Міністерство освіти і науки України
Тернопільський національний технічний університет імені Івана Пулюя

Факультет комп'ютерно-інформаційних систем і програмної інженерії
(повна назва факультету)
Кафедра комп'ютерних наук
(повна назва кафедри)

КВАЛІФІКАЦІЙНА РОБОТА

на здобуття освітнього ступеня

магістр
(назва освітнього ступеня)

на тему: Використання штучного інтелекту для виявлення дезінформації в новинах соціальної мережі Facebook

Виконав: студент VI курсу, групи СНнм-61
спеціальності 122 Комп'ютерні науки
(шифр і назва спеціальності)

_____ Дацик С.В.
(підпис) (прізвище та ініціали)

Керівник _____ Дмитроца Л.П.
(підпис) (прізвище та ініціали)

Нормоконтроль _____ Готович В.А.
(підпис) (прізвище та ініціали)

Завідувач кафедри _____ Боднарчук І.О.
(підпис) (прізвище та ініціали)

Рецензент _____ Яворська Є.Б.
(підпис) (прізвище та ініціали)

Тернопіль
2024

Міністерство освіти і науки України
Тернопільський національний технічний університет імені Івана Пулюя

Факультет комп'ютерно-інформаційних систем і програмної інженерії
(повна назва факультету)

Кафедра комп'ютерних наук
(повна назва кафедри)

ЗАТВЕРДЖУЮ
Завідувач кафедри
_____ Боднарчук І.О.
(підпис) (прізвище та ініціали)
«_29_» _травня 2024 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

на здобуття освітнього ступеня Магістр
(назва освітнього ступеня)

за спеціальністю 122 Комп'ютерні науки
(шифр і назва спеціальності)

Студенту Дацику Станіславу Васильовичу
(прізвище, ім'я, по батькові)

1. Тема роботи «Використання штучного інтелекту для виявлення дезінформації в новинах соціальної мережі Facebook»

Керівник роботи Дмитроца Леся Павлівна, к.т.н., доцент кафедри КН
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

Затверджені наказом ректора від « 24 » листопада 2023 року № _4/7-1100_____

2. Термін подання студентом завершеної роботи 29 травня 2024р.

3. Вихідні дані до роботи базуються на основі вхідних даних та результатах дослідження предметної області, використаних літературних джерелах, розглянутих інтернет-ресурсах на тему штучного інтелекту виявлення дезінформації в новинах соціальної мережі Facebook

4. Зміст роботи (перелік питань, які потрібно розробити)
Вступ. 1 Оцінка сучасного стану використання штучного інтелекту для виявлення дезінформації. 1.1 Історичний контекст та еволюція дезінформації. 1.2 Роль штучного інтелекту в ідентифікації фейкових новин. 1.3 Аналіз існуючих досліджень та підходів боротьби з дезінформацією. 1.4 Виклики та обмеження сучасних систем виявлення дезінформації. 1.5 Висновок до першого розділу. 2 Методи та інструментарій штучного інтелекту проти дезінформації. 2.1 Основні поняття та визначення 2.2 Дослідження методів ML та NLP. 2.3 Аналіз інструментів виявлення дезінформації в новинах Facebook. 2.4 Аналіз теоретичних моделей та експериментальних досліджень вдосконалення AI- інструментів. 2.5 Висновок до другого розділу. 3 Розробка AI-системи виявлення фейкових новин. 3.1 Побудова та оцінка моделі NLP. 3.2 Розробка веб-інтерфейсу користувача. 3.3 Веб- інтерфейс AI-системи для виявлення фейкових новин 3.4 Висновок до третього розділу. 4 Охорона праці та безпека в надзвичайних ситуаціях. 4.1 Безпека з охорони праці та організація робочого місця користувачів ПК. 4.2 Підвищення стійкості роботи об'єктів приладобудування у воєнний час. 4.3 Висновок до четвертого розділу. Висновки. Перелік джерел. Додатки.

5. Перелік графічного матеріалу (з точним зазначенням обов'язкових креслень, слайдів)
1 Титульна сторінка. 2 Тема, Мета, Об'єкт, Предмет дослідження. 3 Завдання дослідження. 4. Актуальність дослідження. 5. Методи штучного інтелекту. 6. Категоризація результатів моделей. 7. Модель Support Vector Classifier. 8. Модель Logistic Regression та Multilayer Perceptron. 9. Інструменти розробки. 10. Архітектура проєкту. 11. Веб-інтерфейс користувача. 12. Виявлення правдивої інформації. 13. Результати інших моделей14. Виявлення дезінформації. 15. Результати перевірки. 16. Форма відгуків. 17. Створення моделей 18 Апробація результатів. 19 Висновки.

6. Консультанти розділів роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
Охорона праці	Сенчишин В.С., доцент		
Безпека в надзвичайних ситуаціях	Клепчик В.М., ст. викладач		

7. Дата видачі завдання 24 листопада 2023 р.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів роботи	Термін виконання етапів роботи	Примітка
1.	Ознайомлення з завданням до кваліфікаційної роботи	04.12.2023	*Виконано*
2.	Підбір наукових джерел про методи штучного	15.12.2023-31.11.2023	*Виконано*
	інтелекту для ідентифікації фейкових новин у		
	соціальній мережі Facebook		
3.	Опрацювання наукових публікацій та збір даних по	15.01.2024-25.02.2024	*Виконано*
	темі роботи		
4.	Виконання дослідження згідно мети кваліфікаційної	26.02.2024-07.04.2024	*Виконано*
	роботи		
5.	Оформлення розділу «Аналіз сучасного стану	15.04.2024-18.04.2024	*Виконано*
	використання штучного інтелекту для виявлення		
	дезінформації»		
6.	Оформлення розділу «Методи та інструменти	19.04.2024-25.04.2024	*Виконано*
	штучного інтелекту проти дезінформації»		
7.	Оформлення розділу «Розробка AI-системи виявлення	26.04.2024-02.05.2024	*Виконано*
	фейкових новин»		
8.	Виконання завдання до підрозділу «Охорона праці»	03.05.2024-07.05.2024	*Виконано*
9.	Виконання завдання до підрозділу «Безпека в	08.05.2024-10.05.2024	*Виконано*
	надзвичайних ситуаціях»		
10.	Оформлення кваліфікаційної роботи	11.05.2024-14.05.2024	*Виконано*
11.	Нормоконтроль	15.05.2024-16.05.2024	*Виконано*
12.	Перевірка на плагіат	17.05.2024	*Виконано*
13.	Попередній захист кваліфікаційної роботи	21.05.2024	*Виконано*
14.	Захист кваліфікаційної роботи	29.05.2024	

Студент _____ Дацик С.В.
(підпис) (прізвище та ініціали)

Керівник роботи _____ Дмитроца Л.П
(підпис) (прізвище та ініціали)

АНОТАЦІЯ

Використання штучного інтелекту для виявлення дезінформації в новинах соціальної мережі Facebook // Кваліфікаційна робота освітнього рівня «Магістр» // Дацик Станіслав Васильович // Тернопільський національний технічний університет імені Івана Пулюя, факультет комп'ютерно-інформаційних систем і програмної інженерії, кафедра комп'ютерних наук, група СНнм-61 // Тернопіль, 2024 // С. 74, рис. – 22, табл. – 0, кресл. – 19, додат. – 4, бібліогр. – 56.

Ключові слова: штучний інтелект, машинне навчання, класифікатор, модель, система, дезінформація, фейкові новини, парсинг.

Кваліфікаційна робота присвячена розробці інструменту штучного інтелекту для виявлення дезінформації в новинах соціальної мережі Facebook.

В першому розділі роботи описано роль штучного інтелекту у виявленні дезінформації, висвітлено використання алгоритмів для аналізу лінгвістичних шаблонів, розглянуто потребу в комбінованій стратегії людини та штучного інтелекту, та проаналізовано ефективність інструментів як Grover у виявленні фейкових новин.

В другому розділі досліджено методи та інструменти штучного інтелекту для виявлення дезінформації, подано огляд інструментів для аналізу текстових даних.

В третьому розділі описано архітектуру розробленої AI-системи для виявлення дезінформації в новинах соціальної мережі Facebook, проаналізовано методи машинного навчання, та проведено експерименту роботу розробленої системи штучного інтелекту.

У четвертому розділі детально розглянуті аспекти охорони праці, включаючи ергономічні вимоги до робочого місця, оформлення робочого кабінету, та необхідні умови для роботи за персональним комп'ютером. Також розглянуто способи забезпечення стійкості роботи об'єктів приладобудування у воєнний час.

ANNOTATION

Use of artificial intelligence to detect misinformation in Facebook social network news // The educational level «Master» qualification work // Datsyk Stanislav Vasylovych // Ternopil Ivan Puluj National Technical University, Faculty of Computer Information Systems and Software Engineering, Department of Computer Science, SNnm-61 group // Ternopil, 2024 // P. 74, fig. - 22, tables - 0, posters - 19, annexes - 4, ref. - 56.

Key words: artificial intelligence, machine learning, classifier, model, system, misinformation, fake news, parsing.

Thesis is devoted to the development of an artificial intelligence tool for detecting misinformation in Facebook news.

The first section of the paper describes the role of AI in detecting misinformation, highlights the use of algorithms to analyze linguistic patterns, considers the need for a combined human and AI strategy, and analyzes the effectiveness of tools like Grover in detecting fake news.

The second section explores the methodology and tools of AI for detecting misinformation, providing an overview of tools for analyzing text data and identifying fake images.

The third section describes the architecture of the developed AI system, analyzes machine learning methods, and conducts a computational experiment. Object of study: the use of AI for news analysis, subject of study: the effectiveness of AI in detecting misinformation.

The fourth section discusses in detail the aspects of occupational health and safety, including ergonomic requirements for the workplace, office design, and the necessary conditions for working at a PC. It also discusses ways to ensure the sustainability of instrumentation facilities in wartime.

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

AI (анг. artificial intelligence) – штучний інтелект.

HTML (анг. hypertext markup language) – мова розмітки гіпертексту

ML (анг. machine learning) – машинне навчання

MVC (анг. model-view-controller) – модель-вид-контролер

NLP (анг. natural language processing) – обробка природної мови.

ORM (анг. object-relational mapping) – об'єктно-реляційне відображення

URL (анг. uniform resource locator) – уніфікований локатор ресурсів

XAI (анг. explainable artificial intelligence) – інтерпретований штучний інтелект

ЗМІ – засоби масової інформації

ШІ – штучний інтелект.

ЗМІСТ

ВСТУП

Актуальність теми. Використання штучного інтелекту для виявлення дезінформації в новинах соціальної мережі Facebook є надзвичайно важливою. Інформаційна війна, яка відбувається в онлайн-середовищі, стає все більш загостреною, і важливо розробляти ефективні методи виявлення та боротьби з дезінформацією.

Штучний інтелект може відігравати ключову роль у цьому процесі. Моделі машинного навчання, такі як Support Vector Classifier, Logistic Regression та Multilayer Perceptron, можуть бути використані для аналізу текстового контенту, виявлення підозрілих новин та відсіювання дезінформації.

Застосування штучного інтелекту для виявлення дезінформації допоможе зменшити поширення фейкових новин та покращити якість інформації, яка доходить до користувачів соціальних мереж. Це важливий крок у боротьбі з інформаційною війною та забезпеченні об’єктивної та достовірної інформації для громадськості.

Мета і задачі дослідження. Метою даної кваліфікаційної роботи освітнього рівня «Магістр» є розробка системи штучного інтелекту для виявлення дезінформації в новинах соціальної мережі Facebook. Для досягнення поставленої мети потрібно виконати ряд завдань, зокрема:

– Провести аналіз досліджень, пов'язаних з виявленням дезінформації у мережі Facebook.

– Вивчити та порівняти існуючі методи штучного інтелекту, що використовуються для виявлення дезінформації на платформах соціальних мереж.

– Проаналізувати ефективність різних методів виявлення дезінформації та порівняти їх, щоб визначити сильні та слабкі сторони.

– Запропонувати модель на основі штучного інтелекту, призначену для виявлення та боротьби з дезінформацією у Facebook.

– Розробити систему штучного інтелекту для виявлення дезінформації в новинах соціальної мережі Facebook.

Об'єкт дослідження. Виявлення дезінформації у новинах, що поширюються в соціальній мережі Facebook.

Предмет дослідження. Модель на основі штучного інтелекту для виявлення дезінформації у Facebook, методи та технології штучного інтелекту, алгоритми машинного навчання, обробка природної мови (NLP), нейронні мережі, методи та технології штучного інтелекту, алгоритми машинного навчання, обробка природної мови (NLP), нейронні мережі.

Наукова новизна одержаних результатів кваліфікаційної роботи полягає в розробці комплексної системи, що поєднує моделі штучного інтелекту, такі як Support Vector Classifier, Logistic Regression та Multilayer Perceptron для виявлення дезінформації у новинах в соціальній мережі Facebook.

Практичне значення одержаних результатів полягає у використанні штучного інтелекту для виявлення дезінформації в новинах соціальної мережі Facebook, користувачі якого можуть краще розрізняти достовірну та неправдиву інформацію, тим самим сприяючи розвитку більш поінформованої та критично обізнаної онлайн-спільноти.

Апробація результатів магістерської роботи. Основні результати проведених досліджень обговорювались на:

– IV науково-технічної конференції «Інформаційні моделі, системи та технології». Тернопільського національного технічного університету імені Івана Пулюя (м. Тернопіль, 2023 р.).

– Світ наукових досліджень. Випуск 28: матеріали Міжнародної мультидисциплінарної наукової інтернет-конференції (м. Тернопіль, Україна, м. Ополе, Польща, 21-22 березня 2024 р.).

Публікації. Основні результати кваліфікаційної роботи опубліковано у трьох працях конференцій (Див. додатки А).

Структура й обсяг кваліфікаційної роботи. Кваліфікаційна робота складається зі вступу, чотирьох розділів, висновків, списку літератури з 53 найменувань та 4 додатків. Загальний обсяг кваліфікаційної роботи складає 90 сторінки, з них 74 сторінки основного тексту, який містить 22 рисунки.

1 АНАЛІЗ СУЧАСНОГО СТАНУ ВИКОРИСТАННЯ ШТУЧНОГО ІНТЕЛЕКТУ ДЛЯ ВИЯВЛЕННЯ ДЕЗІНФОРМАЦІЇ

1.1 Історичний контекст та еволюція дезінформації

У цифрову еру історичний контекст дезінформації значно змінився завдяки появі складних технологій, таких як чат-боти ШІ. Які здатні імітувати людську розмову і стали інструментом для поширення неправдивої інформації із загрозливою швидкістю [1]. Здатність імітувати людську мову дає змогу поширювати спеціалізовані наративи, призначені для маніпулювання громадською думкою, створення розбрату та навіть впливу на політичні результати. Особливе занепокоєння викликає наближення виборів, що створює сприятливі умови для кампаній дезінформації, спрямованих на те, щоб вплинути на сприйняття виборців і підірвати демократичний процес [2]. Незважаючи на те, що це давня проблема, сучасне втілення дезінформації, підкріплене технологіями, випередило традиційні методи перевірки фактів і модерування [3]. Тому вкрай важливо розробити інноваційні стратегії, які можуть подолати нюанси викликів, пов’язаних із дезінформацією, керованою ШІ, особливо в політично напруженій атмосфері навколо виборчих змагань.

Із розширенням цифрового ландшафту зростає складність і вплив дезінформації. У минулому дезінформація могла складатися з чуток або пропаганди, що поширювалася через традиційні ЗМІ, але сьогодні вона перетворилася на більш підступну форму неправдивого маніпулятивного контенту. Цей вміст часто створюється так, ніби він походить із законних джерел, що ускладнює для споживачів розрізнення правди, а що ні. Метою цієї еволюції є не просто дезінформація, а навмисне введення в оману та дезорієнтація споживачів [4]. Створюючи плутанину та недовіру, ті, хто поширює дезінформацію, мають на меті пошкодити саму структуру прийняття обґрунтованих рішень, особливо в політично напружених середовищах, таких як вибори. Цей злий намір може дестабілізувати суспільство, впливаючи на громадську думку та результати демократичних процесів. Тому, оскільки

механізми розповсюдження стають все більш складними з появою штучного інтелекту та інших технологій, боротьба з дезінформацією стає все більш складною, що вимагає пильності та інноваційних рішень для збереження цілісності інформації.

Цифрова трансформація сучасних підприємств є актуальною та важливою проблемою. Зростаюча кількість даних, їх обробка та аналіз вимагають нових підходів та інструментів. Одним з таких інструментів є штучний інтелект, який може виявляти підозрілі публікації та дезінформацію на соціальних мережах, зокрема на платформі Facebook.

Дослідники вже вивчають цифрову зрілість бізнесу та її вплив на ефективність підприємств. Однак, використання ШІ для аналізу даних про цифрову зрілість може допомогти виявляти недостовірні новини та дезінформацію, що може впливати на бізнес-структури [5].

Крім того, інтеграція інтелектуальних технологій у навчальний процес може покращити якість підвищення кваліфікації вчителів. Використання ШІ дозволяє аналізувати дані та виявляти дезінформацію, що може впливати на навчальні заклади [6].

Досліджуючи можливості використання штучного інтелекту для аналізу даних про цифрову зрілість та виявлення дезінформації на платформі Facebook. Також розглядає перспективи інтеграції інтелектуальних технологій у навчальний процес [7].

Спираючись на попередню дискусію про еволюцію дезінформації та її наміри ввести в оману, роль штучного інтелекту у поширенні дезінформації на таких платформах, як Facebook, стає все більш важливою. Поява складних чат-ботів зі штучним інтелектом була визначена як ключовий фактор у поширенні неправдивого контенту, зокрема тому, що вони можуть генерувати та поширювати дезінформацію в безпрецедентних масштабах і швидкості [8]. Боти здатні не тільки створювати переконливі фейкові наративи, але також можуть адаптувати свої повідомлення до конкретної аудиторії, що ускладнює виявлення фейків і боротьбу з ними. Проблема ще більше погіршуються тим фактом, що системи штучного інтелекту можуть навчатися на основі взаємодії користувачів,

постійно підвищуючи свою ефективність у поширенні дезінформації [9]. У відповідь на цю зростаючу загрозу такі організації, як «Internews Ukraine», активізують зусилля з навчання та озброєння медіа-професіоналів навичками, необхідними для виявлення та протидії дезінформації в соціальних мережах [10]. Програми навчання зосереджені на аналізі даних і використанні цифрових інструментів для ідентифікації створеного штучним інтелектом контенту, що є обов'язковим для підтримки цілісності інформації на платформах соціальних мереж. Крім того, такі організації, як Європейська комісія, розглядають кроки щодо впровадження спеціального маркування контенту, створеного штучним інтелектом, таким чином забезпечуючи потенційний механізм попередження користувачів про штучне походження інформації, з якою вони стикаються [11]. Цей багатогранний підхід, що включає як технологічні рішення, так і людський досвід, має важливе значення в триваючій боротьбі з дезінформацією, яку спрощує ШІ, у Facebook і за його межами.

1.2 Роль штучного інтелекту в ідентифікації фейкових новин

Штучний інтелект відіграє ключову роль у стратегії Facebook щодо боротьби з поширенням дезінформації. Одним з ключових аспектів інтеграції штучного інтелекту є його зосередженість на лінгвістичних моделях, характерних для фейкових новинних статей. Система прискіпливо перевіряє текст на виявлення ознак неправдивих повідомлень, таких як сенсаційна лексика чи відсутність надійних джерел, які часто супроводжують оманливий вміст [12]. На додаток до лінгвістичного аналізу, керовані ШІ механізми виявлення Facebook критично оцінюють достовірність джерел, на які посилаються в новинах. Вивчаючи походження інформації, штучний інтелект може оцінити, чи часто базове джерело поширює сумнівний або оманливий вміст, тим самим позначаючи потенційні фейкові новині статті для подальшого розгляду. Крім того, можливості штучного інтелекту поширюються на навчання з великого набору даних. Ця значна підготовка дозволяє штучному інтелекту розпізнавати закономірності та аномалії, які можуть вказувати на наявність пропаганди чи

неправдивої інформації, тим самим підвищуючи його ефективність у підтримці цілісності інформації, що поширюється на платформі [13].

Виявлення фейкових новин, як зазначено, полягає в його комплексному підході до вирішення поширеної проблеми дезінформації шляхом застосування методів обробки природної мови. Відрізняється створенням повного конвеєрного проекту, який включає різноманітні інструменти та фреймворки, такі як Django, а також використання регулярних виразів і функцій керування часом для ефективної обробки та класифікації статей новин на «Правдиві» та «Фейкові» категорії на основі їх змісту На відміну від подібних досліджень, таких як Conroy et al., який зосередився на теоретичній основі виявлення фейкових новин, або Gravanis et al. (2019), який застосовував алгоритми машинного навчання без повного конвеєрного підходу, це дослідження не лише стосується технічних аспектів виявлення фейкових новин, але й підкреслює важливість структурованого та масштабованого налаштування проекту для реальних додатків. Крім того, методологічне включення окремих наборів даних для «правдивих» і «неправдивих» новинних статей з метою навчання та тестування пропонує чіткий прагматичний підхід до підвищення точності моделі в розрізненні справжніх новин від фейкових, підкреслюючи практичні проблеми в боротьбі з дезінформація [2]. Цей детальний і практичний підхід дає унікальне розуміння складності виявлення фейкових новин, створюючи прецедент для майбутніх досліджень у цій галузі.

Спираючись на досягнення ШІ у сфері виявлення фейкових новин, Міністерство закордонних справ і цифрового управління Греції очолює ініціативу з розробки спеціалізованої платформи ШІ, спрямованої на стримування поширення дезінформації. Очікується, що ця платформа використовуватиме розширені можливості штучного інтелекту для ретельного вивчення контексту, в якому представлені новини, тим самим підвищуючи точність ідентифікації фейкових новин. Важливість контексту для розуміння контенту важко переоцінити, оскільки він значно посилює арсенал уряду проти поширення неправдивої інформації. З ефективністю виявлення упереджених і фейкових новин у 65-70% система є значним кроком у боротьбі з

маніпулюванням громадською думкою. Ця ініціатива відображає зусилля таких систем, як Sentinel, яку взяли на озброєння ключові інституції по всій Європі завдяки її вмінню протидіяти загрозам, створеним складними дипфейками та іншими формами цифрової дезінформації. Sentinel працює, дозволяючи надсилати цифрові медіафайли через інтерфейс веб-сайту, які потім ретельно аналізуються штучним інтелектом, щоб переконатися в їх автентичності. Ця захисна платформа Sentinel на базі штучного інтелекту є свідченням дедалі більшої ключової ролі, яку ШІ відіграє у захисті цілісності інформації в сучасну цифрову епоху [14].

У боротьбі з фейковими новинами системи штучного інтелекту використовують складні методи, щоб відрізнити справжній вміст від підробок. Однією з основних стратегій є використання потужності алгоритмів машинного навчання для аналізу величезної кількості даних, встановлення чітких зв’язків у тексті. Цей аналіз є критично важливим, оскільки штучному інтелекту необхідно вивчити приблизно 150 новин, щоб розпізнати закономірності та аномалії, які можуть вказувати на дезінформацію . Крім того, дослідники відточують постійні особливості дипфейків специфічного типу синтетичних носіїв, які часто використовують для створення фейкових новин. Відкриття в цій галузі свідчать про те, що, зосереджуючись на цих узгоджених ознаках, штучний інтелект може значно підвищити довгострокові можливості виявлення цих оманливих матеріалів [15]. Однак складність завдання ускладнює еволюцію фейкового контенту, що вимагає постійної адаптації та вдосконалення інструментів ШІ. Контекстне розуміння, критичний компонент штучного інтелекту, також використовується для підвищення ефективності державних органів у перевірці достовірності контенту новин, потенційно захищаючи цілісність публічної інформації . Тим не менш, оскільки штучний інтелект прагне доповнити зусилля журналістів у цій галузі, очевидно, що підходу, керованого виключно штучним інтелектом, ще недостатньо для повного викорінення фейкових новин, що підкреслює необхідність комбінованої стратегії людини та штучного інтелекту.

Серед безлічі систем штучного інтелекту, розроблених для боротьби з розповсюдженням фейкових новин, Grover виділяється як особливо ефективний

інструмент. Використовуючи величезний набір даних із 120 гігабайт справжніх новинних статей, який був ретельно навчений розрізняти справжні новини та сфабриковані історії. Ця обширна підготовка надала Grover можливість визначати нюанси та закономірності, характерні для фейкових новин, що дозволяє йому виявляти такий вміст із вражаючою точністю понад 92%. Крім того, оцінки роботи Grover показали, що система ШІ часто дає результати, які вважаються більш надійними, ніж результати деяких традиційних ЗМІ [15]. Це вказує на потенціал штучного інтелекту не тільки доповнювати, але й покращувати процеси перевірки фактів, які є важливими для підтримки цілісності інформації, що поширюється серед громадськості.

У сфері виявлення глибоких фейків різноманітні компанії та дослідницькі групи використовують різноманітні методи штучного інтелекту, кожна з яких адаптована до унікальних проблем, які створює ця нова форма цифрового обману. Dessa є фаворитом у вдосконаленні методів виявлення, які є достатньо надійними, щоб ретельно перевіряти відео, що циркулюють в Інтернеті, де найчастіше поширюються дипфейки та можуть мати значний вплив. Цей підхід контрастує з методами, розробленими компанією Google, яка внесла свій внесок у боротьбу з глибокими фейками, надаючи попередньо скомпільовані набори даних, спеціально розроблені для навчання та тестування алгоритмів виявлення глибоких фейків. Набори даних мають вирішальне значення для тих, хто працює над «криміналістикою обличчя», галуззю дослідження, яка зосереджена на виявленні невеликих розбіжностей і порушень у відео, які можуть вказувати на підробку. Суть проблеми полягає в змагальному характері цих технологій; з одного боку, перед складними нейронними мережами поставлено завдання генерувати все більш переконливі штучні обличчя, тоді як з іншого, настільки ж просунуті мережі навчаються розпізнавати тонкі сигнали, які сигналізують про автентичність або фальшивість обличчя [16]. Ця гонка технологічних озброєнь погіршується ще й тим фактом, що багато сучасних дипфейків створюються за допомогою альтернативних методів, крім генеративних змагальних мереж (GAN), які спочатку популяризували це явище. У результаті системи виявлення мають постійно розвиватися, адаптуватись до ландшафту, де інструменти та

методи, що використовуються для створення глибоких фейків, постійно змінюються.

Обмеження та виклики, з якими стикаються системи штучного інтелекту на політичній арені, багатогранні та викликають глибоке занепокоєння. Оскільки потенційне використання ШІ на президентських виборах у США у 2024 році назріває, експерти попереджають про значне збільшення використання ШІ для виробництва пропаганди, що може глибоко підірвати саму структуру демократії. Це виходить за рамки простої тактики кампанії; існує відчутний ризик того, що штучний інтелект може призвести до створення повністю штучних кандидатів, представляючи сценарій, за яким кандидатів-людей замінюють у виборчому процесі. Це може перерости в безпрецедентну ситуацію, коли вибори стануть полем битви не між людськими ідеологіями, а між конкуруючими системами штучного інтелекту, кожна з яких запрограмована на удосконалені іншої в маніпулюванні суспільним сприйняттям. Крім того, здатність ШІ розуміти та потенційно контролювати людські емоції становить ще один рівень складності, що загрожує довірі до результатів виборів [17]. Такі технологічні досягнення можуть змусити виборців поставити під сумнів автентичність того, що вони бачать і почують, що потенційно призведе до повсюдної втрати довіри до виборчого процесу та його результатів. Виклики підкреслюють нагальну потребу в нормативних актах і запобіжних заходах, щоб запобігти неправильному використанню штучного інтелекту в критично важливих демократичних процесах, гарантуючи, що технології підтримують, а не підривають волю людей.

