Fake News Detection System Using Modified Random Forest

Aman Jatain[1], Priyanka Vashisht[2]

[1,2] Dept. of Computer Science and Engineering, Amity University, Haryana Email: amanjatainsingh@gmail.com, Priyanka.vashisht@gmail.com

***Corresponding Author:** (Priyanka.vashisht@gmail.com

Abstract Although fake news has existed since before the World Wide Web, its prevalence has expanded tenfold with the availability of mobile technology and internet services. With an increasing number of internet users, people share billions of posts and articles on various social media platforms like Facebook, Twitter, WhatsApp, and Instagram. As a result, fake news spread quickly among millions of people and everyone began believing in it, so it needs to stop now more than ever. This fake news is meant for readership as a part of phycology warfare and their goal is profiting through clickbait. It can be a propaganda against a particular society or an individual or political party. It is highly challenging for a human to tell whether a news story is true or false, hence deep learning techniques are required to automatically identify fake news using various categorization algorithms. We emphasised the thorough work done by the researchers, the datasets, the overall architecture, and the many evaluation metrics they utilised to assess their model.

Keywords: Fake News Detection, Fake News Dataset, Machine Learning, Classification, Web UI, Deploy, Front-end, Back-end.

1. Introduction

Information dissemination has never before been possible because to the development of the World Wide Web and the rising popularity of social media platforms (like Facebook and Twitter), which are both readily available. More customers than ever are sharing and producing information because to the widespread use of social media platforms. While some of the news and information provided are false or misleading and have no bearing on reality, the question of whether it is true or fake has been raised. Automatic disinformation detection of written content is a challenging task; one strategy is to look at how bogus news is disseminated while keeping real news attractive. When making a determination regarding the veracity of an article, even an expert in a certain field must consider a number of factors. [1]

Exposing fake news is crucial in preventing its detrimental effects on people and society because it is being used to spread incorrect or rumoured information in an effort to alter individual behaviour. Fake news has quickly become a social concern, and with this growing interest, it is imperative. The two main components of the concept of fake news are authenticity and intent. Since authenticity refers to what is real or genuine and what is not, conspiracy theories are excluded from the definition of fake news since they are frequently impossible to show to be accurate or incorrect. The next phrase, "purpose," denotes that the information was published with the intention of misleading the reader for a specific goal, parties seeking to profit from it, or to set up a trap.

Fake news, defined as "a news piece that is purposely and indubitably untrue," is endangering our civilization more and more. As a result, many authors have previously suggested that text mining, which divides data into training and testing sets, is a crucial component of evaluating data mining models, machine learning, deep learning approaches, and textual data to forecast the news trustworthiness. After Facebook CEO Mark Zuckerberg's categorical denial that Meta had a bearing on the election's result, Meta and other online media sites have started to create various techniques for spotting fake news and preventing its spread. Deep learning and machine learning models perform better than classic text mining and machine learning techniques because they have more processing power and can handle enormous volumes of data. The current work has connections to several fields of study, including sentiment analysis and text classification.

Machine learning algorithms can be used to classify information to detect fake news. The essential components of model detection mostly rely on textual qualities that, when used with algorithms, separate bogus information from real. Knowing whether the words and tokens in news articles had a substantial impact on whether the news is real or fake is necessary to develop an accurate fake news detection model, and this can be done by TF-IDF vectorization. [2] In this paper, the classifier is then used to assess the associated performance. We construct a "ON" network to support the notion of utilising machine learning to identify bogus news. [3]

2. Related Work

The work of Rosas et al., Kleinberg et al., Lefevre et al., and Mihalcea et al. covered seven important domains and involved two fresh datasets for the detection of fake news. They provide a dual contribution. They begin by analysing the information on the linguistic distinctions between false and accurate news. Second, via a computer approach, they created bogus news detectors. Here, a Linear SVM (LSVM) classifier and five-fold cross validation were applied. The readability characteristics, followed by the linguistic features, served as the basis for the classifier for the first dataset, while the punctuation features, followed by the syntax features, served as the foundation for the accurate model for the second dataset.Further they tested their fake news detector with humans and compared the accuracy and they found humans are better at detecting fake news. Hence here the model outperforms.[4]

