УДК 004.8-9 DOi 10.30837/ bi.2025.1(102).02

О.В. Лозинська[1], О.О. Марків [2], В.А. Висоцька [3]

[1]НУЛП, м. Львів, Україна, olha.v.lozynska@lpnu.ua, ORCID iD: 0000-0002-5079-0544

[2]НУЛП м. Львів, Україна, oksana.o.markiv@lpnu.ua, ORCID iD: 0000-0002-1691-1357

[3]НУЛП, м. Львів, Україна, victoria.a.vysotska@lpnu.ua, ORCID iD: 0000-0001-6417-3689

МЕТОД ВИЯВЛЕННЯ ДЖЕРЕЛ ДЕЗІНФОРМАЦІЇ НА ОСНОВІ АНСАМБЛЕВИХ МОДЕЛЕЙ МАШИННОГО НАВЧАННЯ

У представленому дослідженні розроблено метод виявлення джерел дезінформації на основі ансамблевих моделей машинного навчання. Проаналізовано сучасні методи боротьби з дезінформацією та виявлення неправдивого контенту. В рамках роботи реалізовано систему ідентифікації фейків, побудовану на ансамблевому підході, а також описано її архітектурну структуру. Детально описано основні етапи очищення текстових даних, отриманих із соціальних мереж і новинних, зокрема нормалізацію категоріальних змінних. Проведено статистичний аналіз тексту та аналіз критеріїв виявлення джерел поширення дезінформації. Здійснено аналіз балансу цільових і допоміжних змінних, що дало змогу виявити залежності між мовою повідомлення та достовірністю. Для моделювання використано два різні типи текстових ембедингів та відповідні моделі класифікації: лінійну регресію та логістичну регресію. Підсумковим етапом стало застосування ансамблю моделей, що дало змогу поєднати прогностичну здатність обох моделей. Результати показали, що комбінація підходів покращує класифікаційну якість, особливо в умовах незбалансованих даних. Використання ансамблю моделей дало змогу збільшити точність з 73% (модель 1) та 71% (модель 2) до 78%.

ДЕЗІНФОРМАЦІЯ, ДАТАСЕТ, МАШИННЕ НАВЧАННЯ, АНСАМБЛЕВІ МОДЕЛІ, ЛІНІЙНА РЕГРЕСІЯ, ЛОГІСТИЧНА РЕГРЕСІЯ, ЕМБЕДИНГ

O.V. Lozynska, O.O. Markiv Oksana, V.A. Vysotska. Method for detecting sources of disinformation based on ensemble machine learning models. The presented study developed a method for detecting sources of disinformation based on ensemble machine learning models. Modern methods of combating disinformation and detecting false content were analyzed. A fake news identification system based on the ensemble approach was implemented as part of the work, and its architectural structure was described. The main stages of cleaning text data obtained from social networks and news are described in detail, in particular, the normalization of categorical variables. Statistical analysis of the text and analysis of the criteria for identifying sources of disinformation distribution are carried out. An analysis of the balance of target and auxiliary variables was performed, which made it possible to identify dependencies between the language of the message and reliability. Two types of text embeddings and corresponding classification models were used for modeling: linear regression and logistic regression. The final stage was the application of an ensemble of models, which made it possible to combine the predictive ability of both models. The results showed that the combination of approaches improves classification quality, especially in conditions of unbalanced data. Using an ensemble of models allowed us to increase the accuracy from 73% (model 1) and 71% (model 2) to 78%.

DISINFORMATION, DATASET, MACHINE LEARNING, ENSEMBLE MODELS, LINEAR REGRESSION, LOGISTIC REGRESSION, EMBEDDING

Вступ

Основним джерелом новин та даних для більшості людей є соціальні мережі та онлайн-новини, оскільки вони є легкодоступні. Однак одночасно вони також сприяють поширенню фейкових новин, тобто повідомлень, які містять дезінформацію та мають значний негативний вплив на суспільство. Такі повідомлення зазвичай мають спільні риси, такі як граматичні помилки, неправдива інформація, використання подібно обмеженого набору слів та містять емоційно забарвлену інформацію, яка впливає на думку читача [1]. Щоб вирішити цю проблему, дослідженням з ідентифікації фейкових новин останнім часом приділяють багато уваги. Незважаючи на численні обчислювальні рішення, доступні наразі для виявлення фейкових новин, відсутність комплексної бази даних фейкових новин стала однією з суттєвих перешкод. Масштабне поширення новин через соціальні мережі унеможливлює ручну перевірку, що сприяє розробці та впровадженню автоматизованих систем виявлення дезінформації [2].

Фейкові новини можуть функціонувати як пропаганда або дезінформація, але вони завжди апелюють до емоцій громадськості та мають намір приховати раціональні реакції, аналіз та порівняння інформації з кількох джерел.

Виявлення фейкових новин – це багаторівневий процес, який включає аналіз змісту новин для визначення їхньої правдивості. Новини можуть містити інформацію в різних форматах, таких як текст, відео, зображення тощо. Комбінації різних типів даних ускладнюють процес виявлення дезінформації. Оскільки фейкові новини утворюють великі, неструктуровані дані [3], попередня обробка таких даних надзвичайно важлива для очищення та структурування даних перед їх використанням у моделі виявлення.

Однією з проблем аналізу фейкових новин є правильна обробка відсутніх даних. Оскільки фейкові

новини не створюються для досліджень, більшість інформації в соціальних мережах та новинах не структурована заздалегідь. Відповідно, виникають відсутні значення, що призводить до суперечливих або упереджених статистичних результатів при застосуванні аналізу або класифікації [4].

У даній роботі запропоновано метод виявлення джерел дезінформації на основі ансамблевих моделей машинного навчання. Розглянуто відомі методи ідентифікації дезінформації та фейкових новин. Наведено основні етапи ідентифікації дезінформації, використовуючи датасет текстових новин з різних українських соцмереж та новинних сайтів. Зокрема, проведено препроцесинг датасету, який включає очищення певних непотрібних полів, нормалізацію категоріальних змінних, уніфікацію, аналіз балансу цільової змінної та моделювання. У межах моделювання застосовано два різні типи ембедингів для представлення текстових даних та відповідні моделі класифікації: лінійну регресію та логістичну регресію. Крім того, використано ансамбль цих двох моделей. Такий підхід може привести до покращення метрик класифікації інформації у різних соцмережах.

