УДК 004.8

Денисенко Б.О., Молчанова М.О., Мазурець О.В.

Хмельницький національний університет

ІНТЕЛЕКТУАЛЬНА СИСТЕМА ВИЯВЛЕННЯ ДЕЗІНФОРМАЦІЇ З ЗАСТОСУВАННЯМ ШТУЧНИХ НЕЙРОННИХ МЕРЕЖ

Результатом дослідження є створена інтелектуальна система виявлення дезінформації з застосуванням штучних нейронних мереж. Під час навчання, систему виявлення дезінформації вдалося натренувати до 99% правдивості результатів на навчальній, та до 91% на тестовій вибірці. Одержані результати демонструють важливість та ефективність використання штучних нейронних мереж для боротьби з дезінформацією, забезпечуючи користувачів надійним інструментом для перевірки правдивості повідомлень.

The result of the research is the creation of intelligent disinformation detection system using artificial neural networks. During training, the disinformation detection system was trained to 99% accuracy of results on the training sample and 91% on the test sample. The results demonstrate the importance and effectiveness of using artificial neural networks to combat disinformation, providing users with a reliable tool for checking the veracity of messages.

Аналіз проблеми дезінформації вимагає глибокого розуміння її контексту та ключових складових [1, 2]. Дезінформація охоплює широкий спектр сфер, зокрема журналістику та соціальні мережі, впливаючи як на саму інформацію, так і на джерела, авторів та канали її розповсюдження [3]. Виявлення дезінформації базується на детальному аналізі текстового контенту для ідентифікації неточностей, упереджених тверджень та оманливих фактів [4].

Ця проблема виходить за межі інформаційних технологій, переплітаючись із соціальними викликами, які пов'язані з розповсюдженням неправдивих відомостей [5]. У галузі журналістики перехід від друкованих ЗМІ до цифрових платформ сприяв швидшому поширенню новин, що, у свою чергу, посилило проблему дезінформації [6, 7]. У соціальних мережах швидке розповсюдження контенту, підкріплене механізмами «вірусного» поширення, також створює сприятливі умови для розповсюдження неправдивих матеріалів [8].

Основне завдання організацій, які займаються протидією дезінформації, полягає у захисті достовірності інформаційних потоків [9, 10]. Вони працюють над виявленням неправдивої інформації та мінімізацією її впливу. До таких організацій належать фактчекінгові агентства, інформаційні платформи та соціальні мережі [11, 12, 13]. В їхній діяльності задіяні різні спеціалісти — від модераторів контенту і аналітиків даних до інженерів із машинного навчання, кожен з яких відіграє важливу роль у боротьбі з дезінформацією [14].

У межах цієї сфери аналізуються різні елементи, такі як текстовий контент, джерела, автори та користувачі [15, 16]. Ключовими характеристиками цих елементів є метадані, наприклад, дата публікації, рівень достовірності джерела та показники взаємодії з аудиторією. Процес виявлення дезінформації [17] включає кілька ітераційних етапів: збір даних, їхню обробку, виділення ознак, навчання моделей та оцінку результатів. Кожен із цих етапів сприяє вдосконаленню підходів до ідентифікації неправдивої інформації.

На сучасному етапі є ряд програмних засобів, які можна використати для виявлення дезінформації.

Наприклад, «ZeroGPT» – інструмент для виявлення вмісту зі штучним інтелектом, який претендує на звання передового і надійного. Він може виявляти вміст, створений штучним інтелектом, з різних джерел, зокрема ChatGPT, GPT-4 і Bard. ZeroGPT пропонує безкоштовний план з обмеженими функціями та преміум-план з більшою кількістю функцій (рисунок 1) [18].