1.3 Аналіз існуючих досліджень та підходів боротьби з дезінформацією

У безперервній боротьбі з дезінформацією Facebook використовує різноманітні методи штучного інтелекту, щоб визначити правдивість вмісту, який поширюється на його платформі. Одним із таких методів є крос-модальна перевірка вмісту, яка аналізує розбіжності між текстом і супровідними зображеннями або відео, щоб виявити невідповідності, які можуть вказувати на фейкові новини. Це особливо ефективно для виявлення вигадок, де текстова

інформація не відповідає візуальному контексту. Крім того, Facebook використовує розвінчання мікротаргетингу, техніку, яка передбачає використання штучного інтелекту для відстеження та боротьби з поширенням фейкових новин серед певних демографічних груп або груп користувачів, які, ймовірно, можуть бути піддані або піддаватимуться впливу. Цей індивідуальний підхід не тільки допомагає безпосередньо протистояти поширенню дезінформації, але й допомагає зрозуміти моделі розповсюдження фейкових новин серед різних сегментів бази користувачів. Крім того, комплексний аналіз соціальних медіа з використанням розширених алгоритмів машинного навчання, таких як обробка природної мови і опорні векторні машини, дозволяє здійснювати широкомасштабний моніторинг і оцінку новин, щоб відфільтрувати ті, які, ймовірно, є неправдивими. [18]. Складні методи ШІ є невід'ємною частиною багатогранної стратегії Facebook щодо підтримки цілісності інформації в соціальній мережі. На рисунку 1.1 подано архітектурний шаблон MVC, який використовується у процесі планування та створення програмного забезпечення.

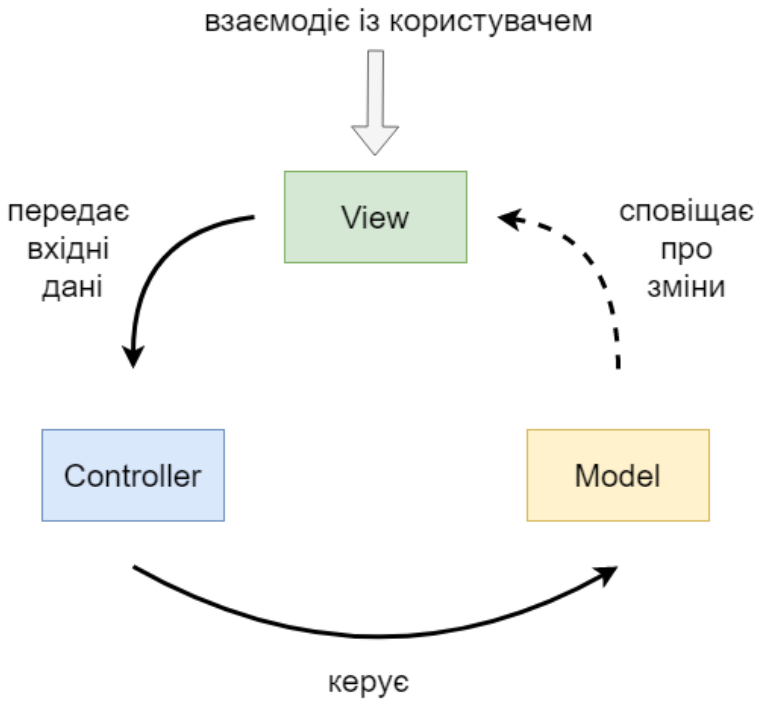

Рисунок – 1.1 Діаграма взаємодії між компонентами шаблону MVC

Роль Explainable AI (XAI) у контексті перевірки фактів важко переоцінити, особливо коли йдеться про підтримку цілісності демократичних процесів. XAI переслідує кілька цілей, серед яких прозорість, причинно-наслідковий зв'язок, конфіденційність, справедливість, довіра, зручність використання та надійність. Це продемонстровано на рисунку 1.2.

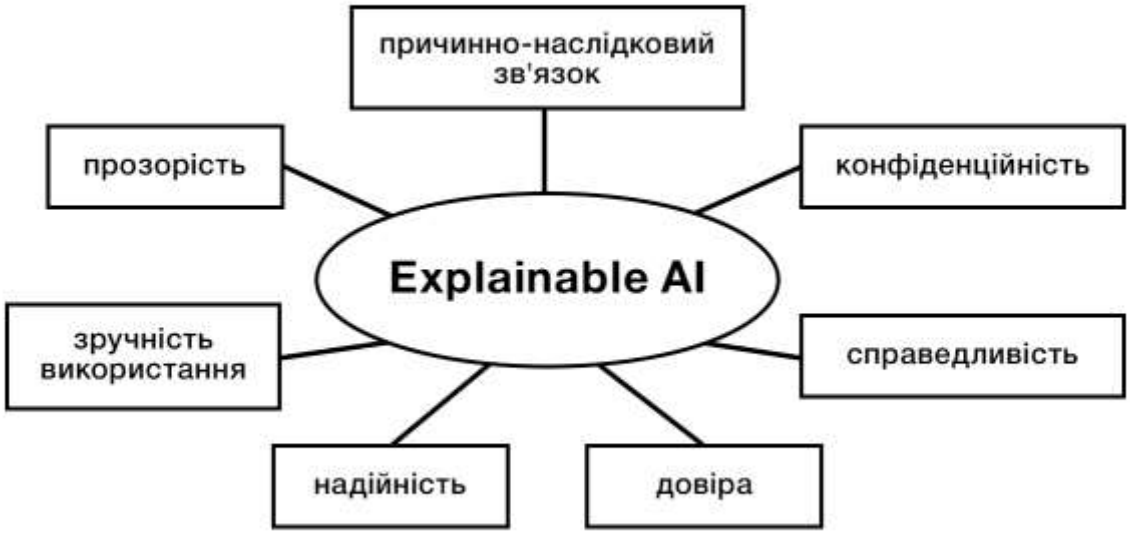

Рисунок 1.2 – Переваги використання XAI

Роль пропонує вікно в міркування алгоритмічних рішень, дозволяючи користувачам зрозуміти і довіряти правдивості інформації, яка їм надається. Ця прозорість має вирішальне значення, коли ШІ використовується як сторожовий пес під час виборів, коли поширення неправдивої інформації може мати значні наслідки. Спостерігачі штучного інтелекту, оснащені передовими алгоритмами машинного та глибокого навчання, старанно виявляють, аналізують і нейтралізують кампанії з дезінформації, забезпечуючи таким чином збереження чесності виборів. Спостерігачі використовують можливості обробки природної мови для аналізу та контекстуалізації даних, дозволяючи детально виявляти та протидіяти неправдивій інформації, що є важливим для захисту демократичних процесів. NLP включає в себе п’ять основних компонентів, які подано на рисунку 1.3.

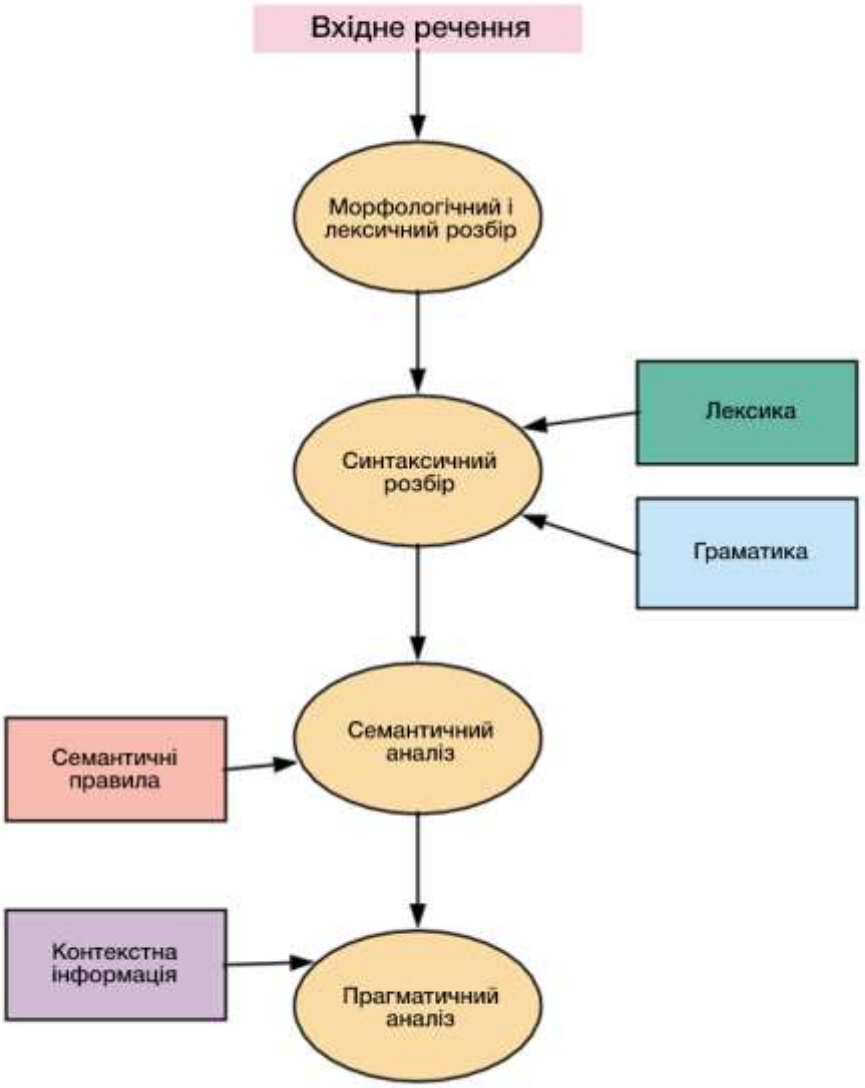

Рисунок 3.3 – Основні складові NLP

Ефективність NLP збільшується завдяки його здатності постійно відстежувати та адаптуватися до нових тактик, які використовують зловмисники, гарантуючи, що навіть коли зловмисники розвивають свої стратегії, системи ШІ залишаються на крок попереду в захисті правди. Разом ці методи штучного інтелекту, підкріплені XAI, утворюють потужний бар'єр проти розповсюдження дезінформації, особливо на чутливій арені політичних виборів [18].

Серед динамічного поля інформаційної битви дослідники стикаються зі значними проблемами під час розгортання штучного інтелекту для критичного завдання виявлення дезінформації. Основна проблема полягає в залежності

штучного інтелекту від баз даних, які можуть не містити останньої оновленої інформації, залишаючи прогалину для нової дезінформації, яка прослизає непоміченою. Крім того, системи штучного інтелекту, стаючи дедалі складнішими, ще не досягли того етапу, коли вони можуть повністю витіснити детальні аналітичні можливості людей-журналістів. Це обмеження погіршується тим фактом, що штучний інтелект може бути введений в оману логічними помилками; одна помилка в міркуванні може змусити штучний інтелект робити неправильні висновки, що призводить до більших відхилень. Неточності є особливо проблематичними в складних сценаріях, що вимагають багатоетапного обґрунтування, коли спостерігалося, що моделі штучного інтелекту «галюцинують» або фабрикують факти, ще більше мутять воду правди. Ця тенденція штучного інтелекту робити помилкові висновки, навіть з успішними моделями, є яскравим нагадуванням про поточні обмеження технологій у сфері достовірності. Незважаючи на те, що алгоритми штучного інтелекту продовжують розвиватися щодня, завдання виявлення дезінформації залишається для дослідників лабіринтом, що вимагає пильності та постійного вдосконалення можливостей штучного інтелекту [19].

Дослідження ролі штучного інтелекту в розповсюдженні та пом’якшенні дезінформації у Facebook призвело до критичних висновків, відображених у значному корпусі наукової літератури. Одним із головних висновків є ідентифікація різних методів, які ШІ може використовувати для виявлення та протидії неправдивій інформації, які поділяються на чотири групи: методи, засновані на наукових знаннях, соціологічні підходи, психологічні тактики та математичні та статистичні методи. Кожна група представляє унікальний ракурс, під яким штучний інтелект аналізує дані, щоб виявити моделі, що вказують на дезінформацію. У свою чергу, соціологічні методи можуть передбачати вивчення поширення інформації через соціальні мережі, тоді як математичні та статистичні методи можуть включати використання алгоритмів для виявлення відхилень у даних, які свідчать про маніпулятивний контент. Крім того, оцінка поточного стану штучного інтелекту в боротьбі з дезінформацією у Facebook була значно покращена завдяки всебічному огляду вітчизняної та міжнародної

науково-технічної літератури, а також патентним пошуком, які висвітлюють практичні рішення, які вже впроваджені, та ті, що знаходяться в розробці. Цей масив робіт надає панорамний огляд інструментів і методів штучного інтелекту, які виявилися ефективними, а також областей, де необхідні подальші інновації, щоб випередити тих, хто генерує та поширює дезінформацію [20].

Спираючись на використання методів штучного інтелекту, таких як крос-модальна перевірка вмісту на таких платформах, як Facebook, очевидно, що потрібен систематичний підхід до виявлення прогалин у виявленні дезінформації. Одним із значних прогалин у дослідженнях є необхідність всебічного моніторингу та збору даних протягом життєвого циклу інформації для ефективного відстеження еволюції та поширення дезінформації. Оцінка поточного стану виявлення дезінформації на основі штучного інтелекту потребує ретельного аналізу як вітчизняної, так і зарубіжної наукової літератури, а також патентних пошуків, щоб забезпечити не лише розуміння проблеми, але й розробку практичних рішень. Крім того, для підвищення надійності цих методів виявлення вкрай важливо проаналізувати характеристики різних методів дослідження, включаючи соціологічні, психологічні та математичні та статистичні методи. Цей мультидисциплінарний підхід забезпечує більш детальне розуміння проблеми, дозволяючи розробити більш просунуті інструменти штучного інтелекту, які можуть адаптуватися до складної тактики, яку використовують розповсюджувачі фейкових новин.

Щоб спиратися на основу, закладену Explainable AI (XAI) для підвищення довіри через прозорість, майбутні дослідницькі зусилля повинні зосередитися на комплексній оцінці поточного стану виявлення дезінформації. Це передбачає ретельний аналіз як вітчизняної, так і зарубіжної науково-технічної літератури, а також патентний пошук для виявлення стратегій, які практично реалізовані, та переконання в ефективності існуючих методів [21]. Така оцінка висвітлила б недоліки та сильні сторони поточного ландшафту, забезпечивши, щоб наступні дослідження ґрунтувалися на добре поінформованому розумінні поточної проблеми. Крім того, вкрай важливо, щоб це дослідження підтримувалося надійними методами наукового дослідження. Класифікуючи ці методи за

групами, наукові знання, соціологічні, психологічні та математичні та статистичні методи, можна культивувати багатогранний підхід, який є важливим для боротьби зі складною природою дезінформації в соціальних мережах. Цей структурований підхід до дослідницьких методів не тільки підвищить точність виявлення дезінформації, але й сприятиме постійному моніторингу та збору даних у всьому спектрі пластикового життєвого циклу дезінформації – від її створення, поширення до кінцевого впливу на громадська думка [22]. Таким чином, майбутні дослідження надають необхідні інструменти для адаптації та вдосконалення систем перевірки фактів на основі штучного інтелекту, що зрештою підвищить стійкість демократичних процесів проти поширеної загрози дезінформації.

У сфері стратегій боротьби з дезінформацією використання досягнень штучного інтелекту є багатообіцяючим шляхом удосконалення поточних методів. Подібно до того, як великі сховища даних про споживання газу зробили революцію в застосуванні методів машинного навчання для уточнення прогнозів використання газу та ефективності роботи, подібні принципи можна застосувати для боротьби з дезінформацією. Використання алгоритмів прийняття рішень, наприклад, продемонструвало значний потенціал у допомозі операторам робити обґрунтований вибір щодо оптимізації процесів споживання газу. Алгоритми продемонстрували свою цінність, спрощуючи моніторинг використання газу, скорочуючи витрати електроенергії та зменшуючи погіршення стану навколишнього середовища.

Крім того, штучний інтелект, зокрема завдяки використанню нейронних мереж, продемонстрував надзвичайну здатність у вирішенні проблем, пов’язаних із забезпеченням надійності систем електропостачання та аналізом моделей споживання. В академічному та технологічному дискурсі питання якості електропостачання та аналізу споживання традиційно розглядаються як окремі та окремі галузі. Однак ці сфери нерозривно пов’язані між собою та мають однакову вагу в гарантуванні надійної роботи інституційних структур використання електроенергії [23].

Нейронні мережі відмінно справляються з широким спектром завдань, включаючи класифікацію, регресію, кластеризацію, асоціацію тощо, використовуючи різні архітектурні конструкції, такі як згорткові, рекурентні, графові, капсульні, генеративні змагальні мережі (GAN) і механізми уваги. Ця універсальність дозволяє адаптувати обробку результатів відповідно до конкретних характеристик даних і вимог завдання [24].

Крім того, нейронні мережі мають здатність моделювати та обробляти різні циклічні сигнали за допомогою цифрових систем, що є ключовим аспектом аналізу сигналів. Моделювання служить критичним інструментом для оцінки ефективності як існуючих, так і нових методів циклічної обробки сигналів на різних етапах, включаючи аналіз і тестування. Крім того, імітація циклічних сигналів полегшує перенавчання нейронних мереж, тим самим підвищуючи їх продуктивність і точність.

Завдяки цим прикладам стає зрозуміло, що застосування штучного інтелекту, зокрема нейронних мереж, виходить за рамки традиційних рамок, пропонуючи інноваційні рішення давніх проблем. У контексті дезінформації потенціал штучного інтелекту аналізувати величезні набори даних, виявляти закономірності та прогнозувати тенденції може мати ключове значення для розробки ефективних стратегій виявлення та протидії неправдивій інформації, тим самим захищаючи цілісність публічного дискурсу [25].

Подібні дослідження в галузі штучного інтелекту включають роботу S.J. Вуллі, який досліджує алгоритмічне виявлення дезінформації за допомогою методів машинного навчання, і дослідження Л. Ф. Насіменто, зосереджуючись на алгоритмах обробки природної мови (NLP), щоб зрозуміти нюанси та шаблони, характерні для поширення фейкових новин. Ці дослідження разом сприяють ширшому розумінню того, як ШІ може бути потужним інструментом у боротьбі з дезінформацією на таких платформах, як Facebook, демонструючи багатогранне застосування штучного інтелекту для збереження цілісності інформації в цифровому суспільстві.

1.4 Виклики та обмеження сучасних систем виявлення дезінформації

Серед набору технологій штучного інтелекту, які використовує Facebook, однією з найважливіших програм є раннє виявлення потенційних загроз, які можуть порушити цілісність інформації, що поширюється на платформі. Системи раннього виявлення, керовані штучним інтелектом, спеціально розроблені для виявлення та пом'якшення ризиків, пов'язаних з кампаніями в соціальних мережах, особливо тими, які можуть включати поширення неправдивої інформації чи пропаганди . Технології не лише є ключовими для захисту цифрової екосистеми, але також служать основою для більш широких ініціатив, спрямованих на освіту громадськості та формування правових рамок для вирішення нових проблем, таких як дипфейки та інші форми цифрових маніпуляцій. У тандемі з цими можливостями виявлення Facebook реалізував удосконалені алгоритми машинного навчання, які відіграють важливу роль у протидії маніпуляціям у соціальних мережах шляхом аналізу шаблонів і поведінки, що вказують на неавтентичні або скоординовані кампанії. Проактивний підхід до виявлення загроз та управління ними підкреслює прагнення Facebook підтримувати надійну та безпечну платформу для своїх користувачів, що є ключовим компонентом, враховуючи широкий охоплення та вплив платформи [26].

У боротьбі з дезінформацією системи штучного інтелекту піднялися як пильні захисники, використовуючи складні методи для виявлення та пом'якшення поширення неправди. Системи, розроблені з можливістю аналізу величезної кількості даних, використовують обробку природної мови для ретельного вивчення семантики та контексту в тексті, дозволяючи їм розрізняти шаблони, які можуть вказувати на дезінформацію [26]. Примітним прикладом цього застосування є розгортання в США системи штучного інтелекту, яка активно виявляє та каталогізує зусилля Росії з дезінформації в Інтернеті. Крім того, ці спостерігачі ШІ не працюють ізольовано; вони втілюють проактивних агентів у забезпеченні цілісності демократичних процесів, таких як вибори, шляхом постійної адаптації до нових тактик дезінформації. Розгортання

алгоритмів машинного навчання має тут вирішальне значення, оскільки вони дозволяють системам штучного інтелекту навчатися на нових даних і покращувати свої можливості виявлення з часом, таким чином формуючи стійкий механізм захисту від складного та постійно мінливого ландшафту дезінформації. Ця технологічна пильність у поєднанні з непохитною відданістю етичним практикам підкреслює необхідність роботи систем штучного інтелекту в межах підзвітності, гарантуючи, що боротьба з дезінформацією випадково не порушує свободу слова [27].

Ефективність штучного інтелекту в боротьбі з дезінформацією додатково підкреслюється проактивними ініціативами великих технологічних компаній і спільними зусиллями міжнародних партнерів. Рішення Meta зібрати спеціалізовану команду, спрямовану на протидію дезінформації та зловживанням генеративним штучним інтелектом перед виборами, є свідченням визнання індустрією загрози, яку становлять складні кампанії з дезінформації. Цей крок також відображає ширшу тенденцію, коли організації визнають необхідність передових інструментів для виявлення та пом’якшення поширення неправдивих наративів. Такі команди мають вирішальне значення для розробки та реалізації керованих ШІ стратегій для виявлення та придушення шкідливого контенту, тим самим захищаючи цілісність публічного дискурсу. Подібним чином бажання ділитися досвідом і стратегіями боротьби з дезінформацією, як висловив Романишин, ілюструє дух співпраці, який є важливим для міжнародних партнерів для ефективного вирішення глобальної проблеми дезінформації. Цей спільний підхід не обмежується інституційними зусиллями, але також поширюється на громадськість, яка має бути добре поінформована про інструменти ШІ, які є в їхньому розпорядженні. Люди повинні розуміти функціональні можливості цих систем штучного інтелекту, як отримати до них доступ і до кого звертатися за допомогою, що підкреслює важливість прозорості та освіти в ширшій стратегії боротьби з дезінформацією, створеною ШІ [28].

Однією з ключових проблем, з якими стикається штучний інтелект при розрізненні законної інформації від дезінформації, є витонченість маніпулятивних повідомлень, створених спеціально для конкретних соціальних

груп. Повідомлення часто неможливо відрізнити від автентичних повідомлень, проблема погіршуються тим фактом, що дезінформаційні компанії, такі як ті, що спостерігалися під час військових конфліктів, таких як російсько-українська війна, стають все більш автоматизованими. Ця автоматизація дозволяє поширювати дезінформацію із загрозливою швидкістю та масштабом. Крім того, алгоритми, що керують цими системами штучного інтелекту, настільки вправні в персоналізації контенту, що можуть доставляти маніпулятивні повідомлення саме цій аудиторії в найвигідніший момент. Це не лише підриває демократичні процеси, вводячи в оману обраних офіційних осіб щодо справжніх поглядів їхніх виборців, але й поляризує суспільство, радикалізуючи певні групи та підриваючи довіру до ЗМІ. Таким чином, перед штучним інтелектом стоїть подвійний виклик: розвивати свою здатність виявляти ці дедалі складніші та цілеспрямовані кампанії дезінформації та робити це таким чином, щоб не відставати від швидкого поширення дезінформації, характерного для сучасних військових конфліктів [29].

Незважаючи на вдосконалені алгоритми, які використовуються такими платформами, як Facebook, для адаптації контенту та виявлення шахрайських дій, властиві обмеження ШІ значно перешкоджають його здатності ефективно боротися з неправдивою інформацією. Однією з головних проблем є схильність штучного інтелекту генерувати переконливий текст і зображення, які можна легко використати для кампаній з дезінформації [30]. Ця здатність не тільки дозволяє створювати неправдиві наративи, але й ускладнює завдання відрізнити автентичний вміст від вигадок, таким чином стираючи межі між фактом і вигадкою та ставлячи під загрозу довіру до справжнього звіту. Хоча системи штучного інтелекту можуть швидко ідентифікувати шаблони, які можуть вказувати на скоординовані зусилля з дезінформації, перевершуючи швидкість людей-модераторів, вони також можуть піддаватися помилкам, починаючи від упередженості та галюцинацій до фундаментальних помилок здорового глузду та математики. Такі недоліки викликають особливе занепокоєння в контексті написання курсових робіт чи статей для авторитетних видань, де висока ставка на точність і надійність. Ефективність штучного інтелекту в цій області також

погіршується відсутністю стандартизованої інфраструктури, спеціально розробленої для виявлення та пом’якшення поширення неправдивої інформації, створеної ШІ. Ця прогалина в технологічній структурі підкреслює потребу в додаткових дослідженнях і потенційно фінансованих урядом ініціативах для підвищення можливостей штучного інтелекту в розпізнаванні та нейтралізації неправдивих наративів [31].

Динамічний ландшафт штучного інтелекту вимагає активного підходу до регулювання та адаптації для захисту інтересів суспільства. Для вирішення проблем, пов’язаних із швидким розвитком технологій штучного інтелекту, важливо не лише підтримувати існуючі стандарти, але й розвивати їх у відповідь на розгортання технологічного ландшафту. У цьому відношенні Європейський Союз зробив похвальний крок, працюючи над законопроектом, спрямованим на регулювання діяльності ШІ, що відображає зростаюче визнання необхідності нагляду та управління впливом ШІ на різні аспекти життя. Положення законопроекту щодо захисту конфіденційності користувачів і забезпечення прозорості джерела даних є критично важливими заходами для збереження індивідуальних прав і зміцнення довіри до систем ШІ. Крім того, обов’язкове маркування медіафайлів, створених за допомогою штучного інтелекту, є свідченням прихильності ЄС до прозорості та підзвітності, які є важливими для боротьби з поширенням дезінформації та підтримки цілісності цифрового спілкування [32]. Заходи разом із активною роллю штучного інтелекту в протидії дезінформації, як зазначено в попередньому абзаці, підкреслюють важливість багатогранного підходу до управління штучним інтелектом, де як технологічні рішення, так і потужна нормативна база працюють у тандемі для вирішення складних проблем.

1.5 Висновок до першого розділу

В першому розділі кваліфікаційної роботи освітнього рівня «Магістр» описано роль штучного інтелекту в боротьбі з дезінформацією.

Штучний інтелект все частіше використовується для боротьби з дезінформацією на платформах соціальних мереж, таких як Facebook. ШІ використовується для аналізу лінгвістичних шаблонів і оцінки достовірності джерел, цитованих у новинах, щоб ідентифікувати потенційні фейкові новині статті для подальшої перевірки. Однак одного тільки штучного інтелекту ще недостатньо для повного викорінення фейкових новин, і потрібна комбінована стратегія людини та штучного інтелекту. Grover є особливо ефективним інструментом, з точністю понад 92% у виявленні фейкових новин.

