A Survey on Fake Content Detection Over the Internet

Shweta Kushwaha [1]

*UG student of
Department of Computer Science
SRMCEM
Lucknow, Uttar Pradesh, India*

shweta_1812210107@srmcem.ac.in

Siddhant Bajpai [2]

*UG student of
Department of Computer Science
SRMCEM
Lucknow, Uttar Pradesh, India*

bajpaisiddhant57@gmail.com

Er. Sandeep Dubey [3]

*Department of Computer Science
SRMCEM
Lucknow, Uttar Pradesh, India*

sandeep@srmcem.ac.in

Abstract— **Fake content is something that is present on social media or on other sites with the objective of misleading the audience or the readers with the increasing reach of social media to the number of people the chances of spreading fake news among them is also increasing. So has always been a burning topic. To control the widespread of fake content and to retain the trust of people on social media it is necessary to perform a fake content check on any material before it is posted on social media. This paper presents a survey on fake news detection. It summarizes the work done previously in this field. In addition, here we propose a machine learning approach to detect fake content. We predict whether the text provided by the user is fake or real using the SVM and Naive Bayes machine learning algorithms.**

Index Terms— *Machine Learning, Logistic Regression, ensemble, Random Forest.*

Fig.1 Sources of Fake and Genuine news.[1]

I. INTRODUCTION

Fake news has always been a matter of concern since the start of civilization. Then rumors were spread physically then but now with the advent of social media, it has become very easy to disseminate any false information. And on the other hand, it has become even tougher to differentiate which news is genuine and which one is fake. This fake news can mislead a huge number of people which can lead to a very serious situation. Thus, it is very important to maintain a check on the content that is being posted on social media platforms.

Detection of fake news online is important in today's society as fresh news content is rapidly being produced as a result of the abundance of available technology.

The project is covered in making a machine learning model that could be used to detect fake news or content for helping the user avoid fake news. It is very important that the solution is found as it will prove useful to both readers and companies involved.

A brief description of the sources of real content and fake content is illustrated in Fig.1.

II. LITERATURE SURVEY

In this section, we summarize some of the existing research works in the field of Machine Learning and Deep Learning to Detect Fake News.

A. In the paper "Fakebuster**:** Fake News Detection System using Logistic Regression Technique in Machine Learning"[2], Mr. Muhammad Syahmi Mokhtar etal proposes a model that uses Logistic Regression to detect fake news. Here all stages of the Model development, Data management, Model Training, Model evaluation, Model integration, and

Model Deployment are described in detail. First of all, the data is cleansed by preprocessing it. Term Frequency-Inverse Document Frequency (TF-IDF) is used for converting the data into a structured format. Then by analyzing the features of the dataset, the appropriate algorithm is chosen and, in this case, it is Logistic Regression. Then the output of the model is evaluated and then it is allotted a score like the F1 score based on the confusion matrix. This model is integrated with the web application in which we can give the text or its URL as an input and obtain the output as real or fake. Here the stance-based approach is used to increase accuracy.

B. In the paper "Analysis of Classifiers for Fake News detection"[3] Mr.Vasu Agarwala etal discusses the approach of Natural Language Processing (NLP) and Machine Learning (ML) to detect fake news. Methods like TF-IDF, n-grams, and bags of words are used for feature selection. Lemmatization, removing stopwords, and stemming have been done to keep only meaningful words in the dataset. A comparative study is performed by feeding each extracted feature to five different classifiers-Naïve Bayes, Logistic Regression Linear SVM, and Random Forest Classifier from the sci-kit library. After evaluation, all five classifiers, SVM and Logistic Regression were observed as the best performers. But the accuracy and prediction are highly dependent on the data. The use of concepts like POS tagging, word2vec, and Topic modeling is suggested for better feature extraction.

C. In the paper "A Machine Learning Approach to Fake News Detection using Knowledge Verification and Natural Language Processing"[4] Mr. Marina Danchovsky Ibrishimova etal discusses a hybrid framework for news detection. The classification model consists of five NLP features along with three knowledge verification features. Where the NLP features include stopword percentage, the ratio of proper nouns to nouns, title length, AR readability, and overall, sentiment analysis. The knowledge verification feature is: - whether the title is similar to a recent title from trusted sources, the number of sources where a similar title is available.

D. In the paper "Fake News Detection Using Machine Learning Ensemble Methods"[5] Mr. Iftikhar Ahmad etal proposes to use a machine learning ensemble approach to detect fake news. Here this technique can detect the news from multiple domains with a higher level of accuracy. Various Ensemble techniques like Random Forest, Bagging ensemble classifier, Boosting Ensemble classifier, and Voting Ensemble Classifier have been used over multiple datasets along with Linguistic Inquiry and Word Count (LIWC) feature set. Two different Voting Classifiers have been used. The first One is an Ensemble of Logistic Regression and Random Forest and KNN whereas the second Voting Classifiers consist of Logistic Regression, Linear SVM, and Classification and Regression Trees (CART). The various algorithms used are discussed in detail. The cross-Validation approach is used to minimize overfitting. The accuracy of the Logistic Regression model is found to be over 90%.

