International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 05 | May www.irjet.net p-ISSN: 2395-0072

Rumor Detection System using Machine Learning

Dr. Anil Kr. Dubey, Apoorv Singhal, Sarthak Gupta

[1]*Associate Professor, Dept. of Computer Science* Engineering, *ABES Engineering College, Uttar Pradesh,*
[2]*India Student, Dept. of Computer Science* Engineering, *ABES Engineering College, Uttar Pradesh, India*
[3]*Student, Dept. of Computer Science* Engineering, *ABES Engineering College, Uttar Pradesh, India*

Abstract - *In a rapidly growing world, everyone has access to the internet and are connected to social media anyhow. A pool of information is floated on these websites but the validation and verification has no legitimate source. This is where rumors come in. Rumors are hoaxes created to decide or bring a radical shift in the opinion of the masses and the effect of it are mostly observed in Politics especially elections and on Social Media. So, in order to get rid of this issue there is a need for a detector of rumors which can precisely convey that the information is fake or real. Through this Rumor Detection System, we worked on the algorithms like Multinomial Naive Bayes, Gradient Boosting and Random Forest with particular datasets to implement them and to get closer to more precise results of a rumor. The accuracy achieved on using Multinomial Naive Baye's is around 90.4% when used Random Forest it was 86.5% on using Gradient Boosting it is around 88.3%.*

***Key Words*: Rumour, Fake News, Prediction, Detection Naive Baye's classifier, Machine Learning Application**

1. INTRODUCTION

There will be a chip in 2000 Rs. note that can trace where the note is with our exact coordinates and will come into existence from 10th November 2016. This rumour spread like wildfire on social media and various news channels as well supported in the same. The effect of it was not only seen in common people but all over the globe.

Government was very much moved by many such rumours regarding currencies, few general awareness rumours regarding Kerala relief funds and many more fake information started floating everywhere around. Thus, affecting the financial and political welfare of the country. This paper gives us a systematic view and the process involved in detecting whether the news is valid or not, algorithms implemented and result acquired is embedded in this paper. We chose a dataset to train other datasets for which after importing the dataset we analyzed the datasets and various classification algorithms such as Multinomial Naive Baye's, Random forest and gradient boosting were used to generate outputs.

1.1 OBJECTIVE

Our center point of focus or main objective is to find rumors, categorizing the content and predicting its legitimacy is a really tough task. It is needed to build a model that can differentiate between the 2R's "Rumors" and "Reality" on social media platforms like Facebook, Instagram, Hike, Twitter etc. These platforms are the biggest carriers of spreading rumors and misleading people about the truth all over the globe.

1.2 PROBLEM IDENTIFICATION

Rumors are where individuals or organizations intentionally publish hoaxes, propaganda and other misinformation and present it as factual. The effect of Rumors has been observed in changing decisions overnight be it an election, or government policies or verdict of a long term case. A misleading statement or video with relevant images and video imbibed are floated as Reality all over the internet. Demonetization was one such event that made us think of this system as common people were highly affected with all the FAKE information. Being a sufferer of it we thought of removing this problem from roots and for it machine learning algorithms proved to be our biggest tool.

1.3 USE OF SOFTWARE

The software used in this rumor detection system is Anaconda by importing python libraries like Tensorflow, Numpy, and Matplotlib.

2. RELATED WORK

Numerous number of attempts have been tried to check the truthfulness of the statement. Around 2018, 3 students who used to study in Mumbai College from a well reputed college named Vivekananda Education Society's Institute of Technology, their research paper on fallacy detection. In this research paper they mentioned about the role of social media in catalyzing rumors all over the globe. Fake news are those piece of information uploaded by individuals or organizations deliberately to with a simple agenda of spreading rumors and get benefited socially or politically.[1] Tools and technologies used by these college students to detect rumors were basic NLP (Natural Language Processing) techniques, artificial intelligence and Machine learning. [1]Moreover, from this [1] literature survey we got the idea of the block diagram to be followed in processing of the whole function.

