Issue 18, part 1, 2025

https://doi.org/10.23939/sisn2025.18.278

УДК 004.89

МЕТОД ВИЯВЛЕННЯ ДЖЕРЕЛ ДЕЗІНФОРМАЦІЇ ТА НЕАВТЕНТИЧНОЇ ПОВЕДІНКИ КОРИСТУВАЧІВ ЧАТІВ

Марія Назаркевич[1], Назар Наконечний[2]

Національний університет "Львівська Політехніка",
кафедра інформаційних систем та мереж, Львів, Україна
[1] E-mail: mariia.a.nazarkevych@lpnu.ua, ORCID: 0000-0002-6528-9867
[2] E-mail: nazar.i.nakonechnyi@lpnu.ua, ORCID: 0009-0000-2456-3498

Розроблено метод виявлення джерел дезінформації та неавтентичної поведінки користувачів чатів у соціальних мережах. Розроблена модель базується на аналізі текстової інформації з використанням сучасних алгоритмів машинного навчання, зокрема класифікаторів (SVM, наївний Байес, дерева рішень тощо) та кластеризаційних методів для виявлення структурних зв'язків між новинами та користувачами. Значна увага приділяється збиранню і балансуванню датасетів, а також візуалізації мереж для оцінки поширення фейкових новин. Результати експериментів засвідчили високий рівень точності класифікації, з найкращими показниками у Naive Bayes, що підтверджує потенціал запропонованого підходу для автоматичного моніторингу та протидії дезінформації у соціальних мережах.

Ключові слова – фейки, дезінформація, користувачі чатів, соціальні мережі.

Постановка проблеми

Розповсюдження та поширення повідомлень у соціальних мережах створили передумови для зручного споживання новин та інформації великій кількості користувачів. Легкість, з якою інформація, сповіщення та попередження, зробила соціальні мережі платформою для поширення інформації, бо можуть бути передані великій кількості реципієнтів за дуже короткий проміжок часу. Таким чином новини у соціальних медіа дозволяють впливати на думки, дії споживачів. Звідси випливає, що неправдива інформація може з такою ж легкістю поширюватися, як і правдива. Новини впливають на думки, дії та настрої людей.

Інтернет надає платформу для швидкого поширення інформації. Позитивний вплив інтернету – люди стають більш обізнаними з місцевими та глобальними новинами, своїми правами, підвищують обізнаність щодо глобальних проблем, включаючи зміну клімату. Протягом останніх кількох років існування дезінформації в Інтернеті та зловмисних агентів, які виступають джерелами такої дезінформації, було визнано та підтверджено.

Більшість фейкових повідомлень, які поширюються в Інтернеті, складаються з кількох типів даних, об'єднаних разом – наприклад, сфабрикованого зображення разом із текстом, пов'язаним із зображенням. Зазвичай це робиться для того, щоб зробити статтю більш правдоподібною для користувача. Крім того, до посту можуть бути коментарі, де деякі користувачі могли зазначити, що стаття виглядає сфабрикованою та фальшивою, таким чином спонукаючи інших користувачів перевірити достовірність джерела.

Таким чином, постановкою проблеми є:

- Розробити та навчити модель виявляти джерела дезінформації та неавтетичної поведінки користувачів чатів.
- Оцінити продуктивність моделі виявлення джерел дезінформації за допомогою стандартних метрик, таких як точність, матриці помилок, точність класифікації та чутливості.
- Порівняти точність з іншими моделями машинного та глибинного навчання.

Аналіз останніх досліджень та публікацій

Багато дослідників сьогодні шукають методи знаходження дезінформації та неетичну поведінку користувачів чатів у соціальних мережах.

Зокрема у (Lozynska O., 2024) проведено аналіз технологій ідентифікації фейкових новин та показана архітектура виявлення дезінформації в соціальних мережах. Визначено множину критеріїв та параметрів виявлення джерел та мереж розповсюдження дезінформації.

У (Masood F, 2019) досліджується вплив повідомлень на користувачів з соціальними сайтами Twitter та Facebook. Twitter допускає надмірну кількість спаму. Фальшиві користувачі надсилають небажані твіти користувачам для просування своїх послуг, які не тільки впливають на користувачів, але й порушують споживання ресурсів. З кожним днем зростає можливість поширення недійсної інформації. У цьому дослідженні представлено підходи до виявлення спаму в Twitter, яка класифікує методи на основі фальшивого контенту, спаму на основі джерел розповсюдження, спам у трендових темах та фальшивих користувачів.

