УДК 004.89:004.8 **DOI: https://doi.org/10.30837/PMW.2026.T1.252**

РОЗРОБЛЕННЯ ІНТЕЛЕКТУАЛЬНОЇ СИСТЕМИ АНАЛІЗУ ДОСТОВІРНОСТІ НОВИН ТА ІДЕНТИФІКАЦІЇ ЇХ ДЖЕРЕЛ З ВИКОРИСТАННЯМ ТРАНСФОРМЕРНИХ МОДЕЛЕЙ

Лозинська О.В.
к.т.н., доцент, кафедра Інформаційних систем та мереж,
НУ «Львівська політехніка»
ORCID ID: 0000-0002-5079-0544
Висоцька В.А.
д.т.н., доцент, кафедра Інформаційних систем та мереж,
НУ «Львівська політехніка»
ORCID ID: 0000-0001-6417-3689
Марків О.О.
к.т.н., доцент, кафедра Інформаційних систем та мереж
НУ «Львівська політехніка»
ORCID ID: 0000-0002-1691-1357
Бахмат К.Ю.
здобувач, кафедра Інформаційних систем та мереж
НУ «Львівська політехніка»

Анотація. *У розділі представлено прототип інтелектуальної системи для автоматизованого виявлення дезінформації та пошуку першоджерел інформації. Обґрунтовано використання трансформерних моделей, зокрема BERT, для класифікації текстових даних. Запропонований підхід базується на методах машинного навчання та опрацювання природної мови і забезпечує ефективний аналіз новин навіть за обмеженого обсягу даних. Застосування підходу, заснованого на оцінці впевненості моделі, підвищує інтерпретованість результатів. Визначено обмеження системи та окреслено напрями її подальшого розвитку.*

Ключові слова: *дезінформація, машинне навчання, опрацювання природної мови, bert-модель, веб-інтерфейс, трансформерні моделі.*

Вступ

У сучасних умовах стрімкого зростання обсягів цифрової інформації та поширення соціальних медіа проблема дезінформації набуває особливої актуальності. Автоматизовані системи перевірки достовірності контенту та ідентифікації першоджерел стають важливими інструментами забезпечення інформаційної безпеки, підвищення якості аналітичних рішень та протидії поширенню недостовірних даних.

У даному дослідженні представлено процес розроблення та реалізації прототипу інтелектуальної системи, яка функціонує як мінімальний життєздатний продукт і демонструє ключові підходи до автоматизованого

аналізу достовірності новин та пошуку їх першоджерел. Розглянуто архітектурні рішення, принципи побудови модульної структури системи, а також особливості інтеграції компонентів машинного навчання та веб-сервісів.

Окрему увагу приділено варіантам розгортання прототипу, його функціональним можливостям, обмеженням, а також опису користувацького інтерфейсу та сценаріїв взаємодії з системою. Представлені результати дають змогу оцінити практичну реалізацію запропонованих підходів та окреслити напрями подальшого вдосконалення системи.

Мета та задачі дослідження

Мета дослідження полягає у розробленні та реалізації прототипу інтелектуальної системи, призначеної для автоматизованого виявлення дезінформації та ідентифікації першоджерел інформації із застосуванням методів машинного навчання та технологій опрацювання природної мови.

Для досягнення поставленої мети було сформульовано наступні задачі дослідження.

1. Проаналізувати сучасні підходи та існуючі рішення у сфері автоматизованого виявлення фейкових новин і визначення достовірності інформації.
2. Дослідити методи опрацювання природної мови та архітектури моделей машинного навчання, зокрема трансформерних моделей, для задач класифікації текстів.
3. Спроєктувати архітектуру інтелектуальної системи з урахуванням принципів модульності, масштабованості та розширюваності.
4. Розробити модуль виявлення фейкових новин на основі моделі машинного навчання для класифікації текстової інформації.
5. Реалізувати модуль пошуку першоджерел інформації із використанням веб-скрейпінгу та інтеграції із зовнішніми пошуковими системами.
6. Провести аналіз обмежень розробленого прототипу та окреслити напрями його подальшого вдосконалення.