Zhou et al., Jain et al., Phoha et al., and Zafarani et al. model represents the news by sets of features, capturing both structural and style language and using these features they built a machine learning model using supervised learning. The experimental result based oncreal-world dataset achieved the accuracy of 88%[5]. On all four datasets, they applied a variety of machine learning algorithms, including Logistic Regression, SVM, MLP, KNN, Ensemble Learners (Random Forest, Bagging, Boosting and Voting Ensemble Classifiers), and Benchmark Algorithm (Perez-LSVM, CNN, Bi-LSTM Networks). They found that Ensemble Learners performed better than individual learners. [6]

Khanam et al, Alwasel et al, Sirafi et al and Rahid et al, estimated various supervised Machine Learning algorithms and made the research on the accuracy and the performance matrix of the predictions. The accuracy of the model was derived using NLP (textual analysis) and performed tokenization and feature extraction of the given text data. They used six machine algorithms (XGBoost, Random Forest, Naive Bayes, KNN, Decision Tree, SVM). The result shows the XGBoost and Random Forest got the same accuracy but Random Forest performance matrix i.e confusion matrix resulted in more false negatives than XGBoost.[7]

Sharma et al, Saran et al and Patil et al, performed the binary classification on the basis of accuracy, confusion (performance) charts. Also, they used two approaches: static and dynamic approach. In the static approach implementation was done using ML algorithms with vector features and found Logistic Regression achieved higher accuracy (65%) and higher precision, recall and f1

score. Also, their study shows that the accuracy of Logistic Regression can be increased by grid search parameter optimization. Whereas in the dynamic approach we used Passive Aggressive and yielded 92% accuracy.[8]

Aldwani et al and Alwahedi et al, designed a tool that detects and eliminates the online sites. The tool will look upon various features of the sites like syntactical structure, number of words associated with wording used in the tiles of sites, monitors the punctuation marks it will also examine the factors linked with sites like bonus rate, Higher the bonus rate the too will designate it as source of fake news. For this the paper used four popular algorithms (Bayes-Net, Logistic Regression , Random Forest, Naive Bayes) Based on the experimental results [9]

Shu et al., Sliva et al. Liu, Wang, and others, as well as Tang, They give a data mining viewpoint in their study on how to spot fake news on social media, which includes how fake news is defined in social and psychological theories. Naive realism and confirmatory bias are the two key elements that explain how easily users accept misleading messaging [10]

Gahirwal et al, analyse the data using two methods, first stance detection which is an important part of NLP which divides the article in various categories having some weights which later help in conclusion. Second Document similarity also known as id-idf -it checks the similarity of document and search results [11]

Kudari et al, V et al, BG et al and R et al, Designed a project that easily understands the difference between real and fake news. For this they performed comparative analysis over features that differentiate the fake news. The classifiers used in this research are naïve bayes and passive aggressive classifiers using TF-IDF Vectorizer and count vectorizer. The result shows that passive aggressive and TF-IDF vectorizer performed well with accuracy of 90% for this model.[13]

3. Methodology

3.1 Proposed Work

The method used is explained in this article. We have three different sections interconnected with each other by technology. On the basis of a machine learning classifier, the initial segment is static. The machine learning model was developed. Sci-kit libraries were used, along with Python. Data preparation was done using Python libraries. For machine learning algorithms, the Sci-kit library is utilised since it has built-in, readily accessible machine learning methods. The second section is our front end. We built the basic UI for our fake news prediction using Html and Css that shows the output in a more presentable way. Third, to connect front end to back end we used Flask, a web framework that connects machine learning model and front end and gives the desirable result.