У (Kirdemir B, 2023) описуються онлайн соціальні мережі, які все частіше змінюються маніпулятивними кампаніями. YouTube є одним з найпопулярніших веб-сайтів у світі. Однак більшість існуючих систем та досліджень для виявлення та характеристики маніпулятивних кампаній впливу зосереджено на інших платформах. У даному дослідженні спостерігаємо за характеристиками підозрілої діяльності каналів YouTube, поєднуючи кілька шарів аналізу кореляції з рухомим вікном, виявлення аномалій, виявлення піків, підходи до неконтрольованої кластеризації. Експериментальний набір даних склав 39 каналів. Результати показують, що канали, які демонструють неавтентичну активність, характеризуються відносно меншою кількістю піків у своїх аномальних патернах.

У (Pathak A, 2021) розроблено серію класифікаторів для вивчення лінгвістичних ознак, що демонструються різними типами фейкових новин та дослідження дезінформаційних заголовків. У цьому дослідженні розроблено модуль логічної перевірки дезінформації. Тому було розроблено набір даних на основі правдивості та доказами достовірності повідомлення. Позначки правдивості були сформульовані на основі вивчення стандартів, що використовуються авторитетними організаціями з перевірки фактів. Розроблено аналіз помилок, який ілюструє проблеми, пов'язані з автоматизованим завданням перевірки фактів, та визначає фактори, які можуть покращити цей процес.

У (Shu K, 2019) досліджується вплив соціальних мереж на поширення фейкових новин. Досліджується взаємозв'язок між профілями користувачів у соціальних мережах та фейковими новинами. Досліджено використання профілів користувачів у соціальних мережах для виявлення фейкових новин. Спочатку виміряно поведінку користувачів щодо поширення інформації та згруповано репрезентативних користувачів, які частіше поширюють фейкові та реальні новини; потім проведено порівняльний аналіз явних та неявних характеристик профілів між цими групами користувачів (Лейзер, Д, 2018) .

Формулювання цілі статті

Ціллю дослідження є розробити модель виявлення дезінформації та неавтентичної поведінки користувачів чатів.

Нехай існує набір даних у повідомленні K={x_1 , x_2, x_3 , … , x_{i-1} , x_i}, де x_i − це перше слово, а x_i – останнє слово в повідомленні. Припустимо, що існує L_d класів, і кожен клас d має λ_d повідомлень.

Набір даних у повідомленні необхідно зрівноважити:

Для деякого класу обчислюємо кількість повідомлень:

$$\lambda_d = \sum_{i=1}^{n} I(y_i = d),$$

де I − індикаторна функція, яка дорівнює 1, якщо y c і 0 в іншому випадку; λ_D – кількість повідомлень.

Виявляємо цільову кількість новин у кожному класі. Нехай N - деяка кількість новин у кожному класі після балансування. N обчислюємо як середнє значення кількості повідомлень у всіх класах:

$$N = \frac{\sum_{j=1}^{H_d} \lambda_d}{H_d}.$$

На третьому етапі здійснюємо балансування класів. Для цього можна здійснювати додавання наявних повідомлень з недостатньо представлених класів та використати метод Undersampling у видаленні деяких повідомлень з надмірно представлених класів.

Для кожного класу c з $nc<N$, додамо $N-nc$ копій новин:

$$D' = D \cup \{(x_i, y_i) | y_i = d, copies = N - \lambda_d\}.$$

Було проаналізовано 1400 новин, які розповсюджувалися в українському контенті від вересня 2024 року по березень 2025 року. Dataset був побудований таким чином, що фіксувалася інформація, яка була опублікована в інтернеті на сайтах, які розповсюджують новини, у соціальних мережах Telegram, fasebook, Instagram, twitter, новинах на сайтах.

Виклад основного матеріалу

Розроблено метод виявлення дезінформації, який полягає в тому, що на вході отримуємо новину, а на виході − категоризацію вхідного тексту (Chang, 2025). Для того, щоб отримати результат, необхідно для заданої частини тексту знайти найбільш значущі речення.

Потім для кожного з найважливіших речень формуємо запит у формі кон'юнкції, де x_i та y_i – ключові слова.

Виконуємо пошук (Vysotska, V., 2024) та проводимо формування результатів для набору запитів (Vysotska, V., 2025).

Створюємо набір текстів-кандидатів, які можуть бути джерелом для аналізованих текстів. Для кожного кандидата порівнюємо його з початковим текстом. Якщо виявлено високу схожість разом із заміненою сутністю, то виявляємо дезінформацію.

Визначаємо відповідність сутностей та їх атрибутів від початкового до вихідного тексту. Виділяємо замінені сутності та атрибути.