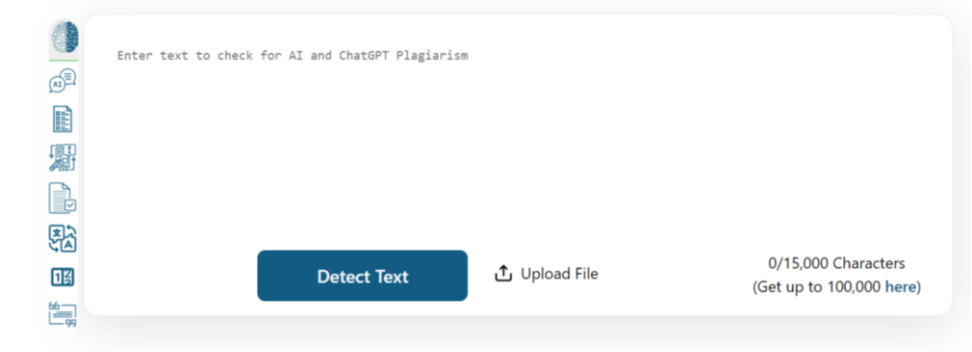

Рисунок 1 – Вигляд інструменту «ZeroGPT» [18]

До переваг можна віднести:

– Може виявляти контент, створений штучним інтелектом, який іноді може бути фейковою інформацією.

– Використовує багатоступеневу модель виявлення, яка навчається на різних наборах даних.

Недоліки:

– Незрозуміло, наскільки добре ZeroGPT може виявляти інші типи фейкової інформації, наприклад, інформацію, яка створена вручну, але фактично не відповідає дійсності.

– Веб-сайт не надає жодної інформації про точність моделі виявлення ZeroGPT.

«Fake text checker» позиціонує себе як найпростішу у світі браузерну утиліта для перевірки тексту на фальшивість. «Завантажте текст, який здається вам

підозрілим, у форму введення зліва, і ви миттєво отримаєте його статус в області виведення». Текст вважається несправжнім, якщо в ньому присутні літери з інших алфавітів. Потужно, безкоштовно і швидко (рисунок 2) [19].

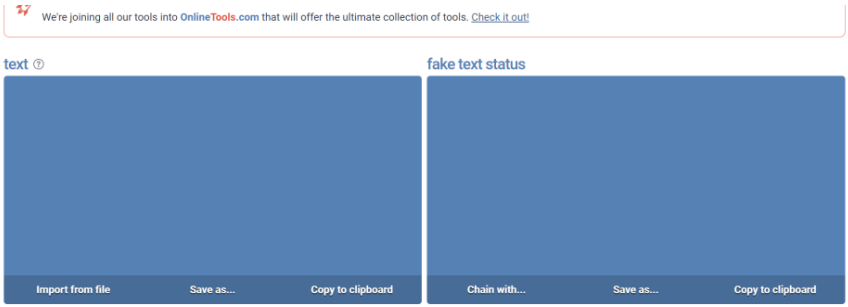

Рисунок 2 – Вигляд інструменту «Fake text checker» [19]

Плюси:

– Простота використання: Веб-сайт є простим інструментом, яким може користуватися будь-хто, щоб перевірити, чи є текст підробленим.

– Безкоштовність: веб-сайт не стягує жодної плати за використання свого інструменту.

Мінуси

– Точність: У статті не обговорюється, наскільки точно інструмент виявляє фейковий текст.

– Обмеження: У статті лише згадується, що інструмент можна використовувати для перевірки реклами. Незрозуміло, чи можна використовувати інструмент для інших цілей, наприклад, для перевірки новин або постів у соціальних мережах.

Загалом, цей веб-сайт видається багатообіцяючим інструментом для перевірки фейкових текстів. Однак потрібно більше інформації про точність і обмеження цього інструменту, перш ніж його можна буде рекомендувати для широкого використання.

«AI Content Detector» – детектор контенту зі штучним інтелектом. Їхній детектор може ідентифікувати вміст, створений ШІ, з точністю 99%. Він також може виявляти код, створений штучним інтелектом. Copyleaks пропонує безкоштовну пробну версію (рисунок 3) [20].

Плюси:

– Висока точність: Copyleaks стверджує, що ідентифікує вміст, згенерований ШІ, з точністю 99%.

– Виявляє код, створений ШІ: Це може бути корисно для розробників, які хочуть переконатися, що їхній код є оригінальним.

– Безкоштовна пробна версія: Copyleaks пропонує безкоштовну пробну версію, тому ви можете спробувати сервіс, перш ніж купувати його.

Мінуси:

– Вартість: Copyleaks є платним сервісом, тому вам потрібно буде врахувати його вартість при прийнятті рішення про те, чи варто ним користуватися.

– Обмежена сфера застосування: Copyleaks виявляє лише контент, створений штучним інтелектом. Він не може виявити плагіат з джерел, написаних людиною.

