ISSN:0975-9646

Fake News Detection and Prediction Using Machine Learning Algorithms

Thasniya KP, Muneer V.K, Mohamed Basheer K.P, Rizwana K T.
Research Scholar in Computer Science
SS College, Areekode.
Kerala, India

***Abstract*—Today, the increased amount of information sources on internet creates the problem of information overflow. Filtering the relevant and genuine information is another challenge social media facing now. Mobile phones and other electronic gadgets became quite common through which people get up-to-date information. Verifying the authenticity of news needs to have prime importance though a difficult task. This paper outlines a new approach for finding the genuineness of news content. This helps to eliminate the rumors from spreading through social platforms. By using the web scraping method, we assemble the news content related to the news posted for checking. The news prediction is done by implementing techniques like TF-IDF, Bag of words and Natural language processing. The experimental results specify that the system shows an accuracy of 90% when tested against a test set.**

Keywords — Fake News, Web scraping, Natural language processing.

I. INTRODUCTION

The rapid change in our communication environment through the internet has brought a lot of changes in our society. Social media is widely considered to be the topmost preferred medium of daily communication. From children to the elderly, social media has had a profound impact on their daily lives. Therefore, the rapid adoption of social media for everything has increased the sharing of information between users without knowing whether it is fake or genuine. People are sharing news without checking the authenticity of the news.

Fake news is of different types such as target misleading information, which is shared through social media utilizing someone's interest, for generating additional attention fake headlines were created which depict bogus facts, moreover, viral posts which are shared through various social media without checking the authenticity of posts. Misinformation has become a daily phenomenon in the changing media landscape, where readers are becoming publishers. Anyone can create misinformation and share this on social platforms. We can say that the main objective of these fake news creators is to mislead the newsreaders [1]. This news has a target, like creating an issue in society, damaging the dignity of an individual or an organization. As a result, detecting fake news and assessing the quality of the news becomes an even more predominant skill. When people share news articles, they must first check the authenticity. Therefore, this is an important factor that can reduce the spread of anonymous information.

Fake news detection helps in finding the truth regarding the news articles spreading through web platforms. This paper is a preliminary attempt to solve the issue of spreading suspicious data. The principal objective of the proposed system is to find out the real and fake news. The system helps to find the authenticity of the articles. In this paper, we developed a methodology for false news detection using the concept of text similarity checking. We have used TF-IDF (text representation model) [4], Bag of words, and natural language processing for implementing the system.

Section I give a brief overview of fake news in social media and in section II the currently used algorithm, methodologies, and their features for the identification of news articles as accurate or inaccurate are given. In sections III and IV, we bring up a new framework for fake news detection and discuss the strategies and algorithms used in the system. In section V, we describe the details of the system and Section VI includes the experimental details of the system. Our conclusions are drawn in the final section VII.

II. LITERATURE SURVEY

In this section, we discuss the conventionally used algorithms, methodologies, and techniques that have been used to execute a fake news detection system. Discussions regarding fake news detection is a dominated research in recent years.

A. Current System

Many investigators have been conducting experiments on implementing fake news identification systems. Prior research substantiates use of machine learning algorithms such as Support vector machine, Naïve based classifier, NLP methods, sentence similarity, classification algorithm, which are most widely used for detecting fake news.

An approach for detecting fake news based on social media [2], reported a data mining technique for fake news

detection. According to their view news verification depends on factors like publisher, content, time of posting on social media websites, number of engagements between different users, and the article. By extracting text features from the news and performing linguistic study on these features they build up a machine learning model for fake news detection. A smart system for fake news detection [3], In which they have proven that by using supervised machine learning algorithms (Naïve Bayes Classification and SVM) and Natural language processing the suggested framework achieved accuracy of up to 93.5% for detecting suspicious data. This seems to be an innovative approach.