2 МЕТОДИ ТА ІНСТРУМЕНТИ ШТУЧНОГО ІНТЕЛЕКТУ ПРОТИ ДЕЗІНФОРМАЦІЇ

2.1 Основні поняття та визначення

Штучний інтелект, як галузь, охоплює низку концепцій і застосувань, які виходять далеко за межі його найпростішого визначення як комп'ютерних систем, здатних виконувати завдання, які зазвичай потребують людського інтелекту. Однією з таких програм є розробка систем штучного інтелекту, які можна запрограмувати на розпізнавання зображень або тексту, створених іншими машинами. Ця здатність має вирішальне значення в сучасному цифровому ландшафті, де генерація синтетичного контенту, ще один фундаментальний термін, пов'язаний зі штучним інтелектом, стала звичним явищем. Синтетичний контент включає засоби масової інформації, створені або змінені системами штучного інтелекту, починаючи від глибоких фейків і закінчуючи алгоритмічно згенерованими статтями, що становить серйозну проблему у вигляді дезінформації. Дезінформація, навмисно вводить в оману інформація, яка поширюється з наміром ввести в оману, є проблемою, якій штучний інтелект як сприяє, так і допомагає її пом'якшити. GPTZero – це програма, розроблена для оцінки ймовірності того, що даний текст був складений штучним інтелектом, з метою вирішення проблеми штучного інтелекту, яка вводить в оману, шляхом виявлення нелюдських шаблонів у тексті [33]. Аспекти штучного інтелекту ілюструють складну та дихотомічну природу технології, де вона одночасно є інструментом і викликом у сфері автентичності інформації.

У сфері штучного інтелекту дезінформація набуває складного характеру, оскільки здатність технології створювати та поширювати оманливий контент перевершує традиційні методи. Незважаючи на визнання цієї проблеми, дане дослідження не дає конкретного визначення дезінформації в контексті штучного інтелекту, залишаючи прогалину в розумінні цього цифрового явища. Ця відсутність чіткого визначення є критичною, оскільки технології штучного

інтелекту мають потенціал для обмеження дезінформації, коли неправда поширюється із загрозливою швидкістю, охоплюючи широку аудиторію з безпрецедентною ефективністю. Такі шторми можуть спровокувати як агенти-людини, так і складні мовні моделі, порівняння, яке розглядається в дослідженні, хоча й без зупинки на точній термінології для окреслення дезінформації, створеної ШІ [34]. Проте зусилля щодо протидії цим штормам тривають, про що свідчить ініціатива Державного департаменту щодо використання агрегатора контенту на основі штучного інтелекту, зосередженого на Україні, який має на меті просіювати інформацію, щоб висвітлити перевірені випадки російської дезінформації. Ця система є прикладом подвійної ролі штучного інтелекту у виявлені дезінформації ландшафті як детектора, причому остання функція підкреслюється здатністю штучного інтелекту надійно ідентифікувати вигадки, створені російськими чат-ботами. Ця подвійність підкреслює необхідність тонкого розуміння терміну «дезінформація» в контексті штучного інтелекту, який є не просто академічною вправою, а необхідною умовою для розробки ефективних заходів протидії цифровому обману.

Розуміння ролі штучного інтелекту у поширенні дезінформації має вирішальне значення через здатність ШІ створювати переконливий і, здавалося б, достовірний контент, який може вводити громадськість в оману. Оскільки штучний інтелект стає все більш досконалим, він може створювати текст, аудіо та зображення, які все важче відрізнити від контенту, створеного людьми, що збільшує ризик дезінформації. Це не лише потенційно може підірвати довіру до авторитетних джерел, але й може маніпулювати громадською думкою, настроями та навіть втручатися в демократичні процеси, такі як вибори. Для вирішення цих проблем надзвичайно важливо розробити інструменти та стратегії, які можуть виявляти та протидіяти дезінформації, створеній ШІ. Інструменти мають бути достатньо складними, щоб ідентифікувати «цифрові сліди» або невідповідності, залишені штучним інтелектом, такі як незвичайне формулювання тексту чи артефакти на зображеннях і відео, які можуть свідчити про контент, створений ШІ. Оснащуючи себе здатністю визнавати роль

штучного інтелекту в дезінформації, ми надаємо людям і установам можливість підтримувати цілісність інформації та захищати демократичні цінності[35].

2.2 Дослідження методів ML та NLP

В основі сучасних програм штучного інтелекту лежать складні алгоритми, які сприяють виконанню безлічі функцій, починаючи від повсякденних і закінчуючи складними. Алгоритми розвинулися настільки, що вони не тільки обробляють величезні обсяги даних, але роблять це з надзвичайною ефективністю та точністю. У сфері безпеки штучний інтелект продемонстрував свою майстерність, використовуючи технологію розпізнавання обличчя для швидкого і точного виявлення злочинців у кримінальних справах, таким чином революціонізувавши спосіб управління безпекою об’єктів. Універсальність штучного інтелекту додатково підтверджується його здатністю виконувати такі творчі завдання, як малювання, письмо та навіть перевірка граматики, демонструючи його адаптивність у різних областях. Крім того, основні методи ШІ не обмежуються одним підходом; натомість вони охоплюють спектр алгоритмів навчання, кожен з яких унікально розроблений для вирішення конкретних типів проблем у галузі, будь то через контрольоване чи неконтрольоване навчання, навчання з підкріпленням або складні рівні моделей глибокого навчання. Такі децентралізовані моделі навчання прокладають шлях для автономних систем, які зміцнюють локальний інтелект, забезпечуючи застосовність ШІ в різних вертикалях і його здатність аналізувати інформацію, генерувати контент і давати рекомендації з високим ступенем точності [36].

Застосування та функціональність методів штучного інтелекту істотно відрізняються, особливо якщо взяти до уваги взаємодію між штучними нейронними мережами та новим використанням блокчейну в просторі штучного інтелекту. Штучні нейронні мережі, модель глибокого навчання на передньому краї розвитку ШІ, залежать від їх здатності навчатися з об’ємних наборів даних, процес, який потребує значних обчислювальних ресурсів [36]. Це інтенсивне навчання дозволяє нейронним мережам виконувати широкий спектр складних

завдань, починаючи від розпізнавання зображень і мови до прогнозування поведінки споживачів. Навпаки, конвергенція ШІ з технологією блокчейн вводить нову парадигму, наголошуючи на безпеці та надійності в обробці даних. Система розподіленої книги Blockchain пропонує незмінний запис транзакцій, що сприяє покращенню процесу аудиту мережі, тим самим підвищуючи безпеку та конфіденційність. Крім того, здатність блокчейну безпечно зберігати інформацію про запити та взаємодії в системі ШІ підвищує надійність і підзвітність цих технологій. Дійсно, оскільки галузь зосереджена на обробці постійно зростаючих обсягів даних і збільшенні обчислювальної потужності, прийняття користувачами підключених систем штучного інтелекту залишається першорядним [37]. Цей подвійний фокус на вдосконаленні обчислювальних можливостей і забезпеченні безпечної конфіденційної обробки даних відображає складну, багатогранну природу програм ШІ та їх функціональність, що розвивається в сучасній технологічній екосистемі.

Спираючись на потребу у широких можливостях обробки даних в обчислювальній потужності, недавні досягнення в області штучного інтелекту справді змінили. Одним із таких нововведень є запровадження Google «Talk to Books», який використовує штучний інтелект, щоб надавати користувачам відповіді на їхні запитання у формі цитат, взятих безпосередньо з книг. Це передове застосування штучного інтелекту для розуміння людських запитів і відповідей на них демонструє прогрес, досягнутий штучним інтелектом у обробці природної мови, і його потенціал кардинально змінити спосіб пошуку інформації та взаємодії з нею. Подібним чином Replika пропонує персоналізований досвід розмови, де штучний інтелект лежить в основі створення індивідуального чат-бота для користувача, здатного брати участь в обговоренні тем, які користувачі вважають цікавими [38]. Це не тільки підкреслює успіхи, досягнуті в машинному навчанні для розуміння та моделювання людської розмови, але й підкреслює емоційний інтелект, який починають демонструвати системи ШІ. Крім того, сфера веселощів і творчості не залишається позаду. AutoDraw від Google є свідченням інноваційного використання штучного інтелекту в графічному дизайні, де він допомагає

користувачам, удосконалюючи їхні рудиментарні ескізи у відшліфовані ілюстрації, роблячи дизайн більш доступним для непрофесійного художника. На додаток до цих додатків розробка мобільних додатків все більше включає штучний інтелект для покращення взаємодії з користувачем, пропонуючи інтуїтивно зрозумілі інтерфейси та персоналізований вміст, що розширює охоплення та інтеграцію штучного інтелекту в повсякденне життя. Розробки є не просто поступовими удосконаленнями, та розглядаються як новаторські досягнення, які встановлюють нові стандарти в ландшафті ШІ.

У пошуках боротьби з поширеною проблемою дезінформації конкретні методи ШІ стали ефективними інструментами. Одним з таких підходів є використання алгоритмів класифікації, заснованих на машинному навчанні, які виявилися вправними у виявленні та фільтрації фейкових новин або чуток на платформах соціальних мереж. Ці алгоритми є частиною ширшого набору стратегій, які включають обробку природної мови, лінійну регресію, k-найближчих сусідів (KNN), метод опорних векторів (XAI) і довгу короткочасну пам’ять (LSTM), усі з яких удосконалити здатність штучного інтелекту ретельно досліджувати та розуміти текстовий вміст на ознаки фальшивості. На рисунку 2.1 проілюстровано графік порівняння ефективності моделей машинного навчання.

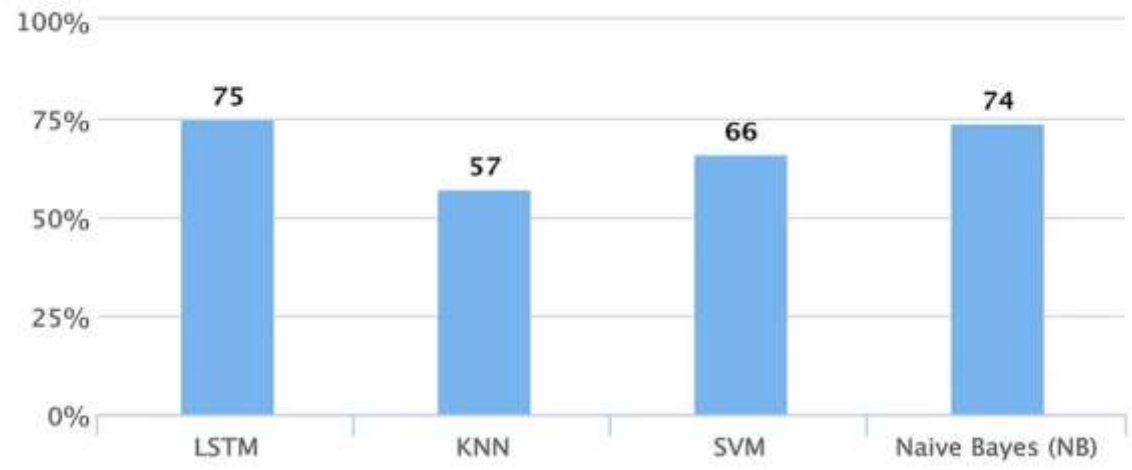

Рисунок 2.1 – Порівняльний графік ефективності моделей

Крім того, системи штучного інтелекту використовують інструменти, засновані на лінгвістиці, для виявлення розбіжностей і вигадок у тексті, часто шляхом аналізу шаблонів і аномалій, які нелегко помітити людям. Ефективність цих методів штучного інтелекту полягає не лише в їхній здатності аналізувати масивні набори даних – завдання, яке є непрактичним для людей, – а й у їхньому постійному навчанні та адаптації, що має вирішальне значення, враховуючи тактику тих, хто поширює дезінформацію, що постійно розвивається. Ця здатність до адаптації підсумовується у твердженні, що «тільки штучний інтелект може впоратися зі штучним інтелектом», наголошуючи на необхідності використання штучного інтелекту для протидії створеним штучним інтелектом або з його допомогою кампаніям дезінформації.

У пошуках неправдивої інформації методи штучного інтелекту особливо вправно орієнтуються в складності намірів, що є критичним компонентом у розрізненні дезінформації та обману. Дослідники з Дартмута розробили складні алгоритми ШІ, які аналізують текстові дані з новин, щоб з’ясувати прихований намір оратора чи автора. Алгоритми навчені розпізнавати шаблони, які дозволяють припустити, чи людина свідомо намагається обдурити аудиторію, що є необхідною умовою для обману. Ця відмінність є надзвичайно важливою, оскільки не вся неправдива інформація є результатом наміру ввести в оману. Особа може ненавмисно поширювати дезінформацію, не знаючи, що ця інформація неправдива, або може просто висловлювати суб’єктивну думку, а не стверджувати фактичні твердження. Таким чином, передові методи штучного інтелекту, які використовуються в дослідженнях Дартмутської команди, зосереджені на складному завданні розпізнати намір, що стоїть за поданою інформацією, з метою категоричного відділення навмисного обману від невинного поширення неправдивої інформації [39]. Цей тонкий підхід підкреслює важливість наміру в оцінці достовірності інформації та є значним прогресом у боротьбі з дезінформацією, керованою ШІ.

Враховуючи розширені можливості штучного інтелекту у створенні високоякісних зображень, завдання виявлення дезінформації стає дедалі складнішим. Правдоподібність цих зображень, створених штучним інтелектом,

створює величезну перешкоду, оскільки відмінність між автентичним і сфабрикованим вмістом стирається, що ускладнює ідентифікацію підробок навіть складним алгоритмом. Щодо перевірки дезінформації з таких зображень підкреслюють необхідність нюансованого підходу, який виходить за межі простого алгоритмічного аналізу та включає контекстуальну оцінку та оцінку на основі джерел. Однак, оскільки онлайн-платформи використовують механізми автоматизації у своєму прагненні керувати величезним обсягом контенту, вони ненавмисно порушують зобов'язання держави щодо захисту свободи слова. Напруга між автоматизованою модерацією вмісту та громадянськими свободами посилюється відсутністю підзвітності платформ, скрутним становищем, яке не тільки підриває довіру до цифрових посередників, але й перешкоджає ефективному виявленню дезінформації. Дані проблеми погіршується тим фактом, що алгоритми часто віддають перевагу сенсаційному контенту над центристським, роблячи останній практично непомітним і, як наслідок, ускладнюючи ідентифікацію дезінформації, сприяючи поляризаційним наративам. Той факт, що кілька платформ диктують логіку поширення інформації, ще більше ускладнює ландшафт, оскільки їхні власні алгоритми розроблені для максимального залучення, а не точності, часто за рахунок суспільної згуртованості. Щоб вирішити ці багатогранні проблеми, необхідні нормативні акти, спрямовані на підвищення прозорості та підзвітності цих платформ, оскільки вони можуть забезпечити більш надійну структуру для ШІ для ефективної боротьби з дезінформацією [40].

2.3 Аналіз інструментів виявлення дезінформації в новинах Facebook

У сфері аналізу цифрового контенту такі інструменти, як Sphere, CrowdTangle, Factmata та NewsGuard, відіграють ключову роль у просуванні величезних обсягів інформації, щоб відрізнити правду від вигадки. Спільним серед цих інструментів є їх здатність аналізувати соціальні медіа, що виявляється важливим для перевірки автентичності онлайн-контенту. Це особливо важливо в епоху, коли дезінформація може швидко поширюватися платформами. Крім

того, їх методи виходять за рамки поверхневих перевірок; вони використовують складні крос-модальні методи перевірки вмісту. Аналізуючи текст у поєднанні з зображеннями, відео та показниками залучення користувачів, ці інструменти можуть ефективно викривати навмисно чи ненавмисно сфабрикований вміст. Крім того, ці платформи вправно протидіють підступній практиці мікротаргетування. Досліджуючи шаблони та тактики, які використовуються для поширення спеціальних кампаній з дезінформації, Sphere, CrowdTangle, Factmata та NewsGuard можуть розвінчати стратегії, призначені для маніпулювання громадською думкою, таким чином підтримуючи цілісність інформації, яка протікає через цифровий ландшафт [41].

Спираючись на можливості інструментів аналізу соціальних медіа, ці механізми виявлення дезінформації глибше досліджують достовірність цифрового контенту. Такі інструменти, як Sentinel, використовують штучний інтелект для ретельного вивчення цифрового медіаландшафту, виявляючи ознаки втручання, які можуть вказувати на наявність дезінформації. Це особливо критично в епоху складних методів, таких як глибокі фейки, де зображення та відео маніпулюють із високим ступенем реалістичності, що потенційно може ввести в оману навіть найвибагливіших глядачів. Інструмент автентифікації відео Microsoft є прикладом цього підходу, надаючи оцінку автентичності в реальному часі, яка допомагає користувачам переконатися в легітимності медіафайлів, які вони споживають [41]. Ретельно вивчаючи метадані та використовуючи передові алгоритми, інструмент може виявляти невідповідності та зміни, які є ознаками підробленого контенту, який часто стратегічно розгортається як частина кампаній з дезінформації. Технологічні досягнення мають ключове значення в боротьбі з дезінформацією, оскільки вони надають окремим особам і організаціям засоби для оскарження автентичності контенту, який може бути використаний зловмисно для обману, маніпулювання громадською думкою або підриву довіри до встановлених установ.

Щоб ефективно оцінити інструменти виявлення дезінформації, необхідно встановити певні критерії, які враховують нюанси природи дезінформації. Зважаючи на те, що дезінформація часто спотворює передбачуване значення

контенту, здатність інструменту ідентифікувати та висвітлювати ці розбіжності є надзвичайно важливою. Це стає особливо складним, коли дезінформація включає зерно правди, яким маніпулюють або виривають його з контексту. Такий інструмент, як YouTube, Data Viewer, розроблений Amnesty International, міг би стати еталоном у цьому відношенні [42]. Цей інструмент допомагає журналістам та іншим слідчим відстежувати походження відео, що є важливим для оцінки достовірності вмісту та перевірки його контексту. Порівнюючи результат різних інструментів із продуктивністю YouTube Data Viewer, можна визначити їхню ефективність у виділенні та виявленні дезінформації. Крім того, ці контрольні показники також повинні враховувати здатність інструменту доповнювати існуючі стратегії, такі як попередження про дезінформацію, які використовуються платформами соціальних мереж, як обговорювалося в попередньому параграфі. Завдяки інтеграції з цими мітками та експертними консультаціями можна сформувати більш надійний і багатогранний підхід до виявлення дезінформації, що підвищить загальну точність і надійність відповідного інструменту.

Заглиблюючись у конкретні методи, які використовують інструменти порівняльного аналізу, важливо розуміти, що метод одновимірного порівняльного аналізу виділяється своїм цілеспрямованим підходом. Цей метод включає в себе пряме порівняння одного або кількох обраних індикаторів, зосереджуючись або на одному об’єкті, або порівнюючи кілька об’єктів на основі лише одного індикатора. Така техніка особливо корисна, коли метою є виділення та ретельний аналіз конкретного аспекту продуктивності без змішувальних ефектів кількох змінних. Однак важливо відзначити, що цей метод навмисно уникає складності оцінки багатьох показників для різних об’єктів, таким чином забезпечуючи чітке, хоча і вузьке, уявлення про предмет дослідження. Ця форма аналізу зазвичай використовується для оцінки ефективності виконання завдань, дотримання договірних зобов’язань або точності прогнозованих контрольних показників [43]. Застосовуючи порівняльний аналіз, компанія може порівняти свої поточні економічні показники з раніше поставленими цілями, історичними даними про ефективність

або навіть з показниками своїх конкурентів, таким чином отримуючи цінну інформацію про свою конкурентну позицію та операційну ефективність.

У сфері виявлення дезінформації такі інструменти, як Sphere, CrowdTangle, Factmata та NewsGuard, використовують потужність аналізу соціальних медіа для перевірки автентичності вмісту, але порівняльний аналіз їхньої ефективності може запропонувати глибше розуміння їхніх унікальних переваг. Надійний алгоритм Sphere які можуть ефективно розвінчувати мікротаргетинг шляхом аналізу моделей поведінки та показників залучення, таким чином пропонуючи превентивний захід проти поширення неправдивих наративів. CrowdTangle, з іншого боку, чудово відстежує неправдивий вміст на платформах, надаючи аналітикам дані в режимі реального часу для швидкого виявлення та боротьби з поширенням дезінформації. Більше того, Factmata використовує вдосконалені методи обробки природної мови для оцінки достовірності джерел новин і ймовірності того, що контент є навмисно оманливим. Таким чином, кожен інструмент вносить свій внесок у багатогранний підхід, необхідний для боротьби з дезінформацією, з їх порівняльним аналізом, що підкреслює додатковий характер їхніх функцій. Використовуючи комбінацію цих спеціалізованих інструментів, організації можуть створити більш повний і стійкий захист від маніпулювання інформацією [44].

Порівняльний аналіз методів виявлення зловмисного програмного забезпечення забезпечує важливу основу для визначення обмежень, притаманних різним інструментам, які використовуються для боротьби з поширенням дезінформації та шкідливих програм в Інтернеті. У сфері обробки зображень і інструментів комп’ютерного зору аналіз AForge.NET, MATLAB і OpenCV показує, що, хоча ці інструменти надійні у своїх відповідних можливостях, вони мають явні недоліки при застосуванні до конкретних завдань. MATLAB відомий своєю великою бібліотекою та потужним набором інструментів, але він може бути надзвичайно дорогим і не найефективнішим для обробки в режимі реального часу, що має вирішальне значення для швидкого виявлення та швидкого реагування на кампанії дезінформації. Незважаючи на те, що OpenCV має відкритий вихідний код і є універсальним, може знадобитися

глибоке розуміння програмування, щоб повністю використовувати його можливості, потенційно обмежуючи доступ до нього лише експертам. Крім того, AForge.NET, з його залежністю від .NET framework, може не забезпечити крос-платформну сумісність, необхідну для широкого використання в різноманітних обчислювальних середовищах. Обмеження підкреслюють важливість розуміння показників ефективності кожного інструменту не лише для виконання завдань і договірних зобов'язань, але й для досягнення прогнозованих показників для ефективної протидії дезінформації [45]. Розуміння нюансів, отримане в результаті такого порівняльного аналізу, є неоціненним для компаній, особливо для платформ соціальних мереж, які прагнуть покращити свої системи виявлення та забезпечити цілісність інформації, що поширюється серед громадськості.

2.4 Аналіз теоретичних моделей та експериментальних досліджень вдосконалення AI-інструментів

Експертні системи, як головна теоретична модель штучного інтелекту, є прикладом чудової здатності ШІ імітувати та потенційно замінювати досвід людини в певних областях. Системи працюють на базі знань, що відображає більш широкий принцип штучного інтелекту, згідно з яким інтелектуальні системи повинні володіти здатністю вивчати та застосовувати інформацію. Цей принцип ґрунтується на епістемологічних результатах кібернетики, які стверджують, що інтелектуальні процеси можна моделювати, якщо їх можна точно описати за допомогою обмеженого лексикону. Наслідки цього є глибокими, припускаючи, що будь-яка психічна функція, коли її однозначно визначено, теоретично може бути відтворена електронною обчислювальною машиною. Ця амбіція узгоджується з основною метою досліджень штучного інтелекту: побудувати модель мозку, яка може розкрити складні процеси мислення. Прагнення до цієї мети підтримується сучасним науковим розумінням, яке стверджує, що інформаційні процеси мозку керуються кінцевим набором правил [46]. Розробка коннекціоністських моделей, таких як нейронні мережі, які наголошують на навчанні та адаптації, і еволюційних моделей, таких

як генетичні алгоритми, є невід’ємною частиною розвитку ШІ. Модель архітектури рекурентних нейронних мереж зображено на рисунку 2.1.

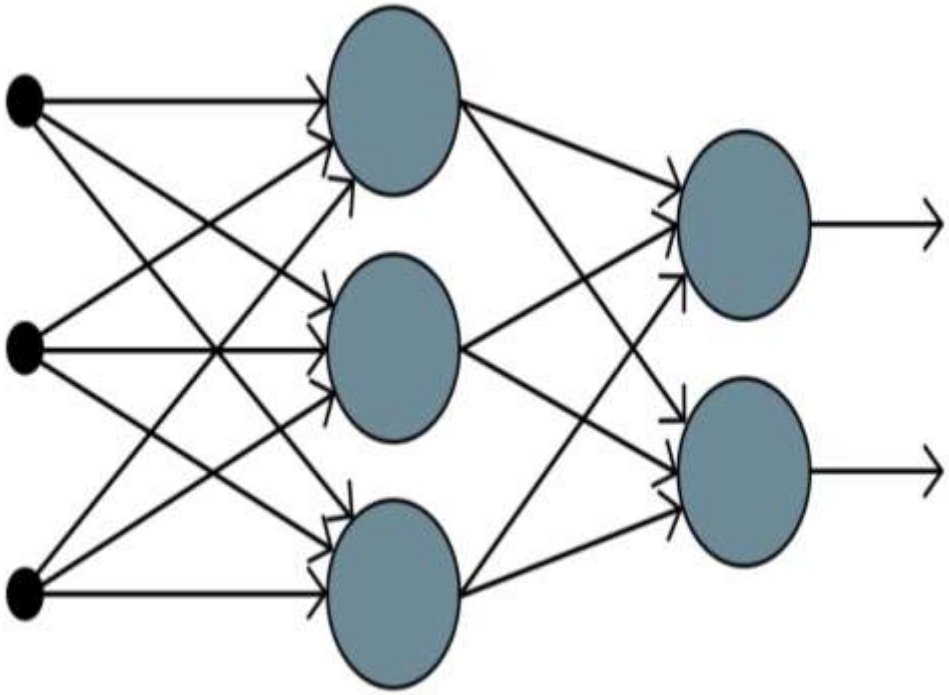

Рисунок 2.2 – Модель архітектури рекурентних нейронних мереж

Ці моделі підкреслюють важливість логічного виведення, особливо в когнітивних моделях, де мислення є центральною здатністю, що моделюється. Когнітивні моделі, з їхньою історичною популярністю в штучному інтелекті, характеризуються своїм підходом, орієнтованим на знання, ще більше посилюючи важливу роль накопичення знань і їх застосування в галузі. На рисунку 2.3 зображено модель трансформаторної нейронної мережі.

Рисунок 2.3 – Модель трансформаторної нейронної мережі

На еволюцію інструментів ШІ суттєво вплинула інтеграція різнорідних моделей, що призвело до розробки гібридних систем. Системи поєднують сильні сторони різних методів штучного інтелекту для створення більш надійних та адаптивних рішень. Поєднання нечітких експертних систем із нейронними мережами призводить до створення інструментів штучного інтелекту, які можуть справлятися з невизначеністю та вивчати дані у спосіб, який нагадує когнітивні процеси людини. Такі гібридні системи розробляються шляхом ретельного процесу інтеграції, де кожен компонент вибирається відповідно до його здатності сприяти загальній функціональності та ефективності інструменту ШІ. Продукційна модель, яка використовує набір правил для керування сортуванням правил за допомогою механізму деривації, виділяється як найпоширеніша мова представлення знань, яка використовується для розробки цих складних систем ШІ. Крім того, моделі нейронних мереж, використання яких у різних галузях штучного інтелекту значно поширилося завдяки прогресу апаратного та програмного забезпечення, є важливим компонентом у створенні цих інтегрованих систем [47]. Така синергія різних моделей не тільки розширює можливості інструментів штучного інтелекту, але й прискорює процес

проектування, роблячи його швидшим і ефективнішим, як це видно у випадку з розробниками чіпів штучного інтелекту, які використовують глибоке підсилювальне навчання, щоб перевершити експертів як у швидкості, так і в продуктивності. Таким чином, внесок цих моделей у розробку інструментів штучного інтелекту відзначається спільним підходом, який поєднує найкращі характеристики кожної моделі для вирішення складних проблем людства.