3.2 System Design

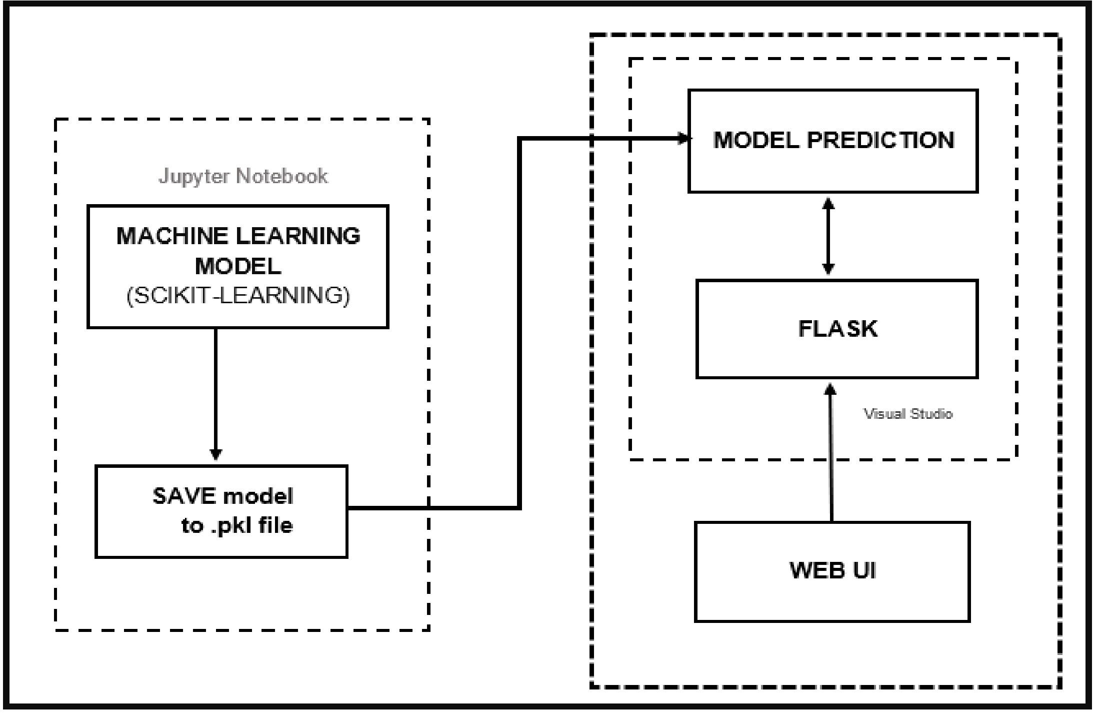

Fig 1: System Design for Fake News Detector Web UI

3.3 System Architecture

The machine learning architecture includes loading of data followed by cleaning or preprocessing of data, extraction of features using vectorizer etc, splitting of data into train-test sets, applying machine learning algorithms after this we can predict or visualise our data. Machine Learning architecture is self explanatory as seen in figure 2 shown below.

