Fake News Detection

A Project Work Synopsis

Submitted in the partial fulfillment for the award of the degree of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE WITH SPECIALIZATION

IN

INFORMATION SECURITY

Submitted by:

CHINNARI ABHISHEK	21BCS3692
MENDA MANMADHA RAO	21BCS3544
PENTAKOTA SRI PRANEETH	21BCS3523
YADALA CHANDU	21BCS11015

Under the Supervision of :

Ms. Komal Mehta (E15888)

CHANDIGARH UNIVERSITY, GHARUAN,

MOHALI, PUNJAB - 140413

March, 2023

Abstract

Keywords - Social Media, Fake News, Scikit-Learn, NLP, Python, Artificial Intelligence, Machine Learning

The widespread dissemination of fake news on social and other media platforms is a major concern due to its potential to cause significant social and national damage. Many researchers are working to detect and combat its spread. This paper analyses research related to fake news detection and explores traditional machine learning models to create a product using a supervised machine learning algorithm that can classify fake news as either true or false. The proposed method utilizes tools such as Python Scikit-Learn and NLP for textual analysis, resulting in feature extraction and vectorization.

The Python Scikit-Learn library is proposed for tokenization and feature extraction of text data because of its useful tools like Count Vectorizer and Tiff Vectorizer. Feature selection methods are then employed to experiment and choose the best-fit features to obtain the highest precision based on the confusion matrix results.

The proposed model utilizes a supervised machine learning algorithm trained on a dataset consisting of true and fake news articles. It classifies new articles as either true or false based on the features extracted from the text data. This enables users to quickly and accurately identify the authenticity of news articles they encounter online.

In summary, this paper contributes to the ongoing efforts to combat fake news by analyzing research related to its detection and proposing an approach that utilizes traditional machine learning models, Python scikit-learn, and NLP for textual analysis. The proposed model achieves high accuracy and precision in detecting fake news, providing a useful tool for users to determine the authenticity of news articles they encounter online.

Table of Contents

1. INTRODUCTION

Fake news refers to false information presented as news, often with the intent of damaging the reputation of individuals or entities or generating advertising revenue. While it was once limited to print media, the advent of social media platforms, particularly the Facebook News Feed, has contributed to its widespread dissemination. Detecting fake news is becoming increasingly difficult because the individuals spreading it are doing so in a highly convincing manner, making it challenging to distinguish from real news. In this paper, we propose a simplistic approach that examines news headlines and attempts to predict whether they are fake or not. While fake news can be a highly effective marketing strategy, it can also be intimidating as it attracts a larger audience than authentic news. It is crucial to recognize that the profits earned from spreading fake news may not justify the harm it causes to individuals and society as a whole.

1.1 Problem Definition

In recent years, social media has become an integral part of people's lives, and it has also become a breeding ground for fake news. False information is rampant in various fields such as politics, democracy, education, finance, and business, putting them at risk. Although fake news is not a new problem, social media's prominence has amplified the spread of deceitful statements and misinformation. In today's world, distinguishing between accurate and misleading news is becoming increasingly challenging, resulting in confusion and complexity. Manual identification of fake news is difficult and requires extensive subject matter expertise. Fake news can cause severe damage to a person's career and can have disastrous effects on nations, citizens, businesses, products, and reputations, especially when it has political motivations.

1.2 Problem Overview

The issue of fake news has led researchers to use both supervised and unsupervised learning algorithms to classify text. However, most studies focus on specific datasets or domains, particularly politics. As a result, algorithms trained on one domain may not work effectively when applied to articles from other domains due to differences in textual structure. To address this problem, we propose a machine-learning ensemble approach for fake news detection. Our approach involves exploring different textual properties that distinguish fake news from real news and training a combination of machine learning algorithms using various ensemble methods that are not well-researched in the current literature. Ensemble learners have been effective in reducing error rates through techniques like bagging and boosting, making it possible to train different machine learning algorithms efficiently and effectively.

1.3 Hardware Specification

1.3.1) RAM – 8GB and above

1.3.2) Graphics Card – 4GB and above

1.3.3) Processor – Intel Core i5 and above

1.4 Software Specification

1.4.1) Python and ML libraries

1.4.2) PyCharm and Jupiter Notebook

1.4.3) NumPy, SciPy.

