https://doi.org/10.31891/2219-9365-2025-82-26
УДК 004.85:004.912:025.4.03

ШУПТА Андрій
Хмельницький національний університет
https://orcid.org/0009-0000-9771-5579
e-mail: andrii.shupta@gmail.com

МЕТОД АДАПТИВНОГО ВИЯВЛЕННЯ ФЕЙКОВИХ НОВИН НА ОСНОВІ УЗАГАЛЬНЕНОГО ВЕКТОРА ТЕКСТОВИХ ОЗНАК

Фейкові новини стали серйозною проблемою останніми роками, оскільки вони можуть швидко поширюватися через соціальні мережі та інші онлайн-платформи. Для виявлення фейкових новин можуть використовуватися різні методи та матеріали. Один підхід полягає в аналізі змісту новин, включаючи текст та супровідні зображення чи відео. Інший підхід передбачає врахування соціального контексту, в якому поширюються новини, наприклад, джерело новин та настрої людей, які ними діляться. У цій роботі представлено метод адаптивного виявлення фейкових новин за допомогою обробки природної мови. Пропонується використовувати вектор ознак, що побудований на основі узагальнених характеристик текстів новин. Також пропонується можливість розширення вектора ознак та наборів навчальних даних для адаптації класифікатора до нових типів фейкових новин. Продемонстровані експериментальні результати якісно (візуальна аналітика) та кількісно (статистичні метрики) демонструють здатність запропонованого методу виявляти фейкові новини з достатньою якістю (90%). Дослідження має на меті сприяти розробці точної та надійної системи для виявлення фейкових новин, що дасть змогу зменшити негативний вплив цієї проблеми в сучасному суспільстві.

Ключові слова: детектування фейкових новин, обробка природної мови, узагальнені текстові ознаки, машинне навчання, аналіз контенту.

SHUPTA Andrii
Khmelnytskyi National University

METHOD OF ADAPTIVE DETECTING FAKE NEWS BASED ON A GENERALIZED VECTOR OF TEXTUAL FEATURES

The rapid spread of "fake news" via social media and online platforms poses a significant threat to informed public discourse and trust in information. While existing detection methods analyze content (text, images) or social context (source, sharer sentiment), they often struggle to adapt to the evolving, sophisticated tactics of misinformation campaigns, losing efficacy as new deceptive forms emerge. This paper presents an innovative, adaptive Natural Language Processing framework designed to tackle this dynamic challenge. Our core strategy involves a feature vector built from generalized textual characteristics, capturing enduring linguistic patterns and structural irregularities indicative of fabricated content, rather than superficial, easily outdated markers. A key aspect is the system's designed evolvability: it supports continuous expansion of this feature vector and retraining of the classifier with new datasets. This ensures sustained responsiveness and effectiveness against novel fake news iterations in a constantly changing information landscape. The system's efficacy is validated through a dual evaluation: qualitative visual analytics offer insights into its decision-making, while quantitative statistical metrics (precision, recall, F1-score) confirm its robustness. Experimental results demonstrate a commendable detection accuracy of approximately 90%, underscoring the power of the generalized features and adaptive learning. Ultimately, this research contributes to the critical development of a more reliable, accurate, and dynamically responsive system for identifying and mitigating the spread of fake news. The development of such sophisticated tools holds profound implications for safeguarding the integrity of information, fostering media literacy, and addressing one of the most pressing informational challenges in contemporary society.

Keywords: fake news detection, natural language processing, generalized text features, machine learning, content analysis.

Стаття надійшла до редакції / Received 16.04.2025
Прийнята до друку / Accepted 11.05.2025

ПОСТАНОВКА ПРОБЛЕМИ У ЗАГАЛЬНОМУ ВИГЛЯДІ ТА ЇЇ ЗВ'ЯЗОК ІЗ ВАЖЛИВИМИ НАУКОВИМИ ЧИ ПРАКТИЧНИМИ ЗАВДАННЯМИ

Фейкові новини стали серйозною проблемою в сучасному суспільстві, оскільки вони можуть швидко поширюватися через соціальні медіа та інші онлайн-платформи, впливаючи на думки та переконання людей. Виявлення фейкових новин стало важливим завданням, яке потребує використання різноманітних методів та технік для точної ідентифікації неправдивої або оманливої інформації [1].

