Issue 18, part 1, 2025

https://doi.org/10.23939/sisn2025.18.098

УДК 004.8-9

МЕТОД ВИЯВЛЕННЯ ДЕЗІНФОРМАЦІЇ НА ОСНОВІ АНАЛІЗУ ТЕКСТОВИХ ДАНИХ ІЗ ЗАСТОСУВАННЯМ TF-IDF ТА КОНТЕКСТНИХ ВЕКТОРНИХ ПРЕДСТАВЛЕНЬ

Ольга Лозинська, Вікторія Висоцька, Оксана Марків, Мар'ян Куспісь

Національний університет "Львівська політехніка",
кафедра Інформаційних систем та мереж, Львів, Україна
E-mail: olha.v.lozynska@lpnu.ua, ORCID: 0000-0002-5079-0544
E-mail: victoria.a.vysotska@lpnu.ua, ORCID: 0000-0001-6417-3689
E-mail: oksana.o.markiv@lpnu.ua, ORCID: 0000-0002-1691-1357
E-mail: marian.kuspis.mnitm.2023@lpnu.ua, ORCID: 0009-0001-0678-784X

У статті розглянуто підхід до виявлення джерел дезінформації у цифровому середовищі за допомогою аналізу текстів із використанням методів машинного навчання та опрацювання природної мови. Запропонований метод базується на гібридному представленні тексту, яке поєднує частотні ознаки (TF-IDF) з контекстними векторними представленнями, отриманими за допомогою моделі IBM Granite. Розроблено повний цикл опрацювання даних, що охоплює етапи дослідницького аналізу (EDA), попереднього опрацювання та токенізації текстів, формування векторних представлень, навчання моделі логістичної регресії та отримання ключових метрик. Основні етапи попереднього опрацювання тексту включали приведення всіх символів до нижнього регістру, видалення URL-адрес і HTML-тегів, очищення від небуквених символів і зайвих пробілів, усунення дублікатів для уникнення повторного навчання, уніфікацію значень певних полів. Для векторизації очищених текстів застосовано поєднання TF-IDF з контекстними векторними представленнями, що дало змогу моделі одночасно враховувати статистичну значущість термінів та їхній семантичний контекст у межах повідомлень. Побудована модель логістичної регресії у поєднанні з гібридним представленням текстових даних продемонструвала високу ефективність, досягнувши загальної точності 82 % та збалансованими значеннями F1-міри для класів "правда" і "фейк". Для ідентифікації найбільш релевантних термінів застосовано аналіз ваг TF-IDF-ознак на основі коефіцієнтів логістичної регресії. Проведений аналіз показав, що модель схильна асоціювати правдиву інформацію з україномовною, нейтральною лексикою, тоді як тексти з ознаками дезінформації часто містять російськомовні елементи, характерні для пропагандистських або маніпулятивних повідомлень. Подальші дослідження будуть спрямовані на розширення набору даних та створення нових ансамблевих моделей для виявлення джерел дезінформації.

Ключові слова – дезінформація, фейк, контекстні векторні представлення, TF-IDF, набір даних, точність, метрика.

Постановка проблеми

У добу цифрових технологій дезінформація є свідомим поширенням неправдивих чи викривлених даних з метою маніпулювання громадською думкою і становить серйозну загрозу для політичної, соціальної та економічної стабільності (Wardle, Derakhshan, 2017). Зважаючи на те, що основна частина інформації поширюється через соціальні мережі, блоги та новинні ресурси, виникає нагальна потреба у створенні автоматизованих інструментів, здатних швидко та ефективно виявляти джерела дезінформації у великомасштабних інформаційних потоках.

Соцмережі є основними каналами поширення інформації, однак водночас і платформами для масового розповсюдження дезінформації та пропаганди. Тому створення ефективних інструментів для моніторингу та протидії таким загрозам має вирішальне значення для захисту інформаційного простору.

Особливу складність становить масштаб та динаміка інформаційних потоків. Дезінформаційні кампанії часто проводяться скоординовано й організовано, з використанням бот-мереж, які маскуються під звичайних користувачів. Для ідентифікації таких акаунтів, їхніх зв'язків та джерел поширення інформації необхідні сучасні автоматизовані інструменти та методи аналізу великих даних.

Технології штучного інтелекту та машинного навчання значно підвищують ефективність процесів виявлення фейкових новин і активності чат-ботів. Їх можна застосовувати для аналізу текстів на предмет ознак фальсифікації, розпізнавання підозрілої поведінки в соцмережах, виявлення фальсифікованих зображень і відео.

Хоча проблема дезінформації є глобальною та стосується багатьох країн, її вирішення на національному рівні може зробити вагомий внесок у загальносвітову боротьбу з інформаційними загрозами.

Більшість сучасних інструментів для протидії дезінформації здебільшого зосереджені на англомовному контенті, тоді як україномовний сегмент інформаційного простору має низку унікальних рис – мовних, культурних та історичних. Створення спеціалізованих засобів, адаптованих саме до україномовного контексту, дасть змогу подолати цю проблему та забезпечити відповідність актуальним вимогам і викликам. Крім створення засобів, велике значення та актуальність має формування україномовних наборів даних, які є цінними для міжнародної наукової спільноти та сприятимуть подальшим дослідженням у цій сфері.

Традиційні підходи до виявлення фейків або ботів дедалі більше втрачають ефективність у зв'язку з постійним удосконаленням засобів дезінформації. Тому виникає потреба у створенні багатофункціональних засобів, здатних в автоматичному режимі опрацьовувати великі обсяги даних, виявляти фейковий контент, аналізувати поведінку користувачів та ідентифікувати скоординовані мережі ботів.

Авторами запропоновано підхід, який включає збір даних із різних цифрових джерел, попереднє опрацювання та очищення контенту, виявлення дезінформаційних повідомлень, аналіз поведінки користувачів для виявлення ботів, формування аналітичних звітів із відповідними рекомендаціями.

