An Efficient FAKE NEWS DETECTOR

Ms. Smita Vinit Bhoir,
Department of Computer Engineering,
Ramrao Adik Institute of Technology,
Affiliated by University of Mumbai
Navi Mumbai, India.
smitapatilbe@gmail.com

***Abstract*- Social media now plays a vital role in shaping the feeling of people in favor or against a government or organization. Therefore, an effective method is an absolute requirement to recognize the feeling of any posting in social media[1]. Fake news is a phenomenon that affects our social life greatly, particularly in the political world. Fake news identification is a growing field area of interest, which faces some difficulties because of the small amount of resources available (i.e. databases, published literature)[2]. The primary objective of the proposed system is to develop an efficient system that can anticipate whether a piece of information is fake based purely on its content, thus addressing the issue from a purely natural language processing (NLP) perspective. The article analyzes different classification methods and suggests an active NLP method to identify FAKE NEWS. The main advantage of the proposed model is that it classifies fake news with good accuracy.**

Keywords*: *Fake NEWS, Natural Language Processing, Support Vector Machine, Naïve Bayes, Random Forest.

I. INTRODUCTION

Social media plays a major role in trying to influence people in favor or against a government or organization and the opinions of the society. Fake news is a phenomenon that affects our social life significantly, particularly in the political world. The enormous amount of information is created with various social media formats on the social networking. Whenever some event has happened, through social networking, many people are discussing it on the web. They are searching for or retrieving and discussing news events as the daily routine. Nonetheless, when searching and downloading, very large volumes of news or posts made users face the problem of data overloading. Inaccurate information sources bring attention to a dose of false news, hoaxes, rumors, theories of conspiracy, and misleading news. When the unexpected events occur, fake news is also being broadcast that creates confusion because of the nature of the events. Some individuals understand the fact from the incident, while most people believe that the news has been transmitted through various media from their credible friends or relatives. The fake news emerges from the source's disinformation, confusion, and amazing material. These are difficult to detect once they receive news information, whether to believe or not. Likewise, when people find series of fake news in society, they also don't believe in real news unless they're part of that event. Thailand, for example, is in a tropical area. Therefore, the rain is almost all year round, causing massive flooding in Thailand. The Thai Meteorological Department broadcasts the forecast information in order to alert the public in advance and protect their property. Nevertheless, with inaccurate misunderstandings, the unpredictable news could be spread rapidly around the world[3]. The reports that the dam is destroyed and overflowing in areas that have not actually occurred may be indicators of flood conditions. It is claimed that the water is not overflowing, but that the city is potentially flooding. Such rumors damage the real disaster preparation[4]. Distributing data before the internet was more expensive and there were far clearer interpretations of what was news and media, making it easier to control and self-regulate. But the advent of social media has degraded several walls that have stopped the dissemination of false or misleading news in democracies. It has allowed anyone to create and disseminate information, especially for those who have proven to be most skilled at how social networks function. Facebook, Twitter and other social media have made it possible for people to exchange information on a much larger scale than ever before, while publishing platforms and online news portals have made it easy for anyone to create a dynamic website. In short, the barriers to fake news creation have been removed. Since its early days, however, hoaxes and falsehoods have been synonymous with the Internet, however orchestrated, systemic disinformation campaigns, sometimes connected to governments, have only arisen in the last two years, and their impact on democracy and culture has been scrutinized. The goal is to build an efficient model that can determine if a piece of news is false based solely on its content, thus approaching the problem from a strictly NLP perspective. An important part of the goal is to analyze and report the results of multiple different design implementation and provide a findings review. From an NLP perspective, this phenomenon offers an interesting and valuable opportunity to identify patterns that can be coded in a classifier. In proposed system, we ignored all other signals (e.g., the source of the news, whether it was reported online or in print, etc.), and instead focus

978-1-7281-4514-3/20/$31.00©2020 IEEE

only the content matter being reported. The main advantage of the proposed system is that the method can classify posts or comments with good accuracy.

