УДК 323.266:004.9
DOI: 10.25140/2411-5363-2020-4(22)-91-95

Володимир Базилевич, Марія Прибитько

СИСТЕМА ВИЯВЛЕННЯ ФЕЙКОВИХ НОВИН ЗА ДОПОМОГОЮ DATA SCIENCE

Актуальність теми дослідження. *На сьогодні актуальним є завдання аналізу правдивості інформації в новинах, якими заповнені всі існуючі канали отримання інформації. Її актуальність пов'язана з необхідністю запобігання паніки через отримання недостовірної інформації, розвінчування псевдонаукових фактів, що можуть загрожувати життю людей, боротьба з політичною пропагандою та інші.*

Постановка проблеми. *Ця робота фокусується на понятті розробки системи виявлення фейкових новин, аналізі існуючих систем та їх принципів роботи, принципів побудови їх алгоритмів та особливостях їх використання.*

Аналіз останніх досліджень та публікацій. *Були розглянуті останні публікації у відкритому доступі, статистичні дані, звіти корпорацій.*

Виділення недосліджених частин загальної проблеми. *Аналіз файлу буде виконаний за допомогою трьох методів/класифікаторів і без використання PassiveAgressive класифікатора. Обчислення та виведення результатів виконується за допомогою побудови матриць помилок та розрахування точності.*

Постановка завдання. *Основною метою роботи є створення на основі розглянутих матеріалів системи виявлення фейкових новин та досягти найбільш можливої точності.*

Виклад основного матеріалу. *Обрано вхідні дані для дослідження, проведена їх підготовка та аналіз. Проведено дослідження даних за допомогою методів/класифікаторів Логістичної регресії, Дерева рішень та Рандомного лісу. Обчислена точність виявлення фейкових новин.*

Висновки відповідно до статті. *Запропонована система дозволяє класифікувати новини як «фейкові» або «правдиві» з точністю 98-99 %.*

Ключові слова*: Python; фейк; фейкова новина; правдива новина; класифікатор.*
Рис.: 6. Табл. 1. Бібл.: 6.

Актуальність теми дослідження. На сьогодні актуальним є завдання аналізу правдивості інформації в новинах, якими заповнені всі існуючі канали отримання інформації. Її актуальність пов'язана з необхідністю запобігання паніки через отримання недостовірної інформації, розвінчування псевдонаукових фактів, що можуть загрожувати життю людей, боротьба з політичною пропагандою та інші.

Постановка проблеми. Відтоді як термін «фейкові новини» потрапив у повне використання та став загальновживаним у 2016 році, ми стали свідками раціонального поширення неправдивої інформації та способів її поширення. У цьому році глядачам та читачам довелося стикнутися з фальшивими відеороликами, що обули подані як дійсно правдиві, блогами, написаними анонімними «тролями», а також сумнівними теоріями, у які повірили навіть у Білому домі США. Це ще не кажучи про численні приклади помилкових або завідомо оманливих даних у ЗМІ або виступах всесвітньовідомих політиків.

Аналіз останніх досліджень і публікацій. Дослідження цієї проблеми проводилися такими вченими та організаціями, як Filip Mishevski, [1], data-flair [2] та інші.

Виділення недосліджених частин загальної проблеми. У цій статті аналіз файлу буде виконаний за допомогою трьох методів/класифікаторів і без використання PassiveAgressive класифікатора. Обчислення та виведення результатів виконується за допомогою побудови матриць помилок та розрахування точності.

Постановка задачі. Проаналізувати існуючі алгоритми та методи виявлення фейкових новин та визначити найбільш ефективний, або їх комбінацію.

Виклад основного матеріалу. «Фейкова новина» визначається як така новина, що «повністю складена і сфабрикована для обману читача, з метою збільшення трафіку і прибутку». Такі новини можна охарактеризувати як елемент інформаційної містифікації або навмисне поширення неправдивих фактів в онлайн- та традиційних медіа або ЗМІ з метою введення споживача в оману або отримання фінансової чи політичної вигоди [3].