Аналіз останніх досліджень та публікацій

Сьогоднішнє цифрове середовище характеризується активним поширенням дезінформації та зростанням використання чат-ботів, що становить серйозний виклик для інформаційної безпеки. Саме тому методи машинного навчання та опрацювання природної мови є важливою складовою в ідентифікації джерел фейкових повідомлень і виявленні неавтентичної поведінки чат-ботів. Застосування цих методів дає змогу виявляти не лише джерела, а й шляхи поширення дезінформаційних наративів. Завдяки алгоритмам машинного навчання можливо автоматизувати процес ідентифікації фейкових повідомлень, зокрема через розпізнавання повторюваних мовних конструкцій та шаблонів, характерних для дезінформації.

Виділяють два ключові підходи до виявлення джерел дезінформації (Tajrian, et al., 2023):

- ручний – передбачає перевірку фактів експертами шляхом ручного аналізу;
- автоматизований – включає використання алгоритмів машинного навчання для автоматичного аналізу великих інформаційних потоків.

Чат-боти, що дедалі частіше використовуються у цифрових комунікаціях, можуть слугувати засобом поширення фейкових повідомлень. Визначення їхньої поведінки, зокрема неавтентичної, є важливим аспектом протидії дезінформації.

Алгоритми машинного навчання демонструють високу ефективність у виявленні дезінформації та її джерел поширення. Вони здатні опрацьовувати дані в реальному масштабі часу, виявляючи нові способи поширення фейків та адаптуючись до змін. Такий підхід дає змогу оперативно виявляти неправдиву інформацію та реагувати на неї.

Системи машинного навчання та опрацювання природної мови (NLP) дають змогу аналізувати великі масиви текстових даних, виявляти закономірності, автоматично інтерпретувати зміст та генерувати мовний контент. Їх застосування охоплює:

- машинний переклад,
- створення чат-ботів,
- аналіз тональності тексту,
- автоматичне резюмування,
- виявлення дезінформації.

Класифікація тексту, тобто визначення належності до певної категорії (фейк/правда, позитивний/негативний тощо), здійснюється з використанням таких методів машинного навчання, як логістична регресія, метод дерев рішень, наївний байєсівський класифікатор (Mohammed, et al., 2024) метод опорних векторів (Sudhakar, Kaliyamurthie, 2024) та методів глибокого навчання (Kozik, et al., 2022).

У дослідженні (Battal, et al., 2024) розроблено моделі класифікації для виявлення фейкових новин за допомогою алгоритмів машинного навчання. Моделі протестовано на наборі даних, який включає приклади фейкових та реальних новин і містить 42 000 прикладів. Для підвищення якості, точності та зручності використання до набору даних були застосовані методи попереднього опрацювання тексту. Моделі були створені для виявлення фейкових новин за допомогою алгоритмів класифікації за сингулярними та ансамблевими ознаками. Найвищі результати продуктивності спостерігалися в алгоритмі класифікації випадкового лісу з коефіцієнтом точності 76 %.

Авторами (Abdulaziz, Marwan, 2020) досліджено використання чотирьох відомих алгоритмів машинного навчання, а саме випадкового лісу, наївного баєсівського класифікатора, нейронної мережі та дерев рішень, щоб окремо перевірити ефективність класифікації у виявленні фейкових новин. Експерименти проводились на публічному наборі даних LIAR (Wang, 2017). Результати експериментів показали, що наївний баєсівський класифікатор значно перевершує інші алгоритми на цьому наборі даних.

У дослідженні (Alikhashashneh, et al., 2024) запропоновано метод, який поєднує алгоритми машинного навчання і глибокого навчання для виявлення фейкової інформації порівняно з правдивою інформацією. Авторами наведено результати застосування алгоритмів та порівняння їх продуктивності. Результати цього дослідження показують, що алгоритми, які використовують обернену частоту документів за частотою термінів TF-IDF, досягли кращих результатів, ніж алгоритми, які використовують Word2Vec. Однак алгоритм довгої короткочасної пам'яті LSTM досягає найкращої продуктивності; точність 99 % – при використанні TF-IDF та 94 % – при використанні Word2Vec.

Актуальні дослідження в галузі опрацювання природної мови (NLP) та машинного навчання пропонують ефективні підходи до автоматизованого виявлення фейкової інформації. У роботі розглядаються техніки аналізу текстового контенту, взаємодій у соціальних мережах і поведінкових характеристик користувачів з метою виявлення дезінформації (Shu, et al., 2020). Автори підкреслюють ключову роль наявності анотованих наборів даних, необхідних для ефективного навчання моделей. У науковій роботі (Su, et al., 2020) запропоновано підхід для автоматизованого виявлення дезінформації з використанням NLP. Цей підхід включає комплексний огляд аналізу, виявлення та запобігання дезінформації, представляючи ключові методи, функції, моделі та набори даних.

Отже, застосування NLP у боротьбі з дезінформацією включає:

- автоматичне розпізнавання фейкових повідомлень,
- оцінку надійності джерел,
- перевірку фактів через порівняння з наборами достовірної інформації.

Для того, щоб моделі машинного навчання мали змогу опрацьовувати тексти, їх потрібно представити у вигляді чисел (векторів). Серед етапів підготовки тексту до опрацювання можна виділити також токенізацію, лематизацію, видалення стоп-слів, векторизацію. Найбільш поширені алгоритми векторизації – алгоритм «мішок слів» (BoW) (Huang, 2020), частотних ознак TF-IDF (Jouhar, et al., 2024), алгоритм векторного представлення слів (Word Embeddings) (Alikhashashneh, et al., 2024), трансформерів BERT (Kozik, et al., 2022) та RoBERTa (Shu, 2024).

Машинне навчання й технології опрацювання природної мови виступають ефективними засобами для автоматизованого аналізу текстової інформації, виявлення дезінформації та ідентифікації ботів. Завдяки сучасним підходам, зокрема методам машинного навчання, значно підвищується якість опрацювання мовного контенту, що відкриває нові перспективи у сфері кібербезпеки та аналітики.