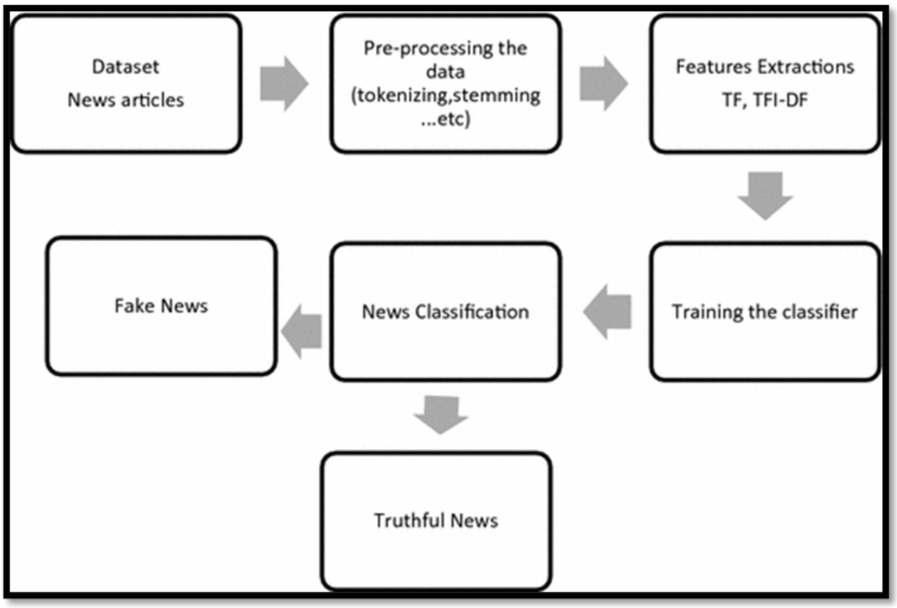

Fig 2: Proposed Architecture of System

3.4 Web UI

The web part here shows us whether the entered news is real or fake. The entered news here is from the data we obtained after preprocessing of news text. The front end is made using Html and Css.

3.5 Flask

It is used as a backend framework for our web application Flask is used to deploy machine learning models. For Flask to connect to machine learning models we need to save the machine learning model using .pkl (pickle) extension and also the vectorizer in .pkl form. And will generate app.py file, main application file where all code resides and it binds everything together.

3.6 Random Forest

Random Forest is a technique used in ensemble learning for classification and regression problems. Also called random decision forests. The random forest method creates a lot of decision trees. Every decision tree develops a single class and finally bootstraps the votes in order to increase the Random Forest method's accuracy.

3.7 Fake News Diagnosis Techniques

Few Techniques are mentioned below:

3.7.1 Removing The Stop-Words

Mainly some Languages use these words to connect words, it makes us know about the tense or indicate the tense of sentences. In addition, using these words in a sentence does not add much to the context of the sentence so even if we remove them, we can still understand the context.

3.7.2 Tokenization

The action of breaking text into smaller pieces is called tokenization. NLP can represent words, special characters, and numbers in sentences as tokens.

3.7.3 Vectorization

Vectorization is the mapping of words to their corresponding real vectors, which is used for word prediction and word similarity/meaning finding.

The most used vectorizers are:

- Count Vectorizer: It's the one that's easiest to understand.
- Hash Vectorizer: It is made to use the least amount of memory feasible. The method's flaw is that after the features are vectorized, it is impossible to restore their names.
- Term Frequency-Inverse Document Frequency Vectorizer, also known as TF-IDF. In other words, rather than only considering a term's frequency within a particular document, the load allocated to each token now ideally considers the term's frequency over the entire corpus.

4 Data Set

Dataset for Fake News are available easily on platforms like kaggle, UCI. For our project we used a dataset from Kaggle “FAKE NEWS” Dataset that predicts whether the news is real or fake. It has 20800 rows and 5 attributes with some null or missing value. Its following attributes are Id, Title, author, text, are features of the dataset and label attribute is our target attribute also our output on the basis of these features.

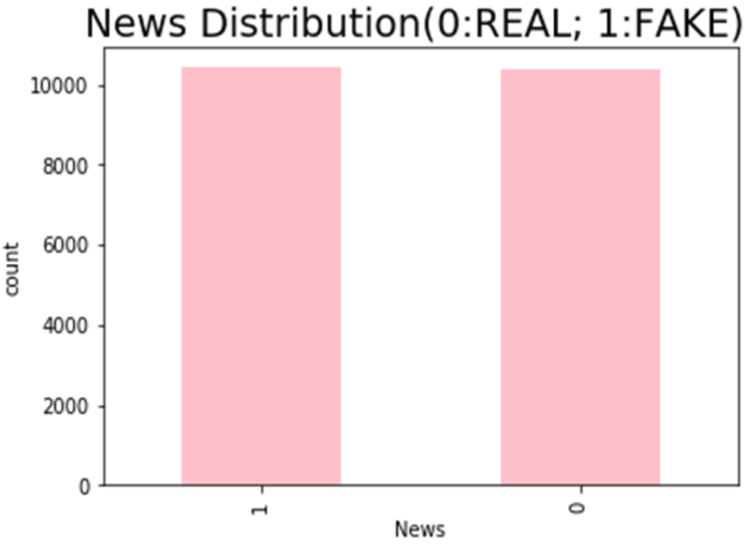

Fig 3: Fake News Distribution as Fake and Real

5 Implementation

ML Model Implementation

In this section we will discuss the implementation of the proposed ML model.

