Fake News Detector: FND

Pravin P. Kharat
Department of IT
Vidyalankar Institute of Technology,
Mumbai, India

Sanyam Sharma
Department of IT,
Vidyalankar Institute of Technology, Mumbai, India

Sayali Tambe
Department of IT,
Vidyalankar Institute of Technology, Mumbai, India

Deepali Vora, PhD
HOD, Department of IT,
Vidyalankar Institute of Technology, Mumbai, India

ABSTRACT
The idea focuses on providing the information on whether the news is real or fake and with distinguished information about the content or news headline provided by the user into the developed system. The user is a client or a customer, gets particular news or headline from any source which can be news providing application, news blog, website and social networking site; and upload the news content in the proposed system which is a web-based application.

After uploading the news content or headline the user clicks the submit button which is available on the website. Then the content is processed accordingly and the metadata of the content is extracted. There are derived parameters on the basis of which calculation of news authenticity is done. The system also uses the Naive Bayes and Term Frequency-Inverse Document Frequency (TFIDF) algorithm which is used to predict the probability of different classes, based on various parameters or attributes. TFIDF i.e. Term Frequency-Inverse Document Frequency is an algorithm used to transform the text into a meaningful representation of numbers. Based upon the parameters and using the respective algorithm the news authenticity is calculated and the result is uploaded. The final result states whether the news is real or fake news and is developed upon the parameters, metadata and algorithm which simultaneously gives the respective result to the user.

Keywords
Fake news, social network, metadata, classification, extraction, Naive Bayes, TF-IDF, crawler.

1. INTRODUCTION
Fake news can be defined as a pseudo-news or junk news which is false information or content spread via different broadcasting media. This false information is mainly created to mislead people or to cause ravage to a person, organization, etc., or to gain financial benefits. Easier ways to get news nowadays are from sources like media outlets, newspapers, journalists where they follow defined strict codes of practice; this is the traditional method. But the increase in internet network has changed the certain aspect of publishing and sharing the news and information with less editorial standards and regulations. On the other hand, the news is being manipulated by various networking sites based on personal opinions or interests.[1] Fake news causes immense damage in different ways such as misleading people with sharing false information, this can create racists ideas or can damage people's sentiments, which can give rise to violence among people i.e. causing real-life impacts, etc. This rapid spread of fake news is a serious problem calling for AI solutions.[3] To overcome this, the system is developed which identifies the news authenticity and identifies it as Fake News or Real News. This is done by analyzing the news content where analysis is done on the basis of defined parameters. In the proposed system the user directly enters the data or the news content (news headlines) and enters into the systems search bar and gets the result. The proposed system works upon the two main algorithms i.e. Naive Bayes algorithm and TFIDF algorithm. The defined parameters help to identify the news content and increases the accuracy of the system. Parameters considered are website data, website validates, the similarity of content, timestamps, reviews, and grammatical analysis. These parameters calculate the news authenticity and generate the final result as fake or real news.

2. LITERATURE REVIEW
Through this section, summarization of some of the existing research work is done to detect the news authenticity and build a model according to the existing applications.

2.1 S. Ananth, Dr. K. Radha, Dr. S. Prema, K. Nirajan, "Fake News Detection using Convolution Neural Network in Deep Learning", *IJIRCCE*, (2019): 1-4
In this paper, the author mainly focuses on classifying the news article into certain or not by identifying the fake news using various models and classifiers and predict the accuracy of the different models. A data set is proposed combined of fake news and real news and implementation of data set are done using technologies like machine learning, natural language processing, deep learning. NLP and machine learning techniques are mainly used in this to implement models and compare which models will give more accurate results.

2.2 Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, Huan Liu, "Fake news detection on social media: A data mining perspective", ACM Explorations Newsletter, (2018): 22-36
The authors of this paper are mainly considering the social media platform. To detect the authenticity of news fake news characterizations on psychology and social theories is done, it also gives a comprehensive review of fake news using evaluation metrics. Auxiliary information i.e. user information is also used and later exploited to check the engagement of the user with the fake news, this produces big data that is incomplete and unstructured and noisy.

2.3 Sneha Singhania, Nigel Fernandez, and Shrisha Rao, "3HAN: A Deep Neural Network for Fake News Detection", International Institute of Information Technology, (2018): 1-5
The author in this paper explains the uses of 3HAN i.e. three-level hierarchical attention network. 3HAN generates a news

vector which is a distinguished formed sentence. 3HAN gives an intelligible result using various models.