Формулювання цілі статті

Метою роботи є розроблення методу виявлення дезінформації на основі аналізу набору даних новин та постів, зібраних із соцмереж. Для цього використано метод логістичної регресії із застосуванням частотних ознак TF-IDF та контекстних векторних представлень.

Виклад основного матеріалу

Перш ніж розпочати моделювання, було виконано попередній аналіз структури вхідного набору даних (Lozynska, et al., 2024), який був доповнений оновленою інформацією, за допомогою методу дослідницького аналізу даних EDA (Komorowski, et al., 2016). Основною метою цього етапу було виявлення ключових закономірностей, можливих аномалій, відсутніх значень, дисбалансу між класами, а також інших особливостей даних, які можуть вплинути на подальший вибір підходів до опрацювання й моделювання.

Під час аналізу було проведено:

- візуалізацію розподілу мовної змінної (“lang”) та цільової змінної “label”, яка позначає приналежність запису до категорії “правда” або “фейк”;
- оцінку типів даних (текстові, категоріальні, числові, дата);
- перевірку на наявність пропущених або дублікатів записів;
- базовий аналіз текстів за довжиною та частотою слів, а також попереднє текстове опрацювання.

Отримані результати дали змогу прийняти обґрунтовані рішення щодо:

- вибору релевантних змінних (наприклад: “text”, “label”, “source”, “lang”, “link”);
- видалення дублікатів текстів та нерелевантних записів;
- побудови гістограм для таких показників, як довжина тексту, кількість слів (на основі поділу за пробілами) та число токенів, сформованих токенізатором моделі.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 965 entries, 0 to 964
Data columns (total 23 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Unnamed: 0    185 non-null    object
 1   №             150 non-null    object
 2   дата          738 non-null    object
 3    час          691 non-null    object
 4   text          960 non-null    object
 5   label         958 non-null    object
 6   Post/Repost   963 non-null    object
 7   Автор/Group   942 non-null    object
 8   source        929 non-null    object
 9   lang          960 non-null    object
 10  Like          258 non-null    object
 11  Поширення     84 non-null     object
 12  link          954 non-null    object
 13  Unnamed: 13   0 non-null      float64
 14  Unnamed: 14   2 non-null      object
 15  Unnamed: 15   0 non-null      float64
 16  Unnamed: 16   0 non-null      float64
 17  Unnamed: 17   0 non-null      float64
 18  Unnamed: 18   0 non-null      float64
 19  Unnamed: 19   0 non-null      float64
 20  Unnamed: 20   0 non-null      float64
 21  Unnamed: 21   0 non-null      float64
 22  Unnamed: 22   1 non-null      float64
dtypes: float64(9), object(14)
memory usage: 173.5+ KB
```

Рис. 1. Загальна інформація про набір даних

На першому етапі було виведено загальну структуру набору даних (рис. 1), а також приклади перших чотирьох рядків (рис. 2). Аналіз показав, що дві перші колонки містили неінформативні значення, але загальна кількість змістовних записів дала змогу продовжити опрацювання. Частина

записів містила повідомлення про недоступність контенту, а деякі – системні повідомлення. Було зафіксовано порожні значення у колонці “текстове повідомлення” (яка була перейменована у “text”), та на основі додаткових джерел вдалося частково відновити зміст для окремих записів.

На наступному етапі здійснено очищення набору даних. Збережено лише ті змінні, що мають практичну цінність для задачі класифікації, зокрема “text”, “label”, “source”, “lang” та “link” (рис. 3). Також було видалено записи з порожнім текстом.

	0	1	2	3
Unnamed: 0	NaN	NaN	NaN	NaN
№	NaN	NaN	NaN	NaN
дата	20.08.2024	20.09.2024	20.08.2024	20.08.2024
час	10:06	9:49	23:04	NaN
текст повідомлень	Оборона ВСУ в Донецкой области осталась без сн...	Оборона бандеровцев в Донецкой области осталас...	Після початку військової операції у Курській о...	Оборона ЗСУ на Донбасі залишилася без снарядів...
мітка	ФЕЙК	ФЕЙК	ФЕЙК	ФЕЙК
Post/Repost	Post	Post	Post	Post
Автор/Group	Einarr Modig	Vlad Brigg	Тетяна Бесараб	Медіавектор
Джерело	Facebook	Facebook	Інформатор України	Медіавектор
Мова	Russian	Russian	Ukrainian	Ukrainian
Like	5	13	24	NaN
Поширення	3	NaN	NaN	NaN
веб-адреса	https://www.facebook.com/permalink.php?story_f...	https://www.facebook.com/permalink.php?story_f...	https://informator.ua/uk/pislya-nastupu-na-kur...	https://mediavektor.org/92604-oborona-vsu-v-do...
Unnamed: 13	NaN	NaN	NaN	NaN
Unnamed: 14	NaN	NaN	NaN	NaN
Unnamed: 15	NaN	NaN	NaN	NaN
Unnamed: 16	NaN	NaN	NaN	NaN

Рис. 2. Загальна структура набору даних

```
df.rename(columns={' мітка': 'label',
                   'текст повідомлень': 'text',
                   'Джерело': 'source',
                   'Мова': 'lang',
                   'веб-адреса': 'link'}, inplace=True)