Identification of fake news using machine learning [4], where they implement a system to classify fake news. Approximately eight data sets are used to run the system. They used machine learning algorithms such as the Naive Bayes Classifier, the Passive-Aggressive Classifier, and the Deep Neural Network. The TF-IDF vectorizer is used to convert news articles' text data into its numeric form.

Famous: fake news detection model [5], In this model, they put forward a new sentence matching model to identify suspicious news that can productively manage sentence matching by retrieving the principal sentence based on the bidirectional LSTM model. They worked on a Korean article data set consists of five layers. The overall average accuracy of the system is 69%.

Fake news detection [6] developed a simple fake news detection method based on one of the machine learning algorithms, like naïve Bayes for applying it on Facebook and labeled it as fake or real. By using a count vectorizer first model used title as their source for word formation and in the second model text is the source. The results were compared according to their AUC score and the second model was found to be better at 0.93 points and 0.912 points in n_grams and beyond respectively.

Fake news detection using Naïve Bayes classifier [7] in this model naïve bayes classifier and artificial intelligence methodologies are used for detection of false news. The developed system was tested on a comparatively new dataset (BuzzFeed news), which allowed evaluating its performance on recent data. The system achieves an accuracy of 74% on the test set.

Weakly supervised learning for fake news on Twitter [8] is based on classifying the fake and non-fake tweets. The Classification is purely based on the source of the post/tweet. It consists of a large-scale training data set collected as trustworthy and untrustworthy sources. Their approach is trained on a large-scale noisy data set by using different machine learning algorithms like Naive Bayes, Decision Trees, Support Vector Machines (SVM), and Neural Networks as basic classifiers. And also, random forest and XG Boost, using parameter optimization on all of those approaches.

Improving spam detection in social networks [9] this model presents a method for detecting the misinformation spreaders in Twitter. Twitter is one of the most well-liked and well-popular social media platforms. They implement an integrated approach for identifying spammers. By combining the three machine learning algorithms such as Naïve Bayes classifier, Decision tree algorithm, and clustering they attain an overall accuracy of 87.9% for detecting spammers. They find better accuracy by integrated approach rather than taking alone these algorithms.

Fake news detection on social media using k -nearest neighbor classifier [10] by using k- nearest neighbor classifier they implement their model and they also consider the secondary information like the social activities of that particular user who spreads this false information. By using five features from the buzz feed data set, they trained a model using different values of K and achieved a classification accuracy of 79% when tested against the Facebook news post data set.

Hoaxy: A Fraud Online Tracking Platform [11] They collect news from social platforms and news websites using web scraping and web syndication. They see user activity by calculating the number of user tweets posted and the popularity of the URL by calculating the total number of people who have posted the tweet. Based on these observations they conclude that rumors-domination is dominated by a few active accounts that carry the burden of informing and disseminating false information, and the spread of fact-checking is a more widespread, grassroots activity.

Fake news detection using deep learning techniques [12] They draw our attention to a fake news detection system based on classification such as s Logistic regression (LR), Naïve bayes (NB), Support vector machine (SVM), Random Forest (RF), and deep neural network (DNN). Then they compare NB, RF, SVM, LR, and DNN based on time, memory, and accuracy, according to comparison results exhibits that DNN Algorithm is improved than the rest algorithm in accuracy and time kind because rest classifiers require more time and give less accuracy.

III. PROPOSED APPROACH

The proposed model is based on an online web platform.

- *News checker*

This is the primary component of our system. Here we can verify whether the news we are given is false or not. The client can input the news, which is wanted to be checked into the website. The authenticity of the news can be found here. We can give Text-based data as input to the system. It comes after several steps for finding the authenticity of the news. The input news, which is given by the client is compared to related or similar news from websites or news sources by the web scraping technique. After that, if there is any news excessively like the client input or the same news is found on any website at that point, we can say that the given news is true. Otherwise, we can say that given news is fake or inaccurate. The

main operation of the system is contributed by the check news button. That part is critically important. Most recent news headlines (Live news headlines) can be viewed on our site. This is by scrapping the news websites and displaying that scrapped news using the web. We can get current news updates here.