2.4 Shuo Yang, Kai Shu, Suhang Wang, Renjie Gu, Fan Wu, Huan Liu, "Unsupervised Fake News Detection on Social Media: A Generative Approach", Department of Computer Science and Engineering, USA, (2019): 1-4

In this paper fake news detection is done in an unsupervised manner, the platform considered is social media. The authenticity of news is performed by considering the truths and users' credibility wherein it treats truths of news and users' credibility as latent random variables, and exploit users' engagements on social media to identify their opinions towards the authenticity of news.[4] Naive Bayes algorithm, user credibility and Gibbs sampling algorithm are used to find the final result.

2.5 Supanya, Prabhas, "Detecting Fake News with Machine Learning Method", Department of Computer Engineering Faculty of Engineering, University Bangkok, (2018): 1-3

In this paper, the author describes the fake news and fake news detection is performed by using various machine learning algorithms. This research use three methods to classify the believable and unbelievable message from Twitter, there are Naïve Bayes, Neural network, and Support Vector Machine (SVM).[5] Classification of data is done after the data is cleaned using the normalization method. Using these algorithms gave high accuracy to the system.

2.6 Shaheen Karodia, "Fake News Detection on Twitter Proposal", University of Cape Town Rondebosch Cape Town, (2017): 1-3

In this, the author mainly trying to detects fake news on the social media platform Twitter. Present two approaches, one based on logistic regression, the other based on Boolean crowdsourcing algorithms.[6] Using user credibility, evaluation and classification method authenticity of news is identified. A labelled data set is developed which contains news and the above methods are performed.

2.7 Eugenio, Gabriele, Marco L.D., Stefano, and Luca A, "Automated Fake News Detection in Social Networks", Technical Report UCSC-SOE-17-05 School of Engineering, University of California, (2018): 1-4

In this author mainly focuses on fake news detection on the social media platform. In this, the fake news detection is performed by using two classification techniques i.e. logistic regression and Boolean crowdsourcing algorithms and the final result is evaluated and defined. These results suggest that mapping the diffusion pattern of information can be a useful component of automatic hoax detection systems.[7]

3. PROPOSED SYSTEM

3.1 Problem statement

To cater to the needs of the identification of fake news or false information, the proposed system uses web crawler, web scraping and text analysis to identify fake news effectively.

3.2 Developed System

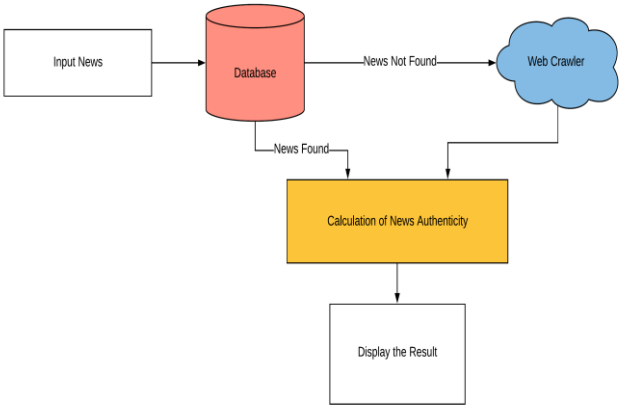

Fig. 1: Components of Proposed System

In the proposed system initially, the news contents are acquired from the user. These news contents are then entered into the system as an input for further processing. This information is further checked into the existing database of the developed system to check whether the same news information is available or not and if the news content is available final result is calculated in further steps. Naive Bayes algorithm is mainly used to identify results from the database. In the database of the system, news data is classified into two types trained and test data, when the data is trained the test data is allotted to the set which has similar characteristics with the set. Feature extraction process is performed to identify the characteristics using various attributes. Based on these content attributes, different kinds of feature representations can be built to extract discriminative characteristics of fake news.[2] Now, the Naive Bayes algorithm is used to detect the accuracy of fake news. Where in TFIDF is mainly used to find the important word into the content and store the news accordingly, it also identifies unimportant words that are not required and each word is allocated with its count. When the news content is not found in the database then the crawler starts searching the news information into the news websites as shown in fig 1.

A detailed description of each of the components in the system diagram:

Input News: The user will give news content as an input as a query in the search box of the web-based application. Then further it will be checked in the database and later the web crawler will do its necessary function.

Database: In the database, it consists of various news articles and stored on the basis of different independent parameters such as timestamp, author, title, text/content and label as fake or real.

Web Crawler: Web crawler crawls through the specific websites: searches the news content using the TFIDF method in which each word in the news content is made to be counted and credibility and similarity of this news is checked. It also extracts the timestamp and details of the news, reviews, website validity, etc.