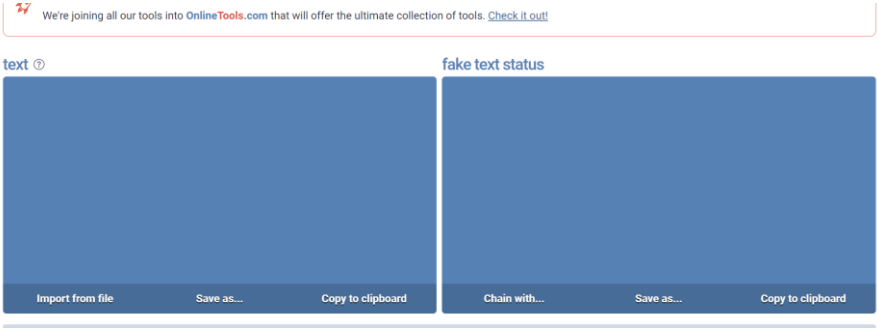

Рисунок 3 – Вигляд інструменту «AI Content Detector» [20]

Загалом, Copyleaks здається хорошим варіантом для людей, яким потрібно ідентифікувати контент, створений ШІ, з високим ступенем точності. Однак, перш ніж прийняти рішення, важливо врахувати вартість сервісу та його обмеження.

Після огляду зразків інструментів для виявлення фейкового контенту, можна зробити висновок, що перспективність розробки схожого застосунку є досить високою, оскільки під час дослідження існуючого програмного забезпечення предметної області, виявлено багато недоліків. Наприклад, працюють лише онлайн, або являють собою малопродуктивні системи з обмеженнями відносно вхідних даних, можуть виявляти тільки текст написаний штучним інтелектом. Тому доцільно розробити застосунок, який буде виявляти фейки не важливо ким написані, не матиме обмежень стосовно вхідних даних, та матиме можливість працювати без доступу до мережі інтернет, водночас зберігаючи здатність працювати на різних системах.

Отож, метою роботи є розробка інтелектуальної системи виявлення дезінформації з застосуванням штучних нейронних мереж.

Відповідно до поставленої мети, було спроєктовано відповідну структуру інтелектуальної системи виявлення дезінформації з застосуванням штучних нейронних мереж., зображену на Рисунку 4. Інформаційна система включає в себе три підсистеми та базу даних. Підсистема спілкування з користувачем інтелектуальної системи виявлення дезінформації з застосуванням штучних

нейронних мереж. відповідає за всі дії які були виконані користувачем. Також включає в себе збереження інформації, яку подав користувач для перевірки та відгук користувача про даний результат.

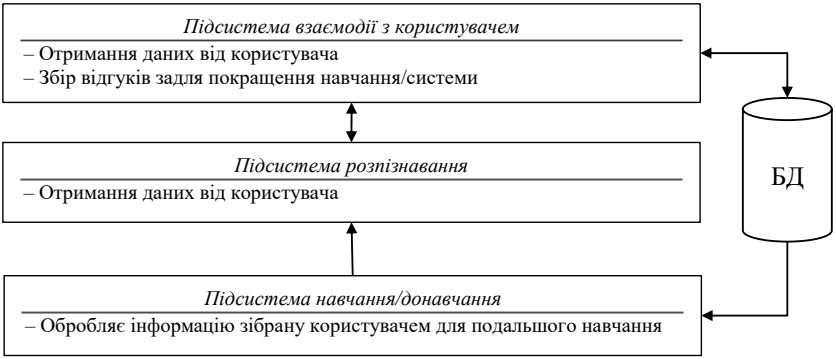

Рисунок 4 – Структура інтелектуальної системи виявлення дезінформації з застосуванням штучних нейронних мереж.

Підсистема розпізнавання – інтелектуальна модель для розпізнавання дезінформації. Підсистема навчання/донавчання необхідна для навчання інтелектуальної моделі інтелектуальної системи виявлення дезінформації з застосуванням штучних нейронних мереж. на готовій вибірці, а пізніше – додаткового донавчання на даних наданих користувачами.

База даних інтелектуальної системи виявлення дезінформації з застосуванням штучних нейронних мереж. містить в собі таблиці Messages та Messages_Score. Таблиця «Messages» зберігає всю інформацію про повідомлення, які надсилаються боту. Таблиця «Messages_Score» зберігає оцінки повідомлень, які надсилаються користувачами після перевірки ботом.