Step 1. Load the dataset (mentioned in 4th section) and clean the data, by cleaning means fill the null values.

Step2. We will preprocess the data using a stemming process.

Stemming Process:- A word is stemmed when its root word is eliminated. Prefixes and suffixes will be eliminated because they don't add much significance to text data. As it cuts down on the number of words as much as possible and improves the performance of our model, this is one of the crucial processes in data pre-processing. The Porter Stemmer Function was imported for the stemming procedure.

Step3. After preprocessing of data, extract the features of data by converting the textual data into numerical form, using the TF-IDF vectorization process.

TF-IDF Vectorization: Number of terms The term "inverse document frequency" refers to a method of counting the number of times a word appears in a document or text paragraph. The repetition of words implies that the word is important or has a special meaning and tf assign a particular numerical value to it and sometimes we have word which has repetition but is not important or any impact so here comes in action it eliminates such words from paragraph or document. Together they make a feature vector.

Step4. After all the preprocessing, split the data into test-train sets 20%-80% respectively. For data to split in similar proportions we set the target variable as stratify.

Step5. Training the model, using Random Forest Classifier. Accuracy obtained is shown in Table 1 given below.

Step 6. Evaluating the model obtained by evaluating its performance using confusion matrix and classification report. Refer Table 2 for confusion matrix and Table 3 for classification report.

Step7. Save the machine learning model created using Random Forest using pickle in a .pkl format and also save the vectorizer created using TF-IDF vectorizer in .pkl format.

Train Data Accuracy	Test Data Accuracy
100%	99.32%

Table 1: ML Model Accuracy

Actual//Predicted	0	1
0	2058	19
1	9	2074

Table 2: Confusion Matrix

	Precision	Recall	F1-score	Support
0	1.00	0.99	0.99	2077
1	0.99	1.00	0.99	2083

Table 3: Classification Report

5.1 ML Model Deployment

After Building machine learning model we will deploy the ML model, creating an end to end application to make it in a more usable form.

5.1.1 Front End

The first section of ml model deployment is to design a UI. The UI here is created using HTML and Css technology. The UI has three main pages 1) HOME PAGE 2) PREDICTION PAGE 3) KNOW MORE PAGE The Home page is the front page giving us a brief overview about the Fake News Detection System Using Machine Learning. From the home page we can lead towards a prediction page where the user can enter the news to ensure it is fake or real. And the third page is a brief overview of developers, who created the project. The images for the same are attached below.