Для попереднього опрацювання повідомлень розроблено метод опрацювання повідомлень для аналізу наявності фейкового повідомлення, який формалізує процес підготовки повідомлення, яке є в чаті чи в інтернеті на наявність фейкової інформації через послідовність визначених кроків:

Завантаження повідомлення в систему опрацювання даних. Кожне повідомлення завантажується у систему, і йому присвоюється унікальний номер. Повідомлення перевіряється на наявність у системі, це свідчитиме про те, що повідомлення є дублікатом. У протилежному випадку, повідомлення додається до масиву повідомлень для подальшого опрацювання. Вважаючи, що множина вхідних даних, що формує множину повідомлень *K* буде наступною:

$$K = \{x_1, x_2, x_3, \ldots, x_{i-1}, x_i\},$$

де для прикладу xi - це слова в повідомленні.

Перевірка на наявність повторення даних допомагає забезпечити точність та надійність моделі перевірка наявності повторення даних у наборі даних, що є наступним етапом. Для повідомлень це особливо важливо, оскільки наявність повторення даних може стати негативним

наслідком на навчанні моделі. Потім необхідно узагальнити модель, а на цьому вже втрачається така здатність. Набір повідомлень описується множиною C . Оскільки є K повідомлень, то C={ K }. Копій може бути безліч, і вони ідентифікуються як одинакові повідомлення. Щоб зберегти повідомлення в множині С та отримати перевірку на наявність повторення даних, використано закономірність:

$$C=\{K_i|i\ 1,2,\ldots,N\}$$

Якщо K кількість унікальних елементів є менша за N, це означає, що у наборі даних є повторення даних. Визначено поріг схожості τ. Якщо $\text{sim}(x_i)\geq\tau$, то x_i вважається копією.

Нерівномірність розподілу даних за класами виникає тоді, коли кількість повідомлень у різних категоріях істотно відрізняється. Щоб запобігти цьому, під час підготовки даних до навчання моделей машинного навчання застосовують балансування вибірки. Його мета полягає у вирівнюванні кількості прикладів у кожному класі, що дозволяє уникнути упередженості алгоритму на користь більш чисельних категорій. У протилежному випадку модель може надмірно зосереджуватися на класах із великою кількістю даних, що призведе до зниження якості класифікації для менш представлених класів.

Навчання моделі

Навчання моделі розпочинається зі збору вибірки даних. Спочатку інформацію поділяють на тренувальний та тестовий набори: приблизно 70–80 % використовується для навчання, а решта 20–30 % — для перевірки якості. Для побудови моделі SVM застосовується тренувальний масив даних. Далі виконується попереднє опрацювання передбачає очищення та стандартизацію даних, щоб вони були придатними для роботи алгоритмів. Текстові описи товарів проходять лінгвістичну обробку, що дає змогу виділити основні ознаки. За допомогою токенізації текст розбивається на окремі слова чи фрази. Потім виконується перетворення текстових даних у числове представлення за допомогою підходів TF-IDF (частотність терміну — обернена частотність документа) або Word2Vec.

Метод кластеризації як виявлення джерел дезінформації та неавтентичної поведінки користувачів чатів

Кластеризація використовується для дослідження структури даних, проте вона не завжди може виконувати функції класифікації, оскільки не забезпечує повної інформації про набір спостережень. Один із найвідоміших методів кластеризації — k-середніх. Його принцип роботи полягає у тому, щоб поділити вибірку на такі групи, в яких відмінності між об’єктами усередині кластерів є мінімальними. Іншими словами, алгоритм намагається зменшити сумарну відстань між елементами та центроїдом їхнього кластера.

Математично формулюється:

$$W\ (C\ )\ (x\ )\ k\ =\ \sum x_i\ \in Ck_{\ i}\ -\mu_k$$

де x_i– об’єкт, який належить кластеру C_k; μ_k – середнє значення (центроїд) усіх точок кластера C_k.

У процесі роботи алгоритм відносить кожне спостереження x_i до того кластера, відстань до центроїда якого є найменшою. У результаті дані групуються так, щоб мінімізувати суму квадратів відстаней від об’єктів до їхніх центрів.

Метод виявлення джерел дезінформації полягає у тому, що здійснюємо задачу кластеризації.

1. Визначаємо k — кількість кластерів, які будуть створені у результаті роботи програми.
2. Вибираємо випадкові k об'єкти з набору даних як початкові центри кластерів.
3. Призначаємо кожне спостереження найближчому центроїду на основі евклідової відстані між об'єктом і центроїдом.
4. Для кожного з k кластерів перераховуємо центроїд кластера шляхом обчислення нового середнього значення усіх точок даних у кластері.
5. Ітеративно мінімізуємо загальну суму в межах площі.

