FAKE NEWS DETECTION SYSTEM USING MACHINE LEARNING

Shipra Srivastava *Ajay Saini* *Ayush Arun* *Satyendra Kumar Yadav* *Nancy Aggarwal*
Asst. Prof. *Student* *Student* *Student* *Asst. Prof.*
DEPTT. OF CSE-IOT, GREATER NOIDA INSTITUTE OF TECHNOLOGY

Abstract- *In this period of digital consumption of informations and news, Fake news has become a serious and major issue before the society, it may affect the social integrity, public opinion, and public trust over media, government, judiciary. This research paper provide an investigation over fake news using machine learning techniques. The main goal is to make a well regulated, well planned, dependable and authentic system so that spreading of fake news can be reduce. The plan investigates the ethical predicaments encompassing the revelation of phony news, addressing matters of equity, clarity, and motivations. The improvement approach consolidates popular strategies and moral rules to guarantee the dependable sending of the disclosure framework.*

Keywords: *Machine Learning, News Prediction, Recommendation, Logistic Regression, Gradient Boosting, Decision Tree, Random Forest.*

1. INTRODUCTION

In this era of digitalization, news spread very rapidly via different platforms and all this has become normal part of our life. This easy and within second access to the information and news leads towards a problem, that is circulation of fake news. Fake news is intentionally prepared news that is circulated for the purpose of breaking the integrity and unity among the citizens or to give threat to the trust in media. As the technology is becoming more and more advance, it should be clear that our information system will detect the fake information. This paper gives a comphrensive review of fake news detection, it target to findout deceptive news by some examining methods of Artificial Intelligence and Machine Learning technologies. By a carefull research and study on the writing pattern, creating and broadcasting or circulation of the fake news. Through thorough assessment, we intend to recognize the constraints of ebb and flow research strategies, assess the ethical ramifications of fighting phony news, and produce novel plans to fortify our safeguards against this unavoidable threat[4]. To battle counterfeit news successfully, it is fundamental to comprehend the techniques utilized by the individuals who spread deception. This paper will dive into the mental cycles that render people vulnerable to counterfeit news and look at the mental and close to home factors that add to its dispersal. Moreover, through discussion investigation, we will reveal explicit attributes of misleading substance, including language designs, utilization of thoughts, and control of feelings. By acquiring understanding into the intricacy of phony news, we can foster more powerful identification strategies[7]. This paper analyzes the possible traps, predispositions, and unseen side-effects related with robotized search calculations. To address these worries, we will introduce a moral structure underlining straightforwardness, responsibility, and the job of innovation chasing truth. By creating algorithmic models to help data education and encouraging joint effort among partners, we expect to extend and actually battle disinformation in the consistently advancing computerized danger scene.

2. LITERATURE REVIEW

2.1 Content-Based Approaches:

- Research and analysis of writing style and sentiment of news to find out similar pattern of fake news. This will help model to distinguish between the news.

2.2 Social Context-Based Approaches:

- To evaluate news authencity leverage the social media context and find the social cues like interaction between users,source credibility and pattern sharing

2.3 Machine Learning Algorithms:

- Logistic Regression, Decesion tree, Gradient boosting and random forest algorithms are used to design this machine learning model.

2.4 Fact-Checking Organizations:

- Organization who are independent have to keep verification, debug false information and find out news quality

2.5 Hybrid Approaches:

- Combine approach is mixing content and social based to achieve high accuracy

3. PROBLEM STATEMENT

3.1 Content Based Approch: To distinguish between Fake and True news by viewing manually is very difficult task and have more chances of error.Very less , labelled data for training has been used ,so it will hinder to find out fake news.

3.2 Early Detection: To distinguish between Fake and True news by viewing manually is very difficult task and have more chances of error. Very less, labelled data for training has been used, so it will hinder to find out fake news.

3.3 Model Generalization: Model should be work in all languages and various domain. The people may manipulate feature of news to deceive our model.

3.4 Resource-Intensive: We need to check fact of news manually using expert and skilled person. In a big news , we can not check each and every paragraph manually.

3.5 Integration Complexity: Content gathered from different sources can be very difficult for our model to distinguish between news, so we need to keep balance between precision and recall.

4. METHODOLOGY

4.1 **Data collection:** Take the various type of data from fake news and true news.

4.2 **Prerequisite:** Eliminate special symbol like question marks, punctuation marks, stop words, link, underscore etc. before clearing and specifying the text. Then convert word to numerical representation.

