ISSN 2707-1898 (print)

Український журнал інформаційних технологій
Ukrainian Journal of Information Technology
http://science.lpnu.ua/uk/ujit
https://doi.org/10.23939/ujit2024.02.105
Article received 09.09.2024 p.
Article accepted 19.11.2024 p.
UDC 004.8:9

Correspondence author
V. A. Vysotska
victoria.a.vysotska@lpnu.ua

В. А. Висоцька, Р. В. Романчук

Національний університет "Львівська політехніка", м. Львів, Україна

СИСТЕМА ПІДТРИМКИ ПРИЙНЯТТЯ РІШЕНЬ ВИЯВЛЕННЯ ДЕЗІНФОРМАЦІЇ, ФЕЙКІВ ТА ПРОПАГАНДИ НА ОСНОВІ МАШИННОГО НАВЧАННЯ

Внаслідок спрощення процесів створення та поширення новин через інтернет, а також через фізичну неможливість перевірки великих обсягів інформації, що циркулює у мережі, значно зросли обсяги поширення дезінформації та фейкових новин. Побудовано систему підтримки прийняття рішень щодо виявлення дезінформації, фейків та пропаганди на основі машинного навчання. Досліджено методику аналізу тексту новин для ідентифікації фейку та передбачення виявлення дезінформації в текстах новин. У зв'язку з цим виявлення неправдивих новин стає критичним завданням. Це не лише забезпечує надання користувачам перевіреної та достовірної інформації, а й допомагає запобігти маніпулюванню суспільною свідомістю. Посилення контролю за достовірністю новин важливе для підтримки надійної екосистеми інформаційного простору. Комбінування IR та NLP дає змогу системам автоматично аналізувати та відстежувати інформацію, щоб виявляти можливі факти дезінформації або фейкові новини. Важливо також враховувати контекст, джерело інформації та інші фактори для точного визначення достовірності. Такі автоматичні методи допомагають у реальному часі виявляти та вирішувати проблеми, пов'язані з поширенням дезінформації в соціальних мережах. Для експерименту ми використали набір даних із загальною кількістю 20 000 статей: 10 000 записів для фейкових новин і 10 000 для нефейкових. Більшість статей пов'язані з політикою. Для обох піднаборів даних виконано основні процедури очищення тексту, такі як зміна тексту на малі літери, видалення знаків пунктуації, очищення тегів розташування та автора, а також видалення стоп-слів тощо. Після очищення виконано токенізацію та лематизацію. Для кращих результатів лематизації кожен токен позначено тегом POS. Використання тегів POS допомагає точніше виконувати лематизацію. Для обох піднаборів даних створено біграми та триграми, щоб краще зрозуміти контекст статей у наборі даних. Виявлено, що у нефейкових новинах використовується офіційніший мовний стиль. Проаналізовано настрої в обох піднаборах даних. Результати показують, що фальшивий субнабір даних містить більше негативних балів, тоді як нефальшивий субнабір даних – переважно позитивні оцінки. Піднабори даних були об'єднані перед створенням моделі прогнозування. Для моделі прогнозування використано функції BOW і Logistic Regression. Оцінка F1 становить 0,98 для обох класів фейк / не фейк.

Ключові слова: розпізнавання дезінформації, розпізнавання фейків, опрацювання природної мови, машинне навчання, класифікація новин.

Вступ / Introduction

У XXI ст. інформація стала ключовим інструментом для національного управління, впливу недержавних суб'єктів та змінила динаміку впливу. Останнім часом спостерігається стрімкий розвиток дезінформації та пропаганди в усьому світі. Противники використовують пропаганду та дезінформацію для маніпулювання громадською думкою, руйнуючи соціально-політичні інституції і, отже, ослаблюючи демократію. Глибоке розуміння інформаційного середовища, його впливу на геополітичні події та поведінку людей стає критичним для підтримання безпеки та прийняття важливих політичних рішень. Основним викликом для сучасного суспільства є захист демократичних процесів та особистості від поширення дезінформації та пропаганди. Внаслідок спрощення процесів створення та поширення новин через інтернет, а також фізичної неможливості перевірки великих обсягів інформації, що циркулює в мережі, істотно зросли обсяги поширюваних дезінформації та фейкових новин [1], [2], [3], [4]. Із огляду на це, виявлення неправдивих новин стає критичним завданням. Це не лише забезпечує надання користувачам перевіреної та достовірної інформації, а й допомагає запобігти маніпуляціям суспільною свідомістю. Посилити контроль над достовірністю новин важливо для підтримання надійної екосистеми інформаційного простору. Дезінформація – це поширення неправдивої або об-

манливої інформації з умислом вводити в оману [5]. Головна мета дезінформації полягає в заподіянні економічної шкоди, маніпулюванні громадською думкою або отриманні прибутку. У наш час дезінформацію часто супроводжують фальсифіковані, висмикнуті з контексту та маніпуляційні зображення або відео. Основні канали поширення дезінформації – інтернет-форуми, новинні сайти та соціальні мережі [6]. Важливо зауважити, що вперше Європейська рада визнала загрозу дезінформації у 2015 р., коли росія посилила свої зусилля в цьому напрямі на території Європейського Союзу. Дезінформаційні кампанії, особливо ті, які проводять треті країни, часто стають складовою гібридної війни, що передбачає кібератаки та злам мереж.

Об'єкт дослідження – методи класифікації новин для виявлення дезінформації та передбачення фейків на основі машинного навчання.

Предмет дослідження – дослідити ефективність класифікації новин та розпізнавання фейкових новин на основі машинного навчання.

Мета роботи – розробити систему підтримки прийняття рішень щодо виявлення дезінформації, фейків та пропаганди на основі машинного навчання.

Для досягнення зазначеної мети визначено такі основні *завдання дослідження*:

- аналіз набору даних;
- розроблення модулів виявлення дезінформації основі машинного навчання;
- аналізування отриманих результатів на точність виявлення фейків.

Аналіз останніх досліджень та публікацій. У "Плані дій ЄС щодо дезінформації" з 2018 р. визначено, що дезінформація становить складову частину гібридних кампаній, спрямованих на роз'єднання європейського суспільства та підрив довіри громадян до демократичних процесів та інститутів ЄС. Автори дослідження "Інформаційний розлад: до міждисциплінарних рамок для досліджень та формування політики", опублікованого у 2017 р. за підтримки Ради Європи, – Клер Вордл (Claire Wardle) і Хосейн Деракшан (Hossein Derakhshan). Згідно з цим дослідженням, інформаційний розлад охоплює три основні концепції: неправдива інформація, дезінформація та шкідлива інформація.

Неправдива інформація (*misinformation*) – це інформація, яка є неточною, але створена без наміру завдати шкоди. Цей тип інформації часто виникає внаслідок ненавмисних помилок, неправильних підписів до фотографій або відео, некоректних дат або статистики тощо.

Дезінформація (*disinformation*) – це навмисно створена неправдива інформація з метою заподіяти шкоду країні, організації, соціальній групі чи особі. Також це стосується цілеспрямовано зміненого або фабрикованого відео/аудіоконтенту, а також створення різноманітних конспіраційних теорій або чуток.

Шкідлива інформація (*malinformation*) – це інформація, яка ґрунтується на справжніх подіях, але застосовується для завдання шкоди країні, організації, соціальній групі чи особі. Це також зумисна зміна контексту тексту повідомлення та навмисне розголошення особистих даних у власних інтересах.

У Європейському Союзі існує проблема з узгодженням загальних термінів, таких як "дезінформація", "пропаганда" і "фейкові новини", через відсутність єдиних визначень, що використовуються в ЄС. Термін "фейкові новини" тлумачать по-різному, але Кембриджський словник надає одне з чітких визначень: "це неправдиві історії, які, виглядаючи як новинні повідомлення, поширюються в інтернеті або використовуються в інших ЗМІ, зазвичай з метою впливу на політичні переконання або як гумор".