Повторюємо кроки 3 і 4, поки центроїди не зміняться або не буде досягнуто максимальної кількості ітерацій (R за замовчуванням використовує 10 як максимальну кількість ітерацій).

Експеримент та результати

```
D:\Labs\IICT\.venv\Scripts\python.exe D:\Labs\IICT\Exa
Завантаження даних з файлу: D:\Labs\IICT\updated_news_
Створення графа...
Граф створено. Виконуємо кластеризацію...
Аналізуємо центри кластерів...
Кластер 1: Центр — повідомлення з індексом 0
Кластер 0: Центр — повідомлення з індексом 128
Кластер 2: Центр — повідомлення з індексом 3
Кластер 4: Центр — повідомлення з індексом 5
Кластер 3: Центр — повідомлення з індексом 10

Аналіз кластерів виділених новин:
Фейкова новина (ID=485) належить до кластеру 2
Фейкова новина (ID=486) належить до кластеру 0
Фейкова новина (ID=487) належить до кластеру 4
Фейкова новина (ID=488) належить до кластеру 2
Фейкова новина (ID=489) належить до кластеру 3
Фейкова новина (ID=490) належить до кластеру 2
Фейкова новина (ID=491) належить до кластеру 4
Фейкова новина (ID=492) належить до кластеру 0
Фейкова новина (ID=493) належить до кластеру 2
Фейкова новина (ID=494) належить до кластеру 2
Фейкова новина (ID=495) належить до кластеру 2
Фейкова новина (ID=496) належить до кластеру 0
Фейкова новина (ID=497) належить до кластеру 1
Фейкова новина (ID=498) належить до кластеру 2
Фейкова новина (ID=499) належить до кластеру 1
Правдива новина (ID=470) належить до кластеру 2
Правдива новина (ID=471) належить до кластеру 3
Правдива новина (ID=472) належить до кластеру 2
Правдива новина (ID=473) належить до кластеру 3
```

Рис. 1: Робота програмного ужитку виявлення джерел дезінформації

Розроблено програму (див. Рис. 1), яка виконує аналіз мережі новин, які зберігаються у файлі Excel, та візуалізують виявлені зв'язки між ними. Спочатку завантажуються дані про повідомлення, їхніх авторів або групи та джерела, після чого створюється граф, де кожне повідомлення подається як окремий вузол. Зв'язки між вузлами встановлюються на основі спільного автора або джерела, що дозволяє побудувати структуру взаємозв'язків між новинами.

Далі програмний ужиток застосовує алгоритм Louvain (Anuar, S, 2021) для виявлення спільнот або кластерів у графі, тобто груп повідомлень, які тісно пов'язані між собою (див. Рис. 2). Для кожного з кластерів визначається центральне повідомлення − таке, яке має найбільшу кількість зв'язків усередині своєї групи. Окремо аналізуються останні 30 новин: 15 з них умовно позначаються як фейкові, а інші 15 − як правдиві. Це робиться виключно для спрощення візуалізації, оскільки графік був би перегруженим через велику кількість записів у датасеті. Для кожної новини виводиться кластер, до якого вона належить, що дозволяє простежити їхнє розташування в загальній структурі.

На завершення програма візуалізує побудований граф. Кожен кластер відображається окремим кольором, центральні повідомлення, фейкові та правдиві новини − виділені різними кольорами рамки. Така візуалізація дозволяє інтуїтивно оцінити структуру мережі новин, їх згрупованість та взаємозв'язки між повідомленнями.

Завантажуємо dataset для визначення фейкових чи правдивих новин.