Ця робота фокусується на понятті розробки системи виявлення фейкових новин, аналізі існуючих систем та принципів їхньої роботи, принципів побудови їхніх алгоритмів та особливостях їх використання.

Для дослідження фейкових новин та отримання якомога точнішого результату, потрібно мати досить великий обсяг вихідних даних. Причому чим більшу кількість новин вдасться проаналізувати, тим більшу точність результатів можна буде отримати. Серед чималої кількості варіантів був обраний сайт https://www.kaggle.com/, який надає можливість знайти багато сетів даних для майже будь яких аналітичних потреб [4]. У нашому випадку був обраний датасет «Fake and real news dataset». Цей набір даних складається із двох файлів Fake.csv та True.csv, які містять відповідно набір фейкових та набір правдивих новин.

У датасеті використані новини, що були опубліковані з 31 березня 2015 року по 19 лютого 2018 року та з 13 січня 2016 року по 31 грудня 2018 року для файлів Fake.csv та True.csv відповідно

Для зручності подальшого аналізу даних, потрібно підготувати файл таким чином:

Додаємо лейбли “fake” та “true” для файлів.

«Зливаємо» обидва файли в один фінальний.

Перемішуємо дані, щоб отримати випадковий порядок новин у файлі.

Прибираємо колонки title та date як такі, що не несуть цінності для дослідження.

Переводимо текст у lowercase.

Прибираємо пунктуацію.

Видаляємо “stopwords” (це такі слова, які не додають особливого значення реченню. Вони можуть бути безпечно проігноровані без шкоди для змісту речення). Після виконання дій, описаних вище, файл можна вважати готовим для подальшого дослідження.

Перед тим як реалізувати алгоритм пошуку фейкових новин, проведемо первинний аналіз вмісту даних у файлі, адже важливо розуміти не лише з яким обсягом даних ми працюємо, а й з якими саме даними ми маємо справу.

Виведемо у вигляді стовпчастої діаграми кількість фейкових та правдивих новин. Для цього використаємо лейбли “true” та “false”, які були встановлені на початку роботи з файлом. Результат підрахунку представлений на рис. 1.

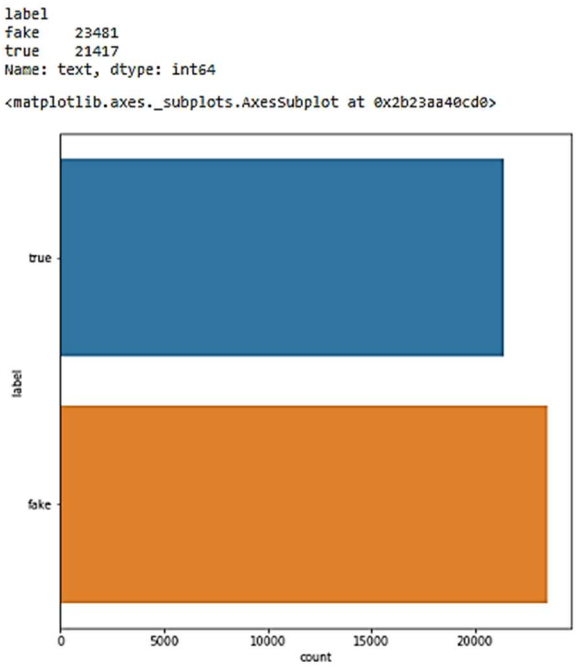

Рис. 1. Результат виведення кількості фейкових та правдивих новин

Вирахуємо також точну кількість найбільш вживаних слів у фейкових та правдивих новинах і виведемо результат, як показано на рис. 2 та 3 відповідно.