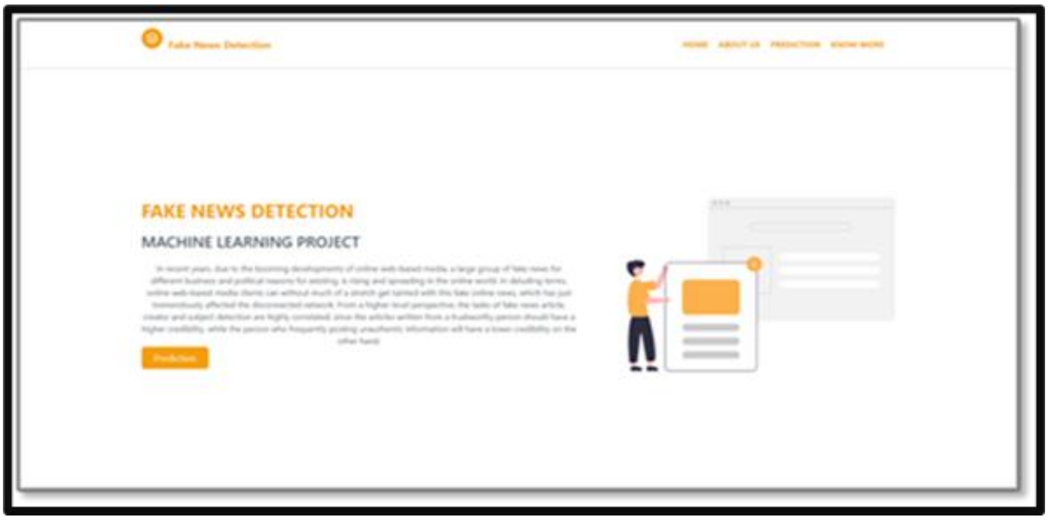

Fig 4: Home Page

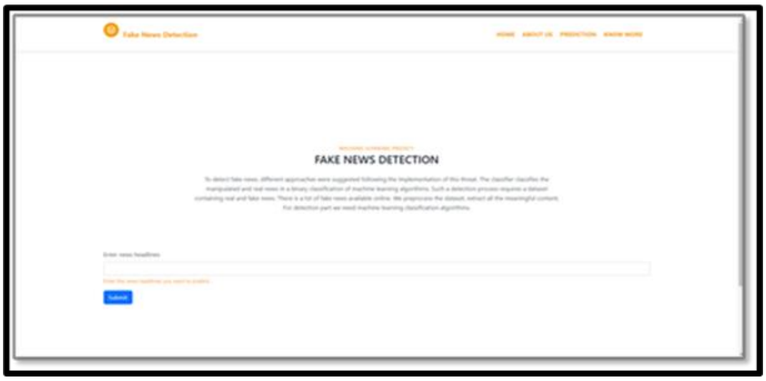

Fig 5: Prediction Page

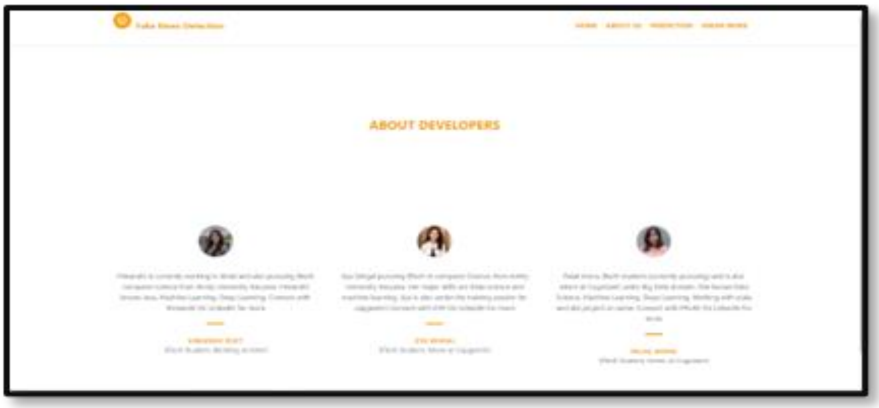

Fig 6: Know More Page

5.2.2 Back End

The flask is used to create the web application. Here it connects machine learning model and UI making it interactive for end users. The following steps show the connection.

Step1. Build the app.py in the project directory. It's a python file and a main application file where the code resides.

Step 2. Import all the needful libraries like Flask, request, render_template etc

Step 3. Create an object of the Flask class named it app (in this project its app) it will handle the request coming from the browser and will provide the appropriate responses.

Step4. Pass the Url in @app.route and map it with the front-end and render the html page.

a. For the index/home page, we passed '/' in app-route, created a home function and rendered the index.html page.

b. For the prediction page, we passed '/prediction' in an app-route that maps the prediction method with '/prediction' url. The prediction method does all the preprocessing , generates the final feature vector and runs the model on it and gives us the final prediction .Since our target variable was in binary form (0 or 1) so here we used if-else to show in text form (real or fake).

c. For the know More page, we passed '/contactus' url in an app-route. It will render to Know More Page(contact-us page).

Step5. To run the code, open the terminal. Create the virtual environment variable: virtualenv news Step6. Activate the virtual environment variable : news/scripts/activate

Step7. Run the app.py file : python app.py

Step8. Follow the link and the website will open in web browser.

To test the website add the news text (we got after pre-processing) and click on submit button it will give us result saying "NEWS HEADLINE IS —> FAKE/REAL".

