UNVEILING MISINFORMATION: DETECTING FAKE NEWS BY CLASSIFICATION METHODS

T Sundararajulu[1], S Guru Sai Prasad[2], K Sai Pranathi[2], N Sathwik[2], E Jaswanth Kumar Reddy[2], C Rohith[2]

[1] *Professor, Department of Computer Science & Information Technology, Siddharth Institute of Engineering & Technology, Andhra Pradesh, India*
[2] *Research Scholar, Department of Computer Science & Information Technology, Siddharth Institute of Engineering & Technology, Andhra Pradesh, India*

ABSTRACT

In the present era, social media platforms such as Facebook, WhatsApp, Twitter, and Telegram are significant sources of information distribution, and people believe it without knowing their origin and genuineness. Social media has fascinated people worldwide in spreading fake news due to its easy availability, cost-effectiveness, and ease of information sharing. Fake news can be generated to mislead the community for personal or commercial gains. It can also be used for other personal benefits such as defaming eminent personalities, amendment of government policies, etc. Thus, to mitigate the awful consequences of fake news, several research types have been conducted for its detection with high accuracy to prevent its fatal outcome. Motivated by the aforementioned concerns, present a comprehensive survey of the existing fake news identification techniques in this project. Then, select Machine Learning (ML) models such as Random Forest (RF), and Naive Bayes (NB), Logistic Regression, Support Vector Machine (SVM) and train them to detect fake news articles on the self-aggregated dataset. Later, we implemented these models by hyper tuning various parameters such as smoothing, drop out factor, and batch size, which has shown promising results in accuracy and other evaluation metrics such as F1-score, recall, precision, and Area under the ROC Curve (AUC) score.

Keyword: *Fake, News, Commercial, Defaming*

1. INTRODUCTION

Fake news is manipulated information that resembles news media content in nature but not in management structure or intent. It is continuously exploded via social media, newspapers, online blogs, forums, and magazines, making it hard to identify reliable news sources. The continuous explosion of fake news increases the need for efficient analytical tools capable of providing insight into the reliability of online content. The false nature of news has a significant impact (negative/positive) on frequent social media users. It must be detected as early as possible to avoid a pessimistic in hence on the readers. Thus, the algorithms and techniques that effectively detect fake news become the focus of intense research. Fake news sources neglect the editorial procedures and standards of the mainstream media to ensure information reliability and trustworthiness.

Fake news primarily draws the attention of the people who are more interested in political talks and stock values and may affect their mental health, which leads to stress, anxiety, and depression-like issues. To mitigate the dissemination of fake news, one should focus on the original stories published by the authorized publishers rather than individual articles. There exist few reports that claims the spread of fake news was in Before Christ (BC) also. But, its wide spreading was initiated with the invention of print media, i.e., the printing press in 1439 Later, the era of social media (Orkut, Facebook, WhatsApp, Twitter, and Telegram) begins in the late 1990s, which has the ability for fast and incredible dissemination of information. It becomes an ideal place for all to create, manipulate, and disseminate fake news. Facebook reported that the malicious actor manipulations accounted for less than one-tenth of 1% of public

content posted on the site. The false rumors on Steve Jobs' health (suffering from a heart attack) reported as authentic had great actuations in the stock exchange of Apple Inc.. For instance, research shows that about *19 million* bot accounts tweeted in support of either Trump or Clinton during the 2016 US presidential election which perfectly demonstrates how social media greatly contributes to the creation and dissemination of fake news.

2. DOMAIN DESCRIPTION

2.1 DATA MINING

The process of extracting information to identify patterns, trends, and useful data that would allow the business to take the data-driven decision from huge sets of data is called Data Mining.

In other words, we can say that Data Mining is the process of investigating hidden patterns of information to various perspectives for categorization into useful data, which is collected and assembled in particular areas such as data warehouses, efficient analysis, data mining algorithm, helping decision making and other data requirement to eventually cost- cutting and generating revenue.

Data mining is the act of automatically searching for large stores of information to find trends and patterns that go beyond simple analysis procedures. Data mining utilizes complex mathematical algorithms for data segments and evaluates the probability of future events.

Data Mining is also called Knowledge Discovery of Data (KDD). Data Mining is a process used by organizations to extract specific data from huge databases to solve business problems. It primarily turns raw data into useful information.

Data Mining is similar to Data Science carried out by a person, in a specific situation, on a particular data set, with an objective. This process includes various types of services such as text mining, web mining, audio and video mining, pictorial data mining, and social media mining. It is done through software that is simple or highly specific. By outsourcing data mining, all the work can be done faster with low operation costs. Specialized firms can also use new technologies to collect data that is impossible to locate manually. There are tones of information available on various platforms, but very little knowledge is accessible. The biggest challenge is to analyze the data to extract important information that can be used to solve a problem or for company development. There are many powerful instruments and techniques available to mine data and find better insight from it.