Основна частина

Аналіз літературних джерел та практичних рішень

Проблема автоматизованого виявлення дезінформації та ідентифікації першоджерел інформації є одним із найбільш актуальних напрямів сучасних досліджень у галузі штучного інтелекту, опрацювання природної мови [1] та інформаційної безпеки. Зростання обсягів цифрового контенту та швидкість його поширення в соціальних мережах зумовлюють необхідність розроблення ефективних методів автоматичної перевірки достовірності інформації.

У науковій літературі останніх років значна увага приділяється використанню моделей глибокого навчання для задач класифікації текстів.

Зокрема, трансформерні архітектури, такі як BERT [2], RoBERTa та їх похідні, демонструють високу ефективність у задачах семантичного аналізу тексту, визначення контексту та виявлення прихованих закономірностей у великих текстових корпусах. Ці моделі стали стандартом для багатьох завдань опрацювання природної мови (NLP), включаючи виявлення фейкових новин.

Окремий напрям досліджень пов'язаний із методами виявлення дезінформації. У роботах [3, 4] розглядаються як контент-орієнтовані підходи (аналіз тексту, стилометрія, лінгвістичні ознаки), так і контекстні методи, що враховують поширення інформації в соціальних мережах. Встановлено, що комбіновані підходи, які поєднують аналіз змісту та поведінкових характеристик інформації, є більш ефективними порівняно з ізольованими методами.

Важливим напрямом є також дослідження методів пошуку першоджерел інформації. У сучасних роботах застосовуються алгоритми зворотного пошуку, аналіз семантичної подібності та інформаційного ранжування. Зокрема, використання пошукових API та методів веб-скрейпінгу дозволяє автоматизувати процес виявлення першоджерел, хоча такі підходи мають обмеження, пов'язані зі змінами структури веб-сторінок та політиками доступу до даних.

У наукових дослідженнях також розглядаються методи виявлення розповсюджувачів дезінформації, зокрема підходи, що базуються на графовому представленні структури соціальних мереж, які дозволяють аналізувати зв'язки між користувачами та визначати ключові вузли поширення недостовірної інформації [5].

У літературі також активно розглядаються питання побудови архітектури інформаційних систем для аналізу великих обсягів даних. Сучасні рішення базуються на мікросервісній архітектурі, REST API та контейнеризації (Docker), що забезпечує масштабованість і гнучкість систем. Окремо підкреслюється важливість розділення системи на модулі опрацювання даних, машинного навчання та представлення результатів користувачу.

Попри значну кількість досліджень, існує низка невирішених проблем. Серед них – недостатня точність моделей у випадках обмеженого контексту, складність визначення першоджерела для перероблених або агрегованих новин, а також залежність систем від зовнішніх джерел даних. Це зумовлює необхідність подальших досліджень та розроблення інтелектуальних систем, які поєднують методи NLP, машинного навчання та веб-інтеграції.

Додатково до цього, у сучасних прикладних дослідженнях підкреслюється практична реалізація подібних систем, що дає змогу швидко перевіряти гіпотези та оцінювати ефективність архітектурних рішень. Зокрема, розроблена авторами система поєднує два ключові напрями: виявлення дезінформації на основі машинного навчання та пошук можливих першоджерел інформації.

Виявлення дезінформації на основі машинного навчання передбачає глибинний аналіз текстів, виявлення маніпулятивних конструкцій, порушень логіки та стилістичних патернів, що можуть свідчити про спотворення

інформації. Результатом такого аналізу є оцінка достовірності матеріалу та визначення ймовірності його належності до дезінформаційних повідомлень.

Другий напрям стосується модулів реверсивного пошуку першоджерел інформації, які дають змогу встановлювати первинний контекст появи матеріалів та аналізувати їхнє поширення в інформаційному середовищі. Це забезпечує можливість перевірки авторитетності джерел і розуміння трансформації контенту в процесі його розповсюдження.