У прагненні зрозуміти обмеження поточних теоретичних моделей, вирішальним аспектом, який слід розглянути, є окремий характер проблем, які вони мають на меті вирішити. Кожна проблема в області штучного інтелекту відзначається різними рівнями загальності, абстрактності та складності, а також етапами розвитку. Це розмаїття відображає безліч викликів, з якими стикаються сучасні теоретичні моделі, оскільки вони спрямовані не на універсальний набір проблем, а на конкретні, часто ізольовані проблеми. Отже, незважаючи на те, що було досягнуто значних успіхів як у практичній, так і в теоретичній сферах, і триваючі інтенсивні дослідження дають важливі результати, ці досягнення не завжди взаємозамінні або застосовні в різних проблемних областях. Притаманні фундаментальні та практичні труднощі, які характеризують кожну проблему, свідчать про те, що універсальна теоретична модель неправдоподібна. Це ще більше погіршується тим фактом, що сфера штучного інтелекту не обертається навколо однієї, чітко визначеної проблеми, яка спрямовує траєкторію його теоретичних основ. Натомість штучний інтелект охоплює безліч спеціалізованих і вузьких проблем, кожна з яких вимагає індивідуальних рішень. Ця фрагментація означає, що хоча еволюційні алгоритми та інші класичні моделі штучного інтелекту виявилися ефективними в певних контекстах, їхня застосовність не є універсальною, таким чином підкреслюючи потребу в різноманітному арсеналі теоретичних моделей, налаштованих на особливості кожної унікальної проблеми в експансивному ландшафті штучного інтелект.

У галузі штучного інтелекту, що розвивається, експериментальні дослідження мають вирішальне значення для оцінки ефективності та практичності інструментів ШІ перед їх інтеграцією в клінічні умови. Хоча системи штучного інтелекту мають потенціал для революції в охороні здоров’я,

вони ще не стали частиною клінічної практики, оскільки їх надійність і точність повинні бути ретельно перевірені. Типи експериментальних досліджень, проведених для перевірки цих інструментів, часто відрізняються, але вони, як правило, поділяються на категорії, які оцінюють можливості пошуку та підвищення продуктивності. Дослідження, пов’язані з пошуком, можуть передбачати перевірку здатності штучного інтелекту переглядати величезні набори даних і медичні записи, щоб ідентифікувати закономірності або діагностувати захворювання швидше, ніж аналоги людини. Водночас дослідження продуктивності часто вимірюють, наскільки добре інструменти штучного інтелекту можуть автоматизувати завдання, тим самим звільняючи медичних працівників від зосередження на більш складних заходах з догляду за пацієнтами. Такі експериментальні дослідження є не лише свідченням зростаючої складності інструментів штучного інтелекту, але й необхідним кроком для забезпечення їх безпечного та ефективного застосування в реальних клінічних середовищах [48]. Дослідження дають безцінне уявлення про сильні сторони та обмеження застосування штучного інтелекту, прокладаючи шлях для їх можливого впровадження в лікарнях, де вони можуть підтримувати практикуючих лікарів і зрештою покращувати результати пацієнтів.

Постійне вдосконалення інструментів штучного інтелекту можна напряму пов’язати з експериментальними результатами, які інформують про їхню еволюцію. У сфері комп’ютерного зору, секторі штучного інтелекту, який зосереджується на тому, щоб дозволити комп’ютерам інтерпретувати та розуміти візуальну інформацію з навколишнього світу, експериментальні дослідження були особливо трансформаційними. Дослідження в цій галузі продемонстрували, що інтеграція гібридного підходу в штучний інтелект, який поєднує різні алгоритми та техніки, призвела до стрибка в здатності системи відстежувати та прогнозувати рух суб’єкта з більшою точністю. Це надзвичайно важливо для таких додатків, як автономні транспортні засоби, доповнена реальність і системи безпеки, де точне виявлення руху має першорядне значення. Крім того, той самий гібридний підхід довів свою цінність, створюючи зображення високої роздільної здатності зі сцен, які зазвичай складно

розшифрувати, що закриті туманом, дощем чи іншими формами поганої погоди [49]. Ці експериментальні результати не тільки демонструють потенціал гібридних систем штучного інтелекту, але й служать каталізатором для подальших досліджень і розробок, гарантуючи, що інструменти штучного інтелекту стають все більш досконалими та вмілими для вирішення складних реальних завдань.

Розглядаючи проблеми, пов’язані з розробкою експериментів для оцінки інструментів штучного інтелекту, важливо визнати величезну здатність штучного інтелекту генерувати безліч ідей і варіантів для розробників. Цей генеративний аспект штучного інтелекту є двосічним мечем; Хоча це надає широкий спектр можливостей, воно також ускладнює завдання розробки експериментів, які можуть точно оцінити ефективність і корисність інструментів ШІ. Творці повинні орієнтуватися в широкому масиві згенерованих ідей, щоб визначити найбільш життєздатні для подальшого розвитку та оцінки. Крім того, експериментальні дослідження, проведені для оцінки інструментів штучного інтелекту, таких як ті, які ще не застосовуються в клінічних умовах, підкреслюють стадію зародження прямого застосування та наступні проблеми в створенні реальних середовищ тестування. Пильний контроль, під яким інструменти штучного інтелекту знаходяться в рамках академічних досліджень, ще більше ускладнює експериментальний план, оскільки існує високий попит на суворі та прозорі методи, які можуть витримати критичний погляд наукової спільноти. Однак ця проблема залишається за межами академічних кіл, розгалужуючись до потреби в інструментах штучного інтелекту, які слід оцінювати таким чином, щоб відображати їхній потенціал революціонізувати не лише бізнес, але й різні інші сектори. Нарешті, очікування, що такі інструменти ШІ, як Polymer, аналізуватимуть дані та покращуватимуть розуміння користувачем без традиційних тривалих процесів, вводить вимогу до експериментів, які мають бути розроблені таким чином, щоб підтверджувати ці твердження про економію часу [50]. Таким чином, розробка експериментів для оцінки інструментів штучного інтелекту є багатогранним завданням, яке включає в себе баланс між дослідженням створених штучним інтелектом

можливостей і вимогами, застосованих у реальному світі та ефективних у часі процесів перевірки.

2.5 Висновок до другого розділу

В другому розділі кваліфікаційної роботи досліджено методи та інструменти штучного інтелекту для виявлення дезінформації в Facebook.

Алгоритми штучного інтелекту можуть аналізувати текстові дані, щоб відрізнити дезінформацію від достовірної. Такі інструменти, як Sphere, CrowdTangle, Factmata та NewsGuard, відіграють ключову роль у просуванні величезних обсягів інформації, щоб відрізнити правду від вигадки. Необхідно створити системи, які спрямовані на підвищення прозорості та підзвітності онлайн-платформ, щоб забезпечити більш надійну основу для ефективної боротьби з дезінформацією для ШІ.

3 РОЗРОБКА AI-СИСТЕМИ ВИЯВЛЕННЯ ФЕЙКОВИХ НОВИН

3.1 Побудова та оцінка моделі NLP

У сфері машинного навчання надмірна кількість статей, як правило, зосереджується виключно на процесі моделювання. Безумовно, етапи побудови та перевірки моделі є надзвичайно важливими та заслуговують на значну увагу. Однак існує потреба у більш повному висвітленні. Тому взято ініціативу провести ретельний огляд власного проекту, охоплюючи не лише необхідні завдання боротьби з даними та розробки моделі, але й створення загальнодоступного інтерфейсу для виявлення дезінформації в новинах соціальної мережі Facebook.

Початкова передумова була зосереджена навколо ідеї, що мова, яка використовується у фейкових новинах, суттєво відрізняється від мови достовірних новин, і що методи машинного навчання можуть ефективно ідентифікувати цю відмінність. Провівши ретельний аналіз фейкових новинних статей у соціальні мережі Facebook, помітно переважання термінів, стратегічно розроблених, щоб викликати обурення. Крім того, стало очевидним, що рівень письменницької майстерності, демонстрований у цих статтях, загалом був значно нижчим за стандарти, які зазвичай спостерігаються в авторитетних джерелах новин. Озброївшись цими спостереженнями, з’явилася можливість побудувати свою модель точної класифікації фейкових новин.

Щоб підтвердити дану гіпотезу, необхідно отримати об’єктивно позначений набір даних, який охоплює приклади як фейкових, так і справжніх новин, які визначено професійними факт-чекерами. Цей набір даних має складатися зі статей новин із відповідними URL-адресами, а також вердикт щодо точності чи хибності кожної статті. Визначено два набори даних, які відповідають цим критеріям. Перший набір даних отримано з діаграми упередженості інтерактивних медіа, наданої Ad Fontes Media. Другий набір даних, який використано – набір даних Fake News Net. Щоб забезпечити узгодженість, внесені певні зміни в набори даних. Це передбачало видалення

будь-яких PDF-файлів або елементів, які явно не містили URL-адреси. Крім того, стандартизовано систему підрахунку балів, призначивши «фальшивий» рейтинг усім записам у файлі politifact_fake.csv Fake News Net і «справжній» рейтинг усім записам у файлі politifact_real.csv.

Початкова мета полягає в тому, щоб об’єднати дані в одну повну таблицю бази даних, з якої можемо отримати необхідний текст для проекту. Щоб досягти цього, використано Django ORM. Крім того, використання Django як повної веб-платформи, MVC спрощує процес надання загальнодоступного інтерфейсу для моделей. Коли запускаємо програму через manage.py, створення таблиці полегшується класами моделей Django.db.

Фрагмент коду який відповідає за використання Django ORM для об’єднання даних подано в лістингу 3.1

Лістинг 3.1 – Інтеграція даних за допомогою Django ORM

```
from django.db import models
class NewsArticle(models.Model):
# Stores the main content of the article
content = models.TextField()
# A numerical score indicating media bias, not utilized in this project
media_bias_score = models.FloatField()
media_bias_category = models.IntegerField()
# This represents the reliability score based on Media Bias Chart
reliability_score = models.FloatField()
# Categorizes articles into four levels of truthfulness, following Politifact's classification
# True = 4, Predominantly True = 3, Predominantly False = 2, False = 1
truthfulness_rating = models.IntegerField()
# URL from where the article was sourced
source_url = models.TextField()
# The name or identifier of the source
source_name = models.TextField()
```

Тепер повернемося до Django's manage.py і виконаємо команди makemigrations і migrate, щоб налаштувати базу даних.

Введемо дані в цю таблицю. Це процес обробки даних, і це одна з найскладніших і довготривалих частин машинного навчання. У великих

середовищах машинного навчання є спеціалісти з обробки даних, які майже нічого іншого не роблять, окрім збору набору даних.

Для вирішення цієї проблеми потрібно:

– Завантажити список URL-адрес і оцінок.

– Завантажити та проаналізувати сторінку з URL-адреси.

– Зберегти опрацьований варіант разом із оцінкою.

– Додати клас для будь-якої точки даних, яка його не має, розподіливши дані рівномірно.

Спочатку розглянемо все, крім частини про парсинг сторінки. Зараз просто потрібно знати, що SoupStrainer буде виконувати все це.

Основний компонент цієї програми – harvester.py. У початковому розділі harvester необхідно виконати попередні конфігурації, щоб програма django могла працювати через командний рядок, а не покладатися на веб-інтерфейс, реалізацію у вигляді коду подано в лістингу 3.2.

Лістинг 3.2 – Автоматизація збору та класифікації даних з Django

```
import os, sys, re, time
proj_path = “/home/jwales/eclipse-workspace/crowdnews/”
os.environ.setdefault(“DJANGO_SETTINGS_MODULE”,
“crowdnews.settings”)
sys.path.append(proj_path)
os.chdir(proj_path)
from django.core.wsgi import get_wsgi_application
application = get_wsgi_application()
from django.contrib.gis.views import feed
```

Наступний крок передбачає завантаження та налаштування моделей та SoupStrainer, який є парсером подано у лістингу 3.3.

Лістинг 3.3 – Налаштування парсера та моделей для обробки даних

```
import pandas as pd
from newsbot.strainer import SoupStrainer
from newsbot.models import NewsModel

ss = SoupStrainer()
print("Запуск ініціалізації словника...")
ss.initialize()
```

Далі продемонстровано подальший підхід до завантаження даних, отриманих від Politifact, лістинг коду подано в додатку В.

Використано Pandas для отримання файлу CSV, що містять URL-адреси та відповідні оцінки. Згодом відправляємо їх в парсер і зберігаємо отриманий текст у змінній body_text. Процедура обробки даних Media Bias Chart відбувається за аналогічною схемою; однак повинні розділити quality_class на кожну чверть показника якості, отриманого з діаграми упередженості медіа.

Однак справжня суть завдання ще не розкрита! Як саме отримуємо дані з цих веб-сайтів? Який формат даних і як їх аналізуємо?

Щоб виконати це завдання, повинні обмежити кількість слів. Потрібно переконатися, що слова є справжніми, зберігши лише основу слова. Такі слова, як програміст, програмування та програма, можна спростити до програмування. Зробивши це, зменшимо розмір навчальних прикладів, що зробить їх можливими для роботи на звичайному ПК. Це завдання включено в клас SoupStrainer, який раніше зустрічали в harvester.

Вже, маємо необхідний набір імпорту. Для парсингу HTML використовується BeautifulSoup, для завантаження веб-сторінок з Інтернету - urllib3, а для виконання стемінгу слів – PorterStemmer. Крім того, ще знадобляться кілька інших речей для обробки цих даних (див. лістинг 3.4).

Лістинг 3.4 – Комплексний підхід до ефективного парсингу даних

```
import urllib3 as url_lib
import re as regex_lib
import string as str_lib
import json as json_lib
import html as html_lib
from bs4 import BeautifulSoup as Soup
from bs4.element import Comment as BComment
from urllib3.exceptions import HTTPError as UrlError
from io import StringIO as SIO
from nltk.stem import PorterStemmer as PStemmer
```

Після цього перейдемо до створення класу та завантаження словника, використовуючи повний словник, який подано в лістингу 3.5

Лістинг 3.5 – Створення класу та завантаження словника

```
class SoupFilter():
   engDict = {}
   hasHeadline = False
   recordedHeadline = ''
   locationToGet = ''
   dataPage = None
   error = None
   bsoup = None
   outputMsg = True
   def initialize(self):
      with open('newsbot/words_dict.json') as json_file:
         self.engDict = json.load(json_file)
```

Далі потрібно зосередитись виключно на видимому тексті на сторінці; отже, наступний етап передбачає побудову фільтра, який може ефективно розрізняти видимі та невидимі теги (див. лістинг 3.6).

Лістинг 3.6 – Розробка фільтра для розрізнення видимого тексту

```
def is_tag_visible(self, elem):
   if elem.parent.name in ['style', 'script',
          'head', 'title', 'meta', '[document]']:
      return False
   if isinstance(elem, BComment):
      return False
   return True
```

Майбутня функція, яка обробляє процес завантаження та аналізу. Спочатку створимо певні елементи та переконаємося, що вони певною мірою нагадують URL-адресу, реалізацію подано в лістингу 3.7.

Лістинг 3.7 –Завантаження та аналіз даних з URL-адреси

```
def setAddress(self, addr):
   self.locationToGet = addr
   self.hasHeadline = False
   httpMatch = regex_lib.compile('.*http.*')
   userAgent = {'user-agent': 'Mozilla/5.0 (Windows NT 6.3;
rv:36.0) Gecko/20100101 Firefox/36.0'}
   stemmer = PStemmer()
   if(httpMatch.match(self.locationToGet) is None):
      self.locationToGet = "http://" + self.locationToGet
```

Далі продовжимо перевіряти, чи можливе успішне завантаження URL-адреси (див. лістинг 3.8).

Лістинг 3.8 – Валідація та завантаження URL-адреси

```
     if(len(self.locationToGet) > 5):
        if(self.outputMsg):
          print("Готовий завантажити дані сторінки для: " +
self.locationToGet +
                "який був отриманий з " + addr)
        try:
           url_lib.disable_warnings(
                  url_lib.exceptions.InsecureRequestWarning)
           httpPool = url_lib.PoolManager(2, headers=userAgent)
           req = httpPool.request('GET', self.locationToGet)
           self.dataPage = req.data
           if(self.outputMsg):
              print("Дані сторінки успішно завантажені")
        except:
           self.error = 'Помилка при HTTP-запиті'
           if(self.outputMsg):
              print("Проблема з завантаженням сторінки")
           return False
```

Процес перевірки проходить успішно, оскільки завантажили веб-сторінку та отримали файл HTML. Наступний крок передбачає виділення тексту, який видимий, видалення будь-якої пунктуації, перевірку його автентичності як справжнє слово, а потім вилучити його, реалізацію у вигляді коду подано в лістингу 3.9.

Лістинг 3.9 – Алгоритм аналізу видимого тексту з HTML-файлів

```
     self.textExtract = ''
     self.recordedHeadline = self.locationToGet
     self.bsoup = Soup(self.dataPage, 'html.parser')
     textElements = self.bsoup.findAll(text=True)
     visibleText = filter(self.is_tag_visible, textElements)
     allVisibleText = u"".join(t.strip() for t in visibleText)
     for word in allVisibleText.split():
        standardizedWord = word.lower()
        standardizedWord = standardizedWord.translate(
             str_lib.maketrans('', '', str_lib.punctuation))
        standardizedWord                                        =
standardizedWord.strip(str_lib.punctuation)
        if(standardizedWord in self.engDict):
           standardizedWord = stemmer.stem(standardizedWord)
           self.textExtract = self.textExtract + standardizedWord +
" "

     return True
```

Після виконання сценарію harvester.py отримаємо вичерпну колекцію прикладів, що складається з фактичних слів із коренем, URL-адреси та індикатора quality_class, який визначає, чи він автентичний чи сфабрикований.

Далі необхідно розглянути модель, яку будемо будувати. Фундаментальний аспект плану передбачає створення матриці екземплярів, де кожен рядок представляє окремий приклад. У рядку серія з 1 і 0 вказує на наявність або відсутність слова, що відповідає цьому конкретному індексу в прикладі. Отже, наступний крок передбачає призначення унікального індексу кожному слову в завантажених прикладах, які зберігаються в базі даних. Тому створення словника є обов'язковим. Першим кроком до досягнення цього є створення моделей даних і таблиць, необхідних для цього завдання. Повертаємося до models.py і додаємо такий код:

```
class WordEntry(models.Model):
 standardWord = models.TextField()
```

У терміналі виконаємо процедуру makemigration/migrate, і коли це буде завершено, тоді можемо продовжити завдання зі створення словника. Щоб досягти цього, потрібно вивчити сценарій dictbuilder.py.

Після імпорту необхідних бібліотек django наступним кроком є створення словника Python, використовуючи існуючі слова в таблиці словників канонічних слів. Цей підхід дозволяє нам швидко визначити, чи слово вже присутнє в словнику, усуваючи необхідність частого пошуку в базі даних. Виконання окремого запиту до бази даних для кожного тесту слова значно знизить продуктивність, принаймні в 10 разів. Тому цей аспект є дуже цінним. Оскільки будемо виконувати це завдання неодноразово протягом усього процесу, створимо файл під назвою util.py і визначимо функцію, яка повертатиме словник канонічних слів, фрагмент коду подано в лістингу 3.10.

Лістинг 3.10 – Оптимізація пошуку слів з Django

```
def loadStandardDict():
    standardDict = WordEntry.objects.all()
    dictSize = standardDict.count() + 1
```

```
    sDict = {}
    for sw in standardDict:
        sDict[sw.standardWord] = sw.pk

    return sDict
```

Нарешті готові побудувати словникову таблицю, що включає всі слова, зібрані з різних прикладів, фрагмент коду подано в лістингу 3.11.

Лістинг 3.11 – Створення словникової таблиці з інтегрованих даних

```
    sample_Articles = ArticleExample.objects.all() #filter(pk__gt
= 2942)
    print("Приклади: " + str(sample_Articles.count()))
    for article in sample_Articles:
       words = article.body_text.split()
       for word in words:
          if(word in wordDict.keys()):
             print('.', end='', flush=True)
          else:
             print('X', end='', flush=True)
             newEntry = WordEntry(standardWord = word)
             newEntry.save()
             wordDict[word] = newEntry.pk
```

Це призначить ідентифікатор первинного ключа для кожного слова, які будуть нумеруватися на сайті. Таким чином, коли будуємо приклад, тоді можемо використати первинний ключ як номер стовпця для оновлення будь-якого конкретного прикладу, і бути впевненими, що той самий стовпець одне й те ж має значення для кожного прикладу.

Зробивши крок назад, маємо повну колекцію даних, які можемо використовувати, а також метод визначення присутності конкретних слів у кожному прикладі. У наборі даних є загалом 20 870 слів і 2 500 окремих прикладів, яким було присвоєно бали. Заклавши цю основу, можемо приступити до процесу навчання моделей.

Завдання полягає в класифікації статей новин за однією з чотирьох окремих категорій:

- Фейкові.
- Підступні.
- Схожі на справжні.

– Справжні.

Це є значною перешкодою для моделі класифікатора. Щоб вирішити дану проблему, будемо використовувати бібліотеку scikit-learn, яка пропонує низку моделей. Мета – протестувати кілька моделей і визначити найбільш ефективну для цього конкретного завдання. Машина опорних векторів або нейронна мережа були б ідеальними для вирішення такого типу проблем.

Для кожної статті отримали спрощений і скорочений варіант тексту, що складається з окремих слів. Кожному слову присвоюється окремий ідентифікатор. Щоб побудувати кожен приклад, представимо його як вектор-рядок numpy. Якщо в статті з'являється слово з ідентифікатором n, example[n] буде встановлено на 1; інакше він дорівнюватиме 0. Потім ці вектори-рядки будуть об'єднані, щоб сформувати велику матрицю з розмірами 20870 x 2500. Крім того, створимо вектор-стовпець для зберігання класу якості, наданого набором даних. Ця матриця та вектор-стовпець будуть використані для навчання та тестування моделі.

Щоб забезпечити ефективну обробку статей у рядки, важливо розробити метод, який можна застосовувати послідовно. Це завдання потрібно буде повторювати кожен раз, коли тестуються нові приклади. Тому доцільно створити службову функцію, спеціально призначену для цієї мети (див. лістинг 3.12).

Лістинг 3.12 – Реалізація функції для послідовної обробки тексту

```
def constructSampleRow(text_body, wordDict):
    dictLength = len(wordDict.keys())
    single_example_vector = np.zeros(dictLength+2)
    words = text_body.split()
    for word in words:
        if(word in wordDict.keys()):
            single_example_vector[wordDict[word]-1] = 1
        else:
            print("Це слово не існує в словнику:" + word)

    return(single_example_vector)
```

Далі продовжуємо створення функції, яка дозволяє отримувати та завантажувати приклади з бази даних, фрагмент коду подано в лістингу 3.13.

Лістинг 3.13 – Отримання та завантаження даних з бази даних

```
    def handleSamples(sample_Examples, wordDict):
        class_vector      =      np.zeros(sample_Examples.count(),
dtype=np.int8)
        class_vec_counter = 0
        samplesMatrix = None

        for sample in sample_Examples:
            class_vector[class_vec_counter]                        =
int(sample.quality_class)
            class_vec_counter = class_vec_counter + 1
            if(samplesMatrix is None):
                samplesMatrix                                      =
constructSampleRow(sample.body_text, wordDict)
            else:
                samplesMatrix = np.vstack(
                   [samplesMatrix,
                   constructSampleRow(sample.body_text,
wordDict)])
                print('.', end='', flush=True)

        return( (class_vector, samplesMatrix))
```

У class_learner.py тепер можемо ефективно та легко встановити необхідні налаштування для навчання (див. лістинг 3.14).

Лістинг 3.14 – Налаштування class_learner.py для навчання моделей

```
    from newsbot.utility import *

    print("Підготовка..")
    wordDict = loadStandardDict()
    sample_Examples                                                =
ArticleExample.objects.filter(quality_class__lt = 5)

    print("Обробка прикладів")
    (class_vector, samplesMatrix) = handleSamples(sample_Examples,
wordDict)
```

Настав час розпочати навчання та оцінку моделей. Дані будуть розділені на два набори для:

– Навчання.

– Тестування.

Продовжимо навчання багатошарового перцентроного класифікатора, який подано в лістингу 3.15.

Лістинг 3.15 – Навчання багатошарового перцептронного класифікатора

```
    trainData,    testData,    trainLabels,    testLabels    =
train_test_split(samplesMatrix, class_vector, test_size=0.2)
    neuralNet = MLPClassifier(hidden_layer_sizes=(128,64,32,16,8),
max_iter=2500)
    neuralNet.fit(trainData, trainLabels)
```

Перевага використання надійної бібліотеки, такої як SciKit, полягає в тому, що основні етапи цієї процедури залишаються узгодженими для різних моделей.

Вже є модель, яка пройшла відповідне навчання, і тепер настав час визначити її ефективність і потенціал для створення чогось значного. Щоб переконатися в цьому, потрібно провести різні тести і ретельно осмислити їх результати. Це означає початок процесу перевірки моделі. Маючи справу з моделлю класифікації, можна отримати цінну інформацію про ефективність класифікації новин, реалізацію подано в лістингу 3.16.

Лістинг 3.16 – Оцінка навченої моделі класифікації

```
print("################")
print("Класифікація базується на:")
print(accuracy_score(predicted, testLabels))
print(confusion_matrix(predicted, testLabels))
print(classification_report(predicted, testLabels))
print("################")
```

Після навчання моделі та її оптимізації, настав час оцінити її ефективність, результат оцінки подано в лістингу 3.17.

Лістинг 3.17 – Оцінки точності класифікаційної моделі

```
print("################")
print("Точність класифікації: 0.546")
print("Матриця помилок: ")
print("[[ 35   8   3   2]")
print(" [ 10  44  28   5]")
print(" [ 21  40  83  58]")
print(" [  3   4  45 111]]")
print("Звіт про класифікацію: ")
print("              точність    відгук  f1-оцінка   підтримка")
print("1       0.51      0.73      0.60         48")
print("           2       0.46      0.51      0.48         87")
print("           3       0.52      0.41      0.46       202")
print("           4       0.63      0.68      0.65       163")
print("точність                           0.55       500")
```

```
print("   середнє       0.53      0.58      0.55       500")
print("зважене середнє      0.54      0.55      0.54      500")
print("###############")
```

Оцінка точності 0,546 – означає, що 54,6% прогнозів були точними. Тому є 4 класи, випадкове введення призвело до точності в 25%, тому це фактично покращення.

Матриця плутанини – набір з 16 чисел дає нам цікаву інформацію. Подано її нижче в більш зрозумілому форматі. Кожен рядок відповідає передбачуваному класу, а кожна колонка має справжній клас. Число на перетині рядка і колонки – кількість звернень, коли справжній клас передбачено як кожен передбачений клас. Число 44, яке знаходиться в другому рядку, другій колонці, є кількістю прикладів, де справжній клас був 2, а передбачене значення також було 2. Число 28, яке знаходиться в наступній колонці того ж рядка, є приклади , які насправді були 2, але модель передбачила 3. Зверніть увагу, що серед статей, які насправді є 1, сильно переважають значення в сторону 1/2 шкали. Аналогічно, у нижньому рядку маємо приклади, які насправді оцінюються як 4, і позначено, що останні 2 колонки переважають перші дві. Це хороші новини!

У звіті про класифікацію кожен рядок відповідає певному класу та відображає точність, запам’ятовування та оцінку F1. Представлені тут числові значення відрізняються за рівнем досконалості. Загалом здатність розпізнавати правдиві новини виявляється сильнішою, ніж здатність виявляти фейкові новини. Однак, навіть якщо можемо досягти першого, все одно можемо працювати ефективно. Продовжуючи прокручувати, потрапимо на загальну середньозважену точність/запам’ятовування/оцінку F1, яка забезпечує більш повну оцінку продуктивності моделі. Бал F1 0,54 забезпечує баланс між точністю та запам’ятовуванням, що робить його надійним показником для оцінки моделей класифікації. Якщо зустрінемо іншу модель зі значно вищим результатом F1, було б доцільно вибрати її. Крім того, оцінка точності служить ще одним цінним показником, і на практиці ці два показники мають тенденцію тісно збігатися один з одним.