1. Методи ідентифікації дезінформації

Для ідентифікації дезінформації та фейкових новин можна ефективно використовувати такі підходи: машинне навчання, аналіз настроїв або їх комбінацію. Ці підходи можна використовувати для двох різних завдань – розпізнавання фейкових новин або їх авторів.

У роботі [5] машинне навчання та аналіз настроїв використовуються як паралельні підходи для ідентифікації не лише фейкових новин, але й їх авторів. Використання методу опорних векторів у цій роботі дало змогу досягнути значення 0,98 для F1-оцінки.

Наукова праця [6] зосереджена на виявленні ненадійних авторів онлайн-дописів за допомогою згорткової нейронної мережі (CNN), а також довгої короткочасної пам'яті (LSTM). Автори використовували дані з Twitter, а також класичні методи машинного навчання, такі як метод опорних векторів (SVM) та метод k-найближчих сусідів (KNN). Використання нейронної мережі дало змогу досягнути точності 0,93. Цей результат виявився кращим на кілька відсотків порівняно з SVM та на 10% кращим, ніж KNN. Іншим прикладом використання комбінації машинного навчання та аналізу настроїв є дослідження [7].

У науковій роботі [8] розглянуто такі методи машинного навчання для виявлення фейкових новин: метод на основі логістичної регресії, метод на основі нейронних мереж (багатошаровий перцептрон (MLP) та згорткові нейронні мережі), методи дерев рішень, баєсівські методи, класифікація С-опорних векторів. Згорткові нейронні мережі були оцінені як найкращий варіант серед інших методів, незважаючи на їх значно довший час навчання та вимогу до більших наборів даних.

У статті [9] запропоновано такі підходи для розпізнавання фейкових новин в онлайн-просторі, як логістична регресія, метод опорних векторів (SVM), k-NN, дерева рішень, випадковий ліс, згорткові нейронні мережі, вентильні рекурентні мережі та довгу короткочасну пам'ять. В експериментах було використано два набори даних, а саме ISOT та KDnugget. Завдяки використанню підходу стекування, авторам вдалося досягти найкращої точності 0,99 для моделі випадковий ліс і набору даних ISOT та 0,96 – для логістичної регресії та набору даних KDnugget.

У інших роботах, таких як [10], додатково досліджувалася методологія виявлення фейкових новин, яка враховує не лише інформацію про зміст новин, але й додаткову інформацію щодо використання соціальних мереж. Для генерації представлення новин було використано тензорну факторизацію. Було представлено порівняльний аналіз трьох підходів: перший – на основі новинних текстів, другий – на основі способу використання соціальних мереж, і третій – на основі їх комбінації. Було виявлено, що комбінований підхід, який базується на контенті та контексті, забезпечує кращі результати. Використання глибокої нейронної мережі забезпечило покращення точності та повноти (recall) порівняно з класифікатором XGBoost.

У дослідженні [11] розроблено архітектуру моделі FakeDetector, яка являє собою графову нейронну мережу. У роботі автори поєднали явні та неявні атрибути, отримані з новинних текстів, для виявлення фейкових новин. FakeDetector – це глибока дифузійна нейронна мережа, яка одночасно представляє та оцінює новини, авторів та тему статті. Таким чином, модель прогнозування може визначати достовірність авторів. У статті представлено нову модель дифузійної одиниці, а саме GDU. Модель являє собою нейронну мережу, засновану на принципі поширення інформації шляхом дифузії. Модель GDU може обробляти кілька вхідних даних одночасно та ефективно комбінувати вхідні дані для генерації виходів. Проведені експерименти на наборах даних показали задовільний рівень ефективності запропонованого підходу у виявленні фейкових новин та їх авторів в онлайн-просторі.

Авторами дослідження [1] застосовано такі методи машинного навчання як метод опорних векторів, метод k-найближчих сусідів, згорткові нейронні мережі, а також ансамблеві методи. Використання ансамблевих навчальних методів дало змогу досягти кращих оцінок за всіма показниками продуктивності порівняно з окремими навчальними методами.

У роботі [12] запропоновано новий підхід DocEmb

для виявлення фальшивих повідомлень за допомогою вбудовування текстових даних. У статті представлені експерименти з різними текстовими представленнями, такими як TF-IDF або з використанням вбудовування слів та трансформаторів: Word2Vec SG та CBOW, FastText SG та CBOW, GloVe, BERT, BART та RoBERTa. Автори навчали моделі на цих текстових представленнях, використовуючи методи машинного навчання, такі як наївний байєсівський метод, градієнтно-підсилені дерева та моделі глибокого навчання – Perceptron, Multi-Layer Perceptron, [Bi]LSTM та [Bi] GRU. Результати дослідження показують, що представлення слів у документі відіграє важливу роль для досягнення більшої точності.

Більшість спроб виявлення фейкових новин зазвичай зосереджені лише на текстовій інформації. Мультимодальні підходи менш поширені та зазвичай класифікують повідомлення як правдиві або хибні. У роботі [13] описано мультимодальний підхід до виявлення фейкових новин. Виявлення проводилося з використанням як унімодального, так і мультимодального підходів. Результати роботи показали, що мультимодальний підхід досяг досить задовільних результатів у випадку, коли він базувався на архітектурі згорткової нейронної мережі, де поєднувалися лише текстові та графічні дані. Таким чином, використання як текстових, так і графічних даних покращує виявлення фейкових новин.

У дослідженні [12] наведено огляд існуючих інструментів, придатних для виявлення фейкових новин, та огляд веб-сайтів, які можна використовувати для перевірки фактів. Вони вказують на важливість донесення до громадськості методів виявлення дезінформації.

Авторами [15] представлено методи глибокого навчання для створення ансамблю моделей. У статті [16] описано підхід з використанням різноманітних ансамблевих стратегій з метою підвищення ефективності виявлення дезінформації для певного набору моделей.

