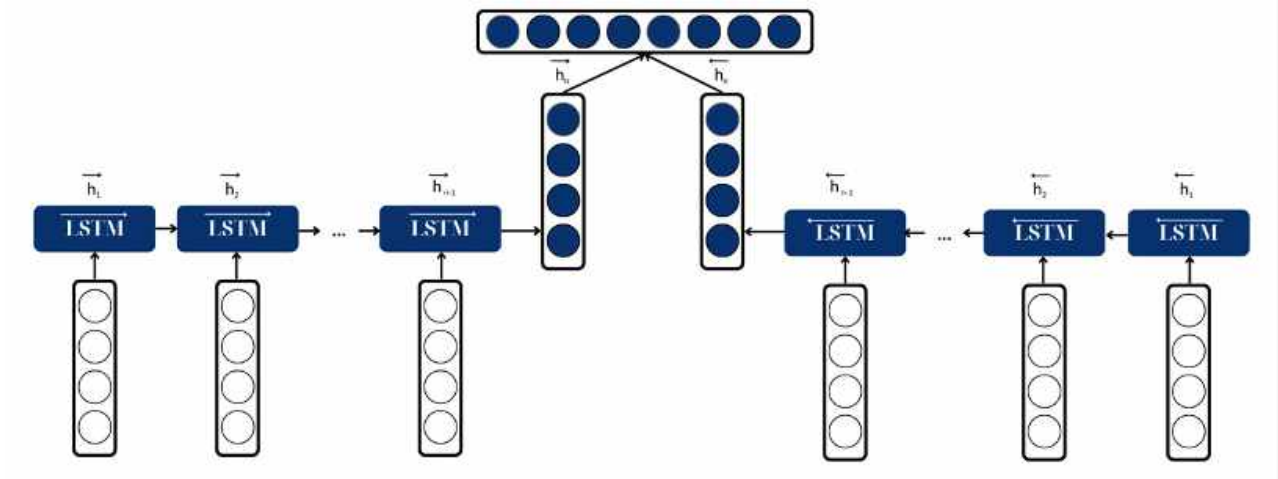

Рисунок 3.7 – Архітектура вдосконаленої моделі BiLSTM

Переваги моделі BiLSTM:

- Захоплення контексту з минулого і майбутнього. BiLSTM може захоплювати контекстуальну інформацію як з минулого, так і з майбутнього на кожному кроці, що є важливим для точної класифікації фейкових новин.
- Краще опрацювання довготривалих залежностей. Обробляючи вхідну послідовність (статтю новин) у прямому і зворотному напрямках, BiLSTM краще захоплює довготривалі залежності в даних, що є критичним для точної класифікації фейкових новин, оскільки це часто включає тонкі підказки і вимагає розуміння контексту в межах послідовності слів або речень.

BiLSTM особливо добре підходить для завдання класифікації фейкових новин, оскільки вимагає захоплення контексту з минулого і майбутнього, як це потрібно, наприклад, у машинному перекладі, розпізнаванні іменованих сутностей та аналізі настроїв. Вони широко використовуються в різних додатках, таких як обробка природної мови, розпізнавання мови і прогнозування часових рядів.

Архітектура запропонованої моделі наведена на рисунку 3.8.

3. Attention-based BiLSTM

Модель на основі двоспрямований довгої короткострокової пам’яті з механізмом уваги (Attention-based BiLSTM) поєднує переваги як BiLSTM, так і механізму уваги, створюючи більш потужну модель для завдань класифікації послідовностей [131].

Модель Attention-based BiLSTM складається з трьох основних компонентів:

- Шар BiLSTM. Він обробляє вхідну послідовність (наприклад, новинну статтю) як у прямому, так і у зворотному напрямках, захоплюючи минулу та майбутню контекстуальну інформацію на кожному часовому кроці.
- Механізм уваги – це ключовий компонент моделі. Механізм уваги дозволяє моделі фокусуватися на різних частинах вхідної послідовності під час формування вихідної послідовності, фактично зважуючи важливість різних частин. У контексті класифікації фейкових новин модель може

фокусуватися на найважливіших словах або реченнях, які вказують на те, чи є новина фейковою чи правдивою.

- Шар класифікації – це фінальний шар моделі, який бере зважену суму виходів BiLSTM (згенерованих механізмом уваги) та формує кінцеву класифікацію (фейкова чи правдива).

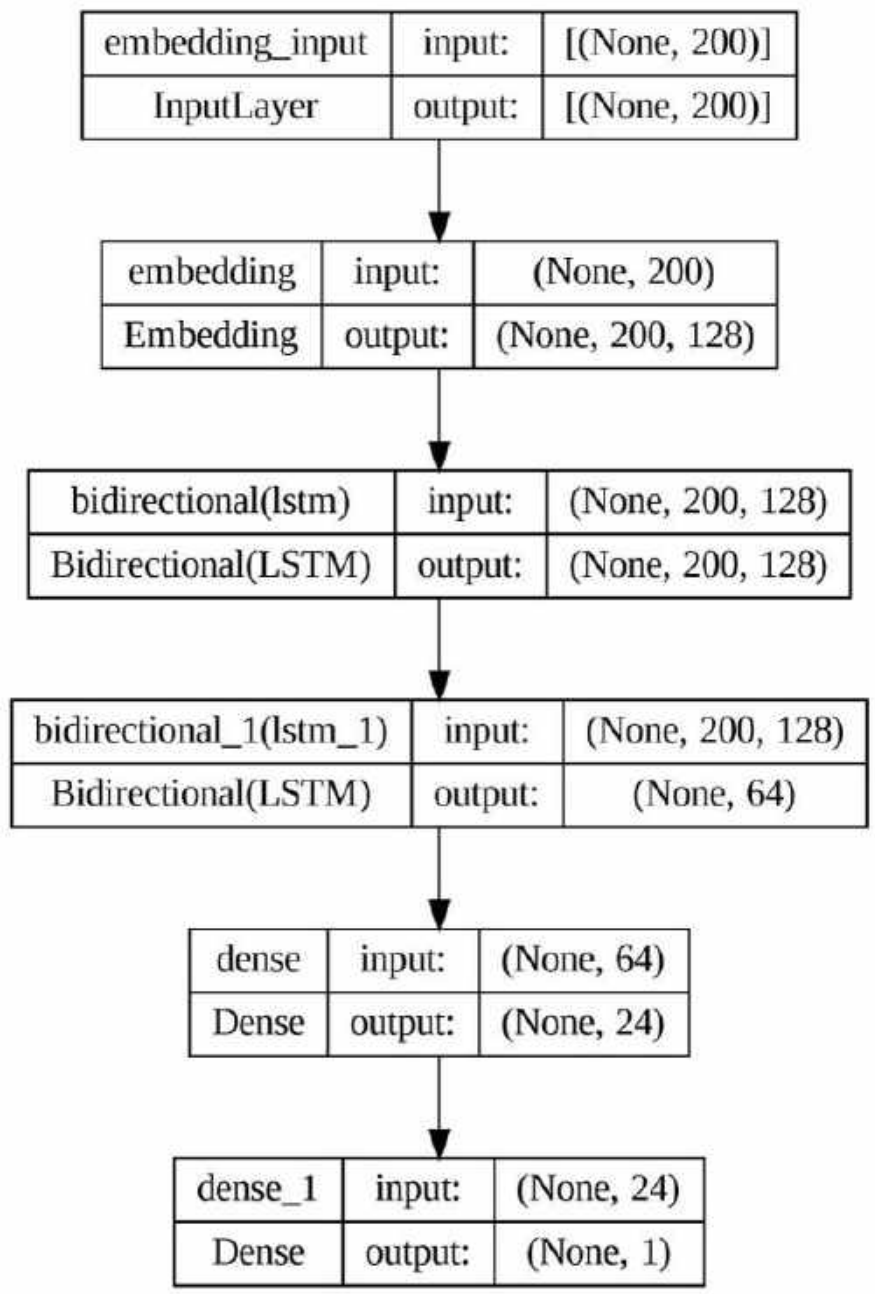

Рисунок 3.8 – Архітектура розробленої моделі

Процес обробки вхідної послідовності в моделі Attention-based BiLSTM відбувається в таких кроках [132]:

- Вхідна послідовність передається через шар BiLSTM, який обробляє її в обидва напрямки (вперед і назад), генеруючи набір прихованих станів для кожного часового кроку.
- Приховані стани, згенеровані шаром BiLSTM, передаються через механізм уваги, який створює зважену суму прихованих станів. Ця зважена сума є єдиним вектором, який узагальнює вхідну послідовність, причому важливіші частини послідовності отримують вищу вагу.
- Зважена сума, згенерована механізмом уваги, передається через шар класифікації, який формує кінцеву класифікацію (фейкова чи правдива).

Архітектура розробленої моделі на базі Attention-based BiLSTM наведена на рисунку 3.9.

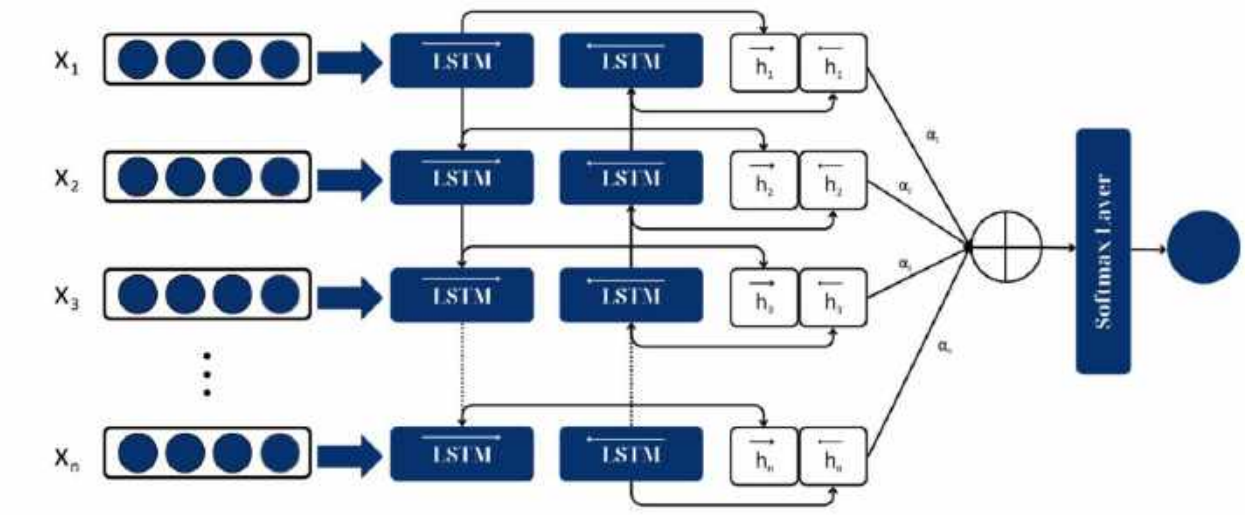

Рисунок 3.9 – Архітектура Attention-based BiLSTM

Переваги:

- Фокусування на важливих частинах вхідних даних. Механізм уваги дозволяє моделі фокусуватися на найважливіших словах або реченнях новинної статті, які вказують на її фейковість або правдивість. Це надзвичайно важливо для

точної класифікації фейкових новин, оскільки часто потрібне розуміння контексту протягом всієї послідовності слів або речень.

- Краще опрацювання довготривалих залежностей. Шар BiLSTM дозволяє моделі захоплювати довготривалі залежності у вхідній послідовності, що є критичним для точної класифікації фейкових новин, оскільки часто потрібне розуміння контексту протягом всієї послідовності.

Модель BiLSTM на основі механізму уваги особливо добре підходить для завдання класифікації текстової дезінформації та пропаганди, оскільки вона вимагає захоплення контексту з минулого і майбутнього, а також фокусування на найбільш критичних частинах вхідної послідовності. Розроблену модель також можна застосовувати до інших завдань класифікації послідовностей, таких як аналіз настроїв, розпізнавання іменованих сутностей і машинний переклад.

Архітектура розробленої моделі показана на рисунку 3.10.

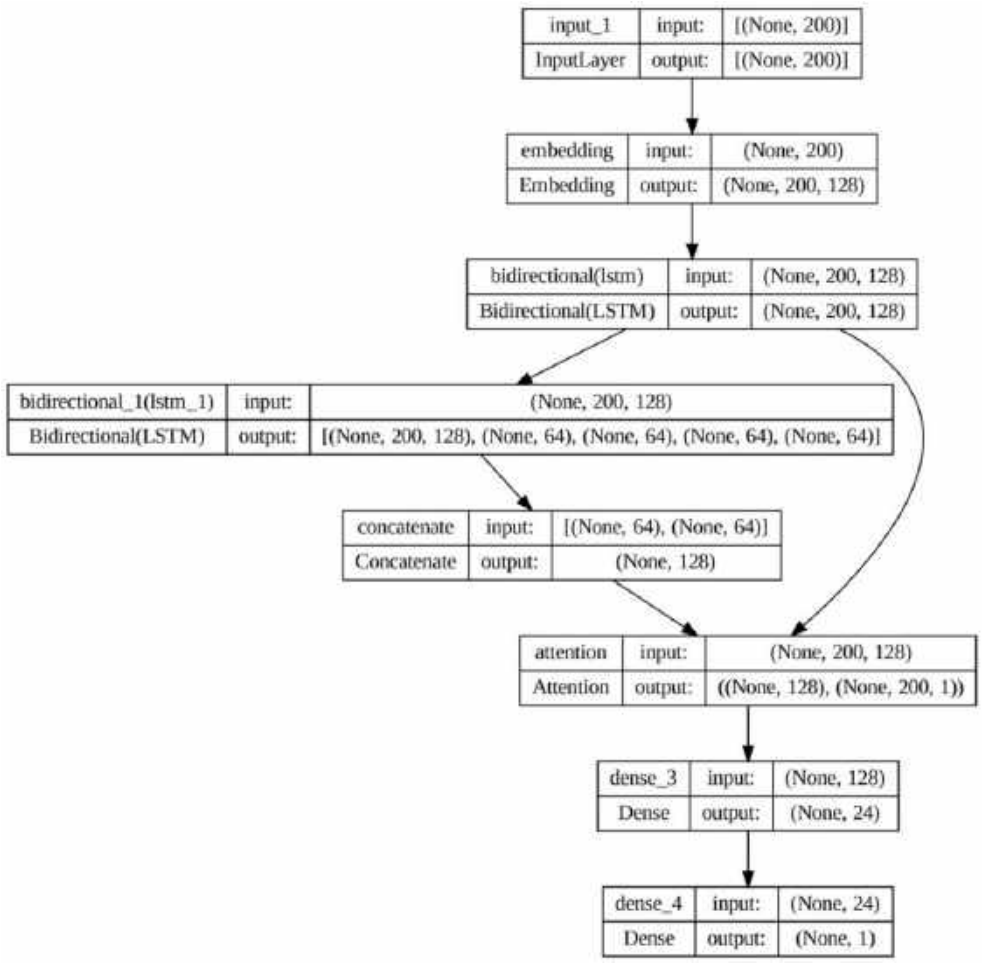

Рисунок 3.10 – Архітектура розробленої моделі на основі архітектури Attention-based BiLSTM

3.3.2. Результати продуктивності моделей на основі архітектур BiLSTM та Attention-based BiLSTM

Результати продуктивності моделей глибокого навчання для кластеризації даних перевірялись на наборі даних WELFake [124], який був використаний для аналізу ансамблевих методів у попередньому підрозділі 2.3.4 та описаний в Додатку Г.

Процес навчання розроблених моделей BiLSTM та Attention-based BiLSTM для класифікації фейкових новин включав кілька ключових етапів. По-перше, набір даних було поділено на навчальний та тестовий набори, при цьому 80% даних використовували для навчання моделі, а 20% – для тестування її продуктивності. Навчальні дані було попередньо оброблено: текст було токенізовано, видалено стоп-слова, а послідовності доповнено для забезпечення однакової довжини.

Для моделі на основі BiLSTM розробка включала створення шарів для ембедингу, двоспрямованої LSTM та щільного виходу. Для моделі на основі Attention-based BiLSTM було додано додатковий шар уваги між BiLSTM та щільним шаром виходу. Модель було скомпільовано, використовуючи оптимізатор, функцію втрат та метрики оцінки для навчання. Модель була натренована на тренувальних даних протягом визначеної кількості епох, використовуючи розмір пакета, який визначав кількість вибірок, що використовуються в кожній ітерації для оновлення ваг моделі. Під час навчання продуктивність моделі контролювалася на валідаційному наборі, який є підмножиною навчальних даних і не використовувався для оновлення ваг моделі. Це допомогло запобігти перенавчанню та забезпечити, що модель добре узагальнюється на нових даних.

Продуктивність моделей показана в Таблиці 3.3.

Незважаючи на те, що обидві моделі показали високу продуктивність на навчальних і тестових наборах даних, модель Attention-based BiLSTM загалом показала невелику перевагу в більшості метрик, особливо на навчальних

даних. Однак різниця між продуктивністю двох моделей на тестових даних була незначною, що свідчить про те, що обидві моделі є надійними та ефективними для класифікації фейкових новин.

Матриця невідповідностей для обох моделей представлена на Рисунку 3.11, Динаміка втрат та точності для обох моделей показана на Рисунку 3.12, порівняння продуктивності обох моделей наведено на Рисунку 3.13.

Порівняльний аналіз моделей BiLSTM і Attention-based BiLSTM для класифікації фейкових новин (Рис. 3.11, 3.12, 3.13) надає цінні уявлення про ефективність цих архітектур глибокого навчання в ідентифікації правдивих новин серед сфабрикованого контенту.

Таблиця 3.3 – Оцінка продуктивності моделей

Метрики	BiLSTM	Att-BiLSTM
Міра F1 (тест)	0.97481	0.97621
Точність (тест)	0.97491	0.97657
Повнота (тест)	0.98406	0.97666
Прецизійність (тест)	0.96738	0.97702
Втрати (тест)	0.07437	0.07227
Міра F1 (тренування)	0.98922	0.99205
Точність (тренування)	0.98927	0.99199
Повнота (тренування)	0.98954	0.99254
Прецизійність (тренування)	0.98968	0.99209
Втрати (тренування)	0.03131	0.02527

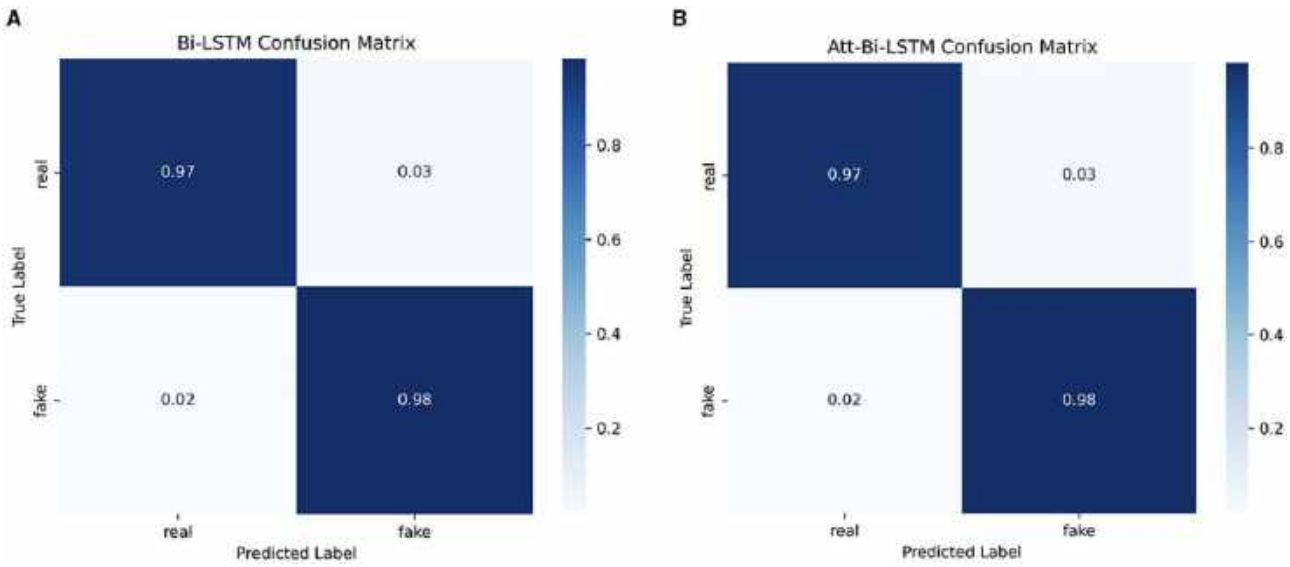

Рисунок 3.12 – Матриця невідповідностей для обох моделей

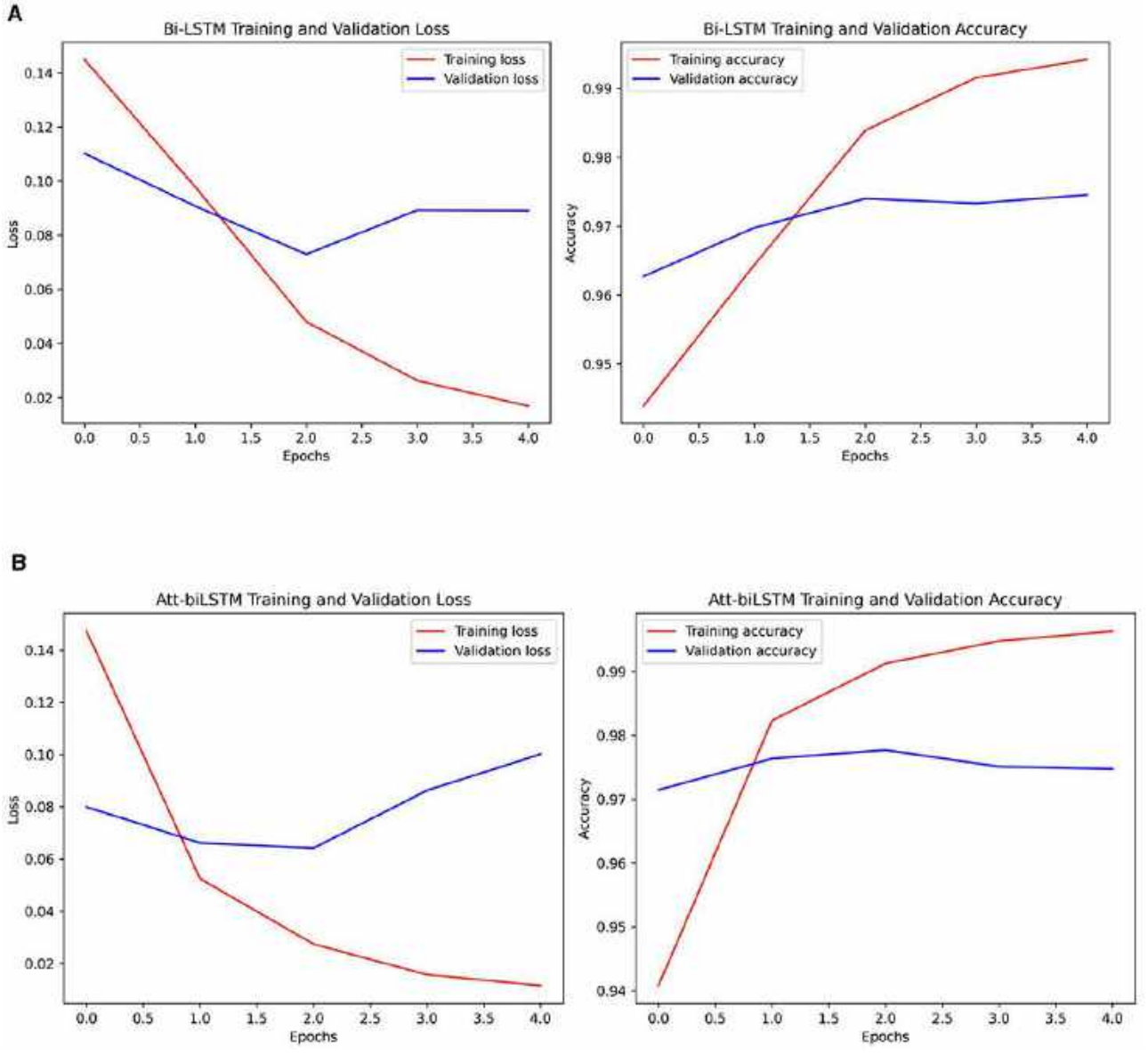

Рисунок 3.12 – Динаміка втрат та точності для обох моделей

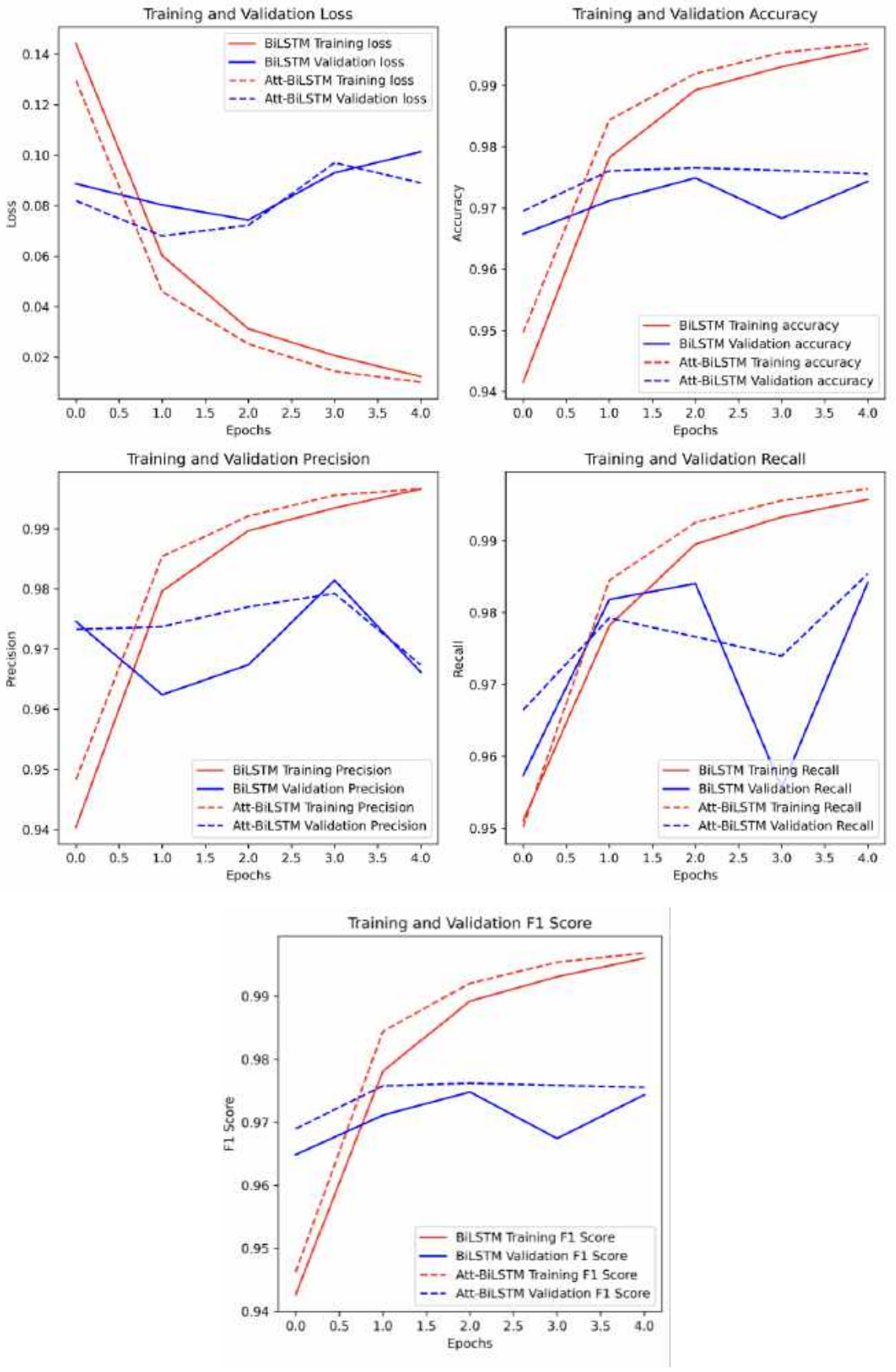

Рисунок 3.13 – Порівняльний аналіз продуктивності моделей

Обидві моделі показали високу продуктивність на тестовому наборі даних. Модель BiLSTM, із чутливістю близько 98.41%, продемонструвала свою перевагу в ідентифікації більшості істиннопозитивних. Однак, незважаючи на дещо нижчу чутливість, Attention-based BiLSTM продемонструвала вищу прецизійність, що свідчить про меншу кількість хибно позитивних результатів. Така прецизійність є вирішальною для виявлення фейкових новин, оскільки хибнопозитивна класифікація може мати суттєві наслідки. Невелика різниця у мірі F1 та точності між двома моделями свідчить про те, що обидві моделі забезпечують збалансоване співвідношення між точністю та чутливістю. Значення втрат додатково підтверджують надійність моделей, причому модель Attention-based BiLSTM має невелику перевагу.

Результати тренувального набору даних підкреслюють здатність моделей навчатися та узагальнювати з тренувальних даних. Обидві моделі досягли високих показників прецизійності та чутливості, при цьому модель Attention-based BiLSTM трохи перевершила BiLSTM. Вища точність та менші втрати моделі Attention-based BiLSTM на навчальних даних свідчать про її підвищену здатність точно узгоджуватися з даними без перенавчання, враховуючи її продуктивність на тестових даних.

Невелика перевага моделі Attention-based BiLSTM може бути пояснена інтеграцією механізму уваги. Механізми уваги дозволяють моделям зосереджуватися на конкретних частинах вхідних даних, які є більш релевантними для завдання. У контексті класифікації фейкових новин модель може надавати більшу вагу певним фразам або шаблонам у змісті новин, які вказують на їхню автентичність. Такий нюансований підхід може пояснити перевагу Attention-based BiLSTM, особливо в прецизійності.

3.4. Запропонована архітектура моделі класифікації текстової дезінформації та пропаганди

3.4.1. Опис запропонованої архітектури моделі класифікації текстової дезінформації та пропаганди

В дисертаційному дослідженні вперше запропонована архітектура моделі класифікації текстових даних для виявлення дезінформації та пропаганди. Вона інтегрує попередньо навчені ембединги, послідовне моделювання, механізми уваги та щільні шари, забезпечуючи надійну структуру для виділення змістовних ознак і генерації точних прогнозів, архітектура моделі зображена на рисунку 3.14.

Вхідний шар моделі приймає токенізовані послідовності, отримані за допомогою токенізатора BERT. Вхідні дані складаються з двох компонентів:

- input_ids – тензор, що містить ідентифікатори токенів вхідної послідовності з максимальною довжиною 128 токенів;
- attention_mask – тензор, що вказує, які токени є значущими (1 для корисних токенів і 0 для заповнювачів).

Ці вхідні дані необхідні для розрізнення фактичного змісту і токенів-заповнювачів, що забезпечує ефективну обробку BERT-моделлю.