Таким чином, подібні системи позиціонуються як інструменти підвищення інформаційної грамотності та підтримки користувачів у складному інформаційному середовищі, включаючи журналістів, фактчекерів та дослідників.

Пошук та визначення наявних рішень для виявлення дезінформації.

Першим кроком у розробленні інтелектуальної системи є вивчення наявних рішень, які частково або повністю виконують схожі функції. Такий аналіз дає змогу визначити рівень зрілості ринку, окреслити актуальні тенденції, виявити технологічні обмеження та зрозуміти, які можливості залишаються незадіяними. Порівняння здійснювалося з провідними світовими рішеннями у сфері протидії дезінформації.

Найбільш відомі практичні рішення включають системи, що безпосередньо працюють із виявленням дезінформації, перевіркою фактів або аналізом контенту. До них належать Logically AI [6], Blackbird AI [7], Primer AI [8] та Twitter/X Community Notes [9], які демонструють різні підходи до автоматизації або напівавтоматизації процесу перевірки інформації.

Розроблення та реалізація прототипу інтелектуальної системи для аналізу достовірності новин

У рамках дослідження було здійснено розроблення діючого прототипу інтелектуальної системи для виявлення дезінформації та ідентифікації першоджерел інформації, реалізованого як мінімальний життєздатний продукт. Прототип призначений для демонстрації основних функціональних можливостей системи автоматизованого виявлення дезінформації та ідентифікації першоджерела інформації. Розробка виконана із застосуванням сучасних технологій інженерії програмного забезпечення, що забезпечують можливість подальшого масштабування системи.

Архітектура прототипу побудована на основі сучасного технологічного стеку, що забезпечує продуктивність, масштабованість та підтримуваність. Вибір технологій здійснено з урахуванням вимог до опрацювання природної мови, інтеграції із зовнішніми сервісами та забезпечення надійності.

До основних технологічних компонентів належать:

– мови програмування та фреймворки, серед яких використано Python 3.11, FastAPI, Uvicorn;

– машинне навчання та NLP на основі Transformers, PyTorch, Scikit-learn, Hugging Face;

– веб-скрейпінг з використанням Requests, BeautifulSoup, Selenium;
– бази даних реалізовано у SQLAlchemy, PostgreSQL, Alembic;
– контейнеризація з використанням Docker, Docker Compose;
– тестування проведено із залученням Pytest, Pytest-cov, HTTPX.

Архітектура системи реалізована за принципами модульного проєктування з чітким розмежуванням відповідальності. Система включає два основні модулі: аналіз достовірності новин та пошук першоджерела.

Модуль аналізу достовірності новин забезпечує класифікацію текстів за допомогою моделі на базі архітектури BERT. Модуль пошуку джерела здійснює ідентифікацію першоджерела через інтеграцію з пошуковими системами.

Шар інтеграції координує взаємодію модулів, а шар представлення реалізовано у вигляді REST API та веб-інтерфейсу. Система підтримує такі функціональні можливості: автоматизоване виявлення фейкових новин, ідентифікацію першоджерела, комбінований аналіз, навчання моделей та інтерактивну взаємодію з користувачем.

Система підтримує гнучке налаштування через змінні середовища, автоматичну ініціалізацію бази даних та логування. Реалізовано базові механізми моніторингу через спеціальні API-ендпоінти.

Разом з тим, прототип має низку обмежень:
– використання базової моделі BERT без спеціалізованого донавчання;
– нестабільність веб-скрейпінгу як механізму пошуку джерел;
– обмежена масштабованість;
– базовий рівень безпеки;
– обмежені можливості моніторингу.

Ці обмеження планується усунути у повноцінній версії системи.

Опис інтерфейсу користувача та процедура роботи з системою

Прототип інтелектуальної системи реалізує веб-інтерфейс користувача, що забезпечує інтуїтивну взаємодію із функціоналом системи.

Інтерфейс організовано як односторінковий додаток, що включає:
– заголовок системи;
– форму введення даних;
– опції аналізу;
– кнопку запуску;
– секцію результатів.