Група високого рівня з фейкових новин та онлайн-дезінформації зробила першу офіційну спробу визначити термін "дезінформація" у своєму звіті [7] у 2018 р., стверджуючи, що різке зростання поширення фейків прямо залежить від темпів розвитку інформаційних технологій, їх активного впровадження в ЗМІ. Ці проблеми також мають маніпулятивні вияви у комунікаційних механізмах.

Для того, щоб вирішити завдання виявлення та боротьби з поширенням фейкової інформації, важливо розуміти принципи, за якими її поширюють. У цьому контексті комп'ютерна лінгвістика використовується для розкриття правил і шаблонів, за якими можна ідентифікувати елементи дезінформації в потоці текстових даних. Дослідження [8] вказують на те, що для створення фейкової інформації використовується певний набір правил, які роблять новину правдоподібною.

Неправдивий контент має конкретні особливості, такі як використання скорочень, передавання обмеженої кількості інформації та наявність негативного спрямування. Методи комп'ютерної лінгвістики дають змогу аналізувати ці особливості, сприяючи виявленню та ідентифікації потенційно фейкової інформації в текстових джерелах.

У зв'язку з масовим поширенням дезінформаційних кампаній все більше дослідників намагаються вирішити завдання, що стосуються боротьби з ними. Зокрема, у [9] автори викладають загальний погляд на "хибні новини" (*fake news*). Також з'являються дослідження, які стосуються швидкого зростання дисбалансу між правдивими і неправдивими новинами у соціальних мережах [10], [11], перевірки викладених фактів під час політичних дебатів [12], [13], перевірки фактів на форумі відповідей на запитання спільноти [14] тощо.

Автори публікації [15] вважають, що фейкові новини можна розділити на клікбейт, впливові та сатиричні. Щоб боротися з поширенням фейкових новин, було використано такі методи, як виявлення спаму, визначення позиції та створення еталонних даних. Далі проаналізовано настрої, що належать до методів опрацювання природної мови. Один із прикладів фейкових новин, що обговорювався, – це історія про "Електорошок" китайського робота з безпеки в аеропорту, що сталася у 2016 р. і призвела до появи понад 12 тисяч фейкових новин в Китаї, які були розміщені на 244 різних вебсайтах як джерела.

Останнім часом дослідники, що вивчають проблеми упередження та дезінформації, частіше звертають увагу на соціальні мережі та інші специфічні онлайн-платформи. У публікаціях [16], [17] автори досліджують достовірність та упередженість повідомлень у Twitter (нині X), способи протидії токсичним коментарям на різноманітних форумах [18]. Разом з тим існують дослідження проблеми правдивості статей і у цифрових та друкованих ЗМІ [19] а також джерел новин для ЗМІ [20].

З погляду виділення ознак більшість попередніх досліджень залежать від стилістичних та мовних особливостей, які можуть бути незалежними від теми чи жанру. Інакше кажучи, незалежно від того, чи подія згадується в конкретній статті новин або від її упередженості, характеристики повинні містити таку інформацію, яка допоможе моделі ML приймати правильні рішення. Зокрема, це основний принцип дизайну, що використовується в профілюванні користувачів, щоб перевірити, чи сумнівний вміст написав той самий користувач, який публікує широкий спектр однотипних матеріалів [21].

Різні ознаки (features) по-різному впливають на кінцевий результат. Наприклад, довжина вмісту та особливості теми відіграють дуже незначну роль у вирішенні таких завдань. Натомість символьні N-грами (char N-grams) істотно впливають на фінальний результат [22]. Помічено, що ознаки цього типу є впливовішими та значущішими для визначення пропаганди та рівнів упередженості. Попередні дослідження виявлення упередженості та пропаганди також досліджували цей тип ознак. Щоб перевірити достовірність тверджень у новинах, автори [23] використовували стилістичні особливості, зокрема появу асертивних, фактивних дієслів, а також дієслів непрямої мови, пом'якшувальних та імплікативних слів, а також маркери дискурсу, видобуті вручну за допомогою створених вручну лексиконів.

У дослідженні [24] автори звернули увагу на значне зростання впливу фейкових новин на повсякденне життя. Вони аналізували три методи для ідентифікації неправдивих новин: наївний метод Баєса, нейронні мережі та метод опорних векторів. Згідно з їх дослідженням, наївний метод Баєса мав точність виявлення фейкових новин 96,08 %, тоді як метод опорних векторів показав точність 99,90 %.

Для виявлення категорично упереджених думок, а саме відповіді "точно так" або "точно ні", корисно звертати увагу на стилометричні характеристики. Наприклад, автори праці [25] досліджували відмінності між правдивими новинами, сатиричними та фейковими новинами. Вони створили набір даних, який складався з новинних статей, які журналісти зібрали з дев'яти різних вебджерел і перевірили вручну. Для ідентифікації вони використовували стилометричну техніку, розроблену в [26] для перевірки авторства, щоб передбачити фактичність і ступінь упередженості. Дослідники припустили, що упереджений вміст має типовий стиль написання. В іншому тематичному дослідженні розроблено модель для виявлення фейкових новин [27]. Науковці зробили висновок, що використання власних імен та структури заголовків є дуже важливими характеристиками для класифікації фейкових та справжніх новин. Запропоновано також інший метод для розрізнення справжніх, фейкових і сатиричних новин за допомогою індикаторів стилю написання та складності. Серед характеристик були різні теги частин мови, лайливі та сленгові слова, стоп-слова, знаки пунктуації та заперечення. Вивчаючи показники складності, дослідники використовували різні індекси читабельності та дійшли висновку, що у "фейкових новин" низькі показники читабельності, менша довжина, використовується проста мова та менше технічних слів. Крім того, вони встановили, що фейкові новини ближче до сатири. Спочатку вибрали понад 130 характеристик, але більшість з них не виявилися корисними.

Колектив авторів статті [28] розробив модель, яка використовує алгоритм машини опорних векторів (SVM) для збирання статей та визначення, чи є вони легітимними, чи фальсифікованими. За допомогою алгоритму SVM для бінарної класифікації модель систематизує статті та класифікує їх як справжні чи фейкові. Розроблені моделі містять три головні компоненти: агрегатор для збирання статей, аутентифікатор для перевірки їх правдивості та систему рекомендацій. Для додаткової перевірки правдивості статей команда використала наївний баєсівський алгоритм, який у поєднанні з SVM і NLP забезпечив точність 93,50 % у класифікації новин.

Для класифікації новинних статей на чотири категорії (серед яких справжні, сатиричні, містифікація та пропаганда) дослідники [29] сформували корпус новинних матеріалів, взятих із англійського Gigaword та семи різних джерел новин. Вони дійшли висновку, що використання словесних N-грам погіршило результати під час тестування на невідомих джерелах. У 2017 р. була розроблена бінарна модель класифікації пропаганди [30] за допомогою загальнодоступного набору даних, що містив інформацію із 15 тисяч ботів. Новинні статті отримали мітки на основі опублікованого вмісту, який могли створити боти або люди. Їхня модель забезпечила 95 % показник AUC за допомогою методу Random Forest. Пізніше дослідники [31] розробили модель ідентифікації, використовуючи 130 ознак на основі контенту, зібраного із літературних джерел.

У дослідженні [32] автори аналізували методи виявлення фейкових новин на основі дописів у Twitter. Вони зосередилися на визначенні, чи є дописи справжніми, чи фальшивими, зокрема, розглядаючи такі події, як землетрус у Чилі 2010 р. та президентські вибори в США. Основний метод виявлення фейкових новин, який запропонували ці дослідники, ґрунтувався на опрацюванні природної мови, що передбачало класифікацію новин перед застосуванням різноманітних моделей машинного навчання для отримання результатів. Автори звернули увагу на підвищення ефективності виявлення фейкових новин у своїх методах у разі використання довжини слова. Під час дослідження було застосовано п'ять різних алгоритмів машинного навчання: наївний метод Баєса, метод опорних векторів (SVM), логістичну регресію, рекурентні нейронні мережі та довготривалу і короткочасну пам'ять. Автори ввели чотири типи векторів ознак: вектори підрахунку, вектори на рівні слів, N-грамові вектори та вектори на рівні символів, серед яких метод опорних векторів виявився найефективнішим для ідентифікації фейкових новин.