У науковій роботі [17] розглянуто метод ідентифікації фейкових новин шляхом використання архітектури BERT на даних із соціальних мереж. Поряд із текстовим вмістом публікацій, автори також включили дев'ять додаткових атрибутів, що стосуються користувача або контенту, створеного користувачем, таких як кількість друзів, перегляди або використання хештегів у публікаціях користувача. Отримані результати дали змогу досягти оцінки F1 приблизно 85%. У випадку [18] виявлення фейкових повідомлень проводилося з використанням архітектури на основі CNN та LSTM. Для векторного представлення автори використовували W2V. Набір даних містив заголовки та текстовий вміст новинних статей, що супроводжувалися пов'язаними статтями, позначеними як джерела. Метою цього дослідження було оцінити, чи мають статті схожість, а також зв'язок з відповідними джерелами. Автори досягли результатів для F1-оцінки, що становить 0,97.

2. Пропонований виявлення джерел дезінформації з використання ансамблевих моделей

Для реалізації методу виявлення джерел дезінформації авторами запропоновано та розроблено систему ідентифікації дезінформації з використання ансамблю моделей, яка включає в себе попередню обробку (препроцесинг) даних датасету, вибір текстових ознак, векторизацію (ембедінг), використання моделей для класифікації тексту та оцінку даних моделей. Архітектура даної системи ідентифікації зображена на рис. 1.

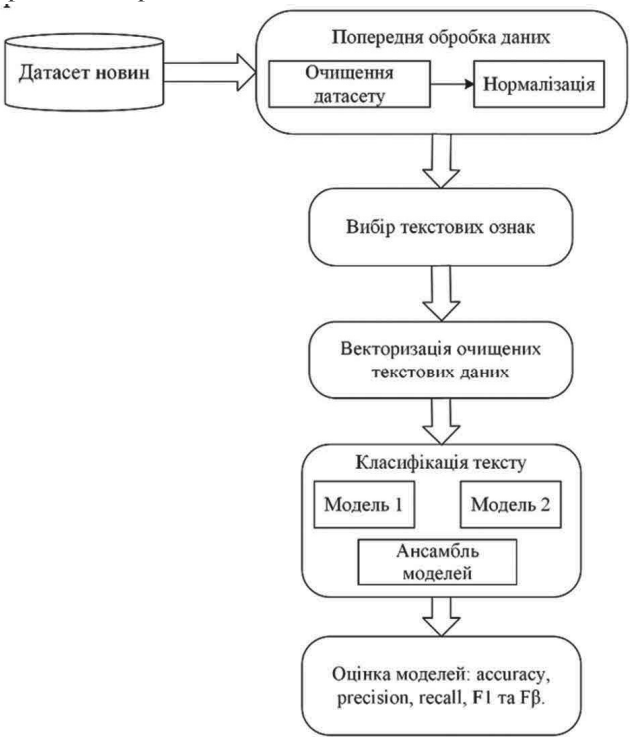

Рис. 1. Архітектура системи ідентифікації дезінформації

2.1. Нормалізація та препроцесинг датасету новин

Для забезпечення достовірності аналізу та точності побудови моделі ідентифікації фейкових новин використовуються нормалізація та препроцесинг датасету україномовних новин. Препроцесинг включає в себе очищення певних полів датасету. Для аналізу використано датасет, розроблений авторами [19] (рис. 2).

Очищення датасету є одним з ключових етапів підготовки даних до аналізу, що включає виявлення та усунення некоректних, відсутніх або зайвих значень. Цей процес необхідний для покращення якості даних, підвищення достовірності результатів та точності побудови моделей машинного навчання.

У межах попередньої обробки даних було виконано наступне:

5 rows × 14 columns

	Unna…	Unnam…	Unnamed: 2	Unnamed: 3	Unnamed: 4	Un…	Un…	Unname…	Unnamed: 8	Unnamed: 9	Unnamed: 10	Unnamed: 11	Unnamed…	Unnamed: 13
0	NaN	№	Дата	Час	Текст повідомле…	Мітка	Post/…	Автор/Gro…	Джерело	Мова	Like	Поширення	Веб-адреса	NaN
1	NaN	NaN	20.08.2024	10:06	Оборона ВСУ в Д…	FALSE	Post	Einarr Mo…	Facebook	Russian	5	3	https://w…	NaN
2	NaN	NaN	23.09.2024	7:01	FLASH COLONEL M…	TRUE	Post	Psn Anima…	Facebook	French	146	12	https://w…	NaN
3	NaN	NaN	11.09.2024	17:37	🔍 Wer kontroll…	FALSE	Post	Wissensch…	Facebook	German	13	5	https://w…	NaN
4	NaN	NaN	12.09.2024	14:01	Астрологический…	FALSE	Post	Магия кот…	Facebook	Russian	332	64	https://w…	NaN

Рис. 2. Датасет після завантаження

Видалення неінформативних колонок: після імпорту датасету, який складався з 14 колонок, залишено лише ті, що є релевантними для подальшого аналізу (текст повідомлень, мітка, post/repost, автор/group, джерело, мова).

Нормалізація категоріальних змінних: оскільки значення деяких категоріальних полів були неоднорідними, їх було приведено до уніфікованого формату — логічних значень типу True/False.

Нормалізацію застосовано до таких колонок як «мітка», «пост/репост» та «мова», які потрібно було привести до стандартизованого вигляду, з огляду на велику кількість варіантів формулювання. На рис. 3-5 зображено значення колонок «мітка», «пост/репост» та «мова» до уніфікації, на рис. 6 – значення колонки «мова» після уніфікації.

```
print(df['Мітка'].unique())
✓ [4] < 10 ms
['FALSE' 'TRUE' 'фейк' 'істина' 'Фейк' 'Істина' ' False' ' True ' 'F' 'T'
 'T' 'F ' ' фейк' 'правда' 'Брехня' 'брехня' nan ' TRUE' 'FAKE' 'TRUE '
 'Fake' 'True' 'False' 'Реальна' 'Фейкова']
```

Рис. 3. Значення колонки "мітка" до уніфікації

```
print(df['Post/Repost'].unique())
✓ [6] < 10 ms
['Post' 'Reel' 'post' 'reporter' 'repost' 'Publication' ' Post' 'Repost'
 'POST' 'post ' 'video' 'Post ' 'REPOST' 'News' 'Пост']['Post' 'Reel' 'post' 'reporter'
  'repost' 'Publication' ' Post' 'Repost'
 'POST' 'post ' 'video' 'Post ' 'REPOST' 'News' 'Пост']
```