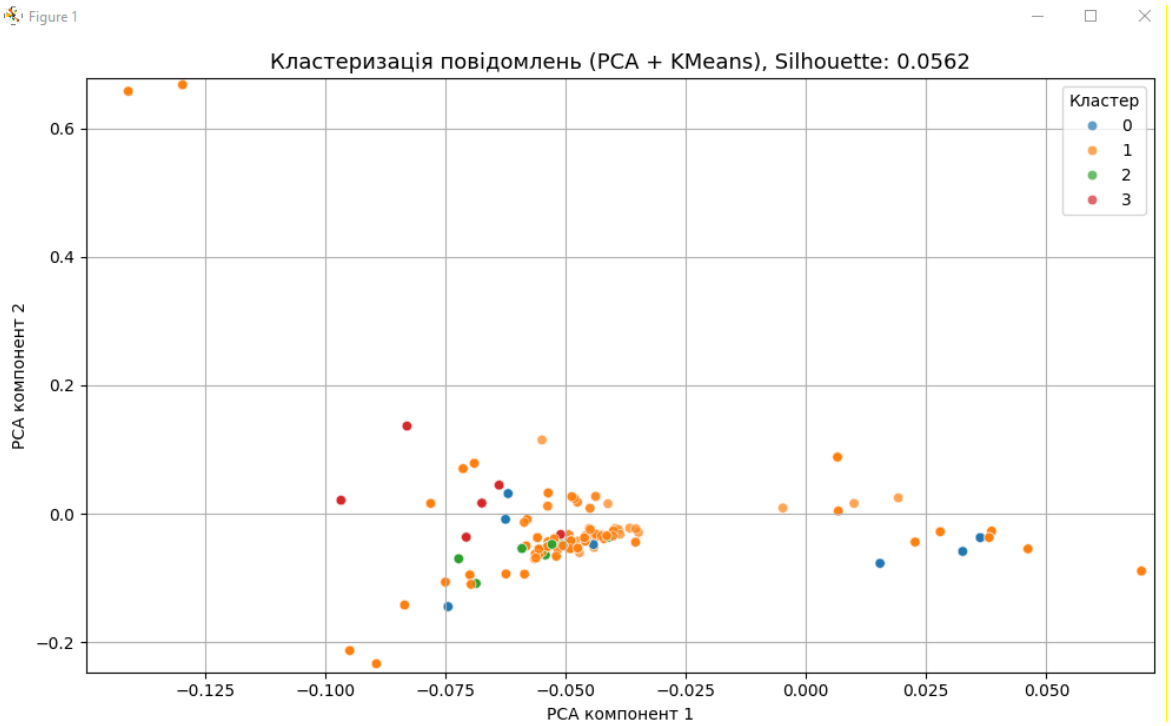

Рис. 2: Візуалізація кластеризації визначення фейкових чи правдивих новин

Робота програмного продукту починається так.

1. Проводимо опрацювання даних. Опрацювання dataset полягає у вилученні зайвих рядків, забирання непотрібних символів, тощо.
2. Здійснюємо візуалізацію dataset (див. Рис.3).

а *б* *в*

г

Рис. 3: Візуалізація dataset по а) fake і нефейкові новини; б) розподіл новин у Facebook; в) розподіл новин у website; г) розподілу в Telegram, Facebook, Instagram, Twitter, Vkontakte, Yandex

3. Застосовуємо навчання на предмет фейковості чи правдивості новини.

Для цього розбиваємо на навчальну та тестову вибірки.

```
X_train, X_test, y_train, y_test = train_test_split(X_combined, y, test_size=0.2, random_state=42)
```

Далі створюємо і навчаємо класифікатори (Сюй Х , 2025). :

```
models = {
   "Logistic Regression": LogisticRegression(max_iter=1000),
   "Decision Tree": DecisionTreeClassifier(),
   "Random Forest": RandomForestClassifier(),
   "Naive Bayes": MultinomialNB(),
   "Support Vector Machine": SVC()}
```

У результаті проведених експериментів отримано такі результати (див. Рис.4 - Рис.15).

Метод k-найближчих сусідів

Матриця плутанини класифікатора k-найближчих сусідів показано на рис. 4. Метод k-найближчих сусідів точність 85 %, що видно з рис. 5.

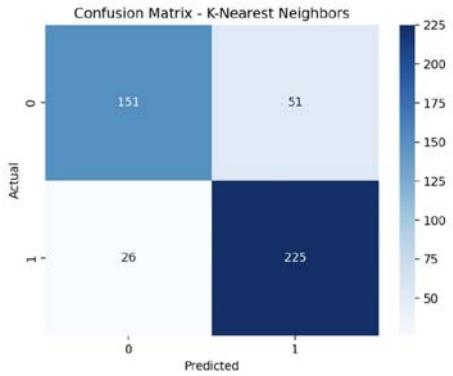

Рис. 4: Матриця плутанини k-найближчих сусідів

```
K-Nearest Neighbors
Accuracy: 0.8300
              precision    recall  f1-score   support

         0.0       0.85      0.75      0.80       202
         1.0       0.82      0.90      0.85       251

    accuracy                           0.83       453
   macro avg       0.83      0.82      0.83       453
weighted avg       0.83      0.83      0.83       453
```

Рис. 5: Результати методу k-найближчих сусідів

Метод машини опорних векторів

Матриця плутанини класифікатора машини опорних векторів показано на рис. 6. Метод Support Vector Machines точність 95%, що видно з рис. 7, що є одним з найкращих показників з досліджених класифікаторів.

