Fake News Detection using Machine Learning Framework

Aishwarya Panicker[1], Prof. Milind Gayakwad [2], Prof. Sandeep Vanjale[3], Prof. Pramod Jadhav[4], Prof. Prakash Devale[5], Prof. Suhas Patil[6]

Department of Information Technology, Bharati Vidyapeeth Deemed University, College of Engineering Pune, Maharashtra, 411043

Abstract *— Social media is nowadays a popular medium for the circulation of real-time news all over the world. Easy and quick information proliferation is one of the reasons as to why it is popular. Social media websites are used by a large number of people of various ages, genders, and societal ideas. Despite these positive aspects, a significant disadvantage comes in the form of fake news, as people usually read and share information without caring about its genuineness. During recent pandemics, various misleading information was disseminated, causing devastation among residents. It is critical to investigate news authentication methods. To address this issue, this article first preprocesses the fetched data from the dataset and validates the data using Metadata, Content and Interaction Features and then applies classification. To validate this approach, this article uses fake and real dataset, which incorporates different data sets to generate an unbiased classification output.*

Keywords **- Fake News, Machine Learning, Social Media-Twitter.**

1. INTRODUCTION

Social Media platforms are one of the most popular as well as plays an important part to keep connected with each other, in this modern time. Not only that, we are able to find news all around the world irrespective of its location, language and truthfulness of that news. With the help of social media platforms information's are spread quickly without even checking if the information shared is true or not. The reliableness of information distributed on the World Wide Web (WWW) is a central issue of modern society. In recent years the spreading of misinformation and fake news on the Internet has drawn increasing attention, and has reached the point of dramatically influencing not only political and social realities but also people's life's. As an example, showed the significant impact of fake news in the context of COVID-19, because of this fake news not only normal people but also front-line workers like doctors, nurse, police, researcher, etc. had faced several issue, leading in increasing growth of COVID cases across the world. As we spend more time connecting online through social media platforms, individuals are increasingly seeking for and consuming news from social media outlets rather than traditional news organizations. Some of the few reasons people are spending their time on social media are- 1). It is often more timely and less expensive to consume news on social media compared with traditional news media, such as newspapers or television; and 2) On social media, it is easy to share, comment on, and discuss the news with friends or other readers. It was also discovered that, as a key news source, social media currently exceeds television. Despite the benefits of social media, the quality of news on these platforms is inferior to that of established news organizations.

Large volumes of fake news, i.e. news pieces with purposely misleading material, are created online for a variety of reasons, including financial and political advantage, because it is inexpensive to provide news online and much faster and easier to disseminate through social media. Nowadays anybody can post any news content over the internet and we are unable to identify if that information is fake or true. Fake news' widespread dissemination has the potential to

harm both individuals and society. First, fake news has the potential to upset the news ecosystem's authenticity balance. Second, fake news is designed to induce customers to believe in biased or incorrect information. Propagandists frequently use fake news to spread political ideas or influence. According to some claims, Russia used phony accounts and social bots to propagate misleading information.

Fake news is spreading widely and fastly. In today's digital era, there are many issues of crises, and fake news is one of them. It can jeopardize a person's reputation, organization, and system. Detecting fake news on social media presents a number of novel and difficult research challenges. The term Fake News was not as popular in earlier decades as it is now, however it was still a major concern in our digital world.

Our contribution:-

In this article, we suggest a solution for detecting fake news using a machine learning approach. This article explores many different linguistic and other content and interaction features that are used to distinguish fake information from real ones.

2. LITERATURE SURVEY

As technology is developing day by day thus spreading fake news without verifying the authenticity of those news. Many portals and websites are allowing people to post these information without checking if they are fake or real. And these information spread very fast causing misunderstanding and havoc among citizens. Many researchers tried to find ways to detect these type of false information and news using different machine learning algorithms.

Most of the fake news is spread on social media, these news have characteristics. The result of [1] study that documents the performance of fake new classifier are presented. Method used are Natural processing language, Attribute classification and Features used are Natural language processing, Fake news detection based on attribute classification. Dataset – Social Media. Gap/Future – the precision rate is less.

Identifying each news one by one is completely infeasible. To avoid such thing [2] proposed a system that can reliably classify fake news. Method used are Naive Bayes, Passive Aggressive Classifier and Deep Neural Networks and Features used are Predictive model, feature extraction.. Dataset- Social Media. Gap/Future- it is time consuming.

A fake news detection model based on Bi-directional LSTM-recurrent neural network is done in [3]. The model's performance is evaluated using two freely available unstructured news article datasets. Method used are Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) and Short-Term Memory (LSTM) and feature such as linguistic modeling, Long Short-Term Memory (LSTM). Dataset- News Articles and Social Media. Gap/Future- Future work related to study include use of deep learning approach for fake new detection.