2.2 MACHINE LEARNING

Machine Learning is said as a subset of artificial intelligence that is mainly concerned with the development of algorithms which allow a computer to learn from the data and past experiences on their own. The term machine learning was first introduced by Arthur Samuel in 1959. We can define it in a summarized way as:
Machine learning enables a machine to automatically learn from data, improve performance from experiences, and predict things without being explicitly programmed.

With the help of sample historical data, which is known as training data, machine learning algorithms build a mathematical model that helps in making predictions or decisions without being explicitly programmed. Machine learning brings computer science and statistics together for creating predictive models. Machine learning constructs or uses the algorithms that learn from historical data. The more we will provide the information, the higher will be the performance.

A machine has the ability to learn if it can improve its performance by gaining more data. A Machine Learning system learns from historical data, builds the prediction models, and whenever it receives new data, predicts the output for it. The accuracy of predicted output depends upon the amount of data, as the huge amount of data helps to build a better model which predicts the output more accurately.

Suppose we have a complex problem, where we need to perform some predictions, so instead of writing a code for it, we just need to feed the data to generic algorithms, and with the help of these algorithms, machine builds the logic as per the data and predict the output. Machine learning has changed our way of thinking about the problem.

2.3 SUPERVISED MACHINE LEARNING

Supervised learning is the types of machine learning in which machines are trained using well "labelled" training data, and on basis of that data, machines predict the output. The labelled data means some input data is already tagged with the correct output. In supervised learning, the training data provided to the machines work as the supervisor that teaches the machines to predict the output correctly. It applies the same concept as a student learns in the supervision of the teacher. Supervised learning is a process of providing input data as well as correct output data to the machine learning model. The aim of a supervised learning algorithm is to find a mapping function to map the input variable(x) with the output variable(y). In the real-world, supervised learning can be used for Risk Assessment, Image classification, Fraud Detection, spam filtering, etc.

2.4 UNSUPERVISED MACHINE LEARNING

Unsupervised learning cannot be directly applied to a regression or classification problem because unlike supervised learning, we have the input data but no corresponding output data. The goal of unsupervised learning is to find the underlying structure of dataset, group that data according to similarities, and represent that dataset in a compressed format.

3. THE EXISTING SYSTEM

3.1 OVERVIEW

In Existing system, The task of fake news detection and mitigation becomes crucial in the digital era, i.e., post the advent of social media to mitigate its adverse impacts. Considerable research has been undertaken for the same by researchers worldwide over time. Manual detection of fake news is challenging, as it seems as good as real news from manual observation. In recent years, various AI approaches have been proposed for fake news detection and have shown promising results. Various surveys have been conducted for ML and DL techniques used in this ld. Recent surveys for fake news detection have analyzed various ML and DL techniques for fake news classification incorporating various datasets available, identified challenges and future scopes for the same and Authors surveyed various fake news detection methods. The authors have analyzed fake news from various perspectives and proposed a survey describing various research conducted for fake news and rumor identification on social media platforms. Surveyed various fake news detection techniques and proposed a novel system for the same. Overall, the survey has reviewed the efcacy of modern AI techniques for fake news detection and identified the societal impacts of fake news dissemination. Though these surveys are much information oriented when we look at the various points which need to be covered, each research seems to miss to incorporate one or the other component such as overview and background of fake news, detailed and comprehensive review of AI techniques used based on various categories.

3.2 DISADVANTAGES EXISTING SYSTEM

- Analyzing the textual content of new articles using natural language processing (NLP) techniques such as sentiment analysis, topic modeling, and linguistic analysis.
- Oversimplify the complexity of fake news detection.

4. THE PROPOSED METHOD

4.1 OVERVIEW

Sometimes dissemination of fake news has severe impacts directly or indirectly related to the financial crisis and mental health. Its widespread is for various purposes, such as political parties spreading fake news to get an advantage in the elections (making the election procedure unfair). Thus, there was an imperative need to develop solutions to combat the problem of fake news dissemination. We were scintillated by the cant prowess of AI, ML, and DL techniques to identify fake news. The present surveys discuss and analyses various AI techniques and Ensemble learning-based approaches. We were motivated by the findings

- We implemented a few approaches and discussed their empirical results. The AI techniques have Evolved to give significant results in terms of their efficacy in the held and the research is ongoing to enhance the AI techniques for even better results**.**

- We present a comprehensive survey and discuss the taxonomy on AI techniques employed for fake news classification and highlight their advancements in the same domain.
- We implemented passive aggressive, LSTM, NB, and random forest algorithms for the fake news classification. Passive aggressive is an ideal algorithm to read data dynamically.
- The performance evaluation section discusses the results and empirical findings of these methods in detail. Finally, we present the research challenges and open issues about the state-of-the-art AI techniques designed for the identification/detection of fake news.