Під час попереднього спостереження за матрицею плутанини виявлено помітне спостереження, що класифікатор часто відхиляється на 1. Крім того,

досліджуючи середню абсолютну похибку моделі, знаходяться додаткові докази на підтримку цього спостереження.

```
absError = mean_absolute_error(testLabels, predicted)
print("Середня абсолютна помилка: " + str(absError))
```

Збираємося отримати:

```
averageAbsoluteError = 0.54
print("Середня абсолютна помилка: " + str(averageAbsoluteError))
```

Це служить додатковим доказом того, що в середньому модель є точною в межах 1 класу порівняно з фактичною відповіддю. Знання будуть керувати майбутньою розробкою моделі та остаточною поставкою. Якщо маємо намір надати кінцевому користувачеві повне розуміння процесу мислення моделі, потрібно створити оцінки ймовірності. У той час як більшість моделей можуть робити це автоматично, класифікатор опорних векторів потребує використання позначки probability=True під час навчання, щоб створити оцінку ймовірності.

Отримавши розуміння інтерпретації матриці плутанини та класифікаційного звіту, а також вибравши один показник для оцінки моделей, можемо перейти до тестування багатьох моделей.

Після тривалого навчання та ретельної каталогізації результатів представляємо кульмінацію зусиль з тестування на безлічі моделей, розроблених для вирішення цієї конкретної проблеми, подано на рисунку 3.1.

Model	Accuracy	Precision	Recall	F1
SVC	65.0%	70%	65%	66%
LogisticRegression	57.2%	59%	57%	58%
MLPClassifier	54.6%	54%	55%	54%
LinearDiscriminantAnalysis	52.0%	55%	52%	52%
GaussianNB	45.0%	52%	45%	47%
KNeighborsClassifier	43.0%	49%	43%	45%
DecisionTreeClassifier	45.0%	45%	45%	45%

Рисунок 3.1 – Категоризація результатів моделей

Логістична регресія посіла друге місце. Однак не можу заперечити, що K-Neighbors і Decision Tree які просто не підходили для цієї конкретної проблеми. Безсумнівно, SVC став беззаперечним переможцем.

Поза межами інтерфейсу є чіткі уявленням про те, наскільки добре або погано працюють моделі. Далі потрібно побачити, як вони себе проявляють на реальних прикладах. Моделі, навіть якщо вони не є повністю правильними, все ж наближаються до правильності. Далі потрібно знати більше про те, чи надають вони необхідні відповіді. Далі потрібно зберегти топ 3 моделей і побудувати щось для їх тестування.

Збереження цінних моделей. Для цієї частини завдання створено новий файл під назвою class_saver.py. Таким чином, можна завантажити дані 1 раз, запустити та перевірити всі три найкращі моделі, а потім зберегти їх за допомогою pickle. Налаштовуємо середовище і помістимо три найкращі моделі в інтерабельний словник Python, фрагмет реалізації коду поданий в лістингу 3.18.

Лістинг 3.18 – Збереження та управління моделями машинного навчання

```
     print("Підготовка..")
     wordDict = loadStandardDict()
     sample_Examples                                                    =
ArticleExample.objects.filter(quality_class__lt = 5)
     print("Обробка прикладів")
     (class_vector, samplesMatrix) = handleSamples(sample_Examples,
wordDict)
     trainData, testData, trainLabels, testLabels =
          train_test_split(samplesMatrix,            class_vector,
test_size=0.2)
     selected_models = {}
     selected_models['newsbot/MLPClassifier_model.sav']           =
MLPClassifier(hidden_layer_sizes=(128,64,32,16,8), max_iter=2500)
     selected_models['newsbot/SVC_model.sav'] =
     SVC(gamma='scale', probability = True)
     selected_models['newsbot/LogReg_model.sav'] =
     LogisticRegression()
```

Далі перейдемо до ітерації вибраних моделей, проведемо навчальні сесії та приймемо рішення щодо їх збереження. Важливо зазначити, що моделі ініціалізуються випадковим чином, і їх багаторазове виконання дає дещо різні результати. Тому доцільно вивчити їх статистику, щоб переконатися у виборі надійної моделі (див. лістинг 3.19).

Лістинг 3.19 – Ітерація та оцінка надійності моделей

```
for modelName, trainedModel in selected_models.items():
    print("Обробка " + modelName)
    trainedModel.fit(trainData, trainLabels)
    predicted = trainedModel.predict(testData)
    print("Звіт про класифікацію: ")
    print(classification_report(predicted, testLabels))
    print("###############")
    saveModel = input("Зберегти " + modelName + "? ")
    if(saveModel == 'т' or saveModel == 'Т'):
        print("Збереження...")
        pickle.dump(trainedModel, open(modelName, 'wb'))
        print("Збережено!")
    else:
        print("Не збережено!")
```

Після завершення навчання моделі та подальшого збереження залишиться трійка файлів. У конкретному випадку три файли зайняли значну кількість місця, загальний розмір приблизно 391 МБ у сукупності.

Створимо інтерфейс командного рядка, який надасть оцінки в реальному часі. Перейдемо до роботи з classify_news.py.

Після завершення звичайного налаштування, яке включає імпорт необхідних компонентів і налаштування Django для доступу до ORM, наступним кроком є завантаження трьох попередньо розроблених моделей, які подано в лістингу 3.20.

Лістинг 3.20 – Оцінювання новин у реальному часі з classify_news.py

```
    print("Завантаження мозку...")
    logistic_model                                          =
pickle.load(open('newsbot/logistic_model.sav', 'rb'))
    support_vector_model                                    =
pickle.load(open('newsbot/support_vector_model.sav', 'rb'))
    multi_layer_perceptron_model                            =
pickle.load(open('newsbot/multi_layer_perceptron_model.sav', 'rb'))
    print("Успішне завантаження мозку.")
```

Після цього необхідно ініціалізувати всі компоненти, щоб перетворити статтю на приклад, як це робили раніше, яке реалізовано в лістингу 3.21.

Лістинг 3.21 – Ініціалізація компонентів для конвертації статей

```
print("Ініціалізація словників...")
```

```
wordDict = loadStandardDict()
soupFilter = SoupFilter()
soupFilter.initialize()
```

Далі перетворюємо статтю на рядок введення для моделей, реалізацію подано в лістингу 3.22.

Лістинг 3.22 – Конвертація статей у формат вхідних даних

```
webAddress = input("URL для аналізу: ")

print("Спроба URL: " + webAddress)
if(soupFilter.setAddress(webAddress)):
    articleSample = constructSampleRow(soupFilter.textExtract,
wordDict)
else:
    print("Помилка URL, вихід")
    exit(0)

articleSample = articleSample.reshape(1, -1)
```

Як тільки створили добре організований вектор-рядок, можемо передбачати та генерувати ймовірності для кожної з моделей, фрагмент коду реалізації подано в лістингу 3.23.

Лістинг 3.23 – Генерація ймовірностей за допомогою моделей

```
logistic_predicted = logistic_model.predict(articleSample)
logistic_probabilities =
logistic_model.predict_proba(articleSample)
svc_predicted = support_vector_model.predict(articleSample)
svc_probabilities =
support_vector_model.predict_proba(articleSample)
mlp_predicted =
multi_layer_perceptron_model.predict(articleSample)
mlp_probabilities =
multi_layer_perceptron_model.predict_proba(articleSample)

print("*** Модель опорних векторів ")
print("Прогноз на цю статтю: ")
print(svc_predicted)
print("Ймовірності:")
print(svc_probabilities)
print("*** Логістична модель ")
print("Прогноз на цю статтю: ")
print(logistic_predicted)
print("Ймовірності:")
print(logistic_probabilities)
print("*** Модель багатошарового перцептрона ")
```

```
print("Прогноз на цю статтю: ")
print(mlp_predicted)
print("Ймовірності:")
print(mlp_probabilities)
```

Переходимо до виконання класифікатора. Процес видалення моделей і завантаження словників займає приблизно 1-2 секунди, що трохи довго, але все ще в допустимих межах. Однак, якщо маємо намір реалізувати це у більшому масштабі, тоді обов'язковим потрібно розробити метод завантаження моделей і словників у пам'ять і забезпечити їх збереження, доки вони явно не будуть видалені. Для розробки цей підхід продовжуватиме працювати за призначенням. Після введення справжньої URL-адреси статті результат буде виглядати наступним чином, як подано в лістингу 3.24.

Лістинг 3.24 – Оптимізації виконання класифікатора

```
print("*** Модель опорних векторів")
print("Прогноз на цю статтю: ")
print([3])
print("Ймовірності:")
print([[0.01111608, 0.0503078, 0.70502378, 0.23355233]])
print("*** Логістична модель")
print("Прогноз на цю статтю: ")
print([3])
print("Ймовірності:")
print([[5.61033543e-04,    5.89780773e-03,    7.63196217e-01,
2.30344942e-01]])
print("*** Модель багатошарового перцептрона")
print("Прогноз на цю статтю: ")
print([4])
print("Ймовірності:")
print([[1.18020372e-04,    1.93965844e-09,    4.88694225e-01,
5.11187753e-01]])
```

Наведені тут ймовірності вказують на ймовірність кожної з 4 класифікацій, причому перша категорія представляє повну фабрикацію, а остання абсолютну автентичність. Згідно з класифікатором опорного вектора існує 93,9% ймовірність того, що стаття є справжньою або переважно автентичною, як визначено комбінованою ймовірністю 70,5% і 23,4%. Інші два класифікатори дають подібні результати. Усі три класифікатори погоджуються, що свідчить про те, що ця стаття виглядає дуже надійною. Щоб додатково оцінити надійність

класифікаторів, переходимо до аналізу джерел новин, відомого розповсюдженням фальшивих і ненадійних новин. На їхній домашній сторінці навмання вибрано статтю, отриманий результат подано в лістингу 3.25.

Лістинг 3.25 – Аналіз оцінки надійності класифікатора

```
print("*** Модель опорних векторів")
print("Прогноз на цю статтю: ")
print([1])
print("Ймовірності:")
print([[0.80220529, 0.18501285, 0.01051862, 0.00226324]])
print("*** Логістична модель")
print("Прогноз на цю статтю: ")
print([1])
print("Ймовірності:")
print([[0.53989611, 0.45269964, 0.0005857, 0.00681855]])
print("*** Модель багатошарового перцептрона")
print("Прогноз на цю статтю: ")
print([1])
print("Ймовірності:")
print([[8.38376936e-01,    1.84104358e-03,    1.59391877e-01,
3.90143317e-04]])
```

Консенсус усіх трьох класифікаторів очевидний: ця стаття є шахрайською. Імовірності, які приписуються цій оцінці. І моделі SVC і Logistic висловлюють високу впевненість у віднесенні до категорії підроблених або сумнівних. Однак варто зазначити, що MLP призначає ймовірність 15,9% того, що стаття буде переважно точною.

3.2 Розробка веб-інтерфейсу користувача

Веб-інтерфейс реалізованого проекту подано в додатку А.

Нарешті, розробили три моделі, які демонструють вражаючу ефективність, і зараз мета – запровадити їх для загального використання. Усвідомлюємо, що в численних випадках, коли стаття повинна бути класифікована як клас 4, моделі, як правило, приділяють значний акцент класам 3 і 4, і ця модель є послідовною і на іншому кінці спектру. Іноді можуть виникати розбіжності, особливо щодо точних ймовірностей. Крім того, що пересічний користувач не матиме труднощів із встановленням Python і безлічі бібліотек лише для використання цього

інструменту. Тому вкрай важливо забезпечити, щоб він був зручним, легким для розуміння та доступним для кожного.

Спочатку розробили цей проект за допомогою фреймворку Django ORM, а це означає, що створення веб-інтерфейсу має бути простим процесом.

Для початку повинні зробити необхідні оновлення у файлі views.py, який був створений Django, коли спочатку використовували команду startapp для ініціювання всього цього процесу.

У даній частині можемо побачити деякий код, який раніше використовували в інтерфейсі командного рядка. Цей код відповідає за конфігурацію моделей, розбір прикладу та генерацію вектора-рядка, який представляє текстовий вміст поданої статті, лістинг коду подано в додатку Д.

Успішно отримали articleX і зберегли його в масиві numpy, що складається з 1 і 0, який точно представляє канонічні слова, знайдені в статті.

На даному етапі вкрай важливо отримати переконливий результат від усіх моделей, який подано в лістингу 3.26.

Лістинг 3.26 – Прогнозування моделей машинного навчання

```
    svc_predicted = support_vector_model.predict(articleSample)
    svc_probabilities                                          =
support_vector_model.predict_proba(articleSample)

    mlp_predicted                                              =
multi_layer_perceptron_model.predict(articleSample)
    mlp_probabilities                                          =
multi_layer_perceptron_model.predict_proba(articleSample)

    logistic_predicted = logistic_model.predict(articleSample)
    logistic_probabilities                                     =
logistic_model.predict_proba(articleSample)
```

Тепер необхідно встановити певні змінні відображення, які можуть бути корисними для шаблону. Щоб звести до мінімуму присутність математичних операцій і логічних міркувань у шаблоні, представимо ймовірності в більш зручному форматі, вираженому у відсотках, а не в традиційній формі з плаваючою точкою в діапазоні від 0 до 1, яку зазвичай використовують статистики. Для кожної з моделей генеруємо як загальну фальшиву метрику, так і загальну реальну метрику таким чином:

Загальна ймовірність бути класифікованим як фейк дорівнює сумі ймовірностей бути класифікованим як фейк (клас 0) і бути класифікованим як хитрий (клас 1).

Загальну реальну вартість можна визначити шляхом об’єднання ймовірності «Здається законним» у класі 3 з ймовірністю «Правда» в класі 3.

Це надає єдине порівняння в шаблоні, що дозволяє визначити, чи маємо представити «Здається справжнім» чи «Здається фальшивим» як остаточний вердикт, фрагмент коду реалізації подано у лістингу 3.27.

Лістинг 3.27 – Відсоткове представлення ймовірностей у моделях

```
     svc_probs            =            (svc_probabilities[0][0]*100,
svc_probabilities[0][1]*100,           svc_probabilities[0][2]*100,
svc_probabilities[0][3]*100)
     svc_totalFake      =      (svc_probabilities[0][0]*100)      +
(svc_probabilities[0][1]*100)
     svc_totalReal      =      (svc_probabilities[0][2]*100)      +
(svc_probabilities[0][3]*100)

     mlp_probs            =            (mlp_probabilities[0][0]*100,
mlp_probabilities[0][1]*100,           mlp_probabilities[0][2]*100,
mlp_probabilities[0][3]*100)
     mlp_totalFake      =      (mlp_probabilities[0][0]*100)      +
(mlp_probabilities[0][1]*100)
     mlp_totalReal      =      (mlp_probabilities[0][2]*100)      +
(mlp_probabilities[0][3]*100)

     log_probs            =            (log_probabilities[0][0]*100,
log_probabilities[0][1]*100,           log_probabilities[0][2]*100,
log_probabilities[0][3]*100)
     log_totalFake      =      (log_probabilities[0][0]*100)      +
(log_probabilities[0][1]*100)
     log_totalReal      =      (log_probabilities[0][2]*100)      +
(log_probabilities[0][3]*100)
```

Далі суть полягає в тому, щоб об’єднати три моделі та вирішити ситуації, коли рішення є спірними. У випадках, коли дві з трьох моделей надають перевагу одному результату, а решта моделі надає перевагу протилежному результату, можемо вирішити суперечку шляхом усереднення їхніх прогнозів. Цей підхід надає розподіл ймовірностей, який служить бажаним єдиним прогнозом у верхній частині сторінки для кінцевого користувача: визначення того, справжній вміст чи підробка, реалізацію подано у лістингу 3.28.

Лістинг 3.28 – Розв’язання спірних прогнозів у машинному навчанні

```
     final_prob                                =                                (
(((svc_probabilities[0][0]*100)+(mlp_probabilities[0][0]*100)+(log
istic_probabilities[0][0]*100))/3),

(((svc_probabilities[0][1]*100)+(mlp_probabilities[0][1]*100)+(log
istic_probabilities[0][1]*100))/3),

(((svc_probabilities[0][2]*100)+(mlp_probabilities[0][2]*100)+(log
istic_probabilities[0][2]*100))/3),

(((svc_probabilities[0][3]*100)+(mlp_probabilities[0][3]*100)+(log
istic_probabilities[0][3]*100))/3) )
     final_totalFake    =    (svc_totalFake    +    mlp_totalFake    +
logistic_totalFake)/3
     final_totalReal    =    (svc_totalReal    +    mlp_totalReal    +
logistic_totalReal)/3
```

Далі об’єднуємо всю цю інформацію у відповідну структуру та надсилаємо до призначеного шаблону, який поданий у лістингу 3.29.

Лістинг 3.29 – Інтеграція даних до візуалізації у шаблони

```
     contextData     =     {'headline':soupFilter.recordedHeadline,
'words': soupFilter.textExtract, 'url' : webAddress,
      'svc_totalFake': svc_totalFake,
      'svc_totalReal': svc_totalReal,
      'svc_predicted': svc_predicted,
      'svc_probabilities': svc_prob,
      'mlp_totalFake': mlp_totalFake,
      'mlp_totalReal': mlp_totalReal,
      'mlp_predicted': mlp_predicted,
      'mlp_probabilities': mlp_prob,
      'log_totalFake': log_totalFake,
      'log_totalReal': log_totalReal,
      'log_predicted': logistic_predicted,
      'log_probabilities': log_prob,
      'final_totalFake': final_totalFake,
      'final_totalReal': final_totalReal,
      'final_probabilities': final_prob
      }
      return render(request, 'newsbot/results.html', contextData)
```

Крім того, необхідно включити розділ у кінці форми, який пропонує користувачеві ввести URL-адресу, якщо вона ще не вказана.

```
     else:
      return render(request, 'newsbot/formURL.html')
```

Щоб досягти цього, потрібно побудувати коротку таблицю у файлі results.html із описом кожного потенційного результату. Таблицю, спеціально розроблену для остаточного рішення подано в додатку Е.

3.3 Веб-інтерфейс AI-системи для виявлення фейкових новин

На початку головної сторінки веб-інтерфейсу можна знайти інформацію про систему виявлення фейкових новин, яка подана на рисунку 3.2. Штучний інтелект аналізує текстові матеріали, щоб виявити можливі ознаки фейкових новин. Використовуючи передові алгоритми обробки природної мови, система визначає маніпулятивний зміст та неточності в статтях.

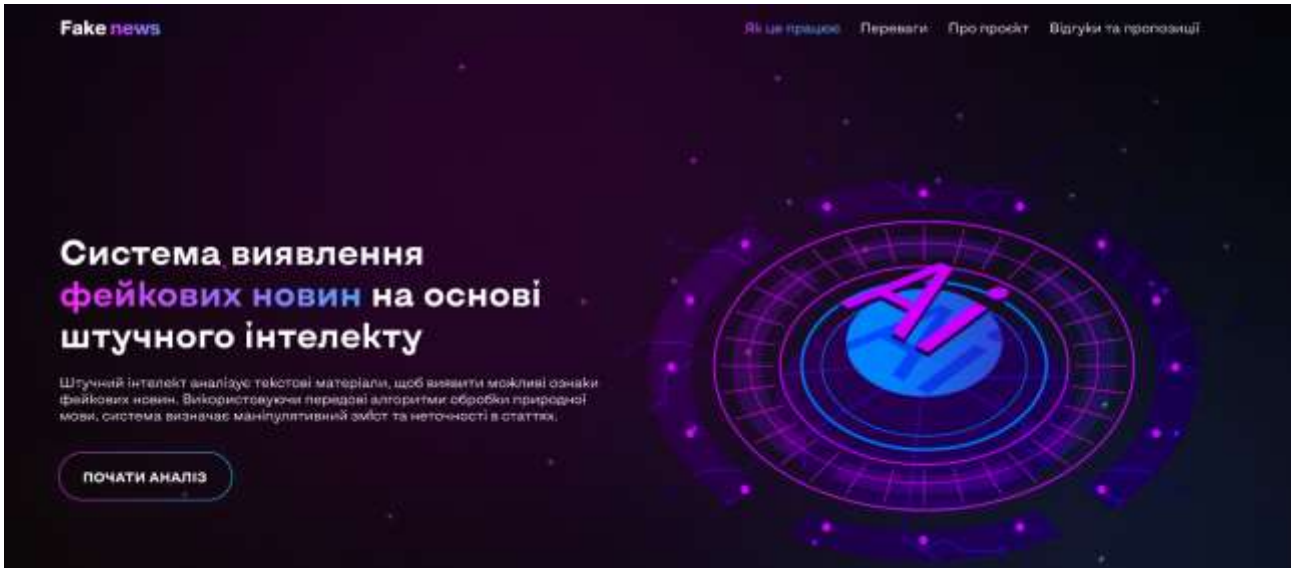

Рисунок 3.2 – Початок головної сторінки

Щоб виявити дезінформації в новинах соціальної мережі Facebook, потрібно дотримуватись наступних етапів:

– Завантаження контенту; користувач може завантажити контент через посилання, вказавши його у відповідну форму. Після цього слід натиснути кнопку «Почати аналіз».

– Аналіз та перевірка джерела; система проводить аналіз завантаженого контенту та перевіряє джерело на достовірність інформації. Використовуючи різні алгоритми та методи для виявлення можливих фейкових новин.

– Результати аналізу; після завершення аналізу користувач отримує звіт про надійність новин. Цей звіт містить інформацію про те, наскільки достовірною є надана новина.

Рекомендації до використання AI-системи подано на рисунку 3.3.

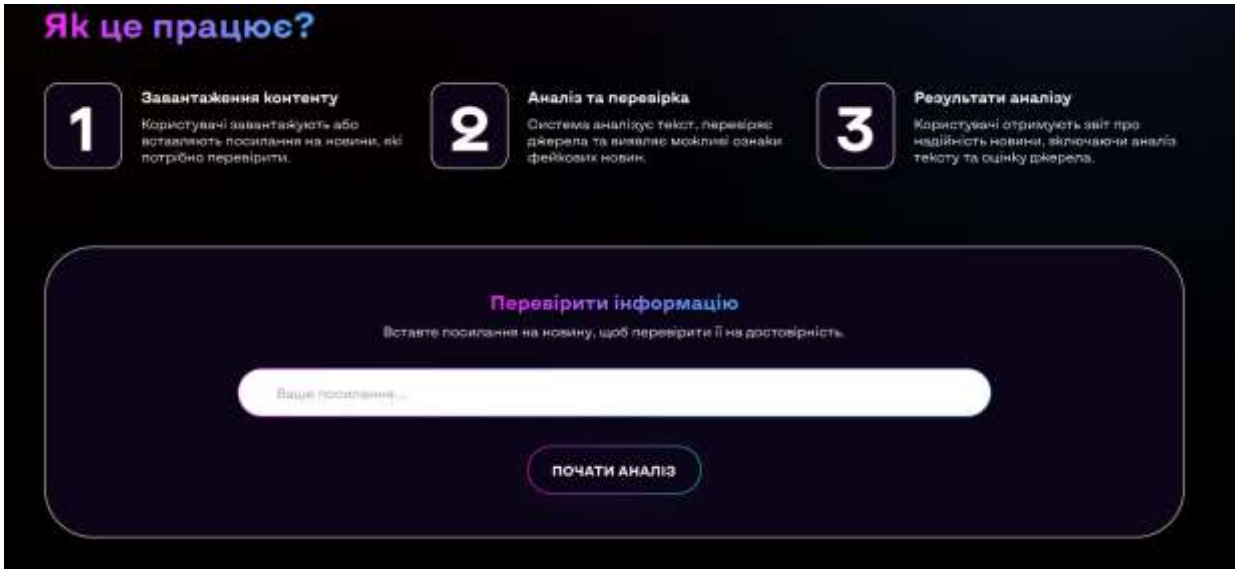

Рисунок 3.3 – Вказівки щодо виявлення фейкових новин

Також на головній сторінці подано інформацію про переваги використання реалізованого застосунку, такі як:

– Швидкість і точність виявлення фейкових новин.

– Користувацький інтерфейс.

– Захист від дезінформації.

Переваги використання розробленого проєкту подано на рисунку 3.4

Рисунок 3.4 – Переваги використання AI-системи

Інформацію про проєкт подано на рисунку 3.5.

Рисунок 3.5 – Інформація про реалізований проєкт

Для удосконалення точності отриманих результатів перевірки фейкових новин, додано можливість користувачам діляться спостереженнями за допомогою форми, яку подано на рисунку 3.6. Виявляти підозрілі матеріали та надавати цінний внесок у боротьбі з дезінформацією. Збільшення кількості відгуків та пропозицій також сприяє покращенню алгоритмів виявлення фейкових новин.

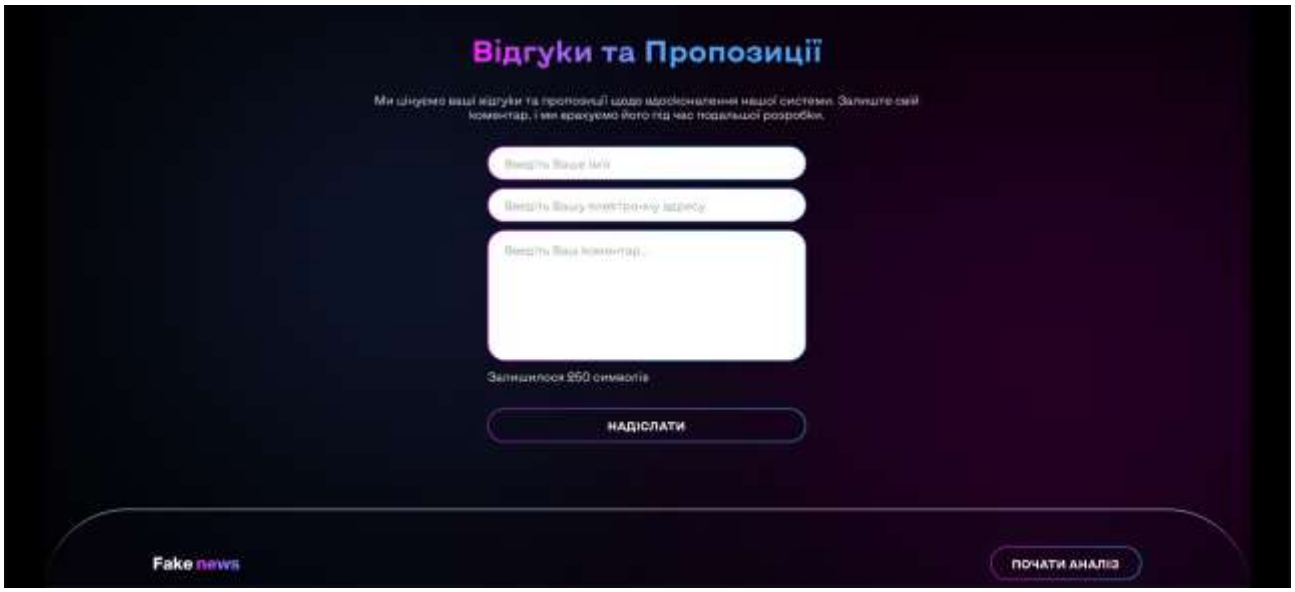

Рисунок 3.6 – Форма відгуків та пропозицій

Щоб перевірити розроблений інструмент штучного інтелекту щодо коректності виявлення дезінформації, використано статтю з Facebook, яку подано на рисунку 3.7.