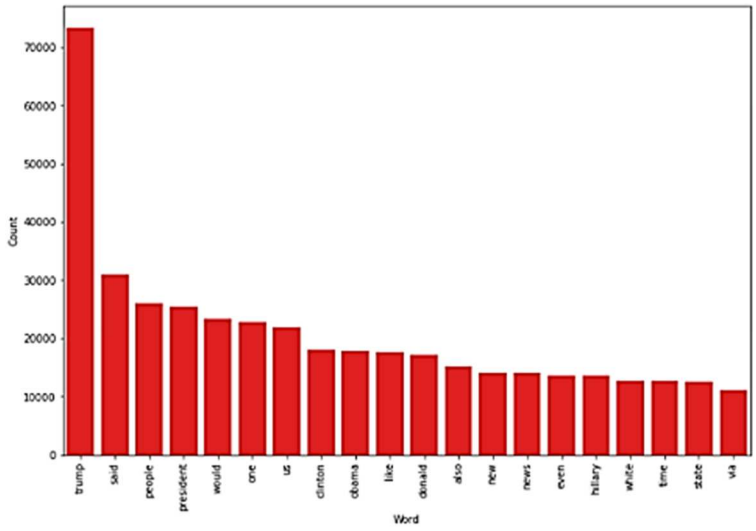

Рис. 2. Результат виведення кількості найбільш вживаних слів у фейкових новинах

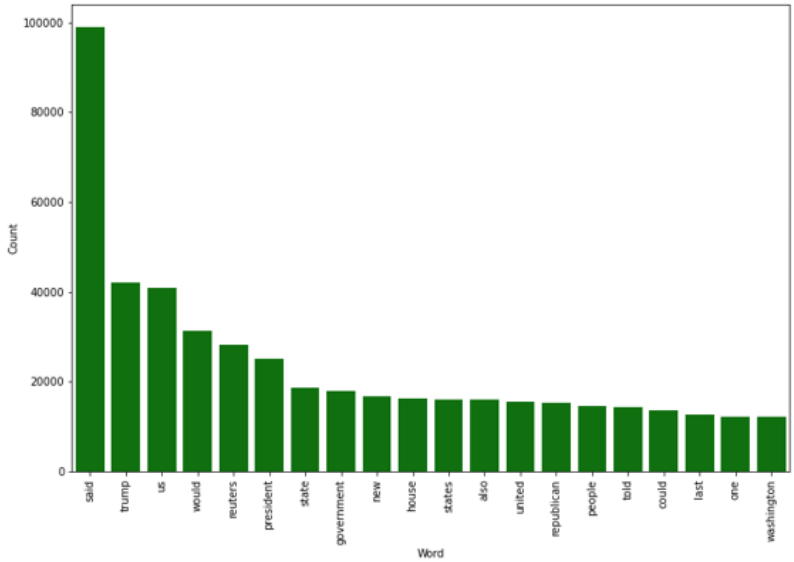

Рис. 3. Результат виведення кількості найбільш вживаних слів у правдивих новинах

Тепер можна зробити припущення про основну тематику новин, що представлені у файлі, а також оцінити належність до певної географічної області (у нашому випадку країни) та опису дій певних груп населення (у нашому випадку можна побачити навіть конкретні прізвища).

Зробивши даний аналіз та розуміючи вміст файлу, можна переходити до створення алгоритму пошуку фейкових новин.

Процес моделювання складається з векторизації корпусу, що зберігається в стовпці „text", потім застосування TF-IDF (frequency-inverse document frequency) і, нарешті, алгоритму класифікації машинного навчання [5].

У контексті даної статті ми виділяємо три методи/класифікатори виявлення, а саме:

1. Логістична регресія.
2. Дерево рішень.
3. Рандомний ліс.

Побудуємо confusion matrix (матриці помилок) моделей для кожного з зазначених вище методів/класифікаторів [6]. На рис. 4 зображена матриця помилок для Логістичної регресії, на рис. 5 – для Дерева рішень та на рис. 6 – Рандомного лісу.