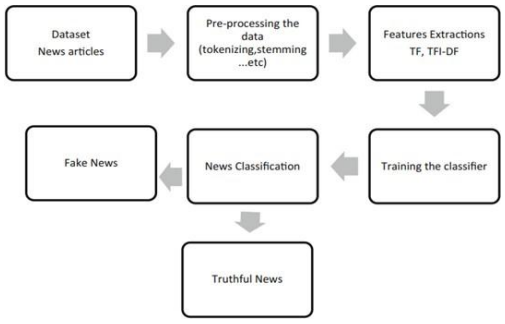

Fig-1: Block Diagram for processing the datasets

Out of handful of options of algorithm we confirmed to use Naïve Baye's [4] and Random Forest [6] as accuracy from this algorithm according to these literature survey was quite high but then too not satisfactory and they mentioned these two to be very efficient.

We thought of not to choose the SVM and Bi-LSTM as in literature survey[5] when they were tested on large range of data sets of around 12.8k which was a Benchmark in world of dataset, a free accessible dataset is available to all for detection of fake news The dataset which was named LIAR was larger than any other dataset that are available for public use.[5] This mainly focused on political statements so that leader's statement could not be molded for any communal fights or crisis rumors.

Dataset Statistics	
Training set size	10,269
Validation set size	1,284
Testing set size	1,283
Avg. statement length (tokens)	17.9
Top-3 Speaker Affiliations	
Democrats	4,150
Republicans	5,687
None (e.g., FB posts)	2,185

Table 1: The LIAR dataset statistics.

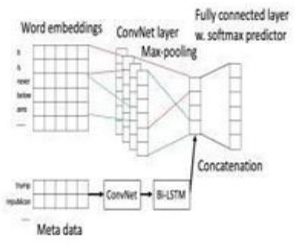

Figure 2: The proposed hybrid Convolutional Neural Networks framework for integrating text and meta-data.

Fig2: LIAR Dataset and CNN framework for text

The student of HCMUT that is Ho Chi Minh City University of Technology which is situated in Cambodia performed their research on rumors back in 2017; the student who performed this was Nyugen Vo. The algorithm and mechanisms used was Bidirectional GRU with Attention mechanism the original founder of the attention mechanism which was used in this project was given by Yang el al. Not only were this some Deep learning algorithms of machine learning also used such as Auto-Encoders, Convolutional Neural Network and Generative Adversarial Network. One of the Stanford University students also contributed to this topic and published a report on fake news detection. Natural Language Processing ideology and some other deep learning algorithms were considered. Signal Media News was the dataset provider which helped in this report. [3] Another work in this field took place when kaggle a dataset website presented a challenge named the "FAKE NEWS CHALLENGE"[7], three students from Chennai took opportunity to use this challenge by implementing algorithms such as Support Vector Machine(SVM), KNN, Decision Tree, Random Forest to bring out better results. While conducting their experiments they used Natural Language Processing (NLP) technology to obtain the outputs.

On implementation following accuracies are achieved, by SVM they managed to get 75.5%, through KNN it was 79.2%, whereas Decision Tree generated an output percentage of 82.7%, above all it was Random Forest that proved to be the best with 90.7% accuracy as shown in the graph plotted below.[7]

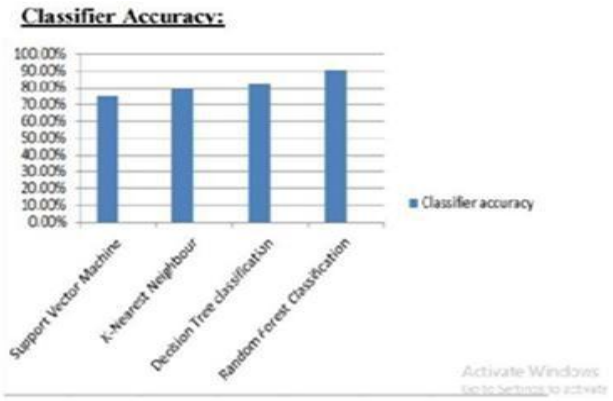

Fig3: Accuracy of different classifier algorithms

In this work of eliminating fake news from the world, three more youngsters from Punjab contributed in this movement. Kaggle supported them in accessing the dataset for their work while they chose to use Support Vector machine (SVM) and Hybrid Classifier as their algorithms for implementation

work. Data split process was taken place in such a manner that 60% of data was used to train the algorithm; hence it is the training dataset while the left out 40% used in testing is the testing dataset.[8]