df = df[['text', 'label', 'lang', 'source', 'link']]
```

Рис. 3. Перейменування та вибір необхідних стовпців

Оскільки деякі дублікати в наборі даних відрізнялися лише незначними деталями (регістром, пробілами), очищення дублікатів відбувається після завершення попереднього опрацювання тексту.

Для забезпечення стабільної та ефективної роботи моделі логістичної регресії реалізовано опрацювання текстової змінної “text” та цільової змінної “label”. Основна мета цього етапу – зменшити варіативність у даних, усунути зайві елементи та нормалізувати значення класів.

Основні етапи попереднього опрацювання тексту:

1. Приведення до нижнього регістру полягає у переведенні усіх символів в полях “text” та “label” у нижній регістр для уникнення дублювання однакових слів з різним написанням (наприклад, «Правда», «ПРАВДА» і «правда»).

2. Очищення тексту передбачає видалення всіх URL-адреси та HTML-теги, що не несуть інформативного навантаження.

3. Вилучення зайвих символів полягає у прибранні з тексту усіх небуквених та нецифрових символів, включно з пунктуацією, емодзі та спецсимволами. Також усунено зайві пробіли.

4. Усунення дублікатів полягає у збереженні лише унікальних записів, що дає змогу уникнути повторного навчання моделі на ідентичних прикладах.

5. Нормалізація міток забезпечує уніфікацію значення змінної “label” до двох валідних варіантів “фейк” та “правда”. Це забезпечує чітке розмежування класів у процесі моделювання.

Результати попереднього опрацювання тексту засвідчили, що всі дублікатні значення успішно усунені, а “label” тепер містить лише два допустимі значення (рис. 4). Оновлена інформація про набір даних після завершення очищення наведена на рис. 5. Після цих кроків дані готові до наступних етапів – токенізації, поділу на тренувальну та тестову вибірки, векторизації та побудови моделі для класифікації повідомлень за достовірністю.

Методи векторизації та класифікаційна модель. У рамках дослідження застосовано комбінований підхід до представлення тексту, що поєднує частотні характеристики (TF-IDF) та контекстні векторні представлення (Granite Embeddings). Така гібридна стратегія дозволяє моделі враховувати не лише кількість і унікальність слів, а й їхній змістовий контекст у межах тексту.

TF-IDF (Term Frequency – Inverse Document Frequency) є методом, що відображає наскільки часто деякі слова трапляються в тексті, а також наскільки ці слова є унікальними для конкретного документа порівняно з усім корпусом. Цей метод не враховує порядок слів чи контекст, але забезпечує стабільне числове представлення, особливо ефективне для коротких текстів.

```
df['label'].unique()
```

```
array(['фейк', 'правда'], dtype=object)
```

Рис. 4. Унікальні значення стовпця "label" після опрацювання

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 483 entries, 0 to 962
Data columns (total 5 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   text    483 non-null    object
 1   label   483 non-null    object
 2   lang    483 non-null    object
 3   source  460 non-null    object
 4   link    473 non-null    object
dtypes: object(5)
memory usage: 22.6+ KB
```

Рис. 5. Оновлена інформація про набір даних

Комбінування частотних і контекстних ознак для класифікації тексту. Для представлення текстових даних було застосовано два підходи з використанням TF-IDF та контекстних векторних представлень, що забезпечило гібридну модель векторизації.

Для TF-IDF обрано такі параметри:

max_features = 4000 – обмеження кількості ознак до 4000 найбільш значущих термінів, що виправдано розміром словника (8816 слів);

tokenizer = lambda x: x.split(' ') – простий поділ тексту на слова за пробілами.

Контекстні векторні представлення. Доповнення TF-IDF реалізовано за допомогою багатомовної моделі ibm-granite/granite-embedding-107m-multilingual, яка належить до трансформерного класу моделей Granite, створених IBM. Ця модель може опрацьовувати до 512 токенів та формувати 384-вимірні векторні подання тексту. Її перевага полягає в здатності враховувати контекст та семантичні взаємозв'язки між словами, незалежно від їхнього порядку, що особливо важливо для виявлення змістовної дезінформації.

Для об'єднання переваг обох підходів створено гібридне представлення шляхом об’єднання TF-IDF та Granite-векторів у розширений вектор ознак.

Логістична регресія як модель класифікації. Як модель класифікації використано модель логістичної регресії, яка є надійним методом для задач бінарної класифікації, особливо в умовах великої кількості вхідних ознак. Вона демонструє хорошу продуктивність, швидкість навчання та простоту інтерпретації.

Використані гіперпараметри для логістичної регресії:

- penalty = 'l2' – регуляризація для уникнення перенавчання;
- solver = 'lbfgs' – оптимізатор, що добре масштабується на задачах з великою кількістю параметрів;
- class_weight = 'balanced' – автоматичне коригування ваг для класів із різною представленістю;
- random_state = 42 – фіксація випадковості для забезпечення відтворюваності результатів.

Завдяки поєднанню частотних та семантичних ознак модель отримала змогу глибше аналізувати текстовий зміст новин, виявляючи ознаки дезінформації. Усі етапи від опрацювання до навчання об'єднано в єдину структуру, що дає змогу застосовувати модель у різних інформаційних середовищах.

Для практичного використання побудовано клас FakeNewsClassifier, який об’єднує всі ключові етапи опрацювання текстових даних. Клас включає функціонал для попереднього опрацювання, векторизації, побудови гібридних ознак, навчання моделі та візуалізації результатів. Це забезпечує цілісність і стабільність процесу класифікації та дає змогу легко повторювати експерименти на нових наборах даних.

Візуалізація балансу класів і текстових характеристик, поділ вибірки та векторизація. Для аналізу розподілу цільової змінної використано метод target_balance(), який будує графік частот класів “фейк” і “правда” (рис. 6). Візуалізація свідчить про відсутність суттєвого дисбалансу між класами і модель має змогу навчатися без схильності до переваги одного з класів, забезпечуючи збалансовану продуктивність.