Існуючі методи виявлення фейкових новин часто навчаються на даних, які є доступними на момент навчання моделей штучного інтелекту, що може бути непридатним для майбутніх подій [2]. Це пов'язано з тим, що багато маркованих зразків, які використовуються для навчання на перевірених фейкових новинах, можуть швидко застаріти з появою нових подій [3]. Крім того, швидкість поширення новин в Інтернеті та постійна еволюція технологій створення фейкових новин створюють додаткові виклики для моделей виявлення [4]. Отже, метою цієї роботи є підвищення точності виявлення фейкових новин через новий метод, який може пристосовуватися до мінливого характеру фейкових новин. Запропонований метод повинен передбачати можливість перенавчання на нових даних та використання попередніх результатів виявлення.

Метод адаптивного виявлення фейкових новин

Індустрія створення фейкових новин постійно розвивається, тому існує потреба в підході, який дав би експерту змогу аналізувати текст за наявними характеристиками, а також надавав інструменти для додавання нових характеристик та перенавчання класифікаторів на нових наборах фейкових новин. Для виявлення фейкових новин експерти використовують набір узагальнених характеристик контенту. Як правило, текст перевіряється на наявність помилкових міркувань, емоційно забарвленого контенту, що маніпулює читачем, тощо. Відповідно запропонований метод ґрунтується на аналізі узагальнених характеристик контенту, а не лише самого тексту.

Метод спрямований на класифікацію новинних повідомлень N_i на два класи $C = \{c_{\text{fake}}, c_{\text{real}}\}$ через аналіз їхніх узагальнених текстових ознак. Основою методу є формування k-вимірного вектора ознак $\boldsymbol{F} \in \mathbb{R}^k$ для кожного повідомлення N_i за допомогою функції вилучення ознак $\Phi: N \to \mathbb{R}^k$. Отже,

$$\boldsymbol{F}_i = \Phi(N_i) = \left(f_{i,1}, f_{i,2}, \dots, f_{i,k}\right), \tag{1}$$

де кожна компонента $f_{i,j}$ відображає кількісну міру певної лінгвістичної, семантичної або стилістичної характеристики тексту, деталізованої нижче в підрозділі "Характеристики текстового контенту".

Вектор (1) ознак слугує вхідними даними для моделі машинного навчання, яка навчається розрізняти фейкові та істинні новини. Процес навчання класифікатора $h: \mathbb{R}^k \to C$ (рис. 1, етап 1) починається з підготовки навчального набору даних $D_{\text{train}} = \{(N_i, y_i)\}_{i=1}^{M}$, де $y_i \in C$ – істинна мітка новини.

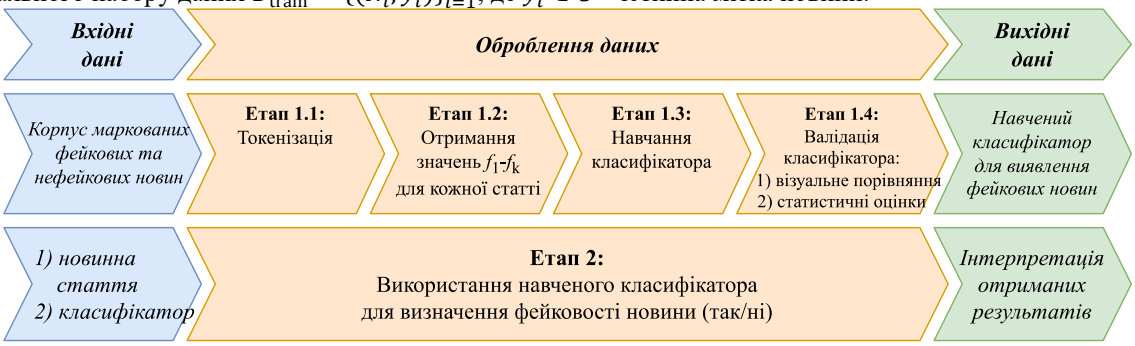


Рис. 1. Схема адаптивного виявлення фейкових новин: етап 1 – метод навчання класифікатора для адаптивного виявлення; етап 2 – процес використання навченої класифікатора

Кожне повідомлення N_i проходить етап попереднього оброблення $\Psi(N_i)$ (токенізація, лематизація, видалення стоп-слів), після чого з нього вилучається вектор ознак $\boldsymbol{F}_i = \Phi\left(\Psi(N_i)\right)$. На основі трансформованого набору $D'_{\text{train}} = \{(\boldsymbol{F}_i, y_i)\}_{i=1}^{M}$ навчається класифікатор, наприклад, за методом опорних векторів (SVM), через мінімізацію відповідної функції втрат $\mathscr{L}$. Валідація моделі проводиться на окремому наборі D_{val} або за допомогою крос-валідації для оцінювання її узагальнюючої здатності за статистичними метриками.