1.4.4) Pandas, Matplotlib & Sklearn

2. LITERATURE SURVEY

2.1 Existing System

In recent years, there has been extensive research into machine learning methods for detecting deception, with a focus on classifying online reviews and social media posts, as well as identifying "fake news" since the 2016 American presidential election. Conroy, Rubin, and Chen have found that simple content-related n-grams and shallow parts-of-speech tagging are inadequate for classification, as they fail to consider crucial contextual information. Instead, more complex methods, such as deep syntax analysis using probabilistic context-free grammars, are more effective in conjunction with n-gram methods. Feng, Banerjee, and Choi have achieved accuracy rates of 85%–91% in deception-related classification tasks using online review data. Feng and Hirst implemented semantic analysis to detect contradictions between object-descriptor pairs and text, which improved their deep syntax model. Similarly, Rubin, Lukoianova, and Tatiana analysed rhetorical structure using a vector space model and obtained positive results. Ciampaglia et al. utilised language pattern similarity networks, which required a pre-existing knowledge base.

2.2 Proposed System

The problem of detecting fake news has been tackled by both supervised and unsupervised learning algorithms, but the literature mostly focuses on specific domains, such as politics, making it difficult to train a generic algorithm that performs well on articles from all domains. To address this, we propose using an ensemble approach to combine different machine learning algorithms trained on various textual properties that can distinguish fake from real content. We explore ensemble methods that have not been thoroughly explored in the literature but have proven effective in reducing error rates in a wide variety of applications. Our proposed technique is validated through extensive experiments on four publicly available datasets using four commonly used performance metrics: accuracy, precision, recall, and F-1 score. Our results show improved performance compared to existing techniques.

2.3 Literature Review Summary **(Minimum 7 articles should refer)**

Fake news detection has become a critical area of research in recent years due to the widespread dissemination of false and misleading information on social media and other online platforms. Here is a literature review on fake news detection :

1. "Detecting Fake News in Social Media Networks" by Ruchika Gupta, Avinash Tiwari, and Vijendra Singh Sengar (2020): This paper reviews the current state of fake news detection techniques and proposes a novel method for detecting fake news using machine learning algorithms.

2. "Fake News Detection on Social Media: A Data Mining Perspective" by Shu-Yu Chen and Kai-Lung Hua (2019): This paper presents a comprehensive review of various data mining techniques that have been used to detect fake news on social media platforms.

3. "A Survey of Fake News Detection Methods: Algorithms, Evaluations, and Future Directions" by Shaurya Rohatgi and Gaurav Arora (2020): This survey paper reviews the recent advances in fake news detection techniques and evaluates their effectiveness. It also discusses the future directions for fake news detection research.

4. "Fake News Detection on Social Media: A Review" by Chuanren Liu and Huan Liu (2018): This paper reviews the various approaches that have been used to detect fake news on social media, including linguistic and network-based techniques.

5. "Combating Fake News: A Survey on Detection and Mitigation Techniques" by Muhammad Bilal, Sheraz Ahmed, and Zeeshan Ahmed (2020): This survey paper reviews the various detection and mitigation techniques that have been proposed to combat fake news, including deep learning and natural language processing.

6. Automatic Detection of Fake News: A Survey (2020) : This paper provides an overview of the state of the art in automatic fake news detection, including different types of fake news, techniques for feature extraction and classification, and evaluation methods. The authors conclude that there is still room for improvement in this area, particularly in detecting fake news that is difficult to distinguish from real news.

7. Machine Learning-Based Fake News Detection: A Systematic Review (2020) : This systematic review examines the use of machine learning techniques for fake news detection. The authors identify several challenges in this area, such as the lack of labelled datasets and the difficulty in detecting fake news that is based on partially true information. They also find that deep learning techniques, such as neural networks, are becoming increasingly popular in this field.

Overall, these papers demonstrate the importance of developing effective techniques for detecting fake news and highlight the potential of machine learning and data mining techniques for addressing this problem.