The major objectives of proposed system are:

1. To use supervised learning in order to build an efficient model.
2. To improve accuracy of prediction whether a news article or piece is real or fake
3. To approach the problem only based on content, by extracting each word from the post or article then match those with a dictionary or bag of words for classification.
4. To identify the sentiments expressed in a new post or comment and extract each word of the posting, then match those with the dictionary words for classification using HYBRID MODEL. Later classify it as FAKE or REAL NEWS using MACHINE LEARNING techniques.

II. LITERATURE REVIEW

A. The paper [4] presented a new method of text classificationin which Naive Bayes and computer genetics have been used. In this, the association rules are used to replace the terms in order to recover the feature sets. Naive Bayes is then applied to the set of features followed by a single Genetic algorithm definition for the final classification test. The result showed that the algorithm proposed requires less training data and less computational time.

B. A classifier combination is proposed in paper[5] using the classifier Multinomial Naive Bayesian (MNB) and the classifier Bayesian Networks (BN). In this, the outcome of two classifiers is combined by taking an average of the individual classifier determined probability distribution. Results showed better performance and greater accuracy.

C. There are many benefits to detecting relationships in the query text as it enables the use of context interpersonal information to produce potential answers and find additional evidence to support the rating of passages[6]. In the DeepQA question-answering system, this paper introduced two approaches to the extraction and scoring of broad-domain relationships, one based on manual pattern description and the other based on statistical methods for pattern elicitation, using a novel transfer learning technique, i.e. related topics. The proposed system was a multidisciplinary template for the detection of fake news and included information from domains such as online news, previous research and social media.

D. In the paper[7], authors conducted a qualitative analysis of the data on about 25,000 tweets of the educational problems of engineering students and studied that many students face problems such as heavy study load, problems of diversity, lack of social commitment, negative emotions and lack of sleep. Taking these analyzes into consideration, researchers introduced a hybrid model by incorporating two of the most popular machine learning algorithms. The example shows that practice time is significantly reduced and shows an increase in classification reliability compared to individual Naive Bayes classifier and Support Vector Machine.

III. ANALYSIS

In table I, analysis of various classification models is given. The analysis is done based on algorithms used, accuracy of classification, features considered, advantages of model, research gaps identified and applications.

IV. PROPOSED WORK.

In paper[7], researchers performed a qualitative analysis of the academic problems of engineering students and also found that many students face issues such as heavy study load, problem of diversity, lack of social interaction, negative emotions and lack of sleep, etc. This design decreased learning time dramatically and increased the accuracy of classification. Compared to Naive Bayes classifier and Support Vector Machine, this hybrid model worked well when implemented separately.

(a) NAIVE BAYES has two main disadvantages-firstly, it makes the assumption that functionality words are independent from each other, which is practically impossible, and that their accuracy is less than that of other machine learning classifiers.

(b) On the other hand, SVM has many benefits, such as noisy data sensitivity, speed and accuracy is much higher than other classifiers for machine learning.

(c) Random forest is a supervised learning algorithm which creates multiple decision trees and merges them in order to improve accuracy and stability of prediction. The advantage being, this algorithm is a handy and easy to use one, since it uses its default parameters to produce a result and it can be used for both classification and regression tasks. But, the more number of trees makes the algorithm slower and incapable of real time problems. They also require more computational resources and are less intuitive.

Any kind of misinformation or hoaxes spread via sources like any social media, news websites or any traditional press, with the intention of misleading people, which is often caused by reporters who pay the various sources for the stories, is called Fake News. Hence, fake news detection is an important research area and the objective of our system is to build an efficient model that can predict whether given news is fake on the basis of its content. Naive Bayes classifier is merged in the proposed system with Support Vector Machine and Random Forest to improve the classifier's performance to address the classification problem.