Основою архітектури є попередньо навчена модель BERT-base-uncased, яка виступає як інструмент для вилучення ознак. Вона обробляє вхідні послідовності й генерує контекстуальні ембединги у вигляді last_hidden_state – тензора розміру (None, 128, 768). Ці ембединги фіксують семантичні та синтаксичні взаємозв’язки між токенами в контексті вхідного тексту. Використання BERT гарантує, що модель отримує переваги великомасштабного попереднього навчання, що забезпечує насичені та повноцінні представлення токенів.

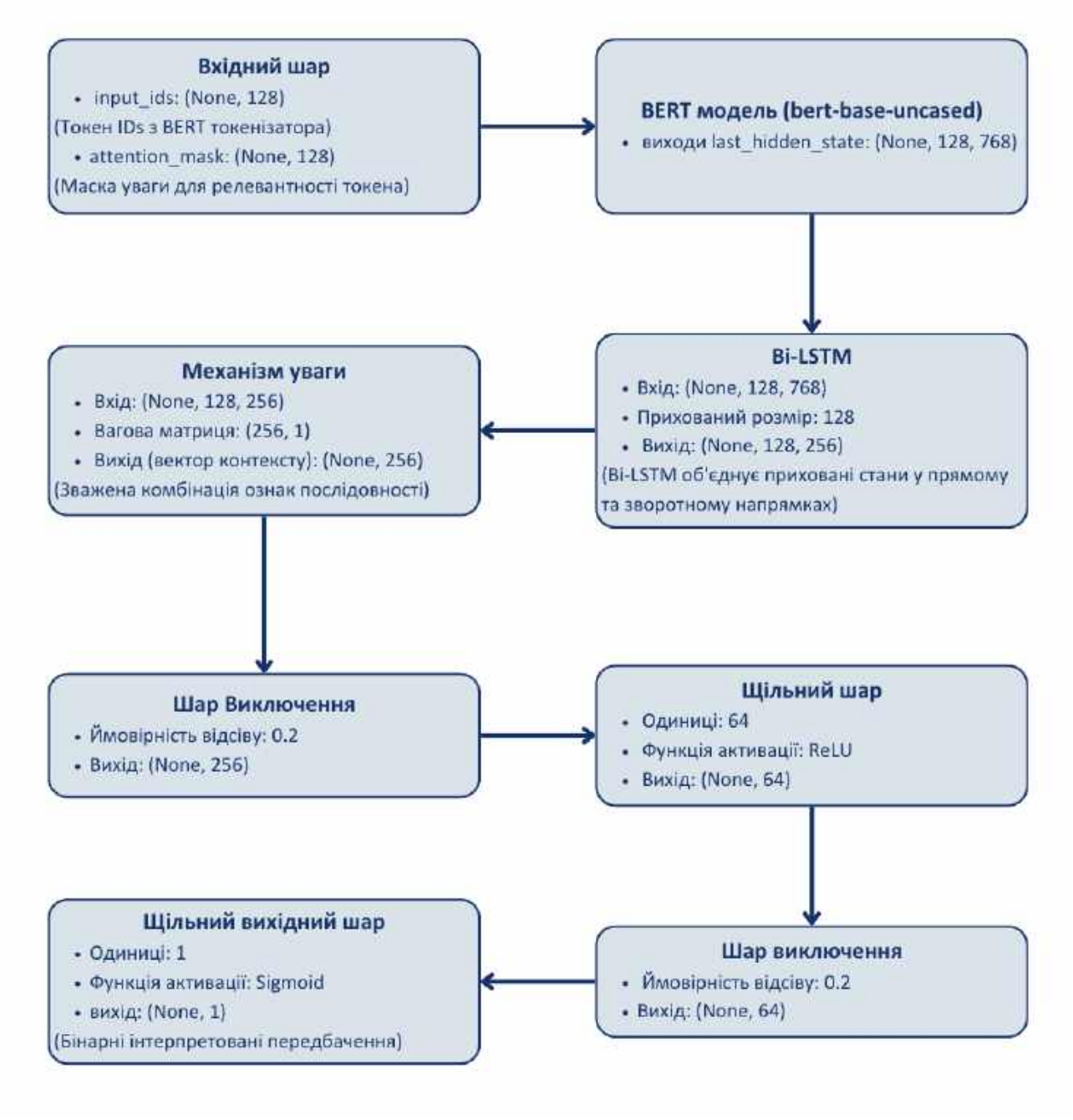

Рисунок 3.14 – Архітектура моделі класифікації текстової дезінформації та пропаганди

Вхідний шар моделі приймає токенізовані послідовності, отримані за допомогою токенізатора BERT. Вхідні дані складаються з двох компонентів:

- input_ids – тензор, що містить ідентифікатори токенів вхідної послідовності з максимальною довжиною 128 токенів;
- attention_mask – тензор, що вказує, які токени є значущими (1 для корисних токенів і 0 для заповнювачів).

Ці вхідні дані необхідні для розрізнення фактичного змісту і токенів-заповнювачів, що забезпечує ефективну обробку BERT-моделлю.

Основою архітектури є попередньо навчена модель BERT-base-uncased, яка виступає як інструмент для вилучення ознак. Вона обробляє вхідні послідовності й генерує контекстуальні ембединги у вигляді last_hidden_state – тензора розміру (None, 128, 768). Ці ембединги фіксують семантичні та синтаксичні взаємозв'язки між токенами в контексті вхідного тексту. Використання BERT гарантує, що модель отримує переваги великомасштабного попереднього навчання, що забезпечує насичені та повноцінні представлення токенів.

Ембединги, згенеровані BERT, передаються у Bi-LSTM шар для покращення розуміння послідовності. Bi-LSTM обробляє ембединги з прихованим розміром 128 у прямому та зворотному напрямках, створюючи вихідний тензор розміру (None, 128, 256). Цей крок дозволяє моделі захоплювати залежності по всій послідовності, включаючи як попередній, так і майбутній контексти, що є критично важливим для розуміння складних зв'язків у текстових даних.

Для подальшого уточнення вихідних даних Bi-LSTM застосовується механізм уваги. Шар уваги бере вихідний тензор Bi-LSTM розміру (None, 128, 256) і обчислює оцінки уваги за допомогою вагової матриці розміру (256, 1). Ці оцінки визначають значущість різних токенів у послідовності, створюючи вектор контексту розміру (None, 256), який концентрується на найважливіших частинах тексту. Цей механізм покращує інтерпретованість моделі та гарантує, що вона надає пріоритет релевантним ознакам для класифікації.

Вектор контексту потім проходить через щільний шар для ущільнення ознак та нелінійної трансформації. Перед щільним шаром застосовується шар виключення із ймовірністю 0.2, щоб зменшити перенавчання. Щільний шар складається з 64 одиниць із функцією активації ReLU, що створює вихідний тензор розміру (None, 64). Після нього додається ще один шар виключення із такою ж ймовірністю для покращення узагальнюючої здатності моделі. Ці

шари забезпечують збереження надійних ознак, зменшуючи ризик перенавчання під час тренування.

Нарешті, модель формує прогноз через вихідний щільний шар, налаштований на бінарну класифікацію. Цей шар містить одну одиницю із сигмоїдною функцією активації, яка відображає вихідне значення у діапазоні ймовірностей [0,1]. Підсумковий тензор розміру (None, 1) представляє ймовірність для вхідної послідовності, що дозволяє моделі робити інтерпретовані бінарні передбачення (наприклад, “правда” або “фейк”).

Гіперпараметри моделі були налаштовані за допомогою Optuna – ефективної платформи оптимізації, яка використовує деревоподібний оцінювач Парцена для вивчення простору пошуку і мінімізації валідаційного збитку. Було проведено пошук за наступними гіперпараметрами:

- Швидкість навчання: Досліджена в межах [5×10^{-6}, 5×10^{-4}] на логарифмічній шкалі, з вибором значення 2×10^{-5} на основі стабільності збіжності та продуктивності узагальнення.
- Розмір партії: Оцінено в межах {16, 32, 64, 128}, з вибором значення 16 для найкращого балансу між стабільністю та обчислювальною ефективністю.
- Рівень виключення: Досліджено в межах [0.1, 0.5] з кроком 0.05, з вибором 0.2 для запобігання перенавчанню при збереженні виразності.
- Розмір прихованого шару Bi-LSTM: Налаштовано в межах {64, 128, 256, 512}, з вибором 128 для оптимального балансу між складністю моделі та продуктивністю.
- Кількість шарів Bi-LSTM: Досліджено в межах {1, 2, 3}, з одним шаром, що показав найкращу продуктивність в контексті узагальнення та обчислювальних витрат.
- Розмір щільного шару: Оцінено в межах {64, 128, 256}, з вибором значення 64 для оптимальної трансформації ознак.
- Вибір оптимізатора: Порівняно Adam, AdamW та RMSprop, з вибором Adam через його вищу швидкість збіжності.

Пошук гіперпараметрів був проведений у 100 спробах з раннім зупиненням на основі валідаційного збитку. Баєсівська оптимізація використовувалася для поступового уточнення обіцяючих конфігурацій. Вибрані гіперпараметри стабільно перевершували альтернативні налаштування на кількох наборах даних.

Ця архітектура поєднує потужність попередньо навчених ембедингів, послідовне моделювання та механізм уваги, забезпечуючи точні та інтерпретовані передбачення для задач класифікації текстової дезінформації та пропаганди.

3.4.2. Продуктивність запропонованої моделі класифікації текстової дезінформації та пропаганди

Запропонована модель була протестована на наборі даних CoVID19-FNIR місінформації [133], який описано в Додатку Г. Модель була тонко налаштована за допомогою ретельно підібраних гіперпараметрів для забезпечення найкращої продуктивності. Використовувався оптимізатор Adam із швидкістю навчання 2×10^{-5}, що сприяло плавній збіжності під час навчання.

Навчання проводилося з розміром пакета 16, що забезпечувало ефективність обчислень, а загальна кількість епох становила 10. Щоб запобігти перенавчанню, було застосовано виключення із ймовірністю 0.2, що покращило узагальнюючу здатність моделі.

Максимальна довжина послідовності була встановлена на 128 токенів, що дозволило зберегти контекст і водночас мінімізувати обчислювальні витрати.

Оцінка продуктивності запропонованої моделі представлена у таблиці 3.4.

Таблиця 3.4 – Оцінка продуктивності запропонованої моделі

Метрика	Значення
Втрати	0.01698
Точність	0.99473
Прецизійність	0.99472
Повнота	0.99472
Міра F1	0.99472

Результати показують, що запропонована модель ефективно виявляє дезінформацію. Втрати на тестових даних – 0.01698, що свідчить про мінімальну похибку передбачення. Точність – 99.47%, що вказує на високу здатність правильно класифікувати вибірки. Прецизійність, повнота та міра F1 – 99.47%, що демонструє збалансовану продуктивність моделі, мінімізуючи як хибно позитивні, так і хибно негативні передбачення. Ці результати підкреслюють надійність і стабільність моделі для завдань з виявлення дезінформації та пропаганди.

Крива втрат (рис. 3.15) демонструє поступове зменшення втрат як на навчальних, так і на валідаційних даних протягом усіх епох. Відсутність різкого зростання втрат свідчить про ефективний процес навчання без ознак перенавчання.

Крива точності на рисунку 3.16 показує стабільне покращення продуктивності протягом навчання. На перших епохах точність швидко зростала, а в пізніх етапах наблизилася до 99%. Валідаційна точність залишалася стабільною та досягла 99.47% до фінальної епохи. Це підтверджує, що модель успішно узагальнює закономірності та добре працює з невидимими даними.

Матриця невідповідностей на рисунку 3.17 демонструє майже ідеальну класифікацію. Лише 2 хибно позитивних та 2 хибно негативних передбачення зі 759 тестових вибірок. Це підтверджує високу точність тестування – 99.47%. Збалансованість між прецизійністю та повнотою робить модель надійною для розрізнення правдивої інформації та дезінформації.

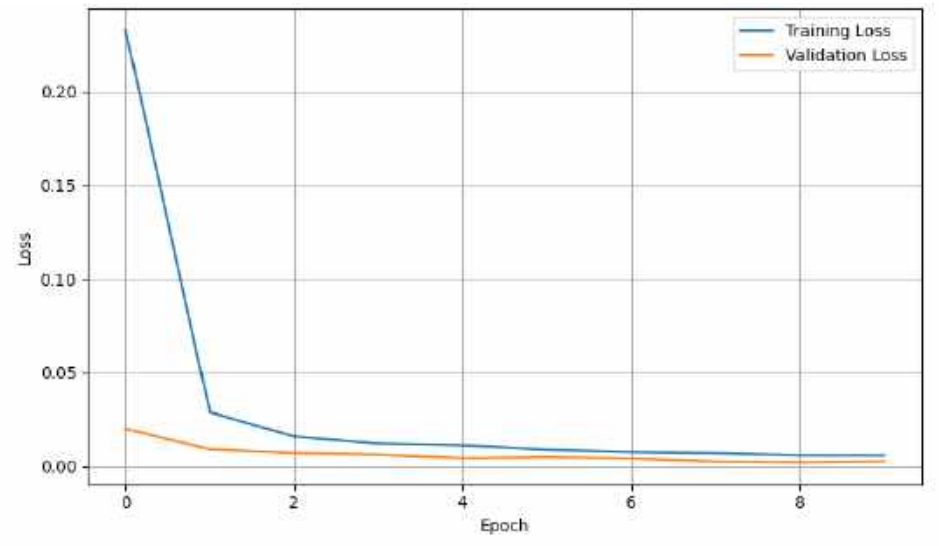

Рисунок – 3.15 Крива втрат для запропонованої моделі

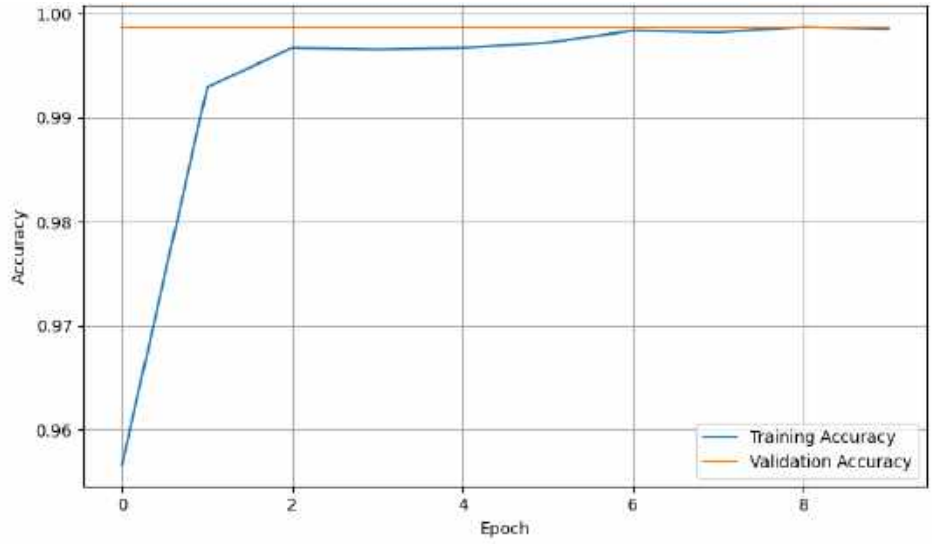

Рисунок 3.16 – Крива точності для запропонованої моделі

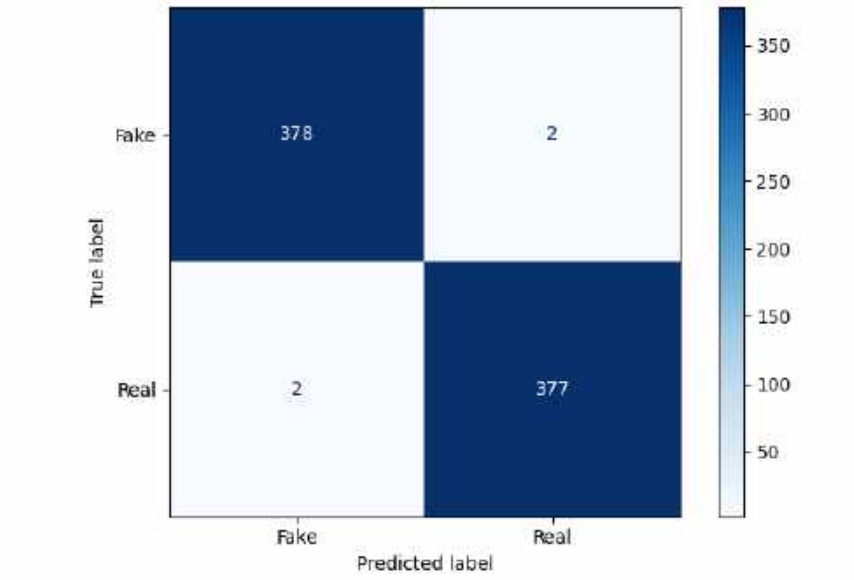

Рисунок 3.17 – Матриця невідповідності для запропонованої моделі

ROC-крива на рисунку 3.18. використовується для оцінки роздільної здатності моделі між позитивними та негативними класами. AUC = 1.0, що підтверджує високу точність у розрізненні між дезінформацією та правдивою інформацією.

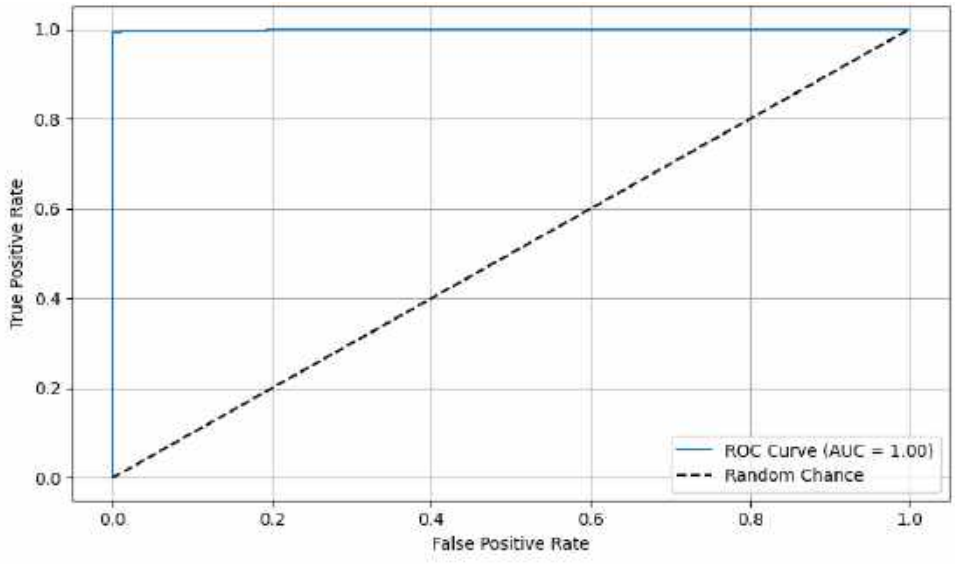

Рисунок 3.18 – ROC-крива для запропонованої моделі

Результати продуктивності моделі показують, що вона не тільки точно класифікує фейкові новини, але й добре узагальнює закономірності, що робить її ефективною для реальних застосувань класифікації текстової дезінформації та пропаганди.

3.5. Висновки до третього розділу

1. Було удосконалено моделі глибокого навчання для класифікації текстової дезінформації та пропаганди, засновані на архітектурах XLNet, Bi-LSTM та Attention-Based Bi-LSTM, які на відміну від існуючих застосовують адаптивне оцінювання важливих фрагментів тексту, що дозволяє забезпечувати високий рівень продуктивності класифікації.

2. Вперше розроблено модель класифікації текстової дезінформації та пропаганди, заснований на поєднанні трансформерів і архітектури Bi-LSTM

для інтеграції лінійних та нелінійних залежностей текстових даних, яка на відміну від існуючих використовує глибокі контекстуальні ембедінги і механізми уваги, що дозволяє підвищити точність моделі. Шляхом змін налаштувань та покращення характеристик моделі запропонована нами архітектура досягла точності 97% у класифікації даних.

3. Проведено експериментальні дослідження моделі машинного навчання BERT для класифікації набору фейкових новин “Fake-Real News".

4. Проведено експериментальні дослідження вдосконаленої моделі машинного навчання XLNet для аналізу російської пропаганди про війну в Україні у X (Twitter).

5. Проведено експериментальні дослідження вдосконалених моделей класифікації текстової дезінформації на основі архітектур LSTM, BiLSTM та Attention Based BiLSTM для аналізу набору даних фейкових та реальних новин “WELFake”.

Основні результати, представлені в цьому розділі, опубліковані в [15], [17], [20].

РОЗДІЛ 4. ЗАСТОСУВАННЯ РОЗРОБЛЕНИХ МЕТОДІВ ТА ФРЕЙМОВКУ ДЛЯ ЗАДАЧ АНАЛІЗУ ТЕКТОВОЇ ДЕЗІНФОРМАЦІЇ ТА ПРОПАГАНДИ НА РЕАЛЬНИХ НАБОРАХ ДАНИХ

4.1. Фреймворк для комплексного аналізу текстової дезінформації та пропаганди

Цей розділ дисертаційного дослідження пропонує комплексний фреймворк, що інтегрує класифікацію та кластеризацію даних через тематичне моделювання для аналізу текстової дезінформації та пропаганди. Фреймворк розроблений для аналізу даних із соціальних мереж, веб-сторінок та інших цифрових платформ з метою виявлення ключових наративів, тем та емоційних відтінків, закладених у даних. Тип інформації (наприклад, правдива чи хибна) визначається за допомогою класифікації, тоді як тематичне моделювання дозволяє виявити тематичні кластери та загальні тенденції. Такий інтегрований підхід надає особам, що приймають рішення інструменти для оцінки інформаційного середовища та розробки стратегічних комунікацій на основі доказових інсайтів, що є критично важливим для реагування на кризи, боротьби з дезінформацією та організації ефективних інформаційних кампаній.

Схема фреймворку, зображена на рисунку 4.1, починається з етапу попередньої обробки даних, щоб забезпечити їх чистоту та підготовленість до аналізу. Цей етап включає нормалізацію тексту, що передбачає приведення до нижнього регістру та стандартизацію для спрощення подальшої обробки. Видаляються HTML-теги та URL-адреси, щоб усунути зайві метадані, а також спеціальні символи та пунктуаційні знаки, залишаючи лише змістовний текстовий контент. Однак хештеги зберігаються, оскільки вони часто містять важливі контекстуальні та тематичні підказки. Потім текст токенізується на менші одиниці, такі як слова або субслова, що є необхідним для моделей машинного навчання. Наступним кроком у схемі попередньої обробки є застосування методів надсемплінгу для вирівнювання розподілу міток та вирішення проблеми дисбалансу класів у наборі даних, що покращує продуктивність моделі.

Рисунок 4.1 – Запропонований фреймворк комплексного аналізу текстової дезінформації

Після попередньої обробки дані проходять етап класифікації, щоб визначити, чи є певна інформація правдивою, хибною або належить до інших заздалегідь визначених дослідником класів. Цей етап починається з застосування попередньо натренованої моделі BERT, яка генерує багатоконтекстні векторні представлення тексту. Ці ембедінги передаються через модель Bi-LSTM, що дозволяє враховувати як попередні, так і наступні залежності у тексті, покращуючи контекстне розуміння. Механізм уваги додатково вдосконалює вихідні дані, виділяючи найбільш релевантні фрагменти тексту. Для покращення узагальнюваності моделі та зменшення перенавчання застосовуються шари виключення перед та після щільного шару, що виконує фінальну класифікацію на заздалегідь визначені категорії.

Після класифікації вибірок, ідентифіковані як дезінформація (або інші класи, необхідні для подальшого аналізу), піддаються тематичному моделюванню для виявлення прихованих наративів і тем. Цей процес розпочинається з кодування тексту за допомогою токенових і позиційних кодувань. Закодовані дані обробляються через шари трансформерного кодування для отримання високоякісних текстових ембедингів. Ці ембедінги агрегуються у векторні представлення фіксованого розміру за допомогою механізму середнього згладжування, що допомагає вловити семантичний зміст тексту. Потім застосовуються методи кластеризації для групування схожих текстових вибірок, що дозволяє виявити домінуючі теми, присутні у наборі даних.

Далі фреймворк переходить до представлення тем, де кожному кластеру призначається змістовний ярлик або дескриптор для узагальнення його змісту. Ієрархічний механізм групування організовує теми в ширші категорії, що сприяє виявленню загальних наративів у вибірках дезінформації. Така ієрархічна структура дозволяє відстежувати потік і зв’язки між різними темами, забезпечуючи структуроване розуміння наративів дезінформації.

Останній етап включає якісний аналіз наративів, що синтезує результати класифікації та тематичного моделювання для отримання практичних

висновків. Інтегруючи результати, цей етап надає цілісне розуміння динаміки дезінформації, висвітлюючи механізми її поширення, домінуючі теми та взаємозв’язки між ними. Висновки з такого аналізу можуть бути використані для розробки комунікаційних стратегій, створення цільових інтервенцій та посилення зусиль у боротьбі з маніпуляцією інформацією, тим самим підтримуючи ухвалення обґрунтованих рішень і формування політик.

Описаний фреймворк є цінним інструментом для аналізу великих неструктурованих наборів даних з метою виявлення прихованих шаблонів і наративів. Перетворюючи текстові дані на семантично насичені ембединги та кластеризуючи їх у інтерпретовані теми, цей фреймворк дозволяє отримувати практичні інсайти з інакше непрозорих масивів даних. Модульний дизайн фреймворку забезпечує гнучкість і можливість адаптації під різні сценарії використання. Кроки попередньої обробки можуть бути налаштовані під конкретні мови чи сфери застосування. Модель SentenceTransformer може бути замінена альтернативними моделями ембедингів для спеціалізованих завдань. Компоненти BERTopic (кластеризація, виділення ключових слів) можуть бути оптимізовані під специфічні потреби, наприклад, об’єднання подібних тем або покращення відбору ключових слів.

Так, в якості класифікаційної моделі фрейморку можуть бути застосовані моделі, запропоновані в цьому дисертаційному дослідженні. В розділі 4.2 наведено аналіз продуктивності ансамблевих та нейромережевих моделей, описаних в розділах 2.3 та 3.3. В розділі 4.3 проведено аналіз продуктивності моделі, запропонованої в розділі 3.4. Моделі, наведені в розділах 4.4 та 4.5 можна використовувати в якості кластеризаційних моделей запропонованого фреймворку. Таким чином, в рамках дисертаційного дослідження запропоновано комплексний підхід для аналізу текстової дезінформації.

4.2. Аналіз продуктивності моделей виявлення фейкових новин

Для аналізу ефективності моделей, запропонованих у розділі 2.3 та 3.3 було порівняно результати продуктивності інших моделей опублікованих у науковій літературі, які використовують набір даних WELFake [124] для верифікації ефективності роботи моделей. В таблиці 4.1 наведено порівняльний аналіз продуктивності запропонованих моделей з результатами інших досліджень.

Таблиця 4.1 – Порівняльний аналіз продуктивності запропонованих моделей з результатами інших досліджень

Метрика	Точність (%)	Міра F1 (%)	Прецизійність (%)	Повнота
Att-BiLSTM (запропонована)	**97.66**	**97.62**	**97.7**	**97.67**
LightGBM (запропонована)	**97.64**	**97.74**	**96.92**	**98.58**
BiLSTM (запропонована)	**97.49**	**97.48**	**96.74**	**98.41**
N-Gram with TF-IDF and BERT [134]	96.8	96.3	96.5	97
SVM [135]	96.73	96.56	94.60	98.61
XGBoost (запропонована)	**96.66**	**96.82**	**95.28**	**98.42**
N-Gram with TF-IDF and LSTM [134]	96	95.8	95.5	96.2
Adaboost [135]	95.32	95.02	91.81	98.46
Bagging [135]	95.31	95.00	91.78	98.46
BRF (запропонована)	**94.33**	**94.46**	**93.96**	**94.96**

Продовження таблиці 4.1.

XGBClassifier [136]	94.00	-	-	-
NB [135]	92.12	91.85	91.45	92.25
BiLSTM [137]	92	91.6	91.89	91.31
MLPClassifier [136]	92.00	-	-	-
KNN [135]	90.16	89.78	89.02	90.55
LSTM [137]	90.15	89.3	92.78	86.07
SVM [137]	90.05	89.74	92.16	87.44
DT [135]	89.92	89.24	86.10	92.62
Random Forest [136]	89.00	-	-	-
Memory-based ensemble model [137]	87.3	86.45	86.17	86.72
NN with Keras [137]	86.74	85.79	86.37	85.22
Non memory-based ensemble model [137]	86.39	85.31	88.37	82.46
Non memory-based ensemble model [137]	86.39	85.31	88.37	82.46
BernoulliNB [136]	86.00	-	-	-
Random Forest [137]	84.37	83.29	83.62	82.96
Naive Bayes [137]	75.89	72.83	82.93	64.92
GaussianNB [135]	74.00	-	-	-

Запропоновані моделі демонструють відмінні результати продуктивності порівняно з тими, які висвітлені у літературі та були застосовані на тому ж наборі даних WELFake [124]. Найкращим підходом серед запропонованих є модель Att-BiLSTM, яка досягла точності 97.66%. Ця модель також демонструє високу міру F1 – 97.62%, прецизійність – 97.7% та повноту – 97.67%, що свідчить про її збалансованість у визначенні правильних класів. Модель LightGBM, яка також є однією з запропонованих, має точність 97.64%, трохи нижчу, ніж у Att-BiLSTM, але перевершує її за мірою F1 –

97.74%, що вказує на ще більш стабільне співвідношення між прецизійністю (96.92%) та повнотою (98.58%). Висока повнота свідчить про її здатність ідентифікувати позитивні випадки, зменшуючи кількість пропущених зразків. Ця модель добре підходить для сценаріїв, де критично важливо мінімізувати помилки пропуску. Ще одна запропонована модель – BiLSTM – продемонструвала точність 97.49%, що трохи поступається двом попереднім моделям, але залишається на дуже високому рівні. Її міра F1 склала 97.48%, що підтверджує стабільну роботу в обох класах, прецизійність – 96.74%, а повнота – 98.41%. Використання двоспрямованого підходу дозволяє цій моделі розглядати послідовність даних у двох напрямках, що особливо важливо для текстової обробки.

Серед запропонованих моделей також виділяється XGBoost, яка досягла точності 96.66%. Вона має міру F1 – 96.82%, прецизійність – 95.28% та повноту – 98.42%. Незважаючи на те, що її точність дещо нижча за BiLSTM та LightGBM, вона демонструє високу повноту, що робить її корисною для виявлення випадків, які важливо не пропустити. Модель BRF, яка також входить до числа запропонованих, показала точність 94.33%, що є нижчим за інші моделі в цьому наборі, проте вона все ще перевершує більшість традиційних підходів. Її міра F1 – 94.46%, прецизійність – 93.96%, а повнота – 94.96%, що свідчить про збалансованість у класифікації.

Розглядаючи моделі, висвітлені у літературі, найвищі результати демонструє підхід N-Gram with TF-IDF and BERT [134], який має точність 96.8%, міру F1 – 96.3%, прецизійність – 96.5% та повноту – 97%. Поєднання традиційних методів векторизації тексту із сучасними трансформерними моделями дає потужний результат, хоча він все ще поступається запропонованим моделям, таким як Att-BiLSTM та LightGBM.