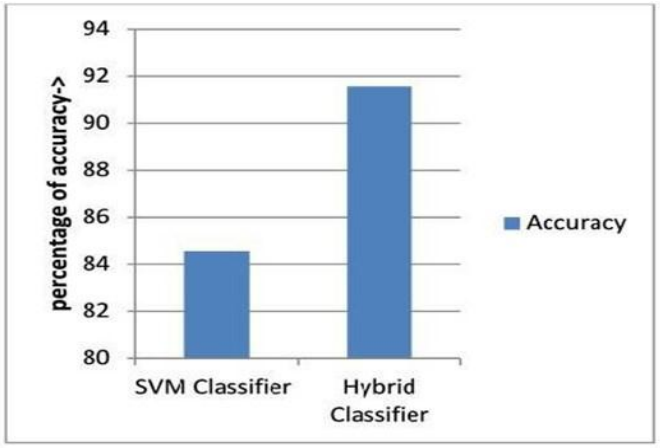

Fig4: Percentage accuracy by SVM and Hybrid Classifier

Surprisingly, the results showed a great difference of 8% between the two algorithms and analyzed that hybrid classification is a great blend of KNN and random Forest tree model which make it much more superior than SVM. California State University student took a charge on this topic of fake news detection and the method used by them are semantic analysis, Naïve Baye's and SVM.[9] Linguistic Cue approach and Network Analysis approach in Linguistic cue approach the study of different communicative behavior helped researchers to detect the deception in the model. Linguistic Cue method consist of data representation, deep syntax, semantic analysis and sentiment analysis, not only this it is also focuses on opinion mining.

Network Analysis Approach mainly targets content and hence it is content based approach. It consists of three fast checking method, first in the line is expert oriented method, second comes computational method while third is crowd sourcing method.

A different approach was observed in a literature survey where neural network was used, the reason of choosing this particular technology was as others were using classification tasks that make it quite difficult to get accuracy of the related news. First of all, they preprocessed the data by STOP WORD REMOVAL followed by Punctuation removal and Stemming. After this they used, word vector representation, moving to Bag of Words lead by TF-IDF vectorizer. When it comes to data splitting this time they used 67% as training data and 33% as the testing data.

A series of three models are used in the research work,[10] first being the "Tf-Idf on unigrams and bigrams with cosine similarity fed into dense neural Network", second on the list is "BoW with multilayer perceptron" and the third is "BoW with cosine similarity fed into dense neural network".

Table 12. Comparison between best performing model and our model

Model Description	Accuracy
Tf-Idf on unigrams and bigrams with cosine similarity fed into dense neural network (our model)	**94.31%**
BoW with multilayer perceptron [6]	92.46%
BoW with cosine similarity fed into dense neural network [5]	88.46%

Fig5: Comparison between best performing models

3. PROPOSED METHODOLOGY

The above papers helped us understand the various opinion mining and sentiment analysis techniques which will help us detect whether the news released is legitimate or not. The rumor detection model is divided into three step as:

Step I: Training the model

Step II: Rumor detection using the model
Step III: Efficiency and Accuracy Display

