

Detailed Analysis of Fake News Detection using Machine Learning

Disha Shetty¹ Disha Gite² Affreen Parveen³

^{1,2,3}Department of Information Technology

^{1,2,3}Vasantdada Patil Prathishtan's College of Engineering, Mumbai, India.

Abstract—Today, social media has become an integral part of our life. We get news in just a click, sitting anywhere in the world. In this world of social media where people can express their views easily, spreading fake news is no big deal. It's hard to believe any random news on social media or the internet. There is a severe need to tackle this problem which has given birth to chaos, cyber-crimes, suicides, death of innocents and even terrorism. Hence, this project has been created in an effort to identify whether the news is real or fake. It uses three different machine learning algorithms, whose combined result is evaluated to determine the accuracy of the news. Its accuracy though not 100%, but is sufficient to rely on.

Keywords: Fake news, Machine Learning, Naive Bayes, Decision Tree, Random Forest

I. INTRODUCTION

Nowadays, "Fake news" has become a common term. In this technology driven world, lots of news surface on the internet. Finding genuine news or fact has become so difficult, just like finding a diamond in the ocean. Many articles, photos and videos are circulated for one news by different sources. Many people create fake news so that they get some money on each click which is called "clickbait". Many crimes such as cyberbullying, compel to suicide, riots, terrorism, etc. are associated with this.

The world of fake news is much bigger than the true news. To clear it from the platform is very difficult and endless. Such news is a piece of misinformation or fabrication of the original one. We cannot stop this but we can work on it to make people aware of it. For this, many organizations and tech companies are working their fingers to the bone. This project is just another effort to make a change in the world.

This project uses Machine learning algorithms to tackle the bizarre problem. Machine learning is the study which is used to perform tasks which are unconventional for the conventional algorithms. In this project, three different machine learning algorithms are used, they are Naive Bayes, Decision Tree and Random Forest. Machine learning has four broad categories of algorithms. All the above mentioned algorithms come under Supervised algorithms in machine learning. In this project, the data is processed through each algorithm obtaining three different results. These results are then combined and the final result with the accuracy is displayed.

II. PROPOSED SYSTEM

In Machine Learning, we first frame the problem and collect the data. This data is preprocessed and then some necessary features are added accordingly and the model is trained, so as to measure its performance and improve it as per requirement so that the system will be ready to be deployed.

Now, to find accuracy of the news using machine learning, we begin with data preprocessing steps where we have to handle missing data for data transformation. We have used the Binning method in order to smooth the data (noisy data). Since we are dealing with articles, our system uses only English articles. We then applied the NLP techniques on it i.e. removing punctuation and stopwords, tokenizing.

The dataset that we used contains text and metadata from different websites. After applying the data preprocessing steps on the dataset, we applied three different machine learning algorithms on the preprocessed dataset. The machine learning algorithms that we used are Naive Bayes algorithm, Decision tree algorithm and Random forest algorithm.

The Naive Bayes algorithm is used for predicting whether the news is true or fake. The Decision tree algorithm, which is a supervised learning algorithm, uses a tree structure, where the attribute is represented by an internal node and the class label is represented by the leaf of the tree. The Random forest is a classifier that contains numerous decision trees based on varied subsets of the dataset on which the algorithm is applied. Rather than relying on only one decision tree, the Random forest algorithm considers the prediction from each tree and based on the majority of votes of the predictions of the trees, it estimates the final output.

The voting classifier is used to combine the results of the algorithms. The voting classifier, which has two types- soft voting and hard voting, combines the output of each algorithm passed into it and estimates the output class or label based on the highest majority of voting. We have used soft voting in our model. Here, the output class or label is based on the average of the probability given to that class.

The implementation is done using Python language and Flask framework for the website. The dataset used in this implementation contains four columns namely id, title, text and label. The dataset initially consisted of 33788 entries which after preprocessing reduced to 27985 entries. The dataset was split into training dataset and test dataset in the ratio of 8:2.

III. METHODOLOGY

In the existing system, public opinion is taken into consideration to tag news as True or Fake. However, presently websites can be a platform to try to detect information manually. So, in order to implement it we have used the Flask framework to build the web application. Here, we have provided input fields for users, where the user has to enter the Title and Text (full article) of the news. After which, our website will display the estimated accuracy and label of the news being "True" or "Fake". Fig. 1 is the conceptual methodology diagram that gives a clear idea of our system. As we have mentioned in our diagram, "News

Data” can be any kind of article, from any source, you want to check.