У 2019 р. здійснено аналіз даних із 293 57 101 статті [33] з використанням нейронних мереж. Результати дослідження охоплюють класифікацію на рівні речень і класифікацію на рівні фрагментів. Отримані дані показали, що класифікація на рівні речення дала результат 60,82 % F1-score, тоді як класифікація на рівні фрагментів – 22,58 % F1-score. Пізніше автори [34] розв'язували задачу ідентифікації підозрілих термінів щодо радикального вмісту за допомогою методів машинного навчання. Їхні результати засвідчили, що Random Forest є найкращим класифікатором і забезпечує точність на

рівні 94 %. Пізніше була розроблена бінарна модель ідентифікації пропаганди [35]. Автори сформували набір даних Qprop, який складається із реальних новинних статей із 104 різних новинних видань. Їх репрезентація ґрунтувалася на порівнянні різних текстових та лінгвістичних моделей. Згідно із оцінками дослідників, символьні N-грами дають найкращі результати у поєднанні з Nela. Подібні підходи описано в [36] та [37], які використовували вектори Tf-IDF, POS, Lexicons based та Word2vec для ідентифікації пропаганди.

У дослідженні [38] розроблено метод ідентифікації фейкових новин на основі їх поширення за допомогою графової нейронної мережі. Автори виконали оцінку своєї моделі як на відомих, так і на невідомих наборах даних. У роботі [41] запропоновано двоетапну систему для ідентифікації пропаганди в новинних статтях. Експерименти з набором даних, що складався з 550 новинних статей, показали, що їх модель перевершує найсучасніші методи. Також автори у [39] запропонували метод, який поєднує методи сентиментного аналізу з використанням word2vec, стверджуючи, що ця комбінація семантичного та емоційного аналізу дає змогу краще вирішувати завдання ідентифікації пропаганди.

Останнім часом набувають популярності великі мовні моделі (LLM). У публікації [40] автори досліджують ефективність сучасних великих мовних моделей (LLMs), таких як GPT-3 та GPT-4, для виявлення пропаганди. Виконано експерименти з використанням набору даних SemEval-2020 завдання 11, який містить новинні статті, позначені 14 техніками пропаганди як проблемою багатоміткової класифікації. Використано п'ять варіантів GPT-3 та GPT-4, із різними стратегіями розроблення підказок та доведення різних моделей до високої точності. Результати показують, що GPT-4 досягає результатів, порівнянних з найкращими сучасними підходами. Крім того, це дослідження аналізує потенціал та виклики LLM у складних завданнях, таких як виявлення пропаганди.

Результати дослідження та їх обговорення / Research results and their discussion

Проблему виявлення фейкових джерел інформації можна розглядати як аналогічну до завдань виявлення спаму, особливо у використанні статистичних методів машинного навчання. Для виявлення спаму вже успішно використовують методи машинного навчання, які застосовують для класифікації тексту, такого як твіти чи електронні листи, щоб визначити, чи є він спамом. Ці методи машинного навчання можуть передбачати попереднє оброблення тексту, визначення ознак (наприклад, за допомогою "bag of words"), та відсікання непотрібних ознак для підвищення точності на тестовому наборі даних. Принцип роботи полягає у використанні навчального набору даних, на якому модель "навчається" розпізнавати ознаки спаму, і подальшому застосуванні цієї моделі для класифікації нового тексту як спаму чи не спаму.

Такий підхід можна адаптувати для виявлення фейкових джерел інформації, використовуючи відповідний навчальний набір даних і розробляючи моделі для визначення характерних ознак фейків.

Коли характеристики вже ідентифіковані, то їх використовують для класифікації за допомогою різних методів машинного навчання, які передбачають навчання з учителем. Зокрема, можна використовувати такі методи, як метод опорних векторів, наївний баєсівський класифікатор, TF-IDF або метод *K*-найближчих сусідів:

$$C_{spam} = f(C_{news}, \theta),$$

де C_{news} – новини для класифікації, який є вектором для коефіцієнтів C_{leg} і C_{spam} (правдиві та хибні повідомлення), де $C_{leg} \cap C_{spam} = 0$ та $C_{news} = C_{leg} \cup C_{spam}$, θ – множина характерних ознак фейків.

Завдання ідентифікації фейків подібне до завдання виявлення спаму, оскільки обидва вони спрямовані на відокремлення чесних текстів від фейкових, неправдивих. Однак важливо розглядати ці завдання окремо через низку відмінностей. Для ідентифікації частин фейку застосовують прийоми, схожі на використовувані для виявлення спаму.

POS, або Part-of-Speech тегінг, – це маркування слів у тексті відповідно до їхніх частин мови (наприклад, іменник, дієслово, прикметник тощо). Цей процес є важливим кроком у багатьох завданнях оброблення природної мови, таких як синтаксичний аналіз, автоматичний переклад та семантичний аналіз. Застосування POS тегінгу допомагає машинам точніше розуміти структуру і значення речень, покращуючи якість і підвищуючи точність оброблення тексту.

Word N-gram – це модель, яка ґрунтується на послідовностях N слів у тексті. Цю техніку використовують з метою оброблення природної мови для аналізу контексту і залежностей між словами. N-грами допомагають покращувати розуміння тексту, сприяючи точнішому перекладу, створенню моделей мовлення і розпізнаванню мовних зразків. Часто вони застосовуються в задачах прогнозування наступного слова та автоматичного генерування тексту, забезпечуючи основу для багатьох сучасних систем оброблення тексту.

Char N-gram – це модель, що використовує послідовності з N символів у тексті для аналізу. Вона дає змогу виявляти закономірності та залежності між символами у мовленні, що може бути корисним для завдань опрацювання природної мови як розпізнавання мови, класифікації тексту, а також для створення мовних моделей та фільтрації спаму. Завдяки Char N-gram можна ефективно аналізувати текст незалежно від його мови та словникового складу, що робить її універсальним інструментом для оброблення текстової інформації.

Word2Vec – це алгоритм для векторизації слів у тексті, який перетворює слова на вектори чисел у просторі. Він використовує нейронні мережі для навчання такого подання слів, щоб семантично близькі слова в просторі векторів були розташовані близько одне до одного. Цей метод дає змогу застосовувати математичні операції для роботи зі словами, такі як віднімання або додавання векторів, щоб відтворювати семантичні відносини між ними. Word2Vec широко використовують в різних завданнях оброблення природної мови, таких як кластеризація текстів, машинний переклад, аналіз настроїв та багато інших.

Bidirectional Encoder Representations from Transformers (BERT) [43] – це модель для подання слів на основі трансформерів, яка здатна досягати надзвичайних результатів у різних завданнях оброблення природ-

ної мови. Вона використовує контекстуальні вектори слів, що дає змогу враховувати залежності між словами у реченні в обидва напрямки. BERT натренували на великому корпусі тексту, здійснивши навчання на двох завданнях: передбачення наступного слова в реченні й передбачення, чи є речення з набору тексту послідовним. Модель можна доопрацювати для виконання різних завдань оброблення природної мови за допомогою дошкільного навчання (fine-tuning) на специфічних даних для конкретного завдання. BERT – одна з найефективніших і популярних моделей у галузі оброблення природної мови, завдяки її здатностям до контекстуального розуміння слів і високій точності виконання різних завдань.