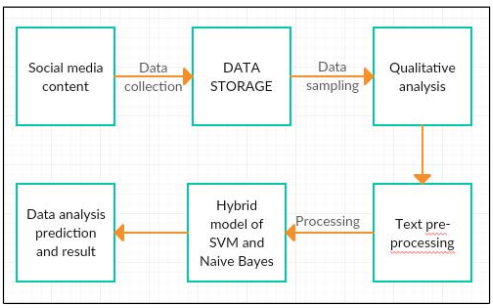

Figure 1. Proposed System

978-1-7281-4514-3/20/$31.00©2020 IEEE

978-1-7281-4514-3/20/$31.00©2020 IEEE

Table I:Literature Review

Sr. no.	Paper Title	Algorithms used	Accuracy	Features considered	Advantages	Research Gaps	Dataset	Applications
1.	Evaluating Machine Learning Algorithms For Fake News Detection [1]	Naive bayes and random forest	67.89%	Bigram Term Frequency-Inverse Document Frequency, Normalized frequency of parsed syntactical dependencies.	This model uses natural language processing and TFID techniques precisely.	They evaluated models using absolute probability thresholds, which may not be the most reliable for models where probability scoring is not well calibrated. While TF-IDF performs better, they are overfitting to topics/ terms important in the news cycle	Signal media	TFID and PCFG probabilistic context free grammar detection.
2.	A Hybrid Model for text classification [7]	Naive bayes and support vector machine	97.01%	Sleep problem, negative emotion, lack of social engagement, heavy study load	Hybrid model gives the highest accuracy, Precision and Recall as compared to that of individual classifiers viz. Naive Bayes Classifier and Support Vector Machine for the data set..	It doesn't work for images videos on social media on twitter but only for text.	Manual dataset of 25000 Tweets	T Text classification and semantic analysis of twitter student posts
3.	Study of Hoax News Detection using Naïve Bayes Classifie r In Indonesian Language [8]	Used naive bayes, SVM algorithm	78.6%	Class index, review,source link ,word	It uses automatic hoax detectionand classifies various accuracy analysis based on % of test and train data ratio, dataset is available freely	The dataset is only available for certain number of year Implementation is time consuming.	Manual dataset of 250 hoax and valid news	Hoax news detection in Indonesia n language.
4.	Automated Fake News Detection Using Linguistic Analysis and Machine Learning Vivek Singh, Darshan Sonagra, Karthik Raman, and Isha Ghosh [3]	Used SVM model for prediction	87%	Word count, clout, authenticity	used multiple features to create comprehensive detection modules for fake news rather than using a single feature	Needs large dataset for greater accuracy levels.	Kaggle fake news dataset	Fake news detection
5.	Fake News Detection Using Naïve Bayes Classifier [2]	Naive bayes classifier	74%	Link of post, text of news and label of text	It uses simple effective naive bayes classifier	It doesn't use hybrid or complex models	Manual dataset of faceboo k news posts	Fake news detection

978-1-7281-4514-3/20/$31.00©2020 IEEE

978-1-7281-4514-3/20/$31.00©2020 IEEE

The proposed system shown in Fig.1 aims to create a FAKE NEWS detection using hybrid model. The process starts with pre-processing of data and removal of stop words and refining data, followed by extracting features from the data to be classified, and then classifying the data as FAKE news or REAL news.

The proposed methodology is summarized as follows:

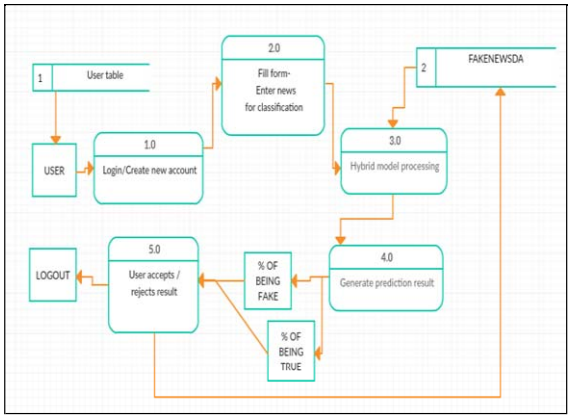

Figure 2. Workflow of proposed system

Pre-processing and refining of data: refinements such as stop-word removal, tokenization, lower case, segmentation of sentences, and removal of punctuation.