4.2 ADVANTAGES OF PROPOSED METHODOLOGY

- Accurate
- Efficient
- Scalable

5. SYSTEM DESIGN

5.1 INTRODUCTION

It is a process of planning a new business system or replacing an existing system by defining its components or modules to satisfy the specific requirements. Before planning, you need to understand the old system thoroughly and determine how computers can best be used in order to operate efficiently.

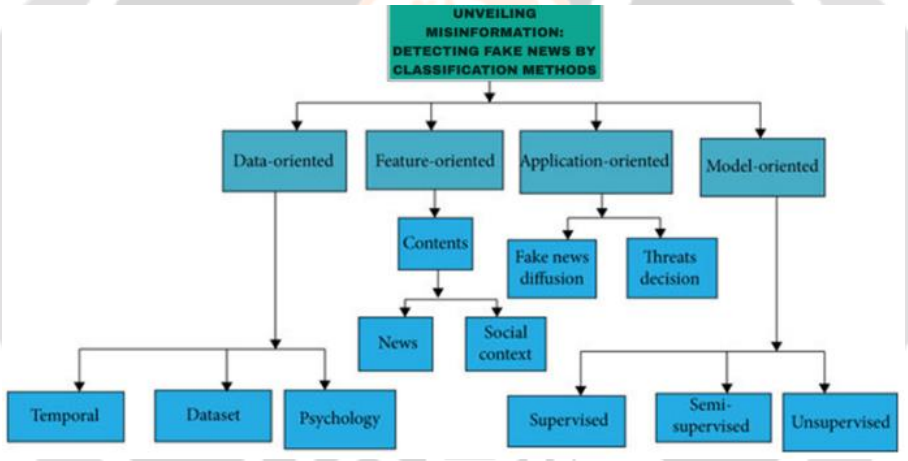

Fig -1 System Architecture

5.2 MODULES DESCRIPTION

a. Service Provider:
In this module, the functionalities are follows :
1. Login
2. Browse & Train & test Data set
3. View trained & tested Accuracy in Bar chart
4. View trained & Tested Accuracy Results
5. View prediction of News type
6. View News type Prediction ratio
7. Download prediction data sets
8. View News Prediction type ratio results
9. View all Remote users
10. Log Out

b. Remote User:
In this module, the functionalities are follows :
1. Register

2. Login
3. Predict News type Detection
4. View Profile
5. Logout.

6. RESULTS & PERFORMANCE

6.1 EXECUTION PROCEDURE

The Execution procedure is as follows :

1. In this research work with data with attributes are observable and then all of them are floating data. And there's a decision class/class variable. This data was collected from Kaggle machine learning repository.
2. In this research 70% data use for train model and 30% data use for testing purpose.
3. Logistic Regression is used as Classifier .
4. In the classification report we were able to find out the desired result
5. In this analysis the result depends on some part of this research. However, which algorithm gives the best true positive, false positive, true negative, and false negative are the best algorithms in this analysis