Calculation of News Authenticity: Depending on the biased percentage of each parameter average (threshold values) of all is calculated and displayed as a single bar with a biased percentage at a particular timestamp.

Display of Result: The calculated result is been displayed on the web-based application.

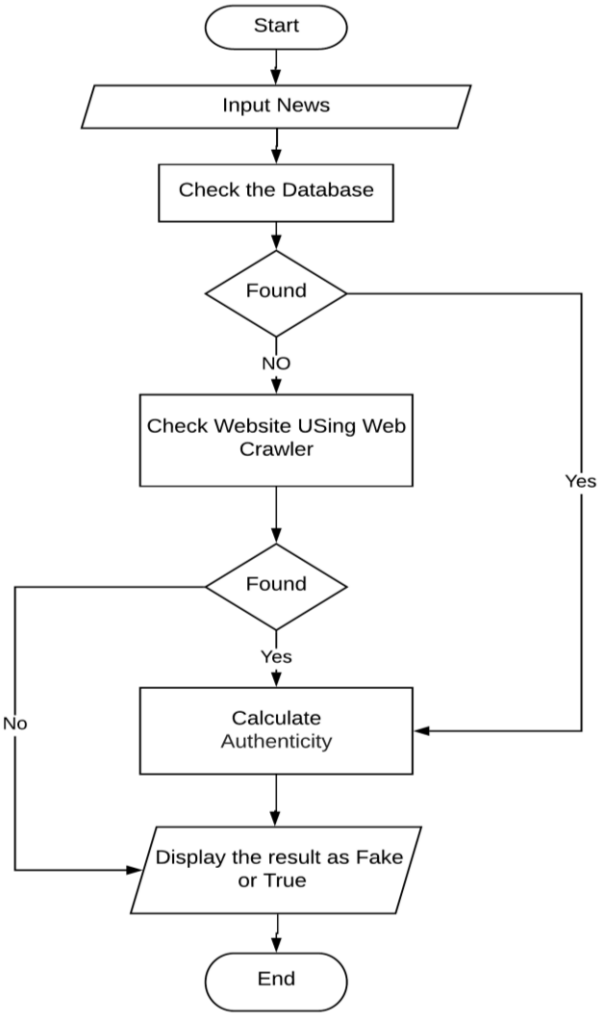

Fig. 2: Flow Chart.

A list of news websites is derived on the basis of credibility given to them and hence the news content is searched in this news providing websites only. When the required information is fetched, the calculation of news authenticity is performed. This is done using various parameters and Naive Bayes Algorithm. While calculating the authenticity, each parameter gives threshold value which later is calculated by summation method and an average value is generated this average value is compared and the final result is generated stating whether the news is Fake or Real.

4. NAIVE BAYES AND TF – IDF

Naive Bayes classifier is a probabilistic classifier in machine learning. It is used to check the authenticity of the news by using Naive Bayes algorithm. This algorithm is used for classification of the text. TF-IDF is used to measure the frequency of words in the content. Using TFIDF count of significant and insignificant text is determined.

Using both Naive Bayes algorithm and TFIDF method, the system will check whether the news is false news or Real news

5. METHODOLOGY

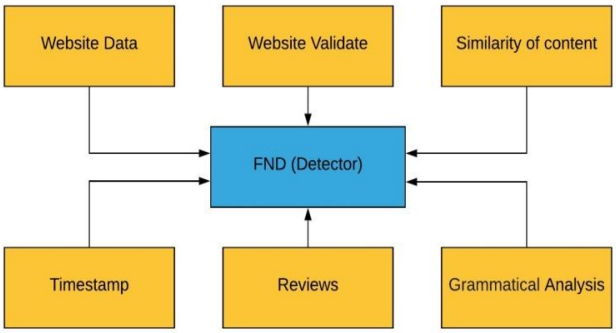

Fig. 3: Parameters used in proposed system.

5.1 The main parameters on which the Fake News Detection system is based on:

Data Stored:

The data is collected from the news websites also data is fetched from different news websites. The crawled data is stored in JSON format and the data from websites is stored in CSV format. The folders or files having fake or real news were divided accordingly. A threshold value is generated in this step.

Website Validation:

N no. of websites are formed into a list that is crawled by the web crawler accordingly when the news is given as an input to check the news authenticity.

Similar Contents:

The news content entered by the user is compared with the news content found on the news websites i.e. comparison of the news articles is performed.

Grammatical Analysis:

Analysis of Text for grammatical errors and for writing patterns is strictly verified into this parameter.

Reviews and Timestamp analysing:

In this parameter, the reviews of the news content are examined and the news article publishing time and date is being classified, compared and stored. In this Naive Bayes algorithm is used for the classification of news.