IV. METHODOLOGY

The methodologies used in the proposed model are given below.

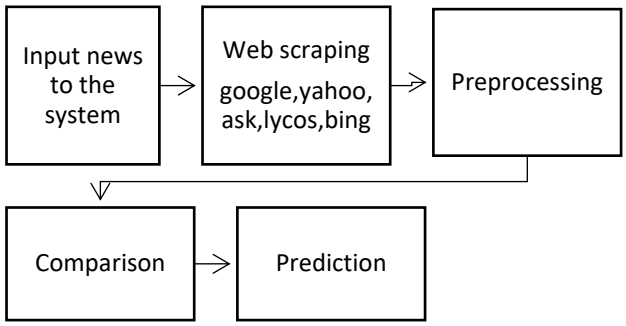

Figure1: System Architecture

Figure1 depicts the architecture of the system. The user can insert the news which wants to find as trustworthy or not. At that point, our system carries out web scraping methods.

4.1. Web scraping

Web scraping [6] is a method used in our system for retrieving news content. We want the news or articles related to the content posted by the user from a trusted website. Web scrapers can bring out all the data on particular websites or the specific data that a user wants. When we need to scrape a site, initially, it has given the URL of the required sites. At that time, it loads all the HTML code for those required sites and a more innovative scraper might even take out all the CSS and JavaScript elements too. After that, the scraper captures the required data from this HTML code and outputs this data in the format specified by the user.

This web scraping method finds out the URL of the page that we want to scrape and extracts the data in our required format. Moreover, we are using five search engines for scraping data from trusted websites.

They are:

1.Google – we are performing searches for our news using google [13]. When the client inputs the news to the website, After the 2 stages of google search crawling and indexing the google tries to bring better results related to the user input considering factors like location, device, etc. We scrape those results using web scraping. *Bing*– We use this search engine also for web scraping purposes, however, it is created and operated by Microsoft. *Ask* , *yahoo* and *Lycos* – are the top search engines used for searching.

After web scraping, the next step is to preprocess the data. Here we applied various steps to make data more suitable.

4.2 Preprocessing

Data should be pre-processed for getting better results. It comprises the removal of URLs, stemming, punctuations, and stop words. At this stage, the Natural language processing method is used. Natural Language Processing allows us to find out the key information from the data.

Preprocessing mainly works on three levels [11].

1. Splitting
2. Stop words removal
3. Stemming

In the initial level of the process, splitting the sentence means separating each sentence from the other part to deal with them separately. And in the second level remove the unimportant words (stop words) like (the, a, an, from, to, for instance, of, etc.) from each part of the sentence. The next level is stemming where each word is returned to its origin and converting into a vector format (using bag of words). A Stemming technique is effectively used to detach suffixes or prefixes from a word.

- *Bag of Words*

Machine learning algorithms are unable to work with the raw text directly. Rather, the text must be converted into vectors of numbers. In natural language processing, a common technique for extracting features of the text is to place entire words that occur in the text in a bucket. This approach is called a bag of words.

4.3. Comparison

After preprocessing, the system will compare the input news with the scraped data. The results of these steps will be input to the TF-IDF vectorizer.

TF-IDF Vectorizer known as the Term frequency-inverse document frequency [4], where the value rises proportionately to the number of times a word becomes visible in the document but is neutralized by the frequency of the word in the collection. Term frequency can be defined as the number of times a word appears in the document divided by the total number of the document. A text similarity check is performed by the system using TF-IDF. Natural language processing is also used for performing sentence matching.

4.4. Prediction

With the completion of the above steps, the next step is prediction. Based on the comparison If there is any similar news that can be found from any of these trusted sites, we can say that the given news is real, otherwise, we can say that the given news is fake. Trustworthy sites do not publish fake news as real. Here we are assuming that search engines, Google, Yahoo, Ask, Bing, and Lycos as trustworthy. Each of the search engines predicts news as fake or real according to the news. The final prediction by the system is by counting the total number

of news predicted as real or fake by each of the search engines. If the news predicted as real is greater than the news predicted as fake, then the final output of the system is real otherwise the output will be fake. This can be shown in the given table.

