УДК 004.9

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ВИЯВЛЕННЯ ДЖЕРЕЛ ДЕЗІНФОРМАЦІЇ ТА НЕАВТЕНТИЧНОЇ ПОВЕДІНКИ КОРИСТУВАЧІВ ЧАТІВ НА ОСНОВІ МЕТОДІВ NLP ТА МАШИННОГО НАВЧАННЯ

Висоцька В. А. – д-р техн. наук, доцент, професор кафедри «Інформаційні системи та мережі», Національний університет «Львівська політехніка», Львів, Україна.

АНОТАЦІЯ

Актуальність. У сучасному цифровому середовищі поширення дезінформації та неавтентичної поведінки користувачів у чатах становить серйозну загрозу для суспільства. Методи опрацювання природної мови та машинного навчання пропонують ефективні підходи для виявлення та протидії таким загрозам.

Метою дослідження є розробка інформаційної технології для автоматичного виявлення розповсюдження джерел україномовних фейкових новин та неавтентичної поведінки користувачів чатів, яка побудована за допомогою методів опрацювання природної мови та реалізована на основі технологій машинного навчання.

Метод. Для реалізації проекту використано такі методи конструювання ознак, як статистичний показник TF-IDF, модель векторизації «Торба слів», розмічування частин мови. Для інших експериментів застосовані моделі векторизації FastText, W2V та Glove word2vec для отримання векторних представлень слів, а також розпізнавання тригерних слів (підсилюючі слова, абсолютні займенники та «блискучі» слова). Ідея полягає в знаходженні подібних за текстом/ значенням (lexical/ semantical) повідомлень, а також аналізі результатів поширення подібних повідомлень в часі та просторі. У якості основних алгоритмів моделювання використані Complement Naïve Bayes, Gaussian Naïve Bayes, HistGradientBoostingClassifier, Multinomial Naïve Bayes та RandomForest для виявлення джерел розповсюдження дезінформації та неавтентичної поведінки чатів.

Результати. У даній статті розглядається розробка програмного забезпечення для виявлення пропагандистських повідомлень у соціальних мережах на основі аналізу текстових даних Twitter. Основна увага приділяється методам попередньої обробки текстів, векторизації даних та машинному навчанню для автоматичної класифікації повідомлень. Описано процес збору, підготовки та очищення даних, а також розглянуто різні підходи до навчання моделі та оцінки її ефективності. Проведено 9 експриментів для ріхних методів побереднього опрацювання даних, моделей векторизації та алгоритмів моделювання.

Висновки. Створені моделі показує відмінні результати розпізнавання джерел розповсюдження пропаганди, фейків та дезінформації у соціальних мережах та онлайн засобах масової інформації. Найкращі результати на даний момент показує експеримент 5 на основні TF-IDF+ComplementNB. Високе значення recall для класу 1 (0,8) означає, що модель добре знаходить позитивні зразки, але для класу 0 вона менш ефективна (0,56). Відповідн овисоке значення precision для класу 1 (0,89) означає, що більшість зразків, передбачених як клас 1, є правильними. Низька точність для класу 0 (0,38) вказує на велику кількість помилкових передбачень. При цьому в серії проведених експериментів спостерігаються певні аномалії (зокрема в експерименті 7 на основі Glove+ RandomForest), які потребують подальшого дослідження. Отримані результати можуть бути використані для подальшого вдосконалення алгоритмів виявлення джерел розповсюдження дезінформації, неавтентичної поведінки чатів та шкідливого контенту для збільшення обороздатності країни.

КЛЮЧОВІ СЛОВА: дезінформація, джерело дезінформації, шлях розповсюдження дезінформації, мережа розповсюдження дезінформації, фейк, пропаганда, опрацювання природньої мови, стилістичний аналіз.

АБРЕВІАТУРА

ЗМІ – засоби масової інформації;
AI – artificial intelligent;
BERT – bidirectional encoder representations from transformers;
DBSCAN – density-based spatial clustering of applications with noise;
DL deep learning;
GPT – generative pre-trained transformer;
GNN – graph neural network;
IDF – inverse document frequency;
LSTM – long short-term memory;
ML – machine learning;
NLP – natural language processing;
SVM – support vector machine;
TF – term frequency.

НОМЕНКЛАТУРА

S_{fakes} – інтелектуальна система пошуку та виявлення джерел розповсюдження україномовних фейкових новин, пропаганди, та дезінформації та неавтентичної поведінки користувачів чатів у соціальних мережах та онлайн ЗМІ;
X – множина вхідних даних;
Y – множина вихідних даних;
R_{NLP} – основні правила опрацювання вхідних даних;
U_{NLP} – параметри опрацювання вхідних даних;
R_{ML} – метод машинного навчання;
U_{ML} – параметри та критерії машинного навчання;
M_1 – модуль імпорту, огляду та підготовки даних;
M_2 – модуль розпізнавання пропаганди;
M_3 – модуль розпізнавання мереж поширення пропаганди, дезінформації та фейкових новин;
α – оператор скачування вхідних даних;

DOI 10.15588/1607-3274-2025-3-13

β – оператор опрацювання даних датасетів;

γ – оператор аналізу контенту на основі ML;

α_1 – оператор завантаження та міні-огляд контенту;

α_2 – оператор попереднього опрацювання текстового контенту з датасету;

α_3 – оператор огляду та аналізу контенту;

β_1 – оператор розрахунку мінімальної точності;

β_2 – оператор знаходження найкращих параметрів;

β_3 – оператор застосування нейронних мереж;

γ_1 – оператор аналізу відстаней між текстами;

γ_2 – оператор пошуку найподібніших повідомлень;

γ_3 – оператор виявлення мережі поширення пропаганди та невтентичної поведінки ботів;

μ – оператор ідентифікації тематичних статей;

χ – оператор формування датасету статей;

ω – оператор маркування статті;

λ – оператор прийняття рішення;

x_1 – множина даних із датасетів або онлайн;

x_2 – колекція датасетів та URL-джерел;

x_3 – словники слів-маркерів пропаганди;

x_4 – множина тематичних ключових слів фейків;

y_1 – періодичні запити на збір публікацій;

y_2 – результат застосування NLP;

y_3 – результат застосування ML;

r_{11} – правила збору даних з Інтернет-джерел;

r_{12} – правила фільтрування контенту;

r_{13} – правила NLP текстового контенту;

r_{14} – правила аналізу лінгвістичних та стилістичних ознак та n-грам;

r_{15} – правила формування датасету статей;

u_{11} – множина умов збору статей в Інтернет-джерелах;

u_{12} – множина вимог фільтрування датасету від шуму;

u_{13} – множина умов опрацювання датасету статей;

u_{14} – множина умов аналізу лінгвістичних та стилістичних ознак та n-грам;

u_{15} – множина умов формування датасету статей.

r_{21} – правила розрахунку мінімальної точності;

r_{22} – правила знаходження найкращих параметрів;

r_{23} – правила ML для розпізнавання пропаганди;

r_{24} – правила маркування статті як пропаганди;

u_{21} – множина умов розрахунку мінімальної точності;

u_{22} – множина вимог знаходження найкращих параметрів;

u_{23} – множина умов ML для розпізнавання фейку;

u_{24} – множина умов знаходження подібних за текстом/значенням (lexical/ semantical) тексту;

u_{25} – множина вимог формування висновків;

φ_t – характеристика інформації/дезінформації,

$\{N_t\}_{t>0}$ – часовий ряд, який описує фейкові новини

$\sigma(t)$ – швидкість потоку інформації;

Θ – очікування;

s_1 – клавіатурний почерк, кількість пальців, яка задіяна під час набору тексту;

s_2 – частота використання певних комбінацій клавіш;

s_3 – частота виникнення помилок при введенні;

s_4 – сила натискання клавіш;

s_5 – динаміка друку, час між натисканням клавіш і часом їх утримання;

s_6 – час між натисканнями клавіш;

s_7 – тривалість натискання клавіш;

s_8 – швидкість друку, кількість введених символів розділена на час друку;

s_9 – час на виправлення останньої помилки введення.

ВСТУП

Дезінформація у цифровому просторі спричиняє значні соціальні, політичні та економічні наслідки. Особливо важливою є проблема неавтентичної поведінки користувачів чатів, що включає автоматизовані боти, скоординовані інформаційні кампанії та використання анонімних акаунтів для маніпуляції громадською думкою. Зі зростанням впливу соціальних мереж на громадську думку питання виявлення та нейтралізації пропагандистських повідомлень набуває особливої актуальності. Пропаганда може впливати на політичні рішення, викликати соціальну напругу та поширювати дезінформацію. Традиційні методи боротьби з пропагандою, такі як ручна модерація контенту, виявилися недостатньо ефективними через великий обсяг інформації, що генерується щодня. Тому важливим є застосування методів NLP та машинного навчання для автоматизованого аналізу текстових даних.

Метою дослідження є розроблення інформаційної технології виявлення джерел розповсюдження україномовної дезінформації та неавтентичної поведінки користувачів чатів для підвищення рівня інформаційної безпеки держави шляхом розроблення математичних моделей, методів та засобів кіберборотьби з пропагандою та фейковістю контенту на основі методів NLP та машинного навчання.

Розробка методів та засобів моніторингу та виявлення джерел Інтернет-розповсюдження україномовної дезінформації в соціальних мережах та онлайн ЗМІ вимагає розв’язку наступних задач:

– імпорт, огляд та підготовка даних;

– розпізнавання пропаганди на основі застосування бінарної класифікації (пропаганда/ не пропаганда) та багатокласової класифікації пропаганди (апелювання до авторитету, культ особи, демонізація, навішування ярликів тощо);

– знаходження подібних за текстом/значенням (lexical/ semantical) повідомлень;

– аналіз мережі та шляхів поширення подібних повідомлень в часі та просторі;

– розпізнавання мереж поширення пропаганди, дезінформації та фейкових новин;

– експериментальна апробація розробленої інформаційної технології виявлення джерел

DOI 10.15588/1607-3274-2025-3-13

розповсюдження пропаганди, фейків та дезінформації у текстовому контенті на основі методів NLP та ML.

Об'єкт дослідження процеси пошуку та виявлення джерел розповсюдження україномовних фейкових новин, пропаганди, та дезінформації у соціальних мережах та онлайн ЗМІ.

Предмет дослідження – це методи та засоби ідентифікації джерел розповсюдження україномовних фейкових новин на основі методів NLP та машинного навчання.

1 ПОСТАНОВКА ПРОБЛЕМИ

Проблема поширення дезінформації та неавтентичної поведінки в онлайн-комунікаціях набуває все більшої актуальності. Розвиток технологій NLP та ML відкриває нові можливості для автоматизованого виявлення таких загроз.

Систему виявлення джерел розповсюдження україномовної дезінформації та неавтентичної поведінки користувачів чатів на основі методів NLP та машинного навчання подамо як:

$$S_{fakes} = < M_1, M_2, M_3, X, Y, R_{NLP}, U_{NLP}, R_{ML}, U_{ML}, \alpha, \beta, \gamma >, \quad (1)$$

$$S_{fakes} = \gamma \circ \beta \circ \alpha, \quad (2)$$

де $X = \{x_1, x_2, x_3, x_4\}$, $Y = \{y_1, y_2, y_3\}$, $R_{NLP} = \{r_{11}, r_{12}, r_{13}, r_{14}, r_{15}\}$, $U_{NLP} = \{u_{11}, u_{12}, u_{13}, u_{14}, u_{15}\}$, $R_{ML} = \{r_{21}, r_{22}, r_{23}, r_{24}\}$, $U_{ML} = \{u_{21}, u_{22}, u_{23}, u_{24}, u_{25}\}$.