ELMo (Embeddings from Language Models) – метод, який використовує контекстуальні вектори слів, навчені на великому корпусі тексту, для подання слів у реченнях. Він ґрунтується на технології LSTM (Long Short-Term Memory) і використовує механізм передавання контексту для того, щоб кожне слово мало векторне представлення, яке враховує його значення у конкретному контексті. Такий підхід дає змогу краще враховувати семантичні зв'язки між словами у реченні. ELMo можна використовувати для різних завдань оброблення природної мови, таких як класифікація тексту, машинний переклад, розпізнавання іменованих сутностей та інших, завдяки його здатності до контекстуального розуміння слів у тексті [44].

TF-IDF (Term Frequency-Inverse Document Frequency) – це статистичний метод, який використовують для визначення важливості термів у документах стосовно корпусу текстів. TF-IDF обчислюють, перемноживши два значення: TF, яке відображає частоту терміна у документі, і IDF, яке відображає обернену частоту документа, що містить цей термін у всьому корпусі. Висока значущість TF-IDF вказує на те, що термін часто трапляється у конкретному документі, але рідко – в інших документах корпусу, що робить його важливим для цього конкретного документа. TF-IDF часто використовують для вагового виділення важливих слів у документах для різних завдань оброблення природної мови, таких як класифікація тексту, кластеризація, індексація тощо [45].

Bag-of-Words (BoW) – простий, але потужний метод для векторизації тексту, який використовують для оброблення природної мови. У цьому методі кожне речення або документ видається як вектор, кожне значення якого відповідає кількості входжень кожного слова із попередньо визначеного словника в документ. BoW використовують для створення "мішка слів", ігноруючи послідовність слів у документі й зосереджуючись тільки на їхній частоті. Цей метод часто використовують для побудови моделей класифікації тексту, кластеризації й аналізу настроїв [46].

Методи машинного навчання, такі як Naive Bayes, Support Vector Machines, TF-IDF або *K*-найближчих сусідів, можуть застосовуватися для аналізу тексту та класифікації його як потенційно дезінформаційного. Аналіз ознак, визначення "bag of words" та відсікання непотрібних ознак також можна використовувати для підвищення точності виявлення.

Важливо враховувати, що виявлення елементів дезінформації у тексті – складніше завдання порівняно із виявленням спаму. Іноді виникають субтильні нюанси, такі як заміна імені людини, що може призвести до зміни смислу тексту. Тому модель потребує навчання на основі широкого спектра інших новин та статей із конкретної теми для досягнення високої точності визначення дезінформації. Крім того, пошук способів поширення фейкових новин може слугувати ефективним інструментом для перевірки історій та визначення їхньої достовірності через зіставлення інформації та перевірку фактів. Розгляду слів ізольовано може виявити недостатньо для ефективного прогнозування фейкових повідомлень. Отже, деякі вчені використовують інші лінгвістичні стратегії, зокрема аналіз граматики та синтаксису природної мови. Один із таких підходів передбачає використання ймовірнісних контекстовільних граматик (PCFG). PCFG застосовують для генерування дерев розбирання речень, які описують їх структуру і перетворюють їх на рекурсивну синтаксичну структуру. Синтаксичний аналіз, який може використовувати PCFG, широко застосовують для семантичного аналізу в сентимент-аналізі. Цей підхід може допомогти виявити синтаксичні аномалії, які можуть свідчити про те, що інформація фейкова або маніпулятивна. Синтаксичний аналіз тексту приводить до перетворення повідомлення на структуру даних, яку зазвичай подають у вигляді дерева. Це дерево відображає синтаксичну структуру повідомлення, надаючи зрозуміле візуальне представлення синтаксичних зв'язків між словами та фразами. Це дерево можна використовувати для подальшого опрацювання тексту. Синтаксичний аналіз зазвичай розглядають як двоетапний:

Лексичний аналіз (Tokenization): На цьому етапі текст розділяють на токени, які є базовими одиницями синтаксичного аналізу. Лексичний аналіз визначає, які слова чи символи становлять токени та як вони пов'язані між собою.

Створення дерева розбирання (Parsing): На цьому етапі використовуються правила граматики мови для створення структури даних у вигляді дерева, що відображає синтаксичну структуру тексту. Дерево розбирання дає змогу визначити відносини між словами та фразами в реченні.

У контексті виявлення фейкових новин або дезінформації синтаксичний аналіз може допомогти виявляти аномалії у структурі тексту, характерні для маніпулятивної інформації. Аналіз коментарів та порівняння їх зі схожими статтями – метод, корисний також для оцінювання правдивості новин чи тексту. Якщо більшість подібних статей не підтверджують дані щодо новини, це часто свідчить, що новина упереджена або фальшива. Аналіз коментарів також може слугувати індикатором. Якщо коментарі вказують на сумніви, відсутність підтверджень або розглядають питання достовірності статті, це може вказувати на неправдиву інформацію. Однак важливо бути обережним, використовуючи цей метод, оскільки деякі фальшиві новини поширюють через соціальні мережі або інші канали, й обсяг коментарів людей, які їх підтримують, іноді значний. Також потрібно враховувати можливість створення штучних коментарів для підтримки фейкових новин. Проте коментарі можуть слугувати важливим джерелом додаткової інформації під час оцінювання достовірності контенту. Для виявлення дезінформації

використовують також методи ручної та автоматичної перевірки. Ручну перевірку фактів можна розділити на дві основні категорії: експертну та краудсорсингову перевірку.

Експертна перевірка фактів:

- Її здійснюють експерти, які мають спеціалізовані знання стосовно конкретної області або предмета.
- Зазвичай використовується для перевірки обмеженої кількості даних або конкретних фактів.
- Забезпечує високу точність, але в разі опрацювання великого обсягу інформації дорога та маломасштабована.

Краудсорсингова перевірка фактів:

- Залучає широкий колектив користувачів чи ентузіастів для перевірки фактів або достовірності інформації.
- Масштабованіша та дає змогу обробляти великі обсяги даних.
- Зазвичай менш точна порівняно з експертною, оскільки передбачає різні рівні експертизи.

```
fake_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 5 columns):
 #   Column   Non-Null Count  Dtype 
---  ------   --------------  ----- 
 0   title    10000 non-null  object
 1   text     10000 non-null  object
 2   subject  10000 non-null  object
 3   date     10000 non-null  object
 4   True     10000 non-null  int64 
dtypes: int64(1), object(4)
memory usage: 390.8+ KB
```

Рис. 1. Атрибути DataFrame для фейкових новин датасету / DataFrame attributes for fake news dataset

Кожен із цих методів має певні переваги та обмеження. Вибір конкретного методу може залежати від обсягу інформації, яку потрібно перевірити, та наявності ресурсів для залучення експертів чи широкого кола користувачів. Часто, враховуючи обсяг та швидкість поширення новин у соціальних мережах, ручна перевірка фактів стає непрактичною для ефективного виявлення та відповіді на дезінформацію. Саме тому розробники використовують автоматичні методи перевірки фактів, основані на технологіях пошуку інформації (IR) та Natural Language Processing (NLP).

NLP:

- Застосовується для розуміння та оброблення мовлення людей.
- Алгоритми NLP можуть аналізувати текст, розпізнавати семантику та виявляти ключові факти та аспекти.