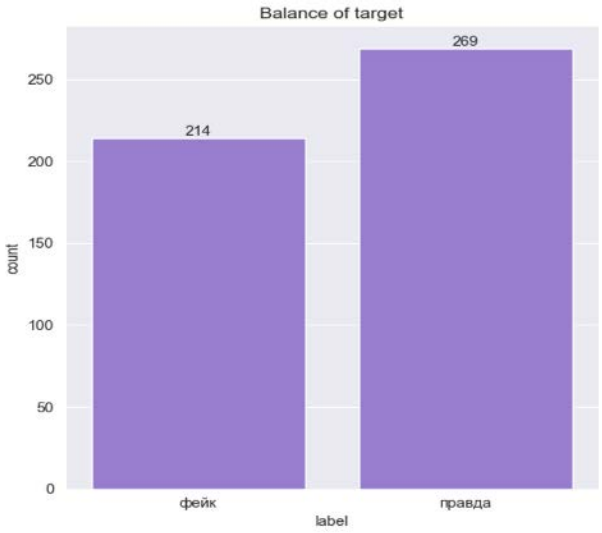

Рис. 6. Баланс цільової змінної

Для вивчення структурних властивостей текстових даних застосовано метод visualize_text_features(), який створює гістограми довжин текстів у символах і кількості слів, з подальшим збереженням зображення (рис. 7). Результати демонструють, що більшість текстів мають від 50 до 500 символів, а кількість слів переважно варіюється від 24 до 41. Спостерігається правостороннє зміщення, тобто значна частина повідомлень є короткими, хоча трапляються й винятки з надмірною довжиною. Ця інформація є важливою для підвищення якості тренування та діагностики моделі.

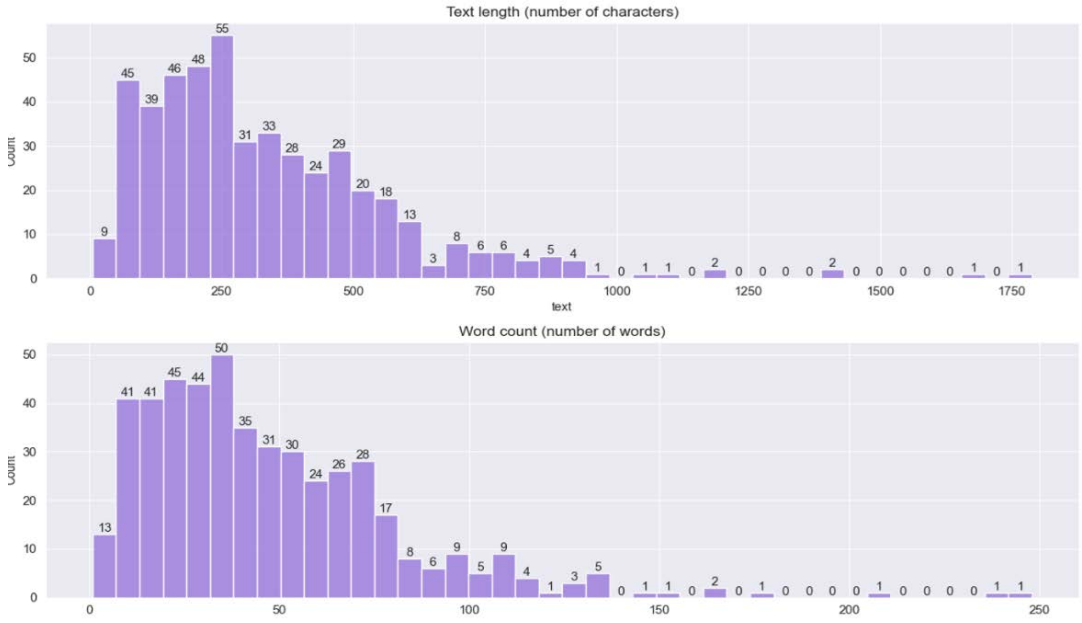

Рис. 7. Візуалізація структурних характеристик тексту

Навчальна та тестова вибірки становлять відповідно 80 % та 20 % усіх записів. У результаті навчальна частина складає 386 записів, а тестова – 97.

Для отримання векторних подань текстів використано гібридний підхід, що поєднує частотні TF-IDF-представлення з контекстними векторними представленнями, створеними за допомогою моделі Granite. У результаті формується матриця ознак розміром (386, 4384), яка об'єднує два типи векторизації в один простір ознак.

Щоб оцінити, як відбувається розділення класів у цьому просторі, застосовано метод, який виконує зменшення розмірності контекстних векторних представлень до двох компонент із використанням t-SNE та будує графічне представлення кластерів для ознак TF-IDF і Granite. Така візуалізація (рис. 8) дає змогу перевірити, наскільки чітко модель здатна розділяти класи “фейк” і “правда” у просторі ознак.

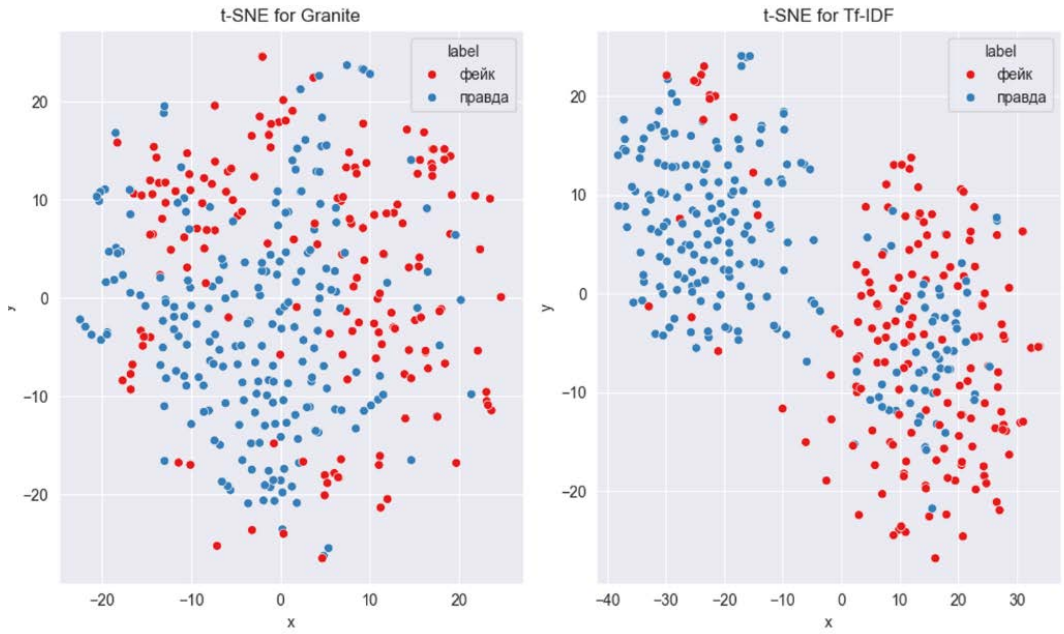

Рис. 8. Візуалізація результатів для обох типів ознак

На рис. 8 видно, що векторні представлення, побудовані на основі TF-IDF, демонструють чітке розмежування між класами "фейк" і "правда". Натомість у випадку контекстних векторних представлень спостерігається часткове перекриття, що може свідчити про більш розмиту межу між класами в контекстному представленні.