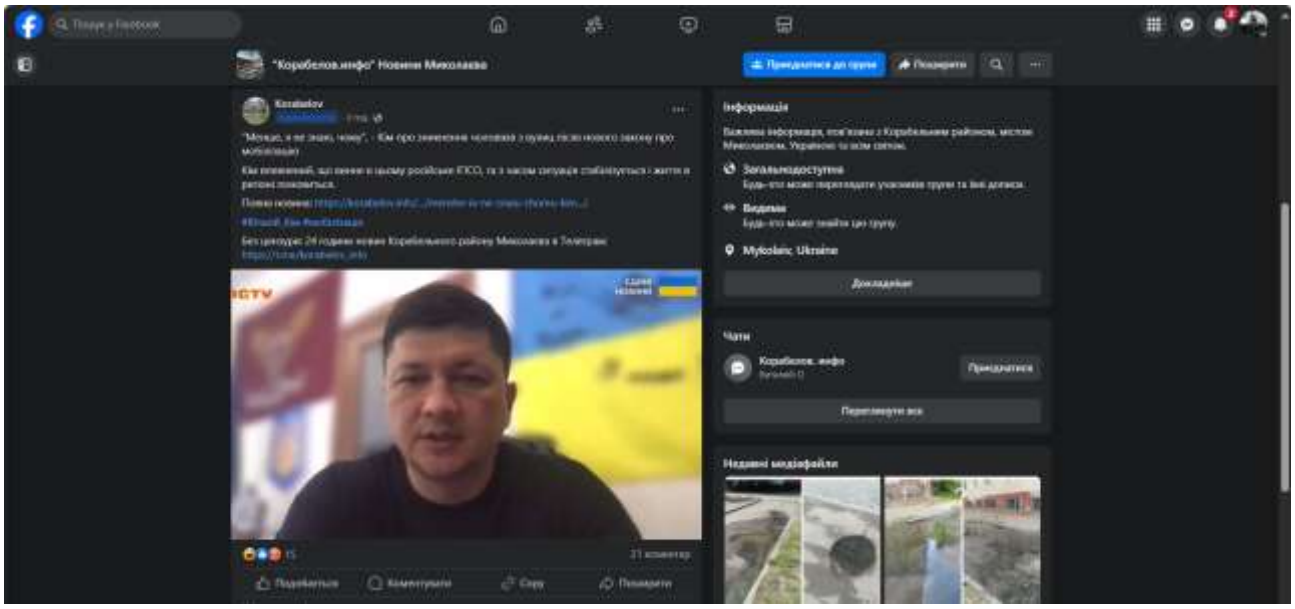

Рисунок 3.7 – Взята стаття з Facebook

Згідно посилання отримано результат у вигляді діаграм, одну з них подано на рисунку 3.8. Загальна діаграма базується на трьох попередніх діаграмах, яка відображає відповідність новинної статті до реальної інформації.

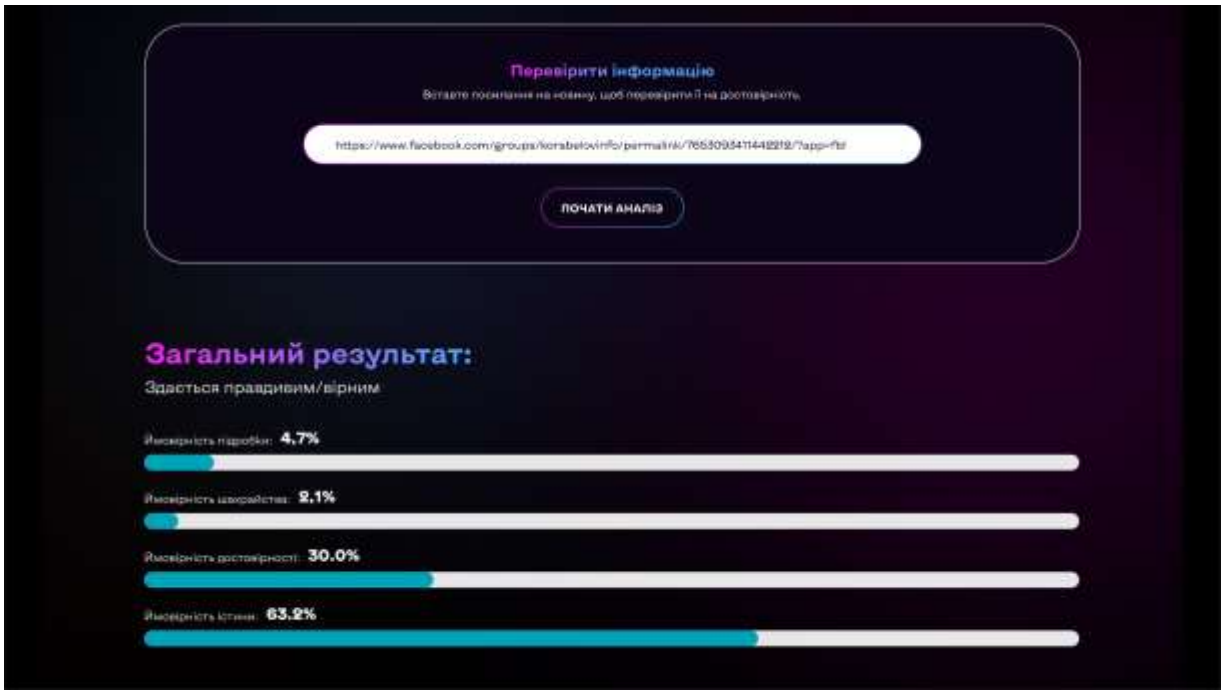

Рисунок 3.8 – Аналіз достовірності інформації

Модель Support Vector Classifier є алгоритмом машинного навчання, який використовується для класифікації даних. Основна ідея полягає в тому, щоб розділяє дані на різні класи за допомогою гіперплощини. Розглянемо деякі ключові аспекти цього алгоритму:

Механіка роботи SVC:

– SVC намагається знайти оптимальну гіперплощину, яка розділяє класи даних.

– Гіперплощина підтримується векторами підтримки, які гарантують, що межа гіперплощини максимально велика.

– Відстань від гіперплощини до найближчих точок називається «маржою». Точки є векторами підтримки.

– Оптимізація гіперплощини відбувається шляхом максимізації маржі.

Згідно результатів даної моделі отримано ймовірність класу істини, що підтверджує достовірність інформації, результат подано на рисунку 3.9.

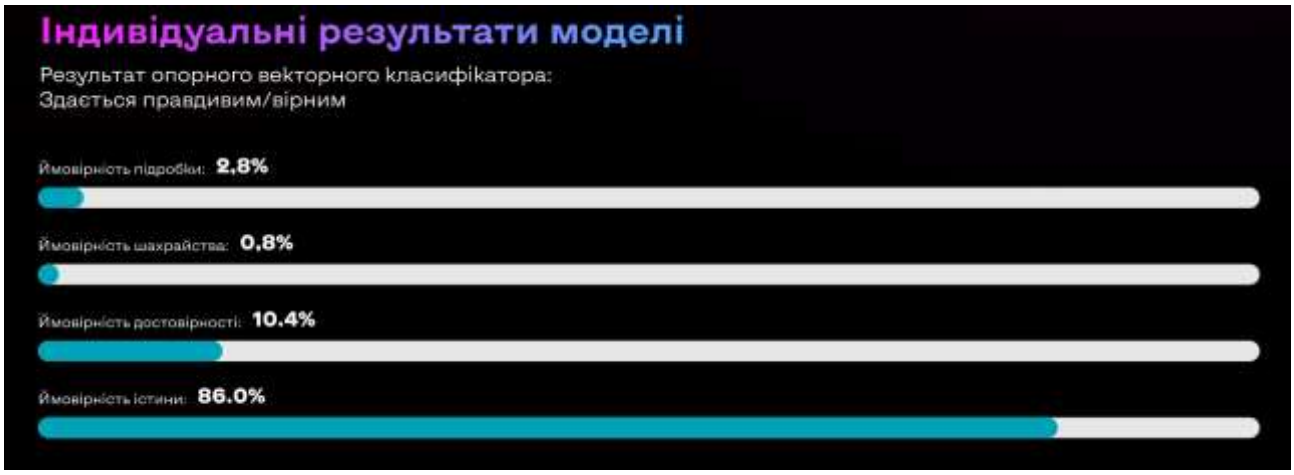

Рисунок 3.9 – Модель Support Vector Classifier

Logistic Regression – метод, який допомагає передбачити результати, коли є два можливих варіанти, правда чи неправда. Ця модель використовує математичні розрахунки, щоб знайти зв’язок між вхідними змінними та відповіддю. Використовуючи логістичну регресію, щоб передбачити, чи стаття є фейковою на основі текстових ознак, яка перетворює числа на ймовірності та допомагає приймати рішення.

Згідно результатів даної моделі отримано ймовірність класу достовірності наближеної до істини, що підтверджує реальність інформації, результат подано на рисунку 3.10.

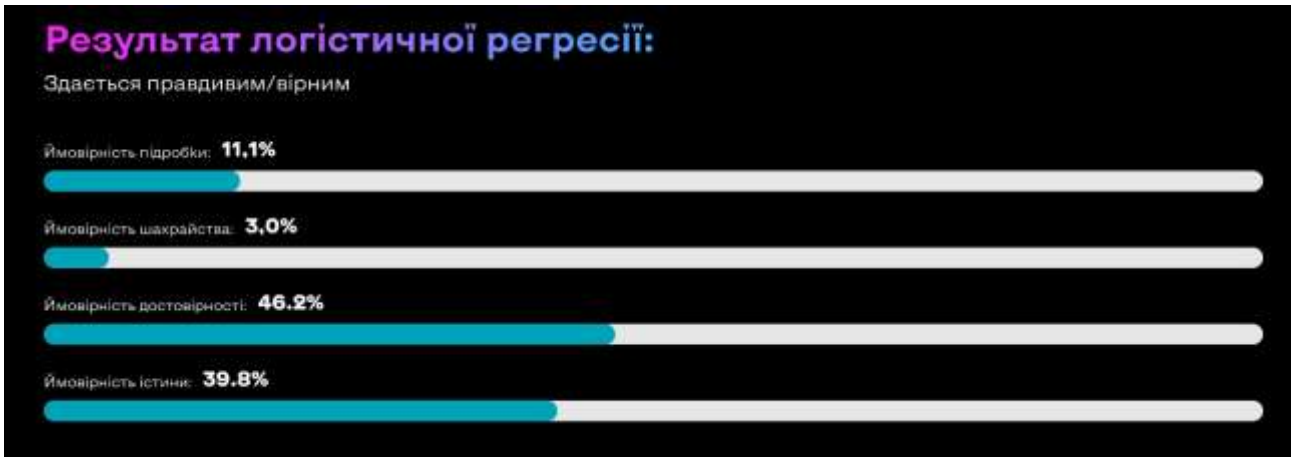

Рисунок 3.10 – Модель Logistic Regression

Multilayer Perceptron – тип штучної нейронної мережі, яка складається з кількох шарів нейронів. Кожен нейрон у MLP використовує нелінійні функції активації, що дозволяє мережі вивчати складні закономірності в даних. MLP використовується для завдань, таких як класифікація, регресія та розпізнавання зразків. ШНМ має вхідний шар, приховані шари та вихідний шар, і нейрони взаємодіють між собою, обробляючи вхідні сигнали та передаючи результати далі.

Згідно результатів даної моделі отримано ймовірність класу істини, що підтверджує реальність інформації, результат подано на рисунку 3.11.

Рисунок 3.11 – Модель Multilayer Perceptron

Для перевірити AI-системи щодо коректності виявлення дезінформації в новинах Facebook, використаємо статтю, яку подано на рисунку 3.12.

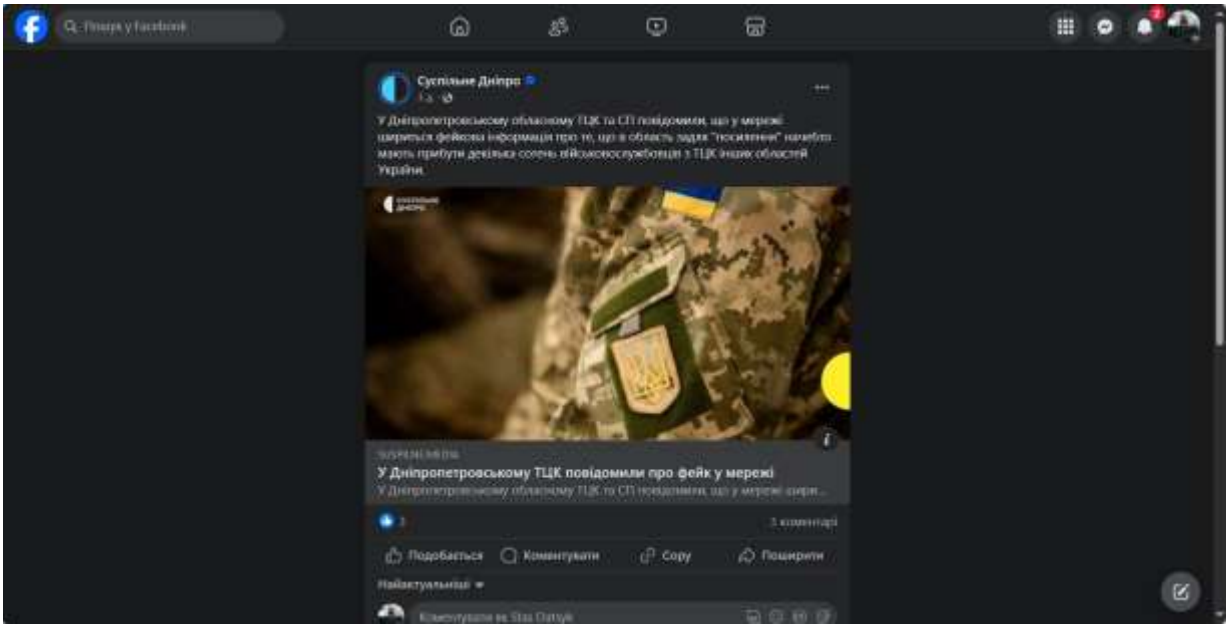

Рисунок 3.12 – Взята стаття з Facebook

Перевірка статті, яка містить дезінформацію, виконується використовуючи вищезгадані моделі. Загальний результат моделей відповідають ймовірності класу підробки, що підтверджує наявність дезінформації, результат подано на рисунку 3.13.

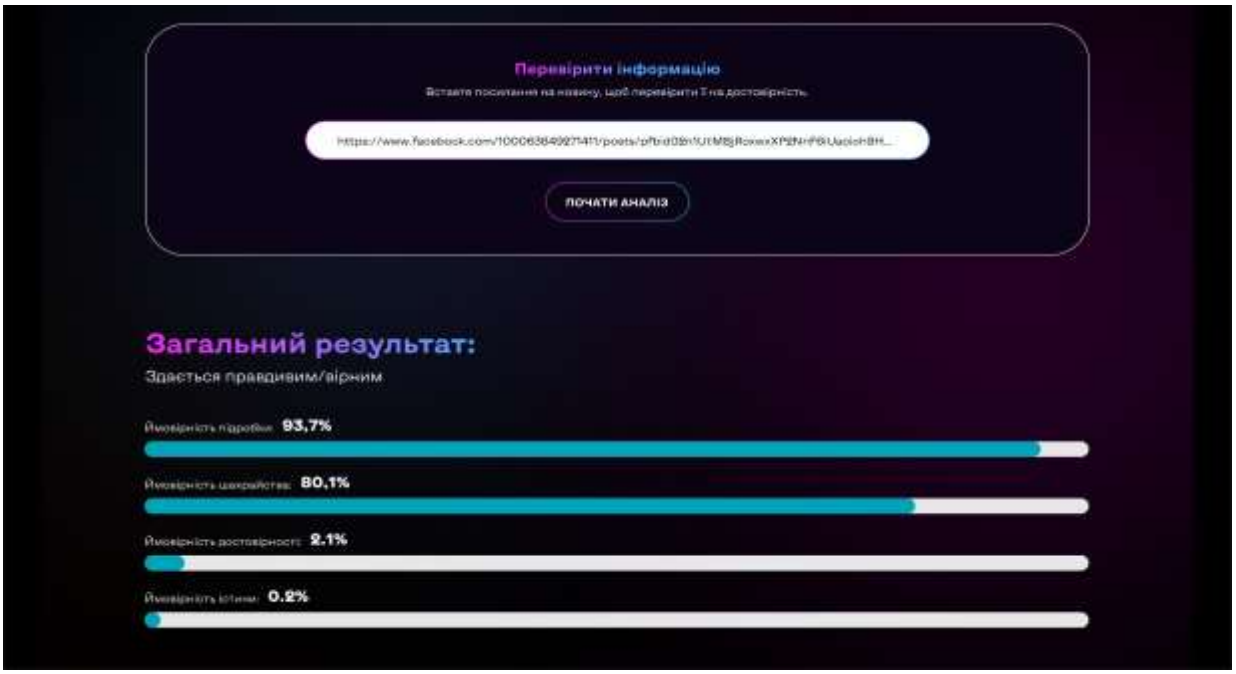

Рисунок 3.13 – Аналіз дезінформації

Згідно результатів моделі SVC отримано ймовірність класу підробки, що підтверджує наявність дезінформації, результат подано на рисунку 3.14.

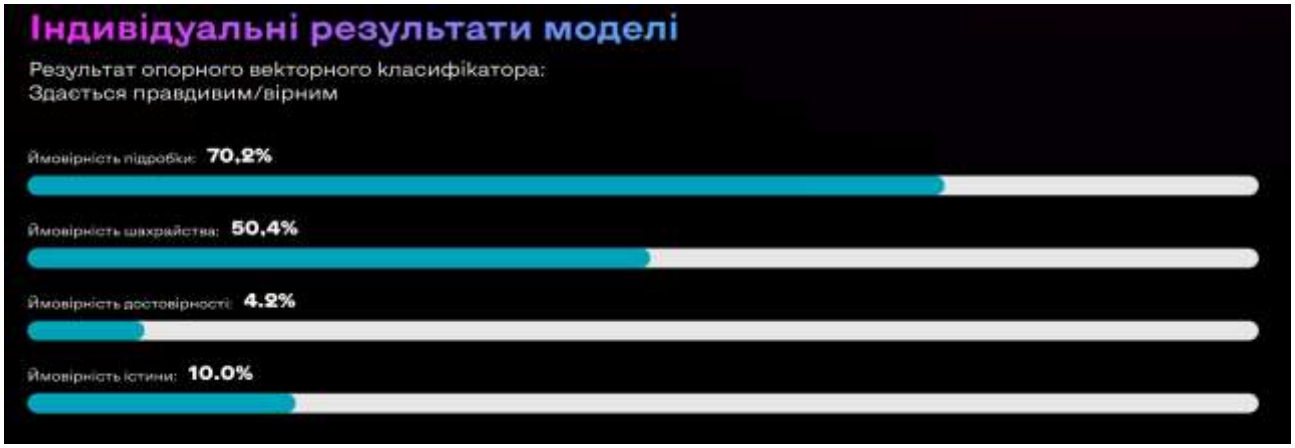

Рисунок 3.14 – Модель Support Vector Classifier

Згідно результатів моделі Logistic Regression отримано ймовірність класу шахрайства, що підтверджує наявність дезінформації, результат подано на рисунку 3.15.

Рисунок 3.15 – Модель Logistic Regression

Згідно результатів моделі Multilayer Perceptron отримано ймовірність класу підробки наближеної до шахрайства, що підтверджує наявність дезінформації, результат подано на рисунку 3.16.

Рисунок 3.16 – Модель Multilayer Perceptron

Таким чином було успішно створено AI-систему виявлення фейкових новин, яка використовує методи NLP. Включаючи кілька етапів, таких як збір даних, попередня обробка, векторизація тексту, навчання моделі та оцінки результатів. Ця система може бути корисною для виявлення дезінформації в новинах соціальної мережі Facebook.

3.4 Висновок до третього розділу

У третьому розділі кваліфікаційної роботи було описано архітектуру розробленої AI-системи для виявлення дезінформації в новинах соціальної мережі Facebook. Досліджено методи машинного навчання та подано опис проведеного експерименту роботи системи. Робота розпочалася з гіпотези, яка вимагає глибокого розуміння машинного навчання та здатності машини до обробки складних даних. Завдяки інноваційному підходу до збору та обробки даних, проект досяг певного успіху у виявленні дезінформації.

Завершення цього проекту не тільки забезпечило розуміння процесу машинного навчання, але й відкрило нові можливості для подальших досліджень та вдосконалення. Переконавшись у потенціалі розробленого класифікатора, особливо з урахуванням його здатності до інтерпретації результатів, можна стверджувати, що можливості застосування системи є безмежними. Це відкриває шлях для нових інноваційних рішень у сфері машинного навчання, які можуть бути використані у різноманітних галузях.

4 ОХОРОНА ПРАЦІ ТА БЕЗПЕКА В НАДЗВИЧАЙНИХ СИТУАЦІЯХ

4.1 Безпека з охорони праці та організація робочого місця користувачів ПК

В даній роботі здійснюється використання штучного інтелекту для виявлення дезінформації в новинах соціальної мережі Facebook. Тому важливе питання вимог до ергономіки робочого місця користувачів ПК, вимог до безпеки з охорони праці та профілактичних медичних оглядів для працівників ПК.

Робота з комп'ютером характеризується значною розумовим та нервово-емоційним навантаженням операторів, високою напруженістю зорової роботи та досить великим навантаженням на м'язи рук при роботі з клавіатурою ПК. Велике значення має раціональна конструкція та розташування елементів робочого місця, що важливо для підтримки оптимальної робочої пози людини оператора [51].

У процесі роботи з комп'ютером необхідно дотримуватися правильного режиму праці та відпочинку. В іншому випадку у персоналу відзначаються значне напруження зорового апарату з появою скарг на незадоволеність роботою, головний біль, дратівливість, порушення сну, втому і хворобливі відчуття в очах, в попереку, в області шиї та руках [51].

Вимоги щодо організації та обладнання робочих місць при розробці програмного забезпечення включає в себе площу, відведену на одне робоче місце не менше 6 м2 . Конструкція робочого місця повинна забезпечувати підтримання оптимальної робочої пози, тобто такої, яка дозволяє працівникові при написанні коду на ПК виконувати роботу з мінімальним напруженням тіла, і яка дозволяє уникнути перевтоми в ході і після закінчення робочого процесу [52].

Раціональна робоча поза має важливе значення для збереження здоров'я працівника, оскільки тривале перебування його в незручній і напруженій позі може призвести до таких захворювань, як сколіоз, варикозне розширення вен, плоскостопість тощо. Установлено, що робота в зігнутому положенні збільшує

затрати енергії на 20%, а при значному нахиленні – на 45% порівняно з прямим 79 положенням корпуса.

За потреби особливої концентрації уваги під час виконання робіт з написання коду для проекту суміжні робочі місця необхідно відділяти одне від одного перегородками висотою 1,5 – 2 м [53]. Забарвлення приміщень та меблів має сприяти створенню сприятливих умов для зорового сприйняття та позитивного настрою.

Джерела світла, такі як світильники та вікна, які дають відображення від поверхні екрана, значно погіршують точність знаків і спричиняють перешкоди фізіологічного характеру, які можуть виразитися у значній напрузі, особливо при тривалій роботі.

Недостатність освітлення призводить до напруги зору, послаблює увагу, призводить до настання передчасної втоми. Надмірно яскраве освітлення викликає засліплення, роздратування та різь в очах. Неправильний напрямок світла робочому місці може створювати різкі тіні, відблиски, дезорієнтувати працюючого. Всі ці причини можуть призвести до нещасного випадку або профзахворювань, тому настільки важливим є правильний розрахунок освітленості.

Відображення, включаючи відображення від вторинних джерел світла, має бути мінімальним. Для захисту від надмірної яскравості вікон можуть бути використані штори та екрани.

Робочі місця слід розташовувати відносно джерела природного світла, тобто вікон, таким чином, щоб світло падало на клавіатуру програміста збоку, переважно зліва. Також робоче місце для роботи на ПК має відповідати сучасним вимогам ергономіки [52]:

• стіл повинен мати висоту поверхні 680 - 800 мм, ширину 600 - 1400 мм і глибину 800 - 1000 мм. Такі параметри забезпечують можливість виконання операцій в зоні досяжності працівника;

• робочий стілець робочий стілець має бути підйомно-поворотним, з можливістю регулювання висоти, бажано зі стаціонарними або змінними

підлікотниками і напівм’якою нековзкою поверхнею сидіння, що легко чиститься і не електризується; 80

• екран комп’ютера має розташовуватися на оптимальній відстані від користувача з урахуванням літерно-цифрових знаків і символів. Стандартне значення становить 600 – 700 мм.

Розміщення принтера або іншого пристрою введення-виведення інформації на робочому місці має забезпечувати добру видимість монітору, зручність ручного керування пристроєм введення-виведення інформації. Техніка безпеки для програміста – це запорука довготривалої та безпроблемної роботи такого фахівця.

Техніка безпеки програмістів регулюється «Інструкцією з охорони праці», де все розкладено за пунктами та дуже докладно описано. Знати її потрібно, якщо програміст працює у великій офісній будівлі, де до комп’ютера мають непрямий доступ кілька людей. У цьому випадку він зобов’язаний слідувати інструкціям техніки безпеки, щоб не наражати на небезпеку своє здоров’я і здоров’я оточуючих його колег. Плюс програміст просто повинен знати, як поводитися під час надзвичайних ситуацій.

Зазвичай техніка безпеки програміста зачитується фахівцями з безпеки праці кожної організації, де працюють програмісти, чи вона доступна прочитання кожним співробітником. А щорічно, іноді й частіше, всі співробітники розписуються, що ознайомлені з технікою безпеки. Фактично техніку безпеки мало хто читає, мало хто знає і мало хто дотримується, тому що все обмежується тим, що програмісти просто розписуються в журналі, ніби вони «ознайомлені» і спокійно працюють далі.

Коли програміст працює віддалено з дому, вся відповідальність за його безпеку лежить на його плечах. Коли програміст працює в офісі, то відповідальність за його безпеку лежить на плечах програміста та окремого спеціаліста, який має слідкувати за дотриманням техніки безпеки. І в тому, і в іншому випадку програміст повинен знати основи охорони праці, описані нижче.

Вся техніка безпеки для програміста поділяється на кілька етапів: до початку роботи, під час роботи, після закінчення роботи; у разі аварійної ситуації.

Програміст перед початком своєї роботи повинен:

1) провести огляд свого робочого місця;

2) провести регулювання освітлення, щоб екран був добре видно і не відображав світло;

3) проконтролювати коректне підключення електричних частин комп’ютера до мережі;

4) проконтролювати відсутність оголених частин проведення на електричних проводах комп’ютера;

5) провести перевірку цілісності столу, стільця, підставки для ніг, висувної частини столу для клавіатури тощо; якщо потрібно, програміст повинен відрегулювати всі ці моменти.

Під час своєї роботи програміст повинен:

1) стежити за чистотою свого робочого місця;

2) не закривати вентиляційні вікна комп’ютера;

3) коректно припиняти роботу комп'ютера, коли це необхідно;

4) стежити за дотриманням свого графіка роботи та відпочинку;

5) правильно та за призначенням експлуатувати комп’ютер та всі його частини;

6) вчасно виконувати фізичні вправи для очей, шиї, рук та тулуба;

7) стежити за своїм розташуванням на робочому місці: правильна постава, відстань до та екрану.