3. PROBLEM FORMULATION

The advent of the World Wide Web and the subsequent widespread adoption of social media platforms such as Facebook and Twitter have revolutionized information dissemination. This has resulted in news outlets being able to provide near-real-time news updates to their subscribers. The evolution of the news media from traditional print formats like newspapers, tabloids, and magazines to digital forms like online news platforms, blogs, social media feeds, and other digital media formats has made it easier for consumers to access the latest news at their fingertips. Facebook is responsible for 70% of traffic to news websites. While these social media platforms have the power to allow users to discuss and share ideas and debate over issues like democracy, education, and health, they are also used negatively by certain entities for monetary gain and to spread biassed opinions, manipulate mindsets, and disseminate satire or absurdity, which is commonly referred to as fake news.

4. OBJECTIVES

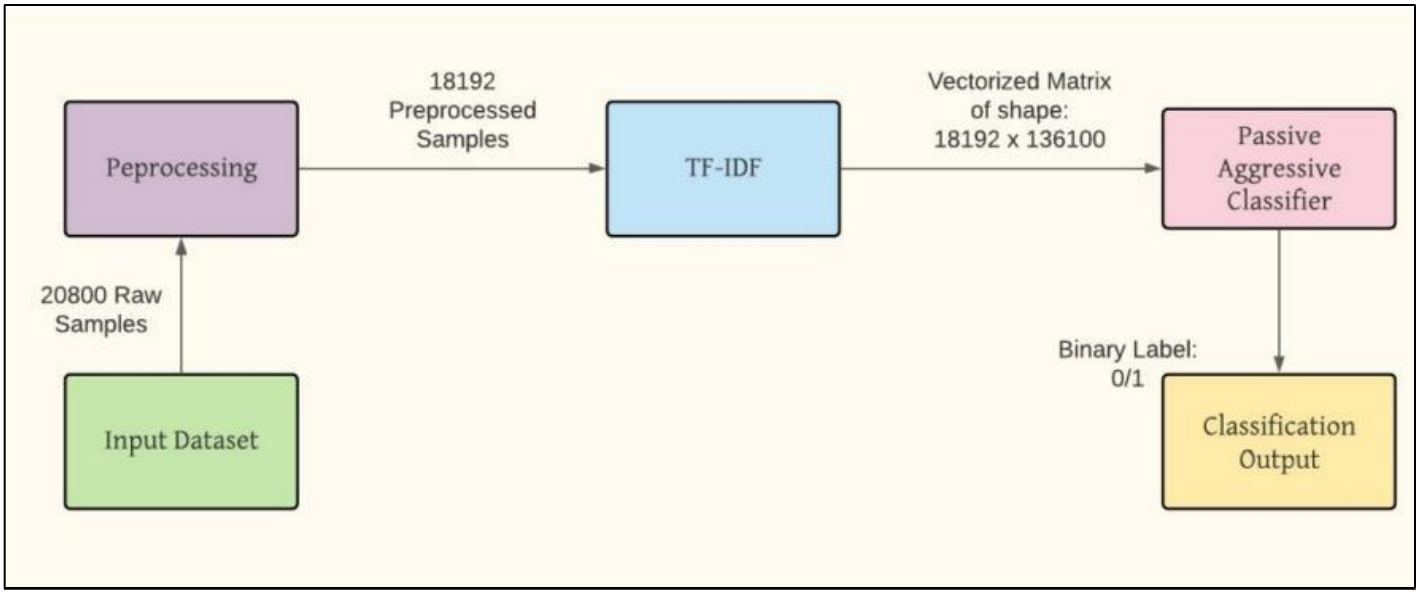
The phenomenon of fake news has existed long before the rise of the internet, but it is currently defined as intentionally fabricated articles meant to deceive readers. Social media and news outlets publish fake news in pursuit of increased readership and as a form of psychological warfare. The ultimate goal is often to profit from clickbait, which uses flashy headlines and designs to entice users to click links and generate advertisement revenue. This paper explores the prevalence of fake news in light of the emergence of social networking sites and seeks to identify a solution to help users detect and filter out sites that contain false and misleading information. The study uses simple and carefully selected features of post titles and content to accurately identify fake posts, achieving 99.4% accuracy using a logistic classifier. However, the detection of fake news still requires the attention of researchers to address key issues, such as identifying the sources of fake news using graph theory and machine learning techniques and developing real-time fake news detection in videos.

