International Journal of Engineering and Management Research e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume-13, Issue-2 (April 2023)
https://ijemr.vandanapublications.com https://doi.org/10.31033/ijemr.13.2.29

Fake News Detection Using Machine Learning: An Exhaustive Review

Arisha Farha[1] and Afsaruddin[2]
[1]PG Student, Department of Computer Science & Engineering, Integral University, Lucknow, U.P., INDIA
[2]Assistant Professor, Department of Computer Science & Engineering, Integral University, Lucknow, U.P., INDIA

[1]Corresponding Author: arishaaz@student.iul.ac.in

Received: 11-03-2023 **Revised:** 27-03-2023 **Accepted:** 27-04-2023

ABSTRACT

Fake news can have serious consequences, from influencing elections to spreading harmful misinformation. Machine learning can be used to help combat the spread of fake news by analyzing large amounts of data and identifying patterns that may indicate the presence of false or misleading information. Here are the steps that can be taken to perform fake news analysis using machine learning. Data Collection, Data Preprocessing, Feature Extraction, Model Training, Model Evaluation, and Model Deployment. That is the reason today we need a PC fake wise based model that can identify any phony news before it is posted. All web-based media stages have worked towards this path, however, in some places it appears to be that their model is deficient to catch such phony news. Since some web-based media organizations have attempted to choose whether the news is phony or not based on some predefined datasets. Furthermore, a few organizations have looked through just the watchwords of the news that the news is phony. This demonstrates that we need a model that depends on the old dataset, and the current news dataset and watchwords. Alongside this, focus on the circumstance, spot, and kind of information, while these things are not dealt with in the current models. So I might want to remember this load of boundaries for my model to assist with distinguishing counterfeit news. On the off chance that we perceive Fake News as the ideal opportunity, we can make the perfect strides at the perfect time. PC based models are not generally exact, so the model ought to likewise have the office to contrast and genuine news. Assuming news is contrasted and current information, 76% of phony news can be distinguished simultaneously. Accordingly, the model ought to likewise have the office of the relative survey

***Keywords*—** Decision Tree Algorithm, Real News, Fake News, Genuine

I. INTRODUCTION

In today's time, 70% of the world has expressed its presence in the virtual world, which means that more than half the world is connected to this virtual world i.e. Internet world in some way or the other. In earlier times, people did not have any open means where they could openly put their ideas in front of the world. Where he can talk about himself, about his society, or his religion and customs. Social media is such a platform in today's time where people can share their problems and get tips to get out of them. By using social media today, people can also raise their voices against the injustice done to them and get public support. In today's time, the governments of countries have started using social media, they are taking their agenda very easily to the people. Political parties are using social media to express their views. Through social media, the work done by him and his party to reach the public so that they can take advantage of this in the elections. Many times people also use social media to bring out someone's talent as we have seen. This changes that person's life overnight. But as we know where there is light there is room for darkness too. And sometimes freedom also brings with it arrogance and people also take wrong advantage of this freedom. What I mean to say is that people use social media to tell about themselves, society, religion, or customs. But sometimes some wrong people start taking advantage of this to spread their wrong feelings which is wrong. Anomaly detection in modern networks is complex because there are so many different kinds of networks, each with unique properties. They know that its effect can leave a bad effect on someone's life, but still, they do such things. When any news is put on social media to hurt the sentiments of any person, society, or religion by using wrong facts, then it is called Fake News. The machine learning method known as "ensemble learning" employs the training of several learners to produce an optimal solution [1]. Whereas those people should understand that it can ruin a life, riots can flare up, which can cause loss of life and property in great quantity. That is why in today's time it becomes the duty of social media companies to take up this responsibility and prevent them from being wrong of their platform. For this, they need to change their platform so that no wrong person can do this. And be recognized even before its fake news is posted. To do this, social media companies will have to design a tool that can do this. First of all, to do this, we have to prepare a dataset in which there is a collection of old and new datasets and keep not only true but also fake news in it so that we can learn the machine to differentiate between both types of news. It has recently been found

e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume-13, Issue-2 (April 2023)
https://doi.org/10.31033/ijemr.13.2.29

that most of the classic methods of anomaly detection use unsupervised approaches to discover anomalies in cases where there is just a limited number of labeled anomalous data and plenty of unlabeled data [2]. After this we should filter the dataset, for which we can do it using an algorithm like NLP, we can filter the dataset by doing it, then we will get a dataset that will be completely correct, while preparing the dataset, we need to keep this point It should be noted that the ratio of true and false news is taken correctly, in general this ratio should be 60:40, which means 60% true news and 40% wrong news. After this, we should use the classification algorithm so that we can extract the features of the news such as the title of the news, timing of the news, source of the news, location of the news, and which class the news belongs to and through the classification we can correct the dataset. We will be able to classify from this and design a correct model through which we will be able to know what are the features of correct news and false news and on the basis of these features we will prepare trained and test dataset which will be used by the machine and correct will be able to decide. Here we need to take Trained and Test dataset in the right ratio, in general, trained dataset is always more than Test dataset, some people keep the ratio of 8: 2 and some people keep the ratio of 7: 3 but according to me, for Fake News Detection For this 8:2 ratio is the best. Once the Trained dataset is ready we can prep the model and then we choose the prediction algorithms, although there are many prediction algorithms in machine learning but in this case the MB algorithm It is the best, the decision tree is also fine for this work. In this way we can identify the fake news and can steam someone's wrong intention in time and the platform will recognize the fake news even before it is posted and necessary steps can be taken. In high dimensional, complex datasets, some of them rely on shallow practices that can't keep up with the numerous interactions between structures and attribute [2].