Модель SVM [135] демонструє точність 96.73%, що є наближеним до найкращих результатів, проте її прецизійність – 94.60%, що є нижчим показником, порівняно з повнотою – 98.61%. Це означає, що модель

ефективніше знаходить позитивні випадки, але робить більше хибнопозитивних класифікацій.

Інша модель N-Gram with TF-IDF and LSTM [134] досягла точності 96%, що нижче запропонованих моделей. Її міра F1 – 95.8%, прецизійність – 95.5% та повнота – 96.2%. Використання LSTM без додаткових механізмів, таких як увага, дещо знижує продуктивність у порівнянні з BiLSTM. Adaboost [135] і Bagging [135] мають майже однакові результати, з точністю 95.32% та 95.31% відповідно, що нижче, ніж запропоновані моделі. Вони мають міру F1 на рівні 95.02% та 95.00%, що є пристойним результатом, але поступається новітнім підходам.

XGBClassifier [136] досяг 94% точності, хоча інші метрики для цієї моделі не наведені. Це свідчить про її ефективність, але недостатню деталізацію її продуктивності у літературі. Модель Memory-based ensemble model [137] досягла точності 87.3%, що є відносно нижчим показником у порівнянні з запропонованими нейромережевими моделями, такими як BiLSTM та LightGBM. Вона має міру F1 на рівні 86.45%, прецизійність – 86.17%, а повноту – 86.72%, що свідчить про її збалансованість, хоча загальний рівень продуктивності поступається більш сучасним методам. Модель NN with Keras [137] показала ще нижчі результати, з точністю 86.74%, F1 – 85.79%, прецизійністю – 86.37% та повнотою – 85.22%, що може вказувати на певні проблеми з узагальненням у цьому підході. Дві реалізації Non memory-based ensemble model [137] демонструють однакові результати, з точністю 86.39%, мірою F1 – 85.31%, прецизійністю – 88.37% та повнотою – 82.46%.

Серед класичних методів Naive Bayes [137] та GaussianNB [135]показали найнижчі результати, з точністю 75.89% та 74.00% відповідно. Їхня продуктивність значно нижча, ніж у нейромережевих та ансамблевих методів, що вказує на обмеженість цих підходів для задач текстової класифікації.

Запропоновані моделі продемонстрували високу ефективність у порівнянні з широким спектром моделей, представлених у літературі, як у

традиційному машинному навчанні, так і в сучасних глибоких нейронних мережах. Найкращими варіантами є Att-BiLSTM та LightGBM, які мають найвищу точність і баланс між прецизійністю та повнотою. BiLSTM та XGBoost також є конкурентоспроможними моделями, особливо у випадках, коли важлива висока повнота.

Набір даних було поділено таким чином: 80% даних використовувалися для навчання моделей, а 20% – для їх тестування. Оцінка продуктивності на валідаційному наборі відіграла важливу роль у налаштуванні гіперпараметрів моделей, тоді як тестовий набір забезпечив об'єктивну оцінку їхньої реальної продуктивності.

Дослідження є значним кроком вперед як у науковому, так і у практичному вимірі виявлення фейкових новин. Наукова новизна роботи полягає у застосуванні передових методів глибокого навчання, зокрема архітектур BiLSTM та BiLSTM з механізмом уваги, які адаптовані для складного завдання розрізнення достовірних та сфабрикованих новин. Такий підхід означає відхід від традиційних методів і пропонує глибший аналіз мовних патернів фейкових новин. Інтеграція механізму уваги у BiLSTM є особливо інноваційною, оскільки вона дозволяє моделі вибірково зосереджуватися на найбільш інформативних частинах даних, що значно підвищує точність та надійність класифікації.

Для розуміння деталей процесу ухвалення рішень моделлю у нашому аналізі було досліджено графіки важливості токенів для моделей XGBoost та LightGBM, що дозволило визначити конкретні слова або фрази, які відіграють ключову роль у процесі класифікації. Ці графіки не лише візуально відображають значущість кожного токена, а й допомагають глибше зрозуміти патерни та характеристики, які моделі вважають важливими для розрізнення правдивих і фейкових новин. Таким чином, вони слугують засобом обґрунтування рішень моделі, що дозволяє дослідникам, зацікавленим сторонам та кінцевим користувачам підвищити довіру до її результатів та розібратися у мовних маркерах, які впливають на класифікацію. Крім того, аналіз цих ключових мовних ознак дає змогу вдосконалити моделі, забезпечуючи їхню здатність орієнтуватися

на найбільш релевантні особливості даних, що підвищує їхню стійкість та точність у реальних умовах.

На рисунку 4.2 представлено важливість токенів для моделі XGBoost, а на рисунку 4.3 – для моделі LightGBM. Виявлення значущості окремих токенів або ознак у процесі класифікації фейкових новин є ключовим для розуміння патернів, які моделі вважають індикаторами дезінформації. Аналіз основних важливих ознак у моделях XGBoost і LightGBM дозволяє виявити лінгвістичні шаблони та ключові слова, які відіграють центральну роль у розрізненні справжніх і фейкових новин.

Для моделі XGBoost найвпливовішими ознаками виявилися слова “breitbart”, “hillary” та “follow”. Значна вага термінів “breitbart” та “hillary” може свідчити про те, що певні джерела новин або політичні фігури частіше асоціюються з певними типами новин – правдивими або фейковими. Інші слова, такі як “2016”, “reuters” та “york”, можуть вказувати на часовий або географічний контекст новинних статей, тоді як терміни “image”, “twitter” та “video” можуть свідчити про мультимедійний характер або платформи розповсюдження новинного контенту.

З іншого боку, у моделі LightGBM ключовою ознакою виявилося слово “said”, за яким слідують “trump” і “new”. Висока значущість слова “said” може свідчити про специфічні лінгвістичні конструкції або стиль викладу, який модель пов’язує з правдивими або фейковими новинами. Також слова “president”, “times” і “people” можуть вказувати на змістовну тематику статей. Повторюваність таких термінів, як “trump”, “hillary”, “reuters” і “york” у обох моделях підтверджує їхню потенційну важливість для класифікації новин. Загалом, ці оцінки важливості ознак дозволяють краще зрозуміти мовні та тематичні елементи, які моделі XGBoost та LightGBM вважають ключовими у виявленні фейкових новин. Вони ілюструють складну взаємодію між джерелами, політичними діячами, мультимедійними елементами та поширеними мовними конструкціями у процесі класифікації дезінформації.

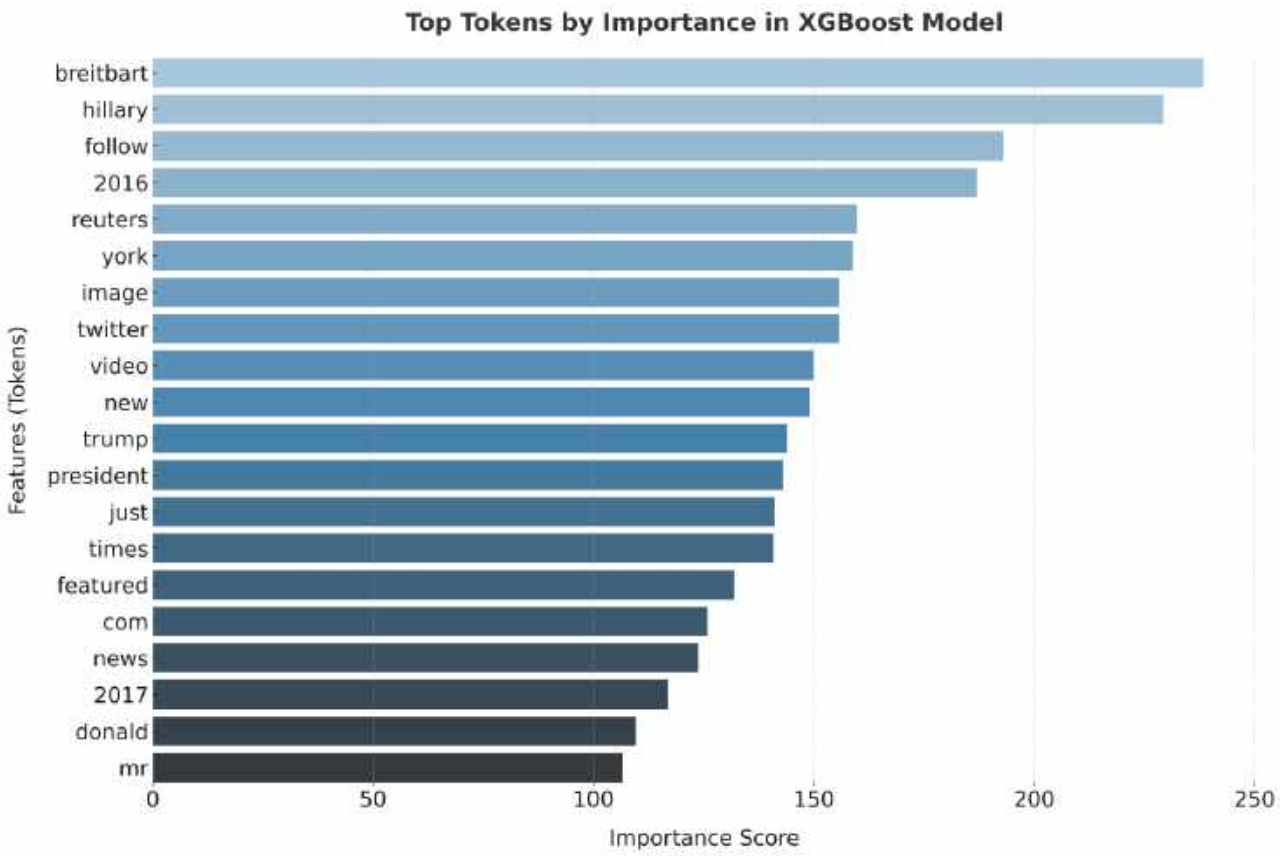

Рисунок 4.2 – Важливість токенів у XGBoost моделі

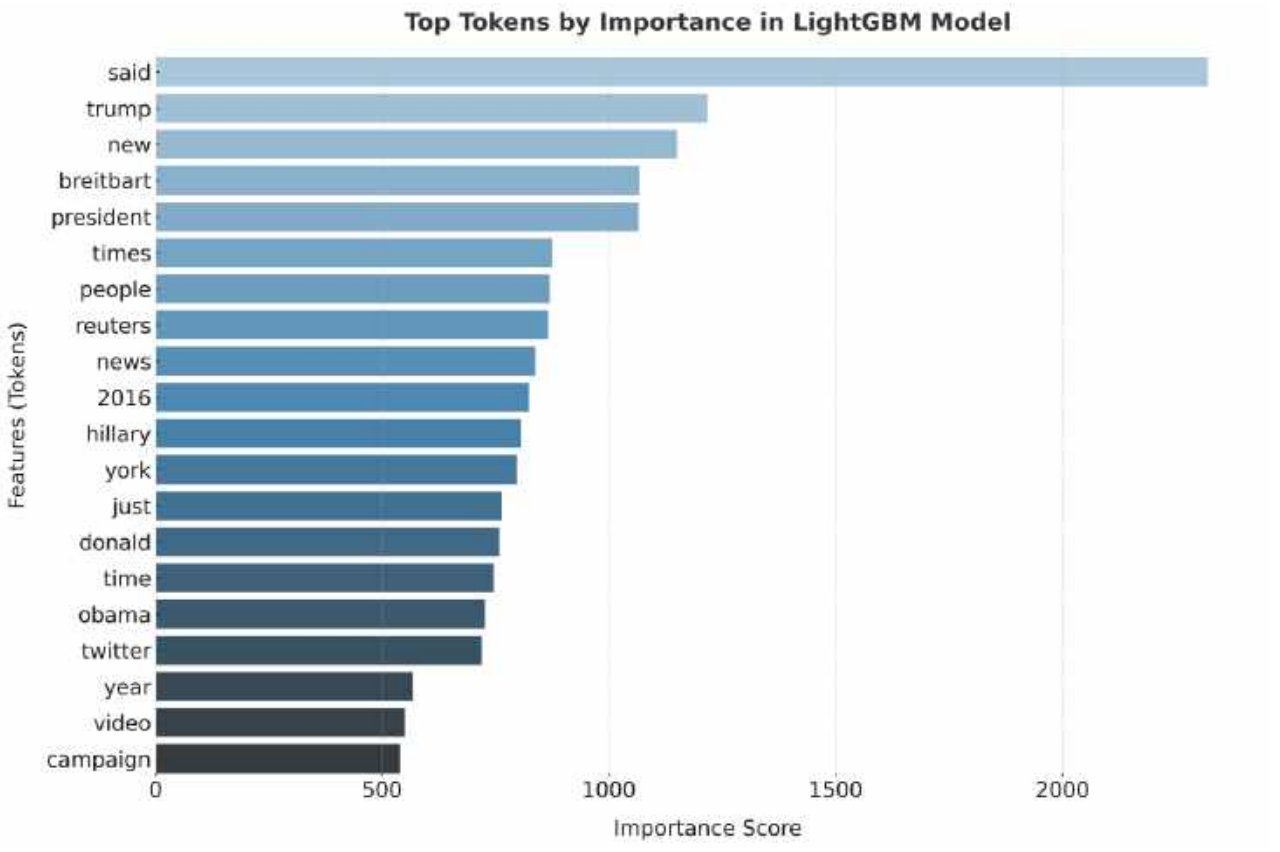

Рисунок 4.3 – Важливість токенів у LightGBM моделі

4.3. Аналіз продуктивності моделей виявлення місінформації, пов’язаної з охороною здоров’я

У 21 столітті дезінформація в галузі охорони здоров’я стала серйозним викликом для громадського здоров’я, ускладненим безпрецедентним охопленням та впливом цифрових комунікаційних платформ. Це питання має глибокі наслідки, не лише для особистих рішень щодо здоров’я, а й для стратегій громадського здоров’я, що вимагають термінових, багатогранних заходів для боротьби з її наслідками.

Наслідки дезінформації в галузі охорони здоров’я не є абстрактними або обмеженими ізольованими інцидентами; вони відгукуються по всьому світу, впливаючи на суспільства та економіки. Пандемія COVID-19 різко продемонструвала всю поширеність і руйнівну природу дезінформації в галузі охорони здоров’я [138]. В перші місяці пандемії поширювалися неправдиві твердження, такі як пропозиції, що пиття відбілювача може вбити вірус або що технологія 5G є причиною поширення хвороби [139]. Ці твердження призвели до трагічних наслідків, включаючи втрати серед населення та широке непорозуміння серед громадськості [140]. Подібно до цього, дезінформація щодо вакцин сприяла сумнівам і відмовам, безпосередньо підірвавши глобальні кампанії вакцинації проти таких хвороб, як кір, поліомієліт і COVID-19 [141].

Окрім безпосереднього впливу на здоров’я, дезінформація сприяє поляризації суспільства та ускладнює зусилля щодо охоплення вразливих груп населення [142]. Наприклад, під час спалахів Еболи в Західній Африці дезінформація щодо походження та поширення хвороби призвела до опору медичним працівникам та нападів на лікувальні установи [143]. Наслідки дезінформації, таким чином, виходять за межі індивідуального здоров’я і впливають на ширші структурні системи, що покликані захищати та підтримувати громадське благополуччя.

Модель, запропонована в розділі 3.4. продемонструвала високу продуктивність для виявлення та аналізу місінформації, пов’язаної з охороною здоров’я. Модель класифікації, застосована до набору даних CoVID19-FNIR [133], досягла високих показників продуктивності, з точністю 99.47%, а також високими значеннями прецизійності, повноти та міри F1. Низька втрата при тестуванні (0.01698) підкреслює здатність моделі до точного прогнозування та мінімальних помилок у задачах класифікації.

Три інші дослідження перевіряли свої моделі класифікації на тому ж наборі даних. Р. Вінай та колеги [144] застосували підхід, заснований на інженерії ознак, використовуючи різні класифікатори машинного навчання. Їхня найкраща модель досягла точності 99,20% і міри F1 0,9920, що підкреслює важливість лінгвістичних, соціальних та статистичних характеристик. Хоча їхній підхід демонструє високу ефективність, запропонована в дисертаційному дослідженні модель перевершує його завдяки використанню глибоких контекстуальних ембедингів і механізмів уваги, що підвищує точність і інтерпретованість моделі. Крім того, архітектура запропонованої моделі забезпечує кращу адаптацію до змінних наративів дезінформації.

У статті М. Сікосана та колег [145] було оцінено різні моделі машинного та глибокого навчання, включаючи гібридні архітектури CNN-LSTM, які досягли точності понад 99% на наборі FNIR. Хоча ці моделі демонструють високу продуктивність, вони значною мірою залежать від складних нейромережевих архітектур, що можуть мати проблеми зі масштабованістю та реальним використанням. Запропонована модель, навпаки, підтримує високу точність при збереженні обчислювальної ефективності, що робить її придатною для великомасштабного розгортання. Хоча точність обох моделей однакова (99.47%), запропонована в дисертаційному дослідженні модель демонструє кращу продуктивність за показником міри F1, яка становить 0.9947 порівняно з 0.9945 у моделі з [145]. Ця, здавалося б, незначна різниця є важливою в академічних оцінках, оскільки міра F1 відображає гармонійне

середнє між прецизійністю та повнотою, забезпечуючи комплексну оцінку продуктивності моделі при роботі з незбалансованими даними. Вища міра F1 запропонованої моделі вказує на кращий баланс між прецизійністю та повнотою, що є особливо критичним для виявлення дезінформації, де мінімізація хибнопозитивних і хибнонегативних випадків має вирішальне значення.

М. Кадіс та А. Ханнан [146] дослідили традиційні моделі машинного навчання, отримавши найкращу точність 91.6% та міру F1 0.92 для моделі RF. Хоча ці моделі є обчислювально ефективними, їхня обмежена здатність враховувати контекстну та послідовну інформацію у тексті обмежує їхню узагальнюваність. Інтеграція механізмів уваги у запропонованій моделі дозволяє подолати ці обмеження, забезпечуючи більш надійне рішення для виявлення дезінформації.

Таблиця 4.2 порівнює продуктивність моделей, представлених у дослідженнях [144], [145], [146], з запропонованою моделлю. Стаття [145] містить різні методи попередньої обробки даних для всіх проаналізованих традиційних класифікаторів машинного навчання. Тому було відібрано найкращу продуктивність для кожної моделі, щоб включити її у порівняльну таблицю.

Таблиця 4.2 – Порівняльний аналіз продуктивності запропонованих ансамблевих моделей з результатами інших досліджень

Модель	Точність	Міра F1	Прецизійність	Повнота
Запропонована модель (розділ 3.4)	99.47	0.9947	0.9947	0.9947
NB [144]	92.49	0.9242	0.9341	0.9233
DT [144]	98.68	0.9868	0.9867	0.9869
RF [144]	99.14	0.9914	0.9913	0.9915
SVM [144]	98.61	0.9863	0.9861	0.9862

Продовження таблиці 4.2

SVM-linear kernel [144]	98.61	0.9861	0.9861	0.9862
SVM-RBF kernel [144]	97.10	0.9709	0.9710	0.9709
SVM-polynomial kernel [144]	86.82	0.8677	0.8783	0.8701
LR [144]	99.20	0.9920	0.9920	0.9921
NB [145]	91.77	0.9194	-	-
Gradient Boosting [145]	91.40	0.9213	-	-
SVM [145]	94.14	0.9441	-	-
DT [145]	89.10	0.8942	-	-
RF [145]	93.58	0.9362	-	-
Bagging [145]	91.14	0.9132	-	-
AdaBoost [145]	91.91	0.9212	-	-
SGD [145]	93.86	0.9408	-	-
LR [145]	93.09	0.9341	-	-
Multilayer Perceptron [145]	93.16	0.9317	-	-
CNN+DistilBERT [145]	52.50	0.5070	-	-
CNN+Word2Vec [145]	99.34	0.9831	-	-
LSTM+Word2Vec [145]	99.47	0.9945	-	-
LSTM+DistilBERT [145]	48.22	0.6507	-	-
NB [146]	89.2	0.89	-	-
Gradient Boosting [146]	85.1	0.86	-	-
SVM [146]	85.1	0.86	-	-
DT [146]	85.8	0.86	-	-
RF [146]	91.6	0.92	-	-
Bagging [146]	87.2	0.87	-	-
AdaBoost [146]	84.8	0.85	-	-
SGD [146]	91.5	0.92	-	-
LR [146]	90.6	0.91	-	-
Multilayer Perceptron [146]	90.4	0.90	-	-

Порівняльний аналіз підкреслює ефективність інтеграції передових методів обробки природної мови з традиційними методами класифікації. Хоча моделі, засновані на інженерії ознак, відзначаються простотою та обчислювальною ефективністю, вони не можуть враховувати глибші семантичні зв’язки. Гібридні моделі глибокого навчання забезпечують високу продуктивність, але можуть викликати проблеми з інтерпретацією та використанням ресурсів. Запропонований підхід знаходить баланс, поєднуючи високу точність з обчислювальною ефективністю та адаптивністю.

4.4. Експериментальне дослідження моделі тематичного моделювання дезінформації українською мовою

Для аналізу наявних наративів у наборах даних дезінформації українською мовою під час повномасштабного вторгнення росії в Україну [147]. було застосовано тематичне моделювання як ключовий підхід для кластеризації даних. Опис набору даних знаходиться у Додатку Г. Запропонований метод дозволив групувати повідомлення на основі спільних тем і наративів, надаючи уявлення про структуру та зміст дезінформації, що поширюється через Telegram.

Для аналізу набору даних зі статтями фейкових новин спочатку було відфільтровано дані, залишивши лише записи з фейковими новинами. Далі текстові дані було очищено, видаливши розділові знаки, числа та стоп-слова. Для перевірки ефективності цього підходу було використано великий набір із 1 983 українських стоп-слів, включаючи числа з набору даних [148].

Також було видалено специфічні фрази, пов’язані з українськими Telegram-каналами та поширенням інформації, такі як: “українські телеграмканали”, “телеграмканали повідомляють”, “повідомляють українські телеграмканали”, “telegram”, “українських телеграмканалах”, “часто згадується”, “інформація розповсюджувалась”, “повідомлення

поширювались”, “новинних телеграмканалів”, “така інформація”, “про це пишуть”, “українські джерела”, “посиланням на твітер”, “повідомлення поширюються”, “це пишуть місцеві”.

Ці фрази не мають особливого значення для видобування тем і можуть більше заплутати модель, ніж забезпечити чіткий результат тематичного моделювання. Було підготовлено дві версії обробленого тексту: одна з видаленими стоп-словами, а інша – зі збереженими, що дозволило оцінити вплив стоп-слів на результати тематичного моделювання.

Для аналізу було адаптовано модель “sentence-transformers/facebook-dpr-ctx_encoder-single-nq-base” для генерації щільних векторних представлень текстових даних [149]. Ця модель, контекстний кодувальник, призначений для Dense Passage Retrieval, кодує речення та абзаци у 768-вимірний векторний простір. Попередньо навчена на наборі даних Natural Questions, вона ефективно захоплює багатий контекстуальний зміст оброблюваних фрагментів тексту. Використовуючи архітектуру на основі трансформерів, модель добре розпізнає взаємозв’язки між словами та їхнім ширшим текстовим контекстом, що робить її ідеальною для завдань глибокого семантичного пошуку.

Щоб зафіксувати контекстуальне значення слів у наборі даних, було використано ембединги на основі трансформерів як вхідні дані для моделі BERTopic. Зокрема, було застосовано попередньо навчену модель BERT для створення векторних представлень українського тексту [21]. Вбудовування отримуються шляхом передавання вхідного тексту через архітектуру BERT, що генерує щільний вектор для кожного слова, де розмірність вектора відображає його семантичний контекст.

Модель BERT використовує механізми самоуваги, які описуються наступним рівнянням:

$$\text{Attention}(\text{Q}, \text{K}, \text{V}) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4.1}$$

де Q представляє матрицю запитів, K – матрицю ключів, V – матрицю значень, а d_k – це розмірність векторів ключів.

Ембедінги BERT створюються за допомогою цього механізму, де представлення кожного слова коригується на основі його оточуючих слів, що дозволяє зберігати семантичні зв'язки. Ці щільні ембедінги слугують основою для подальшого процесу тематичного моделювання.

Основний алгоритм тематичного моделювання, який було використано – це BERTopic, який ґрунтується на ембедінгах, отриманих із трансформерів, для кластеризації схожих текстових фрагментів у теми. BERTopic включає два основні етапи: зменшення розмірності ембедінгів, формування тем шляхом кластеризації.

Спочатку було зменшено розмірність багатовимірних ембедінгів BERT за допомогою методу Uniform Manifold Approximation and Projection (UMAP). UMAP проєктує багатовимірні ембедінги у простір меншої розмірності, зберігаючи при цьому топологічну структуру даних. Алгоритм UMAP мінімізує таку функцію втрат:

$$L_{UMAP} = \sum_{(i,j)\in S} \log\left(\frac{p_{ij}}{q_{ij}}\right), \tag{4.2}$$

де p_{ij} представляє схожість між точками i та j у багатовимірному просторі, q_{ij} – схожість у просторі зниженої розмірності, а S – множина сусідніх точок у вихідному багатовимірному просторі.

Результатом є простір ембедінгів меншої розмірності, де схожі текстові фрагменти розташовані близько один до одного, що полегшує процес кластеризації.

Після зменшення розмірності ембедінгів було застосовано алгоритм Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) для групування схожих текстових фрагментів. HDBSCAN – це алгоритм кластеризації, заснований на щільності, який виявляє кластери різної

щільності, аналізуючи густоту точок у просторі зниженої розмірності. Він визначає кластери, де точки щільно згруповані разом, водночас трактуючи розріджені області як шум.

Основною перевагою HDBSCAN є його здатність ідентифікувати кластери зі змінною щільністю без необхідності заздалегідь задавати кількість кластерів, що робить його особливо корисним для виявлення динамічних і складних тем у кампаніях дезінформації.

Алгоритм кластеризації HDBSCAN базується на обчисленні взаємної досяжності відстані, яка визначається наступним чином:

$$d_{mreach}(\mathrm{x},\mathrm{y}) = \max\left(core\_distance(x), core\_distance(y), d(x,y)\right), \quad (4.3)$$

де d(x, y) – евклідова відстань між точками x і y, core_distance(x) – мінімальна відстань, необхідна для утворення кластера навколо x, d_{mreach} – модифікована відстань, яка використовується для визначення, чи належать точки до одного кластера.

HDBSCAN формує кластери, що відповідають окремим темам, на основі цих взаємних досяжностей.

Для інтерпретації та представлення тем, отриманих у процесі кластеризації, було використано модель представлення, на основі адаптації KeyBERT . KeyBERT витягує найбільш репрезентативні слова або фрази для кожної теми, використовуючи косинусну подібність між ембедингами теми та ембедингами окремих слів.

Косинусна подібність обчислюється за формулою:

$$\mathrm{cosine\_similarity}(\mathrm{A},\mathrm{B}) = \frac{A \cdot B}{\|A\|\|B\|} \quad (4.4)$$

де A і B – це вбудовування двох слів або фраз. KeyBERT ранжує терміни на основі їхньої схожості з ембедингами тем, гарантуючи, що найбільш релевантні слова або фрази вибираються для представлення тем.

Щоб покращити репрезентацію тем, було додатково використано CountVectorizer для виявлення n-грам у діапазоні від уніграм (окремих слів) до триграм (фраз із трьох слів). CountVectorizer створює частотну матрицю термінів у тексті, відстежуючи, як часто n-грами зустрічаються в наборі даних. Для кожного тексту t_i CountVectorizer обчислює частоту кожної n-грами n_j за формулою:

$$\text{freq}(n_j) = \frac{\sum_{t_i} 1(n_j \in t_i)}{N} \tag{4.5}$$

де $1(n_j \in t_i)$ – це індикаторна функція, яка повертає 1, якщо n_j зустрічається у t_i, а N – загальна кількість документів у наборі даних.

Включення аналізу n-грам дозволило виявити як окремі ключові слова, так і багатослівні фрази, що сприяють поширенню дезінформаційних наративів, тим самим покращуючи здатність моделі розпізнавати складні закономірності у даних. Для виявлення основних тем і можливих підтематик були використані як плоскі, так і ієрархічні методи кластеризації.

Після виконання моделювання на основі BERTopic та ідентифікації кластерів, було вручну переглянуто основні ключові слова та фрази кожної теми. Цей якісний аналіз допоміг уточнити розуміння виявлених тем, гарантуючи, що вони є релевантними та зрозумілими у контексті україномовної дезінформації.

Виявлені теми були категоризовані на основі їхніх домінантних характеристик, таких як: емоційна маніпуляція, пропагандистські техніки, політичні наративи.

Для ілюстрації взаємозв’язків між визначеними темами були згенеровані візуалізації, такі як карта міжтематичних відстаней та ієрархічні діаграми кластеризації.

Візуалізація карти міжтематичних відстаней допомагає інтерпретувати результати тематичного моделювання, представляючи зв’язки між темами у

просторі низької розмірності. Вона показує: близькість тем – відстань між темами на графіку, що вказує на їхню подібність. Теми, розташовані ближче одна до одної, мають більше спільних слів і є тематично пов’язаними, тоді як віддалені теми містять менше спільних слів і репрезентують відмінні наративи. Розмір теми – розмір кожного кола на графіку пропорційний частоті теми у наборі даних, тобто наскільки велика ця тема у загальному корпусі текстів. Перекриття тем – ступінь перекриття кіл, що вказує на спільні ключові слова між темами. Чітко розділені кола означають чітко виражені теми, тоді як перекриття може сигналізувати про тематичну близькість.

Візуалізація ієрархічної кластеризації дозволяє відобразити взаємозв’язки між різними даними у ієрархічній структурі. Найпоширеніша візуалізація ієрархічної кластеризації – дендрограма, яка графічно представляє процес злиття або розподілу кластерів. Різноманітність кольорів показує, що дані розподілені між кількома групами, утворюючи менші й чіткіші підкластеризації. Послідовне злиття гілок означає, що кластери поступово об’єднуються, показуючи їхню структурну організацію.

Аналіз двох наборів даних заголовків новин та основного тексту новин, та розподіл їх на чотири піднабори (з урахуванням та без урахування стоп-слів) показав наступні результати. Їх було порівняно у парах початкових двох наборів даних. Розподіл тем представлено за допомогою двох методів візуалізації: карти міжтематичних відстаней та ієрархічної кластеризації. Набір даних “Основний текст новин” без стоп-слів спочатку містив менше тем (34), ніж набір із включеними стоп-словами (43), що свідчить про кращий розподіл тем. На рисунку 4.4 наведено карту міжтематичних відстаней для набору “Основий текст новин” без стоп-слів.

На карті з рисунку 4.4 теми розподілені ширше, утворюючи кілька кластерів, а самі теми розкидані по всій площині. Деякі теми розташовані близько одна до одної, але багато кіл знаходяться на значній відстані, що вказує на чіткий поділ тем із мінімальним перекриттям у розподілі слів. Різний розмір кластерів

може вказувати на поєднання великих і малих тем. На рисунку 4.5 наведено карту міжтематичних відстаней для набору “Основний текст новин” зі стоп-словами.