Для аналізу особливостей токенізації при використанні моделі Granite побудовано гістограму, що відображає кількість токенів у кожному тексті *для* навчальної та тестової вибірок. Це дає змогу оцінити, як модель розбиває тексти на токени перед побудовою контекстних векторних представлень (рис. 9).

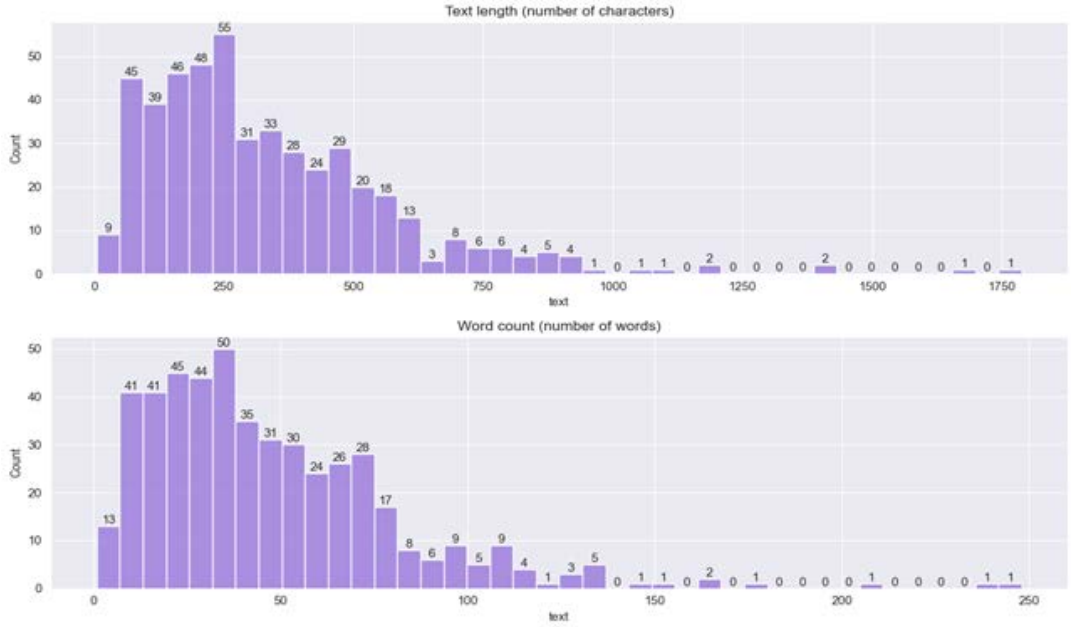

Рис. 9. Візуалізація Granite-токенізації

З рис. 9 видно, що більшість текстів у навчальній вибірці мають довжину від 18 до 42 токенів, з піком у діапазоні 37–42 токенів. Тестова вибірка демонструє більш розсіяний розподіл з менш вираженим піком. Це свідчить про відносно стабільну довжину текстів після токенізації та коректну роботу попереднього опрацювання даних перед їх використанням у моделі Granite.

Навчання моделі класифікації здійснено за допомогою методу логістичної регресії на основі попередньо побудованих гібридних ознак.

У задачах класифікації ефективність моделі доцільно оцінювати за здатністю коректно визначати кожен із класів окремо. Для цього обрано метрики повнота, влучність, точність та F1-міру. Повнота відображає, яку частку об’єктів позитивного класу, що у нашому випадку є класом “фейк”, модель змогла правильно виявити. Натомість влучність відображає, яку частку з усіх передбачених як фейки документів насправді становлять фейкові новини.

За результатами класифікації (табл.) загальна точність моделі (точність) становить 0.82.

Показники метрик для класифікаційної моделі

Метрика	Влучність	Повнота	F1-score
правда	0,84	0,85	0,84
фейк	0,81	0,79	0,80
точність			0,82
Macro avg	0,82	0,82	0,82
Weighted avg	0,82	0,82	0,82

Модель демонструє F1-міру 0.84 для класу "правда" та 0.80 для класу "фейк", що вказує на відносно збалансовану продуктивність за двома класами. Метрики macro avg та weighted avg також перебувають на рівні 0.82, що свідчить про стійку якість передбачень незалежно від розподілу класів.

Для оцінки якості класифікації побудовано три матриці невідповідностей - без нормалізації, нормалізовану за влучністю та нормалізовану за повнотою (рис. 10). Матриця невідповідностей – це таблиця, яка показує, наскільки добре класифікаційна модель справляється з передбаченням (Sathyanarayanan, Roopashri Tantri, 2024). Такий підхід дає змогу всебічно оцінити, як модель справляється з передбаченням кожного класу та наскільки вона схильна до помилок у тій чи іншій формі. Окрім візуалізації, метод додатково обчислює значення Fβ-міри з параметром β = 1.3 за замовчуванням, що зміщує акцент на повноту, відповідно до обраного у дослідженні фокусу на виявлення фейків.

З матриці без нормалізації (рис. 10, графік ліворуч) видно, що модель правильно класифікувала 46 прикладів класу "правда" та 34 приклади класу "фейк", допустивши 8 і 9 помилок відповідно. У нормалізованому вигляді видно, що влучність для класу "фейк" становить 81 %, а повнота – 79 %, що є прийнятним результатом у задачі, де важливо мінімізувати пропущені фейки. Значення Fβ-міри з акцентом на повноту складає приблизно 0.79, що свідчить про хорошу здатність моделі виявляти фейкові новини.