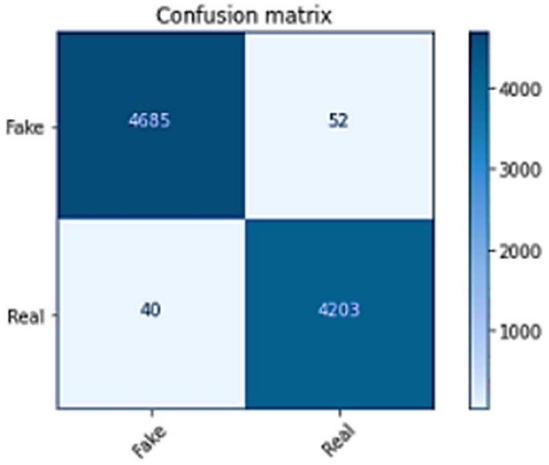

Рис. 4. Матриця помилок (Логістична регресія)

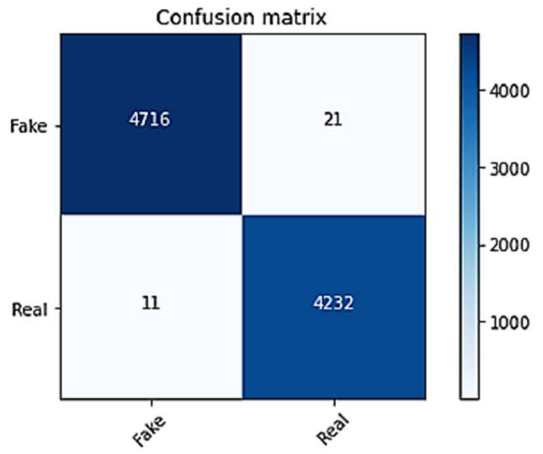

Рис. 5. Матриця помилок (Дерево рішень)

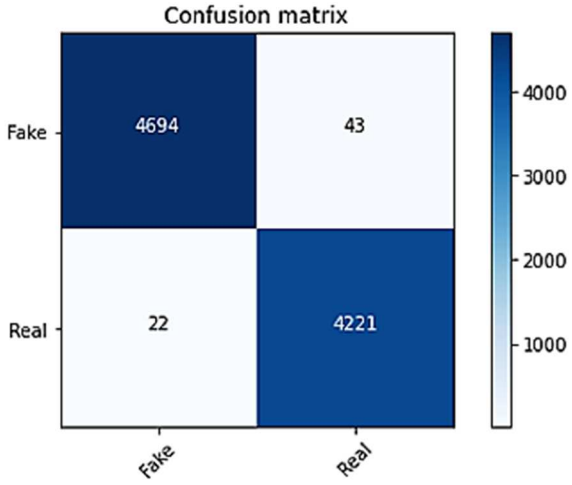

Рис. 6. Матриця помилок (Рандомний ліс)

Висновки відповідно до статті. Для зручності огляду, зробимо порівняльну таблицю отриманих результатів точності. Результат можна представити у вигляді таблиці.

Таблиця

Порівняння результатів обчислення точності за різними методами

Назва методу	**Отримана точність, %**
Логістична регресія	98,98
Дерево рішень	99,64
Рандомний ліс	99,28

Отже, дослідивши існуючі алгоритми та методи виявлення фейкових новин та провівши статистичний аналіз отриманих результатів, можемо зробити висновок, що результат виявлення фейкових новин за допомогою класифікатора Дерева рішень дав найбільший результат.

Список використаних джерел

1. Mishevski F. Detecting Fake News With Python And Machine Learning. URL: https://ichi.pro/ru/obnaruzenie-fejkovyh-novostej-s-pomos-u-python-i-masinnogo-obucenia-272146875193862.