На рисунку 4.5 видно, що кілька тем згруповані дуже щільно. Висока концентрація тем у верхньому правому квадранті вказує на те, що модель виявила схожі теми, які мають спільні слова. Однак, відсутність чіткого розподілу тем у цьому кластері ускладнює інтерпретацію графіка, оскільки теми виглядають хаотично згрупованими. На рисунку 4.6 наведено ієрархічну кластеризацію для набору “Основний текст новин” без стоп-слів

На рисунку 4.6 гілки кластерів рівномірно розподілені, чітко розмежовані різні групи. Велика кількість дрібних кластерів означає, що модель розділила дані на більш точні та дрібні сегменти. Кластери зливаються поступово, що свідчить про чітке розмежування між групами протягом більшої частини процесу кластеризації. На рисунку 4.7 наведено ієрархічну кластеризацію для набору “Основний текст новин” зі стоп-словами.

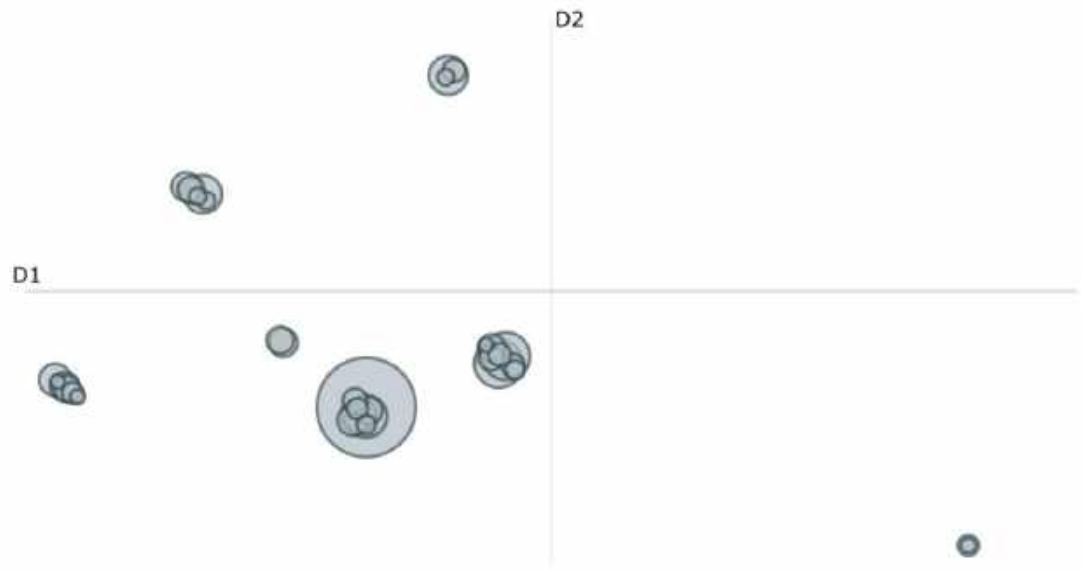

Рисунок 4.4 – Карта міжтематичних відстаней для набору “Основний текст новин” без стоп-слів

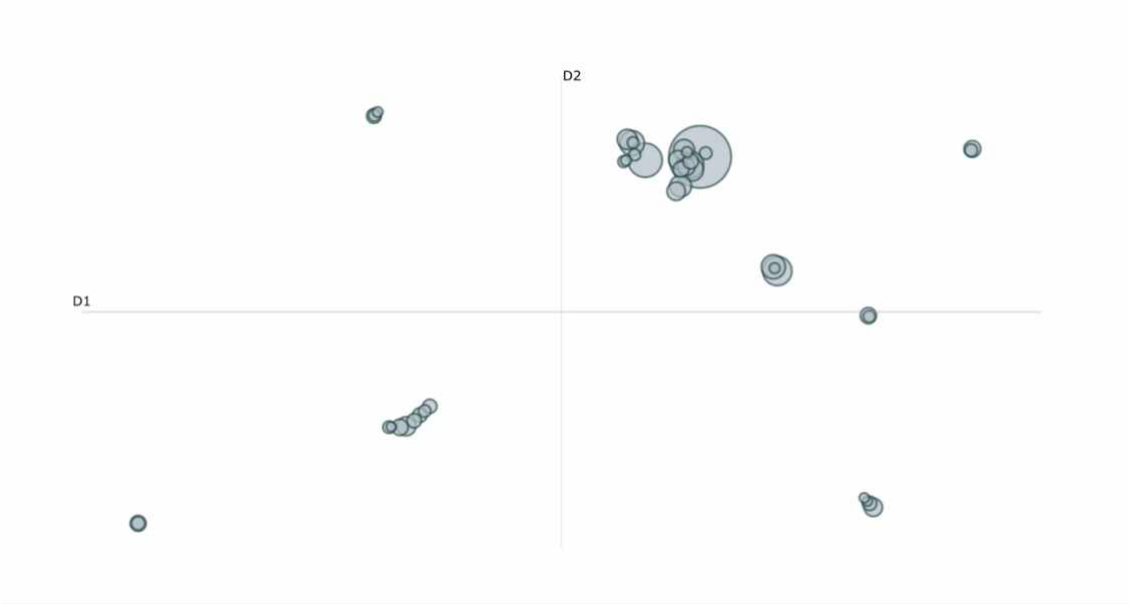

Рисунок 4.5 – Карта міжтематичних відстаней для набору “Основний текст новин” зі стоп-словами

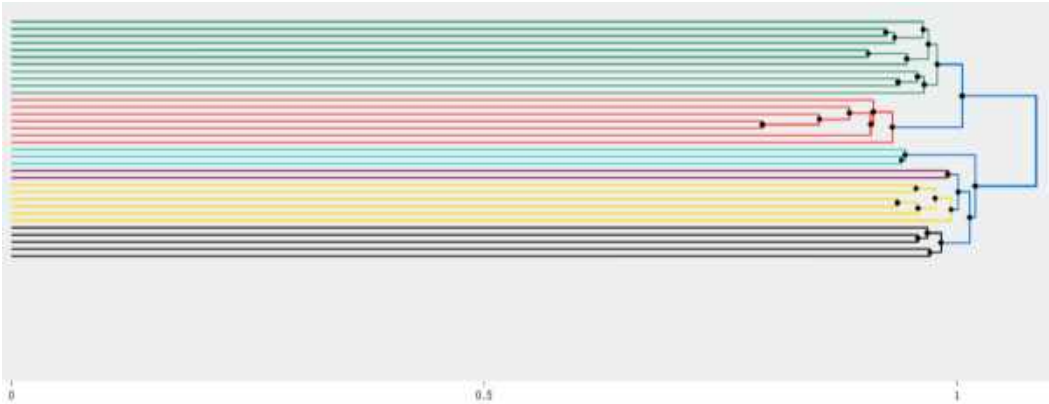

Рисунок 4.6 – Ієрархічна кластеризація для набору “Основний текст новин” без стоп-слів

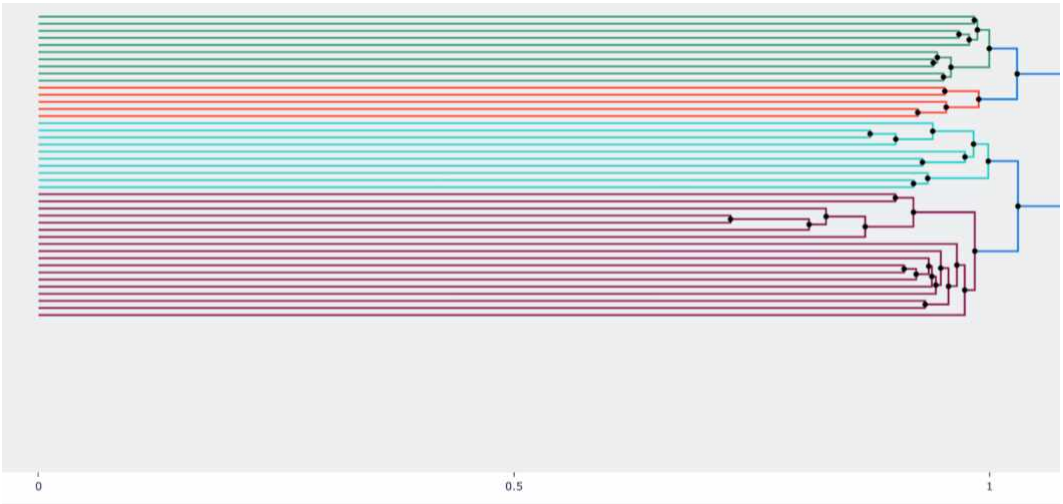

Рисунок 4.7 – Ієрархічна кластеризація для набору “Основний текст новин” зі стоп-словами

На рисунку 4.6 гілки кластерів рівномірно розподілені, чітко розмежовані різні групи. Велика кількість дрібних кластерів означає, що модель розділила дані на більш точні та дрібні сегменти. Кластери зливаються поступово, що свідчить про чітке розмежування між групами протягом більшої частини процесу кластеризації. На рисунку 4.7 наведено ієрархічну кластеризацію для набору “Основний текст новин” зі стоп-словами.

У нижній частині графіка з рисунку 4.5 (велика фіолетова область) структура кластерів більш виражена, тоді як у верхній частині (зелені та червоні сегменти) зберігається відносно рівномірний розподіл. Ранні злиття в деяких областях (особливо у фіолетовій зоні) означають, що ці точки даних мають вищу схожість, що призводить до швидшого формування більших кластерів.

Набір даних “Заголовки новин” без стоп-слів містив 49 тем, а набір із стоп-словами – 54 теми. На рисунку 4.6 наведено карту міжтематичних відстаней для набору “Заголовки новин” без стоп-слів. На рисунку 4.6 теми широко розкидані по всій карті, утворюючи чітко відокремлені кластери. Великі відстані між кластерами свідчать про гарне розмежування тем, що полегшує інтерпретацію. Низький рівень перекриття означає, що теми мають незалежні змісти, без значного накладання. На рисунку 4.7 наведено карту міжтематичних відстаней для набору “Заголовки новин” зі стоп-словами.

На рисунку 4.7 теми згруповані значно щільніше у порівнянні з першою картою. Багато тем розташовані близько одна до одної, що створює щільніші кластери. Через це інтерпретувати карту складніше, оскільки близькість тем означає, що деякі наративи можуть перекриватися. Також можна помітити великий домінуючий кластер, навколо якого розташовані менші. На рисунку 4.8 наведено ієрархічну кластеризацію для “Заголовків новин” без стоп-слів. На рисунку 4.8 кластери чітко розмежовані, з помітною різницею між окремими групами. Кожен колір представляє окрему групу даних, а гілки зливаються поступово, що означає ретельний поділ перед остаточним об’єднанням у більші кластери.

На рисунку 4.9 наведено ієрархічну кластеризацію для “Заголовків новин” зі стоп-словами. Графік на рисунку 4.9 демонструє менш чітку кластеризацію – гілки зливаються раніше, формуючи ширші та менш деталізовані кластери. Це означає, що кластеризація базується на загальних закономірностях, а не на детальному поділі на теми. Порівняно з попереднім графіком, чіткість меж між кластерами зменшена, що ускладнює їхню інтерпретацію.

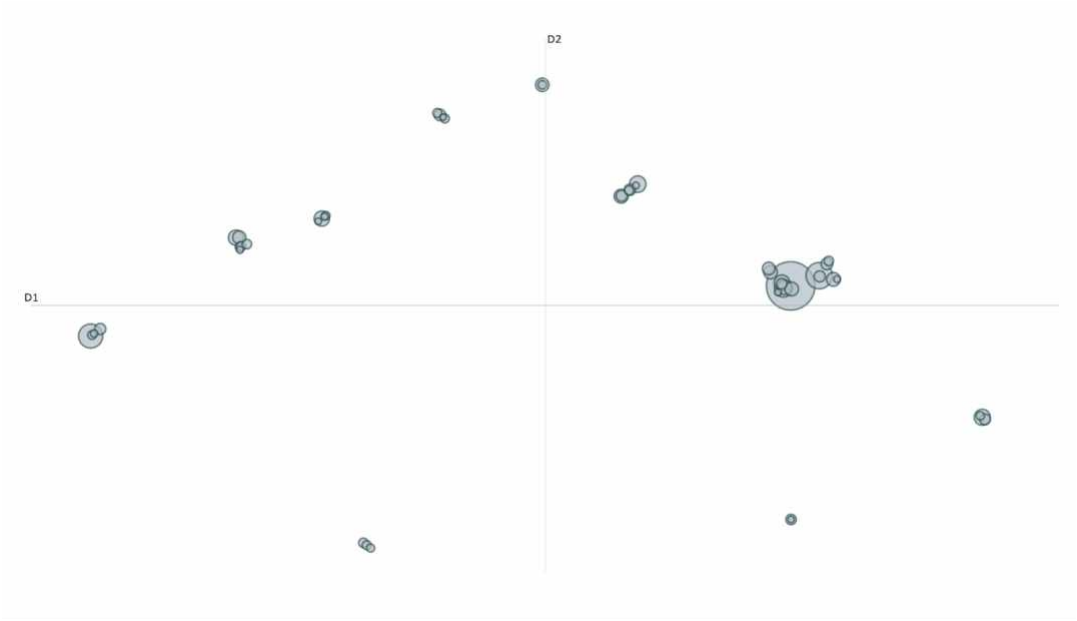

Рисунок 4.8 – Карта міжтематичних відстаней для набору “Заголовки новин” без стоп-слів

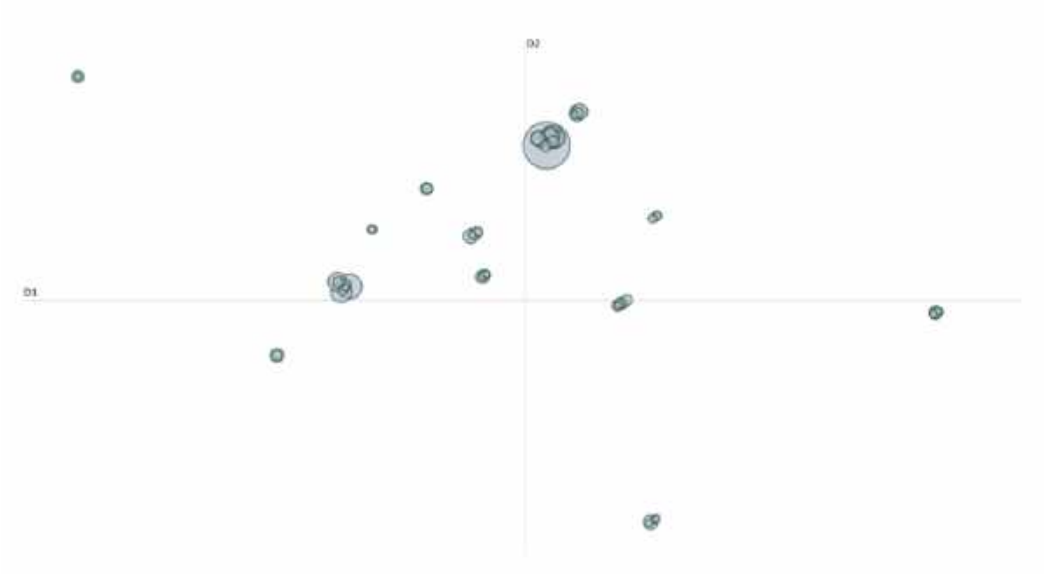

Рисунок 4.9 – Карта міжтематичних відстаней для набору “Заголовки новин” зі стоп-словами

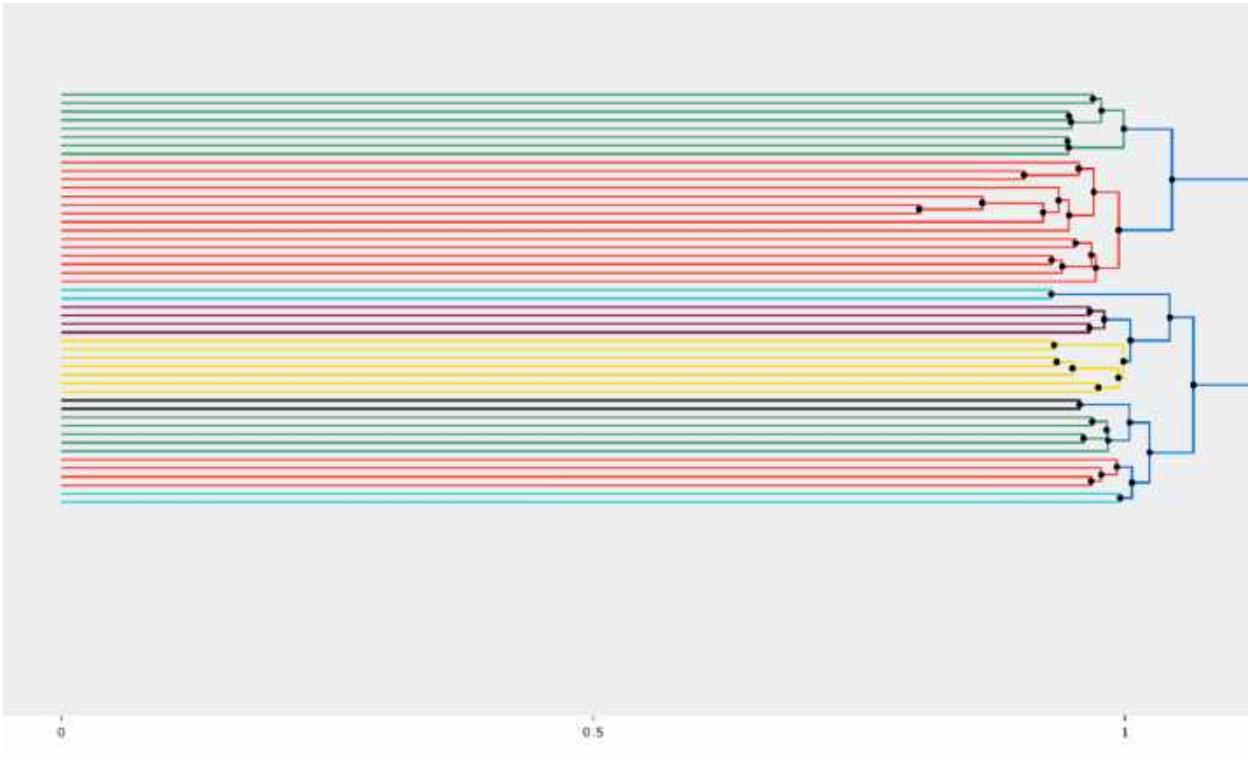

Рисунок 4.10 – Ієрархічна кластеризація для “Заголовків новин” без стоп-слів

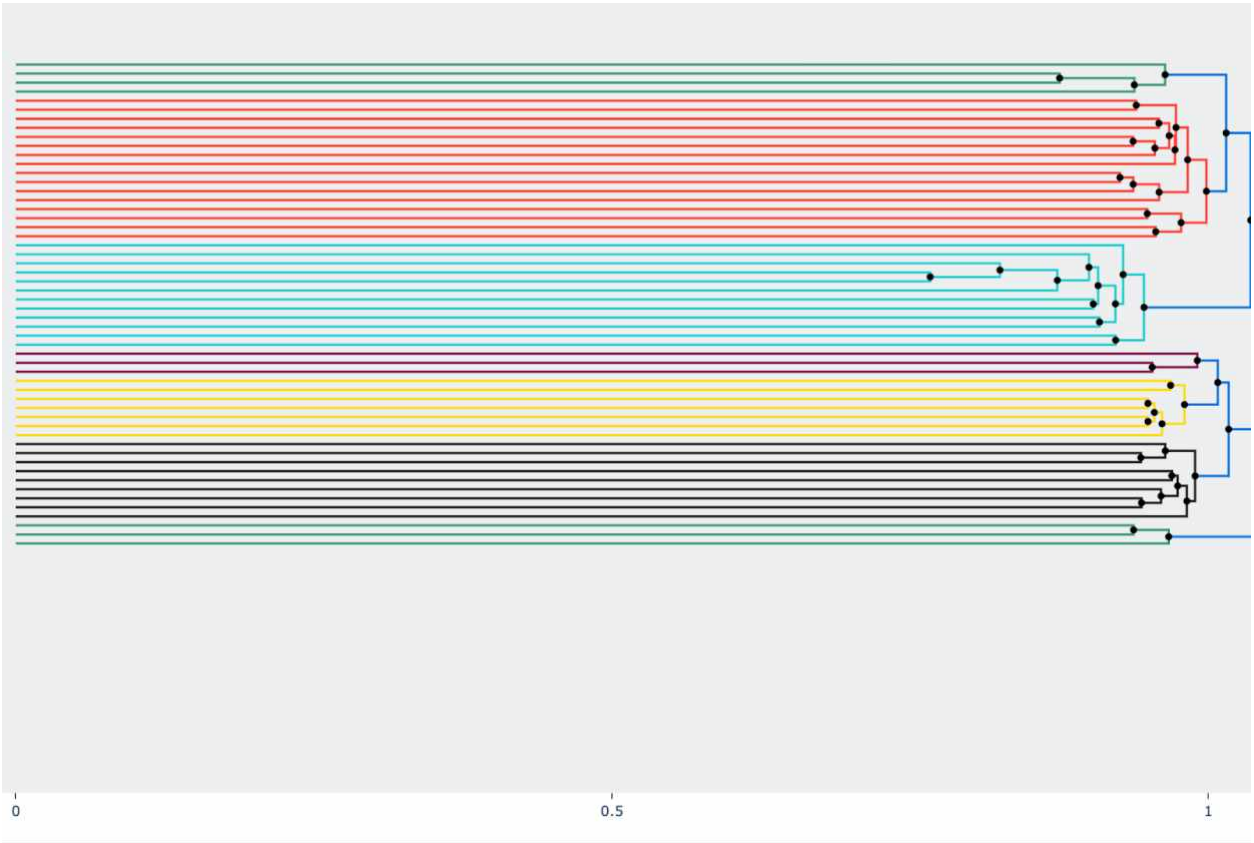

Рисунок 4.11 – Ієрархічна кластеризація для “Заголовків новин” зі стоп-словами

Аналіз міжтематичних відстаней для двох наборів даних виявляє різні особливості у результатах тематичного моделювання. Для основних текстів новин тематичне моделювання зазвичай дає меншу кількість тем, а отримані кластери є більш виразними, що свідчить про чіткіше розмежування між темами. Це пов'язано з багатшим контекстом у текстах новин, порівняно із заголовками, що дозволяє моделі точніше розрізняти теми з меншою кількістю кластерів. Крім того, включення або виключення стоп-слів має більш виражений вплив на кількість сформованих тем, оскільки стоп-слова у повних текстах новин більшою мірою впливають на когерентність тем.

З іншого боку, набори даних, що складаються з заголовків новин, загалом демонструють більшу кількість тем через їхню коротку форму та різноманітність змісту, що призводить до більшої фрагментації кластерів. Незважаючи на більшу кількість тем, різниця у їхній кількості при включенні та виключенні стоп-слів є відносно невеликою – приблизно п'ять тем. Це свідчить про те, що стоп-слова у коротких текстових сегментах, таких як заголовки, менше впливають на загальну структуру тем, ніж у повних текстах новин, де видалення стоп-слів суттєвіше впливає на уточнення кластерів.

Кількісний та якісний аналіз наборів даних показує, що видалення стоп-слів сприяє чіткішому розподілу, який легше проінтерпретувати досліднику на наступних етапах аналізу, порівняно із застосуванням тематичного моделювання до початкового набору даних, що містить стоп-слова. Вплив стоп-слів на тематичне моделювання є суттєвим, оскільки вони створюють шум і знижують чіткість тематичної структури даних.

Карта міжтематичних відстаней для набору даних із видаленими стоп-словами демонструє добре відокремлені та чітко виражені теми. Кожна тема виділяється з мінімальним перекриттям, що дозволяє більш точно ідентифікувати різні наративи. Це особливо важливо при аналізі різноманітних і унікальних тем у корпусі даних.

Відстань між кластерами на Рисунку 4.4 ілюструє, як видалення нерелевантних і часто вживаних слів допомагає виокремити основну лексику,

пов’язану з кожною темою. Це призводить до кращого визначення кластерів, оскільки алгоритм зосереджується на ключових термінах, які справді відображають відмінності між темами, а не на поширених словах на кшталт “або”, “і” тощо.

На відміну від цього, карта міжтематичних відстаней, що включає стоп-слова, демонструє значно менш інтерпретовану та хаотичну структуру. Теми здаються менш чіткими, з великим накладанням кластерів. Це свідчить про те, що стоп-слова ускладнюють здатність моделі ефективно розділяти теми. Повторення тем та хаотичний розподіл на Рисунку 4.5 вказують на те, що стоп-слова приховують ключові наративи, що ускладнює інтерпретацію кластерів. Включення загальновживаних, нетематичних слів робить виявлення значущих закономірностей складнішим, що негативно впливає на якість тематичного моделювання.

Рисунок 4.6, що демонструє ієрархічну кластеризацію набору даних із видаленими стоп-словами, підтверджує цей висновок. Кластери є більш чітко вираженими, що означає, що точки даних згруповані у добре визначені теми. Детальніше розмежування перед злиттям кластерів свідчить про те, що видалення стоп-слів покращує початковий розподіл тем і вдосконалює процес ієрархічної кластеризації.

Рисунок 4.7, який містить стоп-слова, свідчить про те, що точки даних згруповані у більші кластери. Ці більші кластери не мають тонкої деталізації, помітної на Рисунку 4.6. Наявність стоп-слів призводить до меншої кількості малих і чітко виражених груп. Натомість точки даних об’єднуються навколо поширених слів, які можуть не мати значення для ідентифікації тем.

Це призводить до менш інтерпретованих кластерів, де теми можуть здаватися штучно схожими через вплив частотних, нетематичних слів.

На рисунку 4.8 відстань між кластерами в графіку ієрархічної кластеризації демонструє, як видалення стоп-слів дозволяє моделі зосередитися на значущій лексиці. Чітке розділення між кластерами свідчить про те, що точки даних краще поділені на окремі групи. Це відображає основну

лексику, пов’язану з кожною темою, що полегшує їх розмежування. Відсутність часто вживаних, нерелевантних слів, таких як “і”, “або”, дає змогу алгоритму виявити справжні тематичні відмінності.

На відміну від цього, карта міжтематичних відстаней на рисунку 4.9, що включає стоп-слова, призводить до більш хаотичної та менш інтерпретованої структури. Теми виглядають менш чіткими, з помітним перекриттям кластерів. Це свідчить про те, що стоп-слова заважають моделі ефективно розділяти теми, що призводить до повторення тем та недостатнього їх розмежування. Включення цих поширених, нетематичних слів ускладнює виявлення значущих закономірностей у даних, що призводить до нечітких кластерів.

Подібно до цього, рисунок 4.10, який демонструє ієрархічну кластеризацію після видалення стоп-слів, підтверджує ці висновки. Кластери чітко розмежовані, а точки даних згруповані у виразні теми. Детальніше розмежування між кластерами до їхнього злиття показує, що видалення стоп-слів покращує початковий розподіл тем і підвищує загальну якість кластеризації.

На рисунку 4.11, де стоп-слова збережені, точки даних згруповані у більші, менш виразні кластери. Наявність стоп-слів призводить до меншої кількості малих груп, і кластеризація виглядає менш деталізованою. Це призводить до кластерів, які можуть виглядати штучно схожими через включення часто вживаних, нетематичних слів, що знижує інтерпретованість загальної структури та зменшує якість тематичного моделювання.

Без стоп-слів модель формує чіткіші, когерентніші кластери, які легше інтерпретувати. Теми стають більш прозорими, а процес кластеризації точніше відображає справжні тематичні відмінності у даних. У разі збереження стоп-слів модель стикається зі складнощами в розмежуванні тем, що призводить до плутанини та перекриття, коли теми, які повинні бути відокремленими, об’єднуються через вплив нерелевантних слів.

Додатково до кількісних результатів, якісний аналіз кластерів тем показує, що кластери, сформовані після видалення стоп-слів, є більш інтерпретованими та когерентними для людського аналізу. Теми в межах своїх кластерів є більш осмисленими, що дає змогу отримувати інсайти, які не затемнюються наявністю стоп-слів.

Цей аналіз підтверджує важливість застосування тематичного моделювання та ієрархічної кластеризації для аналізу дезінформації в україномовному наборі даних. Аналіз дав змогу зробити кілька висновків щодо ефективності попередньої обробки наборів даних, зокрема у контексті вилучення стоп слів з наборів заголовків та текстів новин.

- Тематичне моделювання показало кращі результати на наборах даних, що містять повні тексти новинних статей, а не лише заголовки. Багатший контекст повного тексту дозволив моделям генерувати більш виразні та узгоджені теми, що підвищило чіткість і точність визначення тем дезінформації. Заголовки самі по собі, будучи короткими та різноманітними, призводили до більш фрагментованих і менш інтерпретованих кластерів, що підкреслює важливість використання повного тексту для точного тематичного моделювання.
- Видалення стоп-слів значно покращило розподіл тем. Усунення частотних та несуттєвих слів дозволило моделі зосередитися на ключових термінах і фразах, які дійсно відображають основні теми в наборі даних. Це покращення стало очевидним як у картах міжтемної відстані, так і у візуалізаціях ієрархічного кластерного аналізу, де було досягнуто чіткішого розділення тем, що сприяло більш зрозумілим та узгодженим кластерам.
- Поєднання ієрархічного кластерного аналізу та тематичного моделювання виявило схожі закономірності в даних, підкреслюючи необхідність використання обох методів для комплексного аналізу. У той час як тематичне моделювання забезпечувало розуміння розподілу тем у даних, ієрархічне кластерування допомагало зрозуміти, як ці теми пов’язані на різних

рівнях деталізації. Такий подвійний підхід підтвердив узгодженість отриманих результатів і підвищив надійність аналізу.

4.5. Тематичне моделювання для аналізу наративів місінформації, пов'язаної зі здоров'ям

Початковий етап архітектури тематичного моделювання, заснованої на удосконаленій моделі all-MiniLM-L6-v2 [150] і представленої на рисунку 4.12, зосереджується на підготовці даних, забезпечуючи очищення та стандартизацію вхідного тексту для подальшої обробки. Сирі текстові дані проходять серію етапів попередньої обробки, включаючи перетворення в нижній регістр, видалення спеціальних символів, пунктуації та зайвих пробілів, якщо це не було виконано раніше на етапі попередньої обробки для класифікації. Цей процес нормалізації зменшує шум у даних і сприяє ефективному створенню ембедингів. Очищений текст потім токенізується, розбиваючи речення на менші одиниці, такі як слова або частини слів, що є критично важливим для моделей на основі трансформерів. Правильна попередня обробка забезпечує надходження до моделі структурованих вхідних даних, що підвищує точність і надійність отриманих ембедингів речень.