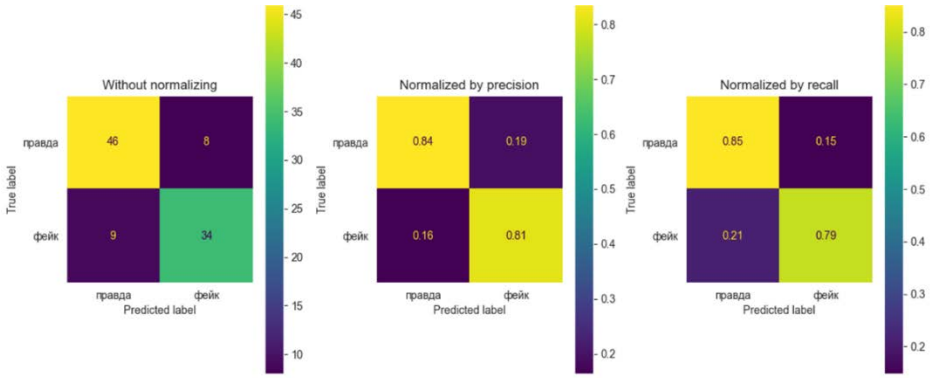

Рис. 10. Матриці невідповідностей

Щоб визначити, які терміни (слова) найбільше впливають на рішення класифікаційної моделі, застосовано метод, який дає змогу проаналізувати вагу TF-IDF-ознак на основі коефіцієнтів логістичної регресії, виявляючи, які слова найсильніше асоціюються з тим чи іншим класом. Для підвищення інформативності візуалізації, відфільтровано 20 найочевидніших термінів (переважно загальні або часто вживані слова), після чого виведено наступні десять за значущістю термінів (рис. 11).

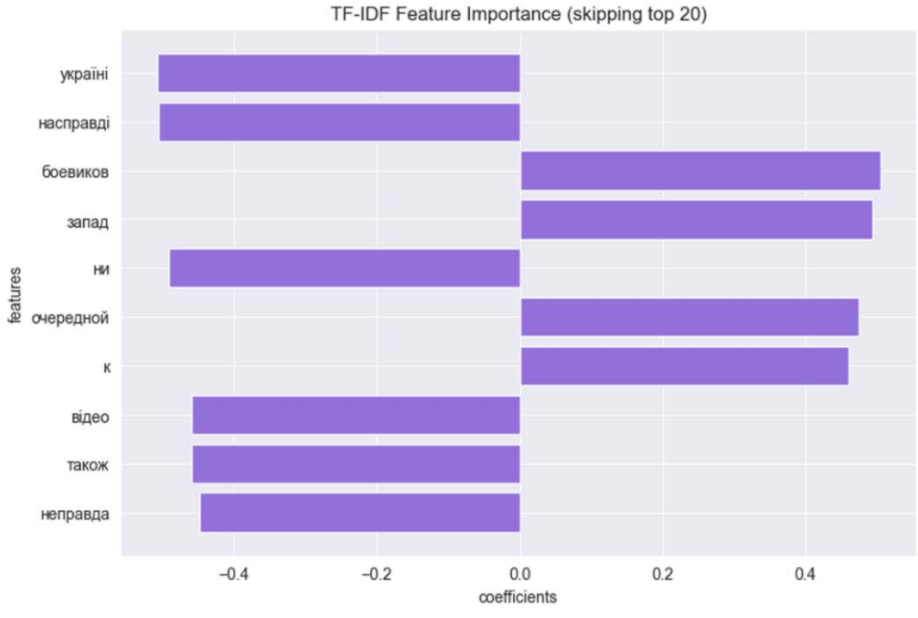

Рис. 11. Найважливіші слова на основі TF-IDF

З цього можна зробити висновок, що позитивні значення (справа) вказують на слова, пов’язані з класом "фейк", наприклад: "боевиков", "запад", "очередной". У той час як негативні коефіцієнти (зліва) відображають терміни, які більше притаманні класу "правда", серед яких: “Україні”, “насправді”, “відео”, “також”, “неправда”.

Отже, аналіз показав, що модель пов’язує достовірну інформацію переважно з україномовною та нейтральною лексикою, тоді як дезінформаційні тексти часто містять російськомовні слова, що мають пропагандистський чи маніпулятивний характер.

Подяка

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

Висновки

У межах дослідження реалізовано підхід до автоматизованого виявлення дезінформації на основі гібридного представлення текстових даних. Запропонована система поєднує застосування частотних ознак (TF-IDF) з контекстними векторними представленнями (IBM Granite), що дало змогу досягти високого рівня точності класифікації завдяки урахуванню як частотних, так і семантичних характеристик текстів.

Проведено попередній аналіз структури набору даних із застосуванням методу дослідницького аналізу даних EDA, що дало змогу виявити нерелевантні записи, пропущені значення та визначити особливості текстових даних. Наведено основні етапи попереднього опрацювання тексту, що включає нормалізацію тексту, очищення від шуму, уніфікацію міток і видалення дублікатів, що забезпечило якісну підготовку даних до навчання моделі.

Для задачі бінарної класифікації застосовано метод логістичної регресії, який продемонстрував стабільні результати за основними метриками: точність – 0.82, F1-міра для класів “фейк” та “правда” – 0.80 і 0.84 відповідно. Основну увагу зосереджено на максимізації метрики повнота, що є критично важливим у контексті виявлення дезінформації.

Запропонований підхід показав свою ефективність та гнучкість, а також забезпечив високий рівень інтерпретованості результатів, що є важливим фактором для подальшого використання моделі в реальних системах медіааналізу та інформаційній безпеці. Подальші дослідження будуть спрямовані на розширення набору даних та створення нових ансамблевих моделей для виявлення джерел дезінформації.