2. Detecting Fake News with Python. URL: https://data-flair.training/blogs/advanced-python-project-detecting-fake-news.

3. Claire Wardle Fake news. It's complicated. URL: https://firstdraftnews.org/latest/fake-news-complicated.

4. Fake and real news dataset. URL: https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset.

5. TF-IDF. URL: https://en.wikipedia.org/wiki/Tf%E2%80%93idf.

6. Confusion matrix. URL: https://scikit-learn.org/stable/auto_examples/model_selection/plot_confusion_matrix.html.

References

1. Mishevski F. Detecting Fake News With Python And Machine Learning. (n.d.). https://ichi.pro/ru/obnaruzenie-fejkovyh-novostej-s-pomos-u-python-i-masinnogo-obucenia-272146875193862.

2. Detecting Fake News with Python. (n.d.). https://data-flair.training/blogs/advanced-python-project-detecting-fake-news.

3. Claire Wardle Fake news. It's complicated. (n.d.). https://firstdraftnews.org/latest/fake-news-complicated.

4. Fake and real news dataset. (n.d.). https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset.

5. TF-IDF. (n.d.). https://en.wikipedia.org/wiki/Tf%E2%80%93idf.

6. Confusion matrix. (n.d.). https://scikit-learn.org/stable/auto_examples/model_selection/plot_confusion_matrix.html.

UDC 323.266: 004.9

Volodymyr Bazylevych, Maria Prybytko

FAKE NEWS DETECTION SYSTEM BASED ON DATA SCIENCE

***Urgency of the research**. Today, the task of analyzing the veracity of information in the news, which filled all existing channels for obtaining information, is relevant. Its urgency is related to the need to prevent panic by obtaining inaccurate information, debunking pseudo-scientific facts that can threaten people's lives, combating political propaganda and others.*

***Target setting** This article focuses on the concept of developing a system for detecting fake news, analysis of existing systems and their principles of operation, principles of construction of their algorithms and features of their use.*

***Actual scientific researches and issues analysis.** Recent open publications, statistics, and corporate reports were reviewed.*

***Uninvestigated parts of general matters defining.** File analysis will be performed using three methods / classifiers and without the use of PassiveAgressive classifier. The calculation and derivation of results is performed by constructing error matrices and calculating accuracy.*

***The research objective.** The main purpose of the work is to create a system for detecting fake news on the basis of the considered materials and to achieve the highest possible accuracy.*

***Presenting main material**. Input data for the study were selected, prepared and analyzed. Data were studied using the methods / classifiers of Logistic Regression, Decision Tree and Random Forest. The accuracy of detecting fake news is calculated.*

***Conclusions.** The proposed system allows to classify news as "fake" or "true" with an accuracy of 98-99 %.*

***Keywords**: Python; fake; fake news; truthful news; classifier.*

Fig .: 6. Table: 1. References: 6.

Базилевич Володимир Маркович – кандидат економічних наук, доцент, завідувач кафедри інформаційних та комп'ютерних систем, Національний університет «Чернігівська політехніка» (вул. Шевченка, 95, м. Чернігів, 14035, Україна).

Bazylevych Volodymyr – PhD in Economics, Associate Professor, Head of the Department of Information and Computer Systems, Chernihiv Polytechnic National University (95 Shevchenka Str., 14035 Chernihiv, Ukraine).

E-mail: bazvlamar@gmail.com

ORCID: http://orcid.org/0000-0001-8935-446X

ResearcherID: G-5764-2014

Scopus Author ID: 57193029322

Прибитько Марія Дмитрівна – здобувачка вищої освіти, Національний університет «Чернігівська політехніка» (вул. Шевченка, 95, м. Чернігів, 14035, Україна).

Prybytko Maria – Phd student, Chernihiv Polytechnic National University (95 Shevchenka Str., 14035 Chernihiv, Ukraine).

E-mail: prybytko.maria@gmail.com

Базилевич В., Прибитько М. Система виявлення фейкових новин за допомогою Data Science. *Технічні науки та технології.* 2020. № 4(22). С. 91-95.