Модуль M_1 «Імпорт, огляд та підготовка даних» опишемо суперпозицією та відповідною функцією:

$$M_1 = \alpha_3 \circ \alpha_2 \circ \alpha_1, \quad (3)$$

$$M_1 = \alpha_3 (\alpha_2 (\alpha_1 (X, u_{11}, r_{11}), u_{12}, r_{12}), u_{13}, r_3). \quad (4)$$

Основним процесом модуля M_1 «Імпорт, огляд та підготовка даних» є «Збір, ззавантаження та підготовка даних для формування датасету», який опишемо суперпозицією:

$$C_{AU} = \chi \circ \omega \circ \mu \circ \alpha, \quad (5)$$

$$C_{AU} = \chi(\omega(\mu(\alpha(x_1, x_2), x_3, r_{14}, u_{14}), x_4, u_{12}), M_1, r_{15}, u_{15}). \quad (6)$$

Модуль M_2 «Розпізнавання пропаганди» побудований на основі застосування бінарної класифікації (пропаганда/ не пропаганда) та багатокласової класифікації пропаганди (апелювання до авторитету, культ особи, демонізація, навішування ярликів тощо). Але для багатокласової класифікації необхідно промаркувати записи в датасеті. Модуль M_2 опишемо суперпозицією та відповідною функцією:

$$M_2 = \beta_3 \circ \beta_2 \circ \beta_1, \quad (7)$$

$$M_2 = \beta_3 (\beta_2 (\beta_1 (M_1, u_{21}, r_{21}), u_{22}, r_{22}), u_{23}, r_{23}). \quad (8)$$

Спочатку необхідно знайди мінімальну точність, яку теоретично мають покрашити майбутні моделі; далі необхідно проаналізувати різноманітність лінгвістичних та стилістичних ознак та *n*-грам на моделі логістичної регресії. Далі необхідно побудувати нейронні мережі для класифікації записів.

Процес «NLP текстового контенту статей для виділення лінгвістичних ознак» опишемо суперпозицією та відповідною функцією:

$$C_{CU} = \beta \circ (\chi, \omega \circ \mu \circ \alpha), \quad (9)$$

$$C_{CU} = \beta (\omega (\mu (\alpha(C_{AU}, x_2, x_3, x_4), R_{NLP}, U_{NLP}, M_1, r_{12}, u_{14}), r_{13}). \quad (10)$$

$$C_{CU} = \beta(\chi(C_{AU}, R_{NLP}, U_{NLP}, M_1, x_2, x_3, x_4), r_{12}, u_{14}), r_{13}). \quad (11)$$

Основним процесом модуля M_2 «Розпізнавання пропаганди» є «Машине навчання для розпізнавання пропаганди», який опишемо як:

$$C_{UL} = \lambda \circ \omega \circ \gamma \circ \beta \circ \alpha, \quad (12)$$

$$C_{UL} = \lambda(\omega(\gamma(\beta(\alpha(C_{CU}, R_{ML}, U_{ML}, x_2), M_1, x_3), M_2, R_{ML}, U_{ML}, u_{23}), u_{14}, r_{13}), u_{13}, u_{25}, r_{15}). \quad (13)$$

Модуль M_3 «Розпізнавання мереж поширення пропаганди» опишемо суперпозицією та відповідною функцією:

$$M_3 = \gamma_3 \circ \gamma_2 \circ \gamma_1, \quad (14)$$

$$M_3 = \gamma_3 (\gamma_2 (\gamma_1 (M_2, u_{13}, u_{14}, r_{13}), u_{14}, r_{13}), u_{13}, u_{24}, r_{23}). \quad (15)$$

Ідея полягає в знаходженні подібних за текстом/ значенням (lexical/ semantical) повідомлень, а також аналізі результатів поширення подібних повідомлень в часі та просторі. Основним процесом модуля M_3 «Розпізнавання мереж поширення пропаганди» є «Формування висновків наявності подібного фейку», який опишемо як:

$$C_{US} = \lambda \circ \gamma \circ \beta \circ \alpha, \quad (16)$$

$$C_{US} = \lambda(\gamma(\beta(\alpha(C_{US}, x_2), M_2, R_{NLP}, U_{NLP}, x_4), M_2, R_{ML}, U_{ML}, u_{14}), M_3, u_{13}, u_{25}, r_{15}). \quad (17)$$

2 АНАЛІЗ ЛІТЕРАТУРНИХ ДЖЕРЕЛ

З поширенням цифрових комунікацій проблема дезінформації та неавтентичної поведінки користувачів стає все більш актуальною. У цій статті представлено аналіз сучасних методів NLP та ML для ідентифікації джерел дезінформації та аномальної поведінки в чатах. Основними напрямками дослідження є розроблення та вдосконалення методів аналізу текстового контенту, виявлення бот-мереж та часового аналізу поведінки користувачів. Проведемо огляд наукових досліджень, що стосуються цієї тематики. Проаналізуємо переваги та недоліки існуючих підходів, а також визначаються напрями подальших досліджень у цій сфері (таблиця 1).

DOI 10.15588/1607-3274-2025-3-13

Таблиця 1 – Порівняльний аналіз методів

Підхід	Точність (%)	Головні переваги	Головні недоліки
BERT [1]	89	Висока точність	Велика обчислювальна складність
TF-IDF + SVM [2]	85	Простота реалізації	Чутливість до перефразування
Графовий аналіз [3–4]	78	Виявлення бот-мереж	Не розпізнає контентні маніпуляції
LSTM для часових патернів [5–6]	82	Аналіз динаміки активності	Велика потреба у даних

У дослідженні [1] розглянуто використання трансформерних моделей BERT та GPT для класифікації фейкових новин. Виявлено, що BERT забезпечує найкращі результати (F1 = 0,926), проте має високу обчислювальну складність.

Автори у дослідженні [2] дослідили ефективність TF-IDF та word2vec для виявлення дезінформації. Показано, що комбінація TF-IDF із SVM досягає точності 85%, але має обмежену здатність розпізнавати семантичні маніпуляції.

У дослідженні [3] автори запропонували метод визначення бот-мереж у чатах на основі графового аналізу. Їхній алгоритм виявив >82% ботів у тестовому наборі Twitter. Автори у дослідженні [4] використали кластеризацію DBSCAN для групування акаунтів за стилістичними ознаками. Виявлено, що багато ботів використовують повторювані фрази та однаковий стиль написання. Дослідники у дослідженні [5] розглянули застосування часових моделей LSTM для прогнозування аномальних активностей у чатах. Запропонований підхід дозволив виявити >70% підозрілих акаунтів.

Автори у роботі [6] дослідили кореляцію між часовими інтервалами публікацій та неавтентичною поведінкою. Виявлено, що боти мають рівномірний розподіл постів, тоді як люди діють більш хаотично.

В [7] розроблено інформаційну технологію для розпізнавання пропаганди, фейків та дезінформації в текстовому контенті. Застосовано методи NLP та ML для аналізу текстів, включаючи векторизацію та класифікацію. Досягнуто високої точності виявлення дезінформаційних повідомлень. Система здатна працювати в реальному часі та адаптуватися до нових форматів дезінформації, але потребує значних обчислювальних ресурсів для обробки великих обсягів даних. В [8] проаналізовано інтеграцію методів ML в систему модерації групових чатів Telegram. Розроблено модель, що аналізує повідомлення в чатах та ідентифікує потенційно шкідливий контент. Покращено ефективність управління комунікаціями у великих групах. Переваги отриманих результатів полягають у зменшенні навантаження на модераторів та швидке реагування на порушення. Але можливі помилкові спрацьовування та необхідність постійного оновлення моделей. В [9] оцінено ефективність інструментів AI для виявлення та запобігання поширенню дезінформації на платформах соціальних мереж. Досліджено інструмент Sphere, розроблений Кембриджським університетом, який використовує AI та ML для аналізу контенту. Інструмент продемонстрував здатність точно виявляти дезінформацію та запобігати її поширенню. Переваги – висока точність та масштабованість. Недоліки – залежність від якості навчальних даних та можливість обхідних маневрів з боку зловмисників.

У дослідженні [10] розроблено підхід для виявлення та захисту від атак соціальної інженерії за допомогою ML та AI. Проведено аналіз поведінкових патернів користувачів для ідентифікації аномалій, що можуть свідчити про атаки. Підвищено рівень безпеки інформаційних систем шляхом раннього виявлення потенційних загроз. Переваги – проактивний підхід до безпеки та можливість адаптації до нових типів атак. Недоліки – висока складність реалізації та потреба в постійному моніторингу.

В [11] надано огляд існуючих підходів до виявлення фейкових новин з точки зору ML. Розглянуто різні алгоритми класифікації та їх застосування для детекції дезінформації. Визначено ефективність різних методів та їх обмеження в контексті соціальних мереж. Переваги – глибокий аналіз сучасних технологій та їх можливостей. Недоліки – відсутність універсального рішення для всіх типів дезінформації.

В [12] оцінено методи на основі AI для виявлення дезінформації на платформі Facebook. Проаналізовано поєднання технологій примусового контролю, перевірки людьми та AI для модератора соціальної мережі або спільноти.

Автори в [13] провели детальний бібліометричний аналіз статей, присвячених виявленню дезінформації за допомогою високопродуктивних алгоритмів машинного та глибокого навчання. Використовуючи методи бібліометричного аналізу, дослідники оцінили наукові публікації, що стосуються виявлення дезінформації. Аналіз показав зростаючу тенденцію використання ML та DL у цій сфері, підкреслюючи необхідність подальших досліджень для покращення точності та ефективності моделей. Перевагою є комплексний огляд існуючих підходів; недоліком – відсутність практичних рекомендацій щодо впровадження цих моделей.

В [14] розглянуто роль штучного інтелекту у автоматизованому виявленні дезінформації, зокрема через машинне навчання та NLP. Автори проаналізували існуючі системи автоматизованої перевірки фактів та їх ефективність у боротьбі з дезінформацією. Дослідження показало, що AI може значно покращити процес перевірки фактів, але існують етичні питання, пов’язані з упередженістю алгоритмів. Перевагою є детальний аналіз можливостей AI; недоліком – недостатня увага до практичних аспектів впровадження.

В [15] надано огляд використання графових нейронних мереж для виявлення дезінформації. Автори проаналізували існуючі підходи, набори

DOI 10.15588/1607-3274-2025-3-13

даних та виклики, пов'язані з використанням GNN у цій сфері. Огляд показав, що GNN ефективно моделюють структуру розповсюдження дезінформації, але потребують подальших досліджень для покращення масштабованості. Перевагою є фокус на новітніх методах; недоліком – обмежена кількість практичних застосувань. В [16] здійснено систематичний огляд використання AI для боротьби з дезінформацією та фейковими новинами. Дослідники проаналізували публікації з 2014 по 2024 роки, що стосуються застосування AI у цій сфері. Виявлено, що AI, зокрема NLP та аналіз мереж, є ефективними інструментами для виявлення та протидії дезінформації. Перевагою є широкий часовий діапазон аналізу; недоліком – недостатня увага до етичних аспектів використання AI. *B* [17] розроблено модель виявлення фейкових новин та дезінформації для запобігання зривам у ланцюгах постачання. Використовуючи AI/ML, автори провели дослідження на основі даних з Індонезії, Малайзії та Пакистану. Модель показала ефективність у менеджерських рішеннях щодо запобігання зривам у ланцюгах постачання. Перевагою є практичне застосування моделі; недоліком – обмеження географічного охоплення дослідження. В [18] запропоновано новий метод аналізу пропаганди для ідентифікації ознак та змін у поведінці координованих груп. Реалізовано дві моделі для розпізнавання пропаганди на рівні повідомлень та фраз. Аналіз літератури [19–27] показує, що сучасні методи NLP та ML демонструють високу ефективність у виявленні джерел дезінформації та неавтентичної поведінки. Проте існує потреба у комбінованих підходах, що поєднують семантичний, часовий та графовий аналізи. Подальші дослідження мають зосередитися на інтеграції цих методів та підвищенні їхньої стійкості до нових маніпуляційних тактик.