Метод класифікації нового, нерозміченого повідомлення N_{new} (рис. 1, етап 2) полягає в його попередньому обробленні $\Psi(N_{\text{new}})$, вилученню вектора ознак, що формально поданий так

$$\boldsymbol{F}_{\text{new}} = \Phi\left(\Psi(N_{\text{new}})\right), \tag{2}$$

та застосуванні навченого класифікатора h для отримання прогнозованої мітки $\hat{y}_{\text{new}} = h(\boldsymbol{F}_{\text{new}})$ – фейк/істинна.

Адаптивність методу забезпечується двома ключовими механізмами: 1) можливістю розширення вектора ознак $\boldsymbol{F} \to \boldsymbol{F}' \in \mathbb{R}^{k+p}$ через додавання нових p характеристик, що відображають нові аспекти фейковості (з відповідною модифікацією функції $\Phi \to \Phi'$); 2) періодичним перенавчанням класифікатора h на оновленому або розширеному наборі навчальних даних $D_{\text{train}} \cup D_{\text{new\_labeled}}$, що дає можливість моделі враховувати еволюцію тактик створення фейкових новин. У підсумку, результатом роботи методу є визначення, чи належить новина до категорії фейкових або нефейкових.

Для аналізу можливостей запропонованого методу до виявлення фейкових новин використовувалася бібліотека spaCy Python NLP [5], яка включає низку інструментів обробки природної мови. Також використовувалася бібліотека scikit-learn [6] для обчислення багатовимірного шкалювання (MDS) та SVM.

Першим кроком у підготовці тексту для його оброблення є очищення тексту та видалення будь-якої нерелевантної або непотрібної інформації. У дослідженні це включало видалення знаків пунктуації, цифр та стоп-слів. У запропонованому методі для видалення стоп-слів використовується вбудований список стоп-слів

з spaCy. Після очищення текст токенізується за допомогою токенізатора spaCy. Кожному токену присвоюється тег частини мови. Далі лематизатор spaCy використовується для зведення кожного токена до його базової форми.

Характеристики текстового контенту (узагальнені текстові ознаки)

Розглянемо набір текстових характеристик, що використовуються в цьому дослідженні. Ватро зазначити, що він не є фіксованим. Запропонований метод є адаптивним, що дає змогу розширювати як набір текстових характеристик, так і навчальні набори даних.

Вектор ознак $F = (f_1, f_2, \dots, f_{10})$ для десяти узагальнених характеристик контенту формується з компонентів, що описані нижче.

1. Переконання та вплив:

– f_1 – «paraphrased_ratio»: коефіцієнт перефразування, що дає змогу знайти відсоток інформації, яка вже була озвучена, але повторюється з певною метою; вимірюється від 0 до 1;

– f_2 – «dehumanizing_language_ratio»: коефіцієнт "дегуманізації"; обчислювальне вимірювання власних іменників (граматична конструкція) у реченні та розбіжності частини мови; вимірюється від 0 до 1;

– f_3 – «subjective_words_ratio»: коефіцієнт суб'єктивних слів, що показує суб'єктивність тексту; визначається за допомогою компонента spacytextblob; вимірюється від 0 до 1.

2. Наратив:

– f_4 – «header_summary_similarity_ratio»: коефіцієнт подібності заголовка статті та її тіла; визначається через порівняння заголовка та тіла статті за допомогою методу подібності бібліотеки spaCy; виміряно від 0 до 1.