The proliferation of fake news has had many negative effects, ranging from political consequences to problems in various other domains, including sports, health, and science. Financial markets have also been affected by fake news, with rumors causing disastrous consequences and potentially bringing markets to a halt. Consumers' worldviews are shaped by the information they consume, and there is growing evidence that consumers have reacted absurdly to news that later proved to be fake. This was particularly evident in the spread of fake news about the novel coronavirus, where false reports led to a worsening of the situation as more people read about it online.

Fortunately, there are several computational techniques that can help identify fake news based on its textual content, including natural language processing and analysis of the propagation of fake news on social networks. However, current resources such as fact-checking websites are limited in their ability to identify fake news across multiple domains and often require human expertise. A more hybrid approach that combines social response analysis with examination of textual features may prove effective in detecting and classifying deceptive articles.

5. METHODOLOGY

The following methodology will be followed to achieve the objectives defined for proposed research work :



Data Collection :

To begin, a dataset containing both real and fake news is required. The proposed system has been tested on a dataset of 6,500 entries previously used by Wang, consisting of a combination of 3,252 fake news stories and 3,259 real news stories.

1. Preprocessing : The data index, which may contain errors, is thoroughly reviewed and corrected to ensure accurate results. Before classifiers are applied, the data is transformed and organised into a proper format. The dataset primarily consists of the English language. To preprocess the data, natural language processing (NLP) techniques are employed, where only English words are selected to enhance accuracy. Following this, text transformations and binary operations are conducted on the dataset to simplify the preprocessing of data.

2. Classifier : After the preprocessed document is obtained, a variety of well-known classifiers are employed to detect features that can identify fake news. These classifiers include support vector machines, K-nearest Neighbours, AdaBoost, Nave Bayesian, neural networks, decision trees, gradient boosting, extreme gradient boosting, random forests, logistic regression, and more.

3. Performance Evaluation : To evaluate the performance of the classifiers, a cross-validation technique is utilised. Specifically, K-fold cross-validation is employed where the dataset is divided into 10 folds. In this step, the model selection function of scikit-learn is used. The Stratified K-Fold subfunction is utilised to split the training dataset into K-folds for cross-validation, and the cross_val_score subfunction is used to observe the cross-validation scores of ML classifiers. Additionally, the GridSearchCV subfunction is utilised to hyper-tune the ML classifiers. After applying all the classifiers, their performance is evaluated based on execution estimations such as test score, ROC score, precision score, recall value, etc. to determine the best classifier.

4. Choosing top 3 Classifiers : The performance of various traditional ML classifiers was evaluated, and the top three were identified. These top three classifiers will undergo further tuning to optimize their output on the dataset. Finally, a voting classifier will be implemented using the best-performing classifiers.

5. Utilizing Ensemble Voting Classifier : The top three classifiers will be used for this voting classification to get the best execution and output.

6. Result : In the final step, the performance of the voting classifier will be evaluated based on various performance metrics such as test score, ROC score, precision score, recall value, etc. The obtained results will be compared with other relevant studies to assess the outcomes.

6.EXPERIMENTAL SETUP

The experiment setup for a fake news detection project typically involves several key steps. First, a dataset of labelled news articles (fake or real) is collected and preprocessed to ensure that the data is in a suitable format for analysis. Then, various feature extraction techniques are applied to the data, such as bag-of-words or word embeddings, to create numerical representations of the text. Next, a machine learning model is trained on the extracted features to classify news articles as either fake or real. To improve the performance of the model, several techniques can be employed, such as hyperparameter tuning, ensemble methods, or using more advanced machine learning models such as neural networks. It is also important to ensure that the model is not overfitting to the training data, so techniques such as cross-validation can be used to estimate the generalization performance of the model.

Overall, the experiment setup for a fake news detection project involves several key steps, including data collection and preprocessing, feature extraction, model training and evaluation, and performance improvement techniques.

7.CONCLUSION

The detection of fake news is a challenging task that necessitates a multifaceted approach. While machine learning algorithms and natural language processing techniques can assist in automatically detecting patterns and characteristics that signify fake news, it is critical to supplement these techniques with human expertise and critical thinking abilities since fake news can be deliberately crafted to deceive these algorithms. Moreover, promoting media literacy and providing education on how to recognise and evaluate sources of information can assist individuals in becoming more astute consumers of news. In conclusion, a combination of technological and human approaches will be required to effectively combat the spread of fake news.