TABLE 1: PREDICTION IN SEARCH ENGINES

	Google	Bing	Ask	Yahoo	Lycos
News Sample	Real	Real	Real	Real	Fake

In the above table1, the news samples given are predicted as real by four search engines and only one predicted it as fake. Since the total count of 'real' prediction is greater than the count of 'fake' prediction the system predicts the sample news as real.

V. IMPLEMENTATION DETAILS

In this subdivision, we consider the implementation details of the system. The system is implemented using python as we know that python contains many python libraries. The python library consists of a large collection of python modules that are organized as a python package. To implement our system, we use libraries like requests, regular expression, Beautiful Soup, NumPy, Flask.

Request module is in python for sending HTTP requests and this HTTP request will return all response data. We are using this request module for scraping. In the scraping method when we give the URL s of the required sites it loads all the HTML code for those required sites, and we want to remove this code and fetch out the required data from that code.

The regular expression module (re module) [14] was used in our system for extracting the data that we needed from the web pages. We are using re module for separating content from Html code. Beautiful soup is another python package that is also used for web scraping. Beautiful Soup is used in our system mainly for getting live news updates. One more python library called NumPy is used while working with arrays. Flask is another python web framework that is used in our system for building the web application.

VI. EXPERIMENTAL RESULTS

We used a specimen model data set for finding the accuracy of the prediction. In the sample model, 0 represents fake news, and 1 represents real news. By using the test set a sample test is implemented in our system. The demonstration is done using python programming and a machine learning algorithm.

The accuracy of predicting fake news, real news, and the mixed data of both fake and real are given in the table below.

TABLE 2: SYSTEM ACCURACY

	Fake	Original	Mixed
Total number of news	32	34	66
Prediction	29	31	60
Accuracy	90.6	91.1	90.9

Table2 highlights that we use 32 fake news for prediction and out of that 29 are predicted as fake. Further, we use 34 original news, and out of that 31 predicted as real. While, mixing both real and fake news, out of 66 news 60 news predicted correctly.

The accuracy of the system can be calculated by using the formula:

$$\text{Accuracy} = \frac{\text{Total no: of accurate prediction}}{\text{Total no: of news prediction}} * 100$$

As stated in the introduction, our main objective is to find out the credibility of news spreading on social media. Our system needs to predict the output as accurately as possible. We can classify the prediction into four categories as given below:

True Fake: When the fake news is predicted as fake.
True Real: When the real news is predicted as real.
False Fake: When the fake news is predicted as real.
False Real: When the real news is predicted as fake.

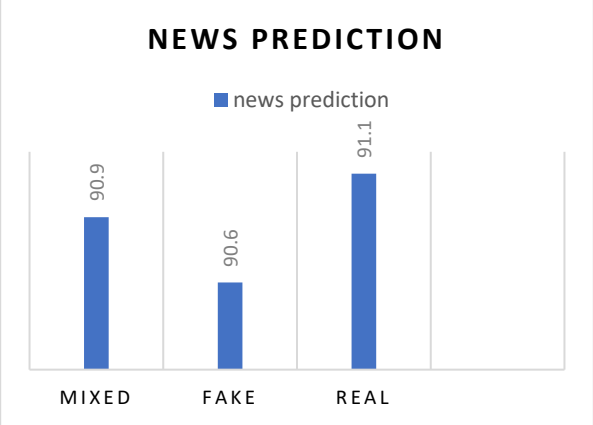

Figure3: Accuracy of system using chart

By analyzing the figure3, we understood that while checking the accuracy of the system considering the prediction of true fake, the system shows an accuracy of 90.6% and the remaining is false fake. When we check the accuracy of genuine news, that is true real the system shows an accuracy of 91.1% here the remaining is false real. However, when we check the accuracy of both real and fake news(mixed) the model has an accuracy of 90.9%. Hence, it is the overall accuracy of the system. These results offer vital evidence for the detection of news as fake or real.