Рисунок 4.12 – Архітектура тематичного моделювання на основі удосконаленої моделі all-MiniLM-L6-v2

Оброблений текст перетворюється в числові представлення за допомогою моделі SentenceTransformer на основі трансформера. Спочатку токени кодуються в щільні вектори (ембединги токенів), що зберігають їхні індивідуальні значення. Далі до цих ембедингів додаються позиційні кодування, що дозволяє моделі зберігати порядок токенів у послідовності. Основу трансформера становлять кілька шарів кодування, у яких механізм багатоголової уваги вловлює контекстні взаємозв'язки між токенами в межах послідовності. Кожен шар уваги супроводжується залишковими з'єднаннями та нормалізацією шару, що забезпечує стабільність моделі та ефективний потік градієнтів під час навчання. Потім глибока нейронна мережа обробляє кожен токен незалежно, покращуючи представлення ознак, після чого застосовуються додаткові залишкові з'єднання та нормалізація. Фінальне представлення речення формується шляхом агрегації виходів шарів кодування, часто використовуючи середнє. Отримані фіксованорозмірні ембединги є насиченим представленням семантичного змісту вхідного тексту.

Після створення ембедингів речень застосовується BERTopic для виділення змістовних тем із тексту. BERTopic починає з кластеризації ембедингів SentenceTransformer, групуючи семантично подібні документи у кластери. Для цього використовуються такі методи: UMAP (для зменшення розмірності простору), HDBSCAN (для кластеризації на основі щільності). Для кожного кластеру застосовується векторизатор (наприклад, CountVectorizer) для виділення ключових слів і фраз, які найкраще представляють кластер. Модель представлення KeyBERTInspired додатково удосконалює ці ключові слова, забезпечуючи їхню семантичну релевантність. За необхідності можна дослідити ієрархічні зв'язки між темами, що дозволяє: об'єднувати схожі теми у ширші категорії та розбивати теми на більш дрібні підтеми.

Хоча BERTopic є основою запропонованого підходу до тематичного моделювання, у цьому дисертаційному дослідженні було введено вдосконалення, що покращують когерентність тем, ефективність кластеризації

та інтерпретованість, роблячи метод більш придатним для аналізу дезінформації. Ці модифікації включають оптимізацію моделі векторів речень та вдосконалення підходу до кластеризації для покращення якості тем, згенерованих моделлю.

Для покращення представлення текстових векторів, було систематично оцінено кілька моделей SentenceTransformer, перевищуючи стандартну модель paraphrase-MiniLM від BERTopic. Серед моделей, які розглядалися, були all-mpnet-base-v2, LaBSE, multi-qa-mpnet-base-dot-v1, multi-qa-MiniLM-L6-cos-v1, clip-ViT, all-roberta-large-v1, stsb-xlm-r-multilingual, sentence-t5-base та sentence-t5-xxl. Модель All-MiniLM-L6-v2 показала найкращий баланс між обчислювальною ефективністю та контекстуальним представленням. Це вибір забезпечив більш когерентні та відокремлені теми, гарантуючи масштабованість для більших наборів даних.

Окрім покращення текстових векторів, було вдосконалено методологію кластеризації, налаштувавши параметри UMAP, зокрема n_neighbors та min_dist, для оптимізації зменшення розмірності. Налаштування n_neighbors дозволило краще контролювати гранулярність тем, забезпечуючи правильне групування тісно пов'язаних тем та уникаючи надмірної фрагментації. За результатами емпіричних оцінок, значення n_neighbors = 15 та min_dist = 0.1 забезпечили найкращий баланс між відокремленням тем та їхньою когерентністю. Крім того, ми досліджували альтернативні стратегії кластеризації, що виходять за межі стандартного HDBSCAN в BERTopic, зокрема Agglomerative Clustering та KMeans. Ці вдосконалення покращили диференціацію перекритих наративів, підвищивши інтерпретованість витягнутих тем і забезпечивши більш надійний аналіз патернів дезінформації. Завдяки цим оптимізаціям, запропонований підхід покращує точність і інтерпретованість моделювання тем, роблячи його більш ефективним для складних завдань, пов'язаних з аналізом дезінформації.

Після налаштування моделей, генеруються візуалізації, що демонструють взаємозв’язки між темами та їхню структуру. Карта відстаней

між темами – 2D-проекція, що дає змогу швидко визначити подібні або унікальні теми. Ієрархічна кластеризація показує, як теми можуть бути об’єднані або поділені, допомагаючи оцінити градулярність тематики. Вихідні дані зберігаються у CSV-файлах та містять інформацію по кожній темі (наприклад, топові слова, розмір теми), інформацію по кожній вибірці даних (наприклад, визначена тема для текстової вибірки) та інші дані.

Для аналізу змісту набору даних [133] було проінтерпретовано карту міжтематичних відстаней, яка зображена на рисунку 4.13. Вона візуально відображає взаємозв’язки між темами, визначеними за допомогою тематичного моделювання на наборі даних. Кожна бульбашка відповідає окремій темі, а її розмір відображає відносну частку документів у наборі даних, віднесених до цієї теми. Чітка кластеризація, спостережувана на карті, свідчить про те, що набір даних містить добре визначені та роздільні теми, що є ідеальним для розуміння тематичної структури даних. Розмір і розташування бульбашок можуть допомогти у подальшому аналізі: більші та центральні теми заслуговують глибшого дослідження для виявлення ключових наративів, тоді як менші або периферійні теми можуть вказувати на нішеві дискусії або нові тренди. Ця візуалізація надає цінні інсайти щодо тематичного ландшафту набору даних, допомагаючи інтерпретувати домінуючі та нові наративи для стратегічної комунікації або планування заходів.

Аналіз наративів включає інтерпретацію CSV-файлу з темами, згенерованими кластеризаційною моделлю. Дані містять такі стовпці: “Count”, що вказує кількість одиниць даних, класифікованих у межах кожної теми; “Name of the topic”; “Representation”, що містить найпоширеніші словосполучення, пов’язані з темою; і “Representative_Docs”, що надає дані, найбільш релевантні для представлення теми. Модель також генерує клас -1, що містить викиди, які не можуть бути класифіковані до жодної з визначених тем.

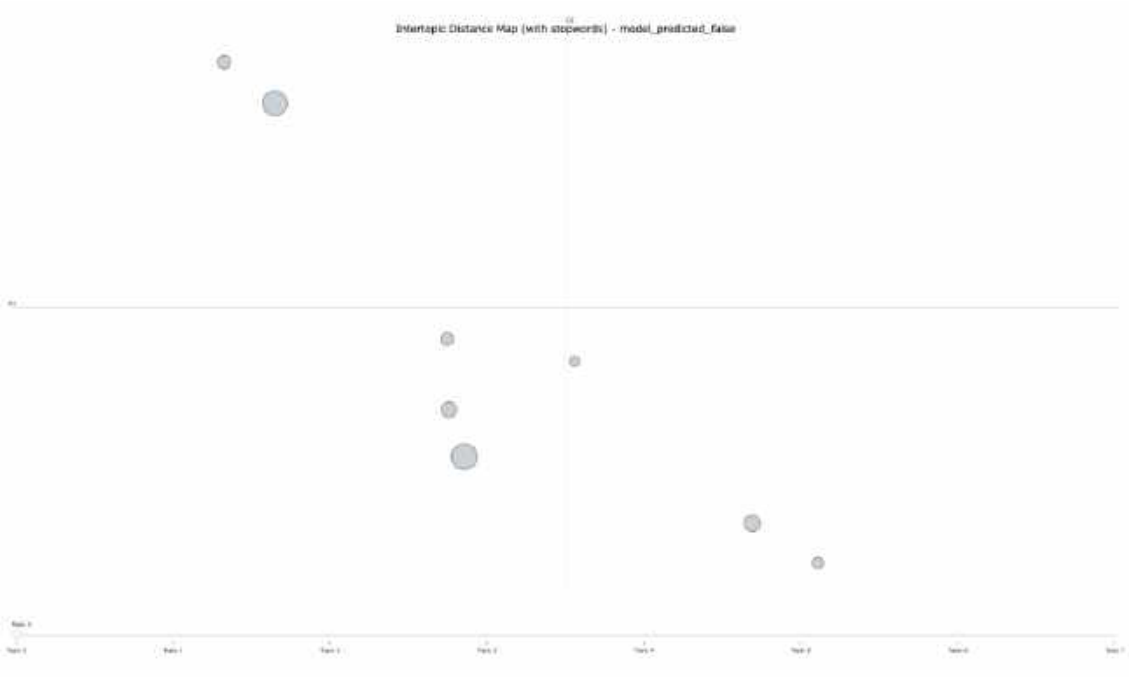

Рисунок 4.13 – Карта міжтематичних відстаней

Для набору даних про COVID-19, який розглянуто в розділі 3.4, інтерпретація восьми тем у дезінформаційному наборі виглядає наступним чином:

- Тема 0: Взаємодія громадськості в Індії. Ця тема зосереджена на наративах, пов’язаних із “соціальним дистанціюванням” та громадською поведінкою в Індії під час пандемії COVID-19. Наративи містять поєднання історій про дотримання та недотримання правил, зокрема “фото людей, які дотримуються соціального дистанціювання на ринку в Маніпурі” та “відео співробітника RPF, який бере хабарі від мігрантів, що повертаються додому під час локдауну в Гуджараті”. Інші приклади висвітлюють порушення, такі як “люди в Мумбаї ігнорують правила дистанціювання під час покупок перед святом Ід”.
- Тема 1: Теорії змови щодо науки про коронавірус. Ця тема містить псевдонаукову інформацію та дезінформацію про вірус, зокрема твердження про несправність ПЛР-тестів у Танзанії та непідтверджені методи лікування, такі як високі дози вітаміну D. Інші наративи стверджують, що коронавірус є штучною мутацією, або поширюють дезінформацію про його стійкість до температури.

- Тема 2: Місцевий італійський контекст. Ця тема зосереджена на подіях, пов'язаних із пандемією в "Італії" та "італійській" реакції, з представницькими фразами, такими як "італійський лікар" та "пандемія COVID". Документи містять історії про виснажених лікарів, які загинули від вірусу, офіційні урядові комунікації щодо заходів локдауну та медіа-спотворення подій. Ці наративи часто висвітлюють ранні труднощі Італії у боротьбі з пандемією та додають людського виміру до кризи через емоційні розповіді, хоча деякі з них виглядають перебільшеними або вигаданими.
- Тема 3: Наративи про Пакистан та індійську політику. Ця тема обговорює "COVID-пацієнтів" і "Пакистан", з особливим акцентом на політичних фігурах і їхніх заявах. Представницькі наративи включають твердження про те, що прем'єр-міністр Пакистану неправильно інтерпретує дані про COVID-19, та заяви індійських лідерів про безкоштовне лікування мільйонів пацієнтів.
- Тема 4: Теорії змови, пов'язані з Біллом Гейтсом. Значна частина набору даних зосереджена на теоріях змови про "Білла Гейтса" і пандемію як "фальшивий привід". Представницькі наративи стверджують, що пандемія була організована для скорочення населення Землі або імплантації мікрочіпів у населення. Ці наративи часто зображують Гейтса як центральну фігуру ширших теорій змови про глобальний контроль.
- Тема 5: COVID-19 і гуманітарна допомога. Ця тема характеризується фразами "гуманітарна допомога" та "випадки COVID". Ключові наративи включають твердження, що ООН висувала умови, такі як легалізація абортів, у пакеті гуманітарної допомоги для Еквадору. Інші повідомлення поширюють псевдонаукові твердження про те, що безсимптомні особи є здоровими і поширюють імунітет, а не вірус.
- Тема 6: Контроверсії навколо масок. Ця тема стосується триваючих дебатів щодо "масок" та їхньої ефективності. Представницькі документи включають твердження про обмеження урядом Франції

виробництва масок та звинувачення у зміні ВООЗ своїх рекомендацій щодо масок, стверджуючи, що вони підвищують ризик інфікування.

- Тема 7: Наративи, пов’язані з Іспанією. Зосереджена на “Іспанії” та її реакції на пандемію, ця тема містить твердження про заходи локдауну, такі як необхідність сертифікатів для виправдання пересування, та звинувачення у порушенні карантинних обмежень політичними лідерами. Представницькі документи також обговорюють чутки про повне закриття Мадрида, які згодом були спростовані.

4.6. Тематичне моделювання для аналізу наративів російської дезінформації про президентські вибори США

Для аналізу наративів була використана база даних російської дезінформації від EU East Stratcom Task Force, EUvsDisinfo [151] Використовуючи відкриту базу даних EUvsDisinfo, яка виявляє, збирає та викриває випадки дезінформації, що походять від прокремлівських ЗМІ, було зібрано два набори даних. Набір даних 1, під назвою “Дезінформація щодо президентських виборів США”, який був відфільтрований за датою з 1 січня 2024 року по 5 листопада 2024 року та за хештегом “US Presidential elections”, що дало 62 випадки статей російської дезінформації. Для того, щоб зрозуміти ширшу когнітивну рамку, у якій росія будує наративи щодо виборів у США, був використаний більший набір даних 2, під назвою “Дезінформація щодо США 2024”, відфільтрований за датою з 1 січня 2024 року по 5 листопада 2024 року та за країною “США”, що дало 616 випадків статей російської дезінформації, що зосереджені на дискурсі навколо Сполучених Штатів Америки.

Проведення попередньої обробки даних та застосування запропонованої у цьому дисертаційному дослідженні моделі тематичного моделювання, яка описана у розділі 4.5 дозволило провести аналіз набору даних. Це виявило п’ять основних тем у даному наборі, як показано в таблиці 4.3. Виходи моделі включають кількість тем, найбільш поширені ключові слова по всіх темах та

список статей, що були призначені моделлю до конкретних тем. Далі ці дані були інтерпретовані здобувачем для створення назви теми, її опису, мети наративу та класифікації цього наративу відповідно до трьох стратегічних цілей російської дезінформації:

- Підрив довіри до західних демократичних інституцій – поширення дезінформації, що ставить під сумнів легітимність демократичних процесів та інститутів.
- Сіяння розбрату та поляризації – використання соціальних напружень для створення конфліктів та поглиблення поділів у цільових суспільствах.
- Заплутування та відволікання аудиторії – заповнення інформаційного простору суперечливими та хибними наративами для створення плутанини.

Оскільки цей набір даних є невеликим, було можливим провести попередній ручний аналіз усіх даних для визначення тем, що містяться в ньому. Машинний аналіз підтвердив ці висновки, виявивши п’ять основних тем у наборі даних, які збігалися з попередньою оцінкою здобувача.

В аналізі набору даних щодо російської дезінформації стосовно президентських виборів США виокремлюється кілька важливих тем, які вказують на основну мету – підірвати довіру громадськості до американських інститутів. Однією з постійно повторюваних тем є наратив про санкції проти росії та її медіа, що зустрічається в кількох темах і зображує Сполучені Штати як несправедливу та авторитарну країну, яка придушує свободу слова. Обговорення конспіративної теорії про глибоку державу (Deep state), чесності виборів і критика еліт США представлені таким чином, що сприяють сумнівам щодо легітимності демократії у США. Нападки на політичні еліти та ліберальні цінності є частиною ширшої кампанії, спрямованої на ослаблення впевненості в американському політичному ландшафті та інституціях, при цьому підкреслюючи ідею переслідування росії “колективним заходом” що узгоджується з дезінформаційними стратегіями, які намагаються дестабілізувати довіру до уряду США. Тематичне моделювання для набору даних 2 виявило 13 різних тем, які були інтерпретовані здобувачем та позначені відповідно до стратегічних комунікаційних цілей росії (табл. 4.4).

Таблиця 4.3 – Аналіз наративів у наборі даних про президентські вибори у 2024 США

Тема	Опис теми	Ключові слова	Мета наративу	Ціль російської пропаганди
1. Глибока держава (Deep state)	Тема зосереджується на Дональді Трампі, теорії глибокої держави та санкцій, особливо проти російських ЗМІ.	‘trump’, ‘deep state’, ‘sanctions media’, ‘election 2020’,‘freedom of speech’	Ці наративи та поширення теорій змови, ймовірно, спрямовані на сіяння дискурсу та плутанини, а також на підсилення заяв про цензуру та переслідування свободи слова.	Заплутування та відволікання аудиторії
2. Демократична партія	Демократична партія та легітимність виборів США 2020. Обговорюються наративи, пов'язані з Україною.	‘democratic party’, ‘russian’, ‘ukraine’, ‘vote’, ‘war’, ‘interference’, ‘evidence’	Ці наративи, мають на меті поставити під сумнів достовірність демократичних процесів та зобразити Демократичну партію як корумповану чи скомпрометовану іноземними акторами.	Підрив довіри до західних демократичних інституцій
3. Еліти США, скандали	Спроби дискредитувати еліти США, зосереджуючи увагу на сенсаційних скандалах. Камала Гарріс є однією з головних фігур, разом із наративами, що стосуються П. Діді, сексуальних скандалів і звинувачень у зловживаннях.	‘elites’, ‘presidency’, ‘harris’, ‘diddy’, ‘covered’, ‘rapists’, ‘scandal’, ‘sex’	Метою є підрив довіри до політичних еліт, асоціюючи їх із скандалами, корупцією та правопорушеннями, підживлюючи ширші антиелітарні настрої.	Підрив довіри до західних демократичних інституцій

Продовження таблиці 4.3

4. Ліберальні цінності	Камала Гарріс та її дії як політичного лідера, включаючи припущення, що Гарріс має на меті просувати ліберальні цінності та збільшити тиск на окремі країни.	‘liberal’, ‘promote’, ‘pressure’, ‘values’, ‘harris kamala’	Тема використовує таке подання, щоб зобразити Гарріс як агресивну та маніпулятивну, створюючи наратив, що зображує її як загрозу традиційним цінностям.	Сіяння розбрату та поляризації; Підрив довіри до західних демократичних інституцій
5. Телеграм та Павло Дуров	Арешт Павла Дурова та збільшення Вашингтоном контролю над медіа особливо напередодні виборів. Байден зображується, як політик, що прагне контролю над інформацією, втручаючись у свободу слова.	‘election’, ‘biden’, ‘pavel’, ‘arrest’, ‘durov’, ‘russia’, ‘control’, ‘washington	Зображення адміністрації США як тиранічної та репресивної, орієнтуючись на впливових медіафігур для маніпулювання публічним дискурсом та сіяння поляризації для політичних вигод.	Сіяння розбрату та поляризації

Таблиця 4.4 – Аналіз наративів у наборі даних про США

Тема	Опис теми	Ключові слова	Мета наративу	Ціль російської пропаганди
1. Україна та Сполучені Штати	Війна в Україні та її взаємини зі Сполученими Штатами в центрі уваги. Наративи включають підготовку хімічної зброї Україною, використання американської хімічної зброї та роль України як інструмента для послаблення росії.	'ukraine prepares', 'ukraine uses', 'ukraine uses american'	Зобразити Україну як об’єкта, а не суб’єкта міжнародних відносин, що знаходиться під контролем західних патронів, в той час як росія подається як жертва маніпуляцій.	Заплутування та відволікання аудиторії
2. Центральна Азія	Захід втручається в Центральну Азію та намагається дестабілізувати вплив росії в регіоні. Наративи включають зусилля із ізоляції Центральної Азії від Москви та спроби дестабілізувати Казахстан через анти-російські настрої.	'to destabilise kazakhstan', 'central asia the', 'russia', 'front against russia', 'kazakhstans information', 'central asia'	Створити наратив про західну агресію, спрямовану на ізоляцію росії від ключових союзників і регіонів, що призводить до недовіри до західних країн.	Підрив довіри до західних демократичних інституцій, Сіяння розбрату та поляризації

Продовження таблиці 4.4

3. Північний потік, Крокус Сіті	Звинувачення Сполучених Штатів у прикритті атак і уникнення відповідальності за саботаж, наприклад, інцидент з Північним потоком та терористичний акт в Крокус Сіті, здійснені "українськими нацистами".	'nord stream sabotage', 'arrest the us', 'putin was', 'crocus hall'	Підрив довіри до Сполучених Штатів і Заходу, звинувачуючи США у співучасті у саботажах і терористичних актам, що приписуються Україні.	Підрив довіри до західних демократичних інституцій, Заплутування та відволікання аудиторії
4. Вірменія, Грузія та кольорові революції	Західне втручання в регіон Кавказу, зокрема в Грузії та Вірменії, де кольорові революції та перевороти нібито організовані Сполученими Штатами.	'protests in georgia', 'and azerbaijan', 'russia', 'armenia', 'caucasus', 'revolution in georgia', 'tbilisi'	Просування ідеї, що революції та політичні протести в регіоні є результатом дій західних держав, що намагаються дестабілізувати уряди і змінити політичні динаміки.	Підрив довіри до західних демократичних інституцій
5. Санкції проти російських медіа	Санкції, накладені на російські медіа, подаються як атака на свободу слова, мотивовані успіхом та ефективністю цих російських “медіа” .	'russia media is', 'sanctions against russian', 'rt sanctions show'	Дискредитація західних санкцій, подаючи їх як безпідставні дії проти успішного поширення “правдивих” наративів російських про-кремлівських платформ.	Сіяння розбрату та поляризації

Продовження таблиці 4.4

6. Амбіції Польщі	Польща нібито прагне розширити свій територіальний вплив на Україну за підтримки США.	'russia poland is', 'poland wants', 'ukraine poland', 'russia poland closed', 'polish foreign minister'	Підрив репутації Польщі шляхом сіяння брехливих намірів щодо її геополітичного місця у Східній Європі, зображуючи Польщу як опортуністичного агресора, готового експлуатувати Україну.	Заплутування та відволікання аудиторії
7. План миру Зеленського	Володимир Зеленський зображується як персонаж, що шантажує своїх “західних патронів” і чий мирний план повен прихованими корисливими мотивами.	'peace zelenskys', 'collective', 'zelenskyys victory plan', 'patrons zelenskyy blackmails'	Зобразити Україну як маріонеткову державу, підірвавши довіру до її керівництва та авторитет президента Зеленського.	Підрив довіри до західних демократичних інституцій, Сіяння розбрату та поляризації
8. Трамп і глибока держава	Дональд Трамп як мішень для змови “глибокої держави” і глобалістів, прив’язуючи спроби його вбивства, як підтвердження цієї теорії.	'trumps assassination attempt', 'deep state', 'trump'	Підсилення теорій змови про те, що Трамп є мішенню глобальних сил, тим самим заплутуючи аудиторію.	Заплутування та відволікання аудиторії

Продовження таблиці 4.4

9. НАТО і росія	Роль НАТО як супротивника Росії, зображення активностей НАТО як форм гібридної війни. Наративи популяризують тезу, що НАТО втрачає свою актуальність через конфлікт з росією.	'russia nato', 'nato plans', 'nato resorting', 'of confrontation nato', 'provocations nato'	Підрив довіри до НАТО, зображаючи його як агресивне формування, що було створене та функціонує аби послаблювати та провокувати росію, та натякаючи на занепад альянсу.	Підрив довіри до західних демократични х інституцій
10. НАТО і напад США на Росію	Сполучені Штати, НАТО та їхні союзники підтримують атаки на російські території та брали участь у проксі-війні проти росії, використовуючи Україну як інструмент.	'nato us', 'us nato', 'russia the us'	Зобразити Захід як безпосередньо залучений у військову агресію проти росії, таким чином легітимізуючи оборонні дії росії.	Підрив довіри до західних демократични х інституцій
11. Гегемо нія США та ЄС	Сполучені Штати та Європейський Союз подаються як ті, що нав'язують ліберальну цензуру та залежать один від одного, що вказує на відсутність автономії в європейському урядуванні.	'us eu', 'hegemony eu bureaucracy', 'military hegemony eu', 'nato europe opposes'	Підрив легітимності європейського урядування, зображаючи його як сильно залежне від інтересів США та вражене репресивними ліберальними ідеологіями.	Підрив довіри до західних демократични х інституцій

Продовження таблиці 4.4

12. Війна Заходу проти росії	Захід, під керівництвом Сполучених Штатів та НАТО, зображується як організатор гібридної війни проти росії, використовуючи Україну як проксі для дестабілізації росії.	'west provokes russia', 'to destroy russia', 'west uses ukraine'	Легітимізація дій росії, консолідація внутрішньої підтримки та зображення західних держав як агресорів, сприяючи недовірі до західних політик та спроби апелювати до солідарності всередині росії.	Підрив довіри до західних демократичних інституцій
13. Біоматеріали США	Сполучені Штати нібито транспортують біоматеріали з України для військових біологічних досліджень, використовуючи маршрути через Молдову та Румунію.	'us exports biomaterials', 'biomaterials from ukraine', 'romania us research'	Створення конспіративних теорій щодо дослідженнях біологічної зброї в Україні для підсилення тривоги, сіяння страху та недовіри до західному та його допомоги Україні.	Заплутування та відволікання аудиторії

Моделювання тем виявляє кілька повторюваних наративів, таких як теми 9, 10 і 12, що зосереджені на нібито агресії Заходу до росії, але з дещо різних точок зору – гібридна війна НАТО, безпосередня участь США та НАТО в атаках і ширша стратегія війни Заходу проти росії. Акцент на Україні, Зеленському та війні відображає сильний фокус на війні в Україні в дезінформаційних наративах, з зусиллями представити українське керівництво як маріонетку західних сил. Часті відсилки до інших країн, таких як Польща, Казахстан, Грузія та Білорусь, вказують на спробу підкреслити вплив, який здійснюється НАТО та Заходом у сусідніх з росією регіонах чи так званих зонах "геополітичного інтересу" росії. Постійно асоціюючи ці країни та їх лідерів з ключовими словами, такими як "терорист", "переворот" і "незаконний", пропаганда намагається підірвати їх легітимність і зобразити їх як інструменти західних інтересів, а не незалежних суб'єктів. Найчастіші цілі російська пропаганда планувала досягти цими наративами передбачали: підрив довіри до західних інститутів та заплутування аудиторії конспіративними теоріями, перебільшеними чи необґрунтованими заявами.

4.7. Висновки до четвертого розділу

1. Вперше запропоновано фреймворк комплексного аналізу текстової дезінформації та пропаганди, заснований на комбінуванні технологій двоспрямованого кодувального представлення з трансформерів (BERT) та двоспрямованої довгої короткочасної пам'яті (Bi-LSTM) для створення інтерпретованих векторних представлень тексту, який на відміну від існуючих технік, об'єднує виявлення та тематичний аналіз дезінформації, що дозволяє забезпечити ефективне виявлення та аналіз текстової дезінформації та пропаганди в соціальних мережах.

2. Удосконалено метод тематичного моделювання, заснований на застосуванні механізмів багатоголової уваги, що на відміну від існуючих

використовує глибокі ембединги для контекстуалізації даних, що дозволяє забезпечити чітке розмежування між кластерами.

3. Проведено експериментальне дослідження розробленого фреймворку на наборі фейкових та реальних новин WELFake.

4. Проведено експериментальне дослідження розробленого фреймворку для набору даних місінформації про COVID CoVID19-FNIR, яка поширювалась під час кризи громадського здоров’я.

5. Проведено експериментальне дослідження розробленого фреймворку для україномовного набору даних дезінформації щодо повномасштабного вторгнення росії в Україну з телеграму.

6. Проведено експериментальне дослідження розробленого фреймворку для аналізу російської дезінформації щодо президентських виборів у США у час кампанії.

Основні результати, представлені в цьому розділі, опубліковані в [14], [21], [23], [29].

ВИСНОВКИ

Дисертаційне дослідження присвячене розробці ефективних моделей та методів для виявлення та аналізу текстової дезінформації та пропаганди в соціальних мережах. Мета дисертаційної роботи, яка полягала в ефективному виявленні та аналізі текстової дезінформації та пропаганди в соціальних мережах, досягнута. Завдання дослідження, зокрема аналіз теоретичної бази понять інформаційної війни, дезінформації та пропаганди, розгляд головних принципів та підходів російської дезінформації, аналіз наявних моделей та методів машинного навчання для аналізу текстової дезінформації та пропаганди, розробка моделей для аналізу дезінформації та пропаганди на основі методів машинного навчання, розробка моделей на основі ансамблевих методів машинного навчання, розробка моделей на основі методів глибокого машинного навчання, розробка фреймворку для комплексного аналізу текстової дезінформації та пропаганди у соціальних мережах, вирішення практичних завчань пов’язаних виявленням та аналізом текстової дезінформації та пропаганди за допомогою розроблених моделей, досягнуті в повному обсязі.

Вперше запропоновано фреймворк комплексного аналізу текстової дезінформації та пропаганди, заснований на комбінуванні технологій двоспрямованого кодувального представлення з трансформерів (BERT) та двоспрямованої довгої короткочасної пам’яті (Bi-LSTM) для створення інтерпретованих векторних представлень тексту, який на відміну від існуючих технік, об’єднує виявлення та тематичний аналіз дезінформації, що дозволяє забезпечити ефективне виявлення та аналіз текстової дезінформації та пропаганди в соціальних мережах.

Проаналізовано теоретичні засади інформаційних маніпуляцій у контексті інформаційних воєн. Розглянуто основні типи інформаційних розладів та неконвенційних воєн. Розглянуто застосування технологій ШІ для протидії дезінформації, інформаційним маніпуляціям, пропаганді та фейкам. Проаналізовано методи машинного та глибокого навчання для протидії

дезінформації, інформаційним маніпуляціям, пропаганді та фейкам. Розглянуто інформаційні технології аналізу дезінформації та пропаганди.

Вперше розроблено модель класифікації текстової дезінформації та пропаганди, заснований на поєднанні трансформерів і архітектури Bi-LSTM для інтеграції лінійних та нелінійних залежностей текстових даних, яка на відміну від існуючих використовує глибокі контекстуальні ембедінги і механізми уваги, що дозволяє підвищити точність моделі.

Удосконалено моделі глибокого навчання для класифікації текстової дезінформації та пропаганди, засновані на архітектурах XLNet, Bi-LSTM та Attention-Based Bi-LSTM, які на відміну від існуючих застосовують адаптивне оцінювання важливих фрагментів тексту, що дозволяє забезпечувати високий рівень продуктивності класифікації.

Удосконалено метод тематичного моделювання, заснований на застосуванні механізмів багатоголової уваги, що на відміну від існуючих використовує глибокі ембедінги для контекстуалізації даних, що дозволяє забезпечити чітке розмежування між кластерами.

Дістали подальшого розвитку моделі для класифікації текстової дезінформації та пропаганди, засновані на ансамблевих методах машинного навчання, які на відміну від існуючих враховують балансування класів і адаптуються до змін вхідних даних, що дозволяє підвищити продуктивність моделей.

Дістали подальшого розвитку моделі для класифікації інформації, засновані на статистичних методах машинного навчання, які на відміну від існуючих адаптовані до класифікації текстової дезінформації та пропаганди, що дозволяє створення систем для виявлення місінформації, дезінформації та малінформації.

Використання розроблених моделей та методів дозволило збільшити точність класифікації виявлення текстової дезінформації та пропаганди у соціальних мережах. Використання розробленого фреймворку забезпечує аналіз наративів який може бути використаний для побудови стратегій

контрзаходів та стратегічних комунікацій. Наприклад, застосування фреймворку для аналізу дезінформації про вибори, COVID-19 та російську війну проти України дозволить отримати цінне розуміння наративів у великих корпусах даних.

Практичне значення дисертаційного дослідження підтверджене апробацією на міжнародних наукових та науково-практичних конференціях, розробкою політик та впровадженням в Balsillie School of International Affairs, Center for International Governance Innovation, Державну установу «Харківський обласний центр контролю та профілактики хвороб при Міністерстві охорони здоров'я», а також у навчальний процес кафедри математичного моделювання та штучного інтелекту та наукову роботу Національного аерокосмічного університету «Харківський авіаційний інститут».