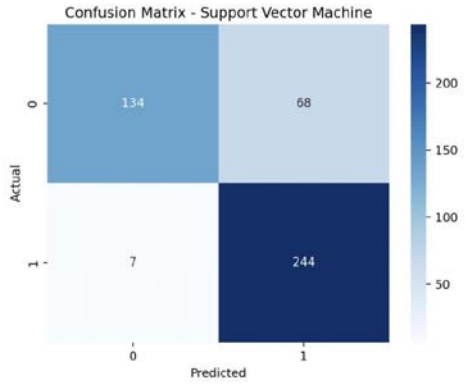

Рис. 6. Матриця плутанини машини опорних векторів

```
Support Vector Machine
Accuracy: 0.8344
              precision    recall  f1-score   support

         0.0       0.95      0.66      0.78       202
         1.0       0.78      0.97      0.87       251

    accuracy                           0.83       453
   macro avg       0.87      0.82      0.82       453
weighted avg       0.86      0.83      0.83       453
```

Рис. 7. Результати методу Support Vector Machines

Класифікатор дерева рішень

Матриця плутанини класифікатора дерева рішень показано на рис. 8. Метод класифікатора дерева рішень показав точність 78 %, що видно з рис. 9, що є найнижчим показників з досліджених класифікаторів.

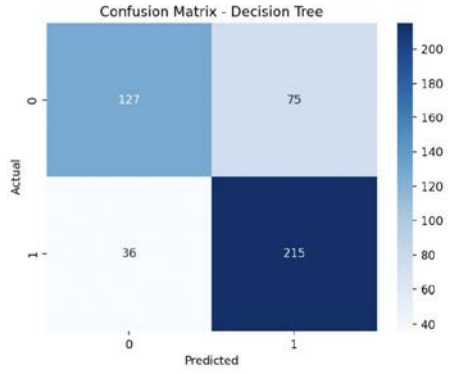

Рис.8: Матриця плутанини класифікатора дерева рішень

```
Decision Tree
Accuracy: 0.7550
              precision    recall  f1-score   support

         0.0       0.78      0.63      0.70       202
         1.0       0.74      0.86      0.79       251

    accuracy                           0.75       453
   macro avg       0.76      0.74      0.75       453
weighted avg       0.76      0.75      0.75       453
```

Рис. 9: Результати методу Decision Tree Classifier

Метод випадкового лісу

Метод випадкового лісу можна використовувати для передбачення фейкових новин, що продемонстровано на рис. 10. Метод випадкового лісу показав точність 93 %, що видно з рис. 11.

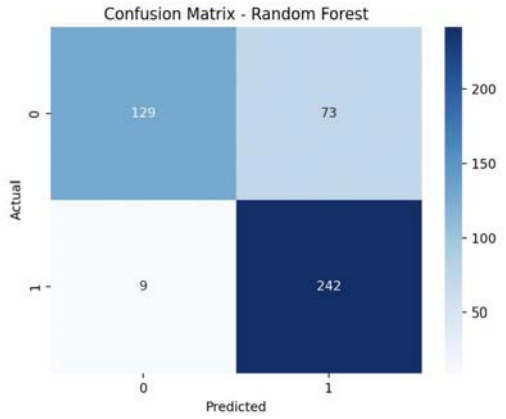

Рис. 10: Матриця плутанини методу випадкового лісу

```
Random Forest
Accuracy: 0.8190
              precision    recall  f1-score   support

         0.0       0.93      0.64      0.76       202
         1.0       0.77      0.96      0.86       251

    accuracy                           0.82       453
   macro avg       0.85      0.80      0.81       453
weighted avg       0.84      0.82      0.81       453
```

Рис. 11: Результати методу методу випадкових лісів

Наївний байесовський метод (Naive Bayes)

Одним з найкращих методів є наївний байесовський метод, у якого машинні результати передбачення фейкових новин становило 243 позиції, а нефейкових 8, що показано на рис. 12. Метод наївного байесовського методу показав точність 94%, що видно з рис. 13.

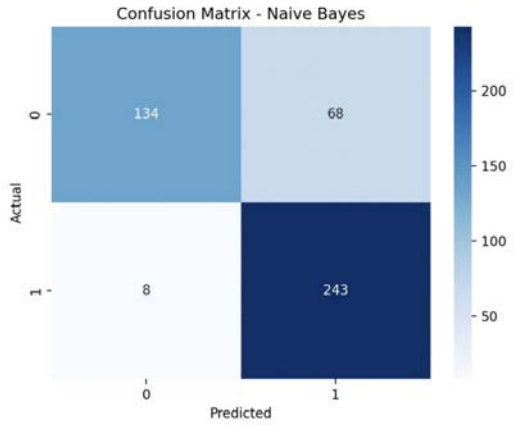

Рис. 12: Матриця плутанини наївного байесовського методу

```
Naive Bayes
Accuracy: 0.8322
              precision    recall  f1-score   support

         0.0       0.94      0.66      0.78       202
         1.0       0.78      0.97      0.86       251

    accuracy                           0.83       453
   macro avg       0.86      0.82      0.82       453
weighted avg       0.85      0.83      0.83       453
```