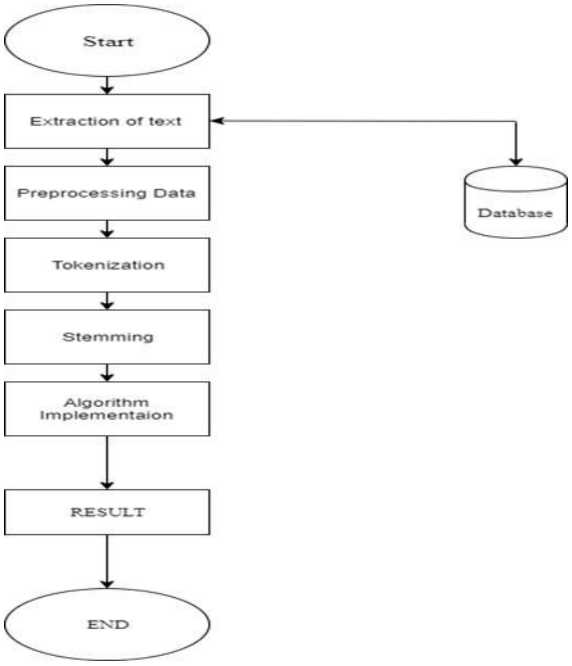

Fig6: Flow Chart of the methodology followed

We implemented all our algorithms and methodology on the software Anaconda Prompt. The data Sets were imported from kaggle and random datasets were trained to bring out results. After training and normalizing the data we used three algorithms of Machine Learning. Steps followed on data sets chosen.

4. RESULT DISCUSSION

After analyzing the literature surveys we read and taking in consideration the various algorithms frequently used in them. The SVM came as a highlight in almost all the cases and it terribly failed as it was unable to hold the noisy data. Moreover, when we thought of using KNN it proved to be a lazy learner and this is what stopped us from choosing that too, lazy learner is all about not learning anything from the training data. Though the decision tree delivered result but it proved to be very unstable, minute changes in data give rise to a massive change in the structure of the optimal decision tree. Moreover they are less precise. A pictorial representation of the same could be seen with the accuracy chart through the figure below.

Machine Learning Technique	TP	FP	FN	TN	Classifier accuracy
Support Vector Machine	884	43	421	553	75.5%
K-Nearest Neighbour	779	184	211	727	79.2%
Decision Tree classification	806	161	167	767	82.7%

Fig7: Result of different research paper with accuracy

These less efficient results inspired us to choose three different algorithms that are Multinomial Naïve Bayes, Gradient boosting and Random Forest.

On implementation of the algorithms we acquired a much higher efficiency on larger datasets with a difference of around 10-12 % on comparing them with SVM and KNN. Multinomial Naïve Baye's proved to be best in detecting the rumors in the category of true and false. As observed, Gradient Boosting is not used by many people researching in this field but on testing significant results prove it to be a tough competitor in this list of algorithms, as accuracy was much higher. Random Forest gave a considerable result of 86.5% which is also a satisfactory result value.

Table -1: Results Acquire by us with accuracy

Algorithm	T P	FP	FN	TN	Efficiency
Multinomial Naïve Bayes	32 24	711	52	39 76	90.4%
Gradient Boosting	38 40	365	71 5	432 7	88.3%
Random Forest	37 43	365	81 7	427 1	86.5%

We plotted these algorithms through matplotlib and get these results

```
Accuracy with Multinomial Naive Bayes:   0.904
```

```
cm = metrics.confusion_matrix(Y_test, pred, labels=['FAKE', 'REAL'])
plot_confusion_matrix(cm, classes=['FAKE', 'REAL'])
```

```
Confusion matrix
```

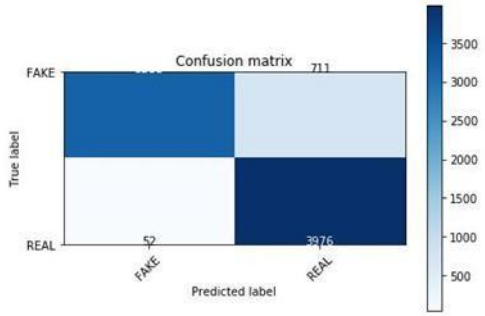

Fig8: Confusion matrix of Multinomial Naïve Bayes

Confusion Matrix of Multinomial Naïve

Bayes True positives (TP) = 3224 False Positive (FP) = 711

False Negatives (FN) = 52
True Negatives(TN) = 3976
Classifier accuracy = (TP+TN)/ (TP+TN+FP+FN) = (3224+3976)/ (7963)=90.4%

```
Accuracy with Gradient Boosting:   0.883
```

```
cm = metrics.confusion_matrix(y_test, pred, labels=['FAKE', 'REAL'])
plot_confusion_matrix(cm, classes=['FAKE', 'REAL'])
```

```
Confusion matrix
```