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] K. Giles, ‘Handbook of Russian Information Warfare’, 2016. [Online]. Available: https://css.ethz.ch/content/dam/ethz/special-interest/gess/cis/center-for-securities-studies/resources/docs/NDC%20fm_9.pdf

[2] P. Pomerantsev, *Nothing Is True and Everything Is Possible : the Surreal Heart of the New Russia*. Publicaffairs, 2014.

[3] O. Fridman, *Russian ‘Hybrid Warfare’*. Oxford University Press, 2018. doi: 10.1093/oso/9780190877378.001.0001.

[4] T. Rid, *Active Measures*. Farrar, Straus and Giroux, 2020.

[5] E. Treyger, J. Cheravitch, and R. Cohen, ‘Russian Disinformation Efforts on Social Media’, 2022.

[6] NATO Strategic Communications Centre of Excellence, ‘Russia’s Information Influence Operations in the Nordic - Baltic Region’, Stratcomcoe.org. [Online]. Available: https://stratcomcoe.org/publications/russias-information-influence-operations-in-the-nordic-baltic-region/314

[7] NATO Strategic Communications Centre of Excellence, ‘Russian information operations outside of the Western information environment’, Stratcomcoe.org. [Online]. Available: https://stratcomcoe.org/publications/russian-information-operations-outside-of-the-western-information-environment-revised-version/316

[8] EUvsDisinfo, ‘Disinformation to Conceal War Crimes: Russia is Lying About Atrocities in Bucha’, EUvsDisinfo. [Online]. Available: https://euvsdisinfo.eu/disinformation-to-conceal-war-crimes-russia-is-lying-about-atrocities-in-bucha/

[9] EUvsDisinfo, ‘Doppelganger operation’, EU DisinfoLab. [Online]. Available: https://www.disinfo.eu/doppelganger-operation/

[10] Atlantic Council, ‘Disinformation’, Atlantic Council. [Online]. Available: https://www.atlanticcouncil.org/issue/disinformation/

[11] U. Reisach, ‘The Responsibility of Social Media in Times of Societal and Political Manipulation’, *Eur. J. Oper. Res.*, vol. 291, no. 3, pp. 906–917, 2020, doi: 10.1016/j.ejor.2020.09.020.

[12] D. A. Broniatowski *et al.*, ‘Weaponized Health Communication: Twitter Bots and Russian Trolls Amplify the Vaccine Debate’, *Am. J. Public Health*, vol. 108, no. 10, pp. 1378–1384, 2018, doi: 10.2105/ajph.2018.304567.

[13] Detector Media, ‘#DisinfoChronicle. Кремлівська пропаганда щодо військового наступу на Україну - Детектор медіа.’, Detector.media. Accessed: Jan. 01, 2025. [Online]. Available: https://disinfo.detector.media/search?search_string=%D0%9D%D0%90%D0%A2%D0%9E&search_tag=

[14] H. Padalko, V. Chomko, S. Yakovlev, and D. Chumachenko, ‘A Novel Comprehensive Framework for Detecting and Understanding Health-Related Misinformation’, *Information*, vol. 16, no. 3, p. 175, 2025, doi: 10.3390/info16030175.

[15] H. Padalko, V. Chomko, and D. Chumachenko, ‘A novel approach to fake news classification using LSTM-based deep learning models’, *Front. Big Data*, vol. 6, 2024, doi: 10.3389/fdata.2023.1320800.

[16] H. Padalko, V. Chomko, S. Yakovlev, and D. Chumachenko, ‘Ensemble machine learning approaches for fake news classification’, *Radioelectron. Comput. Syst.*, no. 4, pp. 5–19, 2023, doi: 10.32620/reks.2023.4.01.

[17] H. Padalko, V. Chomko, S. Yakovlev, and P. P. Morita, ‘Classification of disinformation in hybrid warfare: an application of XLNet during the Russia’s war against Ukraine’, *Radioelectron. Comput. Syst.*, vol. 2024, no. 4, pp. 46–58, 2024, doi: 10.32620/reks.2024.4.04.

[18] M. Lotto *et al.*, ‘Ethical principles for infodemiology and infoveillance studies concerning infodemic management on social media’, *Front. Public Health*, vol. 11, 2023, doi: 10.3389/fpubh.2023.1130079.

[19] A. Fitz-Gerald, D. Chumachenko, H. Padalko, and V. Ganesh, ‘Artificial Intelligence and New Threat Vectors: Using Scenario Planning for Trend Forecasting’, *Balsillie Pap.*, vol. 5, no. 6, 2023, doi: 10.51644/bap56.

[20] H. Padalko, V. Chomko, and D. Chumachenko, ‘Misinformation Detection in Political News using BERT Model’, presented at the CEUR Workshop Proceedings, 2023, pp. 117–127.

[21] H. Padalko, V. Chomko, and D. Chumachenko, ‘The Impact of Stopwords Removal on Disinformation Detection in Ukrainian language during Russian-Ukrainian war’, presented at the CEUR Workshop Proceedings, 2024, pp. 87–101.

[22] H. Padalko, V. Chomko, and D. Chumachenko, ‘COVID-19 Misinformation Detection on Weibo by Support Vector Classifier’, *IEEE Int. Conf. Dependable Syst. Serv. Technol. DESSERT*, p. 197136, 2023, doi: 10.1109/dessert61349.2023.10416460.

[23] H. Padalko and P. Morita, ‘Covid-19 “infodemic” management: russia’s key tool for influencing geopolitics’, presented at the Population Medicine, EU European Publishing, 2023, p. 512. doi: 10.18332/popmed/164626.

[24] H. Padalko, ‘Machine learning approach for classification of russian propaganda’, presented at the Матеріали V Міжнародній науково-практичній конференції IT-професіоналів та аналітиків комп’ютерних систем “ProfIT Conference”, 2023, pp. 96–97.

[25] Г. А. Падалко and С. В. Яковлев, ‘Модель розповсюдження COVID-19 на основі аналізу соціальних мереж’, presented at the Матеріали IV Міжнародній науково-практичній конференції IT-професіоналів та аналітиків комп’ютерних систем “ProfIT Conference”, 2021, p. 106.

[26] Г. А. Падалко, ‘Мультиагентний підхід як інструмент моделювання соціальних мереж’, presented at the Матеріали III Міжнародній науково-практичній конференції IT-професіоналів та аналітиків комп’ютерних систем “ProfIT Conference”, 2020, p. 112.

[27] R. Vysochanskyy, H. Padalko, and A. Ollila, ‘Information Warfare in Canada: Countering AI-backed Misinformation, Disinformation and Malinformation (MDM) Surrounding the Russian Invasion of Ukraine’, 2023. [Online]. Available: https://uwaterloo.ca/defence-security-foresight-group/sites/default/files/uploads/documents/vysochanskyy-et-al_information-warfare-in-canada.pdf

[28] H. Padalko and A. Fitz-Gerald, ‘The Need for a Strategic Approach to Disinformation and AI-Driven Threats’, Rusi.org. [Online]. Available: https://www.rusi.org/explore-our-research/publications/commentary/need-strategic-approach-disinformation-and-ai-driven-threats

[29] H. Padalko, ‘Russian Disinformation about the US Election: AI Analysis of Narratives’, Center for Internation Governance Innovation, 2025. Accessed: Jan. 01, 2025. [Online]. Available: https://www.cigionline.org/static/documents/DPH-paper-Padalko.pdf

[30] World Economic Forum, ‘Global Risks Report 2024’, 2024. [Online]. Available: https://www.weforum.org/publications/global-risks-report-2024/

[31] World Economic Forum, ‘Global Risks Report 2025’, 2025. [Online]. Available: https://www.weforum.org/publications/global-risks-report-2025/

[32] OECD, ‘Facts not Fakes: Tackling Disinformation, Strengthening Information Integrity’, 2025. Accessed: Jan. 01, 2025. [Online]. Available: https://www.oecd.org/en/publications/facts-not-fakes-tackling-disinformation-strengthening-information-integrity_d909ff7a-en/full-report.html

[33] C. Wardle and H. Derakhshan, ‘Information Disorder: Toward an interdisciplinary framework for research and policy making’, 2017. [Online]. Available: https://rm.coe.int/information-disorder-report-version-august-2018/16808c9c77

[34] M. Miron, ‘Russian “hybrid warfare” and the annexation of Crimea: the modern application of Soviet political warfare’, *Int. Aff.*, vol. 99, no. 1, pp. 397–398, 2023, doi: 10.1093/ia/iiac279.

[35] C. Wardle, ‘Understanding information disorder’, First Draft. [Online]. Available: https://firstdraftnews.org/long-form-article/understanding-information-disorder/

[36] T. Nelson, N. Kagan, C. Critchlow, A. Hillard, and A. Hsu, ‘The Danger of Misinformation in the COVID-19 Crisis’, *Mo. Med.*, vol. 117, no. 6, p. 510, 2020.

[37] M. S. Schmidt, ‘Trump Invited the Russians to Hack Clinton. Were They Listening?’, *The New York Times*, 2018. [Online]. Available: https://www.nytimes.com/2018/07/13/us/politics/trump-russia-clinton-emails.html

[38] Ukraine World, ‘Fake History: How Russia Manipulates Ukraine’s past’, Ukraineworld.org. [Online]. Available: https://ukraineworld.org/en/articles/analysis/russia-manipulates-ukraines-past

[39] Canadian Centre for Cyber Security, ‘How to Identify misinformation, disinformation, and Malinformation’, Canadian Centre for Cyber Security. [Online]. Available: https://www.cyber.gc.ca/en/guidance/how-identify-misinformation-disinformation-and-malinformation-itsap00300

[40] EEAS, ‘1st EEAS Report on Foreign Information Manipulation and Interference Threats | EEAS’, www.eeas.europa.eu. [Online]. Available: https://www.eeas.europa.eu/eeas/1st-eeas-report-foreign-information-manipulation-and-interference-threats_en

[41] U.S. Department of Justice, ‘Justice Department Disrupts Covert Russian Government-Sponsored Foreign Malign Influence Operation Targeting Audiences in the United States and Elsewhere’, Justice.gov. [Online]. Available: https://www.justice.gov/opa/pr/justice-department-disrupts-covert-russian-government-sponsored-foreign-malign-influence

[42] T.-C. Hung and T.-W. Hung, ‘How China’s Cognitive Warfare Works: a Frontline Perspective of Taiwan’s Anti-Disinformation Wars’, *J. Glob. Secur. Stud.*, vol. 7, no. 4, 2022, doi: 10.1093/jogss/ogac016.

[43] S. K. Nur and E. Şeker, ‘Cyber Warfare: Terms, Issues, Laws and Controversies’, arXiv.org. [Online]. Available: https://arxiv.org/abs/1905.13532

[44] D. Ventre, *Information Warfare*. Iste Ltd. ; John Wiley & Sons, 2016.

[45] M. Warner, ‘US Cyber Command’s First Decade’, *U. S. Defend Forw. Cyber Strategy*, pp. 33–64, 2022, doi: 10.1093/oso/9780197601792.003.0004.

[46] M. Jaitner and H. Kantola, ‘Applying Principles of Reflexive Control in Information and Cyber Operations’, *J. Inf. Warf.*, vol. 15, no. 4, pp. 27–38, 2016, doi: 10.2307/26487549.

[47] E. Rindzevičiūtė, ‘Reflexive Control’, *Cornell Univ. Press EBooks*, pp. 122–149, 2023, doi: 10.7591/cornell/9781501769771.003.0007.

[48] M. Grisé, A. Demus, Y. Shokh, M. Kepe, J. W. Welburn, and K. Holynska, ‘Rivalry in the Information Sphere: Russian Conceptions of Information Confrontation’, 2022. [Online]. Available: https://www.rand.org/pubs/research_reports/RRA198-8.html

[49] Центр Стратегічних Комунікацій, ‘Терор як покликання: Як з 2014 втілювалася «Доктрина Герасимова»’, Центр Стратегічних Комунікацій. [Online]. Available: https://spravdi.gov.ua/teror-yak-poklykannya-yak-z-2014-vtilyuvalasya-doktryna-gerasymova/

[50] Mi. Kelley, ‘Understanding Russian Disinformation and How the Joint Force Can Address It’, US Army War College - Publications. [Online]. Available: https://publications.armywarcollege.edu/News/Display/Article/3789933/understanding-russian-disinformation-and-how-the-joint-force-can-address-it/

[51] D. Matter, E. Kuznetsova, V. Vziatysheva, I. Vitulano, and J. Pfeffer, ‘Temporally Stable Multilayer Network Embeddings: A Longitudinal Study of Russian Propaganda’, arXiv.org. [Online]. Available: https://arxiv.org/abs/2307.10264

[52] I. Yablokov, ‘Russian disinformation finds fertile ground in the West’, *Nat. Hum. Behav.*, vol. 6, no. 6, pp. 766–767, 2022, doi: 10.1038/s41562-022-01399-3.

[53] Global Affairs Canada, ‘Global Affairs Canada statement on Russian disinformation’, Canada.ca. [Online]. Available:

https://www.canada.ca/en/global-affairs/news/2024/10/global-affairs-canada-statement-on-russian-disinformation.html

[54] United States Department of Justice, ‘Justice Department Leads Efforts among Federal, International, and Private Sector Partners to Disrupt Covert Russian Government-Operated Social Media Bot Farm’, www.justice.gov. [Online]. Available: https://www.justice.gov/opa/pr/justice-department-leads-efforts-among-federal-international-and-private-sector-partners

[55] European Commission, ‘Commission opens formal proceedings against TikTok on election risks under the Digital Services Act’, European Commission. [Online]. Available: https://ec.europa.eu/commission/presscorner/detail/en/ip_24_6487

[56] K. Connolly, ‘Germany unearths pro-Russia disinformation campaign on X’, *The Guardian*, 2024. [Online]. Available: https://www.theguardian.com/world/2024/jan/26/germany-unearths-pro-russia-disinformation-campaign-on-x

[57] Secrétariat général de la défense et de la sécurité nationale, ‘RRN : une campagne numérique de manipulation de l’information complexe et persistante’, Gouv.fr. [Online]. Available: https://www.sgdsn.gouv.fr/publications/maj-19062023-rrn-une-campagne-numerique-de-manipulation-de-linformation-complexe-et

[58] D. Ingram, ‘How Elon Musk turned X into a pro-Trump echo chamber’, NBC News. [Online]. Available: https://www.nbcnews.com/tech/social-media/elon-musk-turned-x-trump-echo-chamber-rcna174321

[59] J. Kaplan, ‘More Speech and Fewer Mistakes’, Meta. [Online]. Available: https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/

[60] X, ‘Non-commercial use of the X API’, X.com. [Online]. Available: https://developer.x.com/en/developer-terms/commercial-terms

[61] N. Bontridder and Y. Poullet, ‘The Role of Artificial Intelligence in Disinformation’, *Data Policy*, vol. 3, no. E32, 2021, doi: 10.1017/dap.2021.20.

[62] G. Bendiab, H. Haiouni, I. Moulas, and S. Shiaeles, ‘Deepfakes in digital media forensics: Generation, AI-based detection and challenges’, *J. Inf. Secur. Appl.*, vol. 88, pp. 103935–103935, 2024, doi: 10.1016/j.jisa.2024.103935.

[63] J. Joseph, ‘Meta: AI Made Less Than 1% of Election Misinformation on Our Apps’, PCMAG. [Online]. Available: https://www.pcmag.com/news/meta-ai-less-than-1-percent-election-misinformation-instagram-facebook

[64] B. Allyn, ‘Deepfake Video of Zelenskyy Could Be “Tip of the Iceberg” in Info war, Experts Warn’, *NPR*, 2022. [Online]. Available: https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia

[65] В. Міський and Н. Соколенко, ‘Технологія ШІ розвивається: Визначати штучно згенеровані зображення стає дедалі складніше’, detector.media. [Online]. Available: https://detector.media/mediumy/article/237046/2025-01-14-tekhnologiya-shi-rozvyvaietsya-vyznachaty-shtuchno-zghenerovani-zobrazhennya-staie-dedali-skladnishe-oksana-polulyakh-u-podkasti-mediumy/

[66] E. Cano-Marin, M. Mora-Cantallops, and S. Sanchez-Alonso, ‘The power of big data analytics over fake news: A scientometric review of Twitter as a predictive system in healthcare’, *Technol. Forecast. Soc. Change*, vol. 190, no. 122386, 2023, doi: 10.1016/j.techfore.2023.122386.

[67] M. Park and S. Chai, ‘Constructing a User-Centered Fake News Detection Model by Using Classification Algorithms in Machine Learning Techniques’, *IEEE Access*, vol. 11, pp. 71517–71527, 2023, doi: 10.1109/ACCESS.2023.3294613.

[68] F. Fifita, J. Smith, M. B. Hanzsek-Brill, X. Li, and M. Zhou, ‘Machine Learning-Based Identifications of COVID-19 Fake News Using Biomedical Information Extraction’, *Big Data Cogn. Comput.*, vol. 7, no. 1, p. 46, 2023, doi: 10.3390/bdcc7010046.

[69] N. Fahad *et al.*, ‘Stand up Against Bad Intended News: An Approach to Detect Fake News using Machine Learning’, *Emerg. Sci. J.*, vol. 7, no. 4, pp. 1247–1259, 2023, doi: 10.28991/esj-2023-07-04-015.

[70] Y. Tohabar, N. Nasrah, and A. M. Samir, ‘Bengali Fake News Detection Using Machine Learning and Effectiveness of Sentiment as a Feature’, *2021 Jt. 10th Int. Conf. Inform. Electron. Vis. ICIEV 2021 5th Int. Conf. Imaging Vis. Pattern Recognit. IcIVPR*, 2021, doi: 10.1109/icievicivpr52578.2021.9564138.

[71] S. A. Yousif and R. Jehad, ‘Classification of Covid-19 fake news using machine learning algorithms’, *AIP Conf. Proc.*, vol. 2483, no. 1, p. 070007, 2022, doi: 10.1063/5.0117133.

[72] Z. Tian and S. Baskiyar, ‘Fake News Detection using Machine Learning with Feature Selection’, IEEE Xplore. Accessed: Jan. 01, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9776346

[73] D. Choudhury and T. Acharjee, ‘A Novel Approach to Fake News Detection in Social Networks Using Genetic Algorithm Applying Machine Learning Classifiers’, *Multimed. Tools Appl.*, vol. 82, no. 6, 2022, doi: 10.1007/s11042-022-12788-1.

[74] E. M. Mercha and H. Benbrahim, ‘Machine Learning and Deep Learning for sentiment analysis across languages: A survey’, *Neurocomputing*, vol. 15, no. 531, 2023, doi: 10.1016/j.neucom.2023.02.015.

[75] M. Yousefi-Azar and L. Hamey, ‘Text summarization using unsupervised deep learning’, *Expert Syst. Appl.*, vol. 68, pp. 93–105, 2017, doi: 10.1016/j.eswa.2016.10.017.

[76] Md. N. Y. Ali, Md. L. Rahman, J. Chaki, N. Dey, and K. C. Santosh, ‘Machine translation using deep learning for universal networking language based on their structure’, *Int. J. Mach. Learn. Cybern.*, vol. 12, no. 8, pp. 2365–2376, 2021, doi: 10.1007/s13042-021-01317-5.

[77] N. Capuano, G. Fenza, V. Loia, and F. D. Nota, ‘Content-Based Fake News Detection With Machine and Deep Learning: a Systematic Review’, *Neurocomputing*, vol. 530, pp. 91–103, 2023, doi: 10.1016/j.neucom.2023.02.005.

[78] Y. Akter and B. Arora, ‘Deep Learning Techniques Used for Fake News Detection: A Review and Analysis’, *Roc Int Conf Recent Innov Comput Lect Electr Eng*, no. 1001, pp. 127–140, 2023, doi: 10.1007/978-981-19-9876-8_11.

[79] M. R. Islam, S. Liu, X. Wang, and G. Xu, ‘Deep learning for misinformation detection on online social networks: a survey and new perspectives’, *Soc. Netw. Anal. Min.*, vol. 10, no. 1, 2020, doi: 10.1007/s13278-020-00696-x.

[80] ‘Дисертація (1)’.

[81] R. Dutta, D. Ranjan, and M. Majumder, ‘A Deep Learning Model for Classification of COVID-19 Related Fake News’, *Lect Electr Eng*, vol. 860, pp. 449–456, 2022, doi: 10.1007/978-981-16-9488-2_42.

[82] H. J. Alshahrani *et al.*, ‘Hunter Prey Optimization with Hybrid Deep Learning for Fake News Detection on Arabic Corpus’, *Comput. Mater. Contin.*, vol. 75, no. 2, pp. 4255–4272, 2023, doi: 10.32604/cmc.2023.034821.

[83] T. H. Vo, T. L. T. Phan, and K. C. Ninh, ‘Development of a fake news detection tool for Vietnamese based on deep learning techniques’, *East.-Eur. J. Enterp. Technol.*, vol. 5, no. 2(119), pp. 14–20, 2022, doi: 10.15587/1729-4061.2022.265317.

[84] D. Salh and R. M. Nabi, ‘Kurdish Fake News Detection Based on Machine Learning Approaches’, *Passer J. Basic Appl. Sci.*, vol. 5, no. 2, pp. 262–271, 2023, doi: 10.24271/psr.2023.380132.1226.

[85] N. Kausar, A. AliKhan, and M. Sattar, ‘Towards better representation learning using hybrid deep learning model for fake news detection’, *Soc. Netw. Anal. Min.*, vol. 12, no. 1, 2022, doi: 10.1007/s13278-022-00986-6.

[86] K. Ivancová, M. Sarnovský, and V. Maslej-Krcšňáková, ‘Fake news detection in Slovak language using deep learning techniques’, IEEE Xplore. Accessed: Jan. 01, 2022. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9378650

[87] M.-A. Ouassil, B. Cherradi, S. Hamida, M. Errami, O. El Gannour, and A. Raihani, ‘A Fake News Detection System based on Combination of Word

Embedded Techniques and Hybrid Deep Learning Model', *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 10, 2022, doi: 10.14569/ijacsa.2022.0131061.

[88] S. A. Althubiti, F. Alenezi, and R. F. Mansour, 'Natural Language Processing with Optimal Deep Learning Based Fake News Classification', *Comput. Mater. Contin.*, vol. 73, no. 2, pp. 3529–3544, 2022, doi: 10.32604/cmc.2022.028981.

[89] B. Nordin, R. Alfred, C. P. Yee, S. H. Tanalol, R. V. Loudin, and Z. Iswandono, 'Malay Fake News Classification Using a Deep Learning Approach', *Lect. Notes Electr. Eng.*, vol. 983, pp. 17–32, 2023, doi: 10.1007/978-981-19-8406-8_2.

[90] World Health Organization, 'Infodemic', www.who.int. [Online]. Available: https://www.who.int/health-topics/infodemic

[91] I. F. Strydom and J. Grobler, 'Transformers for COVID-19 Misinformation Detection on Twitter: A South African Case Study', *Lect. Notes Comput. Sci.*, no. 13810, pp. 197–210, 2023, doi: 10.1007/978-3-031-25599-1_15.

[92] J.-A. Chen, C.-L. Hung, and C.-Y. Wu, 'Attention-Based Deep Learning Models for Detecting Misinformation of Long-Term Effects of COVID-19', *2024 IEEE Conf. Artif. Intell. CAI*, pp. 240–245, 2024, doi: 10.1109/cai59869.2024.00053.

[93] M. Bozuyla and A. Özcift, 'Developing a fake news identification model with advanced deep languagetransformers for Turkish COVID-19 misinformation data', *Turk. J. Electr. Eng. Comput. Sci.*, vol. 30, no. 3, pp. 908–926, 2022, doi: 10.55730/1300-0632.3818.

[94] M. G. Kim, M. Kim, J. H. Kim, and K. Kim, 'Fine-Tuning BERT Models to Classify Misinformation on Garlic and COVID-19 on Twitter', *Int. J. Environ. Res. Public. Health*, vol. 19, no. 9, p. 5126, 2022, doi: 10.3390/ijerph19095126.

[95] A. Mulahuwaish, M. Osti, K. Gyorick, M. Maabreh, A. Gupta, and B. Qolomany, 'CovidMis20: COVID-19 Misinformation Detection System on Twitter Tweets Using Deep Learning Models', *Lect. Notes Comput. Sci.*, no. 13741, pp. 466–479, 2023, doi: 10.1007/978-3-031-27199-1_47.

[96] M. N. Alenezi and Z. M. Alqenaei, ‘Machine Learning in Detecting COVID-19 Misinformation on Twitter’, *Future Internet*, vol. 13, no. 10, p. 244, 2021, doi: 10.3390/fi13100244.

[97] S. S. Birunda, R. K. Devi, M. Muthukannan, and M. M. Babu, ‘ACOVMD: Automatic COVID-19 misinformation detection in Twitter using self-trained semi-supervised hybrid deep learning model’, *Int. Soc. Sci. J.*, vol. 74, no. 252, pp. 713–730, 2023, doi: 10.1111/issj.12475.

[98] I. Alsmadi, N. M. Rice, and M. J. O’Brien, ‘Fake or not? Automated detection of COVID-19 misinformation and disinformation in social networks and digital media’, *Comput. Math. Organ. Theory*, vol. 30, 2022, doi: 10.1007/s10588-022-09369-w.

[99] F. Mlawa, E. Mkoba, and N. Mduma, ‘A Machine Learning Model for detecting Covid-19 Misinformation in Swahili Language’, *Eng. Technol. Appl. Sci. Res.*, vol. 13, no. 3, pp. 10856–10860, 2023, doi: 10.48084/etasr.5636.

[100] H. Xu, M. Curci, S. Ek, P. Liu, Z. Li, and S. Xu, ‘COS2: Detecting Large-Scale COVID-19 Misinformation in Social Networks’, *Lect. Notes Comput. Sci.*, no. 12989, pp. 97–104, 2022, doi: 10.1007/978-3-030-96326-2_7.

[101] C. Maathuis and I. Kerkhof, ‘First Six Months of War from Ukrainian topic and sentiment analysis’, *Eur. Conf. Soc. Media*, vol. 10, no. 1, pp. 163–173, 2023, doi: 10.34190/ecsm.10.1.1147.

[102] R. Marigliano, L. Hui, and K. M. Carley, ‘Analyzing digital propaganda and conflict rhetoric: a study on Russia’s bot-driven campaigns and counter-narratives during the Ukraine crisis’, *Soc. Netw. Anal. Min.*, vol. 14, no. 1, 2024, doi: 10.1007/s13278-024-01322-w.

[103] K. Lipianina-Honcharenko, Y. Bodyanskiy, N. Kustra, and A. Ivasechko, ‘OLTW-TEC: online learning with sliding windows for text classifier ensembles’, *Front. Artif. Intell.*, vol. 7, 2024, doi: 10.3389/frai.2024.1401126.

[104] V. Schmitt *et al.*, ‘Evaluating Human-Centered AI Explanations: Introduction of an XAI Evaluation Framework for Fact-Checking’, pp. 91–100, 2024, doi: 10.1145/3643491.3660283.

[105] M. T. Ijab, M. S. Shahril, and S. Hamid, ‘Infodemiology Framework for COVID-19 and Future Pandemics Using Artificial Intelligence to Address Misinformation and Disinformation’, *Lect. Notes Comput. Sci.*, pp. 530–539, 2021, doi: 10.1007/978-3-030-90235-3_46.

[106] L. Ameli, M. Shah, F. Farid, A. Bello, F. Sabrina, and A. Maurushat, ‘AI and Fake News: a Conceptual Framework for Fake News Detection’, 2022. doi: 10.1145/3584714.3584722.

[107] A. Kumar and J. W. Taylor, ‘Feature importance in the age of explainable AI: Case study of detecting fake news & misinformation via a multi-modal framework’, *Eur. J. Oper. Res.*, vol. 317, no. 2, 2023, doi: 10.1016/j.ejor.2023.10.003.

[108] O. Ibrahim Aboulola and M. Umer, ‘Novel Approach for Arabic Fake News Classification Using Embedding from Large Language Features with CNN-LSTM Ensemble Model and Explainable AI’, *Sci. Rep.*, vol. 14, no. 1, 2024, doi: 10.1038/s41598-024-82111-5.

[109] M. Kumar, Praveen Kumar Rai, and P. Kumar, ‘A Novel Approach for Detecting Deepfake Face Using Machine Learning Algorithms’, 2024. doi: 10.1109/icdt61202.2024.10489036.

[110] W. Antoun, F. Baly, and H. Hajj, ‘AraBERT: Transformer-based Model for Arabic Language Understanding AraBERT: Transformer-based Model for Arabic Language Understanding’, 2020.

[111] I. Z. Hussain, J. Kaur, M. Lotto, Z. A. Butt, and P. P. Morita, ‘Tweeting for Health using Real-Time Mining and AI-Based Analytics: Design & Development of as Misinformation Data Ecosystem for Twitter (Preprint)’, *JMIR*, vol. 25, no. 44356, 2022, doi: 10.2196/44356.

[112] S. T. Smith, E. K. Kao, E. D. Mackin, D. C. Shah, O. Simek, and D. B. Rubin, ‘Automatic detection of influential actors in disinformation networks’, *Proc. Natl. Acad. Sci.*, vol. 118, no. 4, 2021, doi: 10.1073/pnas.2011216118.

[113] M. Sudhakar and K. P. Kaliyamurthie, ‘Fake News Detection Approach Based on Logistic Regression in Machine Learning’, *Lect. Notes Netw. Syst.*, pp. 55–60, 2023, doi: 10.1007/978-981-19-9304-6_6.

[114] M. Granik and V. Mesyura, ‘Fake news detection using naive Bayes classifier’, *2017 IEEE First Ukr. Conf. Electr. Comput. Eng. UKRCON*, 2017, doi: 10.1109/ukrcon.2017.8100379.

[115] A. Kesarwani, S. S. Chauhan, and A. R. Nair, ‘Fake News Detection on Social Media using K-Nearest Neighbor Classifier’, *2020 Int. Conf. Adv. Comput. Commun. Eng. ICACCE*, 2020, doi: 10.1109/icacce49060.2020.9154997.

[116] I. G. Hosea, V. O. Waziri, I. Idris, J. A. Ojeniyi, M. Olalere, and O. Adebayo, ‘A Machine Learning Approach to Fake News Detection Using Support Vector Machine (SVM) and Unsupervised Learning Model’, *Adv. Multidiscip. Sci. Res. J.*, vol. 11, no. 1, pp. 11–18, 2023, doi: 10.22624/aims/csean-smart2023p1.

[117] M. Anjali and G. Jivani, ‘A Comparative Study of Stemming Algorithms’, 2011. Accessed: Jan. 01, 2025. [Online]. Available: https://www.ccs.neu.edu/home/vip/teach/IRcourse/2_indexing_ngrams/lecture_notes/stemmer/0fa35d4ff8a2f925eb955e48d655494bd167.pdf

[118] C. Yang, X. Zhou, and R. Zafarani, ‘CHECKED: Chinese COVID-19 fake news dataset’, *Soc. Netw. Anal. Min.*, vol. 11, no. 1, 2021, doi: 10.1007/s13278-021-00766-8.