Користувач може вводити заголовок та текст новини, обирати тип аналізу та ініціювати обробку.

Процес взаємодії включає наступні етапи (рис. 1).

1. Введення тексту.
2. Вибір типу аналізу.
3. Ініціація обробки.
4. Отримання результатів.

Система підтримує кілька сценаріїв:
– класифікація правдивих новин із високою точністю;
– виявлення фейкових новин;
– опрацювання помилок.

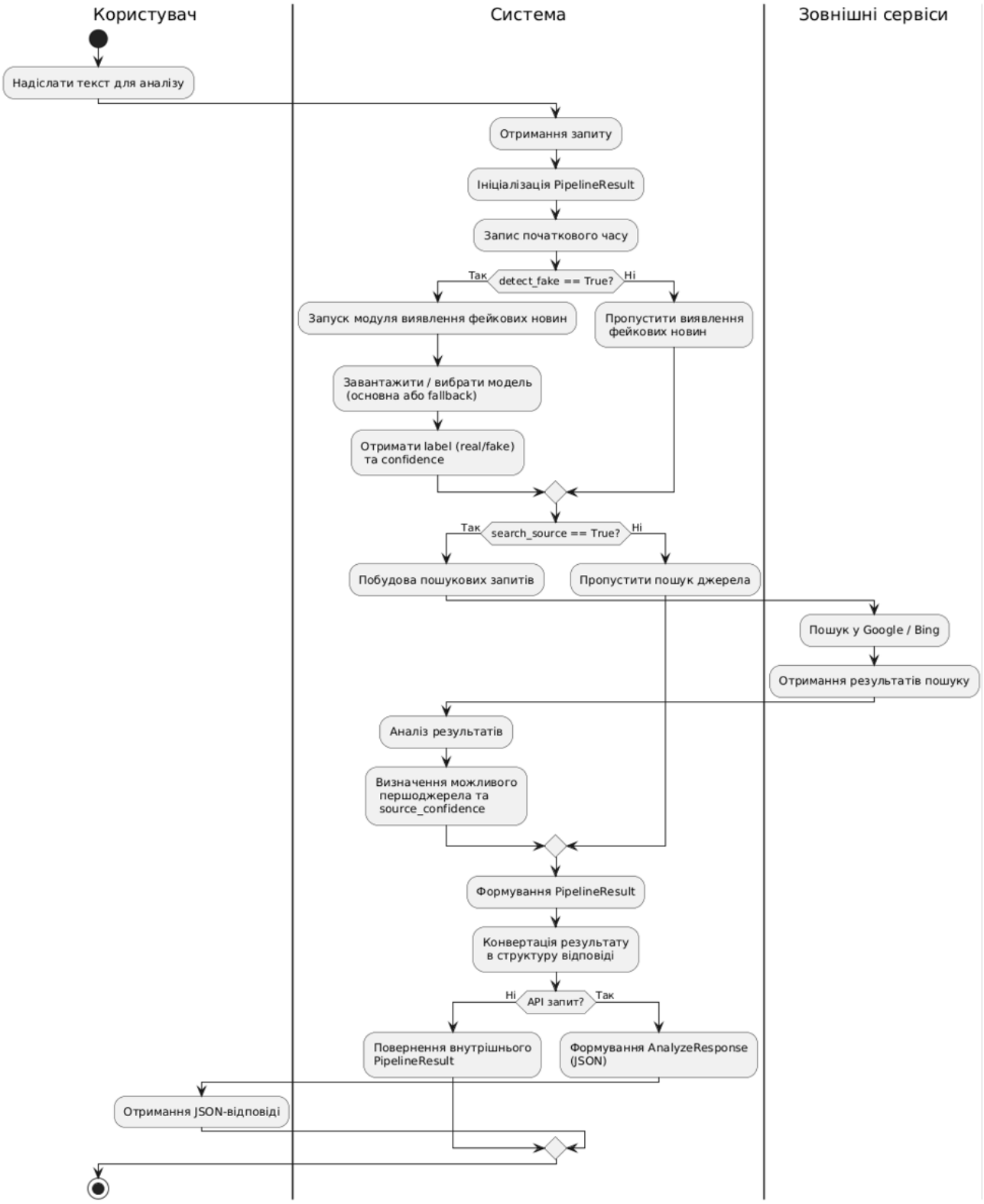

Рисунок 1 – UML-діаграма діяльності основного процесу

Результати досліджень

Результати подаються у зручній візуальній формі із зазначенням F1-метрики та деталізацією ймовірностей.