Після закінчення робочого дня або свого робочого часу програміст повинен:

1) правильно завершити роботу всіх запущених програм та пристроїв;

2) перевірити відсутність у дисководах дисків чи дискет;

3) відключити системний блок від електромережі;

4) вимкнути додаткові пристрої від електромережі;

5) оглянути своє робоче місце і привести його до ладу, якщо це необхідно.

Перш за все, відповідно до ст. 169 Кодексу законів про працю України та ст. 17 Закону України «Про охорону праці» від 2002 року роботодавець зобов’язаний за свої кошти організувати проведення попереднього (при прийнятті на роботу) і періодичних (протягом трудової діяльності) медичних оглядів працівників, зайнятих на важких роботах, роботах зі шкідливими чи 82 небезпечними умовами праці або таких, де є потреба у професійному доборі, а також щорічного обов’язкового медичного огляду осіб віком до 21 року.

Метою проведення обов’язкових профілактичних медичних оглядів є запобігання розповсюдженню інфекційних та небезпечних захворювань, динамічне спостереження за станом здоров'я працюючого населення.

Такий вид оглядів передбачений статтею 21 Закону України «Про захист населення від інфекційних хвороб» та статтею 26 Закону України «Про забезпечення санітарного та епідемічного благополуччя населення».

Таким чином працівники окремих професій, виробництв та організацій, діяльність яких пов’язана з обслуговуванням населення і може призвести до поширення інфекційних хвороб, повинні проходити обов’язкові попередні (до прийняття на роботу) і періодичні профілактичні медичні огляди.

4.2 Підвищення стійкості роботи об’єктів приладобудування у воєнний час

Проблематика стійкості об’єктів приладобудування у воєнний час є критичним фактором, який впливає на безпеку та ефективність систем, що покладаються на точні дані та контроль.

За загальною теоретикою, стійкість об’єктів приладобудування залежить від багатьох факторів: якість компонентів, процедури ремонту та обслуговування, аналіз навантаження та надмірностей, застосування методів надання послуг, захищеність від надзвичайних ситуацій, здатність інженерно-технічного комплексу протистояти руйнуванню, надійність постачання енергетики та інших ресурсів, підготовлення до аварійно-рятовних та відновлюваних робіт. У воєнний час це означає, що об’єкти приладобудування

повинні бути стабільними та надійними у будь-яких умовах, навіть якщо вони 83 знаходиться під загрозою, або пошкоджене ворожим нападом, або умовами навколишнього середовища.

Моделі забезпечення стійкості об’єктів приладобудування – це способи описати та аналізувати поведінку і функціонування таких об’єктів у різних умовах і ситуаціях. Вони можуть мати різну складність і глибину, але загальною метою є визначити потенційні ризики і надзвичайні ситуації, що можуть загрожувати стабільності і надії об’єктів приладобудування.

Існує ряд методів покращити стабільність об’єктів приладобудування у воєнний час.

Проектування інженерно-технічних заходів цивільного захисту (ІТЗЦ), які передбачають врахування можливих надзвичайних ситуацій та їх наслідків для об’єкта, а також розробку ефективних дій для запобігання або зменшення їх впливу [54].

Використання спеціальних захисних споруд і особливих конструкцій на радіаційно-, вибухо- і пожежонебезпечних об’єктах, будівництво дамб і обвалування в районах можливих затоплень, укріплення схилів у районах з підвищеним ризиком зсувів.

Також до конструкційної стійкості можна віднести розташування об’єктів важливих для оборони у місцях з природними або штучними перешкодами, які ускладнюють атаку. Використання твердих матеріалів та конструкцій, які можуть витримувати великі навантаження. Застосування технік маскування для затруднення визначення місцезнаходження об’єкта [55].

Крім того, можна розташування багато функціональних об’єктів поруч з цільовим об’єктом для створення плутанини та ускладнення визначення цілей.

Застосування нових технологій, матеріалу для покращення якості ефективності обладнання і систем приладобудування, наприклад: електронне управління, автоматизації, механизмизації, оптимизації.

Використання прогресивного управління для передбачення потенційних проблем на полісі і запровадження заходів для їх запобігання. Наприклад,

регулярний аналіз споживання запасних частин і перевірка їх сумісності з всіма об’єктами на полісі.

Надання послуг для оптимізації ремонту та обслуговування приладів. Наприклад, автоматизована інспекція та моніторинг для виявлення та усунення несправностей. Сюди ж можна додати застосування мобільних та легко пересувних систем для уникнення ворожих атак та ускладнення визначення місцезнаходження. Використання техніки та транспортних засобів з підвищеною мобільністю.

Моделювання для аналізу та оптимізації параметрів і поведінки приладів. Наприклад, використовувати математичні методи та комп'ютерне моделювання для проектування та симуляції систем.

У воєнний час забезпечення економічної та соціальної стійкості об’єктів приладобудування може виявитися складним завданням через можливі економічні труднощі та соціальні виклики. Деякі моделі та методи для забезпечення стійкості у воєнний час включають економічне планування та ресурсне управління, соціальну політику та соціальний захист, створення резервів та запасів, господарську диверсифікацію, інфраструктурну стійкість, економічне збалансування та управління ризиками. Застосування цих моделей спрямоване на створення комплексного підходу до забезпечення стійкості у воєнний час, охоплюючи економічний та соціальний аспекти. Специфічні заходи можуть залежати від конкретних умов та вимог об’єкта приладобудування.

Метою економічного планування та управління ресурсами є забезпечення оптимального використання ресурсів і фінансів для підтримки стабільності економічної системи.

Соціальна політика та соціальний захист спрямовані на зменшення негативного впливу воєнних подій на населення та соціальну сферу. Це відбувається через запровадження систем соціального забезпечення, надання гуманітарної допомоги, розробки програм зайнятості та навчання [56].

Створення резервів і техніки спрямоване на забезпечення доступу до стратегічних ресурсів і техніки у воєнний час. Результат підходу може бути 85

досягнутий шляхом створення національного запасу критично важливих ресурсів, які можна використовувати для подолання економічних труднощів.

Економічна диверсифікація має на меті зменшити залежність від окремих галузей чи ринків і забезпечити економічну стабільність. Це досягається за рахунок розвитку різних секторів економіки, просування інновацій та відкриття нових ринків.

Відмовостійкість інфраструктури спрямована на забезпечення функціональності критичної інфраструктури під час війни. Для цього необхідні такі заходи, як захист критично важливих об’єктів, розробка планів аварійного відновлення та створення резервних систем зв’язку та електроенергії.

Економічний баланс і управління ризиками спрямовані на мінімізацію економічних ризиків і забезпечення стабільності в неспокійному економічному середовищі. Це досягається шляхом проактивного управління ризиками, аналізу та адаптації до мінливих економічних умов.

4.3 Висновок до четвертого розділу

У четвертому розділі кваліфікаційної роботи були детально розглянуті основні аспекти охорони праці, що включають ергономічні вимоги до робочого місця, оформлення робочого кабінету, а також необхідні умови для комфортної та безпечної роботи за ПК. Було висвітлено фактори, які можуть мати як позитивний, так і негативний вплив на працівників під час їхньої діяльності, а також обговорено важливість дотримання вимог безпеки та регулярного проведення профілактичних медичних оглядів для операторів ПК.

Окрім того, було розглянуто способи забезпечення стійкості роботи об’єктів приладобудування у воєнний час, що є актуальним та важливим аспектом у сучасних умовах. Це дозволяє не тільки забезпечити безпеку та здоров’я працівників, але й підтримувати високий рівень продуктивності та надійності роботи обладнання.

ВИСНОВКИ

Штучний інтелект є важливим для ідентифікації дезінформації в соцмережах, забезпечуючи доступ до перевіреної інформації.

В кваліфікаційної роботі освітнього рівня «Магістр»:

– Подано огляд ролі штучного інтелекту в боротьбі з дезінформацією та фейковими новинами.

– Розглянуто використання штучного інтелекту для аналізу лінгвістичних шаблонів у соціальних мережах.

– Висвітлено необхідність комбінованої стратегії людини та штучного інтелекту.

– Проаналізовано ефективність інструментів, таких як Grover, у виявленні фейкових новин.

– Досліджено важливість регулювання поширення дезінформації та запобіжні заходи для правильного використання штучного інтелекту в інфопросторі.

– Описано методи штучного інтелекту для виявлення дезінформації.

– Досліджено інструменти штучного інтелекту для аналізу текстових даних та ідентифікації дезінформації.

– Подано порівняльний опис інструментів штучного інтелекту Sphere, CrowdTangle, Factmata та NewsGuard виявлення фекових новин.

– Розроблено архітектуру AI-системи виявлення дезінформації в новинах Facebook.

– Запропоновано модель на основі штучного інтелекту, призначену для виявлення та боротьби з дезінформацією у Facebook.

– Розроблено та протестовано систему штучного інтелекту для виявлення дезінформації в новинах соціальної мережі Facebook

У розділі «Охорона праці та безпека в надзвичайних ситуаціях» проаналізовано ергономічні вимоги до робочого місця. Описано важливість дотримання вимог безпеки та проведення профілактичних медичних оглядів.

ПЕРЕЛІК ДЖЕРЕЛ

1. Ідеальний брехун і вибори. Як ШІ стає зброєю масового зброєю [Електронний ресурс] // Надія Баловсяк. – 2024. – Режим доступу до ресурсу: www.ukr.net/news/details/technologies/102367264.html

2. ШІ у боротьбі з фейками: чи помічник він журналісту?. [Електронний ресурс] // Милослава Карпенко. – 2024. – Режим доступу до ресурсу: https://mediakrytyka.lnu.edu.ua/

3. Нейромережа виявлятиме фейкові новини, згенеровані штучним інтелектом. ?. [Електронний ресурс] // MediaSapiens. – 2019. – Режим доступу до ресурсу: https://ms.detector.media/it-kompanii/post/23362/2019-08-19-neyromerezha-vyyavlyatyme-feykovi-novyny-zghenerovani-shtuchnym-intelektom/

4. Фейки за допомогою штучного інтелекту та маніпуляції про мобілізацію: добірка тижня ?. [Електронний ресурс] // Павло Новик. – 2023. – Режим доступу до ресурсу: https://ye.ua/syspilstvo/66671_Feyki_za_dopomogoyi_shtuchnogo_intelektu_ta_manipulyaciyi_pro_mobilizaciyi__dobirka_tizhnya.html

5. I. Strutynska, H. Kozbur, L. Dmytrotsa, O. Hlado, I. Kozbur, N. Gashchyn: Analysis of the SMEs' Digitalization State Using HIT Index and Machine Learning Technique. 13th International Conference on Advanced Computer Information Technologies (ACIT). Publisher: IEEE. Institute of Electrical and Electronics Engineers Inc. Wroclaw,Poland. - p. 332-337 (Scopus). URL: https://ieeexplore.ieee.org/document/10275519

6. I. Strutynska, L. Dmytrotsa, H. Kozbur, L.Melnyk, R. Sherstiuk: The Unification of Approaches to Measuring the Digital Maturity of Business Structures (International and Domestic Approaches Volume I: Main Conference, PhD Symposium, and Posters, Kherson, Ukraine, September 28 - October 2, 2021. CEUR Workshop Proceedings. ICTERI 2021: pp. 10-23. URL: https://ceur-ws.org/Vol-3013/20210010.pdf

7. I.. Strutynska, L. Dmytrotsa, H. Kozbur, L. Melnyk: The Digital Business Transformation Index Determining and Monitoring: Development of a National Online

Platform. Theoretical and Applied Problems, Ternopil, Ukraine, November 16-18, 2021. CEUR Workshop Proceedings 3039, CEUR-WS.org ITTAP 2021: pp. 327-334. URL: https://ceur-ws.org/Vol-3039/short33.pdf

8. Штучний інтелект створює гостру проблему фейкових відгуків [Електронний ресурс] // Backstories. – 2023. – Режим доступу до ресурсу: https://www3.nhk.or.jp/nhkworld/uk/news/backstories/2582/

9. Усе, що потрібно знати про реакцію Google на AI-контент. [Електронний ресурс] // Ranktraker – 2022. – Режим доступу до ресурсу: https://www.ranktracker.com/uk/blog/everything-you-need-to-know-about-google%E2%80%99s-reaction-to-ai-content/

10. Удосконалений метод виявлення фейкових новин [Електронний ресурс] // ХНУ. – 2023. – Режим доступу до ресурсу: https://elar.khmnu.edu.ua/statistics/items/2108fdbc-9020-4738-9ab5-e16fda203679

11. Аналіз методів навчання та інструментів нейромереж [Електронний ресурс] // Cybersecurity. – 2023. – Режим доступу до ресурсу: https://csecurity.kubg.edu.ua/index.php/journal/article/view/464/369

12. 5 найкращих інструментів і методів виявлення Deepfake. [Електронний ресурс] // Unite.ai. – 2024. – Режим доступу до ресурсу: www.unite.ai/uk/best-deepfake-detector-tools-and-techniques/

13. Аналіз методів виявлення дезінформації в соціальних мережах за допомогою машинного навчання. [Електронний ресурс] // Максим Марценюк. – 2023. – Режим доступу до ресурсу: https://csecurity.kubg.edu.ua/index.php/journal/article/view/537

14. Відповідальний ШІ: Вирішальна роль спостерігачів ШІ у протидії дезінформації про вибори. [Електронний ресурс] // Доктор Асад Аббас. – 2023. – Режим доступу до ресурсу: http://surl.li/tscun

15. ШІ у боротьбі з фейками: чи помічник він журналісту?. Медіакритика. [Електронний ресурс] // Вікторія Стень. – 2023. – Режим доступу до ресурсу: https://mediakrytyka.lnu.edu.ua/novi-tehnologii-media/shi-u-borotbi-z-feykamy-chy-pomichnyk-vin-zhurnalistu.html

16. OpenAI запропонує новий метод навчання моделей штучного інтелекту для боротьби з дезінформацією. [Електронний ресурс] // Getty Images. – 2024. – Режим доступу до ресурсу: https://forbes.ua/news/openai-zaproponue-noviy-metod-navchannya-modeley-shtuchnogo-intelektu-dlya-borotbi-z-dezinformatsieyu-01062023-13956

17. Медіаклони й дезінформація. Лайфхаки та інструменти для боротьби з фейками. [Електронний ресурс] // Лілія Мицко. – 2023. – Режим доступу до ресурсу: https://mediamaker.me/mediaklony-j-dezinformacziya-lajfhaky-ta-instrumenty-dlya-borotby-z-fejkamy-5540/

18. ШІ виявлення ворожих висловлювань для боротьби зі стереотипами та дезінформацією [Електронний ресурс] // Хазіка Саджид. – 2023. – Режим доступу до ресурсу: http://surl.li/rrojp

19. США розробили систему на основі ШІ для протидії російській дезінформації [Електронний ресурс] // Голос Америки.. – 2023. – Режим доступу до ресурсу: https://texty.org.ua/fragments/109623/ssha-rozrobyly-systemu-na-osnovi-shi-dlya-protydiyi-rosijskij-dezinformaciyi/

20. Meta створить команду для боротьби з дезінформацією [Електронний ресурс] // Антіна Прасад. – 2024. – Режим доступу до ресурсу: https://forbes.ua/news/meta-stvorit-komandu-dlya-borotbi-z-dezinformatsieyu-ta-zlovzhivannyami-shi-na-viborakh-u-es-26022024-19465

21. Україна готова ділитися з партнерами досвідом [Електронний ресурс] // Укрінформ. – 2024. – Режим доступу до ресурсу: https://www.ukrinform.ua/rubric-society/3829489-ukraina-gotova-dilitisa-z-partnerami-dosvidom-borotbi-z-rosijskou-dezinformacieu.html

22. Дезінформація та штучний інтелект: (не)видима загроза сучасності [Електронний ресурс] // Ольга Петрів. – 2023. – Режим доступу до ресурсу: https://cedem.org.ua/analytics/dezinformatsiya-shtuchnyi-intelekt/

23. Gas Consumption Forecasting Using Machine Learning Methods and Taking into Account Climatic Indicators. Shymchuk, G., Lytvynenko, I., Hromyak, R., Lytvynenko, S., Hotovych, V. The 1st International Workshop on Computer

Information Technologies in Industry 4.0, CITI 2023. Ternopil 14 -16 June 2023. Vol. 3468, pp. 156-163. URL: https://ceur-ws.org/Vol-3468/short8.pdf

24. Simulation of cyclic signals (generalized approach). Lupenko, S., Lytvynenko, I., Hotovych, V. 4th International Conference on Informatics and DataDriven Medicine, IDDM 2021. Valencia. 19 November 2021. CEUR Workshop Proceedings. Vol. 3038, P. 86-92. ISSN 1613-0073 URL: https://ceur-ws.org/Vol3038/short2.pdf

25. Lupenko S., Lytvynenko Ia., Hotovych V., Zozulia A., Chizoba N., Volyanyk O. (2021) Concept of design, requirements and generalized architectures of components of the integrated onto-oriented information environment of simulation and processing of cyclic signals. Scientific Journal of TNTU (Tern.), vol 102, no 2, pp. 147–160.

26. Штучний інтелект на службі пропаганди. [Електронний ресурс] // Юлія Лавришин. – 2023. – Режим доступу до ресурсу: https://ms.detector.media/kiberbezpeka/post/33149/2023-10-07-shtuchnyy-intelekt-na-sluzhbi-propagandy/

27. Поміж Сціллою та Харібдою. Як штучний інтелект бореться з дезінформацією і поширює її. [Електронний ресурс] // Гала Скляревська. – 2022. – Режим доступу до ресурсу: https://www.stopfake.org/uk/pomizh-stsilloyu-ta-haribdoyu-yak-shtuchnij-intelekt-boretsya-z-dezinformatsiyeyu-i-poshiryuye-yiyi/

28. ШІ може боротися з дезінформацією та упередженістю в новинах. [Електронний ресурс] // Алекс МакФарланд. – 2022. – Режим доступу до ресурсу: https://www.unite.ai/uk/ai-can-combat-misinformation-and-bias-in-news/

29. Дезінформація штучним інтелектом – нам потрібен новий захист. [Електронний ресурс] // Rodion Shkurko. – 2024. – Режим доступу до ресурсу: https://rodionshkurko.com/blog/dezinformaciya-shtuchnim-intelektom-nam-potriben-novij-zahist

30. Європа переймається ШІ та ризиками дезінформації. [Електронний ресурс] // Розі Біргардт. – 2023. – Режим доступу до ресурсу: https://www.dw.com/uk/evropa-perejmaetsa-si-ta-rizikami-dezinformacii/a-65869238

31. Які загрози несе ChatGPT і як розпізнати текст, написаний ШІ. [Електронний ресурс] // МЕДІАЛАЙФХАК. – 2023. – Режим доступу до ресурсу: https://internews.ua/opportunity/ChatGPT-threats-and-how-to-recognize-AI-written-text

32. У США розробили систему на основі ШІ для протидії російській дезінформації. [Електронний ресурс] // Юлія Поліковська. – 2023. – Режим доступу до ресурсу: https://ms.detector.media/internet/post/31909/2023-05-11-u-ssha-rozrobyly-systemu-na-osnovi-shi-dlya-protydii-rosiyskiy-dezinformatsii/

33. Де журналісти можуть залучити штучний інтелект – кейси світових видань. [Електронний ресурс] // Вероніка Нановська. – 2023. – Режим доступу до ресурсу: https://mediamaker.me/de-zhurnalisty-mozhut-zaluchyty-shtuchnyj-intelekt-kejsy-svitovyh-vydan-4797/

34. Дезінформація, ШІ, війна в Україні: Всесвітній економічний форум назвав ризики найближчих 2 років. [Електронний ресурс] // Радіо Свобода. – 2024. – Режим доступу до ресурсу: https://www.radiosvoboda.org/a/news-dezinformatsia-viina-ryzyky/32768617.html

35. Людям складно виявляти дезінформацію, створену ШІ, – дослідження. ms.detector.media [Електронний ресурс] // Юлія Поліковська. – 2023. – Режим доступу до ресурсу: https://ms.detector.media/internet/post/32314/2023-06-30-lyudyam-skladno-vyyavlyaty-dezinformatsiyu-stvorenu-shi-doslidzhennya/

36. Комплексний огляд блокчейну в ШІ. [Електронний ресурс] // Кунал Кейрівал. – 2023. – Режим доступу до ресурсу: www.unite.ai/uk/a-comprehensive-review-of-blockchain-in-ai/

37. 8 способів використання штучного інтелекту (AI) в розробці мобільних додатків. [Електронний ресурс] // Cases. – 2023. – Режим доступу до ресурсу: https://cases.media/article/8-sposobiv-vikoristannya-shtuchnogo-intelektu-ai-v-rozrobci-mobilnikh-dodatkiv

38. Інженери ШІ розробили метод, який може виявити наміри тих, хто поширює дезінформацію. [Електронний ресурс] // Деніель Нельсон. – 2022. – Режим доступу до ресурсу: https://www.unite.ai/uk/ai-engineers-develop-method-that-can-detect-intent-of-those-spreading-misinformation/

39. 5 найкращих інструментів і методів виявлення Deepfake. [Електронний ресурс] // Алекс МакФарланд. – 2024. – Режим доступу до ресурсу: www.unite.ai/uk/best-deepfake-detector-tools-and-techniques/

40. Штучний інтелект і дезінформація: можливості та ризики в умовах війни [Електронний ресурс] // УКРІНФОРМ. – 2023. – Режим доступу до ресурсу: https://www.ukrinform.ua/rubric-technology/3691961-stucnij-intelekt-i-dezinformacia-mozlivosti-ta-riziki-v-umovah-vijni.html

41. Поради та інструменти для виявлення дезінформації про напад Росії на Україну. [Електронний ресурс] // Досвід Інновацій. – 2023. – Режим доступу до ресурсу: https://redactor.in.ua/2022/05/11/porady-ta-instrumenty-dlya-vyyavlennya-dezinformacziyi-pro-napad-rosiyi-na-ukrayinu/

42. Порівняльний аналіз систем виявлення і запобігання дезінформації [Електронний ресурс] // Олександр Риков. – 2019. – Режим доступу до ресурсу: https://openarchive.nure.ua/server/api/core/bitstreams/381a4953-9306-4362-a6d1-e7199f2f425a/content

43. Порівняльний аналіз ефективності методів виявлення дезінформації [Електронний ресурс] // Даценко, М. Д. – 2020. – Режим доступу до ресурсу: https://ir.nmu.org.ua/handle/123456789/157864

44. Вивчення ролі штучного інтелекту в академічних [Електронний ресурс] // Джесіка Аббадія. – 2023. – Режим доступу до ресурсу: https://mindthegraph.com/blog/uk/ai-in-academic-research/

45. Інтерактивна діаграма зміщення медіа [Електронний ресурс] // Ad Ad Fontes Media. – 2024. – Режим доступу до ресурсу: https://adfontesmedia.com/interactive-media-bias-chart/

46. Набір даних для виявлення фейкових новин [Електронний ресурс] // FakeNewsNet. – 2021. – Режим доступу до ресурсу: https://github.com/KaiDMML/FakeNewsNet

47. Meta VS ШІ: як компанія планує боротися з дезінформацією – [Електронний ресурс] // Владислав Гринів. – 2024. – Режим доступу: https://speka.media/meta-vs-si-yak-kompaniya-planuje-borotis-z-dezinformacijeyu-p6m8kw

48. НАТО Ревю – Протидія дезінформації: посилення цифрової стійкості Альянсу – [Електронний ресурс]] // Johns Hopkins. – 2021. – Режим доступу: https://www.nato.int/docu/review/uk/articles/2021/08/12/protidya-deznformats-posilennya-tsifrovo-stjkost-al-yansu/index.html

49. Боротьба з дезінформацією у соцмережах: погляд зі Сполучених Штатів – [Електронний ресурс] // Павло Будряк. – 2024. – Режим доступу: https://cedem.org.ua/analytics/borotba-dezinformatsiya-usa/

50. Дезінформація у соцмережах: ЄС попереджає X, Meta і TikTok – [Електронний ресурс] // Люсія Шультен. – 2024. – Режим доступу: https://www.dw.com/uk/dezinformacia-u-socmerezah-es-poperedzae-h-meta-i-tiktok/a-67120427

51. В.І. Голінько, М.Ю. Іконніков, Я.Я. Лебедєв. Охорона праці в галузі інформаційних технологій / В.І. Голінько, М.Ю. Іконніков, Я.Я. Лебедєв. – Дніпропетровськ: НГУ, 2015 – 97 с.

52. Робота за комп’ютером, наслідки та поради [Електронний ресурс] // АПАУ. – 2021. – Режим доступу до ресурсу: https://cutt.ly/twDgKv06

53. Організація робочого місця оператора з обробки інформації та програмного забезпечення [Електронний ресурс] // Сонгрова О. В. – 2021. – Режим доступу до ресурсу: https://naurok.com.ua/organizaciya-robrchogo-miscyaoperatora-z-obrobki-informaci-ta-programnogo-zabezpechennya-249780.html

54. Організація робочого місця оператора ПК [Електронний ресурс] // Studcon. – 2021. – Режим доступу до ресурсу: http://studcon.org/organizaciyarobochogo-miscya-operatora-pk 93 52. Охорона праці в офісі. Вимоги до робочого місця офісного працівника [Електронний ресурс] // AGN International. – 2021. – Режим доступу до ресурсу: https://gc.ua/uk/oxorona-praci-v-ofisi-vimogi-do-robochogo-miscya-ofisnogopracivnika/

55. Указ президента України 479/2021 [Електронний ресурс] // Зеленський В. – 2021. – Режим доступу до ресурсу: https://www.president.gov.ua/documents/4792021-40181

56. Указ президента України [Електронний ресурс] // Зеленський В. – 2021. – Режим доступу до ресурсу: https://zakon.rada.gov.ua/laws/show/479/2021#Text

ДОДАТКИ

Додаток А

Тези конференцій

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

ТЕРНОПІЛЬСЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ ІМЕНІ ІВАНА ПУЛЮЯ

М А Т Е Р І А Л И

XI НАУКОВО-ТЕХНІЧНОЇ КОНФЕРЕНЦІЇ

«ІНФОРМАЦІЙНІ МОДЕЛІ, СИСТЕМИ ТА ТЕХНОЛОГІЇ»

13-14 грудня 2023 року

ТЕРНОПІЛЬ
2023

УДК 316.77:004.896:17

Л.П. Дмитроца, канд.техн.наук; С.В.Дацик
(Тернопільський національний технічний університет імені Івана Пулюя, Україна)

АНАЛІЗ ІНСТРУМЕНТІВ ШТУЧНОГО ІНТЕЛЕКТУ ДЛЯ ВИЯВЛЕННЯ ДЕЗІНФОРМАЦІЇ В НОВИНАХ FACEBOOK

L.P. Dmytrotsa Ph.D, S.V. Datsyk

ANALYSIS OF ARTIFICIAL INTELLIGENCE TOOLS TO DETECT DISINFORMATION IN FACEBOOK NEWS

Поширення дезінформації на платформах соціальних мереж стало серйозною проблемою в епоху цифрових технологій. Facebook, будучи однією з найбільш поширених платформ соціальних медіа, перебуває під пильною увагою через його роль у поширенні дезінформації. Для вирішення цієї проблеми існує кілька інструментів для виявлення дезінформації в новинах Facebook. У цьому досліджені здійснено порівняння та зіставлено чотири із них, а саме: Sphere, CrowdTangle, Factmata та NewsGuard. Досліджуючи їхні переваги та недоліки, можемо краще зрозуміти ефективність цих інструментів у виявленні дезінформації в новинах Facebook.