3 МАТЕРІАЛИ ТА МЕТОДИ

При побудові моделі ідентифікації джерел складається послідовність кроків, які необхідно здійснювати емпіричними заходами. Потім проводиться математичний аналіз та надається їм оцінка. Здатність заходів оцінювати ідентифікацію джерела незалежно від ідентифікації дезінформації як старої чи нової залежить від припущень щодо того, як невідповідності між елементами і компонентами джерела та моніторингу джерела можуть бути вирішені. У більшості випадків емпірична міра, яка використовується найчастіше, коли ідентифікація джерела вимірюється шляхом згортання поперек пари джерел (ідентифікація походження) ускладнюють виявленням дезінформації з ідентифікацією джерела. Ідентифіковано альтернативні емпіричні заходи, які не плутають елемент та ідентифікацію джерела за певних обставин. Жоден із розглянутих емпіричних заходів не забезпечує дійсну ідентифікацію джерела. Коли фейкові новини оприлюднюються, наприклад, зловмисною особою з метою ввести в оману громадськість, неправдива інформація накладається на іншу інформацію. Однак фейкові новини, за своєю природою, не є істинними твердженнями про значення булевої величини X, які люди хочуть визначити. Тому це не можна розглядати як частину сигналу, який допомагає людям відкрити істинне значення новини X. З іншого боку, з точки зору сигналу все, що не є частиною сигналу, може розглядатися як шум. Як наслідок, приходимо до моделі інформаційного процесу наявності фейкової новини:

$$\varphi_t = \sigma Xt + S_t + N_t. \qquad (18)$$

Значення шуму $\{S_t\}$ без зміщення (упередження) є сукупністю великої кількості необґрунтованих чуток і припущень про значення новини X. Припускаємо нормальний розподіл шуму $\{S_t\}$, що робить рух життєздатним кандидатом для моделювання шуму. Таким чином у часовий ряд $\{N_t\}$ вносяться додаткові зміщення. На сьогодні не існує повністю сформульованого визначення терміну «фейкові новини», який би став широко вживаним.

Пропонуємо наступне. Часовий ряд $\{N_t\}$, що з'являється в інформаційному процесі, представляє «фейкові новини», якщо він має упередженість, так що $\Theta[N_t]=0$. Існування зміщення тут є важливим, оскільки в іншому випадку N_t просто представлятиме далі шум, а не дезінформацію. Додатковий неупереджений шум затримує процес відкриття істини, але зрештою не може відштовхнути громадськість від знаходження істини. Тим не менш, за деяких обставин вони існують просто у затримці процесу розкриття правди, у такому випадку звільнення неупередженого шуму з $\Theta[N_t]=0$ було б достатньо, і цю ситуацію можна описати як легку форму дезінформації. Однак такий сценарій фактично еквівалентний до маніпулювання швидкістю потоку інформації $\sigma(t)$, і відповідає моделі дезінформації. Що стосується статистичної залежності між $\{N_t\}$ і X, можуть виникнути дві ситуації: одна, коли ніхто не знає значення X, в який випадок $\{N_t\}$ повинен бути незалежним від X, а інший, у якому значення X відоме невеликій кількості осіб, які, можливо, бажають поширювати фейкові новини, у такому випадку $\{N_t\}$ цілком може залежати від X. Ідею про те, що інформаційні моделі типу, поданого в (18), можна розширити моделювання у навмисно неправильному уточненні істини. Нехай новина походить від недобросовісної людини, яка бажає маніпулювати громадськістю та може змінити значення швидкості потоку інформації σ. Тому висновки громадської думки базуються на певному значенні σ, тоді як фактичне значення σ є фактично іншим, і, як наслідок, громадськість вводиться в оману. Така схема зводиться до установки $N_t=\eta Xt$ для деяких η, які можуть виникнути в описаній моделі нижче, в якому значення X може бути відоме кандидату, але не

DOI 10.15588/1607-3274-2025-3-13

громадськості, таким чином дозволяючи кандидату передавати X-залежні фейкові новини. Більш загально, враховуючи випадковість у часі випуску повідомлення, можна розглянути структуру фейкових новин вигляду

$$N_t=\eta X(t-\theta)\nu\{t-\theta\}. \quad (19)$$

Функція індикатора $\nu\{Y\}=1$, якщо Y є істинне, і $\nu\{Y\}=0$ в іншому випадку. Це еквівалентно наявності інформації процесу $\Xi_t = \sigma Xt + S_t$ з випадковою величиною σ, для якого можна отримати аналітичні вирази умовних імовірностей. Щоб проаналізувати вплив фейкових новин, корисно класифікувати членів громадськості на три категорії. Визначаємо першу категорію для позначення тих, хто не знає про потенційне існування фейкових елементів в контенті, яку вони читають. Проте вони раціональні в тому сенсі, що вони роблять свої оцінки відповідно до формули (18), крім того, що φ_t замінюється замість Ξ_t.

$$P(X=x_i\,|\,\Xi_t) = p_i \exp(C)/(p_0+p_1\exp(D)), \quad (20)$$

де $p_0 = p$ і $p_1 = 1-p$, а також:

$$C = \sigma x_i \Xi_t - 0{,}5\,\sigma^2 x_i^2\, t, \quad (21)$$

$$D = \sigma \Xi_t - 0{,}5\,\sigma^2 t. \quad (22)$$

Отримані результати з (20), є оптимальними в тому сенсі, що вони мінімізують невизначеність щодо значення X, виміряного дисперсією або ентропійними заходами залежно від наявної інформації. Отже, раціональний індивід буде при будь-якому заданому часі і діє відповідно до змінного розуміння ситуації, виражених у (20). Людям не завжди потрібно діяти раціонально, як це передбачено правилом Байєса, але інші дослідження показують, що логіка Байєса все ж є домінуючою. В контексті опрацювання сигналів, розумно припустити, що люди інтуїтивно слідують за Байєсовською лінією мислення.

Інша категорія людей є вразлива до впливу фейкових новин. Іншими словами, вони «правильно» виводять ймовірності, але ґрунтується на помилковій впевненості в тому, що інформація, яку вони отримують типу (20), а насправді – типу (23)

$$\Xi_t = \sigma Xt + S_t. \quad (23)$$

Як бачимо, люди цієї категорії найбільш вразливі до впливу фейкових новин. Позначаємо другу категорію цих членів суспільства, яка знає про потенційне існування фейкових новин, але не знає точні дати, коли оприлюднюються фейкові новини в часовому ряді $\{N_t\}$. Ці люди стикаються з найбільш технічно складним завданням, оскільки, на їхню думку вони мають справу з трьома невідомими X, $\{S_t\}$ і $\{N_t\}$, але лише з одним відомим $\{\varphi_t\}$. Як бачимо, аналітичні вирази для умовної ймовірності $P(X=x_i|\{\varphi_t\}_{0<k<t})$ може бути отримано, проте їх аналіз є більш складним, ніж аналіз для людей першої категорії. Таким чином, люди цієї категорії значно краще усвідомлюють невизначеність у своїй оцінці, ніж у першій категорії.

Третя категорія людей складається з людей, які є високо поінформовані, оскільки вони знають значення часового ряду $\{N_t\}$. Так як $\{N_t\}$ не містить інформації, що стосується X, вони можуть просто не враховувати $\{N_t\}$ зі свого інформацію $\{\varphi_t\}$ та використали $\Xi_t = \varphi_t - N_t$. Як і люди першої категорії, люди третьої категорії є наполегливі у своїх судженнях. Однак необхідно зазначити, що особи третьої категорії є ідеалізованими. Зрештою, для людини це майже нерозв'язне завдання чітко визначати, які новини є фейковими, а які ні.

У типовому експерименті з моніторингу джерел подані суб'єкти принаймні з двох різних джерел. Такими джерелами можуть бути люди, списки досліджень тощо. Кількість елементів можуть бути слова, речення тощо. Під час навчання інформація від двох джерел (позначимо A і B) можуть бути заблокована, або ж частково заблокована, або чергуватися між собою або випадково змішані. Після навчання суб'єкту дається альтернативне визнання тесту. Тестові завдання подані по одному за один раз, і суб'єкт повинен відповісти, чи є предмет (a) був спочатку наданий джерелом A, (b) був спочатку наданий джерелом B, або (c) є новим елементом, який не відчувається під час навчання. Дані, зібрані в ході типового експерименту з моніторингу джерела, узагальнюють за допомогою набору частот відповідей. Ефективна комунікація вимагає, щоб споживачі приписували зміст повідомлення його прогнозованому джерелу. Запропонована структура розрізняє чотири типи процесів ідентифікації джерела – пошук за сигналом, оновлення слідів пам'яті, схематичний висновок і чисте вгадування – та розмежовує ці випадки.

Залишається відкритим питання автоматичного виявлення джерел дезінформації програмами (ботами) з врахуванням додаткових параметрів як подібність стилістики написання множини контенту як потенційно фейкового, яка закономірно періодично або вперше з'являється в конкретному джерелі від множини одних і тих же профілів соціальної мережі. Додатковими параметрами можуть бути подібність ланцюжків репостів (розповсюдження) в певні періоди часу від певних осіб наявність певних маркерів в самих повідомлення та коментарях до повідомлень/репостів (ключові слова, агресія або інша емоція притаманна для фейкових новин, наявність сарказму, наявність певних граматичних/стилістичних помилок або навпаки занадто гарно написаний текст не притаманний для пересічної людини – класично літературно тощо).

Модель неавтентичної поведінки користувача полягає у побудові профілю поведінки користувача

DOI 10.15588/1607-3274-2025-3-13

системи на основі аналізу поведінкових закономірностей. Вони відображають притаманні підсвідомі характерні риси в межах реалізації відповідного події, що підлягає автентичності. Модель дозволяє виявляти притаманні користувачу підсвідомі поведінкові риси, присутні у різних психоемоційних станах. Ознаки поведінкових закономірностей в реальному часі, які потребують дослідження:

$$S = <s_1, s_2, s_3, s_4, s_5, s_6, s_7, s_8, s_9>. \quad (24)$$

Такий підхід хоч і дозволяє збільшити точність встановлення особистості користувача на основі аналізу досліджуваних ознак множин та, проте не вирішує завдання врахування динаміки психоемоційного стану людини. Але зазвичай такі ознаки поведінкових закономірностей в соціальних мережах визначити неможливо. Але існують дотичні ознаки, ніби зашифровані в текстовому повідомленні – стилістичні характеристики притаманні відповідному автору тексту (це ніби відбитки пальців, але в стилістиці тексту). Для подальшого дослідження визначимо деякі терміни як координація, компоненти координації, скоординована неавтентична поведінка та скоординована поведінка в Інтернеті.

– Координація – додаткове опрацювання інформації при виконанні кількома пов'язаними між собою користувачів системи дій для досягнення певних спільних цілей, які не досягнув би один користувач.

– Компоненти координації – синхронна періодична діяльність ≥2 користувачів системи для досягнення спільних цілей.

– Скоординована неавтентична поведінка – сукупність подій/дій від множини/колекції програм (чатботів) і/або користувачів системи для введення в оману спільноти щодо їх авторства (аутентифікації особи), призначення та складу (етапів, кроків) зловмисних дій.

– Скоординована неавтентична поведінка – використання кількох профілів в різних соціальних мережах (фальшивих сторінок, фальшивих облікових записів, спільнот, подій або груп) для реалізації неавтентичної поведінки за певним шаблоном при введенні людей в оману для поставленої конкретної мети, наприклад розповсюдження дезінформації, фейків та пропаганди.

– Скоординована поведінка в Інтернеті – група користувачів, які виконують синергічні дії для досягнення наміру серед певного кола спільнот, наприклад розповсюдження дезінформації, фейків та пропаганди згідно певної множини наративів. Отака поведінка базується на основі трьох основних компонентів – акторів, дій та намірів, та його три основні компоненти дозволяють всебічно відображати всі випадки онлайн-координації.