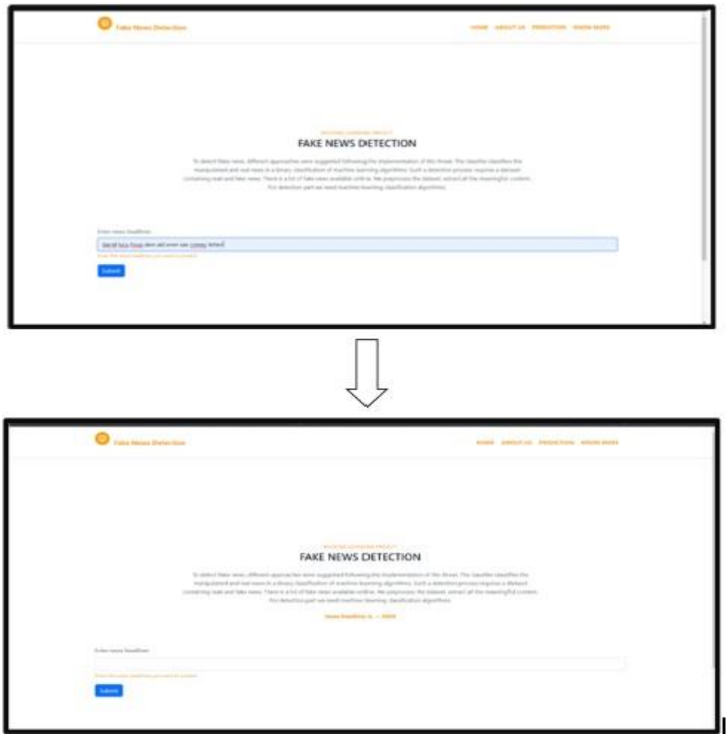

Fig 7: Output

6. Conclusion

A sizable portion of readers prefer to read news on social media rather than in traditional news media due to social media's rise in popularity and usage over the past several years. With this in mind, a lot of publishers use social media and the Internet in general as a breeding ground for swiftly spreading rumours and propaganda, which has a bad effect on the neighbourhood. The proposed work shows the working of a fake news detection system and deploying it on the web. We trained the dataset, taken from kaggle "Fake news Detection Dataset" using Random Forest algorithm. Random Forest is capable of predicting the outcome with an accuracy of 99.32%. The model we created is imported in flask via pickle and through flask we are able to deploy our pre-processed vector and model to a web using front end technology. Presenting the outcome on the web is an attractive way to showcase our model in an attractive way and in future we can make the website user friendly where users can put any news for that we have fetch data using APIs and build proper backend for that. We may also increase the dataset amount for better performance and may use other different Machine Learning Algorithms and compare the accuracy and performance as well.

References

[1]. Anjali Jain, Harsh khatter, Avinash Shakya, "A Smart System For Fake News Detections Using Machine Leaning," 2019 International Conference on Issue and Challenges in Intelligent Computing Techniques(ICICT), doi:10.1109/ICICT46931.2019.8977659.

[2]. KushalAgarwalla, Shubham Nandan, Varun Anil Nail, D.Deva Hema, ''Fake News Using Machine Learning and Natural Language Processing", International Journal of Recent Technology and Engineering ISSN:2277-3878,Volume-7,Issue-6,March 2019.

[3]. [3] Y. Wang et al., "EANN in FND-2018.pdf," pp. 849–857, 2018.

[4]. Veronica Perez-Rosas, Bennett Kleinberg, Alexandra Lefevre and Rada Mihalcea. "Automatic Detection of Fake News." ResearchGate,2017.

[5]. Xinyi Zhou, Atishay Jain, Vir V. Phoha, and Reza Zafarani, Syracuse University, USA. "Fake News Early Detection: A Theory-driven Model." Digit. Threat.: Res. Pract. 1, 2, Article 12 (June 2020), 25 pages.

[6]. Iftikhar Ahmad, Muhammad Yousaf, Suhail Yousaf, Muhammad Ovais Ahmad and Suhail Yousaf. "Fake News Detection Using Machine Learning Ensemble Methods." Volume 2020, Article ID 8885861, 11 pages

[7]. Z Khanam, B N Alwasel, H Sirafi and M Rashid. "Fake News Detection Using Machine Learning Approaches." IOP 2020.