Рис. 4. Значення колонки «пост/репост» до уніфікації

```
print(df['Мова'].unique())
✓ [8] < 10 ms
['Russian' 'French' 'German' 'English' 'Ukrainian' 'Russian ' 'українська'
 'російська' 'польська' nan 'russian' 'Українська' ' Українська'
 'англійська' 'Український' 'Російська' 'УКРАЇНСЬКА' 'РОСІЙСЬКА' 'Чеська'
 'Англійська' 'RUSSIAN' 'UKRAINIAN' 'Polish' 'укр' 'Slovak' 'Poland'
 'Ukranian' 'ukranian' 'Укріїнська' 'Польська' 'Російська '
 'Ukrainian (the original language \nis Russian)' 'ukrainian' 'Латвійська'
 'Французька' 'Українська переклад з російської']
```

Рис. 5. Значення колонки «мова» до уніфікації

```
print(df['Мова'].unique())
✓ [9] < 10 ms
['RUS' 'FRA' 'GER' 'ENG' 'UKR' 'POL' 'CZE' 'SVK' 'LVA']
```

Рис. 6. Значення колонки «мова» після уніфікації

У результаті виконаних трансформацій датасет було очищено від зайвих колонок та нормалізовано всі основні категоріальні змінні (рис. 7).

5 rows × 6 columns

	Текст повідомлень	Мітка	Post/Repost	Автор/Group	Джерело	Мова
1	Оборона ВСУ в Донецкой области оста…	False	Post	Einarr Modig	Facebook	RUS
2	FLASH COLONEL MACGREGOR: La Contre-…	True	Post	Psn Animation Média	Facebook	FRA
3	🔍 Wer kontrolliert die Algorithmen…	False	Post	Wissenschaftswelle.de	Facebook	GER
4	Астрологический прогноз для тебя от…	False	Post	Магия которая работает	Facebook	RUS
5	Увлекательные факты о городе Днепр\…	True	Post	Сергей Владимирович Пискун	Facebook	RUS

Рис. 7. Датасет після етапів очищення та нормалізації

На наступному кроці було здійснено очищення полів імен авторів та назв джерел, що є критично важливим під час обробки текстових даних, особливо якщо джерела – відкриті або неструктуровані. Часто імена чи назви одного об'єкта представлені у різних варіаціях, містять спеціальні символи, емоційні вирази чи граматичні відхилення. У нашому випадку це стосується поля «автор/group», що містить значну кількість варіацій, включно зі смайлами, знаками «@», переносами рядків, порожніми або нечіткими значеннями.

Для уніфікації:

– видалено зайві символи;

– текст переведено у нижній регістр;

– порожні значення замінено на стандартне значення «unknown»;

– об'єднано поля «джерело» та «веб-сторінка» у нове поле з витягнутим доменом як репрезентативним ідентифікатором джерела;

– застосовано правила перетворення, наприклад, заміна «t» на «telegram».

2.2. Аналіз категоріальних змінних

Цей етап спрямований на вивчення якісних характеристик даних, що представлені у вигляді категорій. Для комплексного аналізу було проведено:

1) аналіз дублікатів – за допомогою функції duplicated() виявлено 10 повторюваних записів;

2) аналіз пропущених значень – використано функцію isnull().sum(), результати якої показують відсутність пропусків у колонках (рис. 8);

3) оцінку балансу категорій – виконано візуалізацію для полів «пост/репост», «мова» та «мітка»

(target), що виявила суттєву диспропорцію (рис. 9);

4) розподіл значень відносно поля «мітка» — здійснено аналіз залежності між мовою публікації та ймовірністю її правдивості (рис. 10);

5) крос-табуляцію — побудовано таблиці для оцінки співвідношення категоріальних змінних до цільової ознаки (рис. 11 та рис. 12).

```
df.isnull().sum()

Текст повідомлень    0
Мітка                0
Post/Repost          0
Автор/Group          0
Джерело              0
Мова                 0
dtype: int64
```

Рис. 8. Кількість пропущених значень у кожній з колонок

Побудова крос-таблиць дала змогу оцінити числове співвідношення між категоріальними змінними. Цей метод є ефективнішим у порівнянні з графічною візуалізацією у випадках, коли маркування на діаграмах не є чітким. Було сформовано крос-таблиці для зв'язків між полями «мова» та «мітка», а також «пост/репост» та «мітка».

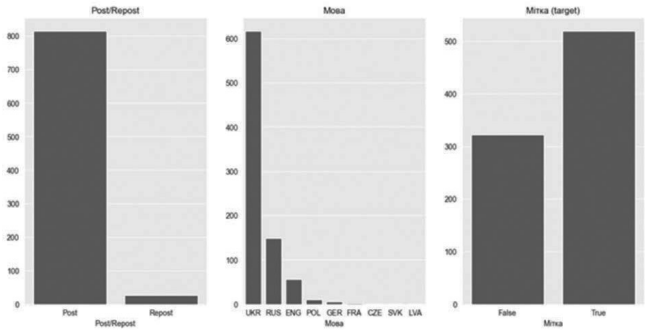

Рис. 9. Баланс категорій для трьох колонок: «пост/репост», «мова», «мітка»

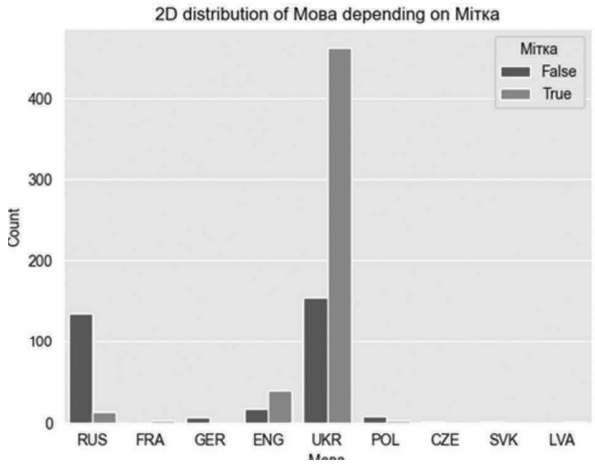

Рис. 10. Розподіл значень поля «мова» відносно значень поля «мітка»

Мова / Мітка	CZE	ENG	FRA	GER	LVA	POL	RUS	SVK	UKR
False	1	17	0	6	0	8	135	1	154
True	0	39	2	0	1	2	13	0	462

Рис. 11. Співвідношення між кількістю значень полів «мова» та «мітка»

Post/Repost / Мітка	Post	Repost
False	311	11
True	503	16

Рис. 12. Співвідношення між кількістю значень полів «мітка» та «пост/репост»

2.3. Оцінка балансу цільової змінної

Аналіз балансу цільової змінної («мітка») дає змогу візуально ідентифікувати можливий дисбаланс класів, що може призводити до упередженості у роботі моделей машинного навчання. Цей крок важливий для прийняття рішень щодо використання технік балансування перед навчанням моделей. На рис. 13 зображено кругову діаграму, яка показує переважання правдивих повідомлень у вибірці.