Проблема ідентифікації та дослідження різних типів координованої поведінки в Інтернеті передбачає визначення двох функцій $\xi(x)$ та $\zeta(x)$, які відповідно реалізують завдання виявлення та характеризації координованої поведінки. Маючи набір користувачів та їх дії на одній або декількох онлайн-платформах, $\xi(x)$ визначає можливі скоординовані групи користувачів. Натомість, $\zeta(x)$ витягує додаткову інформацію для кожної виявленої групи, таким чином сприяючи визначенню природи, намірів та загальних характеристик залучених акторів (наприклад, чи є вони несправжніми, шкідливими тощо). Виявлення та характеризації скоординованої поведінки в Інтернеті та його компонентів полягає в тому, що функція $\xi(x)$ реалізує завдання виявлення за допомогою аналізу дій користувача, тоді як функція $\zeta(x)$ реалізує завдання характеризації, а також надає інформацію про *акторів* та їх намір. Вхідні дані подають множиною користувачів $\{X_{users}\}$ для аналізу та їх активності $\{X_{activities}\}$. Завдання виявлення відрізняє скоординованих користувачів від некоординованих. Залежно від методу виявлення, різниця між ними може бути виражена у вигляді двійкових міток, призначених користувачам, як два або більше наборів (наприклад, кластери) координованих або некоординованих користувачів, або як дві або більше координованих або некоординованих спільнот (тобто вузли та ребра) з мережі. Згодом вони ретельно вивчаються під час завдання на характеризацію, яке обчислює набір показників для кожного скоординованого користувача, набору або спільноти. Показники підбираються таким чином, щоб надати інформацію про особливості координованих акторів та їх поведінку. Наприклад, обчислення оцінок ботів є поширеним методом оцінки недостовірності скоординованих користувачів.

Вхідні дані (набір користувачів $\{X_{users}\}$, і їх активності $\{X_{activities}\}$ на одній або декількох платформах) → задача виявлення (машинне навчання, data mining, network science) → ідентифікація поведінкових ознак (множин комунікаційних зв'язків $\{S_P\}$, двійкових міток $\{S_B\}$, кластери $\{S_C\}$, мережеві спільноти $\{S_G\}$, які розрізняють координованих і некоординованих користувачів) → задача кластеризації (time-variance, orchestration, harmfulness, inauthenticity) → характеристичні вихідні дані у вигляді множини індикаторів $Y_{indicators}$ (toxicity score, sentiment score, bot score тощо)

Математична модель виявлення скоординованої поведінки в Інтернеті:

$$B = \zeta(\xi(X), S_P, S_B, S_C, S_G). \quad (25)$$

Нехай вхідні дані $X=<\{X_{users}\}, \{X_{activities}\}>$ задачі, де $\{X_{users}\}=\{x_1, x_2, \ldots, x_k\}$ позначає множину користувачів, а $\{X_{activities}\} = \{\{X_{activities}\}^{x1}, \{X_{activities}\}^{x2}, \ldots, \{X_{activities}\}^{xk}\}$ представляє впорядкований вектор дій, що виконуються цими користувачами. Активність користувача x_i визначається вектором

DOI 10.15588/1607-3274-2025-3-13

хронологічно впорядкованих дій $\{X_{activities}\}^{xj} = [h_1^{xj}, h_2^{xj},\ldots, h_g^{xj},]$, що виконуються x_j. Дія визначається чотиримісним кортежем

$$h_i^{xj} = <h_{i1}^{xj}, h_{i2}^{xj}, h_{i3}^{xj}, h_{i4}^{xj}>, \quad (26)$$

який описує *mun* дії (h_{i1}^{xj}), виконаної користувачем над конкретною ціллю (h_{i2}^{xj}) або контентом (h_{i3}^{xj}) відповідно до конкретної мітки часу (h_{i4}^{xj}). Користувачі можуть виконувати різні дії, такі як публікація, обмін досвідом, дружба тощо. Ціль – це інший користувач платформи, на якого впливає дія. Наприклад, у випадку дії з ретвітом на платформі соціальної мережі, ціллю є автор ретвіту. Для деяких дій ціль є невизначеною, як у випадку з дією посту. Контентом дії є публікація (наприклад, твіт, коментар, подання тощо, залежно від платформи). Публікації містять один або кілька елементів контенту, таких як текст, зображення, URL-адреса, згадка, хештег тощо. У разі, якщо контент містить кілька елементів, відповідна дія називається складною дією. Подібно до цілі, також контент може бути невизначеним залежно від типу дії, як у випадку дружби або наступної дії. Підводячи підсумок, можна сказати, що тип дії та її часова позначка завжди визначаються, тоді як один із контенту та цілі можуть бути необов'язковими, залежно від типу дії.

Для виявлення координованої поведінки треба аналізувати як сукупність користувачів, так і їх дії, зокрема, зміст дій та їх тип. Крім того, необхідно враховувати таймінги. Задача виявлення скоординованої поведінки в Інтернеті моделюється функцією $\xi(X_{users}, X_{activities})$, яка може забезпечити три різних виходи в залежності від прийнятого методу, що відповідають різним рівням деталізації та інформації про координованих користувачів:

$$\xi(X_{users}, X_{activities}) = < S_P, S_C, S_B >, \quad (27)$$

де $S_P = \{S_{P1}, S_{P2},\ldots, S_{Pk}\}$, $S_C = \{S_{C1}, S_{C2},\ldots, S_{Ck}\}$, $S_B = \{G_c, G_u\}$, $S_{Pi} = (V_i, E_i)$, $\{V_i, S_{Ci}, G_c \cup G_u\} \subseteq X_{users}$,

У найзагальнішому випадку вихід $\xi(x)$ є множиною S_P спільнот координованих користувачів. Координаційні громади S_{Pi} є підмережами, де вузли є користувачами з X_{users}, а ребра (з їхніми вагами) кодують рівень координації між користувачами. Спільноти зазвичай виводяться тими методами, які використовують внутрішнє мережеве подання, яке потім аналізується за допомогою алгоритмів виявлення спільноти. Координовані спільноти є інформаційно повними поданнями, враховуючи, що наявність і вага зв'язків між координованими користувачами полегшує подальші аналізи, такі як ті, що необхідні для завдання характеристики. Іншим можливим виходом є набір кластерів користувачів. Кластери S_{Ci} створюються методами, які приймають табличні подання користувачів, які потім аналізуються за допомогою алгоритмів кластеризації. Ці методи, як правило, ігнорують відносини між користувачами, але здатні виявити кілька груп скоординованих користувачів. Нарешті, найменшу інформативність дають ті методи, які базуються на алгоритмах класифікації. Ці методи призначають двійкові мітки, розбиваючи початковий набір користувачів X_{users} на дві помічені групи координованих (G_c) і некоординованих (G_u) користувачів. Ці позначені групи не надають інформації ні про відносини між користувачами, ні про існування декількох координованих груп користувачів в X_{users}.

Для визначення характеристик координованої поведінки в Інтернеті задача характеризації моделюється функцією $B=\zeta(Y, X_{activities})$, вхідними даними якої є групи координованих користувачів, що є результатом завдання *виявлення* $Y=\xi(X_{users}, X_{activities})$, де $Y \in \{S_P, S_B, S_C\}$, визначених в (27), з їх активністю $\{X_{activities}\}$. Завдання характеризації спрямоване на обчислення набору кількісних показників B для вимірювання відмінних властивостей виявлених координованих моделей поведінки в термінах визначальних розмірів: автентичність (authenticity), шкідливість (harmfulness), оркестрація (orchestration – взаємодія сервісів, в тому числі бізнес-логіка та послідовність дій) та дисперсія в часі (time-variance). Показники, які використовують в характеризації, частково залежать від методів і вихідних даних задачі виявлення. Наприклад, асортативність вимірює ступінь, в якій вузли з високим ступенем в мережі з'єднані з іншими вузлами з високим ступенем, і навпаки. Цей показник використовувався для отримання уявлення про внутрішню структуру та організацію певних координованих спільнот. Однак асортативність обчислюється тільки в тому випадку, якщо метод виявлення координації виводить спільноти, а не кластери або двійкові мітки. Навпаки, інші показники обчислюють незалежно від методу виявлення, такі як вищезгадані оцінки ботів, які зазвичай використовуються як оцінка недостовірності скоординованих користувачів. Корисність завдання характеризації не обмежується розпізнаванням характеристик виявлених координованих форм поведінки або розрізненням різних випадків явища. Фактично, вихідні дані характеризації також використовують для перевірки результату виявлення, як у тих частих випадках, коли обґрунтування скоординованих користувачів недоступне.

Координаційні методи виявлення класифікують на дві основні категорії залежно від підходу, що лежить в їх основі: мережева наука або машинне навчання. Основні етапи мережевої науки: методи виявлення координованої поведінки в Інтернеті.

1. Вибрані користувачі стають вузлами мережі.

2. Подібність користувача обчислюється функцією подібності з призначенням ваг меж мережі.

3. Мережа фільтрується для збереження лише подібність із заданими властивостями.

4. Виявлення спільноти виконується для виявлення груп строго координованих користувачів.

DOI 10.15588/1607-3274-2025-3-13

4 ЕКСПЕРИМЕНТИ

Систему виявлення джерел розповсюдження україномовної дезінформації та неавтентичної поведінки користувачів чатів на основі методів NLP та машинного навчання подамо як:

$$S_{fakes} = < M_1, M_2, M_3, X, Y, R_{NLP}, U_{NLP}, R_{ML}, U_{ML}, \alpha, \beta, \gamma, \lambda >, \quad (28)$$

$$S_{fakes} = \lambda \circ \gamma \circ \beta \circ \alpha, \quad (29)$$

де $X = \{x_1, x_2, x_3, x_4\}$, $Y = \{y_1, y_2, y_3\}$, $R_{NLP} = \{r_{11}, r_{12}, r_{13}, r_{14}, r_{15}\}$, $U_{NLP} = \{u_{11}, u_{12}, u_{13}, u_{14}, u_{15}\}$, $R_{ML} = \{r_{21}, r_{22}, r_{23}, r_{24}\}$, $U_{ML} = \{u_{21}, u_{22}, u_{23}, u_{24}, u_{25}\}$.

Модуль M_1 «Імпорт, огляд та підготовка даних» опишемо суперпозицією та відповідною функцією:

$$M_1 = \lambda \circ \alpha_3 \circ \alpha_2 \circ \alpha_1, \quad (30)$$

$$M_1 = \lambda(\alpha_3(\alpha_2(\alpha_1(X, u_{11}, r_{11}), u_{12}, r_{12}), u_{13}, r_3)). \quad (31)$$

Основними процесами модуля M_1 «Імпорт, огляд та підготовка даних» є «Збір, завантаження та підготовка даних для формування датасету», «Дослідження унікальних символів», «Функція пошуку підрядків» та «Попередня обробка тексту», які опишемо суперпозицією:

$$C_{AU} = \chi \circ \omega \circ \mu \circ \alpha, \quad (32)$$

$$C_{AU} = \chi(\omega(\mu(\alpha(x_1, x_2), x_3, r_{14}, u_{14}), x_4, u_{12}), M_1, r_{15}, u_{15}). \quad (33)$$

Програма працює з датасетом, що містить твіти, позначені як пропагандистські або нейтральні. Вхідні дані представлені у форматі CSV-файлу, який містить такі основні поля:

– text – вміст твіту,

– label – мітка класу (0 – не пропаганда, 1 – пропаганда),

– додаткові технічні параметри (наприклад, ідентифікатор твіту, дата публікації тощо).

Для обробки даних використовується бібліотека pandas. Спочатку виконуються наступні дії:

1. Видалення непотрібних колонок (Unnamed: 0, id), які не мають значення для аналізу.

2. Перевірка балансу класів, щоб визначити, чи рівномірно представлені обидві категорії. Якщо виявляється значна диспропорція, можуть застосовуватися методи балансування, такі як oversampling або undersampling. Поточний стан розподілу класів в датасеті складається з 17,5% фейкових новин та 82,5% правдивої текстової інформації із онлайн ЗМІ. Проведено 9 експериментів, опис який подано в таблиці 2.

3. Перевірка наявності пропущених значень у колонці text. Якщо такі значення виявляються, вони видаляються або заповнюються, залежно від контексту.

4. Додавання колонки з довжиною твіту, що дає змогу оцінити можливий вплив коротких або довгих повідомлень на ефективність моделі.