Рис. 13: Результати наївного байесовського методу

Метод логістичної регресії

Нижче приведено результати методу логістичної регресії, які продемонстровано на рис. 14, які показують високу точність розробленого методу. Метод логістичної регресії показав точність 95 %, що видно з рис. 15.

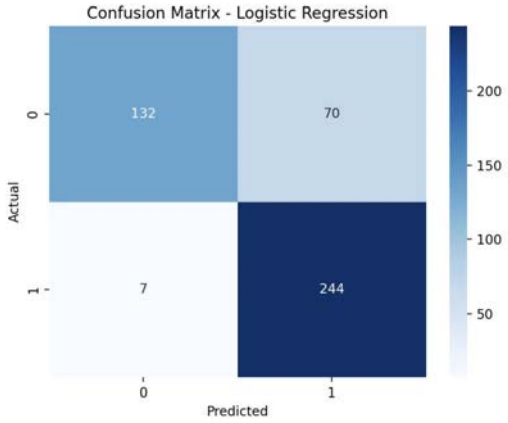

Рис. 14: Матриця плутанини логістичної регресії

```
Logistic Regression
Accuracy: 0.8300
              precision    recall  f1-score   support

         0.0       0.95      0.65      0.77       202
         1.0       0.78      0.97      0.86       251

    accuracy                           0.83       453
   macro avg       0.86      0.81      0.82       453
weighted avg       0.85      0.83      0.82       453
```

Рис.15: Результати методу логістичної регресії

Було порівняно результати вище наведених класифікаторів, які показано на рис.16 і доведено, що найкращої точності можна досягнути класифікатором випадкового лісу.

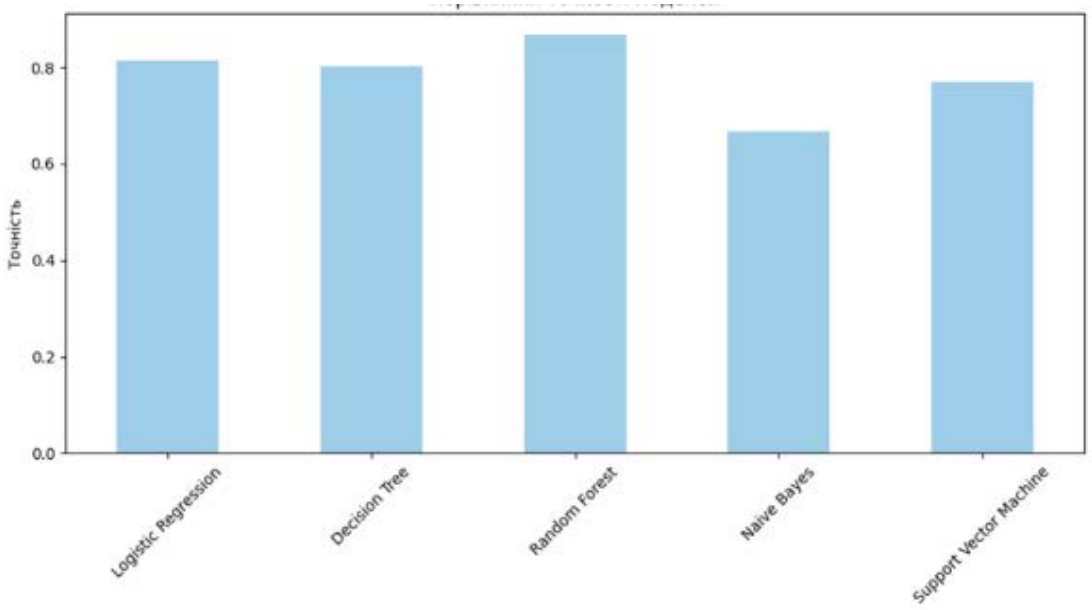

Fig.16: Порівняння точності оцінок класифікаторів.

Висновки

Розроблено метод виявлення джерел дезінформації та неавтентичної поведінки користувачів чатів у соціальних мережах. з використанням алгоритмів машинного навчання, класифікаторів SVM, наївний Байес, дерева рішень тощо) та кластеризаційних методів для виявлення структурних зв’язків.