[119] Z. P. Agusta and A. Adiwijaya, ‘Modified balanced random forest for improving imbalanced data prediction’, *Int. J. Adv. Intell. Inform.*, vol. 5, no. 1, p. 58, 2018, doi: 10.26555/ijain.v5i1.255.

[120] T. Chen and C. Guestrin, ‘XGBoost: a Scalable Tree Boosting System’, *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. - KDD 16*, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.

[121] A. Natekin and A. Knoll, ‘Gradient boosting machines, a tutorial’, *Front. Neurorobotics*, vol. 7, no. 21, 2013, doi: 10.3389/fnbot.2013.00021.

[122] G. Ke *et al.*, ‘LightGBM: A Highly Efficient Gradient Boosting Decision Tree’, 2017. [Online]. Available:

https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

[123] Optuna, ‘Optuna - A hyperparameter optimization framework’, Optuna. [Online]. Available: https://optuna.org/

[124] P. K. Verma, P. Agrawal, and R. Prodan, ‘WELFake: Word Embedding Over Linguistic Features for Fake News Detection’, *IEEE Trans. Comput. Soc. Syst.*, vol. 8, no. 4, pp. 881–893, 2021, doi: 10.1109/tcss.2021.3068519.

[125] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ‘BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding’, *Proc. 2019 Conf. North*, vol. 1, 2019, doi: 10.18653/v1/n19-1423.

[126] K. Yadav, ‘Fake-Real News’, *www.kaggle.com*. [Online]. Available: https://www.kaggle.com/datasets/techykajal/fakereal-news

[127] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, ‘XLNet: Generalized Autoregressive Pretraining for Language Understanding’, arXiv.org. [Online]. Available: https://arxiv.org/abs/1906.08237

[128] S. Hochreiter and J. Schmidhuber, ‘Long Short-Term Memory’, *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[129] E. Levin, ‘A recurrent neural network: Limitations and training’, *Neural Netw.*, vol. 3, no. 6, pp. 641–650, 1990, doi: 10.1016/0893-6080(90)90054-o.

[130] A. Graves and J. Schmidhuber, ‘Framewise phoneme classification with bidirectional LSTM and other neural network architectures’, *Neural Netw.*, vol. 18, no. 5–6, pp. 602–610, 2005, doi: 10.1016/j.neunet.2005.06.042.

[131] P. Zhou *et al.*, ‘Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification’, ACLWeb. Accessed: Jan. 01, 2020. [Online]. Available: https://www.aclweb.org/anthology/P16-2034

[132] C.-W. Chen, S.-P. Tseng, T.-W. Kuan, and J.-F. Wang, ‘Outpatient Text Classification Using Attention-Based Bidirectional LSTM for Robot-Assisted Servicing in Hospital’, *Information*, vol. 11, no. 2, p. 106, 2020, doi: 10.3390/info11020106.

[133] J. A. Saenz, S. R. Kalathur Gopal, and D. Shukla, ‘Covid-19 Fake News Infodemic Research Dataset (CoVID19-FNIR Dataset)’. IEEE Dataport, Jul. 22, 2021. doi: 10.21227/b5bt-5244.

[134] N. Kausar, A. AliKhan, and M. Sattar, ‘Towards better representation learning using hybrid deep learning model for fake news detection’, *Soc. Netw. Anal. Min.*, vol. 12, no. 1, 2022, doi: 10.1007/s13278-022-00986-6.

[135] P. K. Verma, P. Agrawal, I. Amorim, and R. Prodan, ‘WELFake: Word Embedding Over Linguistic Features for Fake News Detection’, *IEEE Trans. Comput. Soc. Syst.*, vol. 8, no. 4, pp. 881–893, 2021, doi: 10.1109/tcss.2021.3068519.

[136] A. Gupta, A. Batla, C. Kumar, and G. Jain, ‘Comparative Analysis Of Machine Learning Models For Fake News Classification’, IEEE Xplore. [Online]. Available: https://ieeexplore.ieee.org/document/10205870

[137] V. S. Nirban, T. Shukla, P. S. Purkayastha, N. Kotalwar, and L. Ahsan, ‘The Role of AI in Combating Fake News and Misinformation’, *Lect. Notes Netw. Syst.*, no. 649, pp. 690–701, 2023, doi: 10.1007/978-3-031-27499-2_64.

[138] M. M. F. Caceres *et al.*, ‘The impact of misinformation on the COVID-19 pandemic’, *AIMS Public Health*, vol. 9, no. 2, pp. 262–277, 2022, doi: 10.3934/publichealth.2022018.

[139] W. Ahmed, J. Vidal-Alaball, J. Downing, and F. López Seguí, ‘COVID-19 and the 5G Conspiracy Theory: Social Network Analysis of Twitter Data’, *J. Med. Internet Res.*, vol. 22, no. 5, p. e19458, 2020, doi: 10.2196/19458.

[140] Heidi Larson, *Stuck : how vaccine rumors start - and why they don’t go away*. Oxford University Press, 2020.

[141] E. Dubé, C. Laberge, M. Guay, P. Bramadat, R. Roy, and J. A. Bettinger, ‘Vaccine Hesitancy’, *Hum. Vaccines Immunother.*, vol. 9, no. 8, pp. 1763–1773, 2013, doi: 10.4161/hv.24657.

[142] Sezer Kisa and Adnan Kisa, ‘A Comprehensive Analysis of COVID-19 Misinformation, Public Health Impacts, and Communication Strategies: Scoping Review’, *J. Med. Internet Res.*, vol. 26, no. e56931, 2024, doi: 10.2196/56931.

[143] P. N. Howard, R. Kleis Nielsen, R. Fletcher, N. Newman, and J. S. Brennen, ‘Navigating the “infodemic”: how people in six countries access and rate news and information about coronavirus’, Reuters Institute for the Study of Journalism. [Online]. Available: https://reutersinstitute.politics.ox.ac.uk/infodemic-how-people-six-countries-access-and-rate-news-and-information-about-coronavirus

[144] R. Vinay, B. Premjith, D. Shukla, and K. P. Soman, ‘Feature Engineering and Selection for the Identification of Fake News in Social Media’, *Lect. Notes Electr. Eng.*, no. 1026, pp. 291–301, 2023, doi: 10.1007/978-981-99-1410-4_24.

[145] M. Sikosana, O. Ajao, and S. Maudsley-Barton, ‘A Comparative Study of Hybrid Models in Health Misinformation Text Classification’, *ArXiv Cornell Univ.*, vol. 23, pp. 18–25, 2024, doi: 10.1145/3677117.3685007.

[146] M. Qadees and A. Hannan, ‘Cross comparison of COVID-19 fake news detection machine learning models’, *2023 17th Int. Conf. Open Source Syst. Technol. ICOSST*, pp. 1–7, 2023, doi: 10.1109/icosst60641.2023.10414227.

[147] Zepopo, ‘Ukrainian news’. [Online]. Available: https://www.kaggle.com/datasets/zepopo/ukrainian-fake-and-true-news/data?select=news_data.csv

[148] Skupriienko, ‘GitHub - skupriienko/Ukrainian-Stopwords: the list of ~2000 ukrainian stopwords (with numbers)’, GitHub. [Online]. Available: https://github.com/skupriienko/Ukrainian-Stopwords

[149] V. Karpukhin *et al.*, ‘Dense Passage Retrieval for Open-Domain Question Answering’, *Proc. 2020 Conf. Empir. Methods Nat. Lang. Process. EMNLP*, 2020, doi: 10.18653/v1/2020.emnlp-main.550.

[150] Sentence-transformers/all-MiniLM-L6-v2, huggingface.co. [Online]. Available: https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[151] EEAS, ‘Disinformation Cases’, EU vs DISINFORMATION. [Online]. Available: https://euvsdisinfo.eu/disinformation-cases/

ДОДАТОК А
СПИСОК ПУБЛІКАЦІЙ ЗДОБУВАЧА

1. H. Padalko, V. Chomko, S. Yakovlev, and D. Chumachenko, “A Novel Comprehensive Framework for Detecting and Understanding Health-Related Misinformation,” Information, vol. 16, no. 3, p. 175, Mar. 2025, doi: https://doi.org/10.3390/info16030175. (Scopus, Q2)

2. H. Padalko, V. Chomko, D. Chumachenko, “A novel approach to fake news classification using LSTM-based deep learning models,” Frontiers in big data, vol. 6, 2024, https://doi.org/10.3389/fdata.2023.1320800. (Scopus, Q2)

3. H. Padalko, V. Chomko, S. Yakovlev, D. Chumachenko, “Ensemble machine learning approaches for fake news classification,” Radioelectronic and computer systems, no. 4, pp. 5–19, 2023, https://doi.org/10.32620/reks.2023.4.01. (Scopus, Q3, Категорія А)

4. H. Padalko, V. Chomko, S. Yakovlev, P. P. Morita, “Classification of disinformation in hybrid warfare: an application of XLNet during the Russia’s war against Ukraine,” Radioelectronic and computer systems, vol. 2024, no. 4, pp. 46–58, 2024, https://doi.org/10.32620/reks.2024.4.04. (Scopus, Q3, Категорія А)

5. M. Lotto, Th. Hanjahanja-Phiri, H. Padalko, A. Oetomo, Z.A. Butt, J. Boger, J. Millar, Th. Cruvinel, P.P. Morita “Ethical principles for infodemiology and infoveillance studies concerning infodemic management on social media,” Frontiers in Public Health, vol. 11, 2023. , https://doi.org/10.3389/fpubh.2023.1130079 (Scopus, Q1)

6. A. Fitz-Gerald, D. Chumachenko, H. Padalko, V. Ganesh, “Artificial Intelligence and New Threat Vectors: Using Scenario Planning for Trend Forecasting” Balsillie Papers, vol. 5, no. 6, 2023, https://doi.org/10.51644/bap56.

7. H. Padalko, V. Chomko, D. Chumachenko, “Misinformation Detection in Political News using BERT Model,” CEUR Workshop Proceedings, vol 3641, Proceedings of the 3rd International Workshop of IT-professionals on Artificial

Intelligence (ProfIT AI 2023) 2023 (Вотерлу, Канада, 20-22 листопада 2023), 2023, pp. 117–127. https://doi.org/10.5281/zenodo.15014455 (Scopus)

8. H. Padalko, V. Chomko, D. Chumachenko, “The Impact of Stopwords Removal on Disinformation Detection in Ukrainian language during Russian-Ukrainian war,” CEUR Workshop Proceedings, vol. 3777, Proceedings of the 4th International Workshop of IT-professionals on Artificial Intelligence (ProfIT AI 2024) 2024 (Кембрідж, США, 25-27 вересня 2024), 2024, pp. 87–101. . https://doi.org/10.5281/zenodo.15014498 (Scopus)

9. H. Padalko, V. Chomko, D. Chumachenko, “COVID-19 Misinformation Detection on Weibo by Support Vector Classifier,” IEEE International Conference on Dependable Systems, Services and Technologies (DESSERT) (Афіни, Греція, 13-15 жовтня 2023), 197136, 2023, https://doi.org/10.1109/dessert61349.2023.10416460 (Scopus)

10. H. Padalko, P. Morita, “Covid-19 ‘infodemic’ management: russia’s key tool for influencing geopolitics,” Population Medicine, vol. 5 (Supplement):A1809, 17th World Congress on Public Health (Рим, Італія, 2-6 травня 2023), 2023, p. 512. doi: https://doi.org/10.18332/popmed/164626.

11. H. Padalko “Machine learning approach for classification of russian propaganda”, Матеріали V Міжнародній науково-практичній конференції IT-професіоналів та аналітиків комп’ютерних систем “ProfIT Conference” (Харків, 28-30 червня 2023), 2023, С. 96-97.

12. Г.А. Падалко, С.В. Яковлев “Модель розповсюдження COVID-19 на основі аналізу соціальних мереж”, Матеріали IV Міжнародній науково-практичній конференції IT-професіоналів та аналітиків комп’ютерних систем “ProfIT Conference” (Харків, 13-15 грудня 2021), 2021, С. 106.

13. Г.А. Падалко “Мультиагентний підхід як інструмент моделювання соціальних мереж”, Матеріали III Міжнародній науково-практичній конференції IT-професіоналів та аналітиків комп’ютерних систем “ProfIT Conference” (Харків, 8-10 грудня 2020), 2020, С. 112.

14. R. Vysochanskyy, H. Padalko, A. Ollila, “Information Warfare in Canada: Countering AI-backed Misinformation, Disinformation and Malinformation (MDM) Surrounding the Russian Invasion of Ukraine,” Defence and Security Foresight Group. AI, Ethics, and Warfare Series, 2023, 10 p. [Online]. Available: https://uwaterloo.ca/defence-security-foresight-group/sites/default/files/uploads/documents/vysochanskyy-et-al_information-warfare-in-canada.pdf

15. A. Fitz-Gerald, H. Padalko, “The Need for a Strategic Approach to Disinformation and AI-Driven Threats,” The Royal United Services Institute (RUSI), 2024. [Online]. https://www.rusi.org/explore-our-research/publications/commentary/need-strategic-approach-disinformation-and-ai-driven-threats

16. H. Padalko, “Russian Disinformation about the US Election: AI Analysis of Narratives,” Center for International Governance Innovation, 2025. [Online]. Available: https://www.cigionline.org/static/documents/DPH-paper-Padalko.pdf

ДОДАТОК Б
ВІДОМОСТІ ПРО АПРОБАЦІЮ РЕЗУЛЬТАТІВ ДИСЕРТАЦІЇ

1. 4th International Workshop of IT-professionals on Artificial Intelligence (ProfIT AI 2024) 2024 (Кембрідж, США, 25-27 вересня 2024) – дистанційна участь.

2. 3rd International Workshop of IT-professionals on Artificial Intelligence (ProfIT AI 2023) 2023 (Вотерлу, Канада, 20-22 листопада 2023) – дистанційна участь.

3. IEEE International Conference on Dependable Systems, Services and Technologies (DESSERT) (Афіни, Греція, 13-15 жовтня 2023) – дистанційна участь.

4. V Міжнародна науково-практична конференція ІТ-професіоналів та аналітиків комп'ютерних систем "ProfIT Conference" (Харків, 28-30 червня 2023), 2023, С. 96-97. – дистанційна участь.

5. 17th World Congress on Public Health (Рим, Італія, 2-6 травня 2023) – очна участь.

6. IV Міжнародна науково-практична конференція ІТ-професіоналів та аналітиків комп'ютерних систем "ProfIT Conference" (Харків, 13-15 грудня 2021), 2021, С. 106. – дистанційна участь.

7. III Міжнародна науково-практична конференція ІТ-професіоналів та аналітиків комп'ютерних систем "ProfIT Conference" (Харків, 8-10 грудня 2020), 2020, С. 112. – дистанційна участь.

ДОДАТОК В

АКТИ ВПРОВАДЖЕННЯ РЕЗУЛЬТАТІВ ДОСЛІДЖЕНЬ

BALSILLIE SCHOOL
OF INTERNATIONAL AFFAIRS
67 Erb Street West, Waterloo, Ontario Canada N2L 6C2

March 18, 2025

Subject: Implementation Act of the results of dissertation research for the PhD degree of Halyna Padalko into research & development work of Balsillie School of International Affairs

The undersigned confirm that the findings of Halyna Padalko's doctoral research have been effectively incorporated into the research analysis conducted by the Balsillie School of International Affairs.

Specifically, the framework for in-depth analysis of textual disinformation and propaganda, which integrates Bidirectional Encoder Representations from Transformers (BERT) and Bidirectional Long Short-Term Memory (Bi-LSTM) to generate interpretable vector representations of text, was utilized to examine Russian election-related disinformation in the EU. The insights derived from this analysis contributed to the conference paper titled "Doppelganger Tactics: Analyzing Russian Information Operations in the Context of EU Elections". By implementing this framework, a narrative analysis of Russian disinformation surrounding the EU elections was conducted.

Yours sincerely,

Dr. Ann Fitz-Gerald
Director, Balsillie School of International Affairs
✉ afitz-gerald@balsillieschool.ca

CIGI | Digital Policy Hub

Implementation Act

of the results of dissertation research for the PhD degree

Of Halyna Padalko into research work of Digital Policy Hub

Centre for International Governance Innovation

The undersigned confirm that the results of Halyna Padalko's doctoral research have been successfully integrated into the research analysis conducted by the Center for International Governance Innovation's Digital Policy Hub.

In particular, the framework for comprehensive analysis of textual disinformation and propaganda, based on the combination of Bidirectional Encoder Representations from Transformers (BERT) and Bidirectional Long Short-Term Memory (Bi-LSTM) for generating interpretable vector representations of text, was employed to analyze election-related Russian disinformation in the United States. The findings contributed to the policy brief titled "Russian Disinformation about the US Election: AI Analysis of Narratives."

The implementation of this framework enabled a narrative analysis of Russian disinformation concerning the U.S. elections, facilitating the development of policy recommendations to enhance Canada's resilience in the upcoming 2025 federal elections.

Director, Program Management

Centre for International Governance Innovation

Dianna H. English

CIGI acknowledges that we live and work on the traditional territory of the Neutral, Anishinaabeg and Haudenosaunee peoples. CIGI is situated on the Haldimand Tract, the land promised to the Six Nations that includes 10 km on each side of the Grand River.

ЗАТВЕРДЖУЮ:
Генеральний директор
державної установи «Харківський
обласний центр контролю та
профілактики хвороб Міністерства
охорони здоров'я України»

Любов МАХОТА
«24» грудня 2024 р.

АКТ

впровадження результатів наукових досліджень дисертаційної роботи
Падалко Галини Анатоліївни
виконаної на здобуття наукового ступеня доктор філософії

Комісія у складі голови – Карлової Т. О., завідувача відділення організації епідеміологічних досліджень, секретар комісії – Лешко О. П., лікар-епідеміолог, склала даний акт про впровадження наукових результатів дисертаційної роботи Падалко Г.А. в практичну діяльність відділу комунікації та інформаційно-роз'яснювальної роботи державної установи «Харківський обласний центр контролю та профілактики хвороб Міністерства охорони здоров'я України», а саме результати комплексного аналізу місінформації, пов'язаної з захворюваністю на COVID-19, отримані за допомогою фреймворку комплексного аналізу текстової дезінформації та пропаганди, заснованому на комбінуванні технологій двоспрямованого кодувального представлення з трансформерів (BERT) та двоспрямованої довгої короткочасної пам'яті (Bi-LSTM) для створення інтерпретованих векторних представлень тексту.

Використання цих результатів дозволило підвищити ефективність комунікаційної та інформаційно-роз'яснювальної роботи щодо захворюваності на COVID-19.

Голова комісії Т. О. Карлова

Секретар комісії О. П. Лешко

Затверджую
Проректор з науково-педагогічної роботи
Національного аерокосмічного університету
«Харківський авіаційний інститут»
к.т.н., доцент
_____ Андрій ГУМЕННИЙ
2025

АКТ ВПРОВАДЖЕННЯ
результатів дисертаційної роботи,
виконаної на здобуття наукового ступеня доктора філософії
Падалко Галини Анатоліївни, у навчальному процесі
кафедри математичного моделювання та штучного інтелекту

Комісія у складі: голови комісії – декана факультету систем управління літальними апаратами д.т.н., проф. Заболотного О.В., членів комісії – в.о. завідувача кафедри математичного моделювання та штучного інтелекту к.ф.-м.н., доц. Карташова О.В, професора кафедри математичного моделювання та штучного інтелекту д.т.н., проф. Скоба Ю.О., встановила, що наукові результати, а саме:

1. Модель класифікації текстової дезінформації та пропаганди, заснований на поєднанні трансформерів і архітектури Bi-LSTM для інтеграції лінійних та нелінійних залежностей текстових даних.

2. Моделі глибокого навчання для класифікації текстової дезінформації та пропаганди, засновані на архітектурах XLNet, Bi-LSTM та Attention-Based Bi-LSTM.

3. Метод тематичного моделювання, заснований на застосуванні механізмів багатоголової уваги, що на відміну від існуючих використовує глибокі ембединги для контекстуалізації даних.

Ці доробки були реалізовані у навчальному процесі кафедри математичного моделювання та штучного інтелекту у вигляді методичних та програмних засобів в рамках курсів «Інтелектуальні системи», «Обробка зображень та мультимедіа» для спеціальностей 113 Прикладна математика та 122 Комп'ютерні науки першого (бакалаврського) рівня вищої освіти, «Теорія та методи оптимізації складних систем», «Інтелектуальні програмні комплекси» для спеціальностей 113 Прикладна математика та 122 Комп'ютерні науки другого (магістерського) рівня вищої освіти.

За результатами дисертаційної роботи було розроблено теоретичне, методичне та програмне забезпечення.

Голова комісії Олександр ЗАБОЛОТНИЙ

Члени комісії Олексій КАРТАШОВ

Юрій СКОБ

Затверджую
Проректор з наукової роботи
Національного аерокосмічного університету
«Харківський авіаційний інститут»
д.т.н., професор
_____ Володимир ПАВЛІКОВ
_____ 2025

АКТ ВПРОВАДЖЕННЯ
результатів дисертаційної роботи
на здобуття наукового ступеня доктора філософії
Падалко Галини Анатоліївни у науково-дослідницьку роботу
Національного аерокосмічного університету «Харківський авіаційний інститут»

Комісія у складі: голови комісії – декана факультету систем управління літальними апаратами д.т.н., проф. Заболотного О.В., членів комісії – в.о. завідувача кафедри математичного моделювання та штучного інтелекту к.ф.-м.н., доц. Карташова О.В, професора кафедри математичного моделювання та штучного інтелекту д.т.н., проф. Скоба Ю.О., встановила, що наукові результати, а саме:

1. Фреймворк комплексного аналізу текстової дезінформації та пропаганди, заснований на комбінуванні технологій двоспрямованого кодувального представлення з трансформерів (BERT) та двоспрямованої довгої короткочасної пам'яті (Bi-LSTM).

2. Модель класифікації текстової дезінформації та пропаганди, заснований на поєднанні трансформерів і архітектури Bi-LSTM.

3. Моделі глибокого навчання для класифікації текстової дезінформації та пропаганди, засновані на архітектурах XLNet, Bi-LSTM та Attention-Based Bi-LSTM.

4. Метод тематичного моделювання, заснований на застосуванні механізмів багатоголової уваги.

Ці доробки були реалізовані у вигляді розробок, використаних при виконанні науково-дослідних робіт кафедри математичного моделювання та штучного інтелекту.

Це дозволило збільшити точність класифікації виявлення текстової дезінформації та пропаганди у соціальних мережах. Використання розробленого фреймворку забезпечило аналіз наративів для побудови стратегій контрзаходів та стратегічних комунікацій для аналізу дезінформації про вибори, COVID-19 та російську війну проти України, що забезпечило отримання цінного розуміння наративів у великих корпусах даних.

Голова комісії Олександр ЗАБОЛОТНИЙ

Члени комісії Олексій КАРТАШОВ

Юрій СКОБ

ДОДАТОК Г

Г.1. Набір даних 1 (Російська пропаганда у X (Twitter))

Процес збору твітів включав сценарій вилучення даних на основі Python, який в першу чергу залежав від двох файлів: twitter-keys.txt, що містить маркер носія для авторизації, і twitter_meta.txt, що містить інформацію про кожен місячний часовий проміжок для запиту API. Спеціальний доступ до API Twitter V2 надається науковцям-дослідникам, що дає їм дослідницький рівень доступу до загальнодоступних даних у реальному часі та історичних даних, додаткових функцій і функцій для збору точних, повних і неупереджених наборів даних. Необроблені результати були збережені як файли JSON і завантажені в призначений контейнер для зберігання двійкових великих об’єктів Azure. Згодом необроблені дані, що зберігаються в репозиторії, попередньо обробляються відповідно до вимог користувача. Було обчислено основний показник ефективності, який призначав вагу загальнодоступним показникам, таким як номер, кількість відповідей, кількість цитувань і кількість ретвітів, дозволяючи фільтрувати повторювані твіти, зберігаючи твіти з найвищим показником ефективності. Після перевірки було створено набір даних із “N” екземплярів найефективніших твітів, які інформували про навчання моделей класифікації пропаганди.

Twitter API використовувався для збору 42 тисяч твітів англійською мовою з вересня 2022 року. Цей конкретний місяць було обрано через активність на полі бою та значні гуманітарні кризи, які стимулювали збільшення спроб Росії заповнити інформаційну сферу дезінформацією та проросійськими наративами. Знаковими подіями вересня 2022 року стали контрнаступ України, який призвів до звільнення величезних територій, виявлення масових поховань в Ізюмі із 447 тілами українців, російський референдум у тимчасово окупованих регіонах України, вибух на газопроводі Nordstream, вбивство російської пропагандистки

Дар'ї Дугіної (ключової фігури в ідеологіях російської пропаганди), а також великий обмін полоненими. У цьому обміні Україна обміняла проросійського олігарха Віктора Медведчука, відомого своїми близькими зв'язками з Путіним, на командирів батальйону "Азов" та інших українських військових.

Прагнучи проаналізувати статті, звіти, дослідження та журналістські розслідування в ЗМІ, було сформульовано комплексну стратегію пошуку Таблиця Г.1., щоб охопити проросійські наративи та відфільтрувати твіти, які можуть демонструвати проукраїнські настрої. Щоб ефективно фіксувати проросійські наративи, було застосовано комбінацію цільових ключових слів і операторів розширеного пошуку.

Збір даних із Twitter дозволив отримати 42 тис. твітів, 5 тис. з яких були вручну розмічені та розподілені на 2 класи: 1 – проросійська пропаганда та 2 – нейтральні твіти та твіти, що виражають проукраїнську позицію. Для уникнення упередженості, у процесі розмітки даних було задіяно 2 дослідника, які розподілили 5 тисяч твітів на 2 групи, порівнюючи результати.

Графіки хмар слів для проукраїнського та проросійського класів у цьому аналізі є не дуже допоміжними у аналізі, оскільки найпопулярніші слова в обох класах відносяться до спільних тем, таких як Україна, росія, війна, НАТО. Однак можна помітити, що в лівому класі є багато наборів слів, які більш очевидно можна віднести до проросійського класу, наприклад, "proxy war", "nazi" тощо. Хмара слів для набору даних російської дезінформації у твітері представлена на рисунку Г.1.

Аналізуючи найчастіші хештеги (рисунок 5), можна побачити, що серед найбільш згадуваних хештегів у цьому наборі даних є #Russia та #Ukraine. Друге місце за популярністю займає хештег #NATO, що свідчить про його значущість у дискурсі. Також серед найпоширеніших хештегів присутні #nazi та #azov, що є складовими популярного російського наративу про "нацистів в Україні", які нібито захопили владу, а також батальйон "Азов", який на початку свого існування складався з ультраправих футбольних фанатів, але був реформований, очищений від радикальних елементів і інтегрований до складу Збройних сил України у 2016 році. Також очікувано популярними є хештеги #EU, #USA, #China, що вказує на

глобальних партнерів та впливових гравців, які так чи інакше впливають або можуть впливати на хід подій на полі бою. Серед помітних хештегів можна помітити назви українських міст #Kherson #Donetsk #Odessa. Ці міста знаходяться у прифронтових або окупованих зонах. Популярність хештега #Kherson пояснюється контрнаступом українських військ у вересні 2022 року, значними успіхами армії на правобережжі Херсонщини та обговореннями щодо деокупації обласного центру. Зрештою, місто Херсон було звільнене силами ССО 11 листопада 2022 року. Рисунок 2 ілюструє аналіз найчастіших хештегів.

Таблиця Г.1. Пошукова стратегія

Included hashtags	Included keywords	Excluded hashtags
#nazi #ukrainenazis #ukronazi #Ukronazist #denazification #azov #DenazifyUkraine #zelenzkywarcrime #ZelenskyyWarCrimin al #Ukrainianwarcrime #Natowarcrime #NATORUSSIAWAR #stopsentweapontoUkr aine #Crimeaisrussia #ministryoffakes #istandwithPutin #istandwithrussia #fuckukraine #russiawillwin #glorytorussia #slavarossiya #longliveputin #Specialmilitaryoperati on #SpecialOperations #fukraine #NATORUSSIAWAR #NATO #proxywar #nonato #abolishNATO #stopnato #NATOisLosingTheW ar #SayNoToNATO	‘special military operation’; ‘special operation’; ‘bio laboratories’ and (‘US’ or ‘NATO’ or ‘USA’ or ‘Ukraine’); ‘fake’ and (‘graves’ or ‘Izyum’or ‘Izium’); ‘denazification’; ‘Azov’; ‘hohly’; ‘khohly’; khokhly’; ‘People of Donbas’; ‘8 years’ and ‘Donbas’; ‘Kyiv regime’; ‘Kiev regime’; ‘Zelenskiy regime’; ‘West hegemony’ and (‘Russia’ or ‘Ukraine’); ‘Westerners’ and (‘Ukraine’ or ‘Russia’); ‘Western regime’ and (‘Russia’ or ‘Ukraine’); ‘Palestine’ and ‘Ukraine’ and Russia; ‘Iraq’ and ‘Ukraine’ and (‘West’ or ‘Wester regime’ or NATO); ‘Global South’ and ‘West’ and (‘Ukraine’or ‘Russia’); ‘Racism’ and ‘Ukraine’ and ‘support’; ‘NATO threat’; ‘Bandera’; ‘NATO proxy war’; ‘US proxy war’; ‘Great Patriotic War’; ‘Russian Red lines’; ‘collective West’;	#RussiaIsATerroristState #RussiaIsANaziState #RussiaInvadedUkraine #RussiaIsATerroristState #RussiaIsANaziState #RussianWarCrimesInUkraine #RussiaTerrorist #GenocideOfUkrainians #RussianWarCrimes #Ukrainewillwin #StopRussia #NaziRussia #Russland #Terrorismus #RussiaTerroristState #RussiaIsATerroristState #RussiaIsANaziState #GenocideOfUkrainians #RussianWarCrimes #RussiaTerrorist #RussianAggression #StopPutin #SlavaUkraini #PedoputinUnion #UkraineUnderAttack #RussiaIsATerroristState #StopPutinNOW #RemovePutin #PutinWarCrimes #RussianNazis #RussianChildKillers #RusoThieves #ArmUkraineNow #StopPutinNOW

AND ‘Russia’ OR ‘Ukraine’	‘RUSSIA IS WARNING’; ‘NATO War Russia’.	#visabanforrussians #fakereferendum

Рисунок Г.1 – Хмари слів для набору даних про-російської дезінформації (зліва) та про-українського меседжингу (справа)

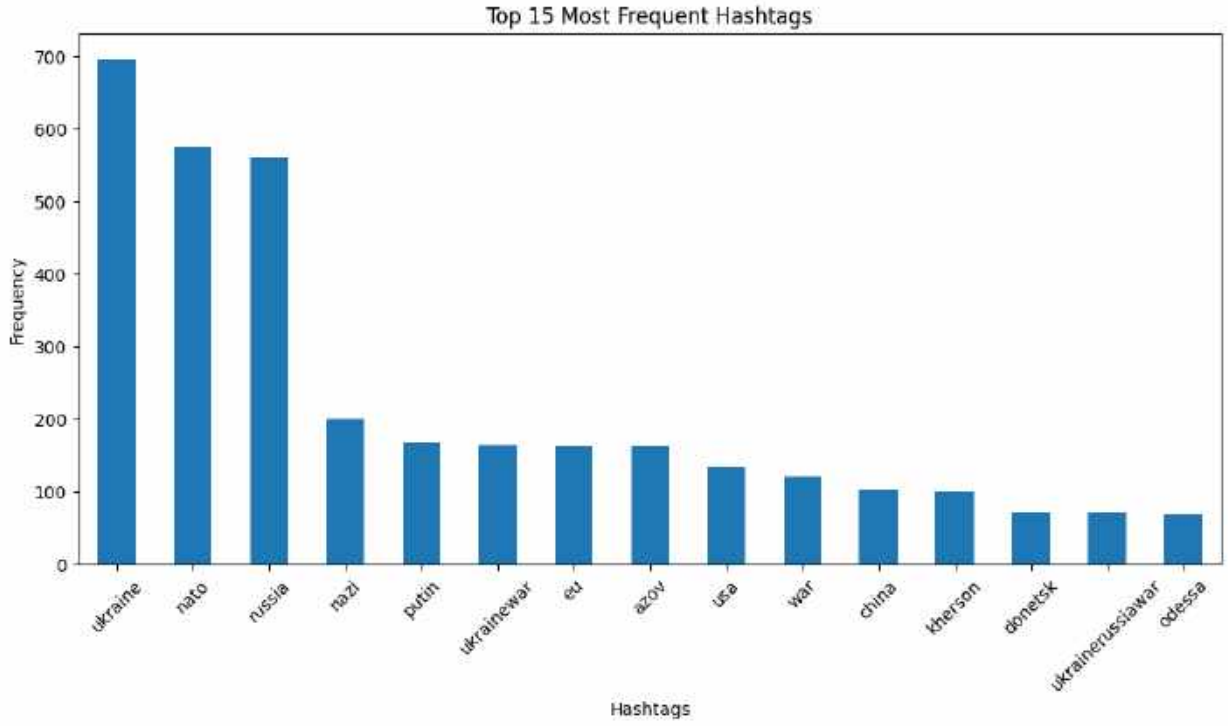

Рисунок Г.2. – Аналіз найпопулярніших хештегів набору даних

Аналіз 15 найпопулярніших слів (рисунок 3) незначно відрізняється від топ-15 хештегів. Помітно, що слово “war” зустрічається значно частіше у тілі твітів (>1000 разів), ніж у вигляді хештега (<200 разів). Це пояснюється тим, що слово “війна” вживається для опису самого процесу, який є постійним і не має яскраво вираженої поляризації.