This [4] provides a detailed review of various fake news detection techniques used by different researchers, the datasets they have worked upon and various evaluation parameters used by them for performance evaluation of their models. Method such as Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Decision tree (DT) are used and feature like Natural language processing, data analysis are used.Dataset- Social media platforms along with twitter, Facebook, YouTube and other social networking web sites.Gap/Future- As the new technology gets introduced this work can be expanded.

Identifying Fake News manually is tremendous problem for all, [5] estimate a model that intuitively distinguishes fake news from news articles. Method and Features like Support Vector Machine (SVM), Computational modeling, linguistic/ stylometric features, a bag of words TF and BOW TF-IDF vector are used. Dataset- Social Media .

Gap/future- In future, perform machine learning models on combined both feature set stylometric features and word vector features to achieve better results rather than proposed work.

Information credibility is becoming an important part of information sharing in society. With new features, in study [6] the credibility of information provided in social media is increasing significantly indicating better accuracy compared to existing one. Method such as classification of Naive Bayes (NB), Support Vector Machine(SVM), Logistic Regression (Logit) and J48 Algorithm (J48) and Features New 17 features for Twitter and 49 features for Facebook. Spam and sentiment are also considered are used. Dataset- 10,000 samples from Twitter and Facebook 56 accounts with 23489 messages. Gap/Future- Advanced feature which would be introduced in future would help in increasing the accuracy.

Sending and receiving email is the easiest way to communicate and spread spam. This [7] different machine learning technique is discussed for controlling spamming problem .Method like Support vector machine (SVM), Decision Tree, Random Forest(RF) are used and Header Based Techniques, Content Based Technique, OCR Based Technique are the features used. Dataset- Email dataset. Gap/Future- There is no machine learning method that achieved 100% accuracy. Identifying the best algorithm is an important task as their strengths need to be weighed against their limitations.

Using [8] tests NL tweets to check the impact of explore to the false information. News literacy, Support Vector Machine are the method used and content based feature. Dataset- Social Media like twitter. Gap/Future- NL message are able to alter misinformation but no single message.

The [9] address the problem of detecting misleading information related to COVID-19. Decision Tree (DT), k-Nearest Neighbor (kNN), Logistic Regression (LR), Bernoulli Naïve Bayes (BNB), Perceptron, Neural Network (NN), Ensemble Random Forest (ERF), and Extreme Gradient Boosting classifiers (XGBoost) are used and feature like Term Frequency-Inverse Document Frequency (TF-IDF) is also used. Dataset- Information from World Health Organization (WHO),UNICEF and United Nations. Gap/Future- use the proposed framework for detecting misleading information, shared or re-tweeted on twitter in near real-time manner

This [10] deals with a review of existing machine learning algorithms for detecting and reducing fake news. Method like- Neural Network, Support Vector Method (SVM),Naive Bayes and feature like N-gram analysis are used. Dataset- Twitter, Facebook, YouTube and other social networking sites. Gap/Future- Existing System are not that much efficient in detecting fake news because lack of fake news dataset

3. PROPOSED APPROACH

News can be found online from a variety of places, including news agency homepages, search engines, and social media websites. Data collection mainly includes data labeling, data acquisition using existing dataset. In this the two dataset for kaggle is used. The data fetched for the dataset are then pre-processed and features like metadata feature, content feature and interaction feature are applied.

System Diagram:

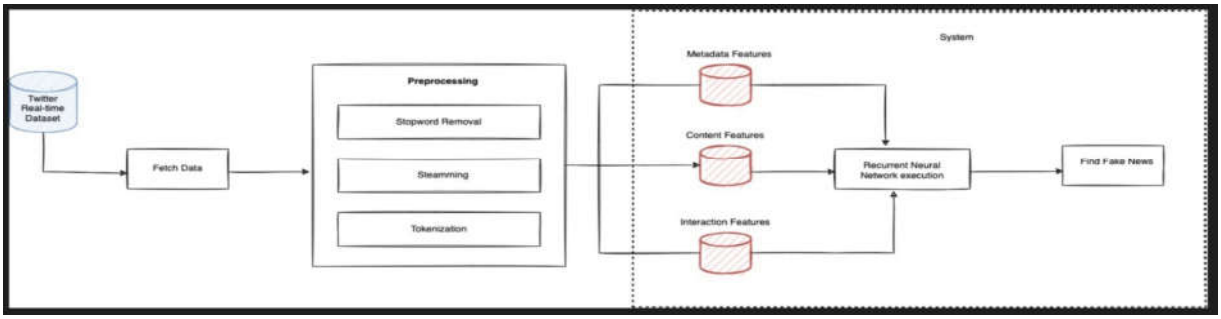

Figure 1. System Architecture

List of Features:-

1. Metadata feature
2. Content feature
3. Interaction feature

Metadata feature- metadata feature is used so that information about data for dataset, finding and working with particular instance of data becomes easier.