E. In the paper "Fake News Detection Using a Deep Neural Network"[6] Mr. R. K. Kaliyar discusses the Machine Learning Models like Naïve Bayes, K nearest neighbors, Decision Trees, Random Forest, and Deep Learning networks like Shallow Convolutional Neural Networks (CNN)[7], Very Deep Convolutional Neural networks (VDCNN)[8], Long Short-Term Memory Network (LSTM)[9], Gated Recurrent Unit Network (GRU)[10], a combination of Convolutional Neural networks with Long Short-Term Memory (CNN-LSTM)[11], and Convolutional Neural Network with Gated Recurrent Unit (CNN-GRU)[12].
The benefit of feature extraction, n-gram, TF-IDF, word embedding, word2vec features in Deep Neural networks, and the use of select best and chi2 for feature extraction in the Machine learning model have been explored.

III. PROBLEM STATEMENT

False or misleading information is sometimes intentionally generated and propagated as news for various malicious intentions such as to cause reputation damage, create panic/ unrest and suppress real news. With the rise of social media, the prevalence of fake news has increased rapidly. We aim to build a solution that can identify fake content published over the internet. URLs reflecting a mix of potentially fake news and genuine news shall be provided as input during evaluation, and the solution shall be expected to identify the ones depicting fake news. The objective of this problem statement is to identify fake & misleading content published online.

IV. PROPOSED SOLUTION

We propose a way to detect fake news over the internet with the help of machine learning algorithms of SVM and Naive Bayes to build their model which could predict whether the content given as input is real or fake. This input is provided to the machine learning model through a web-based API which acts as a user interface to take input and display results. In the machine learning model for preprocessing of data, several techniques are used like count vectorizer, TF-IDF, etc. Now the algorithms of SVM and Naive Bayes do the prediction to a greater extent and they are much more efficient when used together.

The trajectory on which our model will be working is shown in the Fig. 2.

V.　CONCLUSION

According to the various papers studied and analyzed, we conclude that the Machine Learning approach is the appropriate solution to detect fake news. As observed from the researchers and papers of various authors Logistic Regression is the algorithm in machine learning that gets the best performance as well as it is simple to implement. SVM also shows similar performance. However, the best approach is the Ensemble method in Machine Learning. It gives the best result as well, despite the fact that the detection is highly dependent on Dataset

[8] Very Deep Convolutional Networks for Text Classification-Yann Lecun, Loc Barrault, Holger Schwenk, Alexis Conneau - 2016

[9] Long short-term memory (LSTM) recurrent neural network (RNN) for discharge level prediction and forecast in Cimandiri river, Indonesia-Y Sudriani1, I Ridwansyah1 and H a Rustini1

[10] Dey, Rahul & Salem, Fathi. (2017). Gate-Variants of Gated Recurrent Unit (GRU) Neural Networks.

[11] A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images-Md. ZabirulIslamMd. MilonIslam AmanullahAsraf

[12] Zhang, Ziqi & Robinson, David & Tepper, Jonathan. (2018). Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network. 10.1007/978-3-319-93417-4_48.
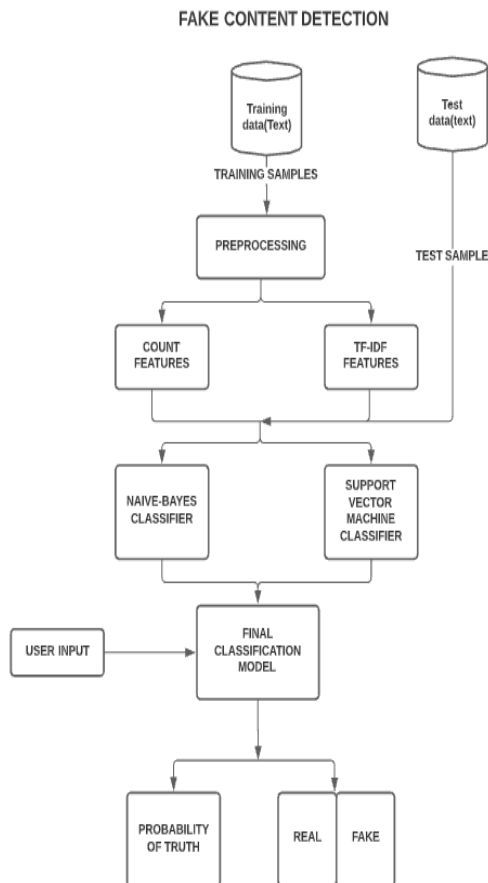
Fig. 2 Process Flow of our Proposed Solution.

REFERENCES

[1] https://www.westfordk12.us/wa/wright-library/pages/how-spot-fake-news

[2] Fakebuster: Fake News Detection System Using Logistic Regression Technique in Machine Learning- Muhammad Syahmi Mokhtar, Yusmadi Yah Jusoh, NoviaAdmodisastro, NorainiChe Pa, Amru Yusrin Amruddin

[3] Analysis of Classifiers for Fake News Detection- VasuAgarwala, H. ParveenSultanaa, SrijanMalhotraa *, AmitrajitSarkar

[4] A Machine Learning Approach to Fake News Detection Using Knowledge Verification and Natural Language Processing- Marina Danchovsky Ibrishimova and Kin Fun Li

[5] Fake News Detection Using Machine Learning Ensemble Methods-Iftikhar Ahmad, 1 Muhammad Yousaf,1 Suhail Yousaf, 1 and Muhammad Ovais Ahmad 2

[6] R. K. Kaliyar, "Fake News Detection Using a Deep Neural Network," 2018 4th International Conference on Computing Communication and Automation (ICCCA), 2018, pp. 1-7, DOI: 10.1109/CCAA.2018.8777343.

[7] Shyam, Radhey. (2021). Convolutional Neural Network and its Architectures. 12. 2021. 10.37591/JoCTA.