5.2 The working of the system is explained in the following steps as shown in Fig 7:

Step 1:

The user will give the input news content; it will consist of content of the news article or headlines of the news article and the source of the news article which is discretionary. Similarly, the trust score of the news source will be calculated and will be stored within the system for further calculations

Step 2:

The designed system FND will further inspect the grammatical errors and writing pattern of the input news content (headline), and certain threshold values will be stored of this parameter which will be further used to calculate the authenticity of the news.

Step 3:

Simultaneously, the system will search for similar articles on different URLs from the predefined list of websites. After Scraping the URLs, the contents found are stored into the database of the system, and further grammatical errors and writing style is checked of the stored content.

Step 4:

Further, the input news article is compared with the scraped news article which is stored into the database and a new threshold value is generated and stored.

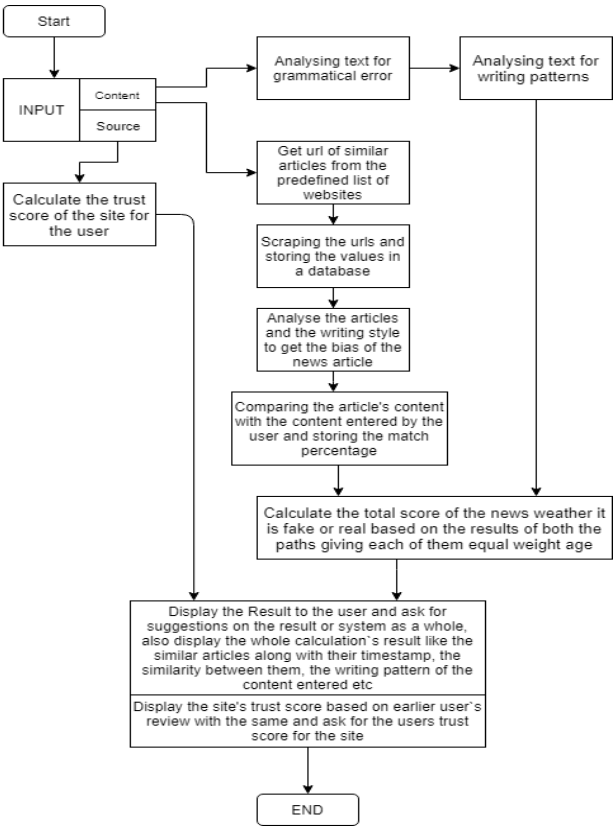

Fig. 4: Detailed Block Diagram.

Step 5:

Date, time and place where the news article is published is also compared with the fetched news article and a new threshold value is again generated for this parameter and is stored.

Step 6:

Reviews of the people (comments) and reviews of the news publishing expertise are also stored into the database and are used to detect whether the news is Fake or Real. The whole process for each parameter to generate each new threshold value is performed simultaneously.

Step 7:

Later the summation of all the threshold values is calculated and stored into the system and displayed accordingly to the user using a web-based application. The result will display every aspect which is compared with the news article and the percentile of news authenticity will be displayed stating whether the news is Fake or Real.

6. OUTPUTS

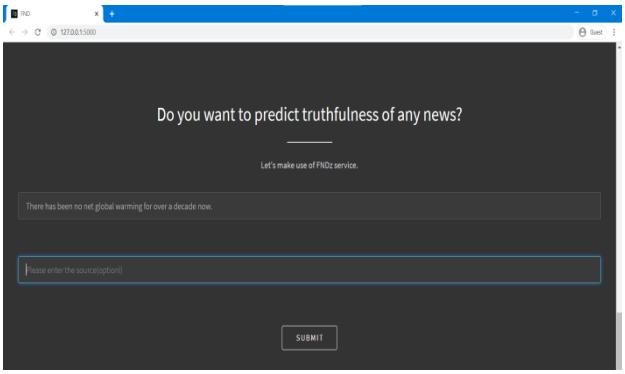

Fig. 5: Screenshot of query search bar

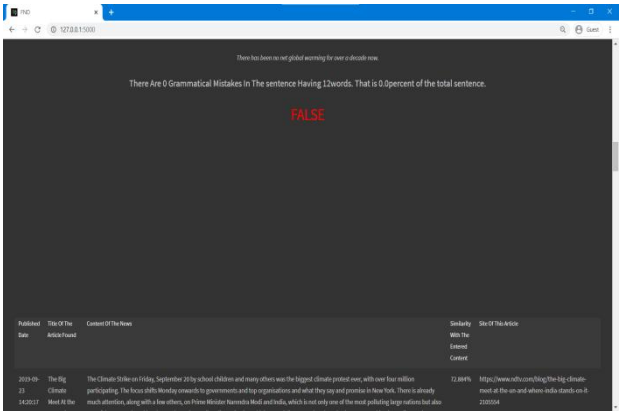

Fig. 6: Screenshot of Actual Output displayed on the User Interface

In the above fig 6. The output of the searched query is derived with the result declared as False news i.e. fake news. It also gives information about the related news its published date, similarity contents, the website of the article, title and content of the news article.