Було проведено класифікацію новин на фейкові та правдиві, використовуючи різні алгоритми, такі як методи k-найближчих сусідів, опорних векторів, дерева рішень, випадкового лісу, наївного баєсівського, лінійний дискримінантний аналіз та логістична регресія. Було здійснено оцінку ефективності моделей та порівняння їх результатів. Найкращі показники точності продемонстрував класифікатор випадкових лісів, що робить його доцільним для використання.

Подяка

Дослідження здійснено 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».

СПИСОК ЛІТЕРАТУРИ

Anuar, S. H. H., Abas, Z. A., Yunos, N. M., Zaki, N. H. M., Hashim, N. A., Mokhtar, M. F., & Nizam, A. F. (2021, December). Comparison between Louvain and Leiden algorithm for network structure. *Journal of Physics: Conference Series, 2129(1), 012028.* https://doi.org/10.1088/1742-6596/2129/1/012028

Chang, T. P., Hsiao, T. C., Chen, T. L., & Lo, T. C. (2025). A framework for detecting fake news by identifying fake text messages and forgery images. *Enterprise Information Systems,* 19, 3-4, 2436495. https://doi.org/10.1080/17517575.2024.2436495

Kirdemir, B., & Adeliyi, O. (2023, April). Towards characterizing coordinated inauthentic behaviors on YouTube. *In Proceedings of the 2nd Workshop on Reducing Online Misinformation through Credible Information Retrieval (ROMCIR 2022), European Conference on Information Retrieval pp. 1–16.* https://par.nsf.gov/servlets/purl/10422775

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Zittrain, J. L. (2018). The science of fake news. . https://doi.org/10.1126/science.aao2998

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

Masood, F., Almogren, A., Abbas, A., Khattak, H. A., Din, I. U., Guizani, M., Zuair, M. (2019). Spammer detection and fake user identification on social networks. *IEEE Access, 7, 68140–68152*. https://doi.org/10.1109/ACCESS.2019.2918196

Pathak, A., Srihari, R. K., & Natu, N. (2021). Disinformation: Analysis and identification. *Computational and Mathematical Organization Theory, 27(3), 357–375.* https://doi.org/10.1007/s10588-021-09336-x

Shu, K., Zhou, X., Wang, S., Zafarani, R., Liu, H. (2019, August). The role of user profiles for fake news detection. In Proceedings of the 2019 IEEE/ACM *International Conference on Advances in Social Networks Analysis and Mining (pp. 436–439). IEEE.* https://doi.org/10.1145/3341161.3342891

Vysotska, V., Nazarkevych, M., Vladov, S., Lozynska, O., Markiv, O., Romanchuk, R., & Danylyuk, V. (2024). Development of a method for detecting information threats in the cyberspace of Ukraine based on machine learning. Eastern-European Journal of Enterprise Technologies, 132(2), 36-48. https://doi.org/10.15587/1729-4061.2024.317456 .

Vysotska, V., Prokipchuk, O., Nazarkevych, M., Romanchuk, R. (2025). Information technology for data authentication based on blockchain. Electronics and Information Technologies, 30, p.59–74. https://doi.org/10.30970/eli.30.5

Xu, H., Yu, P., Xu, Z., & Wang, J. (2025, January). A hybrid attention model for fake news detection using large language models. *In Proceedings of the 5th International Conference on Neural Networks, Information and Communication Engineering (NNICE 2025) (pp. 587–590). IEEE.* https://doi.org/10.48550/arXiv.2501.11967

METHOD FOR DETECTING SOURCES OF DISINFORMATION AND INAUTHENTICAL BEHAVIOR OF CHAT USERS

Maria Nazarkevych[1], Nazar Nakonechny[2]

National University “Lviv Polytechnic”,
Department of Information Systems and Networks, Lviv, Ukraine
[1] E-mail: mariia.a.nazarkevych@lpnu.ua, ORCID: 0000-0002-6528-9867
[2] E-mail: nazar.i.nakonechnyi@lpnu.ua, ORCID: 0009-0000-2456-3498

A method for detecting sources of disinformation and inauthentic behavior of chat users in social networks has been developed. The developed model is based on the analysis of text information using modern machine learning algorithms, in particular classifiers (SVM, Naive Bayes, decision trees, etc.) and clustering methods to identify structural relationships between news and users. Considerable attention is paid to the collection and balancing of datasets, as well as the visualization of networks to assess the spread of fake news. The results of the experiments showed a high level of classification accuracy, with the best indicators in Naive Bayes, which confirms the potential of the proposed approach for automatic monitoring and counteraction to disinformation in social networks.

Keywords – fakes, disinformation, chat users, social networks.