Content based feature- the content based features are the features that focus on the substance of the tweet. The content based features includes- linguistic feature-Linguistic features indicating the basic language element analysis and sentence structure of the news content language. These features are good indicators and cues for detection of false or fake news which was written to mislead or create havoc within the people. Like, from this dataset, words like Government News, politicsNews, etc. Other content based characteristics are number of positive words from the dataset, number of negative words, time span- like the interval time between creation of an account and the time when the post was tweeted from that account.

Interaction feature-This focus on characteristics of tweet and the online users like if the account user is verified, number of followers to the particular account, if the post or tweet have description and the length of the description, number of hashtag and retweet fraction.

Outcome:

The dataset used for this project are fake.csv and true.csv. Each row contains information about a new: title, text, and subject and publication date. Each news was collected and classified into different subjects (news, politics, left-news, government news, US news, and middle-east). The dataset consists of two files: "True.csv" with 21417 news considered as true; "Fake.csv" with 23481 fake news, as shown in figure 2.

Amount of fake news: 23481

Amount of true news: 21417

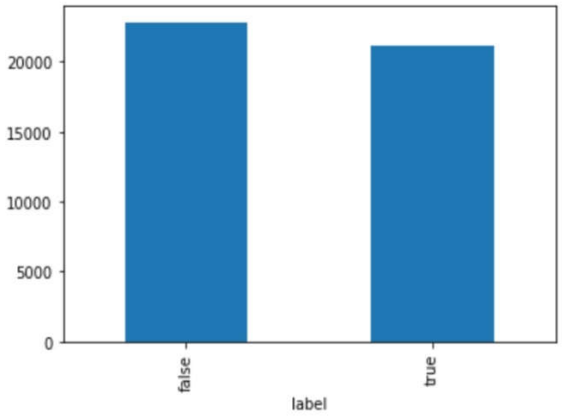

Figure 2: Fake and True news dataset comparison

Data cleaning and preprocessing:

The process of reviewing raw data and reducing it down to a more useful form is known as data cleansing.Here I have removed the title, found the missing value and converted the text to lowercase. As well removed punctuation and stop words. And grouped the dataset based on Subject.

Checking for missing values:-

title False
text False
subject False
date False
isReal False
dtype: bool

As you can see there are no missing values, so our data is clean.
subject
Government News 1570
Middle-east 778
News 9050
US_News 783
left-news 4459
politics 6841
politicsNews 11272
worldnews 10145
Name: text, dtype: int64

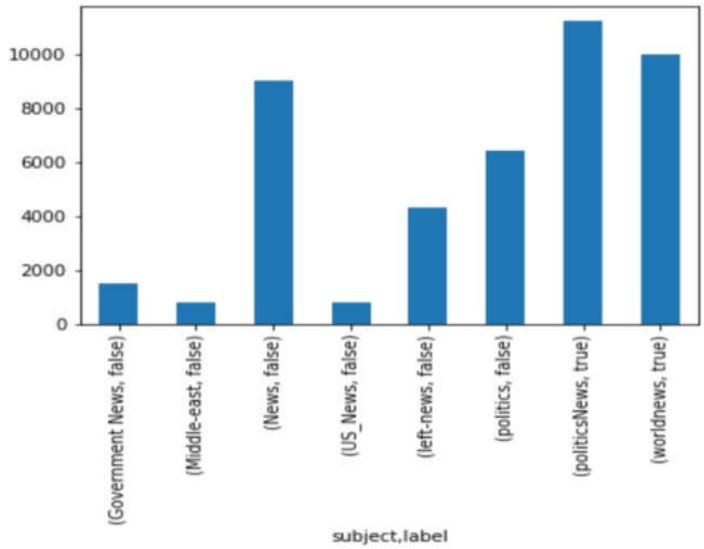

Figure 3: Data exploration

As you can see in Figure 3, there are two categories for politics. In the true news, we have politicsNews while in the fake news we have only politics. If I consider the subject as a feature in our model, to be more realistic, I will have to join both of them.