IR:

- Використовує методи пошуку для виявлення інформації в текстових джерелах.
- Алгоритми пошуку дають змогу швидко виділяти та видобувати релевантні факти з різних джерел.

```
true_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 5 columns):
 #   Column   Non-Null Count  Dtype 
---  ------   --------------  ----- 
 0   title    10000 non-null  object
 1   text     10000 non-null  object
 2   subject  10000 non-null  object
 3   date     10000 non-null  object
 4   True     10000 non-null  int64 
dtypes: int64(1), object(4)
memory usage: 390.8+ KB
```

Рис. 2. Атрибути DataFrame для правдивих новин датасету / DataFrame attributes for true dataset news

Комбінування IR та NLP дає змогу системам автоматично аналізувати та відстежувати інформацію, щоб виявляти можливі факти дезінформації або фейкових новин. Важливо також враховувати контекст, джерело інформації та інші фактори, щоб точно визначити достовірність. Такі автоматичні методи можуть допомогти в реальному часі виявляти та вирішувати проблеми, пов'язані із поширенням дезінформації в соціальних мережах. Для експерименту ми використовуємо набір даних із загалом 20 000 статей: 10 000 записів для фейкових новин (рис. 1) і 10 000 для нефейкових (рис. 2). Більшість статей пов'язані із політикою. Цей набір даних сформований із розміщеного в мережі популярного набору даних WELFake (https://www.kaggle.com/datasets/saurabhshahane/fake-news-classification). Автори набору – P. K. Verma та ін. [47], він складається з 72134 новин із розподілом: 35028 нефейкових і 37106 фейкових. Цей набір є об'єднанням чотирьох інших наборів новин – Kaggle, McIntire, Reuters, BuzzFeed Political. Для нашого набору з обох класів новин ми довільно вибрали по 10000 новин для кращої збалансованості набору. Збалансованість набору даних для задач класифікації допомагає уникнути упередженості моделі. Моделі, навчені на незбалансованих наборах даних, частіше можуть брати до уваги характеристики більшої підгрупи даних, хоча часто у таких випадках характеристики меншості важливіші. Це може впливати на правильність метрик (відгук, точність) – модель може забезпечувати хороші результати, переважно прогнозуючи клас більшості, й мати проблеми під час прогнозування класу меншості. Для обох піднаборів даних виконано основні процедури очищення тексту, такі як зміна тексту на малі літери, видалення знаків пунктуації, очищення тегів розташування та автора, а також видалення стоп-слів тощо.

Після очищення було виконано токенізацію та лематизацію. Для кращих результатів лематизації кожен токен був позначений тегом POS. Використання тегів POS допомагає точніше виконувати лематизацію.

Для обох піднаборів даних ми створили біграми та триграми, щоб краще зрозуміти контекст статей у наборі даних.

Як бачимо на рис. 3–6, у нефейкових новинах використовується офіційніший мовний стиль. Наприклад, верхня триграма із підробленого піднабору даних – це

“(donald, j, trump)”, тоді як для непідробленого піднабору даних – “(president, donald, trump)”. Після цього ми проаналізували настрої на обох піднаборах даних (рис. 7–8). Результати показують, що фальшивий субнабір даних містить більше негативних балів, тоді як нефальшивий – переважно позитивні оцінки.

Піднабори даних були об’єднані перед створенням моделі прогнозування. На рис. 9 подано головну частину отриманого набору даних.

Для моделі прогнозу використано алгоритм векторизаціх текстових даних BoW і статистичний метод Logistic Regression.

BoW – один із основних методів векторизації текстових даних. Він не враховує граматичну структуру та послідовність слів. Для роботи алгоритму створюють словник унікальних токенів, що виявлені у всіх документах з набору даних. Потім кожен документ видається у вигляді вектора, в якому кожен елемент відповідає частоті появи конкретного слова із словника у документі. Оскільки поставлена задача – це підзадача бінарної класифікації, вибрано саме цей алгоритм векторизації, адже він ефективно працює із великими наборами текстових даних, простий у реалізації, завдяки відносній простоті зменшує час оброблення даних, а також може у майбутньому поєднуватись з іншими методам.

Logistic Regression (Логістична регресія) – статистичний метод, який застосовують у задачах пошуку бінарних залежностей. Розроблений спеціально для бінарних класифікацій, швидкий та простий для розуміння. Також логістична регресія надає додаткові дані – імовірності класифікації, що може бути корисним для майбутніх досліджень і експериментів.

Результати моделі наведено на рис. 10.

```
bigrams = (pd.Series(nltk.ngrams(tokens_clean_fake, 2)).value_counts())
print(bigrams[:10])
```

```
(donald, trump)       3353
(hillary, clinton)    2097
(united, state)       2080
(white, house)        1989
(new, york)           1404
(president, obama)    1210
(president, trump)    1031
(fox, news)            995
(can, not)             682
(barack, obama)        677
Name: count, dtype: int64
```

Рис. 3. Частовживані біграми із фейкових новин датасету / Frequently used 2-grams from the fake news dataset

```
trigrams = (pd.Series(nltk.ngrams(tokens_clean_fake, 3)).value_counts())
print(trigrams[:10])
```

```
(donald, j, trump)              643
(21st, century, wire)           542
(j, trump, realdonaldtrump)     523
(new, york, time)               488
(news, 21st, century)           397
(black, life, matter)           377
(subscribe, become, member)     333
(become, member, 21wiretv)      323
(video, screen, capture)        306
(image, video, screen)          297
Name: count, dtype: int64
```

Рис. 4. Частовживані триграми із фейкових новин датасету / Frequently used 3-grams from the fake news dataset

```
bigrams = (pd.Series(nltk.ngrams(tokens_clean_true, 2)).value_counts())
print(bigrams[:10])
```

```
(united, state)       23395
(donald, trump)        4644
(white, house)         3852
(president, donald)    2699
(north, korea)         2682
(prime, minister)      1969
(official, say)        1959
(say, statement)       1874
(say, would)           1717
(told, reuters)        1706
Name: count, dtype: int64
```

Рис. 5. Частовживані біграми із нефейкових новин датасету / Frequently used 2-grams from the non-fake news dataset

```
trigrams = (pd.Series(nltk.ngrams(tokens_clean_true, 3)).value_counts())
print(trigrams[:10])
```

```
(president, donald, trump)    2668
(united, state, president)    1592
(state, president, donald)    1123
(president, barack, obama)     917
(say, united, state)           830
(united, state, senator)       714
(united, state, official)      643
(united, state, senate)        581
(united, state, house)         507
(white, house, say)            484
Name: count, dtype: int64
```

Рис. 6. Частовживані триграми із нефейкових новин датасету / Frequently used 3-grams from the non-fake news dataset

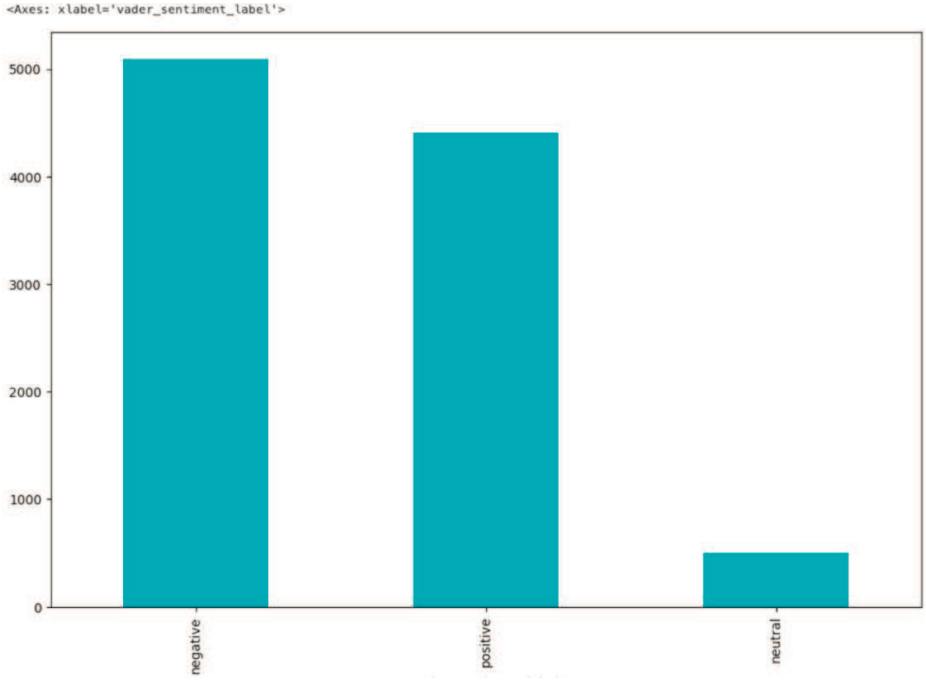

Рис. 7. Розподіл міток настрою для фейкових новин датасету / Distribution of sentiment labels for fake news dataset