Інтерфейс має адаптивний дизайн, підтримує візуальну індикацію станів та забезпечує детальне представлення результатів аналізу, що підвищує зручність використання системи.

У межах дослідження було реалізовано експериментальну модель для задачі автоматизованого виявлення дезінформації на основі методів глибокого навчання. Основною метою експерименту було оцінювання ефективності підходу до класифікації новинних текстів із використанням трансформерної архітектури та аналізу текстово-контекстних ознак (рис. 2).

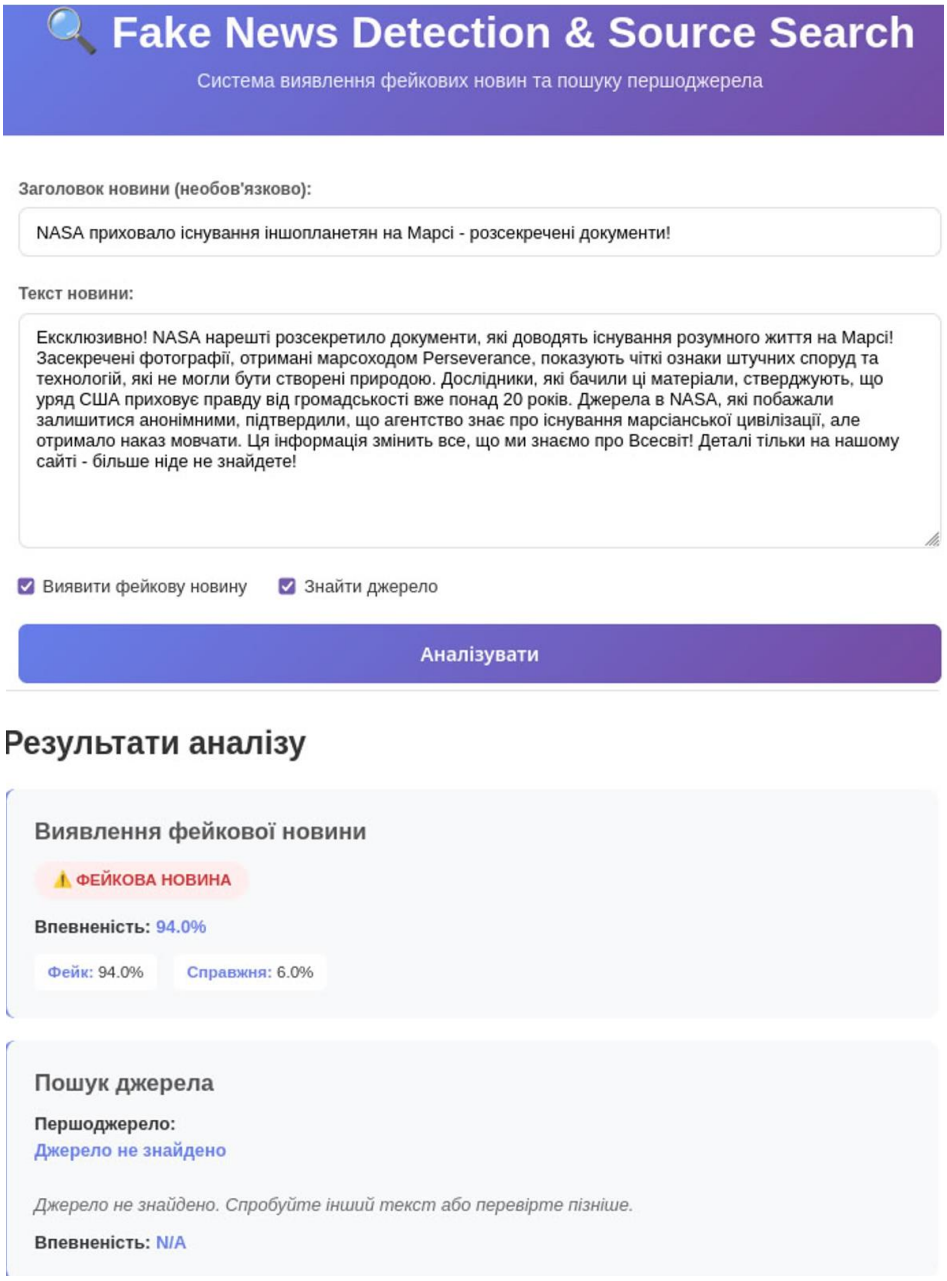

Рисунок 2 – Результати роботи системи

На рис. 2 подано інтерфейс роботи інтелектуальної веб-системи та результати автоматизованого аналізу введеного тексту новини.

У першій частині інтерфейсу користувач вводить заголовок та текст новини, після чого активує процес аналізу. У наведеному прикладі введено текст, що містить ознаки дезінформації, пов'язаної з нібито розсекреченням даних NASA про існування позаземного життя. У результатах аналізу система відобразила такі показники: результат класифікації (новину визначено як фейкову), рівень впевненості моделі, який становить 94.0%, ймовірнісний розподіл класів (фейк – 94.0%, справжня новина – 6.0%).

Для проведення навчання та тестування моделі було використано один датасет, що містив збалансовані класи:

– fake news – 100 прикладів;

– real news – 100 прикладів.

Таким чином, загальний обсяг вибірки становив 200 записів.

Датасет є структурований і містить такі атрибути новин:

– title (заголовок);

– text (основний текст);

– url (посилання на джерело);

– top_img (головне зображення);

– author (автор матеріалу);

– source (джерело публікації);

– publish_date (дата публікації);

– images (додаткові зображення);

– canonical_link (канонічне посилання);

– meta_data (метадані сторінки).