Розроблений бот для запуску потрібує наступні npm-пакети: node-telegram-bot-api, mysql2/promise, axios, також слід пересвідчитись в наявності MySQL 8 сервера на машині. Маючи запущеними модуль для виявлення дезінформації та телеграм бот, можна відкрити соціальну мережу.

На рисунку 5 наведено інтерфейс соціальної мережі та бота, що містить:

1. Поле для пошуку.

2. Поле для введення повідомлення для валідації на предмет дезінформації.

Після надсилання повідомлення, користувач отримає результат перевірки (рисунок 6).

Отже, результатом дослідження є створена інтелектуальна система виявлення дезінформації з застосуванням штучних нейронних мереж. Під час

навчання, систему вдалося натренувати до 99% правдивості результатів на навчальній, та до 91% на тестовій вибірці.

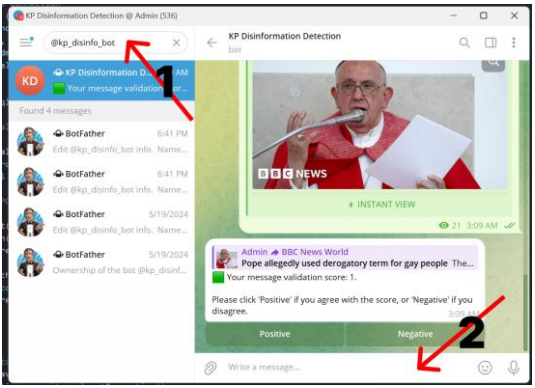

Рисунок 5 – Інтерфейс соціальної мережі та бота для виявлення дезінформації

Рисунок 6 – Результат валідації повідомлення

Створена інтелектуальна система є досить зручною в використанні, завдяки широкому поширенню застосунку для якого розроблявся бот, а також інтуїтивно зрозумілого інтерфейсу самого бота.

У нас час дослідження несе в собі неабияку користь. Маючи ворога, який витрачає такі колосальні зусилля на створення фейків, потрібно мати інструмент для якісної фільтрації повідомлень, що було й отримано в ході цього дослідження. Достатньо лише переслати повідомлення з новинного каналу, і одразу маємо результат перевірки.

Таким чином, результати цього дослідження демонструють важливість та ефективність використання штучних нейронних мереж для боротьби з дезінформацією, забезпечуючи користувачів надійним інструментом для перевірки правдивості повідомлень.

Перелік посилань

1. Sadia J. Artificial intelligence and journalistic practice: The crossroads of obstacles and opportunities for the Pakistani journalists. 2021.Journalism Practice 15: 1400–22.
2. Itai Himelboim, Guy Golan, July 2019. A Social Networks Approach to Viral Advertising: The Role of Primary, Contextual, and Low Influencers. DOI:10.1177/2056305119847516
3. Fake News in Digital Media: https://www.researchgate.net/publication/334167548_Fake_News_in_Digital_Media
4. Krak I., Didur V., Molchanova M., Mazurets O., Zalutska O., Manziuk E., Barmak O. Method for Political Propaganda Detection in Internet Content Using Recurrent Neural Network Models Ensemble. CEUR Workshop Proceedings, 2024, vol. 3806, pp. 312-324.
5. Zalutska O., Molchanova M., Sobko O., Mazurets O., Pasichnyk O., Barmak O., Krak I. Method for Sentiment Analysis of Ukrainian-Language Reviews in E-Commerce Using RoBERTa Neural Network. CEUR Workshop Proceedings, 2023, vol. 3387, pp. 344–356.
6. Залуцька О.О., Молчанова М.О., Віт Р.В., Мазурець О.В. Конфігурування нейронної мережі для класифікації емоційної тональності текстової інформації за показниками семантичної зв'язності. Збірник наукових праць за матеріалами XV Всеукраїнської науково-практичної конференції «Актуальні проблеми комп'ютерних наук АПКН-2023». Хмельницький, 2023. с. 102-107.
7. Mazurets O., Tymofiiev I., Dydo R. Approach for Using Neural Network BERT-GPT2 Dual Transformer Architecture for Detecting Persons Depressive State. Ricerche scientifiche e metodi della loro realizzazione: esperienza mondiale e realtà domestiche. Raccolta di articoli scientifici con gli atti della VI Conferenza scientifica e pratica internazionale. 15 novembre, 2024. Bologna, Repubblica Italiana. 2024. Pp. 147-151.
8. Мазурець О.В., Віт Р.В. Дослідження ефективності методу виявлення цільових об'єктів предметної області. Інформаційні технології і автоматизація. Матеріали XVII міжнародної науково-практичної конференції. Одеса, ОНТУ. 2024. С.650-653.
9. Mazurets O., Molchanova M., Klimenko V., Prosvitliuk M Practice Implementation of Neural Network Model BART-Large-CNN for Text Annotation. Prospects of Scientific Research in the Conditions of the Modern World. Proceedings of XXVII International scientific and practical conference. Rotterdam, Netherlands. 2024. Pp. 97-102.