3. Аналіз тональності та лінгвістичний аналіз:

– f_5 – «unusual_inappropriate_language_ratio»: коефіцієнт незвичайної недоречної мови, що показує, скільки незвичайних слів присутньо в тексті; визначається через перевіряння токенів на відповідність стандартним категоріям (is_alpha, is_punct, є у словнику); вимірюється від 0 до 1;

– f_6 – «awkward_text_ratio»: коефіцієнт незграбних, складних або заплутаних конструкцій речень; визначається з урахуванням та відніманням залежностей лінгвістичного тегування в тексті ("amod", "compound", "nsubj", "dobj", "pobj"); вимірюється від 0 до 1;

– f_7 – «avg_sentiment»: коефіцієнт середньої тональності тексту; визначається за допомогою компонента spacytextblob; вимірюється від -1 (негативна) до 1 (позитивна), 0 – нейтральна;

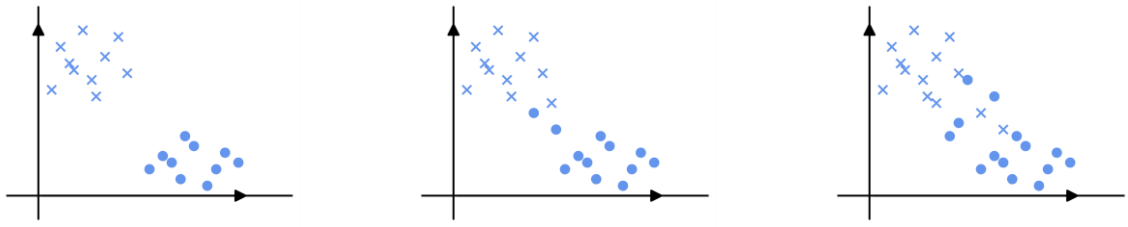
– f_8 – «positive_ratio»: коефіцієнт позитивності тексту; визначається відніманням кількості позитивних слів від усього тексту; вимірюється від 0 до 1;

– f_9 – «neutral_ratio»: коефіцієнт нейтральності тексту; вимірюється від 0 до 1;

– f_{10} – «negative_ratio»: коефіцієнт негативності тексту; вимірюється від 0 до 1.

Оцінка валідності запропонованого вектора ознак

Для оцінювання якості запропонованого вектора ознак для задач класифікації пропонується використовувати MDS. Він призначений для зменшення розмірності до рівня, який можна візуалізувати (3D або 2D). Критерієм зменшення розмірності є евклідова відстань між векторами. Запропоновано візуальні критерії для оцінки якості моделювання (рис. 2).



а) **б)** **в)**

Рис. 2. Якість вектора ознак для задачі класифікації: а) – ідеальна; б) – прийнятна; в) –задовільна

1) Критерій 1 – ідеальний вектор ознак для класифікації тексту; на рис. 2а два класи чітко розділені.

2) Критерій 2 – прийнятний вектор ознак для класифікації тексту; на рис. 2б два класи перетинаються, але об'єкти члени класів є чітко віддаленими.

3) Критерій 3 – задовільний рівень моделі для класифікації тексту; на рис. 2в два класи дещо перекриваються.

Побудований класифікатор оцінено за такими метриками: точність (precision), повнота (recall) та F1-міра.

Результати експериментальних досліджень

У роботі використано набір даних [7], який містить понад 20 000 справжніх та фейкових новин, маркованих та категоризованих. Результати застосування методу MDS до вхідних даних (узагальнених ознак текстів новин) у 2-D просторі показані на рис. 3.