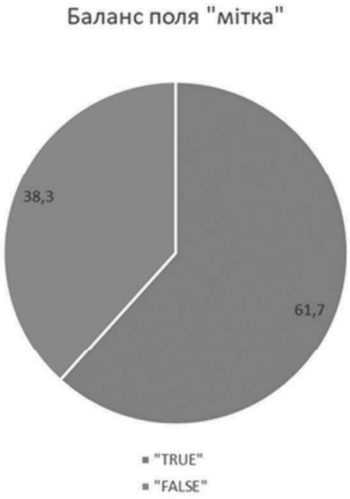

Рис. 13. Діаграма балансу поля «мітка»

Для коректної обробки тексту було проведено видалення всіх зайвих елементів, таких як емоджі, посилання, хештеги, тощо. Це забезпечує чистоту вхідних даних та підвищує якість подальшого аналізу.

2.4. Статистичний аналіз тексту

Використання статистики тексту дає змогу кількісно охарактеризувати зміст повідомлень. Було оцінено:

- довжину текстів,
- кількість слів,
- розподіл за реченнями,
- загальну складність тексту.

За допомогою гістограм із KDE-графіком візуалізовано щільність розподілу довжин текстів. Для

виявлення відмінностей між класами застосовано Boxplot, що дало змогу оцінити типову кількість слів, розкид та наявність аномалій (рис. 14-16).

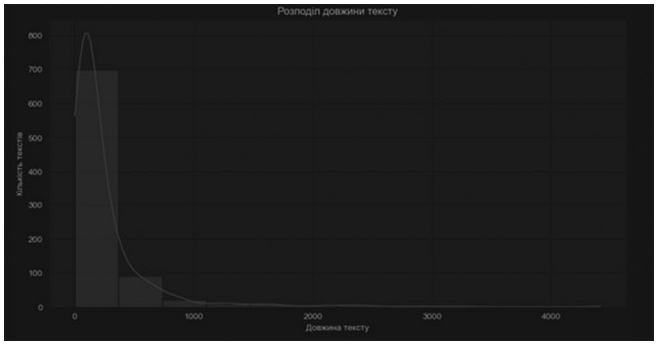

Рис. 14. Гістограма розподілу довжини тексту

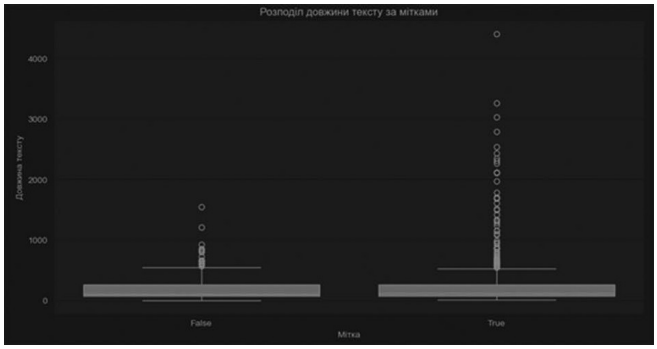

Рис. 15. Діаграма розподілу довжини тексту за мітками

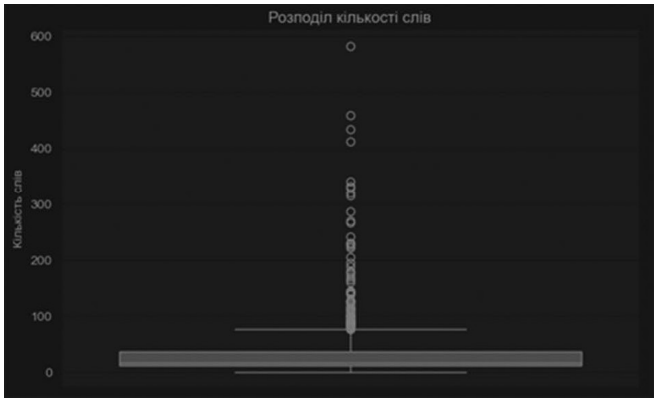

Рис. 16. Діаграма розподілу кількості слів

На рис. 17 зображено хмару слів для ілюстрації найчастіше вживаних слів у корпусі, що дає змогу інтуїтивно ідентифікувати основні теми та контекст повідомлень.

Рис. 17. Хмара слів

2.5. Критерії та параметри для виявлення джерел дезінформації

Крім статистичного аналізу тексту проведено аналіз критеріїв та параметрів, які можуть впливати на виявлення саме джерел дезінформації. Серед критеріїв, які дають змогу виявляти джерела поширення дезінформації обрано основні наративи пропаганди (у датасеті таких 40). Проведено аналіз датасету і обрано найпоширеніші наративи: «НАТО, ЄС», «БІЛЯ ДНІПРА», «АЕС», «ОХМАТДИТ», «КУРСЬК», «МОБІЛІЗАЦІЯ». Діаграма розподілу даних наративів у датасеті зображена на рис. 18. Проведений аналіз дає змогу окреслити наступні висновки:

– якщо у новині використано наратив «НАТО, ЄС» і мова даної новини російська – то дана новина потенційно є фейком;

– якщо більшість новин (постів) на сторінці джерела фейкові, швидше за все дане джерело чи аккаунт є також фейковим

– потрібно відслідковувати час і дату поширення фейкової новини від першоджерела.