VII. CONCLUSION

Prior to believe in bogus news and sharing it through web-based media, we should discover the reality behind it. In our work, we use a text similarity finding algorithm for predicting whether the news is fake or accurate. We use five search engines, initially, we use only one search engine for scraping, even though the accuracy of prediction is not so great. Then, we also included four additional search engines for scraping the data. Although it enhances the accuracy of the system. Our work has led us to the conclusion that including more search engines it improves the system to get better results. The findings of this study indicate that the system achieved accuracy of up to 90% in detecting news as fake or real. Our study provides a new framework for a new way of detecting the authenticity of the news. Hence, future studies on the current topic are therefore needed to improve the accuracy of the system by adding some more search engines for scraping, fetching more data from social media, and using another text similarity checking algorithm.

REFERENCES

[1] "Verizon.com", Accessed on May 22 2021. [online], https://www.verizon.com/info/technology/fake-news-on-social-media/

[2] K. Shu, A. Sliva, S. Wang, J. Tang, H. Liu, Fake news detection on social media: A data mining perspective,2017,ACM SIGKDD Explorations Newsletter, https://doi.org/10.1145/3137597.3137600.

[3] A. Jain, A. Shakya, H. Khatter and A. K. Gupta, "A smart System for Fake News Detection Using Machine Learning," 2019 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), 2019, pp. 1-4, doi: 10.1109/ICICT46931.2019.8977659.

[4] 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.

[5] N. Kim, D. Seo and C. Jeong, "FAMOUS: Fake News Detection Model Based on Unified Key Sentence Information," 2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS), 2018, pp. 617-620, doi: 10.1109/ICSESS.2018.8663864.

[6] A. Jain and A. Kasbe, "Fake News Detection," 2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), 2018, pp. 1-5, doi: 10.1109/SCEECS.2018.8546944.

[7] M. Granik and V. Mesyura, "Fake news detection using naive Bayes classifier," 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering, 2017,pp.900-903,doi:10.1109/UKRCON.2017.8100379.

[8] S. Helmstetter and H. Paulheim, "Weakly Supervised Learning for Fake News Detection on Twitter," 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2018, pp. 274-277, doi: 10.1109/ASONAM.2018.8508520.

[9] A. Gupta and R. Kaushal, "Improving spam detection in Online Social Networks," 2015 International Conference on Cognitive Computing and Information Processing (CCIP), 2015, pp. 1-6, doi: 10.1109/CCIP.2015.7100738.

[10] A. Gupta and R. Kaushal, "Improving spam detection in Online Social Networks," 2015 International Conference on Cognitive Computing and Information Processing (CCIP), 2015, pp. 1-6, doi: 10.1109/CCIP.2015.7100738.

[11] Shao, Chengcheng & Ciampaglia, Giovanni & Flammini, Alessandro & Menczer, Filippo. (2016). Hoaxy: A Platform for Tracking Online Misinformation. WWW '16 Companion: Proceedings of the 25th International Conference Companion on World Wide Web. 10.1145/2872518.2890098.

[12] C. K. Hiramath and G. C. Deshpande, "Fake News Detection Using Deep Learning Techniques," 2019 1st International Conference on Advances in Information Technology (ICAIT), 2019, pp. 411-415, doi: 10.1109/ICAIT47043.2019.8987258.

[13] https://developers.google.com/search/docs/basics/how-search-works. "Google Search Central" Accessed on: May 22,2021.

[14] https://towardsdatascience.com/web-scraping-regular-expressions-and-data-visualization-doing-it-all-in-python-37a1aade7924, "towards data science" Accessed on: May 22,2021. [online].