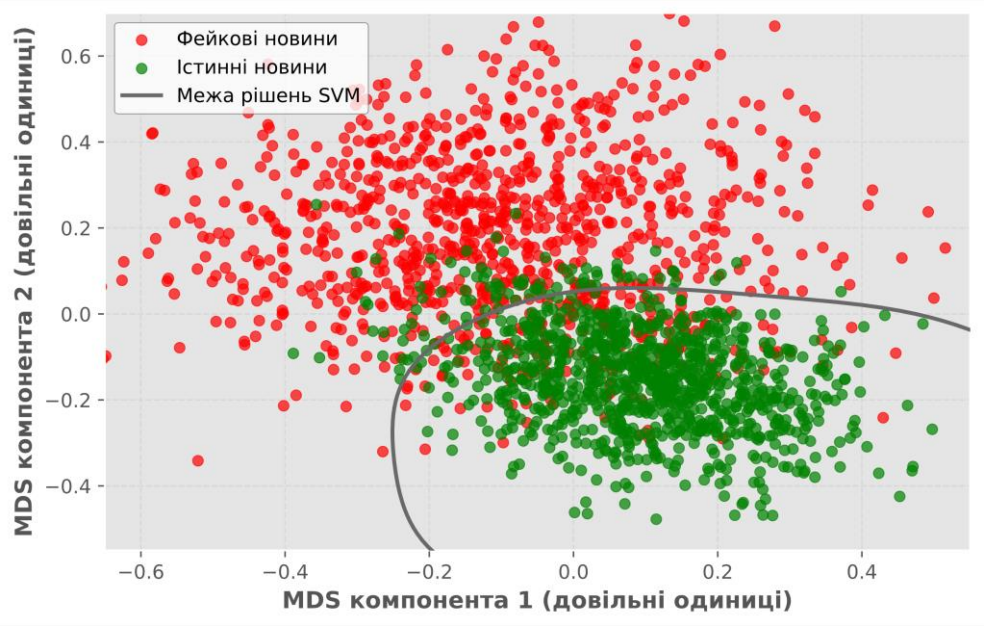


Рис. 3. Результати MDS для 2000 статей
(червоний – фейкові новини, зелений – справжні новини) з межею прийняття рішень SVM

Як видно за відображенням пониження розмірності на рис. 3, результат класифікації є задовільним, класифікація була успішною для більшої кількості текстів з навчального набору. Аналіз невеликої кількості неправильно класифікованих текстів показав, що наявні справжні статті, написані з гіршою якістю тексту, і навпаки. На рис. 3 також показано межу прийняття рішень SVM для 2 000 елементів. Межа визначається опорними векторами, які є точками даних, що розташовані найближче до гіперплощини.

Після обчислення MDS дані були розділені на навчальні та тестові вибірки. З огляду на використання SVM бібліотеки scikit-learn, було отримано наступні результати класифікації, що подані в таблиці 1.

Таблиця 1

Порівняння метрик для задачі класифікації (у %)

Кількість новин	Precision	Recall	F1-score
20	100	100	100
200	88	82	85
2000	93	92	93

Кількісний аналіз (таблиця 1) показує, що для 2 000 новин модель SVM досягла Precision 93%, Recall – 92% та F1-міра – 93%, що свідчить про високу точність запропонованого методу. Якісний аналіз за допомогою MDS (рис. 3) підтверджує задовільну роздільність класів у просторі ознак, хоча й спостерігається невелике перекриття, що вказує на складні випадки, де текстові характеристики фейкових та істинних новин можуть бути подібними. Аналіз помилок часто виявляє фейки, майстерно замасковані під об’єктивні новини, або справжні новини з емоційною лексикою. Проте варто відзначити, що результативність методу залежить від якості навчальних даних, особливо для української мови. Поточний набір з 10 ознак може не охоплювати всі новітні тактики створення фейків.

У майбутньому планується розширення вектора ознак, впровадження автоматизації відбору ознак, розроблення онлайн-навчання та використання накопичувальної адаптації. Потенційними напрямами є дослідження стійкості до змагальних атак, інтеграція мультимодальних даних та розроблення мультимовних моделей, а також застосування методів пояснювального штучного інтелекту до інтерпретації рішень.

ВИСНОВКИ З ДАНОГО ДОСЛІДЖЕННЯ І ПЕРСПЕКТИВИ ПОДАЛЬШИХ РОЗВІДОК У ДАНОМУ НАПРЯМІ

У цій роботі запропоновано метод адаптивного виявлення фейкових новин з використанням засобів оброблення природної мови та алгоритмів машинного навчання. Було проведено ретельний огляд пов’язаних

робіт для забезпечення новизни та ефективності запропонованого методу. Десять різних параметрів (узагальнених текстових ознак) використано для моделювання тексту, а багатовимірне шкалювання (MDS) – застосовано до навченої моделі SVM для отримання візуальної аналітики як одного з критеріїв оцінки якості запропонованого методу. Класифікатор за SVM навчено для класифікації тексту за різними категоріями. Результати дослідження показують, що запропонований метод є на рівні або перевершує існуючі підходи за точністю виявлення фейкових новин (загальна точність – понад 90%). Проте обмеженням методу є брак високоякісних маркованих наборів даних для успішного навчання класифікаторів.

Майбутні вдосконалення методу будуть спрямовані на досягнення більшої інтерпретованості та розуміння результатів класифікації. Це може включати інтеграцію зовнішніх прикладних програмних інтерфейсів для збору більш детальної інформації та перевірки фактів, а також розширення для виявлення контенту, що створено штучним інтелектом.