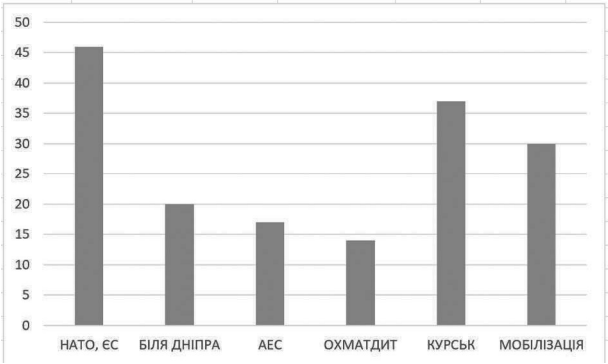

Рис. 18. Діаграма розподілу найпоширеніших наративів у датасеті

3. Експериментальні дослідження

У межах моделювання застосовано два різні типи ембедингів для представлення текстових даних:

1. intfloat/multilingual-e5-base
2. sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

На основі кожного з ембедингів було побудовано окремі моделі класифікації із застосуванням лінійної та логістичної регресії, які прогнозують ймовірність правдивості тексту. Для покращення якості класифікації результати обох моделей було інтегровано, а також проведено порівняння їх ефективності за допомогою відповідних метрик.

3.1. Модель 1 (лінійна регресія)

На першому етапі ембединги були згенеровані за допомогою інструменту intfloat/multilingual-e5-base. Тексти було попередньо підготовлено, де кожен рядок отримав префікс «query:», відповідно до специфікації моделі. В результаті обробки було отримано вектори з 768 ознаками.

Для зменшення розмірності векторів застосовано метод головних компонент (PCA). Розподіл значень у двох головних компонентах подано на рис. 19. Аналіз двовимірної проєкції показав значне перекриття між класами, що свідчить про неефективність PCA у даному випадку. Тому у подальших етапах навчання було збережено повну розмірність ембедингів.

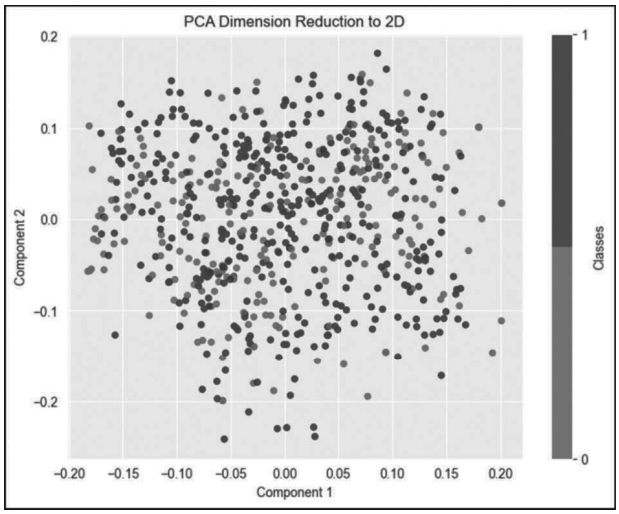

Рис. 19. Графічне зображення зменшення вимірності до 2D з використанням PCA

У процесі навчання моделі лінійної регресії було застосовано метод оптимізації Randomized Search, який виконує випадковий підбір гіперпараметрів із заздалегідь визначеного діапазону. Для забезпечення стабільності та достовірності результатів модель оцінювалася за допомогою крос-валідації. Визначена конфігурація найкращих параметрів була використана для фінального навчання моделі, що позитивно вплинуло на її здатність до узагальнення та точність ідентифікації дезінформації.

Для оцінки роботи моделі побудовано три типи матриці неточностей (confusion matrix):

1) класична (абсолютна) — відображає кількість випадків кожного типу класифікації;

2) нормалізована за рядками — показує, наскільки точно модель класифікує правдиві та фейкові тексти;

3) нормалізована за стовпцями — дає змогу оцінити точність передбачень по кожному з прогнозованих класів.

Ці візуалізації (рис. 20) надають змогу глибше зрозуміти сильні та слабкі сторони моделі.

Оцінка результатів виявила високий рівень помилок першого роду (False Negatives) — велика кількість фейкових повідомлень була неправильно класифікована як правдиві. Це критично у контексті інформаційної безпеки, адже модель пропускає потенційно шкідливий контент. Водночас рівень помилок другого роду (False Positives) є низьким — менше 1% правдивих новин помилково віднесено до фейкових. Це вказує на відносну стриманість моделі у «звинуваченні» достовірних джерел.

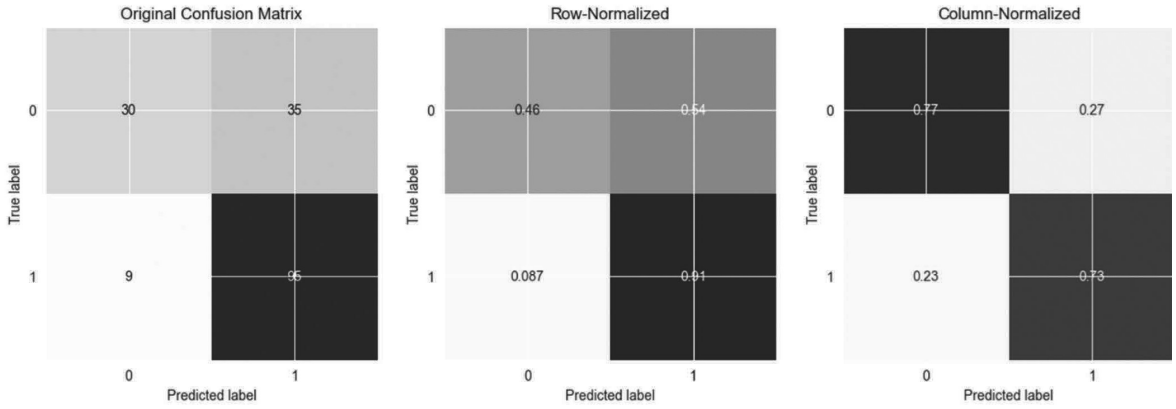

Рис. 20. Матриці невідповідностей для першої моделі

Оцінка метрик класифікації. З огляду на незбалансованість класів у вибірці, загальний показник точність (Accuracy) є малорелевантним. Натомість для оцінки ефективності було використано такі метрики:

- Precision, що відображає точність визначення фейкових новин;
- Recall показує здатність моделі виявити усі фейкові повідомлення;
- F1-метрику, яка показує узагальнений баланс precision та recall;
- Fβ-метрику, адаптовану для підвищеної чутливості до precision.

Отримане значення F1 = 0.81 свідчить про загалом ефективну роботу моделі, попри наявність певних упущень. Додаткова Fβ-метрика допомогла акцентувати увагу на важливості мінімізації помилок другого типу. Результати класифікаційних метрик наведено у табл. 1.