4.3 **Feature engineering:** Now find out features and different matrices like sentiment & readability. Then find the frequency of each words.

4.4 **Data classification:** To Take 75% of data for training the model and 25% of data for testing.

4.5 **Model selection:** Train the model using all the algorithms.

4.6 **Hyperparameter tuning**: Hyperparameter of each model done using cross-validation & grid search method.

4.7 **Model evaluation:** Evaluate accuracy, precision, recall and F1 score of each model.

4.8 **Pooling Methods:** Then combine all the model to enhance the performance of model.

4.9 **Importance analysis:** Compare and analyze each model & find out which model is forecasting best.

4.10 **Interpretive Model:** Review each prediction for more interpretive model, so that we can increase trust and ensure fairness of the decision.

5. ALGORITHM

We use the following learning methods with our plan to test the effectiveness of fake news detection.

5.1 **Logistic Regression:** The Logistic Regression is one of the type of machine learning algorithm, which is mainly used for classification of binary tasks. Basically it determines the possibility of true or fake substances, which belonging to their category or not. It gives their output with the value of 0 and 1.It ensures their value lies between 0 and 1.If the output exceeds in these range, then output 1 is obtained otherwise 0 is produced. This algorithms comes in different types, which is binomial, multinomial etc.

5.2 **Decision Tree:** This is one of the type of machine learning algorithm. It follows the tree like structures. Their prediction makes valuable. So that the decision tree algorithm one of

the accurate algorithm for predict the value.

5.3 **Gradient Boosting:** It is one of the important machine learning algorithm used for describe the information.It predict the output according to previously given information. It combines the one model to another, The prediction is balanced in these algorithms balances prediction shrinkage, making it a go-to choice for predictive modeling.

5.4 **Random Forest:** Random Forest combines the collective wisdom of multiple decision trees to enhance prediction performance. During training, it creates a multitude of decision trees, each constructed using a random subset of the dataset and measuring a random subset of features in each partition. This randomness reduces overfitting risk and improves overall prediction accuracy. In prediction, Random Forest aggregates the results of all trees, providing stable and precise outcomes. Widely used for classification and regression, it handles complex data, reduces overfitting, and offers reliable forecasts in diverse environments".

6. DATASETS

The information we used in this decision is open source and publicly available online. This document contains fake and honest news from various sources. Honest news sources contain factual information about real-world events, while fake news sites contain claims inconsistent with reality. The consistency of the legal question in which many of these articles make their claims can be checked by fact checking such as politifact.com and snopes.com. We use three different data sets in this decision, briefly described below. The first data is called the “ISOT Fake News Dataset” [23] (will be called DS1 in the future), which contains fake and fake news retrieved from the World Wide Web. The real article was probably taken from reuters.com, a well-known news site, while the fake article was taken from another site, usually politifact.com, a published site. The database contains a total of 44,898 sentences, of which 21,417 are true and 23,481 are false. The entire book contains words taken from many sources, but the most prominent one is about political news. The database is available on Kaggle [24] (hereafter referred to as DS2) and contains a total of 20,386 articles for preparation and 5,126 items for testing. A third document is also available on Kaggle ; There are a total of 3,352 items, both fake and real.

7. RESULT AND DISCUSSION

Logistic Regression:

Precision: Logistic Regression can have decent precision, especially when the classes are wellseparated.

Recall: Recall for Logistic Regression can be good, but it depends on the threshold set for classification and the nature of the problem.

Accuracy: Logistic Regression generally provides good accuracy, particularly when the features are

linearly separable.

	Precision	Recall	F1-Score	Support
0	0.99	0.99	0.99	5846
1	0.99	0.99	0.99	5374
Accuracy			0.99	11220
Macro Avg.	0.99	0.99	0.99	11220
Weighted Avg.	0.99	0.99	0.99	11220

Table.1 Classification Report of Logistic Regression

Decision Tree:

Precision: Decision Trees may have lower precision compared to ensemble methods like Random Forest or Gradient Boosting because they are prone to overfitting.

Recall: Decision Trees can have good recall depending on the depth and complexity of the tree. Overfitting can lead to high recall on the training data but may generalize poorly.

Accuracy: Decision Trees can achieve decent accuracy, but they are more susceptible to overfitting, especially on noisy datasets.

	Precision	Recall	F1-Score	Support
0	0.99	0.99	0.99	5846
1	0.99	0.99	0.99	5374
Accuracy			0.99	11220
Macro Avg.	0.99	0.99	0.99	11220
Weighted Avg.	0.99	0.99	0.99	11220

Table.2. Classification Report of Decision Tree

Gradient Boosting:

Precision: Gradient Boosting typically achieves high precision as it sequentially builds trees to correct errors made by previous trees.