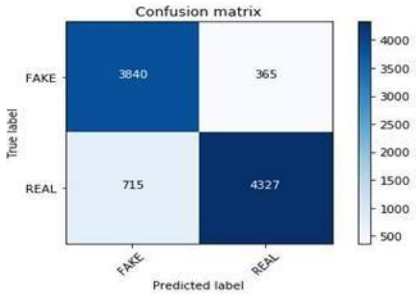

Fig9: Confusion matrix of Gradient Boosting
Confusion Matrix of Gradient

Boosting True positives (TP) =
3840 False Positive (FP) = 365
False Negatives (FN) = 715
True Negatives(TN) = 4327

Classifier accuracy = (TP+TN)/ (TP+TN+FP+FN) = (3840+4327)/ (9247) = 88.3%

```
Accuracy with RandomForestClassifier:   0.865
```

```
cm = metrics.confusion_matrix(Y_test, pred, labels=['FAKE', 'REAL'])
plot_confusion_matrix(cm, classes=['FAKE', 'REAL'])
```

```
Confusion matrix
```

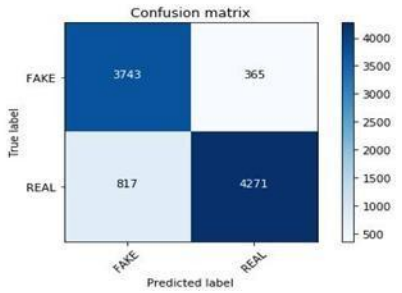

Fig10: Confusion matrix of Random Forest
Confusion Matrix of Random

Forest True positives (TP) =
3743 False Positive (FP) = 365
False Negatives (FN) = 817
True Negatives(TN) = 4271

Classifier accuracy = (TP+TN)/ (TP+TN+FP+FN)=(3743+4271)/ (9196) = 86.5%

5. CONCLUSIONS

Two algorithms were not satisfactory to give the desired results. For more precision we used Random Forest as our implementation algorithm and the reason to choose this particular algorithm was proposed by one of the literature surveys read by us. Gradient Boosting turned out to be the game changer and performed exceptionally well in all type of data sets either it was noisy or a dataset with a high number of entries.

In future, much higher results could be achieved by working on granular aspects of the data and more advance techniques.

REFERENCES

1. Gahirwal, M., Moghe, S., Kulkarni, T., Khakhar, D. and Bhatia, J., 2018. Fake News Detection. International Journal of Advance Research, Ideas and Innovations in Technology, 4(1), pp.817-819.

2. Wu, Liang, and Huan Liu. "Tracing Fake-News Footprints: Characterizing Social Media Messages by How They Propagate." (2018)

3. Working to stop misinformation and false news (https://www.facebook.com/facebookmedia/blog/working-to-stop-misinfomation-and-false-news).

4. Granik, Mykhailo, and Volodymyr Mesyura. "Fake news detection using naive Bayes classifier." Electrical and Computer Engineering (UKRCON), 2017 IEEE First Ukraine Conference on. IEEE, 2017.

5. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection William Yang Wang Department of

6. Computer Science University of California, Santa Barbara Santa Barbara, CA 93106 USA william@cs.ucsb.edu

7. Evaluating Machine Learning Algorithms for Fake News Detection" by Shlok Gilda in the 2017 IEEE 15th Student Conference on Research and Development (SCOReD), December 2017, pp.110-115.

8. International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075,Volume-8, Issue-10, August- 2019 An Effecient Fake News Detection System Using Machine Learning A.Lakshmanarao, Y.Swathi, T. Srinivasa Ravi Kiran

9. International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249-8958, Volume-8 Issue-5, June 2019 Hybrid Text Classification Method for Fake News Detection

10. Fake news detection in social media Kelly Stahl *B.S. Candidate, Department of Mathematics and Department of Computer Sciences, California State University Stanislaus, 1 University Circle, Turlock,CA 95382Received 20 April, 2018; accepted 15 May 2018 Thota,

11. Aswini; Tilak, Priyanka; Ahluwalia, Simrat; and Lohia, Nibrat (2018) "Fake News Detection: A Deep Learning Approach,"SMU Data Science Review: Vol. 1 : No. 3 , Article 10.