Для оцінки здатності моделі розрізняти між класами False Negatives та False Positives було використано ROC AUC і побудовано ROC-криву (див. рис. 21). Хоча значення ROC AUC не досягло високих показників, це свідчить про потенціал до подальшого вдосконалення моделі, зокрема в аспектах зменшення частоти хибнопозитивних значень.

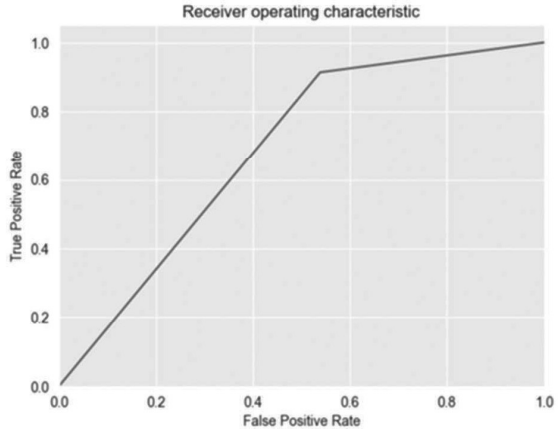

Рис. 21. ROC крива для моделі 1

3.2. Модель 2 (логістична регресія)

У другій моделі як інструмент для векторизації тексту було використано sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2.

Після генерації ембедингів дані поділено на тренувальну та тестову вибірки у співвідношенні 80:20.

Наступним кроком став підбір гіперпараметрів за допомогою Grid Search, після чого оптимальні параметри було застосовано до логістичної моделі. Отримана точність (Accuracy) моделі становить 0.716.

Також було побудовано ROC-криву (рис. 22), яка підтвердила, що загальна якість моделі потребує покращення (низький ROC AUC).

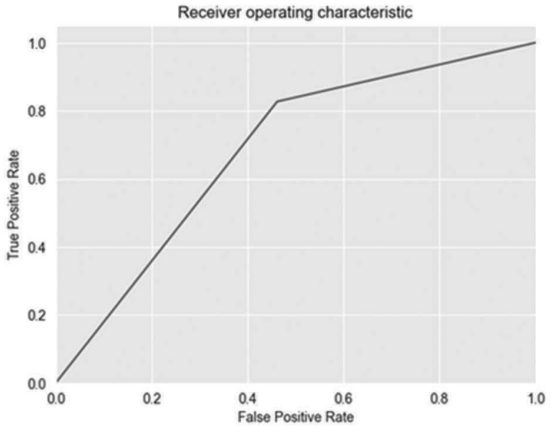

Рис. 22. ROC-крива для моделі 2

Побудована матриця невідповідностей показала (рис. 23):

a) високий рівень помилок першого роду – значна частина фейкових новин була визначена як правдиві;

b) низький рівень помилок другого роду – менше 2% достовірних новин класифіковано як фейки.

Додатково обчислено основні метрики класифікації: Recall становить 0.8269, Precision – 0.7414, метрика F1= 0.7818, метрика Fβ-score = 0.7570.

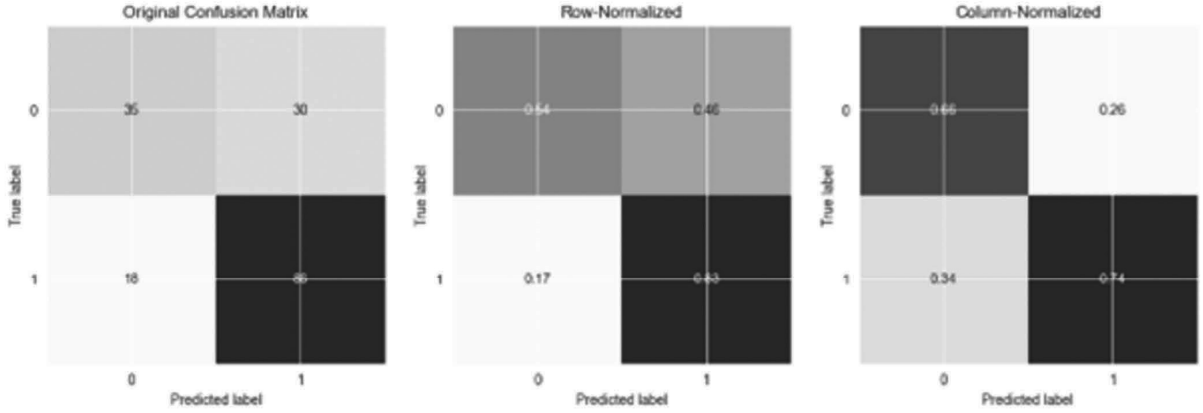

Рис. 23. Матриці невідповідностей для другої моделі

3.3. Використання ансамблю моделей

Для підвищення ефективності ідентифікації фейкових новин реалізовано метод стекування моделей. Для цього об'єднано результати двох попередньо навчених моделей – без повторного навчання, лише на основі тестової вибірки.

Попередньо для кожної моделі обчислено ймовірності приналежності об'єктів до позитивного класу. Після цього ці ймовірності було усереднено і утворено новий комбінований прогноз. На наступному кроці проведено оптимізацію порогу класифікації (threshold) у діапазоні 0.1–0.9 шляхом вибору значення, при якому оцінка F1 є максимальною. Після цього було обчислено загальну оцінку якості ансамлевої моделі, яка включає такі основні метрики, як Accuracy, Recall, Precision, оцінки F1 та Fβ, ROC AUC.

Усі значення були зібрані у порівняльну таблицю 1 для аналізу відносної ефективності кожної моделі та їх комбінації. Результати свідчать, що стекування дало змогу підвищити якість класифікації новин, уникаючи окремих слабких сторін кожної з моделей.