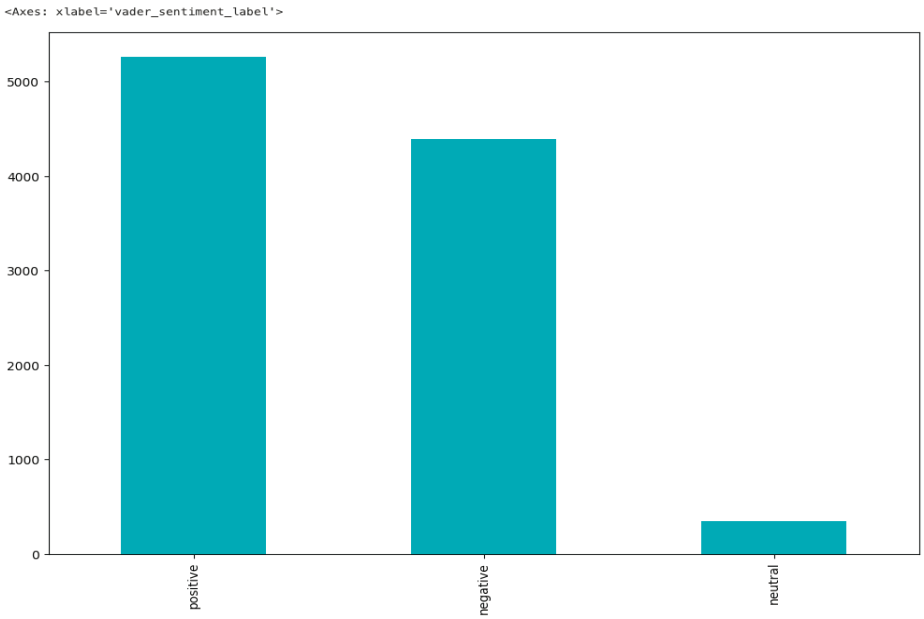

Рис. 8. Розподіл міток настрою для нефейкових новин датасету / Distribution of sentiment labels for fake news dataset

```
result_input.head()
```

	title	text	subject	date	True	text_clean	vader_sentiment_score	vader_sentiment_label
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017	1	[head, conservative, republican, faction, unit...	0.9847	positive
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017	1	[transgender, people, allow, first, time, enli...	-0.5363	negative
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017	1	[special, counsel, investigation, link, russia...	-0.6808	negative
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017	1	[trump, campaign, adviser, george, papadopoulo...	-0.2201	negative
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017	1	[president, donald, trump, call, united, state...	0.6683	positive

Рис. 9. Датасет після очищення та аналізу настроїв / Dataset after cleaning and sentiment analysis

```
print(classification_report(y_test, y_pred_lr))
```

	precision	recall	f1-score	support
0	0.98	0.99	0.98	2971
1	0.99	0.98	0.98	3029
accuracy			0.98	6000
macro avg	0.98	0.98	0.98	6000
weighted avg	0.98	0.98	0.98	6000

Рис. 10. Результати моделі машинного навчання / Results of the machine learning model

Оцінка *F*1 становить 0,98 для обох класів (0 – підробка, 1 – не підробка). Такі хороші результати можна пояснити "лабораторною" якістю набору даних. У подальших експериментах плануємо зосередитися на перевірці моделі на новинах у реальному часі.

Обговорення результатів дослідження. Розроблено систему підтримки прийняття рішень стосовно виявлення дезінформації, фейків та пропаганди на основі машинного навчання. Досліджено методику аналізу тексту новин для ідентифікації фейку та передбачення виявлення дезінформації в текстах новин. Для експерименту ми використали набір даних із 20 000 статей: 10 000 записів для фейкових новин і 10 000 для нефейкових. Більшість статей пов'язані з політикою. Для обох піднаборів даних виконано основні процедури очищення тексту, такі як зміна тексту на малі літери, видалення знаків пунктуації, очищення тегів розташування та автора, а також видалення стоп-слів тощо. Після очищення виконано токенізацію та лематизацію. Для поліпшення результатів лематизації кожен токен позначено тегом POS. Використання тегів POS допомагає точніше виконувати лематизацію. Для обох піднаборів даних ми створили біграми та триграми, щоб краще зрозуміти контекст статей у наборі даних. Виявлено, що у нефейкових новинах використовується офіційніший мовний стиль. Проаналізовано настрої в обох піднаборах даних. З'ясовано, що фальшивий субнабір даних містить більше негативних балів, тоді як нефальшивий – переважно позитивні оцінки. Піднабори даних були об'єднані перед створенням моделі прогнозування. Для моделі прогнозу використано функції BOW і Logistic Regression. Оцінка F1 становить 0,98 для обох класів фейк / не фейк.

Отже, за результатами виконаної роботи можна сформулювати наукову новизну та практичну значущість результатів дослідження.

Наукова новизна отриманих результатів дослідження – розвинено методи розпізнавання дезінформації, фейків та пропаганди на основі методів BOW і Logistic Regression.

Практична значущість результатів дослідження – досягнуто показників точності розпізнавання 98 %.

Висновки / Conclusions

На сучасному етапі існують різноманітні методи виявлення фейкових новин. Перспективний напрям – використання моделей машинного навчання, які можна інтегрувати в інформаційні портали / системи моніторингу новин для ідентифікації дезінформації у текстах. Такий підхід може істотно знизити поширення дезінформації, покращити якість інформації для користувачів та збільшити довіру до ЗМІ.

У світі інформаційного суспільства, коли велика кількість інформації постійно доступна через різні комунікаційні канали, проблема фейкових новин стає надзвичайно актуальною. Вони можуть впливати на громадську думку, політичні рішення та міжнародні відносини. Отже, створення моделей машинного навчання для виявлення фейкових новин у текстах – ключове завдання для наукових досліджень та практичного використання.

У сучасному світі, де соціальні мережі – нелімітоване джерело інформації та впливу, проблема дезінформації стає надзвичайно актуальною. Завдяки швидкому поширенню новин та персональній рекомендаційній системі соціальні мережі стали не лише майданчиком для обміну інформацією, але й потужним інструментом маніпуляції громадською думкою. Саме тут стає важливим створення систем виявлення дезінформації.

По-перше, системи виявлення дезінформації є необхідним інструментом для збереження довіри користувачів до соціальних мереж. Завдяки розвитку технологій глибокого навчання та аналізу великих даних з'явилася можливість автоматизованого виявлення шаблонів та характеристик дезінформації. Це дає змогу оперативно розпізнавати та вилучати недостовірну інформацію, забезпечуючи кращу якість контенту та захищаючи громадську думку.

По-друге, системи виявлення дезінформації допомагають у збереженні демократичних процесів. Особливо

це актуально у час виборів, коли поширення фейкових новин може істотно вплинути на виборчі рішення та стабільність суспільства. Забезпечення достовірності інформації у соціальних мережах стає невід'ємною частиною забезпечення чесності та прозорості в політичних процесах.

По-третє, системи виявлення дезінформації дають змогу зберігати безпеку та відсіювати загрози для громадської безпеки. Враховуючи роль соціальних мереж у мобілізації суспільства та поширенні екстремістських ідей, важливо вчасно розпізнавати та придушувати небезпечний контент. Це сприяє підтриманню стабільності та безпеки інтернет-спільноти.

На основі розробленої системи підтримки прийняття рішень щодо виявлення дезінформації, фейків та пропаганди з використанням машинного навчання експериментально апробовано реалізовані модулі виявлення фейку на основі методів BOW і Logistic Regression. Метрика F1 становить 0,98 для класу 0 –підробка та класу 1 – не підробка. Таке високе значення метрики F1 можна пояснити збалансованістю датасету та високою якістю набору даних від експертів. У подальших експериментах плануємо зосередитися на перевірці моделі на новинах у реальному часі з різних джерел інформації, особливо з інтернету – із соціальних мереж.