Література

1. Fake news detection revisited: An extensive review of theoretical frameworks, dataset assessments, model constraints, and forward-looking research agendas / S. Harris et al. *Technologies*. 2024. Vol. 12, no. 11. P. 222. URL: https://doi.org/10.3390/technologies12110222

2. Radiuk P., Pavlova O., Hrypynska N. An ensemble machine learning approach for Twitter sentiment analysis. *Proc. of the 6th Int. Conf. on Comput. Ling. and Intel. Syst. (COLINS 2022). Volume I: Main Conference* : CEUR-Workshop Proceedings, Gliwice, Poland, 12–13 May 2022 / ed. by V. Lytvyn et al. Aachen, 2022. P. 387–397. URL: https://ceur-ws.org/Vol-3171/paper32.pdf

3. An adaptive approach to detecting fake news based on generalized text features / A. Shupta et al. *Proceedings of the 7th Proc. of the 6th Int. Conf. on Comput. Ling. and Intel. Syst. Volume I: Machine Learning Workshop* : CEUR-Workshop Proceedings, Kharkiv, Ukraine, 20–21 April 2023. Aachen, 2023. P. 300–310. URL: https://ceur-ws.org/Vol-3387/paper23.pdf

4. Wierzbicki A., Shupta A., Barmak O. Synthesis of model features for fake news detection using large language models. *Proceedings of the 8th Proc. of the 6th Int. Conf. on Comput. Ling. and Intel. Syst. Volume IV: Computational Linguistics Workshop* : CEUR-Workshop Proceedings, Lviv, Ukraine, 12–13 April 2024. Aachen, 2024. P. 50–65. URL: https://ceur-ws.org/Vol-3722/paper5.pdf

5. Honnibal M., Montani I. SpaCy: Industrial-strength natural language processing in Python. *spacy.io*. URL: https://spacy.io/

6. Scikit-learn: Machine learning in Python / F. Pedregosa et al. *J. Mach. Learn. Res*. 2011. Vol. 12. P. 2825–2830. URL: https://dl.acm.org/doi/10.5555/1953048.2078195 (date of access: 05.05.2025).

7. KaiDMML/FakeNewsNet: This is a dataset for fake news detection research / K. Shu et al. *GitHub*. URL: https://github.com/KaiDMML/FakeNewsNet

References

1. Fake news detection revisited: An extensive review of theoretical frameworks, dataset assessments, model constraints, and forward-looking research agendas / S. Harris et al. *Technologies*. 2024. Vol. 12, no. 11. P. 222. URL: https://doi.org/10.3390/technologies12110222

2. Radiuk P., Pavlova O., Hrypynska N. An ensemble machine learning approach for Twitter sentiment analysis. *Proc. of the 6th Int. Conf. on Comput. Ling. and Intel. Syst. (COLINS 2022). Volume I: Main Conference* : CEUR-Workshop Proceedings, Gliwice, Poland, 12–13 May 2022 / ed. by V. Lytvyn et al. Aachen, 2022. P. 387–397. URL: https://ceur-ws.org/Vol-3171/paper32.pdf

3. An adaptive approach to detecting fake news based on generalized text features / A. Shupta et al. *Proceedings of the 7th Proc. of the 6th Int. Conf. on Comput. Ling. and Intel. Syst. Volume I: Machine Learning Workshop* : CEUR-Workshop Proceedings, Kharkiv, Ukraine, 20–21 April 2023. Aachen, 2023. P. 300–310. URL: https://ceur-ws.org/Vol-3387/paper23.pdf

4. Wierzbicki A., Shupta A., Barmak O. Synthesis of model features for fake news detection using large language models. *Proceedings of the 8th Proc. of the 6th Int. Conf. on Comput. Ling. and Intel. Syst. Volume IV: Computational Linguistics Workshop* : CEUR-Workshop Proceedings, Lviv, Ukraine, 12–13 April 2024. Aachen, 2024. P. 50–65. URL: https://ceur-ws.org/Vol-3722/paper5.pdf

5. Honnibal M., Montani I. SpaCy: Industrial-strength natural language processing in Python. *spacy.io*. URL: https://spacy.io/

6. Scikit-learn: Machine learning in Python / F. Pedregosa et al. *J. Mach. Learn. Res*. 2011. Vol. 12. P. 2825–2830. URL: https://dl.acm.org/doi/10.5555/1953048.2078195

7. KaiDMML/FakeNewsNet: This is a dataset for fake news detection research / K. Shu et al. *GitHub*. URL: https://github.com/KaiDMML/FakeNewsNet