Recall: Gradient Boosting tends to have good recall because of its ability to focus on difficult-to-classify instances.

Accuracy: Gradient Boosting often achieves high accuracy due to its iterative nature of improving upon the weaknesses of previous models.

	Precision	Recall	F1-Score	Support
0	1.00	1.00	1.00	5846
1	0.99	1.00	1.00	5374
Accuracy			1.00	11220
Macro Avg.	1.00	1.00	1.00	11220
Weighted Avg.	1.00	1.00	1.00	11220

Table.3. Classification Report of Gradient boosting

Random Forest:

Precision: The Random Forest algorithm gives a better characteristics, because of its nature, which is helpful to reduce various things.

Recall: The Random Forest algorithms have best recall, it record complex boundary by classifying multiple decision tree algorithms.

Accuracy: The Random Forest algorithm mostly have high accuracy in comparision to others algorithm, because it merge multiple algorithms in single manner. The performance of this algorithm is also well mannered in comparison to others.

	Precision	Recall	F1-Score	Support
0	0.99	0.99	0.99	5846
1	0.99	0.99	0.99	5374
Accuracy			0.99	11220
Macro Avg.	0.99	0.99	0.99	11220
Weighted Avg.	0.99	0.99	0.99	11220

Table.4. Classification Report of Random Forest

The precision and detail are enhanced when numerous decision trees are combined. The model can discriminate between real and fake news because it can locate the ideal hyperplane.

The ability to interpret decision trees Interpretability is offered by decision trees, however marginally lessso than by other models. The decision tree approach sheds light on the characteristics that influence classification.

8. TEST CASES

We have taken four news articles from various sourses and checked the accuracy of the implemented algorithms in this ML models.

	Logistic Regression	Decision Tree	Gradient Boosting	Random Forest	Result
NEWS 1	Fake News	Not A Fake News	Not A Fake News	Not A Fake News	True News
NEWS 2	Fake News	Not A Fake News	Fake News	Fake News	Fake News
NEWS 3	Not A Fake News	Fake News	Fake News	Fake News	Fake News
NEWS 3	Not A Fake News	Not A Fake News	Not A Fake News	Not A Fake News	True News

Figure.1. Output Of Some News

9. CONCLUSION

With the use of machine learning algorithms such as random forest, decision tree, gradient boosting etc. we find a truthfulness the information, that is fake or true. In today's era various misinformation is reaches from one palace to another palace, due to this the privacy of users leaks on internet. These machine learning techniques help to the find the correctness of information. These all machine learning algorithm combines together and gives the highest accuracy about an information, which is true or false.

10. REFERENCES

[1] A. Douglas, "News consumption and the new electronic media," *The International Journal of Press/Politics*, vol. 11, no. 1, pp. 29–52, 2006.

[2] J. Wong, "Almost all the traffic to fake news sites is from facebook, new data show," 2016.

[3] D. M. J. Lazer, M. A. Baum, Y. Benkler et al., "The science of fake news," Science, vol. 359, no. 6380, pp. 1094–1096, 2018.

[4] S. A. Garc´ıa, G. G. Garc´ıa, M. S. Prieto, A. J. M. Guerrero, and C. R. Jime´nez, "The impact of term fake news on the scientific community scientific performance and mapping in web of science," Social Sciences, vol. 9, no. 5, 2020.

[5] A. D. Holan, 2016 Lie of the Year: Fake News, Politifact, Washington, DC, USA, 2016.

[6] S. Kogan, T. J. Moskowitz, and M. Niessner, "Fake News: Evidence from Financial Markets," 2019, https://ssrn.com/ abstract=3237763.

[7] A. Robb, "Anatomy of a fake news scandal," Rolling Stone, vol. 1301, pp. 28–33, 2017.

[8] J. Soll, "The long and brutal history of fake news," Politico Magazine, vol. 18, no. 12, 2016.

[9] J. Hua and R. Shaw, "Corona virus (covid-19) "infodemic" and emerging issues through a data lens: the case of China," International Journal of Environmental Research and Public Health, vol. 17, no. 7, p. 2309, 2020.

[10] H. Jwa, D. Oh, K. Park, J. M. Kang, and H. Lim, "exBAKE: automatic fake news detection model based on bidirectional encoder representations from transformers (bert)," Applied Sciences, vol. 9, no. 19, 2019.