СПИСОК ЛІТЕРАТУРИ

Abdulaziz, A., & Marwan, A. (2020). An Empirical Comparison of Fake News Detection using different Machine Learning Algorithms. International Journal of Advanced Computer Science and Applications (IJACSA), 11(9), 146-152.

Alikhashashneh, E., Nahar, K., Abual-Rub, M., & Alkhaldy, H. (2024). A robust method for detecting fake news using both machine and deep learning algorithms. Indonesian Journal of Electrical Engineering and Computer Science, 36, 1816-1826. https://doi.org/10.11591/ijeecs.v36.i3.pp1816-1826.

Battal, B., Yıldırım, B., Dinçaslan, Ö., & Cicek, G. (2024). Fake News Detection with Machine Learning Algorithms. Celal Bayar University Journal of Science, 20(3), 65-83. https://doi.org/10.18466/cbayarfbe.1472576.

Huang, J. (2020). Detecting Fake News With Machine Learning. Journal of Physics: Conference Series, 1693. https://doi.org/10.1088/1742-6596/1693/1/012158.

Jouhar, J., Pratap, A., Tijo, N., & Mony, M. (2024). Fake News Detection using Python and Machine Learning. Procedia Computer Science, 233, 763-771. https://doi.org/10.1016/j.procs.2024.03.265.

Komorowski, M., Marshall, D., Salciccioli, J., & Crutain, Y. (2016). Exploratory Data Analysis. Secondary Analysis of Electronic Health Records, 185-203. https://doi.org/10.1007/978-3-319-43742-2_15.

Kozik, R., Kula, S., Choraś, M. & Wozniak, M. (2022). Technical solution to counter potential crime: Text analysis to detect fake news and disinformation. Journal of Computational Science, 60(101576). https://doi.org/10.1016/j.jocs.2022.101576.

Lozynska, O., Markiv, O., Vysotska, V., Romanchuk, R., & Nazarkevych, M. (2024). Information technology for developing and populating a disinformation dataset using intelligent deepfakes and clickbait search. Herald of Khmelnytskyi National University. Technical Sciences, 343(6(1)), 158-167. https://doi.org/10.31891/2307-5732-2024-343-6-24.

Mohammed, S., Al-Aaraji, N., & Al-Saleh, A. (2024). Knowledge Rules-based Decision Tree Classifier model for effective fake accounts detection in social networks. International Journal of Safety and Security Engineering, 14(4), 1243-1251. https://doi.org/10.18280/ijsse.140421.

Tajrian, M., Rahman, A., Kabir, M.A. & Islam, M. R. (2023). A Review of Methodologies for Fake News Analysis. IEEE Access, 11, 73879-73893, https://doi.org/10.1109/ACCESS.2023.3294989.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22–36. https://doi.org/10.1145/3137597.3137600.

Shu, X. (2024). BERT and RoBERTa for Sarcasm Detection: Optimizing Performance through Advanced Fine-tuning. Applied and Computational Engineering, 97, 1-11. https://doi.org/10.54254/2755-2721/97/20241354.

Su, Q., Wan, M., Liu, X., & Huang, C.-R. (2020). Motivations, methods and metrics of misinformation detection: An NLP perspective. Natural Language Processing Research, 1, 1–13. https://doi.org/10.2991/nlpr.d.200522.001.

Sudhakar, M., & Kaliyamurthie, K.P. (2024). Detection of fake news from social media using support vector machine learning algorithms. Measurement: Sensors, 32(101028). https://doi.org/10.1016/j.measen.2024.101028

Wang, W. Y. (2017). “Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (pp. 422–426).

Wardle, C., & Derakhshan, H. (2017). Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making. Council of Europe report DGI09.

Sathyanarayanan, S., & Roopashri Tantri, B. (2024). Confusion Matrix-Based Performance Evaluation Metrics: African Journal of Biomedical Research, 27, 4023-4031. https://doi.org/10.53555/AJBR.v27i4S.4345.

METHOD FOR DETECTION OF DISINFORMATION BASED ON TEXT DATA ANALYSIS USING TF-IDF AND CONTEXTUAL VECTOR REPRESENTATIONS

Olga Lozynska, Victoria Vysotska, Oksana Markiv, Marian Kuspis

Lviv Polytechnic National University, Lviv, Ukraine
E-mail: olha.v.lozynska@lpnu.ua, ORCID: 0000-0002-5079-0544
E-mail: victoria.a.vysotska@lpnu.ua, ORCID: 0000-0001-6417-3689
E-mail: oksana.o.markiv@lpnu.ua, ORCID: 0000-0002-1691-1357
E-mail: marian.kuspis.mnitm.2023@lpnu.ua, ORCID: 0009-0001-0678-784X

The article considers an approach to detecting fake news in the digital environment through text analysis using machine learning and natural language processing methods. The proposed method is based on a hybrid text representation combining frequency features (TF-IDF) and contextual embeddings obtained using the IBM Granite model. A complete data processing cycle was developed, covering the stages of exploratory analysis (EDA), text preprocessing and tokenization, forming vector representations, training a logistic regression model, and obtaining key metrics. The main stages of text

preprocessing included converting all characters to lowercase, removing URLs and HTML tags, cleaning from non-letter characters and excess spaces, eliminating duplicates to avoid re-training, and unifying the values of specific fields. A combination of TF-IDF with contextual embeddings was used to vectorize the cleaned texts, which allowed the model to simultaneously consider the statistical significance of terms and their semantic context within the messages. The constructed logistic regression model combined with a hybrid representation of text data demonstrated high efficiency, achieving an overall accuracy of 82 % and balanced F1-measure values for the “true” and “fake” classes. An analysis of TF-IDF feature weights based on logistic regression coefficients was applied to identify the most relevant terms. The study showed that the model tends to associate truthful information with Ukrainian-language, neutral vocabulary, while texts with signs of disinformation often contain Russian-language elements characteristic of propaganda or manipulative messages. Further research will be aimed at expanding the dataset and creating new ensemble models to identify sources of disinformation.

Keywords - disinformation, fake, contextual vector representations, TF-IDF, dataset, accuracy, metric.