Попри наявність багатьох полів, основну увагу в моделі приділено текстовим ознакам (title та text), які є найбільш інформативними для задачі класифікації дезінформації.

Навчання моделі здійснено із застосуванням стандартного підходу для задач бінарної класифікації текстів.

У процесі оптимізації використано:

– функцію втрат Cross-Entropy Loss;

– оптимізатор AdamW з L2-регуляризацією.

Застосування AdamW дало змогу стабілізувати процес навчання та зменшити ефект перенавчання за рахунок коректної реалізації вагової декомпозиції (weight decay). Додатково використано механізми адаптивного навчання:

– поступове збільшення learning rate на початкових етапах (warm-up strategy);

– подальше обмеження градієнтів (gradient clipping) для стабілізації оновлення ваг;

– навчання малими пакетами (mini-batch training);

– валідація після кожної ітерації.

Для оцінювання якості моделі використано F1-метрику як найбільш збалансований показник для задач класифікації, особливо у випадках потенційної диспропорції помилок першого та другого типу.

Також у процесі навчання застосовано ймовірнісну оцінку впевненості моделі (confidence-based evaluation), що дає змогу інтерпретувати вихід моделі як ступінь її впевненості у прийнятому рішенні. Такий підхід підвищує інтерпретованість результатів і є важливим для практичних систем аналізу інформації.

Отримані результати свідчать про високу впевненість моделі у класифікації тексту як недостовірного, що відповідає характеру навмисно сенсаційного та маніпулятивного контенту, представленого у вхідних даних.

Додатково в межах другого функціонального блоку системи – пошуку першоджерела інформації – було здійснено спробу реверсивного пошуку відповідного джерела. Результат роботи модуля показав відсутність достовірного першоджерела (“джерело не знайдено”), що є типовим для фейкових новин, які не мають підтвердження в авторитетних інформаційних ресурсах.