Подяка

Статтю підготовано завдяки грантовій підтримці 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".

References

1. Tyshchenko, V., & Muzhanova, T. (2022). Disinformation and fake news: features and methods of detection on the internet. *Cybersecurity: education, science, technique,* 2(18), 175–186. https://doi.org/10.28925/2663-4023.2022.18.175186
2. Myronyuk, O. (2024). Misinformation: how to recognize and combat it [Misinformation: how to recognize and combat it]. Retrieved from: https://law.chnu.edu.ua/dezinformatsiia-yak-rozpiznaty-ta-borotysia/
3. Reuter, C., Hartwig, K., Kirchner, J., & Schlegel, N. (2019). Fake news perception in Germany: A representative study of people's attitudes and approaches to counteract disinformation. Retrieved from: https://aisel.aisnet.org/wi2019/track09/papers/5/
4. Luchko, Y. I. (2023). The role of artificial intelligence technologies in spreading and combating disinformation [The role of artificial intelligence technologies in spreading and combating disinformation]. Countermeasures against disinformation in the conditions of Russian aggression against Ukraine: challenges and prospects: theses addendum. participants of the international science and practice conf. (Ann Arbor – Kharkiv, December 12–13, 2023), 104–106. https://doi.org/10.32782/PPSS.2023.1.26
5. Комісарів, М. (2023). Проблема поширення недостовірної інформації у ЗМІ та соціальних медіа. Retrieved from: https://www.osce.org/representative-on-freedomof-media.
6. Marchi, R. (2012). With facebook, blogs, and fake news, teens reject journalistic "objectivity". *Journal of communication inquiry,* 36(3), 246–262. https://doi.org/10.1177/0196859912458700
7. Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. *SIGKDD Explorations Newsletter*, 19(1), 22–36. https://doi.org/10.1145/3137597.3137600
8. Zhou, Z.; Guan, H.; Bhat, M. & Hsu, J. (2019). Fake News Detection via NLP is Vulnerable to Adversarial Attacks. Proceedings of the International Conference on Agents and Artificial Intelligence, 2, 794–800. https://doi.org/10.5220/0007566307940800
9. Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, Cass. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). *The science of fake news. Science, 359*(6380), 1094–1096. https://doi.org/10.1126/science.aao299
10. Vosoughi, S, Roy, D, & Aral, S. (2018). The spread of true and false news online. *Science, 359*, 1146–1151. https://doi.org/10.1126/science.aap9559
11. Zuo, C., Karakas, A. I., & Banerjee, R. (2018). A hybrid recognition system for check-worthy claims using heuristics and supervised learning. CEUR workshop proceedings, 2125. Retrieved from: https://ceur-ws.org/Vol-2125/paper_143.pdf
12. Hansen, C., Hansen, C., Simonsen, J. G., & Lioma, C. (2018). The Copenhagen Team Participation in the Check-Worthiness Task of the Competition of Automatic Identification and Verification of Claims in Political Debates of the CLEF-2018 CheckThat! Lab. CEUR Workshop Proceedings, 2125. Retrieved from: https://ceur-ws.org/Vol-2125/paper_81.pdf
13. Thorne, J, & Vlachos, A. (2018). Automated fact checking: Task formulations, methods and future directions. arXiv preprint arXiv:180607687. https://doi.org/10.48550/arXiv.1806.07687.
14. Mihaylova, T, Karadjov, G, Atanasova, P, Baly, R, Mohtarami, M, & Nakov, P. (2019). SemEval-2019 task 8: Fact checking in community question answering forums. arXiv preprint arXiv:190601727. https://doi.org/10.48550/arXiv.1906.01727
15. O'Brien, N. (2018). Machine learning for detection of fake news. Retrieved from: https://dspace.mit.edu/handle/1721.1/1197279
16. Canini, K. R, Suh, B, & Pirolli, P. L. (2011). Finding credible information sources in social networks based on content and social structure. International Conference on Privacy, Security, Risk and Trust and International Conference on Social Computing, Boston, MA, USA, 1–8. https://doi.org/10.1109/PASSAT/SocialCom.2011.91
17. Derczynski, L., Bontcheva, K., Liakata, M., Procter, R., Hoi, G. W. S., & Zubiaga, A. (2017). SemEval–2017 Task 8: RumourEval: Determining rumour veracity and support for rumours. arXiv preprint arXiv:170405972. https://doi.org/10.48550/ arXiv.1704.05972
18. Baly, R., Karadzhov, G., Alexandrov, D., Glass, J., & Nakov, P. (2018). Predicting factuality of reporting and bias of news media sources. arXiv preprint arXiv:181001765. https://doi.org/10.48550/ arXiv.1810.01765
19. Hardalov, M., Koychev, I., & Nakov, P. (2016). In Search of Credible News. In: Dichev, C., Agre, G. (eds) Artificial Intelligence: Methodology, Systems, and Applications. AIMSA 2016. Lecture Notes in Computer Science, 9883. Springer, Cham. https://doi.org/10.1007/978-3-319-44748-3_17
20. Enayet, O., & El-Beltagy, S. R. (2017). NileTMRG at SemEval-2017 task 8: Determining rumour and veracity support for rumours on Twitter. https://doi.org/10.18653/v1/S17-2082
21. Juola, P. (2012). An Overview of the Traditional Authorship Attribution Subtask. CEUR Workshop Proceedings, 1178. Retrieved from: https://ceur-ws.org/Vol-1178/CLEF2012wn-PAN-Juola2012.pdf.
22. Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology,* 60(3), 538–556. https://doi.org/10.1002/asi.21001
23. Popat, K., Mukherjee, S., Strötgen, J., & Weikum, G. (2017). Where the truth lies: Explaining the credibility of emerging claims on the web and social media. Proceedings of the 26th in-