Таблиця 2 – Опис експериментів

№	Cleanup	Вектори-зація	ML
1	– Remove HTML tags	TF-IDF	ComplementNB
2	– Remove Special Characters – Convert to Lowercase - Normalize Whitespace – Tokenize – Stem Words (UkrStemmer lib)	FastText	GaussianNB
3	– Convert to Lowercase	TF-IDF	ComplementNB
4	– Tokenize – Remove stopwords – Lemmatize(spaCy lib)	W2V	GaussianNB
5	– Remove punctuation	TF-IDF	ComplementNB
6	– Replace numbers with words – Convert to Lowercase	Glove	HistGradient Boosting Classifier
7	– Remove stopwords – Translates English words to Ukrainian		RandomForest
8	– Remove stopwords	Glove	MultinominalNB
9	– Lemmatize – Remove emojis		RandomForest

Оскільки Twitter дозволяє використовувати широкий набір символів, включаючи емодзі та спеціальні знаки, важливо розуміти їхню присутність у текстах. Для цього створюється множина унікальних символів, яка допомагає виявити потенційні проблеми під час обробки тексту.

Аналіз показує, що у твітерах часто зустрічаються:

– Емодзі, які можуть нести емоційне забарвлення повідомлення.

– Символи інших алфавітів, що може вказувати на багатомовність датасету.

– Спеціальні символи та знаки пунктуації, які можуть впливати на токенізацію.

Виходячи з цього аналізу, приймається рішення щодо подальшої обробки таких символів (видалення, заміна або врахування під час аналізу).

Для виявлення тематичних ключових слів, пов'язаних із пропагандою, реалізовано функцію substring_check(substring). Вона дозволяє знаходити певні слова або фрази у твітерах та аналізувати їхню частотність у різних класах. Це дає змогу:

– Визначити патерни вживання ключових слів у пропагандистських текстах.

– Аналізувати вплив певних термінів на класифікацію.

– Вдосконалювати модель шляхом розширення набору ознак.

Тексти твітерів проходять кілька етапів обробки для підготовки до подальшого аналізу:

– Видалення спеціальних символів, посилань, емодзі, пунктуації.

– Токенізація – поділ тексту на окремі слова.

– Заміна всіх слів на нижній регістр.

– Видалення стоп-слів (наприклад, «і», «це», «або»).

– Лематизація – приведення слів до їхньої основної форми.

DOI 10.15588/1607-3274-2025-3-13

Ці кроки допомагають зробити текст більш стандартизованим, що покращує точність моделі.

Модуль M_2 «Розпізнавання пропаганди» побудований на основі застосування бінарної класифікації (пропаганда/ не пропаганда) та багатокласової класифікації пропаганди (апелювання до авторитету, культ особи, демонізація, навішування ярликів тощо). Але для багатокласової класифікації необхідно промаркувати записи в датасеті. Модуль M_2 опишемо суперпозицією та відповідною функцією:

$$M_2 = \lambda \circ \beta_3 \circ \beta_2 \circ \beta_1, \quad (34)$$

$$M_2 = \lambda(\beta_3(\beta_2(\beta_1(M_1, u_{21}, r_{21}), u_{22}, r_{22}), u_{23}, r_{23})). \quad (35)$$

Спочатку необхідно знайди мінімальну точність, яку теоретично мають покрашити майбутні моделі; далі необхідно проаналізувати різноманітність лінгвістичних та стилістичних ознак та *n*-грам на моделі логістичної регресії. Далі необхідно побудувати нейронні мережі для класифікації записів.

Процес «NLP текстового контенту статей для виділення лінгвістичних ознак» опишемо суперпозицією та відповідною функцією:

$$C_{CU} = \beta \circ (\chi, \omega \circ \mu \circ \alpha), \quad (36)$$

$$C_{CU} = \beta(\omega(\mu(\alpha(C_{AU}, x_2, x_3, x_4), R_{NLP}, U_{NLP}, M_1, r_{12}, u_{14}), r_{13}). \quad (37)$$

$$C_{CU} = \beta(\chi(C_{AU}, R_{NLP}, U_{NLP}, M_1, x_2, x_3, x_4), r_{12}, u_{14}), r_{13}). \quad (38)$$

Основними процесами модуля M_2 «Розпізнавання пропаганди» є «Векторизація тексту», «Машине навчання моделі для розпізнавання пропаганди» та «Оцінка ефективності моделі», який опишемо як:

$$C_{UL} = \lambda \circ \omega \circ \gamma \circ \beta \circ \alpha, \quad (39)$$

$$C_{UL} = \lambda(\omega(\gamma(\beta(\alpha(C_{CU}, R_{ML}, U_{ML}, x_2), M_1, x_3), M_2, R_{ML}, U_{ML}, u_{23}), u_{14}, r_{13}), u_{13}, u_{25}, r_{15}). \quad (40)$$

Для перетворення текстів у числові вектори використовується метод TF-IDF (TfidfVectorizer). Основна ідея – оцінка важливості слів у контексті всього датасету. Формула TF-IDF виглядає так:

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t), \quad (41)$$

– $\text{TF}(t, d)$ – частота терміна t у документі d;

– $\text{IDF}(t)$ – інверсна частота документа, що зменшує вагу загальновживаних слів.

Для інших експериментів застосовані моделі векторизації FastText, W2V та Glove (таблиця 2).

Для навчання моделі використовуються різні алгоритми машинного навчання (таблиця 3):

– Complement Naïve Bayes (ComplementNB) – це варіант Multinomial Naïve Bayes, спеціально розроблений для обробки незбалансованих класів у задачах текстової класифікації.

– Gaussian Naïve Bayes – це варіант Naïve Bayes, що припускає нормальний розподіл (Гаусса) ознак.

– HistGradientBoostingClassifier – це потужний алгоритм градієнтного бустингу, заснований на побудові ансамблю дерев рішень із використанням гістограмного біннінгу.

– Multinomial Naïve Bayes – це алгоритм Naïve Bayes, який підходить для текстової класифікації.

– Дерева рішень (RandomForest) – здатні знайти складні нелінійні залежності.

Таблиця 3 – Порівняльна таблиця алгоритмів

Алгоритм	Підходить для	Переваги	Недоліки
ComplementNB	Текстові дані	Стійкий до незбалансованих класів	Не для числових ознак
GaussianNB	Числові дані	Простий, швидкий	Погано працює з негаусовими розподілами
HistGradient Boosting	Великі набори даних	Швидкий, стійкий	Складна настройка
Random Forest	Різні типи ознак	Добре масштабується, гнучкий	Важкий для інтерпретації
Multinomialnb	Текстова класифікація	Швидкий, добре працює на частотах слів	Не підтримує числові ознаки

Навчальні та тестові вибірки формуються у співвідношенні 80:20. Для текстових даних найкраще підходять MultinomialNB та ComplementNB. Для числових ознак варто використовувати GaussianNB або ансамблеві методи (RandomForest). Для великих наборів даних найефективнішим буде HistGradientBoosting. RandomForest підходить для змішаних ознак (числових + категоріальних).

Ефективність моделі оцінюється за метриками: Accuracy – загальна точність передбачень; Recall – частка коректнот передбачених пропагандистських твітів; F1-міра – середнє між точністю та повнотою.

Запропонований модуль демонструє високу ефективність у виявленні пропаганди. Подальше вдосконалення можливе шляхом розширення датасету та адаптації моделі до багатомовного аналізу. Модуль M_3 «Розпізнавання мереж поширення пропаганди» опишемо суперпозицією та відповідною функцією:

$$M_3 = \lambda \circ \gamma_3 \circ \gamma_2 \circ \gamma_1, \quad (42)$$

$$M_3 = \lambda(\gamma_3(\gamma_2(\gamma_1(M_2, u_{13}, u_{14}, r_{13}), u_{14}, r_{13}), u_{13}, u_{24}, r_{23})). \quad (43)$$

Ідея полягає в знаходженні подібних за текстом/ значенням (lexical/ semantical) повідомлень, а також аналізі результатів поширення подібних повідомлень в часі та просторі. Основним процесом модуля M_3 «Розпізнавання мереж поширення пропаганди» є «Формування висновків наявності подібного фейку», який опишемо як:

$$C_{US} = \lambda \circ \gamma \circ \beta \circ \alpha, \quad (44)$$

$$C_{US} = \lambda(\gamma(\beta(\alpha(C_{US}, x_2), M_2, R_{NLP}, U_{NLP}, x_4), M_2, R_{ML}, U_{ML}, u_{14}), M_3, u_{13}, u_{25}, r_{15}). \quad (45)$$

DOI 10.15588/1607-3274-2025-3-13

5 РЕЗУЛЬТАТИ

Отримані результат подано на рис. 1–9. Ці зображення містять по три матриці (Confusion Matrix, Matrix by Recall та Matrix by Precision), що використовуються для оцінки продуктивності класифікаційної моделі. Матриця помилок відображає кількість правильно та неправильно класифікованих зразків. По діагоналі (верхній лівий і нижній правий квадранти) розташовані правильні передбачення для випадкового набору (вибірки) записів з датасету. Позадіагональні значення показують помилки.

Матриця за Повнотою відображає значення, які вказують на рівень повноти (recall) для кожного класу. Наприклад, для рис. 1 для класу 0 recall є 0,38 (тобто лише 38% зразків класу 0 було правильно класифіковано). Для класу 1 recall є 0,91 (91% зразків класу 1 правильно класифіковано). Висока повнота для класу 1 означає, що модель добре знаходить позитивні зразки, але для класу 0 вона неефективна.