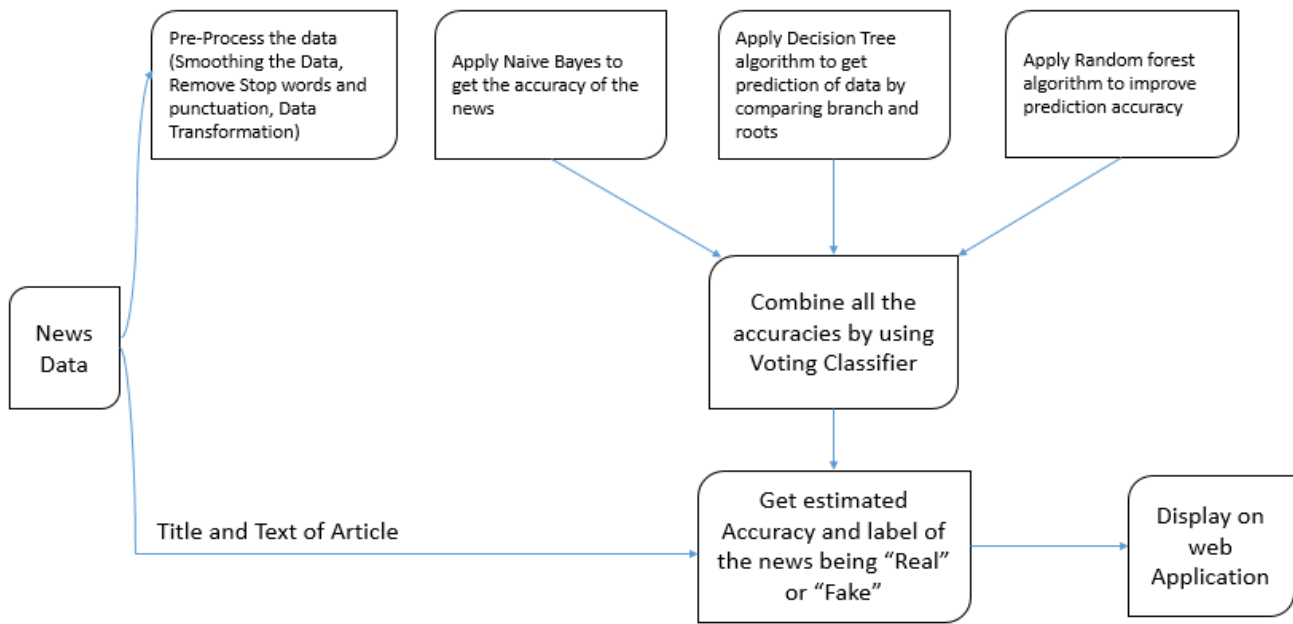


Fig. 1: Conceptual methodology diagram

IV. RESULTS

On implementing the tree machine learning algorithms, we compared the accuracies of the algorithms and found that Random Forest gives the highest accuracy.

For calculating accuracy, we used the formula:

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN)$$

Where TP is True Positive, TN is True Negative, FP is False Positive and FN is False Negative.

Machine learning algorithm	TP	FP	TN	FN	Accuracy
Naive Bayes	1753	682	3037	125	85.58%
Decision Tree	2072	363	2761	401	86.34%
Random Forest	1896	539	3006	156	87.58%

Table. 1: Result of various algorithms

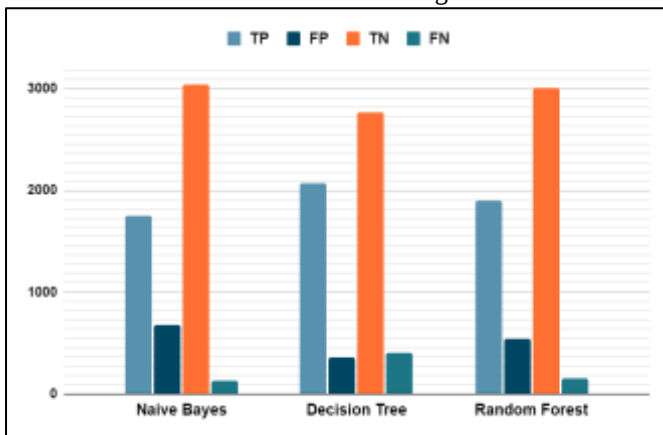


Fig. 2: Analysis of algorithms

V. CONCLUSION

By using Naive Bayes, Decision tree and Random Forest Algorithms we found that Random Forest gives the highest

accuracy which is reliable. In Machine Learning projects, using only one algorithm is not sufficient, minimum three algorithms should be used. We cannot rely on the result of just a single algorithm. Also, the result of each algorithm varies. So, if different sets of algorithms are used the result may differ slightly. However, machine learning has the capability to give solutions to human driven acts.

REFERENCES

- [1] Fake News Detection by Akshay Jain, Amey Kasbe, Maulana Azad National Institute of Technology, Bhopal, India, IEEE 2018 International Students' Conference on Electrical, Electronics, and Computer Sciences.
- [2] Fake News Detection Using Machine Learning approaches: A systematic Review, Syed Ishfaq Manzoor, Dr Jimmy Singla, Nikita, Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI 2019).
- [3] Media-Rich Fake News Detection: A Survey, Shivam B. Parikh and Pradeep K. Atrey, 2018 IEEE Conference on Multimedia Information Processing and Retrieval.
- [4] Supervised Learning for Fake News Detection, Julio C. S. Reis, Andre Correia, Fabricio Murai, Adriano Veloso, and Fabricio Benevenuto, Universidade Federal de Minas Gerais, 2019 IEEE.
- [5] Detecting Fake News Articles, Jun Lin, Glenna Tremblay-Taylor, Guanyi Mou, Di You, Kyumin Lee Computer Science and Engineering Louisiana State University Baton Rouge, LA, USA.
- [6] A smart system for fake news detection using machine learning, Anjali Jain, Avinash Shakya, Harsh Katter, Amit Kumar Gupta, 2019 IEEE 2nd International conference on ICICT.
- [7] An Efficient Fake News Detection System Using Machine Learning, A. Lakshmanarao, Y. Swathi, T. Srinivasa Ravi Kiran, International Journal of Innovative

Technology and Exploring Engineering (IJITEE) ISSN:
2278-3075, Volume-8, Issue-10, August- 2019.

- [8] Fake News Detection Using Naive Bayes Classifier,
Mykhailo Granik, Volodymyr Mesyura, 2017 IEEE First
Ukraine Conference on Electrical and Computer
Engineering.