Table 1: Performance comparison on dataset using Confusion matrix.

Predict actual	False	True	Barely - True	Half – True	Mostly – True	Pants – Fire	Total
False	345	17	9	390	176	0	937
True	219	29	3	345	239	0	829
Barely – True	284	7	17	374	152	0	834
Half – True	290	21	4	508	234	0	1057
Mostly – True	249	20	11	401	269	0	950
Pants – Fire	189	2	0	161	52	1	405
Total	1576	96	44	2179	1116	1	5012

Table 2: Precision, Recall, and F1-Score.

Actual	Precision	Recall	F1 – Score	Support
False	0.27	0.27	0.27	249
True	0.17	0.00	0.01	208
Barely - True	0.14	0.00	0.01	212
Half - True	0.22	0.68	0.34	265
Mostly – True	0.17	0.15	0.16	241
Pants – Fire	0.00	0.00	0.00	92

7. APPLICATIONS OF FND

The proposed system will help the society to deal with the problem of growing number of cases of fake news and act as a way to combat the effects of these cases, like riots, misinformation, hatred towards a community, etc. FND system is much more user – friendly than the existing system where FND is a web-based application and can be accessed from anywhere and anytime. Developed system not only states the news authenticity but it also gives overall details about the news content entered by the user. After each search of news content and display of the result, news data is stored into the database of the system for future use. For the next search of the same news, it is searched into the database which makes it a more efficient and faster web-based application.

8. CONCLUSION

In this paper, system proposed is a fake news detector which is using different parameters and algorithms. The proposed system can be efficiently used by the user since the system is user – friendly and gives instant results. It also reduces disputes, riots, misunderstandings, created due to the fake news. As compared to the existing system, the generated result from the proposed system is faster, accurate and also provides additional information about the searched query(news) as shown in Fig. 5 and Fig. 6. The final results demonstrate the effectiveness of the proposed system and it can be used widely by users to detect whether the news is Fake or Real news.

9. FUTURE SCOPE

This system is considered for searching news authenticity as required by the user. This system can also be updated and can be used in real-time system application such as twitter, Facebook, Instagram, etc. wherein it will give result by fetching the data directly from any of the mentioned application and will state whether the news showing on the application is real or fake on the same application which is being used. This will need to access the data of the application and at the same time synchronization is required to be maintained. Even more parameters and models can be used for defining the final results.

10. ACKNOWLEDGEMENT

We wish to express our sincere gratitude to Dr. Deepali Vora, HOD of Information Technology Department, for her guidance and encouragement in carrying out this project.

We also thank faculty of our department for their help and support during the completion of our project. We sincerely thank the Principal of Vidyalankar Institute of Technology for providing us the opportunity of doing this project.

11. REFERENCES

[1] S. Ananth, Dr. K. Radha, Dr. S. Prema, K. Nirajan, "Fake News Detection using Convolution Neural Network in Deep Learning", *IJIRCCE*, (2019): 1-4

[2] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, Huan Liu, "Fake news detection on social media: A data mining perspective", *ACM Explorations Newsletter 19.1* (2018): 22-36

[3] Sneha Singhania, Nigel Fernandez, and Shrisha Rao, "3HAN: A Deep Neural Network for Fake News Detection", *International Institute of Information Technology*, (2018): 1-5

[4] Shuo Yang, Kai Shu, Suhang Wang, Renjie Gu, Fan Wu, Huan Liu, "Unsupervised Fake News Detection on Social Media: A Generative Approach", *Department of Computer Science and Engineering*, USA, (2019): 1-4

[5] Supanya, Prabhas, "Detecting Fake News with Machine Learning Method", *Department of Computer Engineering Faculty of Engineering, University Bangkok*, (2018): 1-3

[6] Shaheen Karodia, "Fake News Detection on Twitter Proposal", *University of Cape Town Rondebosch Cape Town*, (2017): 1-3.

[7] Eugenio, Gabriele, Marco L.D., Stefano, and Luca A, "Automated Fake News Detection in Social Networks", *Technical Report UCSC-SOE-17-05 School of Engineering, University of California*, (2018): 1-4.