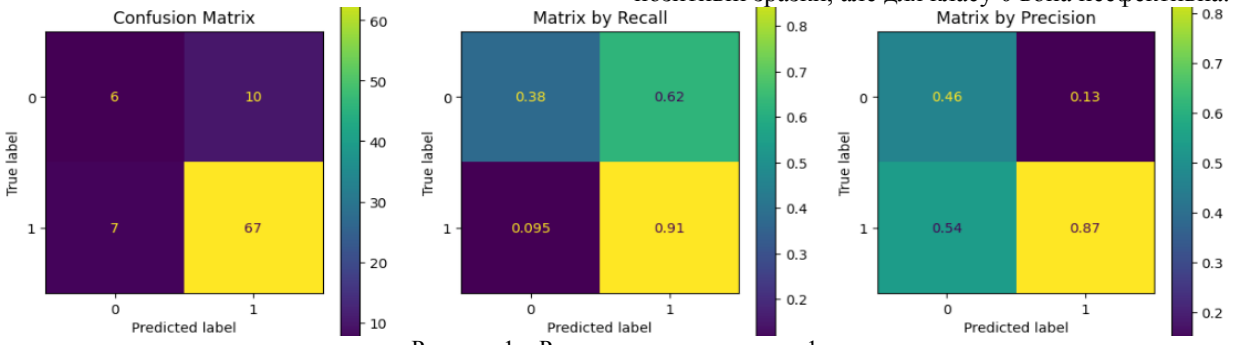

Рисунок 1 – Результати експерименту 1

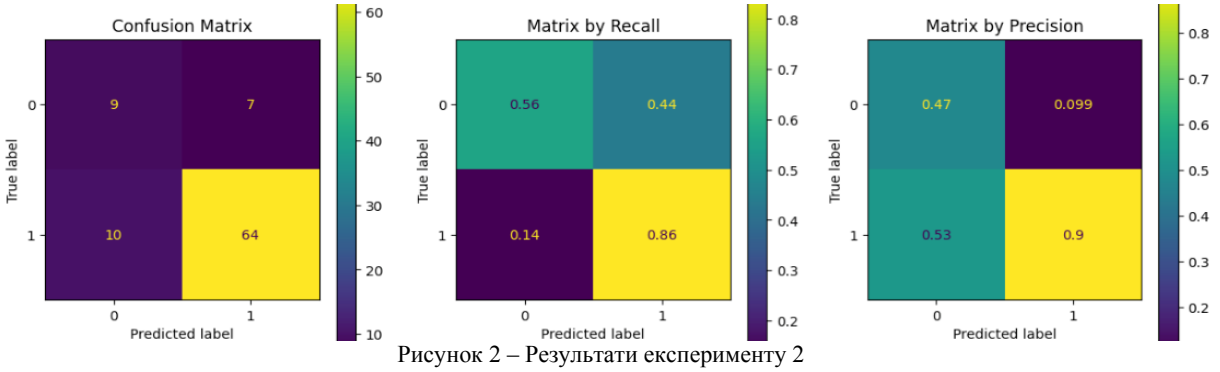

Рисунок 2 – Результати експерименту 2

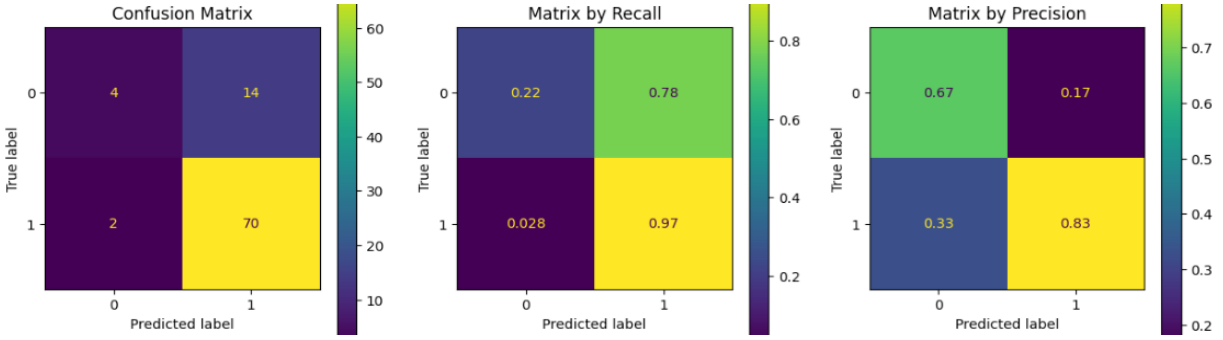

Рисунок 3 – Результати експерименту 3

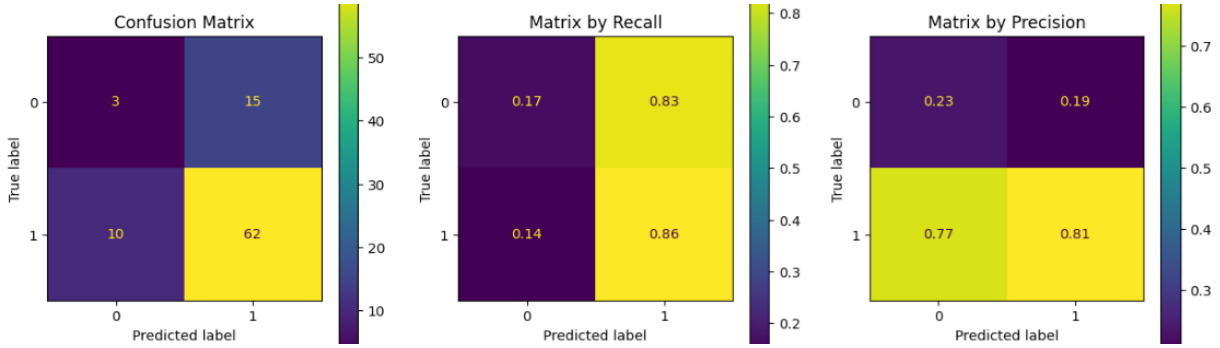

Рисунок 4 – Результати експерименту 4

DOI 10.15588/1607-3274-2025-3-13

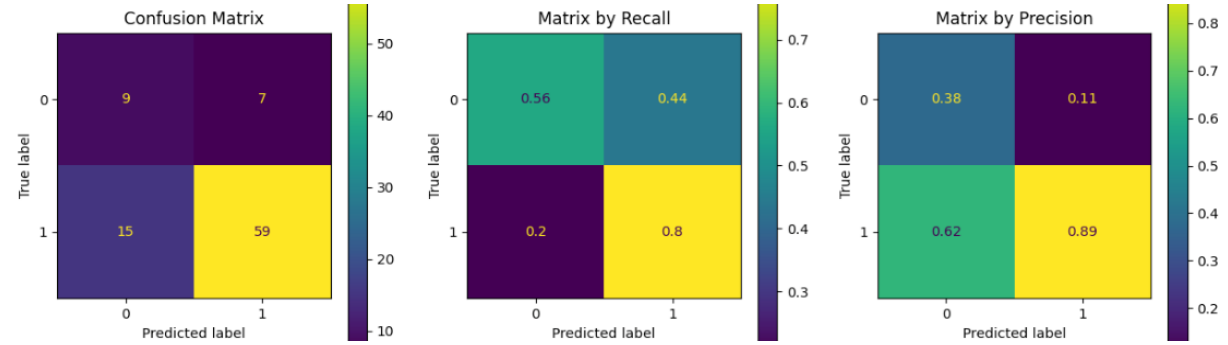

Рисунок 5 – Результати експерименту 5

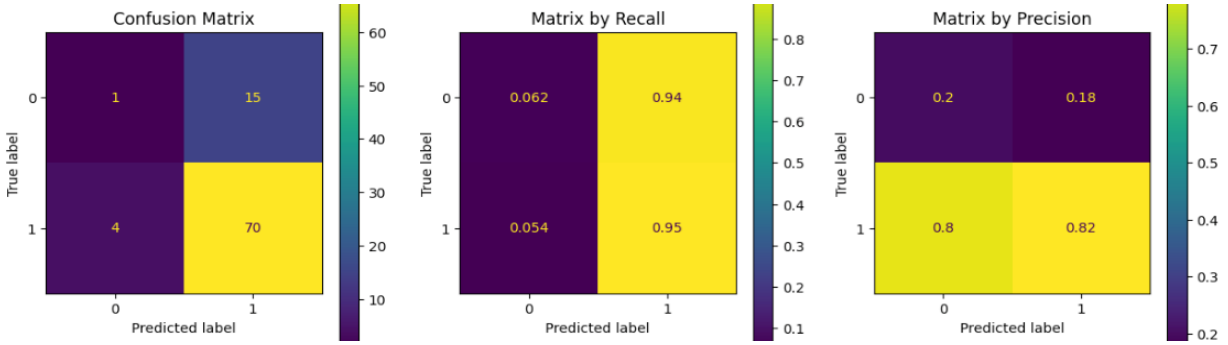

Рисунок 6 – Результати експерименту 6

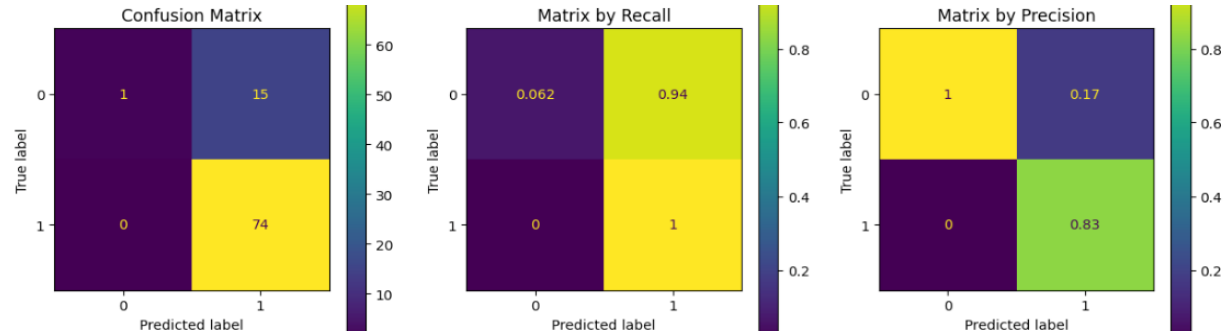

Рисунок 7 – Результати експерименту 7

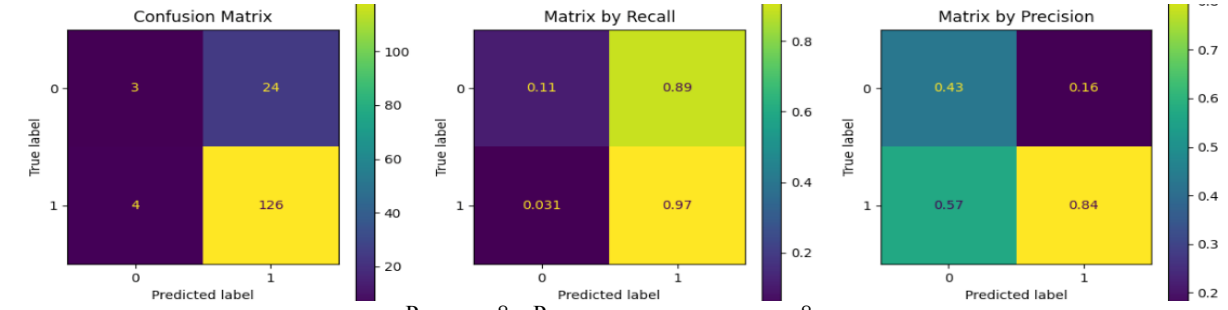

Рисунок 8 – Результати експерименту 8

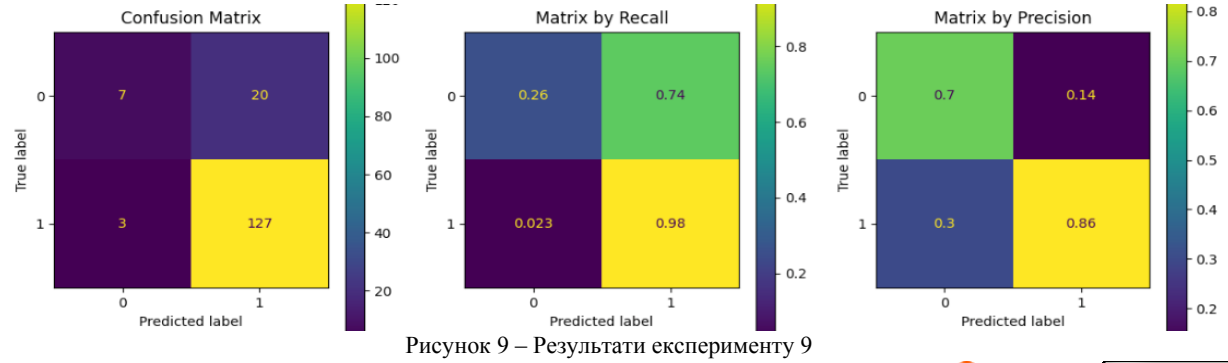

Рисунок 9 – Результати експерименту 9

DOI 10.15588/1607-3274-2025-3-13

Матриця за precision відображає значення, які показують рівень точності, наприклад, для рис. 1 для класу 0 precision є 0,46 (46% передбачень класу 0 були правильними). Для класу 1 precision є 0,87 (87% передбачень класу 1 були правильними). Висока точність для класу 1 означає, що більшість зразків, передбачених як клас 1, є правильними. Низька точність для класу 0 (0,46) вказує на велику кількість помилкових передбачень. Моделі добре передбачає клас 1 (високі recall та precision). Для класу 0 точність і повнота значно нижчі, що може свідчити про класовий дисбаланс або необхідність покращення моделі. У разі потреби можна використати стратегії балансування класів або оптимізувати поріг прийняття рішень для покращення продуктивності.

6 ОБГОВОРЕННЯ

Найкращі результати на даний момент показує експеримент 5 на основні TF-IDF+ ComplementNB (рис. 5). Для класу 0 recall є 0,56 (тобто лише 56% зразків класу 0 було правильно класифіковано). Для класу 1 recall є 0,8 (80% зразків класу 1 правильно класифіковано). Висок а повнота для класу 1 означає, що модель добре знаходить позитивні зразки, але для класу 0 вона менш ефективна (рис. 10). Для рис. 5 для класу 0 precision є 0.38 (38% передбачень класу 0 є правильними). Для класу 1 precision є 0.89 (89% передбачень класу 1 є правильними). Висока точність для класу 1 означає, що більшість зразків, передбачених як клас 1, є правильними. Низька точність для класу 0 (0.38) вказує на велику кількість помилкових передбачень. При цьому в серії експериментів спостерігаються певні аномалії (зокрема в експерименті 7 на основі Glove+ RandomForest – рис. 7), які потребують подальшого дослідження. Підсумовуючі результати по класу F (Фейк) подані на рис. 10–12. Проведення наступних експериментів (комбінацій методів які себе краще показали) а також конструювання нових фічерів (зокрема оцінки сентименту).