6.2 SCRENSHOTS

Fig -2. Home Page

Fig -3 Remote User Login

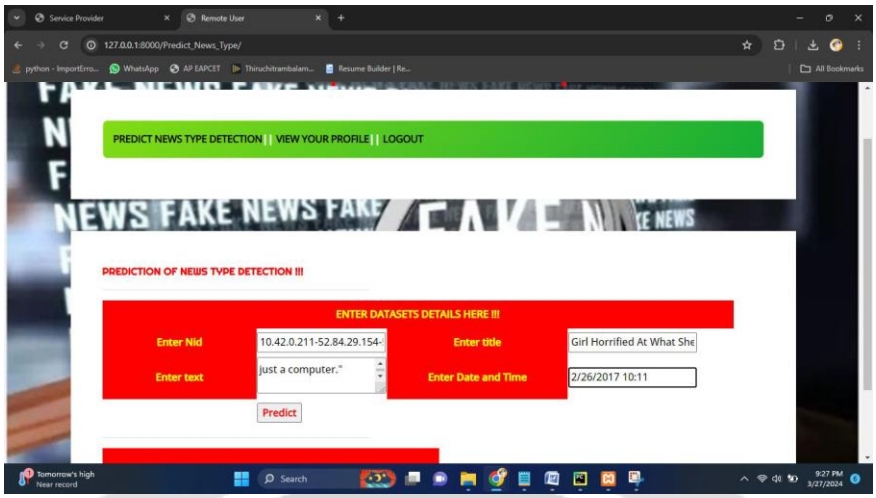

Fig -4 Enter Details For Prediction

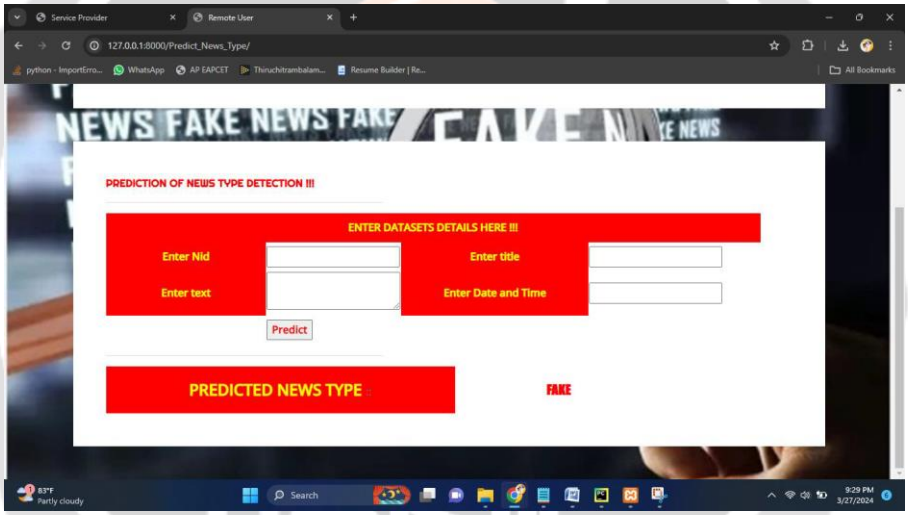

Fig -5 Prediction Result

7. CONCLUSION

In this project, Social media has become pervasive and more prevalent in recent years. People now prefer to read news more from social media platforms than traditional mainstream news channels. This led to an increase in the dissemination of fake news in social media, as it is much easier to share information on social media without any verification. The adverse impact of fake news is also dangerously increasing, like its impact on the 2016 US election. This can harm people lives. This paper presented a comprehensive, analytical, and evidential survey covering all AI techniques like supervised learning, unsupervised learning, semi-supervised learning, reinforcement learning, and ensemble learning for the fake news detectionby overcoming the limitations of the existing state-of-the-art surveys. We implemented four state-of-the art AI techniques for fake news classification for brevity, namely Random Forest (RF), and Naive Bayes (NB), Logistic Regression, Support Vector Machine (SVM). The discussion of how to optimally design hyper parameters is also carried out in each implemented algorithm. At last, some key suggestions from the proposed model is represented, along with the challenges and future scope in this direction. Below are some proposed insights for fake news classification using the techniques discussed in this project. For the detection of fake news, a subset of articles classified as fake by NB (using TF-IDF vectorizer) can be made from the original dataset. This subset will coveralmost all the fake news articles as this technique has a very high recall.

8. REFERENCES

[1] D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, et al., "The science of fakenews," Science, vol. 359, no. 6380, pp. 1094–1096, 2018.
[2] V. Perez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea, "Automatic detection offake news," arXiv preprint arXiv:1708.07104, 2017.
[3] "Newsmedialit fakenews."https://d1e2bohyu2u2w9.cloudfront.net/education/sites/default /files/tlr-asset/ news media lit_fake news time line_8. 5x11.pdf.
[4] J. SOLL, "The long and brutal history of fake news," 2016.
[5] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media:A data mining perspective," ACM SIGKDD Explorations Newsletter, vol. 19, no. 1, pp.22–36, 2017.
[6] J.Weedon,W. Nuland, and A. Stamos, "Information operations and facebook," Retrieved from Facebook: https://fbnewsroomus.files.wordpress.com/2017/04/facebook-and-information-operations-v1. pdf, 2017.
[7] J. Vora, S. Tanwar, S. Tyagi, N. Kumar, and J. J. P. C. Rodrigues, "Homebased exercise system for patients using iot enabled smart speaker," in 2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom), pp. 1–6, 2017.
[8] "Citizen "journalist" hits apple stock with false (steve jobs) heart attack rumor | techcrunch." https://techcrunch.com/2008/10/03/citizenjournalist-hits-apple-stock- with-false-steve-jobs-heart-attack-rumor/. (Accessed on 07/04/2020).
[9] K. Stahl, "Fake news detection in social media," California State University Stanislaus,vol. 6, 2018.
[10] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media:A data mining perspective," SIGKDD Explor. Newsl., vol. 19, p. 22–36, Sept. 2017.