10. Krak I., Zalutska O., Molchanova M., Mazurets O., Bahrii R., Sobko O., Barmak O. Abusive Speech Detection Method for Ukrainian Language Used Recurrent Neural Network. CEUR Workshop Proceedings, 2024, vol. 3688, pp. 16-28.
11. Мазурець О.В., Молчанова М.О., Кліменко В.І., Собко О.В., Супрун П.К. Даталогічна модель бази даних для виявлення гендерної приналежності за SVM-аналізом дописів інтернет-мереж з використанням об'єктно-орієнтованого проєктування. Науковий журнал «Вісник Хмельницького національного університету» серія: Технічні науки. Хмельницький, 2024. №3, Т.2 (337). С. 197-204.
12. Sobko O., Mazurets O., Didur V., Chervonchuk I. Recurrent Neural Network Model Architecture for Detecting a Tendency to Atypical Behavior Of Individuals by Text Posts. Theoretical and Practical Aspects of Modern Research. Proceedings of XXVI International scientific and practical conference. International Scientific Unity. Ottawa, Canada. 2024. Pp. 113-117.
13. Mazurets O.V., Sobko O.V., Molchanova M.O., Zalutska O.O., Yurchak A.V. Practical Implementation of Neural Network Method for Stress Features Detection by Social Internet Networks Posts. Global Science: Prospects and Innovations. Proceedings of the II International Scientific and Theoretical Conference «Scientific Review of the Actual Events, Achievements and Problems». Berlin, Federal Republic of Germany: International Center of Scientific Research. 2024. Pp. 160-167.
14. Blazhuk V., Mazurets O., Zalutska O. An Approach to Using the mBERT Deep Learning Neural Network Model for Identifying Emotional Components and Communication Intentions. The Impact of Scientific Research on the Development of the Modern World. Proceedings of the XLIV International scientific and practical conference. Dubrovnik, Croatia. 2024. Pp. 79-84.
15. Molchanova M., Mazurets O., Sobko O., Boiarchuk I. Object-Oriented Approach for Ethnic Enmity Detection in Text Messages by NLP. Proceedings of XXI International Scientific and Practical Conference «Scientific Achievements and Innovations as a Way to Success». Vilnius, Lithuania. 2024. Pp. 73-77.
16. Залуцька О.О., Кліменко В.І., Гладун О.В. Нейромережева модель для визначення емоційного стану людини у режимі реального часу. Інформаційні технології і автоматизація. Матеріали XVII міжнародної науково-практичної конференції. Одеса, ОНТУ. 2024. С.614-617.
17. Молчанова М.О., Мазурець О.В., Собко О.В., Кліменко В.І., Андрощук В.І. Метод нейромережевого виявлення кібербулінгу з використанням хмарних сервісів та об'єктно-орієнтованої моделі. Науковий журнал «Вісник Хмельницького національного університету» серія: Технічні науки. Хмельницький, 2024. №2 (333). С. 200-206.
18. ZeroGPT. URL: https://www.zerogpt.com.
19. Fake Text Checker. URL: https://onlinetexttools.com/check-if-text-is-fake.
20. AI Content Detector: URL: https://copyleaks.com/ai-content-detector.
21. Бармак О.В., Молчанова М.О., Денисенко Б.О. Нейромережева модель для виявлення дезінформації в текстовому контенті. Інформаційні технології і автоматизація. Матеріали XVII міжнародної науково-практичної конференції. 31 жовтня – 1 листопада 2024 р. Одеса, ОНТУ. 2024. С.583-585.