Sphere – інструмент, розроблений дослідниками з Кембриджського університету, який використовує штучний інтелект (AI) і машинне навчання (ML) для виявлення та пом'якшення поширення дезінформації на платформах соціальних мереж, таких як Facebook [1]. Однією з переваг Sphere є його здатність точно виявляти дезінформацію та запобігати її поширенню серед широкої аудиторії. Однак покладатися виключно на штучний інтелект і машинне навчання для виявлення дезінформації може бути недостатньо. У міру розвитку тактики дезінформації може знадобитися включення інших методів, наприклад, перевірка фактів людьми, щоб забезпечити точне виявлення дезінформації [2].

CrowdTangle – інструмент публічного аналізу, розроблений Facebook, який допомагає видавцям, журналістам, дослідникам, перевіряючим факти та іншим слідкувати, аналізувати та повідомляти про те, що відбувається в соціальних мережах [3]. Однією з переваг CrowdTangle є його здатність надавати дані про поширення дезінформації в режимі реального часу, що може допомогти ідентифікувати джерело та запобігти її подальшому поширенню. Однак CrowdTangle може бути неефективним у виявленні дезінформації, оскільки він значною мірою покладається на створений користувачами контент, який не завжди може бути надійним [4].

Factmata – інструмент, який використовує ШІ та обробку природної мови (NLP) для ідентифікації та позначення дезінформації на платформах соціальних мереж, таких як Facebook [5]. Однією з переваг Factmata є його здатність аналізувати контекст і настрої публікацій, щоб точно виявляти дезінформацію. Крім того, Factmata розробила продукт Narrative Monitoring, який поєднує дві унікальні технології: кластеризацію тем і оцінку вмісту. Однак алгоритми Factmata можуть бути не в змозі виявити більш складні форми дезінформації, такі як глибокі фейки [6].

NewsGuard – інструмент, який надає незалежні, аполітичні рейтинги довіри для онлайн-джерел новин, що може допомогти користувачам визначити надійні джерела інформації [7]. Однією з переваг NewsGuard є його здатність надавати користувачам

простий і легкий у використанні інструмент для оцінки довіри до джерел новин у Facebook. Проте рейтинги NewsGuard можуть бути не зовсім точними, оскільки вони значною мірою покладаються на людське судження та можуть залежати від особистих упереджень [8].

Хоча кожен із інструментів має свої переваги та недоліки, очевидно, що жоден інструмент не може повністю усунути дезінформацію у Facebook. Натомість для ефективної боротьби з дезінформацією на цій платформі може знадобитися поєднання інструментів і стратегій [9]. Алгоритми машинного навчання, класифікаційний аналіз, регресія, кластеризація даних, розробка функцій і зменшення розмірності – інструменти, які можна використовувати для виявлення дезінформації у Facebook [10]. Зрештою, найефективнішим інструментом буде той, який зможе точно виявляти дезінформацію, а також буде зручним і доступним для широкого кола користувачів.

Висновок. Боротьба з дезінформацією в соціальних мережах є постійною проблемою, і такі інструменти, як Sphere, CrowdTangle, Factmata та NewsGuard, є важливими для боротьби з цією проблемою. Хоча кожен інструмент має свої сильні сторони та обмеження, усі вони відіграють вирішальну роль у виявленні та усуненні дезінформації у Facebook.

Література

1. Detecting fake news and disinformation using artificial intelligence and machine learning to avoid supply chain disruptions – [Електронний ресурс] – Режим доступу: https://link.springer.com/article/10.1007/s10479-022-05015-5

2. Why the Government Should Not Regulate Content Moderation of Social Media – [Електронний ресурс] – Режим доступу: https://www.cato.org/policy-analysis/why-government-should-not-regulate-content-moderation-social-media

3. A tool from Meta to help follow, analyze, and report on what's happening across social media – [Електронний ресурс] – Режим доступу: https://www.crowdtangle.com/

4. What data is CrowdTangle tracking? – [Електронний ресурс] – Режим доступу: https://help.crowdtangle.com/en/articles/1140930-what-data-is-crowdtangle-tracking

5. Artificial Intelligence in Automated Detection of Disinformation: A Thematic Analysis – [Електронний ресурс] – Режим доступу: https://www.mdpi.com/2673-5172/4/2/43

6. Factmata Narrative Monitoring. – [Електронний ресурс] – Режим доступу: https://ircai.org/top100/entry/factmata-narrative-monitoring/

7. NewsGuard Ratings. – [Електронний ресурс] – Режим доступу: https://www.newsguardtech.com/solutions/newsguard/

8. Number of Unreliable AI-Generated Sites, Labeled 'UAINs – [Електронний ресурс] – Режим доступу: https://www.newsguardtech.com/press/newsguard-now-identifies-125-news-and-information-websites-generated-by-ai-develops-framework-for-defining-unreliable-ai-generated-news-and-information-sources/

9. The Future of Truth and Misinformation Online – [Електронний ресурс] – Режим доступу: https://www.pewresearch.org/internet/2017/10/19/the-future-of-truth-and-misinformation-online/

10. Machine Learning: Algorithms, Real-World Applications – [Електронний ресурс] – Режим доступу: https://link.springer.com/article/10.1007/s42979-021-00592-x

УДК 004.7:004.8:004.9

Л.П. Дмитроца, канд.техн.наук, С.В.Дацик
(Тернопільський національний технічний університет імені Івана Пулюя, Україна)

ЗАСТОСУВАННЯ МЕТОДІВ ШТУЧНОГО ІНТЕЛЕКТУ ДЛЯ ВИЯВЛЕННЯ ТА ПРОТИДІЇ ДЕЗІНФОРМАЦІЇ У FACEBOOK

L.P. Dmytrotsa Ph.D, S.V. Datsyk

APPLICATION OF ARTIFICIAL INTELLIGENCE METHODS TO DETECT AND COUNTERACT DISINFORMATION ON FACEBOOK

В останні роки проблема дезінформації на платформах соціальних мереж стала серйозною проблемою. Як одну з найбільших платформ соціальних медіа, Facebook зазнав критики за нездатність ефективно виявляти та запобігати поширенню дезінформації. Щоб вирішити цю проблему, Facebook запровадив методи на основі ШІ для виявлення та позначення неправдивої інформації. Однак методи не позбавлені проблем. У цьому досліджені виконаємо огляд методів на основі штучного інтелекту для виявлення дезінформації у Facebook, обговоримо проблеми виявлення дезінформації за допомогою розглянутих методів і дослідимо стратегії підвищення їх точності та ефективності.

Штучний інтелект (AI) став критично важливим інструментом для виявлення та запобігання поширенню шкідливого контенту у Facebook. Згідно із заявою Facebook у листопаді 2020 року, штучний інтелект допомагає масштабувати роботу експертів-людей і підвищити ефективність модерації контенту. Facebook використовує поєднання технологій примусового контролю, перевірки людьми та методів на основі ШІ для виявлення та видалення неправдивої інформації. Використання методів штучного інтелекту для виявлення дезінформації на платформах соціальних мереж в останні роки набирає обертів. Дослідження, проведене Santos et al. у 2023 році [1] проаналізував потенційні переваги автоматизованого виявлення дезінформації з точки зору інформаційних наук. Однак, незважаючи на переваги ШІ у виявленні дезінформації, існують також значні проблеми, які необхідно вирішити.

Однією з основних проблем у виявленні дезінформації за допомогою методів на основі штучного інтелекту є поширення фейкових новин, що досі є складною невирішеною проблемою. Пандемія COVID-19 також висвітлила проблему дезінформації на платформах соціальних мереж із поширенням неправдивої інформації про вірус та його лікування. Крім того, використання змагальних прикладів, які є спеціально створеними вхідними даними, призначеними для обману моделей машинного навчання, також може стати проблемою для методів на основі ШІ для виявлення дезінформації [2]. Виклики підкреслюють необхідність стратегій для підвищення точності та ефективності методів на основі ШІ для виявлення дезінформації на платформах соціальних мереж.

Щоб підвищити точність і ефективність методів на основі ШІ для виявлення дезінформації у Facebook, були запропоновані різні стратегії. Сантос та ін. у 2023 році [2] проаналізував низку підходів, включаючи перевірку фактів, лінгвістичний аналіз, аналіз настроїв та використання систем людського циклу. Баласубраманіам та ін. у 2023 році [3] запропонував систематичний і структурований спосіб визначення вимог пояснювання систем штучного інтелекту, що може покращити прозорість та інтерпретацію методів на

основі штучного інтелекту. Лі та ін. [4] провели дослідження наслідків коментування посту дезінформації у Facebook, яке показало, що втручання агентства користувачів може бути ефективним у зниженні поширення неправдивої інформації. Попередні стратегії можуть допомогти вирішити проблеми з виявленням дезінформації за допомогою методів на основі штучного інтелекту та підвищити їх точність і ефективність.

Хоча методи на основі штучного інтелекту пропонують багатообіцяюче рішення для виявлення та запобігання поширенню дезінформації у Facebook, є також етичні міркування, які слід брати до уваги. Однією з головних проблем є можливість зміщення в алгоритмах, які використовуються в цих методах. Як зазначив Діпак у 2021 році [5], використання автоматизації на основі даних для виявлення фейкових новин може викликати етичні та нормативні міркування. Флорес у 2022 році [6] далі досліджує етичні міркування використання штучного інтелекту, наголошуючи на необхідності прозорості та підзвітності в застосовуваних методах ШІ. Крім того, Лауер у 2021 році [7] підкреслює питання свободи слова та потенціал цензури при виявленні дезінформації. Етичні міркування необхідно ретельно розглянути та розглянути, щоб переконатися, що методи на основі ШІ використовуються відповідально та етично.

Висновок. методи на основі ШІ пропонують багатообіцяюче рішення для виявлення та запобігання поширенню дезінформації у Facebook. Однак існують значні проблеми та етичні міркування, які необхідно брати до уваги. Поширення фейкових новин, використання суперечливих прикладів і потенціал упередженості в алгоритмах – все це виклики, які потребують вирішення. Такі стратегії, як перевірка фактів, лінгвістичний аналіз і втручання на основі користувачів, можуть допомогти підвищити точність і ефективність методів на основі ШІ. Крім того, прозорість і підзвітність мають бути пріоритетними для вирішення етичних міркувань, таких як потенційна упередженість і цензура.

Література

1. Artificial Intelligence in Automated Detection of Disinformation: A Thematic Analysis – [Електронний ресурс] – Режим доступу: https://www.mdpi.com/2673-5172/4/2/43

2. An Adversarial Benchmark for Fake News Detection Models – [Електронний ресурс] – Режим доступу: https://arxiv.org/pdf/2201.00912.pdf

3. Transparency and explainability of AI systems – [Електронний ресурс] – Режим доступу: https://www.sciencedirect.com/science/article/pii/S0950584923000514

4. User agency–based versus machine agency–based misinformation interventions – [Електронний ресурс] – Режим доступу: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10113910/

5. Ethical Considerations in Data-Driven Fake News Detection – [Електронний ресурс] – Режим доступу: https://link.springer.com/chapter/10.1007/978-3-030-62696-9_10

6. Ethical Considerations in the Application of Artificial Intelligence – [Електронний ресурс] – Режим доступу: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9406274/

7. Facebook's ethical failures are not accidental; they are part of the business model – [Електронний ресурс] – Режим доступу: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8179701/

МІЖНАРОДНІ МУЛЬТИДИСЦИПЛІНАРНІ
НАУКОВІ ІНТЕРНЕТ-КОНФЕРЕНЦІЇ

www.economy-confer.com.ua

Світ наукових досліджень

Збірник наукових
публікацій міжнародної
мультидисциплінарної наукової
інтернет-конференції

Випуск 28

21-22 березня 2024 р.

ISSN 2786-6823 (print)

Google Scholar

AKADEMIA NAUK STOSOWANYCH
WYŻSZA SZKOŁA ZARZĄDZANIA I ADMINISTRACJI W OPOLU

Тернопіль, Україна – Ополе, Польща
2024

УДК 001 (063)

Світ наукових досліджень. Випуск 28: матеріали Міжнародної мультидисциплінарної наукової інтернет-конференції (м. Тернопіль, Україна, м. Ополе, Польща, 21-22 березня 2024 р.) / за ред. : О. Патряк та ін. ГО "Наукова спільнота", WSZIA w Opolu. Тернопіль: ФО- П Шпак В.Б. 2024. 191 с.

Збірник наукових публікацій укладено за матеріалами доповідей наукової мультидисциплінарної інтернет-конференції «Світ наукових досліджень. Випуск 28», які оприлюднені на інтернет-сторінці www.economy-confer.com.ua

Адреса оргкомітету:
46005, Україна, м. Тернопіль, а/с 797
тел. +380977547363 e-mail: economy-confer@ukr.net

ISSN 2786-6823 (print)

Список літератури та джерел:

1. Artificial intelligence. URL: https://www.britannica.com/technology/artificial-intelligence/Reasoning
2. Artificial intelligence in medicine. Pavel Hamet, Johanne Tremblay, T 69, S36-S40; DOI: 10.1016/j.metabol.2017.01.011, [PubMed];[Google Scholar]
3. What is artificial intelligence in medicine? URL: https://www.ibm.com/topics/artificial-intelligence-medicine
4. Priorities for the priority review voucher. DB Ridley, *Am J Trop Med Hyg*, 2017 Jan 11; 96 (1): S14-S15, DOI: 10.4269/ajtmh.16-0600, [PubMed];[Google Scholar]

ШТУЧНИЙ ІНТЕЛЕКТ ПРОТИ ДЕЗІНФОРМАЦІЇ: СТРАТЕГІЇ, ВИКЛИКИ ТА ВПЛИВ НА СУСПІЛЬСТВО В УМОВАХ ІНФОРМАЦІЙНОЇ ВІЙНИ

Дмитроца Леся Павлівна
кандидат технічних наук,
Тернопільський національний технічний університет імені Івана Пулюя

Дацик Станіслав Васильович
студент, Тернопільський національний технічний університет імені Івана Пулюя

Інтернет-адреса публікації на сайті:
https://www.economy-confer.com.ua/full-article/5425/

У сучасному світі поширення дезінформації стало серйозною проблемою для суспільств у всьому світі, особливо в контексті інформаційної війни. Використання штучного інтелекту (ШІ) у боротьбі з дезінформацією в останні роки привертає увагу як багатообіцяюча стратегія. Ця дослідницька стаття має на меті вивчити стратегії, які використовує штучний інтелект для виявлення та протидії дезінформації, з акцентом на їх ефективність у контексті інформаційної війни в Україні. Стаття розпочнеться з огляду проблеми дезінформації та ролі ШІ у її вирішенні. Далі буде розглянуто ключові стратегії, які використовує штучний інтелект для боротьби з дезінформацією, і оцінено їх ефективність, з особливим акцентом на їх вплив на суспільство. Таким чином, ця стаття має на меті сприяти глибшому розумінню потенціалу штучного інтелекту в боротьбі з дезінформацією та її наслідків для суспільства.

Штучний інтелект (ШІ) швидко стає потужним інструментом для виявлення та протидії кампаніям дезінформації. Традиційні кадрові ресурси часто не в змозі досить швидко виявляти кампанії дезінформації та реагувати на них; ШІ потрібен, щоб заповнити цю прогалину, особливо у випадку дипфейків та інших складних тактик дезінформації [1]. Технічні компанії розробляють системи виявлення, спрямовані на виявлення дезінформації, як тільки вона з'являється. Mantis Analytics – це платформа на основі штучного інтелекту, яка прослуховує інформаційний простір у режимі реального часу, зосереджуючись

на управлінні фізичними та інформаційними ризиками для боротьби з російською дезінформацією. Компанія надає інструменти, які можуть допомогти тим, хто займається стратегічними комунікаціями та управлінням ризиками, швидко реагувати та боротися з дезінформацією. Одним із таких інструментів є Microsoft Video Authenticator, інструмент на базі штучного інтелекту, який аналізує зображення та відео, щоб визначити ймовірність їх підробки. Інструмент може виявляти тонкі зміни в зображеннях, які можуть бути невидимі для людського ока. ШІ також можна використовувати для фільтрації та виявлення дезінформації в ЗМІ, яку можуть швидко спростувати незалежні журналісти. Крім того, системи штучного інтелекту можуть виявляти початок поширення наративу, дозволяючи негайно вжити заходів, інформуючи партнерів у рамках щоденного звітування про ворожі інформаційні кампанії [2]. Роль штучного інтелекту у виявленні та протидії дезінформації є значною та зростаючою, і ці інструменти призначені для підтримки відповідних фахівців у здійсненні ефективної протидії дезінформації, а не безпосередньо брати участь у протидії.

Дезінформація – виклик, з яким світ продовжує боротися, особливо враховуючи те, що для її створення часто використовується штучний інтелект (ШІ). Ризики значні, зокрема підрив довіри до демократичних інститутів і процесів [3]. У відповідь на це розробляються різні стратегії боротьби з дезінформацією. НАТО бере участь у діяльності, спрямованій на розбудову стосунків і сприяння розвитку стійкості серед аудиторії, яка може стати мішенню для дезінформації. Одним із ключових аспектів використання ШІ для боротьби з дезінформацією є підзвітність, як підкреслив представник ОБСЄ з питань свободи слова. Крім того, штучний інтелект використовується для боротьби з ненавистю та стереотипами за допомогою методів, які сприяють безпечнішій онлайн-спільноті. Ще одне застосування штучного інтелекту полягає у виявленні дипфейків, які вимагають використання алгоритму під назвою кодувальник для аналізу тисяч різних зображень обличчя. Соціальне прослуховування з використанням штучного інтелекту також є ефективною стратегією виявлення та протидії дезінформації, особливо на етапі виконання. ШІ розглядається як потужний інструмент у боротьбі з дезінформацією, який може посилити людські можливості протистояти пропаганді. В Україні видано посібник, який допоможе авторам контенту, журналістам і користувачам соціальних мереж розпізнавати та протистояти російській пропаганді [4]. Загалом, використання штучного інтелекту в боротьбі з дезінформацією має важливе значення для всього процесу, від виявлення джерел дезінформації до протидії її наслідкам.

Боротьба з дезінформацією на платформах соціальних мереж, особливо на Facebook, є актуальною в умовах інформаційної війни, оскільки ці платформи використовуються для швидкого та масштабного поширення неправдивої інформації. Facebook, з його величезною аудиторією, стає ідеальним інструментом для маніпулювання громадською думкою та поширення дезінформації, що може мати серйозні наслідки для національної безпеки та суспільної стабільності. Тому важливо мати надійні інструменти для її виявлення та блокування [5].

Штучний інтелект може стати потужним інструментом у боротьбі з дезінформацією, оскільки він здатний аналізувати великі обсяги даних та виявляти підозрілі шаблони. Алгоритми машинного навчання можуть виявляти дезінформацію за допомогою розпізнавання маніпулятивних текстів та зображень, що дозволяє оперативно реагувати на спроби маніпулювання інформацією. Використання AI для моніторингу соціальних мереж допомагає виявляти та припиняти дезінформаційні кампанії, перш ніж вони зможуть суттєво вплинути на суспільство. Тому для протидії дезінформації в Україні та інших країнах необхідні більш ефективні стратегії.

Використання ШІ у боротьбі з дезінформацією має велике значення для виявлення та протидії. Необхідно розробити стратегії для відповідального та етичного використання ШІ. Для детального опису пропозиції розробки інструменту, який би використовував штучний інтелект (ШІ) для боротьби з дезінформацією на соціальних платформах в контексті інформаційної війни, можна розглянути наступні аспекти:

– Використовувати алгоритми машинного навчання для аналізу змісту новин з даними з надійних новинних баз, таких як Reuters, Associated Press, та інших міжнародних та українських новинних агентств. Це дозволить визначити, чи є відображена інформація достовірною.

– Використання аналізу сентименту допоможе визначити, чи має стаття надмірно негативний або позитивний тон, що може вказувати на упередженість або спробу маніпулювати громадською думкою. Це може бути особливо корисним для виявлення статей, які мають на меті викликати емоційну реакцію.

– Інтеграція функції перевірки фактів, яка звірятиме твердження в статті з відомими фактами та даними з перевірених баз даних, забезпечить додатковий рівень відповідальності та точності.

– Розробка інструментів для аналізу зображень та відео допоможе виявити дипфейки та інші види візуальної дезінформації. Це може включати використання алгоритмів глибокого навчання для аналізу візуальних ознак, які вказують на можливу підробку.

– Простий та інтуїтивно зрозумілий інтерфейс дозволить користувачам легко вставляти посилання на новини та отримувати швидкий аналіз. Це може включати візуальні індикатори, які показують рівень достовірності інформації.

– Партнерство з організаціями, які займаються перевіркою фактів, забезпечить точність та актуальність інформації, яка використовується для аналізу.

– Включення освітніх матеріалів та ресурсів допоможе користувачам розуміти, як визначити дезінформацію та як вона може впливати на їхнє сприйняття новин.

– Можливість для користувачів залишати зворотний зв'язок про новини допоможе покращити алгоритми виявлення дезінформації та зробити інструмент більш чутливим до потреб спільноти.

– Підтримка кількох мов забезпечить, що інструмент буде корисним для широкого кола користувачів, не залежно від їхньої рідної мови.

– Прозорість у функціонуванні алгоритмів дозволить користувачам розуміти, як відбувається аналіз, та забезпечить довіру до інструменту.

Ці аспекти можуть бути використані для створення інструменту, який не тільки ефективно виявлятиме дезінформацію, але й буде легким у використанні та доступним для широкої аудиторії.

Висновок. Застосування штучного інтелекту (ШІ) у боротьбі з дезінформацією стає все більш важливим. Платформа Mantis Analytics, заснована на ШІ, прослуховує інформаційний простір у режимі реального часу та бореться з російською дезінформацією. Однак необхідно розробити стратегії, щоб гарантувати відповідальне та етичне використання ШІ в цілях протидії дезінформації. Штучний інтелект (ШІ) має потенціал стати ключовим інструментом у боротьбі з дезінформацією, особливо в контексті, де швидкість та точність є критично важливими. Розробка інструменту, який використовує ШІ для аналізу новинних джерел, може значно підвищити здатність суспільства виявляти та протидіяти дезінформації.

Література:

1. Діпфейки: як розпізнати та захиститися – [Електронний ресурс] – Режим доступу: https://netfreedom.org.ua/article/dipfejki-yak-rozpiznati-ta-zahistitisya
2. Протидія російській дезінформації через державно-приватне партнерство та міжнародну координацію – [Електронний ресурс] – Режим доступу: https://tsn.ua/ukrayina/protidiya-rosiyskiy-dezinformaciyi-cherez-derzhavno-privatne-partnerstvo-ta-mizhnarodnu-koordinaciyu-2438341.html
3. Дезінформація та штучний інтелект: (не)видима загроза сучасності – [Електронний ресурс] – Режим доступу: https://cedem.org.ua/analytics/dezinformatsiya-shtuchnyi-intelekt/
4. Підхід НАТО у галузі боротьби з дезінформацією – [Електронний ресурс] – Режим доступу: https://www.nato.int/cps/uk/natohq/topics_219728.htm
5. Як штучний інтелект бореться з дезінформацією і поширює її – [Електронний ресурс] – Режим доступу: https://www.stopfake.org/uk/pomizh-stsilloyu-ta-haribdoyu-yak-shtuchnij-intelekt-boretsya-z-dezinformatsiyeyu-i-poshiryuye-yiyi/

Додаток Б

Метод завантаження даних з Politifact

```
    def harvest_Politifact_data():
       print("Ready to harvest Politifact data.")
       input("[Enter to continue, Ctl+C to cancel]>>")
       print("Reading URLs file")
       # Read the data file into a pandas dataframe
       df_csv                                                =
pd.read_csv("newsbot/politifact_data.csv",
       error_bad_lines=False,              quotechar='"',
thousands=',',
       low_memory=False)
       for index, row in df_csv.iterrows():
          print("Attempting URL: " + row['news_url'])
          if(ss.loadAddress(row['news_url'])):
             print("Loaded OK")
       # some of this data loads 404 pages b/c it is a
little old,
       # some load login pages. I've found that
       # ignoring anything under 500 characters is a decent
       # strategy for weeding those out.
             if(len(ss.extractText)>500):
                ae = ArticleExample()
                ae.body_text = ss.extractText
                ae.origin_url = row['news_url']
                ae.origin_source = 'politifact data'
                ae.bias_score  =  0  #  Politifact  data
doesn't have this
                ae.bias_class = 5 # 5 is 'no data'
                ae.quality_score = row['score']
                ae.quality_class = row['class']
                ae.save()
                print("Saved, napping for 1…")
                time.sleep(1)
             else:
                print("**** This URL produced insufficient
data.")
          else:
             print("**** Error on that URL ^^^^^")
```

Додаток В

Налаштування моделей, аналіз прикладі та повернення вектора рядка

```
    from django.shortcuts import render
    import pandas as pd
    import numpy as np
    import pickle

    from .models import *
    from .forms import *
    from newsbot.strainer import *
    from newsbot.util import *

    def index(request):

        url = request.GET.get('u')
        if((url is not None) and (len(url) > 5)):
            print("Setting up")
            svc_model                                    =
pickle.load(open('newsbot/svc_model.sav', 'rb'))
            mlp_model                                    =
pickle.load(open('newsbot/MLPC_model.sav', 'rb'))
            log_model                                    =
pickle.load(open('newsbot/log_model.sav', 'rb'))
            cDict = loadCanonDict()
            ss = SoupStrainer()
            ss.init()
            print("Setup complete")
            print("Attempting URL: " + url)
            if(ss.loadAddress(url)):
                articleX                                 =
buildExampleRow(ss.extractText, cDict)
            else:
                print("Error on URL, exiting")
                return  render(request,  'urlFail.html',
{'URL', url})
            articleX = articleX.reshape(1, -1)
```

Додаток Д

Створення інтерфейсу таблиці з кожним результатом

```
     <h3 style="text-align: center;">
      Combined Result:<br>
      {% if fin_totFake >= fin_totReal %}
      Fake/Dodgy
      {% else %}
      Seems Legit/True
      {% endif %}
      </h3>

      <br>
     Probability of Fake: {{ fin_probabilities.0|floatformat }}%
chance of Fake
     <div class="progress">
      <div          class="progress-bar"          id="fakeProb_bar"
role="progressbar" aria-valuenow="{{ fin_probabilities.0 }}"
      aria-valuemin="0"     aria-valuemax="100"     style="width:{{
fin_probabilities.0 }}%"></div>
     </div>

      <br>
     Probability of Dodgy: {{ fin_probabilities.1|floatformat }}%
chance of Dodgy
     <div class="progress">
      <div          class="progress-bar"          id="MfakeProb_bar"
role="progressbar" aria-valuenow="{{ fin_probabilities.1 }}"
      aria-valuemin="0"     aria-valuemax="100"     style="width:{{
fin_probabilities.1 }}%"></div>
     </div>

      <br>
     Probability of Mostly True: {{ fin_probabilities.2|floatformat
}}% chance of Mostly True
     <div class="progress">
      <div          class="progress-bar"          id="MtrueProb_bar"
role="progressbar" aria-valuenow="{{ fin_probabilities.2 }}"
      aria-valuemin="0"     aria-valuemax="100"     style="width:{{
fin_probabilities.2 }}%"></div>
     </div>

      <br>
     Probability of True: {{ fin_probabilities.3|floatformat }}%
chance of True
     <div class="progress">
      <div          class="progress-bar"          id="trueProb_bar"
role="progressbar" aria-valuenow="{{ fin_probabilities.3 }}"
      aria-valuemin="0"     aria-valuemax="100"     style="width:{{
fin_probabilities.3 }}%"></div>
     </div>
```