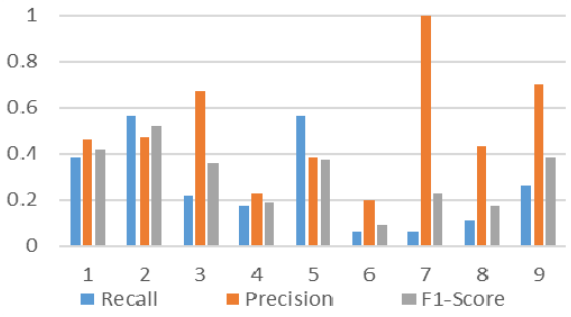

Рисунок 11 – Recall, Precision та F1-Score

ВИСНОВКИ

Стаття описує алгоритм роботи програми, що виконує автоматичне виявлення пропагандистських повідомлень у Twitter, джерел розповсюдження езінформації та неавтентичної поведінки чатів. Основна увага приділяється методам збору та підготовки даних, попередній обробці тексту, векторизації, навчанню моделі та оцінці її ефективності. Методи NLP та ML дозволяють виявляти такі загрози шляхом аналізу стилю авторів, часових закономірностей публікацій та графових зв'язків між користувачами. Описано процес збору, підготовки та очищення даних, а також розглянуто різні підходи до навчання моделі та оцінки її ефективності. Ідея полягає в знаходженні подібних за текстом/ значенням (lexical/ semantical) повідомлень, а також аналізі результатів поширення подібних повідомлень в часі та просторі. У якості основних алгоритмів моделювання використані Complement Naïve Bayes, HistGradientBoostingClassifier, Gaussian Naïve Bayes, Multinomial Naïve Bayes та RandomForest для виявлення джерел розповсюдження дезінформації та неавтентичної поведінки чатів. Основна увага приділяється методам попередньої обробки текстів, векторизації даних та машинному навчанню для автоматичної класифікації повідомлень. Проведено 9 експериментів для різних методів попереднього опрацювання даних, моделей векторизації та алгоритмів моделювання. Найкращі результати на даний момент показує експеримент 5 на основні TF-IDF+ComplementNB. При цьому в серії проведених експериментів спостерігаються певні аномалії (зокрема в експерименті 7 на основі Glove+ RandomForest), які потребують подальшого дослідження. Отримані результати можуть бути використані для подальшого вдосконалення методів виявлення джерел розповсюдження дезінформації, неавтентичної поведінки чатів та шкідливого контенту для збільшення обороздатності країни.

Наукова новизна полягає у розробленні методів:

– ідентифікація схожих за стилістикою фейкових новин для виявлення шляхів розповсюдження дезінформації в часі та просторі;

– стилістичного опрацювання фейкових новин для виявлення спільних лінгвістичних характеристик текстового контенту на основі NLP;

– виявлення неавтентичної поведінки ботів в чатах на основі аналізу скоординованої поведінки користувачів у соціальних мережах та онлайн ЗМІ.

Практична цінність полягає у розробленні системи підтримки прийняття рішення для пошуку та виявлення джерел розповсюдження україномовних фейкових новин, пропаганди, та дезінформації у соціальних мережах та онлайн ЗМІ, а також експериментальна апробація для розрахунку точності отриманих результатів на основі реалізації модуля:

– інтелектуального пошуку, збору, лінгвістичного аналізу, попереднього опрацювання, маркування та класифікації текстового контенту для формування датасету та підготовки даних для виявлення дезінформації та джерел розповсюдження;

– розпізнавання україномовної дезінформації, фейкових новин та пропаганди для виявлення стилістично та змістовно подібного текстового

DOI 10.15588/1607-3274-2025-3-13

контенту при ідентифікації джерел розповсюдження та неавтентичної поведінки ботів;

– розпізнавання мереж поширення пропаганди на основі знаходженні подібних за текстом/ значенням повідомлень, а також аналізі результатів поширення подібних повідомлень в часі та просторі.

Очікувані результати виконання проєкту:

– розроблено метод стилістичного аналізу та лінгвістичного опрацювання текстового контенту на основі NLP та ML для формування інформаційного портрету генератора фейкового повідомлення та подібних до нього за множиною наративів.

– запропоновано моделі та основні принципи інформаційної технології виявлення джерел дезінформації та неавтентичної поведінки користувачів чатів, що дозволить своєчасно виявляти інформаційні загрози в кіберпросторі країни.

– запропоновано параметри та критерії поведінки користувачів чатів для моделі виявлення неавтентичної поведінки ботів, характерні для відповідної групи. Модель неавтентичної поведінки користувача полягає у побудові профілю поведінки користувача системи на основі аналізу поведінкових закономірностей. Вони відображають притаманні підсвідомі характерні риси в межах реалізації відповідного події, що підлягає автентичності. Модель дозволяє виявляти притаманні користувачу підсвідомі поведінкові риси, присутні у різних психоемоційних станах.

ПОДЯКИ

Дана стаття підготована завдяки грантової підтримки Національного Фонду Досліджень України, реєстраційний номер проєкту 187/0012 від 1/08/2024 (2023.04/0012) «Розроблення інформаційної системи автоматичного виявлення джерел дезінформації та неавтентичної поведінки користувачів чатів» за конкурсом «Наука для зміцнення обороноздатності України».

ЛІТЕРАТУРА

1. BERT Based Fake News Detection Model / [Y. Zhang, Y. Shao, X. Zhang et al.] // Training. – 2022. – Vol. 1530. – P. 383.
2. Cahyani D. E. Performance comparison of TF-IDF and word2vec models for emotion text classification / D. E. Cahyani, I. Patasik // Bulletin of Electrical Engineering and Informatics. – 2021. – Vol. 10(5). – P. 2780–2788. DOI: 10.11591/eei.v10i5.3157
3. Bhosale S. Identifying Bots on Twitter with Benford's Law / S. Bhosale. – Access mode: https://scholarworks.sjsu.edu/etd_projects/1041/
4. Ghaemi Z. A Varied Density-based Clustering Approach for Event Detection from Heterogeneous Twitter Data / Z. Ghaemi, M. Farnaghi // ISPRS International Journal of Geo-Information. – 2019. – 8(2). – P. 82. DOI: 10.3390/ijgi8020082
5. Lazebnik T. Temporal graphs anomaly emergence detection: benchmarking for social media interactions / T. Lazebnik, O. Iny // Applied Intelligence. – 2024. – Vol. 54. – P. 12347–12356. DOI: 10.1007/s10489-024-05821-3
6. Do Social Bots (Still) Act Different to Humans? – Comparing Metrics of Social Bots with Those of Humans / [S. Stieglitz, F. Brachten, D. Berthelé et al.] // Lecture Notes in Computer Science. – 2017. – Vol. 10282. – P. 379–395. DOI: 10.1007/978-3-319-58559-8_30
7. Vysotska V. Information technology for recognizing propaganda, fakes and disinformation in textual content based on nlp and machine learning methods / V. Vysotska // Radio Electronics, Computer Science, Control. – 2024. – Vol. 2. – P. 126. DOI: 10.15588/1607-3274-2024-2-13
8. Мокрицька О. В. Використання алгоритмів машинного навчання для автоматизації процесу модерації контенту в групових чатах месенджерів / О. В. Мокрицька, Ю. М. Мочернюк // Scientific Bulletin of UNFU. – 2024. – Том 34(7). – С. 52–59. DOI: 10.36930/40340707
9. Дмитроца Л. П. Аналіз інструментів штучного інтелекту для виявлення дезінформації в новинах Facebook / Л. П. Дмитроца, С. В. Дацик // Інформаційні моделі, системи та технології. – 2023. – С. 35–36. – Access mode: https://elartu.tntu.edu.ua/bitstream/lib/44384/2/IMSTT_2023_Dmytrotsa_L_P-Analysis_of_artificial_35-36.pdf
10. Семенюк А. В. Використання методів машинного навчання та штучного інтелекту для захисту від влпиву соціальної інженерії при кібератаках / А. В. Семенюк. – Access mode: http://ir.lib.vntu.edu.ua/handle/123456789/41797
11. Аналіз методів виявлення дезінформації в соціальних мережах за допомогою машинного навчання / [М. С. Марценюк, В. А. Козачок, О. Богданов, З. М. Бржевська] // Кібербезпека: освіта, наука, техніка. – 2023. – Том 2(22). – С. 148–155. – Access mode: https://elibrary.kubg.edu.ua/id/eprint/48271/
12. Дмитроца Л. П. Застосування методів штучного інтелекту для виявлення та протидії дезінформації у Facebook / Л. П. Дмитроца, С. В. Дацик // Інформаційні моделі, системи та технології. – 2023. – С. 37–38. – Access mode: https://elartu.tntu.edu.ua/bitstream/lib/44385/2/IMSTT_2023_Dmytrotsa_L_P-Application_of_artificial_37-38.pdf
13. Machine Learning and Deep Learning Applications in Disinformation Detection: A Bibliometric Assessment / A. Sandu, L.-A. Cotfas, C. Delcea et al.] // Electronics. – 2024. – Vol. 13(22). – P. 4352. DOI: 10.3390/electronics13224352
14. Santos F. C. C. Artificial Intelligence in Automated Detection of Disinformation: A Thematic Analysis / F. C. C. Santos // Journalism and Media. – 2023. – Vol. 4(2). – P. 679–687. DOI: 10.3390/journalmedia4020043
15. Lakzaei B. Disinformation detection using graph neural networks: a survey / B. Lakzaei, Haghir M. Chehreghani, A. Bagheri // Artificial Intelligence Review. – 2024. – Vol. 57. – P. 52. DOI: 10.1007/s10462-024-10702-9
16. Artificial intelligence in the battle against disinformation and misinformation: a systematic review of challenges and approaches / [H. R. Saeidnia, E. Hosseini, B. Lund et al.] // Knowledge and Information Systems. – 2025. DOI: 10.1007/s10115-024-02337-7
17. Detecting fake news and disinformation using artificial intelligence and machine learning to avoid supply chain disruptions / [P. Akhtar, A.M. Ghouri, H.U.R. Khan et al.] // Annals of Operations Research. – 2023. – Vol. 327. – P. 633–657. DOI: 10.1007/s10479-022-05015-5
18. Disinformation, Fakes and Propaganda Identifying Methods in Online Messages Based on NLP and Machine Learning

DOI 10.15588/1607-3274-2025-3-13

Methods / [V. Vysotska, K. Przystupa, L. Chyrun et al.] // International Journal of Computer Network and Information Security(IJCNIS). – 2024. – Vol. 16(5). – P. 57–85. DOI:10.5815/ijcnis.2024.05.06
19. Prokipchuk O. Ukrainian Language Tweets Analysis Technology for Public Opinion Dynamics Change Prediction Based on Machine Learning / O. Prokipchuk, V. Vysotska // Radio Electronics, Computer Science, Control. – 2023. – № 2(65). – P. 103–116. DOI: 10.15588/1607-3274-2023-2-11
20. Vysotska V. NLP Tool for Extracting Relevant Information from Criminal Reports or Fakes/Propaganda Content / V. Vysotska, S. Mazepa, L. Chyrun, O. Brodyak, I. Shakleina, V. Schuchmann, // Computer Sciences and Information Technologies : 17th International Conference, Lviv, 2022, November. – Lviv: IEEE, 2021. – P. 93–98. DOI: 10.1109/CSIT56902.2022.10000563
21. Information technology for identifying disinformation sources and inauthentic chat users' behaviours based on machine learning / [V. Vysotska, L. Chyrun, S. Chyrun, I. Holets] // CEUR Workshop Proceedings. – 2024. – Vol. 3723. – P. 466–483.
22. Іосіфов Є. Порівняльний аналіз методів, технологій, сервісів та платформ для розпізнавання голосової інформації в системах забезпечення інформаційної безпеки / Є. Іосіфов, В. Соколов // Кібербезпека: освіта, наука, техніка. – 2024. – Том 1(25). – С. 468–486. DOI: 10.28925/2663-4023.2024.25.468486
23. Аналіз методів виявлення дезінформації в соціальних мережах за допомогою машинного навчання / [М. Марценюк, В. Козачок, О. Богданов та ін.] // Кібербезпека: освіта, наука, техніка. – 2023. – Том 2(22). – С. 148–155. DOI: 10.28925/2663-4023.2023.22.148155
24. Інтелектуальний метод виявлення джерел мультилінгвальної дезінформації / [М. Комар, Х. Ліп'яніна-Гончаренко, І. Кіт та ін.] // Measuring and computing devices in technological processes. – 2023. – Том 2. – С. 221–230. DOI: 10.31891/2219-9365-2023-74-31
25. Prytula M. Detection of aggressive rhetoric in text using machine learning algorithms / M. Prytula, I. Olenych // Electronics and information technologies. – 2023. – Vol. 22. DOI: 10.30970/eli.22.4
26. Deep learning for misinformation detection on online social networks: a survey and new perspectives / [M. R. Islam, S. Liu, X. Wang, G. Xu] // Social Network Analysis and Mining. – 2020. – Vol. 10. – P. 82. DOI: 10.1007/s13278-020-00696-x
27. Detecting and responding to hostile disinformation activities on social media using machine learning and deep neural networks / [B. Cartwright, R. Frank, G. Weir, K. Padda] // Neural Computing and Applications. – 2022. – Vol. 34. – P. 15141–15163. DOI: 10.1007/s00521-022-07296-0

Стаття надійшла до редакції 07.03.2025.
Після доробки 09.06.2025.