Таблиця 1

Показники метрик для різних моделей

Метрика	Модель 1	Модель 2	Модель 3
Accuracy	0,739645	0,715976	0,781065
Recall	0,913462	0,826923	0,913462
Precision	0,730769	0,741379	0,772358
F1	0,811966	0,781818	0,837004
Fβ	0,869963	0,808271	0,881262
ROC AUC	0,687500	0,682692	0,741346

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 цій роботі запропоновано метод виявлення джерел дезінформації на основі ансамблевих моделей машинного навчання. Проведено огляд актуальних методів ідентифікації дезінформації та неправдивого контенту. У межах дослідження реалізовано систему ідентифікації дезінформації з використанням ансамблевої моделі. Наведено архітектуру даної системи ідентифікації. Описано основні етапи аналізу текстових даних, отриманих із соціальних мереж і новинних ресурсів, зокрема: очищення датасету від нерелевантних полів, нормалізацію категоріальних змінних, уніфікацію даних, вивчення балансу цільової змінної та побудову моделей. Початкові етапи очистки дали змогу отримати структурований датасет із нормалізованими значеннями, що підвищило якість подальшого аналізу. Для моделювання було застосовано два підходи до генерації ембедингів тексту, а також реалізовано класифікацію за допомогою лінійної та логістичної регресії. Моделювання за допомогою двох типів ембедингів і різних класифікаторів виявило сильні сторони кожної моделі, зокрема високе значення повноти (recall), проте наявність false positives і false negatives вказує на необхідність додаткових підходів до балансування.

На завершальному етапі реалізовано комбінування моделей через метод ансамблювання, що дало змогу підвищити якість класифікації новинного контенту у соціальних медіа. Використання ансамблю моделей дало змогу збільшити точність з 73% до 78%. Результати дослідження демонструють ефективність поєднання класичних підходів до попередньої обробки тексту із сучасними методами машинного навчання для задач інформаційної безпеки.

Перспективним продовженням досліджень є розширення ансамблю моделей та використання на інших датасетах запропонованої моделі.

Список літератури:

[1] Ahmad, I., Yousaf, M., Yousaf, S., & Ahmad, M. (2020). Fake news detection using machine learning ensemble methods. Complexity 2020, 1–11.

[2] Harb, J.G., Ebeling, R., & Becker, K. (2020). A framework to analyse the emotional reactions to mass violent events on Twitter and influential factors. Inform Process Manag, 57(6).

[3] Akinyemi, B. (2020). An improved classification model for fake news detection in social media. International Journal of Information Technology and Computer Science, 12(1), pp. 34–43. https://doi.org/10.5815/ijitcs.2020.01.05.

[4] Maniruzzaman, M., Rahman, M.J., Al-MehediHasan, M., Suri, H.S., Abedin, M.M., et al. (2018). Accurate diabetes risk stratification using machine learning: role of missing value and outliers. J Med Syst, 42(5), pp. 92. https://doi.org/10.1007/s10916-018-0940-7.

[5] Machova, K., Mach, M., & Vasilko, M. (2022). Comparison of Machine Learning and Sentiment Analysis in Detection of Suspicious Online Reviewers on Different Type of Data. Sensors, 22, 155.

[6] Sansonetti, G., Gasparetti, F., D'aniello, G., & Micarelli, A. (2020). Unreliable Users Detection in Social Media: Deep Learning Techniques for Automatic Detection. IEEE Access, 8, 213154–213167.

[7] Kandasamy, V., Trojovský, P., Machot, F.A., Kyamakya, K., Bacanin, N., Askar, S., & Abouhawwash, M. (2021). Sentimental Analysis of COVID-19 Related Messages in Social Networks by Involving an N-Gram Stacked Autoencoder Integrated in an Ensemble Learning Scheme. Sensors, 21, 7582.

[8] Papakostas, D., Stavropoulos, G., & Katsaros, D. (2022). Evaluation of Machine Learning Methods for Fake News Detection. In Combating Fake News with Computational Intelligence Techniques, Studies in Computational Intelligence; Lahby, M., Pathan, A.K., Maleh, J., Shafer-Yafooz, W.M., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, Volume 1001, pp. 163–183.

[9] Jiang, T., Li, J.P., Haq, A.U., Saboor, A., & Ali, A. (2021). A Novel Stacking Approach for Accurate Detection of Fake News. IEEE Access 2021, 9, 22626–22639.

[10] Kaliyar, R.K., Goswami, A., & Narang, P. (2021). DeepFakE: Improving fake news detection using tensor decomposition-based deep neural network. J. Supercomput. 77, 1015–1037.

[11] Zhang, J., Dong, B., & Yu, P.S. (2020). FakeDetector: Effective fake news detection with deep diffusive neural network. In Proceedings of the International Conference on Data Engineering, Dallas, USA, pp. 1826–1829.

[12] Truică, C.O., & Apostol, E.S. (2023). It's All in the Embedding! Fake News Detection Using Document Embeddings. Mathematics, 11, 508.

[13] Deepak, P., Tanmoy, C., & Cheng, L. (2021). Santhosh Kumar, G. Multi-modal Fake News Detection. Inf. Retr. Ser., 42, 41–70.

[14] Sharma, D.K., Garg, S., & Shrivastava, P. (2021). Evaluation of tools and extension for fake news detection. In Proceedings of the International Conference of Innovative Practices in Technology and Management (ICIPTM 21), India, pp. 227–232.

[15] Hrúz, M., Gruber, I., Kanis, J., Boháˇcek, M., Hlaváˇc, M., & Krˇnoul, Z. (2022). One Model is not Enough: Ensembles for Isolated Sign Language Recognition. Sensors, 22, 5043.

[16] Atitalah, S.B., Driss, M., & Almomani, I. (2022). A Novel Detection and Multi-Classification Approach for IoT-Malware Using Random Forest Voting of Fine-Tuning Convolutional Neural Networks. Sensors, 22, 4302.

[17] Heidari, M., Zad, S., Hajibabaee, P., Malekzadeh, M., Hekmati Athar, S., Uzuner, O., & Jones, J.H. (2021). BERT Model for Fake News Detection Based on Social Bot Activities in the COVID-19 Pandemic. In Proceedings of the IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York, USA, pp. 0103–0109.

[18] Umer, M., Imtiaz, Z., Ullah, S., Mehmood, A. Choi, G.S., & On, B.W. (2020). Fake News Stance Detection Using Deep Learning Architecture (CNN-LSTM). IEEE Access 2020, 8, 156695–156706.

[19] Лозинська, О., Марків, О., Висоцька, В., Романчук, Р., & Назаркевич, М. (2024). Інформаційна технологія розроблення та наповнення датасету дезінформації з використанням інтелектуального пошуку дипфейків та клікбейтів. Herald of Khmelnytskyi National University. Technical Sciences, № 343, т. 6(1), с. 158-167. DOI: 10.31891/2307-5732-2024-343-6-24.

Надійшла до редколегії 11.02.2025