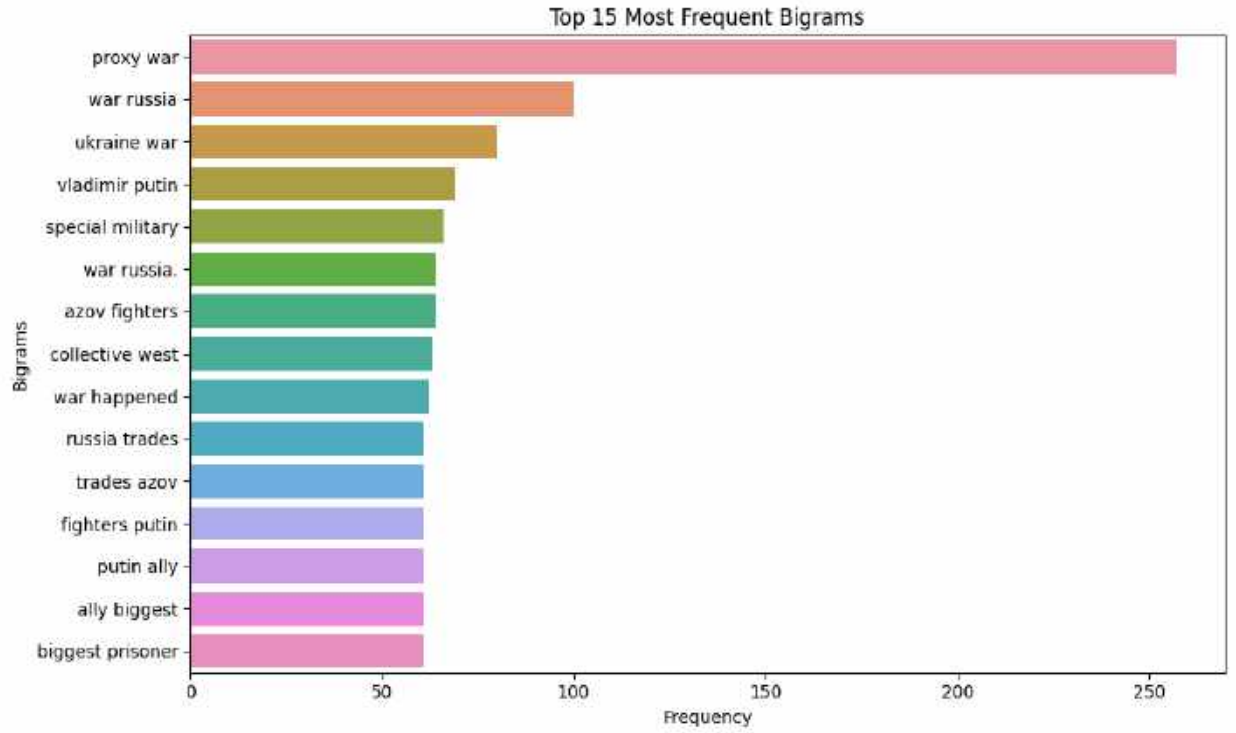

Рисунок Г.3. – Аналіз найчастіших слів.

Результати аналізу біграм (рисунок Г.4.) для класу з проросійськими настроями є цікавими. Біграмний аналіз розглядає найпоширеніші комбінації слів у наборі даних. У нашому випадку помітно значне використання біграми “proxy war”, що відповідає одному з ключових наративів російської пропаганди:

“Російсько-українська війна – це проксі-війна Америки проти Росії”. Це спостереження відповідає характеристикам сучасної російської пропаганди: швидкий та безперервний інформаційний потік, постійне повторення наративів, відсутність прихильності до логічної послідовності

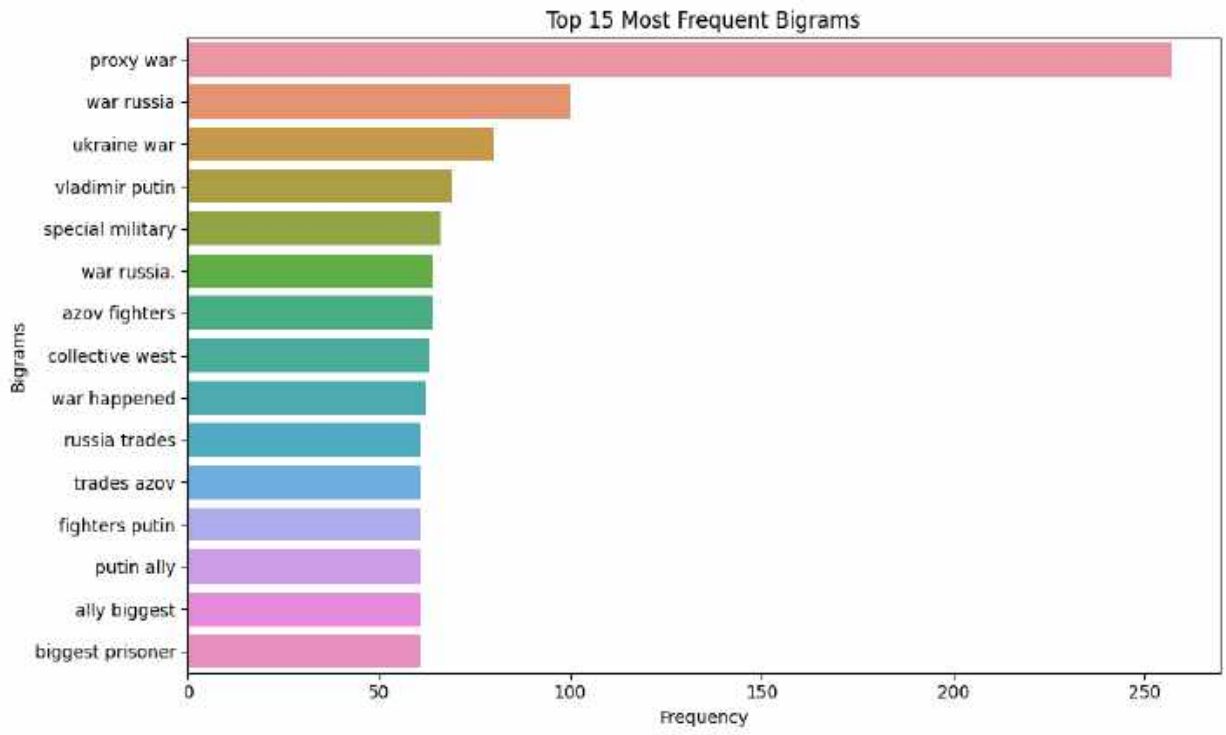

Рисунок Г.4. – Аналіз найчастіших біграм у класі проросійської пропаганди.

Г.2. Набір даних 2 (CHEKED)

Набір даних складається з двох основних категорій мікроблогів: фейкові новини та реальні новини, кожна з яких містить кілька записів. Кожен запис у наборі даних включає різні компоненти, такі як ID мікроблогу, мітка (реальна або фейкова новина), дата публікації, ID користувача, текстовий і візуальний контент мікроблогу, кількість коментарів, репостів і лайків, детальну інформацію про коментарі та репости користувачів, а також експертний аналіз і обґрунтування мітки (лише для фейкових мікроблогів). Крім записів мікроблогів, набір даних включає список ключових слів, які використовуються для визначення актуальності мікроблогу щодо COVID-19.

Набір даних CHECKED є цінним ресурсом для дослідників, які працюють над виявленням дезінформації, зокрема у китайському мовному та контекстуальному середовищі COVID-19. Він надає надійну та конфіденційну колекцію реальних і фейкових мікроблогів, експертний аналіз, еталонні результати та додаткові ресурси для полегшення подальших досліджень у цій важливій сфері.

Г.3. Набір даних 3 (WELFake)

Набір даних WELFake – це багата компіляція з 72,134 новинних статей, серед яких 35,028 є справжніми, а 37,106 класифіковані як фейкові. Ця всеосяжна колекція була створена шляхом об'єднання чотирьох відомих наборів новин: Kaggle, McIntire, Reuters і BuzzFeed Political. Ці набори даних були інтегровані, щоб знизити ризик перенавчання класифікатора та забезпечити ширшу текстову базу для більш ефективного навчання моделей машинного навчання. Кожен запис у наборі даних має унікальний ідентифікатор 0 або 1, що відповідає за правдивість контенту. Крім того, кожна стаття характеризується заголовком, детальним вмістом та міткою, яка вказує на її достовірність.

Основні спостереження щодо набору даних:

- Новинні статті, що містять від 450 до 550 слів, мають більшу надійність.
- Загальні тенденції вказують, що коротші, але змістовні новини частіше є правдивими.
- Фейкові новини мають нижчий рівень читабельності порівняно з реальними новинами.
- Фейкові новини мають вищий рівень суб'єктивності, ніж реальні.
- Кількість статей, що представляють реальні новини, є більшою, ніж кількість фейкових статей.

Ці спостереження надають цінну інформацію про особливості реальних та фейкових новин, що може бути корисним для розробки та вдосконалення алгоритмів виявлення фейків.

Рисунок Г.5. демонструє хмару слів, що відображає найчастотніші слова в наборі даних. Рисунок Г.6. показує збалансований розподіл фейкових та реальних новин у наборі WELFake. Рисунок Г.7. ілюструє розподіл по довжинах новин.

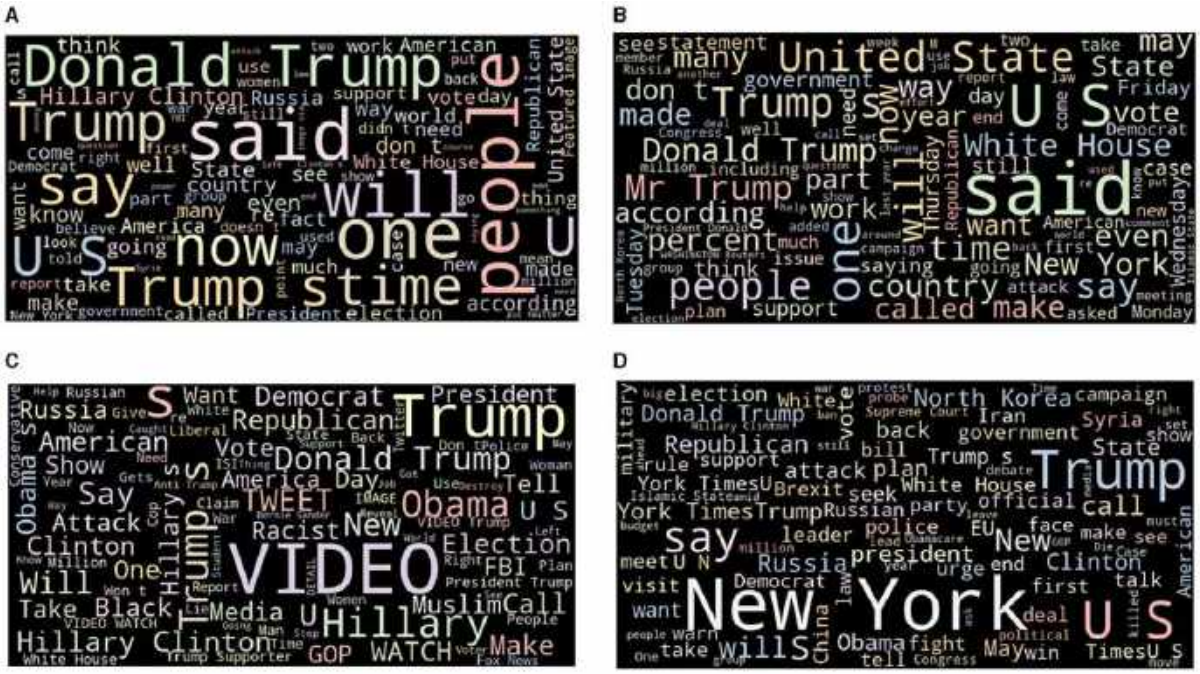

Рисунок Г.5 – Хмари слів набору даних:(A) Фейкові новини; (B) Правдиві новини; (C) Фейкові заголовки; (D) Правдиві заголовки.

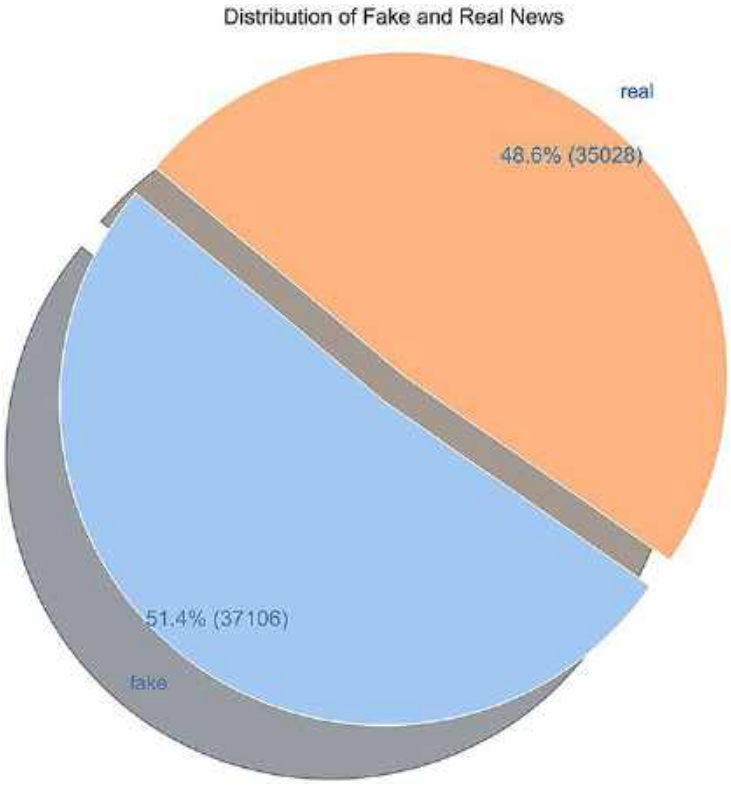

Рисунок Г.6. – Розподіл міток у наборі даних

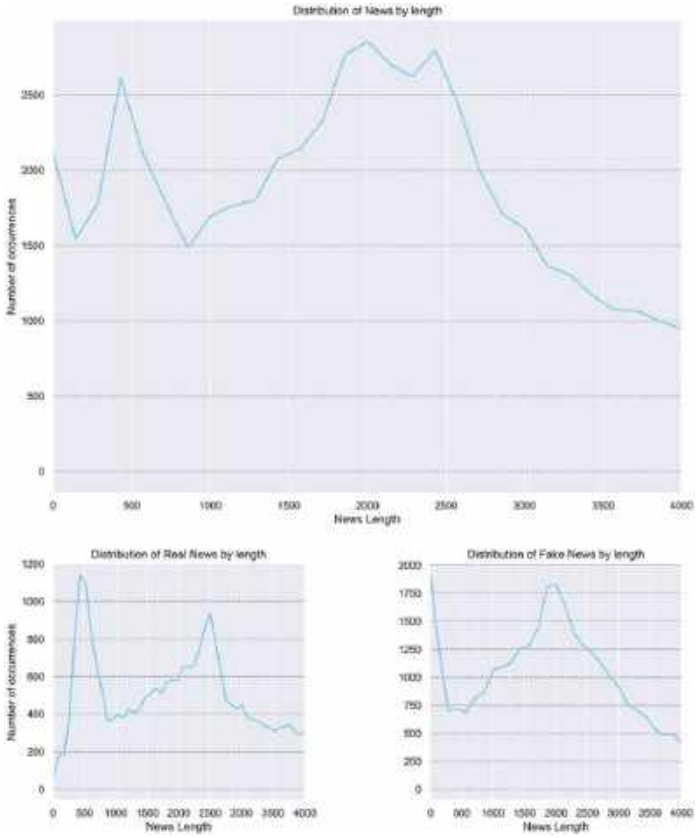

Рисунок Г.7. – Розподіл по довжині новин

Розуміння лінгвістичних нюансів і моделей, притаманних оманливим наративам, є важливим завданням у дослідженні фейкових новин. Біграми та триграми (послідовності двох і трьох слів відповідно) дозволяють отримати детальніший погляд на синтаксичні та семантичні структури, які часто використовуються в справжніх і сфабрикованих новинах.

Аналіз цих послідовностей дозволяє виявити повторювані фразові закономірності, що можуть свідчити про правдивість або фейковість новини. Реальні новини зазвичай дотримуються журналістських стандартів і певного стилю, що відображається у специфічних словосполученнях. Натомість фейкові новини можуть містити характерні повторювані патерни, спричинені сенсаційністю чи іншими маніпулятивними намірами.

Порівняння баграм і триграм обох категорій дає змогу отримати глибший мовний аналіз, що сприяє створенню більш точних та надійних моделей класифікації фейкових новин. Такий підхід не лише покращує ефективність моделі, а й допомагає пояснити її прогнози з лінгвістичної точки

зору, створюючи міст між обчислювальними методами та реальними мовними особливостями. Рисунок Г.8. демонструє найчастіші біграми та триграми у наборі даних.

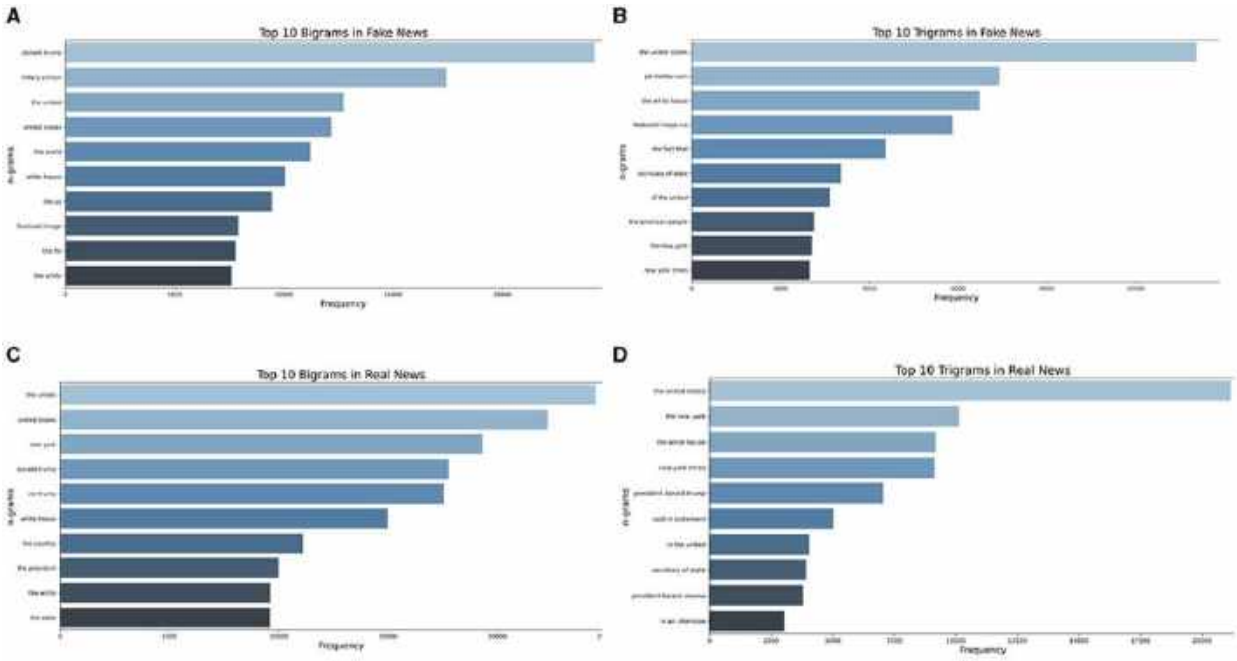

Рисунок Г.8. – Аналіз набору даних: (A) біграми фейкових новин, (B) триграми фейкових новин, (C) біграми правдивих новин, (D) триграми правдивих новин

Набір даних був очищений для усунення потенційних джерел шуму, таких як HTML-теги та URL-адреси, залишивши лише основну текстову інформацію. Крім того, ми забезпечили однорідність тексту, перетворивши весь вміст на малі літери.

Між процесами очищення даних та векторизації важливо зазначити, що набір даних був ретельно підготовлений, щоб забезпечити збалансований розподіл між справжніми та фейковими новинними статтями. Така рівновага має велике значення для завдань машинного навчання, оскільки вона запобігає упередженості моделі та гарантує, що класифікатор не буде схильний до будь-якого конкретного класу, тим самим підвищуючи загальність і надійність подальшого аналізу.

Для векторизації було обрано метод частоти термінів – оберненої частоти документа (TF-IDF). На відміну від простої частоти термінів, TF-IDF кількісно визначає важливість слова в документі відносно його частоти в усіх документах. Це допомагає підкреслити слова, які є більш унікальними для конкретного документа, що полегшує розрізнення між справжніми та фейковими новинами. Перетворення було виконано за допомогою TfidfVectorizer з бібліотеки scikit-learn, виключаючи англійські стоп-слова для зменшення розмірності та видалення поширених слів, які не мають значного внеску у диференціацію контенту. Зокрема, мітка "0" позначає фейкову новину, а "1" – справжню. Загалом, завдяки своєму різноманітному та збалансованому складу, набір даних WELFake є цінним ресурсом для тих, хто досліджує сферу виявлення фейкових новин.

Г.4. Набір даних 4 (Fake-Real News)

“Fake-Real News” набір даних є важливим ресурсом для досліджень у сфері виявлення дезінформації, спеціально розробленим для тренування та оцінювання моделей машинного навчання у розрізненні справжніх і фейкових новин. Він є незамінним інструментом для практиків та дослідників в області обробки природної мови, з акцентом на виявлення фейкових новин.

Структурований у табличному форматі, набір даних складається з двох окремих файлів: один містить фейкові новини, а інший – справжні. Кожен файл ретельно організований і включає колонки, що представляють різні атрибути новинних статей, такі як заголовок, текст і тема. Колонка заголовку охоплює заголовок кожної статті, що є критично важливим аспектом, оскільки заголовки часто створюються для того, щоб привернути увагу і можуть містити елементи сенсаційних меседжів. Колонка тексту надає повний зміст кожної статті, що дозволяє проводити комплексний лінгвістичний аналіз для оцінки контексту, стилю та детального наративу, які є ключовими для визначення достовірності статті. Додатково, колонка “тема” класифікує

статтю за різними змістовними галузями, такими як політика або світові події, що дає змогу оцінити потенційний вплив тематики на достовірність новин.

Набір даних вирізняється своїм об’ємним зібранням статей, що забезпечує багатий і різноманітний матеріал для аналізу. Це різноманіття тем підвищує універсальність набору даних, роблячи його застосовним у різних новинних галузях. Основне призначення цього ресурсу – допомога у розробці та тестуванні моделей машинного навчання для виявлення фейкових новин. Він підтримує широкий спектр аналітичних підходів, від класифікації тексту до аналізу настроїв і розпізнавання лінгвістичних патернів.

Втім, користувачі цього набору даних повинні бути уважними до його джерела та часових рамок новинних статей, щоб підтримувати актуальність моделі та її адаптивність до сучасних тенденцій у виробництві новин. Також важливо перевірити набір даних на наявність можливих упереджень або дисбалансу, що може вплинути на ефективність і узагальнюваність моделей, розроблених за його допомогою.

Г.5. Набір даних 5 (CoVID19-FNIR)

Набір даних CoVID19-FNIR є спеціалізованим ресурсом, створеним для сприяння дослідженням із виявлення фейкових новин, пов’язаних із пандемією COVID-19. Він містить перевірені новинні статті з різних регіонів, включаючи Індію, США та Європу, зібрані у період з лютого по червень 2020 року.

Набір даних включає два основні файли: fakeNews.csv і trueNews.csv, які містять відповідно фейкові та достовірні новини. Фейкові новини отримані з фактчекінгового сайту Poynter та містять детальні пояснення, джерела та мітки. З іншого боку, достовірні новини походять із перевірених Twitter-акаунтів авторитетних видавців, таких як The New York Times, The Hindu та BBC News. Така подвійна структура забезпечує відображення різноманітних

мовних, культурних та географічних контекстів, що робить цей набір цінним для навчання стійких моделей машинного навчання.

Згідно з розподілом даних (Рисунок Г.10.), 50% записів мають мітку “False” (3,795 зразків), а 50% – “True” (3,793 зразки), що забезпечує збалансованість набору даних для навчання моделей машинного навчання.

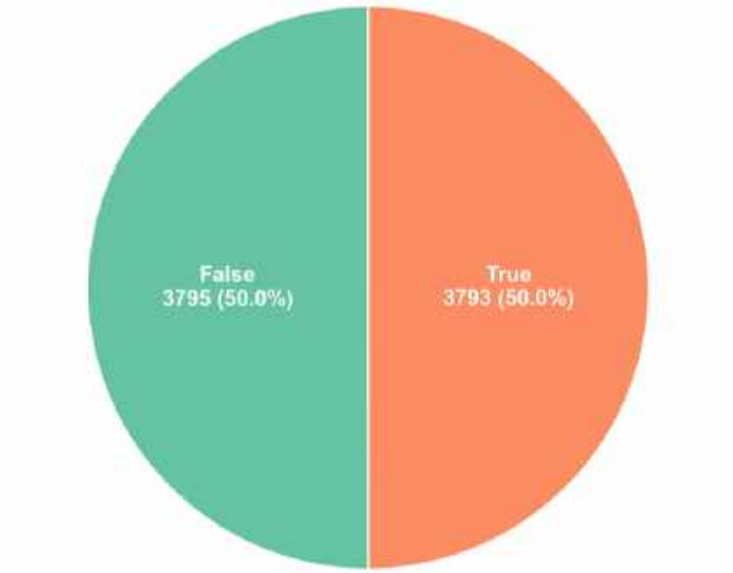

Рисунок Г.10. – Розподіл міток у наборі даних CoVID19-FNIR

Набір даних пройшов ретельне попереднє опрацювання, включаючи видалення непотрібних символів та інформації для підвищення його зручності у використанні. Файл fakeNews.csv містить такі стовпці, як дата публікації, текст статті, пояснення неправдивості, мітки для багатокласової та бінарної класифікації. Такий рівень деталізації допомагає зрозуміти джерела та причини поширення дезінформації. Водночас файл trueNews.csv містить перевірені твіти з надійних джерел, організовані за такими стовпцями, як дата, текст, видавець і Twitter-акаунт, що забезпечує надійну основу для ідентифікації автентичного контенту.

Завдяки такій структурованій та перевіреній інформації, набір даних CoVID19-FNIR є важливим інструментом для дослідників у боротьбі з дезінформацією та підвищенні надійності автоматизованих систем верифікації новин.

Г.6. Набір даних 6 (Фейкові новини у Телеграмі)

Для нашого аналізу ми використали два україномовні набори даних із Telegram: один містив лише заголовки, а інший – повні тексти повідомлень каналів Telegram. Цей великий набір даних є найбільшим відкритим джерелом новинного контенту, що містить приблизно 50 000 новинних статей і 10 700 заголовків новин, ретельно позначених як “фейк” або “правда”. Це забезпечує збалансовану основу для якісного та кількісного аналізу, а також застосування моделей машинного навчання.

Набір даних охоплює період із 24 лютого 2022 року до 11 грудня 2022 року, що є відрізком, позначеним важливими політичними та соціальними подіями в Україні. Такий часовий охват гарантує, що набір даних включає різноманітні теми та наративи, відображаючи як різні аспекти новинного контенту, так і еволюцію стратегій дезінформації.

Кожен запис у наборі даних містить чотири ключові атрибути: унікальний ідентифікатор (id), заголовок статті (title), повний текстовий вміст (text) і мітку класифікації (label), що вказує на автентичність новинного матеріалу, “True” – якщо новина підтверджена, “False” – якщо новина є фейковою.

Розподіл міток у наборі даних:

- 2498 фейкових новин (“False”);
- 8237 підтверджених новин (“True”).

Г.7. Набір даних 7 (Російська дезінфрмація від EUvsDisinfo)

Для аналізу наративів було використано базу даних російської дезінформації від EU East Stratcom Task Force, EUvsDisinfo. EUvsDisinfo ідентифікує, збирає та викриває випадки дезінформації, що походять із прокремлівських джерел.

Для цього аналізу було використано два набори даних. Набір даних 1, під назвою “Дезінформація про вибори президента США”, було відфільтровано за датами з 1 січня 2024 року по 5 листопада 2024 року та за хештегом “US Presidential elections”, у результаті чого було отримано 32 випадки російської дезінформації.

Для розуміння когнітивної платформи, на якій росія будує виборчі наративи, було використано більший набір даних 2, під назвою “Дезінформація щодо США 2024”. Він був відфільтрований за датами з 1 січня 2024 року по 5 листопада 2024 року та за країною США, що дозволило виокремити 616 випадків російських дезінформаційних статей, зосереджених на дискурсі навколо Сполучених Штатів.