Feature extraction using 5 feature types: Five different features selection methods, namely

a) Word Count

b) Authenticity

c) Clout

d) Length

e) Tone

V. RESULTS AND DISCUSSION

The experimental study of a dataset carried out on each classifier independently and then applied on proposed model and results obtained as shown in Table II.

Table II: Experimental Results

Parameters / Models	SVM	Naïve Bayes	Random Forest	Proposed Model
FALSE Positive	0.53	0.58	0.63	0.68
True Positive	0.60	0.62	0.62	0.62
Accuracy	57%	60%	63%	78%

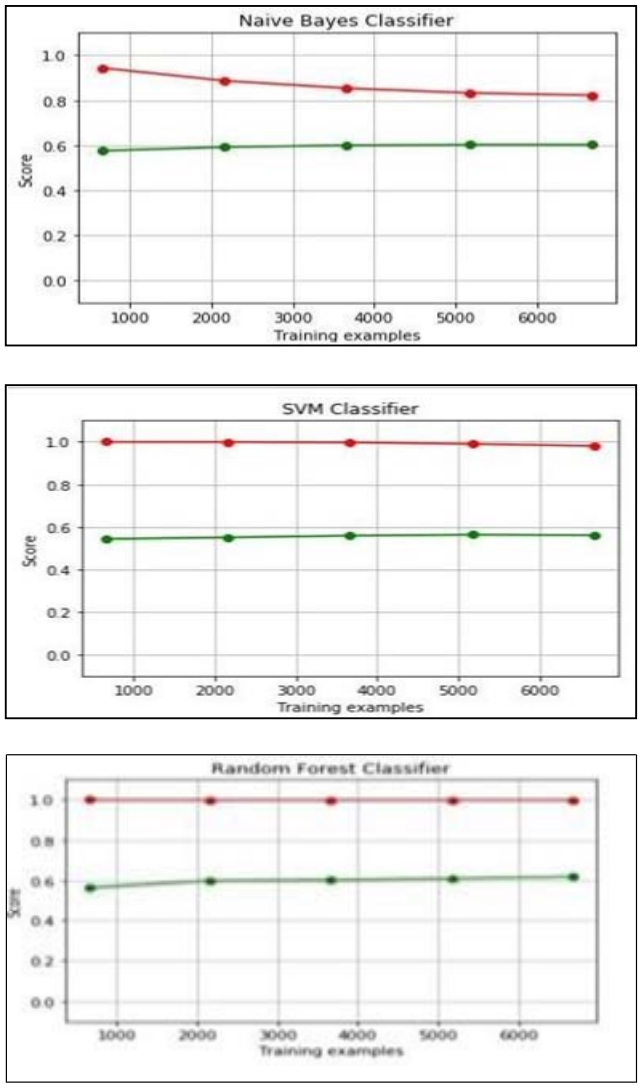

Fig.3 Resultant graphs of Individual Classifier

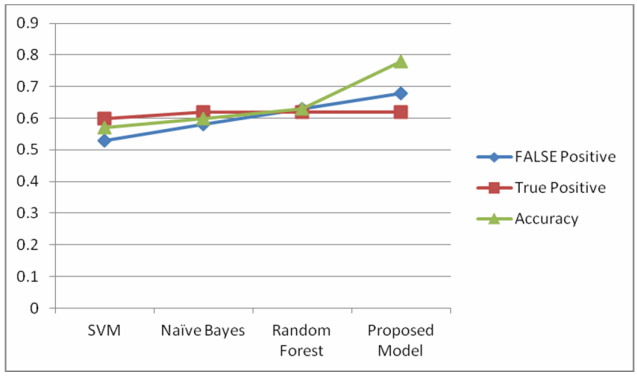

Fig.4 Comparison of different classifiers with proposed system

From experimental results we have analyzed the proposed model is efficient for classification of fake news with better accuracy.