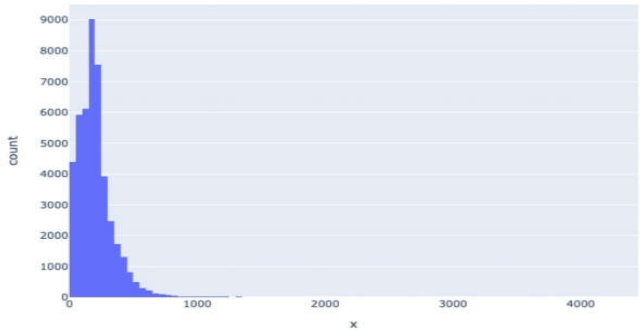

Figure 4: Histogram graph based on distribution of number of words in a text

As shown in figure 4, it is a histogram graph which represents the distribution of the number of words in a text.

Now I will create a new column called Year from Date and Analyze whether fake or true news has any correlation in the timeline as shown in figure 5.

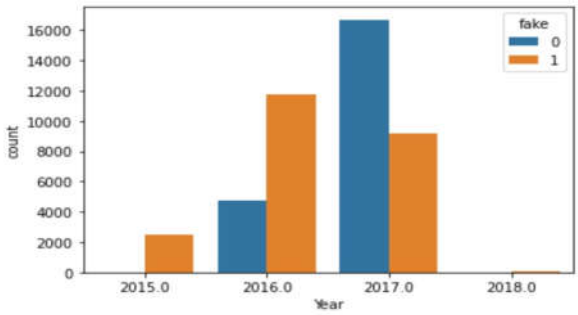

Figure 5: Feature Engineering

Figure 6: Word Count

In figure 6, shows the word count of the most used words which was found using bilinear interpolation.

Conclusion

The popularity of social media is growing rapidly, more and more people consume social media news instead of traditional media. However, social media have also been used to disseminate false news, which has strong negative impacts on individual users and the wider society. Here to explore the problem of false news by reviewing existing literature in two phases: characterization and detection is used. In the characterization phase, we introduce the basic concepts and principles of false news in social media. In the detection phase, we reviewed the current false news detection approaches from a machine learning perspective, including feature extraction and model building. We also discuss evaluation metrics, and future promising directions in fake detection research and expand the field to other applications.

REFERENCES

1. T. Traylor, J. Straub, Gurmeet and N. Snell, "Classifying Fake News Articles Using Natural Language Processing to Identify In-Article Attribution as a Supervised Learning Estimator," *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, 2019, pp. 445-449, doi: 10.1109/ICOSC.2019.8665593.

2. R. R. Mandical, N. Mamatha, N. Shivakumar, R. Monica and A. N. Krishna, "Identification of Fake News Using Machine Learning," 2020 IEEE International Conference on Electronics, Computing and Communication Technologies *(*CONECCT*)*, 2020, pp. 1-6, doi: 10.1109/CONECCT50063.2020.9198610.

3. Pritika Bahad, Preeti Saxena, Raj Kamal, "Fake News detection using Bi-directional LSTM-Recurrent Neural Network",2019 DOI:10.1016/j.procs.2020.01.072

4. S. Kumar and B. Arora, "A Review of Fake News Detection Using Machine Learning Techniques," *2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC)*, 2021, pp. 1-8, doi: 10.1109/ICESC51422.2021.9532796.

5. M. Kumar Jain, D. Gopalani, Y. Kumar Meena and R. Kumar, "Machine Learning based Fake News Detection using linguistic features and word vector features," *2020 IEEE 7th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, 2020, pp. 1-6, doi: 10.1109/UPCON50219.2020.9376576.

6. Erwin B. Setiawan, Dwi H. Widyantoro, Kridanto Surendro, ”Measuring information credibility in social media using combination of user profile and message content dimensions International Journal of Electrical and Computer Engineering (IJECE) “ doi:10.11591/ijece.v10i4.pp35

7. Abdulmalik D. Mohammed, Opeyemi O. Abisoye,“A Review on Machine Learning Techniques for Image Based Spam Emails Detection”, 2021 doi: 10.1109/CYBERNIGERIA51635.2021.9428826

8. Melissa Tully, Emily K. Vraga & Leticia Bode (2020) “Designing and Testing News Literacy Messages for Social Media”, 2019
Doi: 10.1080/15205436.2019.1604970

9. M. K. Elhadad, K. F. Li and F. Gebali, "Detecting Misleading Information on COVID-19," in *IEEE Access*, vol. 8, pp. 165201-165215, 2020, doi: 10.1109/ACCESS.2020.3022867.

10. Murari Choudhary, Shashank Jha, Prashant, Deepika Saxena Ashutosh Kumar Singh,“A Review of Fake News Detection Methods using Machine Learning”, 2021, doi: 10.1109/INCET51464.2021.9456299