ternational conference on world wide web companion, 1003–1012. https://doi.org/10.1145/3041021.3055133
24. Aphiwongsophon, S., & Chongstitvatana, P. (2018). Detecting fake news with machine learning method. International conference on electrical engineering/electronics, computer, telecommunications and information technology (ECTI-CON), 528–531. https://doi.org/10.1109/ECTICon.2018.8620051
25. Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., & Stein, B. (2017). A stylometric inquiry into hyperpartisan and fake news. arXiv preprint arXiv:1702.05638. https://doi.org/10.48550/arXiv.1702.05638
26. Koppel, M., Schler, J., & Bonchek-Dokow, E. (2007). Measuring differentiability: Unmasking pseudonymous authors. *Journal of Machine Learning Research,* 8(6). Retrieved from: https://www.jmlr.org/papers/volume8/koppel07 a/koppel07 a.pdf.
27. Horne, B., & Adali, S. (2017). This Just In: Fake News Packs A Lot In Title, Uses Simpler, Repetitive Content in Text Body, More Similar To Satire Than Real News. Proceedings of the International AAAI Conference on Web and Social Media, 11(1), 759–766. https://doi.org/10.1609/icwsm.v11i1.14976
28. Jain, A., Shakya, A., Khatter, H., & Gupta, A. K. (2019). A smart system for fake news detection using machine learning. In 2019 International conference on issues and challenges in intelligent computing techniques (ICICT), 1, 1–4. https://doi.org/ 10.1109/ICICT46931.2019.8977659
29. Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., & Choi, Y. (2017). Truth of varying shades: Analyzing language in fake news and political fact-checking. Proceedings of the conference on empirical methods in natural language processing, 2931–2937. https://doi.org/10.18653/v1/D17-1317
30. Shao, C., Ciampaglia, G. L., Varol, O., Flammini, A., & Menczer, F. (2017). The spread of fake news by social bots. arXiv preprint arXiv:1707.07592, 96(104), 14. Retrieved from: ttps://cs.furman.edu/~tallen/csc271/source/viralBot.pdf
31. Horne, B., Khedr, S., & Adali, S. (2018). Sampling the News Producers: A Large News and Feature Data Set for the Study of the Complex Media Landscape. Proceedings of the International AAAI Conference on Web and Social Media, 12(1). https://doi.org/10.1609/icwsm.v12i1.14982
32. Mahir, E. M., Akhter, S., & Huq, M. R. (2019). Detecting fake news using machine learning and deep learning algorithms. International conference on smart computing & communications (ICSCC), 1–5. IEEE. https://doi.org/10.1109/ICSCC.2019.8843612
33. Da San Martino, G., Seunghak, Y., Barrón-Cedeno, A., Petrov, R., & Nakov, P. (2019). Fine-grained analysis of propaganda in news article. Proceedings of the conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP–IJCNLP), 5636–5646. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1565
34. Nouh, M., Nurse, J. R., & Goldsmith, M. (2019, July). Understanding the radical mind: Identifying signals to detect extremist content on twitter. International conference on intelligence and security informatics (ISI). *IEEE,* 98–103. https://doi.org/10.1109/ISI.2019. 8823548
35. Barrón-Cedeño, A., Jaradat, I., Da San Martino, G., & Nakov, P. (2019). Proppy: Organizing the news based on their propagandistic content. Retrieved from: https://wwwsciencedirectcom/science/article/abs/pii/S0306457318306058:16. https://doi.org/10.1016/j.ipm.2019.03.005
36. Oliinyk, V. A., Vysotska, V., Burov, Y., Mykich, K., & Basto-Fernandes, V. (2020). Propaganda Detection in Text Data Based on NLP and Machine Learning. CEUR Workshop Proceedings, 2631, 132–144. Retrieved from: https://ceur-ws.org/Vol-2631/paper10.pdf
37. Altiti, O., Abdullah, M., & Obiedat, R. (2020). Just at semeval-2020 task 11: Detecting propaganda techniques using bert pre-trained model. Proceedings of the Fourteenth Workshop on Semantic Evaluation, 1749–1755. https://doi.org/10.18653/v1/2020.semeval-1.229
38. Han, Y., Karunasekera, S., & Leckie, C. (2020). Graph neural networks with continual learning for fake news detection from social media. arXiv preprint arXiv:2007.03316. https://doi.org/10.48550/arXiv.2007.03316
39. Polonijo, B., Šuman, S., & Šimac, I. (2021). Propaganda detection using sentiment aware ensemble deep learning. *International Convention on Information, Communication and Electronic Technology,* 199–204. https://doi.org/10.23919/MIPRO52101.2021.9596654
40. Sprenkamp, K., Jones, D. G., & Zavolokina, L. (2023). Large language models for propaganda detection. arXiv preprint arXiv:2310.06422. https://doi.org/10.48550/arXiv.2310.06422
41. Li, W., Li, S., Liu, C., Lu, L., Shi, Z., & Wen, S. (2022). Span identification and technique classification of propaganda in news articles. *Complex & Intelligent Systems,* 8(5), 3603–3612. https://doi.org/10.1007/s40747-021-00393-y
42. Martseniuk, M., Kozachok, V., Bohdanov, O., & Brzhevska, Z. (2023). Analysis of methods for detecting misinformation in social networks using machine learning. *Electronic Professional Scientific Journal "Cybersecurity: Education, Science, Technique",* 2(22), 148–155. https://doi.org/10.28925/2663-4023.2023.22.148155
43. Ravichandiran, S. (2021). Getting Started with Google BERT: Build and train state-of-the-art natural language processing models using BERT. Packt Publishing Ltd.
44. Xiao, Y., & Jin, Z. (2021). Summary of research methods on pre-training models of natural language processing. *Open Access Library Journal, 8*(7), 1–7. https://doi.org/10.4236/oalib.1107602
45. Ibrahim, M., & Murshed, M. (2016). From tf-idf to learning-to-rank: An overview. Handbook of research on innovations in information retrieval, analysis, and management, 62–109. https://doi.org/10.4018/978-1-4666-8833-9.ch003
46. Omar, M., Choi, S., Nyang, D., & Mohaisen, D. (2022). Robust natural language processing: Recent advances, challenges, and future directions. *IEEE Access*, 10, 86038–86056. https://doi.org/10.1109/ACCESS.2022.3197769
47. Verma, P. K., Agrawal, P., & Prodan, R. (2021). WELFake Dataset for Fake News Detection in Text Data (Version: 0.1) [Data Set]. Genéve, Switzerland: Zenodo.

V. A. Vysotska, R. V. Romanchuk
Lviv Polytechnic National University, Lviv, Ukraine

DECISION SUPPORT SYSTEM FOR DISINFORMATION, FAKES AND PROPAGANDA DETECTION BASED ON MACHINE LEARNING

Due to the simplification of the processes of creating and distributing news via the Internet, as well as due to the physical impossibility of checking large volumes of information circulating in the network, the volume of disinformation and fake news distribution has increased significantly. A decision support system for identifying disinformation, fakes and propaganda based on machine learning has been built. The method of news text analysis for identifying fakes and predicting the detection of disinformation in news texts has been studied. Due to the simplification of the processes of creating and distributing news via the Internet, as well as due to the physical impossibility of checking large volumes of information circulating in the network, the volume of disinformation and fake news distribution has increased significantly. In this regard, detection of fake news becomes a critical task. This not only ensures the provision of verified and reliable information to users, but also helps prevent manipulation of public consciousness. Strengthening control over the credibility of news is important for maintaining a reliable ecosystem of the information space. The combination of IR and NLP allows systems to automatically analyse and track information to detect potential misinformation or fake news. It is also important to consider context, source of information, and other factors to accurately determine credibility. Such automated methods can help in real-time detection and resolution of problems related to the spread of misinformation in social networks. For our experiment, we use a dataset with a total number of 20,000 articles: 10,000 entries for fake news and 10,000 for non-fake news. Most of the articles are related to politics. For both subsets of the data, basic text cleaning procedures such as changing text to lowercase, removing punctuation marks, cleaning location and author tags, and removing stop words, etc., were performed. After cleaning, tokenization and lemmatization were performed. For better lemmatization results, each token is labelled with a POS tag. Using POS tags helps perform lemmatization more accurately. For both subsets of the data, bigrams and trigrams were created to better understand the context of the articles in the data set. It was found that non-fake news uses a more formal language style. Next, we performed sentiment analysis on both subsets of the data. The results show that the fake sub-dataset contains more negative scores, while the non-false sub-dataset has mostly positive scores. Subsets of the data were combined before building the prediction model. BOW and Logistic Regression functions were used for the forecast model. The F1 score is 0.98 for both fake/non-fake classes.

Keywords: disinformation recognition, fake recognition, natural language processing, machine learning, news classification.

Інформація про авторів:

Висоцька Вікторія Анатоліївна, д-р техн. наук, професор, кафедра інформаційних систем та мереж.
Email: victoria.a.vysotska@lpnu.ua; https://orcid.org/0000-0001-6417-3689

Романчук Роман Васильович, аспірант, кафедра інформаційних систем та мереж.
Email: roman.v.romanchuk@lpnu.ua; https://orcid.org/0009-0004-4352-1073

Цитування за ДСТУ: Висоцька В. А., Романчук Р. В. Система підтримки прийняття рішень виявлення дезінформації, фейків та пропаганди на основі машинного навчання. *Український журнал інформаційних технологій*. 2024, т. 6, № 2. С. 105–116.

Citation APA: Vysotska, V. A., & Romanchuk, R. V. (2024). Decision support system for disinformation, fakes and propaganda detection based on machine learning. *Ukrainian Journal of Information Technology, 6*(2), 105–116. https://doi.org/10.23939/ujit2024.02.105