978-1-7281-4514-3/20/$31.00©2020 IEEE

978-1-7281-4514-3/20/$31.00©2020 IEEE

The user interface of our proposed system looks like the following:

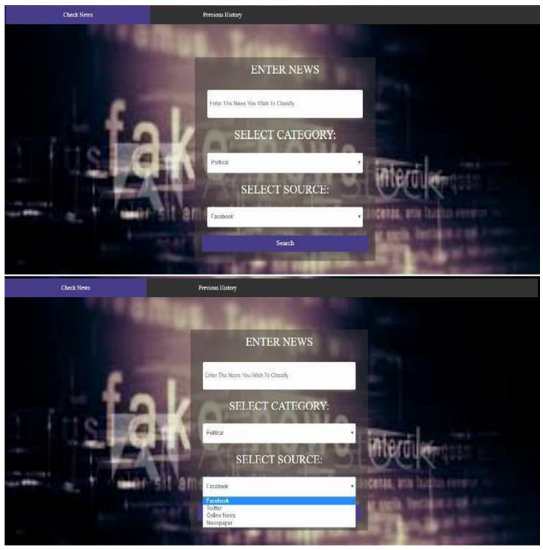

Fig 5. Graphical User Interface

VI. CONCLUSION

Over the past decades, the intensity of Internet use has greatly increased, thus giving hands to Fake news, which is a growing problem as it is the rumors that make it too difficult to identify the fact in content. The panacea is therefore to have a well-efficient and complete proof system that identifies the different patterns in news that can help us identify whether or not it is genuine. The proposed system aims to identify fake news using hybrid model with the help of machine learning.

REFERENCES

[1] Shlok Gilda, "Evaluating Machine Learning Algorithms For Fake News Detection", IEEE Conference, 2017.

[2] Mykhailo Granik, Volodymyr Mesyura, "Fake News Detection Using Naïve Bayes Classifier", 2017 IEEE First Ukraine Conference On Electrical And Computer Engineering, (UKRCON), 2017.

[3] D. S. K. R. Vivek Singh, Rupanjal Dasgupta and I. Ghosh, "Automated fake news detection using linguistic analysis and machine learning," in International Conference on Social Computing, Behavioral-Cultural Modeling, & Prediction and Behavior Representation in Modeling and Simulation (SBPBRiMS), 2017, pp. 1–3.

[4] Kamruzzaman, S. M., and Farhana Haider, "A hybrid learning algorithm for text classification," 3rd International Conference on ElectricalComputer Engineering (ICECE 2004), pp.1009.4574,2010.

[5] A. Rahman and U. Qamar, "A Bayesian classifiers- based combination model for automatic text classification," In Software Engineering and Service Science (ICSESS), 2016 7th IEEE International Conference,2016.

[6] J.W. Murdock, J. Fan, A. Lally, H. Shima, and B. K. Boguraev, "Textual evidence gathering and analysis," IBM Journal of Research and Development, vol. 56, no. 3.4, pp. 8:1–8:14, May 2012.

[7] Priyanka Ingole, Smita Bhoir, A. V. Vidhate, "Hybrid Model for.Text Classification", 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), March 2018.

[8] Inggrid Yanuar, Risca Pratiwi, et.a al.,"Study of Hoax News Detection using Naïve Bayes Classifier In Indonesian Language," International Conference on Information Communication Technology and System, 2017.

AUTHORS PROFILE

Ms. Smita Vinit Bhoir.
M.E Computer Engineering, PhD (Pursuing). She is working as Assistant Professor, Department of Computer Engineering Ramrao Adik Institute of Technology, Navi Mumbai, India. She is having 10 years of R & D experience. Published 11research papers, delivered 22 technical talks, organized 17 workshops and training programmes. Her area of specialization Machine Learning, Data Science, Web and Networking.

978-1-7281-4514-3/20/$31.00©2020 IEEE

978-1-7281-4514-3/20/$31.00©2020 IEEE