[8]. Uma Sharma, Sidarth Saran, Shankar M. Patil. "Fake News Detection using Machine Learning Algorithms". NTASU,2020 Conference Proceedings, International Journal of Engineering Research & Technology (IJERT).

[9]. Monther Aldwairi, Ali Alwahedi. "Detecting Fake News in Social Media Networks". The 9th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN 2018).

[10]. Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang and Huan Liu. "Fake News Detection on Social Media: A Data Mining Perspective". ACM SIGKDD Explorations Newsletter, 2017.

[11]. Manisha Gahirwal, Sanjana Moghe, Tanvi Kulkarni,Devansh Khakhar and Jayesh Bhatia. "Fake News Detection". International Journal Of Advance Research, Ideas And Innovations In Technology (IJARIIT) ISSN:2454-132X, Impact factor:4.295 (Volume 4, issue 1).

[12]. Victoria L. Rubin, Niall J. Conroy, Yimin Chen, and Sarah Cornwell, "Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News." Proceedings of NAACL-HLT 2016, pages 7– 17, San Diego, California, June 12-17, 2016.

[13]. Jayashree M Kudari, Varsha V, Monica BG, Archana R "Fake News Detection using Passive Aggressive and TF-IDF Vectorizer". International Research Journal of Engineering and Technology(IRJET) Volume: 07| issue sep 2020.

[14]. Jatain, A. Chaudhary, S. Batra, P. Bhaskar, S. (2021) Rest web services: An elementary learning. Research Journal of Engineering and Technology, 12(3), 75-78.

[15]. Mor, P.   Bhaskar, S. (2021) Enabling Technologies and Architecture for 5G-Enabled IoT. Blockchain for 5G-Enabled IoT: The new wave for Industrial Automation, 223-259.

[16]. Jatain, A. Chaudhary, S. Nagpal, P. Bhaskar, S. (2021) Cloud Storage Architecture: Issues, Challenges and Opportunities. International Journal of Innovative Research in Computer Science & Technology (IJIRCST) ISSN, 2347-5552.

[17]. T Joy, D. Kaur, G. Chugh, A. Bhaskar, S. (2021) Computer Vision for Color Detection. International Journal of Innovative Research in Computer Science & Technology (IJIRCST) ISSN, 2347-5552.

[18]. Jaglan, V.   Bhaskar, S. (2021) Locking Paradigm in Hierarchical Structure Environment. Advances in Mechanical Engineering: Select Proceedings of CAMSE 2020, , 653-661.

[19]. Nanda, A. Gupta, S. Bhaskar, S. (2020) A Comprehensive Survey of Machine Learning in Scheduling of Transactions. 2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(2020), 740-745.

[20]. Bhaskar S.   Bhaskar, S. (2020) Study of locking protocols in database management for increasing concurrency. 2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(2020), 556-560.

[21]. Bhaskar S.   Bhaskar, S. (2020) Reducing Complexity of Graph Isomorphism Problem. International Journal of Innovative Research in Computer Science & Technology (IJIRCST) ISSN, 2347-5552.

[22]. Bhaskar, S. (2020) CGVL:An Hierarchical Locking Mechanism . International Journal of Control and Automation, 12(6), 725-743.

[23]. Agarwal, R. Bullah, H.R. Prabhakar, A. Jatain, A. Jaglan, V. Bhaskar, S. (2020) Parkinson‘s Disorder: Taking a Step towards Homogenizing Machine Learning and Medical Science. International Journal of Psychosocial Rehabilitation, 24(4),

[24]. Bhaskar, S. (2020) A Review on the Concept of Deep Learning. International Journal of Innovative Research in Computer Science & Technology (IJIRCST), ISSN, , 2347-5552.

[25]. Jaglan, V. Sethi, N. Bhaskar, S. (2020) Distortion Free Image Generation. Grenze International Journal of Engineering and Technology, 15(8), 460-466.