Таким чином, система продемонструвала коректну роботу обох основних модулів: класифікації тексту та пошуку першоджерела, що підтверджує ефективність запропонованого підходу до автоматизованого аналізу інформації.

У процесі навчання система автоматично зберігає найкращі ваги моделі відповідно до значення F1-метрики. Додатково реалізовано механізм ранньої зупинки (early stopping), який припиняє навчання у випадку, якщо показники якості перестають покращуватися протягом певної кількості ітерацій.

Такий підхід дав змогу уникнути перенавчання моделі, скоротити час навчання та забезпечити стабільність результатів.

Отримані результати свідчать про те, що навіть при використанні відносно невеликого та збалансованого датасету модель здатна навчатися розрізненню між справжніми та фейковими новинами на основі текстових ознак.

Застосування підходу, заснованого на оцінці впевненості моделі, дало змогу додатково інтерпретувати результати моделі у вигляді ймовірнісної оцінки достовірності, що є важливим для практичних систем аналізу інформації, орієнтованих на кінцевого користувача.

Висновки

У дослідженні представлено розроблений прототип інтелектуальної системи, що реалізує функції автоматизованого виявлення дезінформації та пошуку першоджерел. Проаналізовано наукові підходи до виявлення фейкових новин, визначено доцільність використання трансформерних моделей, зокрема архітектури BERT. Запропонований авторами підхід базується на використанні сучасних методів машинного навчання та опрацювання природної мови, що дало змогу забезпечити ефективну класифікацію текстових даних.

Експериментальна частина дослідження продемонструвала, що навіть за умов використання обмеженого за обсягом датасету (200 прикладів) модель здатна забезпечити достатній рівень точності класифікації. Застосування підходу, заснованого на оцінці впевненості моделі, дало змогу підвищити інтерпретованість результатів та надати користувачу додаткову інформацію щодо надійності прогнозів.

Разом з тим, дослідження виявило низку обмежень, зокрема використання невеликого датасету, відсутність спеціалізованого донавчання моделі на доменно-орієнтованих даних, а також обмеженість механізмів інтеграції із зовнішніми сервісами. Це визначає перспективні напрями подальших досліджень, серед яких – розширення навчальної вибірки, інтеграція з соціальними платформами та підвищення рівня безпеки і масштабованості системи.

Подяка. Дослідження було проведено за грантової підтримки Міністерства освіти і науки України «Методи та засоби виявлення дезінформації у соціальних мережах на основі технологій глибинного навчання» в рамках проєкту № 0125U001852.

Список літератури.

1. Бондаренко, О.В., & Ковальчук, І.О. (2021). Методи обробки природної мови в задачах аналізу текстової інформації. Наукові праці НУ «Львівська політехніка». Серія: Комп'ютерні науки та інформаційні технології, 5(12), 45-52.

2. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, 4171-4186. https://doi.org/10.18653/v1/N19-1423.

3. Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., & Liu, Y. (2019). Combating fake news: A survey on identification and mitigation techniques. ACM Transactions on Intelligent Systems and Technology, 10(3), 1-42. https://doi.org/10.1145/3305260.

4. Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. ACM Computing Surveys, 53(5), 1-40. https://doi.org/10.1145/3395046.

5. Лозинська, О.В., Марків, О.О., & Висоцька, В.А. (2025). Метод виявлення розповсюджувачів дезінформації на основі графового представлення структури соціальної мережі. Центральноукраїнський науковий вісник. Технічні науки. 11 (42), 70-78.

6. Logically AI. (n.d.). AI-powered misinformation detection platform. https://logically.ai/.

7. Blackbird.AI. (n.d.). Narrative intelligence and risk detection platform. https://blackbird.ai/.

8. Primer AI. (n.d.). AI-powered text and intelligence analysis platform. https://www.primer.ai/.

9. X. (n.d.). Community Notes: crowdsourced context system. https://x.com/i/communitynotes.