UDC 004.9

INFORMATION TECHNOLOGY FOR DETECTION OF DISINFORMATION SOURCES AND INAUTHENTICAL BEHAVIOR OF CHAT USERS BASED ON NLP AND MACHINE LEARNING METHODS

Vysotska V. – PhD, Professor of Information Systems and Networks Department, Lviv Polytechnic National University, Lviv, Ukraine.

ABSTRACT

Context. In the modern digital environment, the spread of disinformation and inauthentic behaviour of users in chat rooms poses a serious threat to society. Natural language processing and machine learning methods offer effective approaches to detecting and countering such threats.

Objective of the study is to develop information technology for automatically detecting the spread of sources of Ukrainian-language fake news and inauthentic behaviour of chat users, which is built using natural language processing methods and implemented, based on machine learning technologies.

Method. To implement the project, such feature construction methods as the TF-IDF statistical indicator, the Bag of Words vectorization model, and part-of-speech mark-up were used. For other experiments, the FastText, W2V, and Glove word2vec vectorization models were used to obtain vector representations of words, as well as to recognize trigger words (reinforcing words, absolute pronouns, and "shiny" words). The idea is to find similar messages in terms of text/meaning (lexical/semantical), as well as analyse the results of the distribution of similar messages in time and space. Complement Naïve Bayes, Gaussian Naïve Bayes, HistGradientBoostingClassifier, MultinomialNB and Random Forest were used as the main modelling algorithms to identify sources of disinformation and inauthentic chat behavior.

Results. This article discusses the development of software for detecting propaganda messages in social networks based on the analysis of Twitter text data. The main attention is paid to the methods of text pre-processing, data vectorization and machine learning for message classification. The process of collecting, preparing and cleaning data is described, and various approaches to training the model and evaluating its effectiveness are considered. 9 experiments were conducted for the selected methods of post-processing data, vectorization models and modelling algorithms.

Conclusions. The created models show excellent results in recognizing sources of propaganda, fakes and disinformation in social networks and online media. The best results so far are shown by experiment 5 on the main TF-IDF + Complement Naïve Bayes. The high recall value for class 1 (0.8) means that the model finds positive samples well, but for class 0 it is less effective (0.56). The correspondingly high precision value for class 1 (0.89) means that most of the samples predicted as class 1 are correct. The low precision for class 0 (0.38) indicates a large number of false predictions. At the same time, certain anomalies are observed in the series of experiments (in particular, in experiment 7 based on Glove + Random Forest), which require further research. The results obtained can be used to further improve the algorithms for detecting sources of disinformation, inauthentic chat behaviour and malicious content to increase the country's transparency.

KEYWORDS: disinformation, source of disinformation, way of disinformation dissemination, disinformation dissemination network, fake, propaganda, natural language processing, stylistic analysis.

DOI 10.15588/1607-3274-2025-3-13

REFERENCES

1. Zhang Y., Shao Y., Zhang X., Wan W., Li J., Sun J. BERT Based Fake News Detection Model, *Training,* 2022, Vol. 1530, P. 383.
2. Cahyani D. E., Patasik I. Performance comparison of TF-IDF and word2vec models for emotion text classification, *Bulletin of Electrical Engineering and Informatics,* 2021, Vol. 10(5), pp. 2780–2788. DOI: 10.11591/eei.v10i5.3157
3. Bhosale S. Identifying Bots on Twitter with Benford's Law. Access mode: https://scholarworks.sjsu.edu/etd_projects/1041/
4. Ghaemi Z., Farnaghi M. A Varied Density-based Clustering Approach for Event Detection from Heterogeneous Twitter Data, *ISPRS International Journal of Geo-Information,* 2019, 8(2), P. 82. DOI: 10.3390/ijgi8020082
5. Lazebnik T., Iny O. Temporal graphs anomaly emergence detection: benchmarking for social media interactions, *Applied Intelligence,* 2024, Vol. 54, pp. 12347–12356. DOI: 10.1007/s10489-024-05821-3
6. [Stieglitz S., Brachten F., Berthelé D., Schlaus M., Venetopoulou C., Veutgen D. Do Social Bots (Still) Act Different to Humans? – Comparing Metrics of Social Bots with Those of Humans, *Lecture Notes in Computer Science,* 2017, Vol. 10282, pp. 379–395. DOI: 10.1007/978-3-319-58559-8_30
7. Vysotska V. Information technology for recognizing propaganda, fakes and disinformation in textual content based on nlp and machine learning methods, *Radio Electronics, Computer Science, Control*, 2024, Vol. 2, P. 126. DOI: 10.15588/1607-3274-2024-2-13
8. Mokrytska O. V., Mochernyuk YU. M. Using machine learning algorithms to automate the content moderation process in messenger group chats, *Scientific Bulletin of UNFU*, 2024, Vol. 34(7), pp. 52–59. DOI: 10.36930/40340707
9. Dmytrotsa L. P., Datsyk S. V. Analysis of artificial intelligence tools for detecting disinformation in Facebook news, *Information models, systems and technologies*, 2023. pp. 35–36. Access mode: https://elartu.tntu.edu.ua/bitstream/lib/44384/2/IMSTT_2023_Dmytrotsa_L_P-Analysis_of_artificial_35-36.pdf
10. Semenyuk A. V. Using machine learning and artificial intelligence methods to protect against social engineering in cyberattacks. Access mode: http://ir.lib.vntu.edu.ua/handle/123456789/41797
11. Martsenyuk M. S., Kozachok V. A., Bogdanov O., Brzhevska Z. M. Analysis of methods for detecting disinformation in social networks using machine learning, *Cybersecurity: education, science, technology*, 2023, Vol. 2(22), pp. 148–155. Access mode: https://elibrary.kubg.edu.ua/id/eprint/48271/
12. Dmytrotsa L. P., Datsyk S. V. Application of artificial intelligence methods to detect and counter disinformation on Facebook, *Information models, systems and technologies*, 2023, pp. 37–38. Access mode: https://elartu.tntu.edu.ua/bitstream/lib/44385/2/IMSTT_2023_Dmytrotsa_L_P-Application_of_artificial_37-38.pdf
13. Sandu A., Cotfas L.-A., Delcea C., Ioanăş C., Florescu M.-S., Orzan M. Machine Learning and Deep Learning Applications in Disinformation Detection: A Bibliometric Assessment, *Electronics*, 2024, Vol. 13(22), P. 4352. DOI: 10.3390/electronics13224352
14. Santos F. C. C. Artificial Intelligence in Automated Detection of Disinformation: A Thematic Analysis, *Journalism and Media*, 2023, Vol. 4(2), p. 679–687. DOI: 10.3390/journalmedia4020043
15. Lakzaei B., Chehreghani Haghir M., Bagheri A. Disinformation detection using graph neural networks: a survey, *Artificial Intelligence Review*, 2024, Vol. 57, P. 52. DOI: 10.1007/s10462-024-10702-9
16. [Saeidnia H.R., Hosseini E., Lund B., Tehrani M. A., Zaker S., Molaei S. Artificial intelligence in the battle against disinformation and misinformation: a systematic review of challenges and approaches, *Knowledge and Information Systems*, 2025. DOI: 10.1007/s10115-024-02337-7
17. Akhtar P., Ghouri A. M., Khan H. U. R., Haq M. A., Awan U., Zahoor N., Khan Z., Ashraf A. Detecting fake news and disinformation using artificial intelligence and machine learning to avoid supply chain disruptions, *Annals of Operations Research*, 2023, Vol. 327, pp. 633–657. DOI: 10.1007/s10479-022-05015-5
18. Vysotska V., Przystupa K., Chyrun L., Vladov S., Ushenko Y., Uhryn D., Hu Z. Disinformation, Fakes and Propaganda Identifying Methods in Online Messages Based on NLP and Machine Learning Methods, *International Journal of Computer Network and Information Security(IJCNIS),* 2024, Vol.16(5), pp. 57–85. DOI:10.5815/ijcnis.2024.05.06
19. Prokipchuk O., Vysotska V. Ukrainian Language Tweets Analysis Technology for Public Opinion Dynamics Change Prediction Based on Machine Learning, *Radio Electronics, Computer Science, Control,* 2023, № 2(65), pp. 103–116. DOI: 10.15588/1607-3274-2023-2-11
20. Vysotska V., Mazepa S., Chyrun L., Brodyak O., Shakleina I., Schuchmann V. NLP Tool for Extracting Relevant Information from Criminal Reports or Fakes/Propaganda Content, *Computer Sciences and Information Technologies : 17th International Conference, Lviv, 2022, November.* Lviv, IEEE, 2021, pp. 93–98. DOI: 10.1109/CSIT56902.2022.10000563
21. Vysotska V., Chyrun L., Chyrun S., Holets I. Information technology for identifying disinformation sources and inauthentic chat users' behaviours based on machine learning, *CEUR Workshop Proceedings,* 2024, Vol. 3723, pp. 466–483.
22. Iosifov E., Sokolov V. Comparative analysis of methods, technologies, services and platforms for voice information recognition in information security systems, *Cybersecurity: education, science, technology*, 2024, Vol. 1(25), pp. 468–486. DOI: 10.28925/2663-4023.2024.25.468486
23. Martsenyuk M., Kozachok V., Bogdanov O., Iosifov E., Brzhevska Z. Analysis of methods for detecting disinformation in social networks using machine learning, *Cybersecurity: education, science, technology*, 2023, Vol. 2(22), pp. 148–155. DOI: 10.28925/2663-4023.2023.22.148155
24. Komar M., Lipyanina-Honcharenko H., Kit I., Madarash R., Yurkiv H. An intellectual method for identifying sources of multilingual disinformation, *Measuring and computing devices in technological processes,* 2023, Vol. 2, pp. 221–230. DOI: 10.31891/2219-9365-2023-74-31
25. Prytula M., Olenych I.Detection of aggressive rhetoric in text using machine learning algorithms, *Electronics and information technologies,* 2023, Vol. 22. DOI: 10.30970/eli.22.4
26. Islam M.R., Liu S., Wang X., Xu G.Deep learning for misinformation detection on online social networks: a survey and new perspectives, *Social Network Analysis and Mining,* 2020, Vol. 10, P. 82. DOI: 10.1007/s13278-020-00696-x
27. Cartwright B., Frank R., Weir G., Padda K. Detecting and responding to hostile disinformation activities on social media using machine learning and deep neural networks, *Neural Computing and Applications,* 2022, Vol. 34, pp. 15141–15163. DOI: 10.1007/s00521-022-07296-0

DOI 10.15588/1607-3274-2025-3-13