8. TENTATIVE CHAPTER PLAN FOR THE PROPOSED WORK

CHAPTER 1: INTRODUCTION

In this chapter, we provide an overview of our project, which focuses on detecting fake news using machine learning libraries and Python. As fake news is designed to spread false claims in news content, our goal is to identify and verify the accuracy of key claims made in news articles to determine their truthfulness.

CHAPTER 2: LITERATURE REVIEW

This chapter provides an overview of the existing literature on fake news detection. We have used a published research paper to explain the approach we have adopted in our project, which involves using machine learning techniques and Python programming. We have also identified other research papers and blogs that have been helpful in completing our project. Additionally, we have discussed some of the challenges we have encountered and the strategies we have employed to overcome them.

CHAPTER 3: OBJECTIVE

The prevalence of digital news and its widespread use have contributed to the rise in the dissemination of hoaxes and disinformation online. Fake news can be found on various popular platforms, including social media and the Internet. Despite the use of artificial intelligence tools for detecting fake news, the articles are often intended to convince the reader to believe false information, making them difficult to perceive. Moreover, the rapid rate of production of digital news, occurring every second of every day, poses a significant challenge for machine learning algorithms to effectively detect fake news.

CHAPTER 4: METHODOLOGIES

This paper addresses the challenge of distinguishing genuine news from hoaxes on social media, especially when the source of the news is unknown. Instead of relying on related metadata, the paper focuses on detecting fake news solely based on the text features. The proposed approach combines stylometric features with text-based word vector representations using ensemble methods, resulting in an accuracy of up to 95.49% in predicting fake news.

CHAPTER 5: EXPERIMENTAL SETUP

We will be utilising various libraries such as NumPy, SciPy, Scikit-Image, Chumpy, OpenCV, TensorFlow, ipdb, pyrender, trimesh, absl, and Matplotlib in Python 2.7 and Jupiter for our project on fake news detection. These libraries will provide hands-on experience with the Python language.

CHAPTER 6: CONCLUSION AND FUTURE SCOPE

In the digital age, most tasks are performed online, including reading news articles on platforms such as Facebook, Twitter, and online news outlets. However, the rise of fake news has made it challenging to determine the authenticity of news stories and has the potential to influence people's opinions and attitudes towards digital technology. To address this issue, we have developed a fake news detection system that utilises various natural language processing (NLP) and machine learning techniques. The system is trained on a suitable dataset and evaluated using different performance metrics. The best-performing model, the passive aggressive classifier, achieved an accuracy of 87% for static search, which was later improved to 93% using grid search parameter optimization. The proposed approach uses TF-IDF, a passive aggressive classifier, and NLP. Overall, our system helps to classify news headlines and articles as real or fake, thereby combating the spread of misinformation.

REFERENCES

(1) D. Holan, 2016 Lie of the Year: Fake News, PolitiFact, Washington, DC, USA, 2016.

(2) D. M. J. Lazer, M. A. Baum, Y. Benkler et al., “The science of fake news,” Science, vol. 359, no. 6380, pp. 1094–1096, 2018.View at: Publisher Site | Google Scholar

(3) S. Kogan, T. J. Moskowitz, and M. Niessner, “Fake News: Evidence from Financial Markets,” 2019, https://ssrn.com/abstract=3237763.View at: Google Scholar

(4) A. Robb, “Anatomy of a fake news scandal,” Rolling Stone, vol. 1301, pp. 28–33, 2017.View at: Google Scholar

(5) J. Soll, “The long and brutal history of fake news,” Politico Magazine, vol. 18, no. 12, 2016.

(6) N. K. Conroy, V. L. Rubin, and Y. Chen, “Automatic deception detection: methods for finding fake news,” Proceedings of the Association for Information Science and Technology, vol. 52, no. 1, pp. 1–4, 2015.

(7) N. K. Conroy, V. L. Rubin, and Y. Chen, “Automatic deception detection: methods for finding fake news,” Proceedings of the Association for Information Science and Technology, vol. 52, no. 1, pp. 1–4, 2015.

(8) F. T. Asr and M. Taboada, “Misinfotext: a collection of news articles, with false and true labels,” 2019.