1.1 **Data Collection**: The step one in any machine learning(ML) project is to collect relevant data. In the case of fake news analysis, this means collecting a large dataset of news articles from various sources. There are several publicly accessible datasets that is ready for use, such as LIAR dataset, which contains labeled examples of true and false statements made by politicians. The distribution of crowd density in photographs of packed crowds is seldom uniform because of the differences in viewpoint and scene that exist in these photos[3] . Due of this, it is nonsensical to try to count the individuals in the crowd while simultaneously taking in the whole sight. The divide-count-sum mechanism was implemented into our system after it was updated as a direct result of this issue[3].

1.2 **Data Preprocessing**: Once the data has been collected, it needs to be preprocessed to prepare it for machine learning. This may involve tasks such as cleaning the data, removing stop words, and converting the text into numerical vectors[4].

1.3 **Feature Extraction**: Next, features need to be extracted from the data. This involves identifying patterns in the text that may indicate the presence of fake news. Some examples of features that can be extracted include the use of emotive language, the presence of logical fallacies, and the use of exaggerated or sensationalist headlines[5].

1.4 **Model Training**: After the features have been extracted, a ML model can be trained on the data. There are many different types of models that can be used for fake news analysis, including decision trees, support vector machines, and neural networks. The model is trained on a subset of the data and validated on another subset to ensure that it is accurately identifying fake news[6].

1.5 **Model Evaluation**: Once the model has been trained, it needs to be evaluated to determine how well it is performing. This can be done using metrics such as precision, recall, and F1 score. These metrics measure the model's ability to correctly identify fake news and avoid false positives.

1.6 **Model Deployment**: Finally, the trained model can be deployed in a production environment to analyze new news articles and identify any that may be fake. This can be done using an API that accepts text input and returns a binary value indicating whether the article is likely to be fake or not.

II. RELATED WORK

On the basis of extensive literature survey related to Fake News Analysis Using Machine Learning has been taken into consideration in this section.

Julio C. S. Reis, et. al., (2019) In this paper, various points have been worked to detect Fake News. For example, the title of the news and the source of the news are given as parameters. And the SVM classifier is used for classification and the model is built on the basis of title and sauce. And Lustum algorithm is used for prediction. And the author has given the accuracy of his algorithm as 87%.But if we analyze this structure carefully, then we finds that the correct comparison has not been done at many places. Which seems as an obstacle in coming to the right conclusion. Because fake news cannot be determined solely on the basis of the title and source of the news[7].

Adrian M.P. et. al. (2019) Detecting fake news is a big challenge in itself. The author has faced this

challenge to a large extent and has designed his model to catch the fake news in its initial steps. For this, the author has done it in a better way using the NLP algorithm. NLP is such a computerized natural language processing algorithm, through which we can easily analyze language differences. And with the help of this author has prepared a dataset of current news and through that a Trained and Test dataset has been prepared .Because Trained and test datasets make a huge contribution to machine learning. In this algorithm, the writer has given the accuracy of 79%.Which is insufficient according to the need to detect fake news in social media. The biggest reason for the decrease in accuracy is that only current news has been said in this algorithm. Whereas fake news can also be spread about old facts. That's why the author should have noticed this as well[8].

William Yang Wang (2018) The author has based this research on the basic word and has worked to identify fake news on the basis of the same. Keyword based searching means that no such news is being posted in the news or post. Since the Chinese government wants to keep an eye on every activity of its citizens, a big example of this is seen in this research. In this algorithm, the author first breaks the news into keywords and then stores it in a lexicon array. And at the same time, a dataset of old and current news has been prepared, then it is divided in the ratio of 7: 3 in the test and trained dataset. And the model has been prepared on the basis of cover, time, title and source. Then an attempt is made to predict the fake news using the decision tree. If we talk about the accuracy, then the author is telling the accuracy of this algorithm to be 92%.That sounds right. But if the algorithm is understood properly, then it turns out that it can be very difficult because of keyword wise searching. This can increase the time complexity of the algorithm and if such a tool is used as a social media tool, then there will be a big problem. And anyway it is okay to stop people from posting wrong news but it is wrong to restrict some words[9].

Costin BUSIOC et. al., (2020) Fake News a Social Media Platform Has Been Explained As A Curse By The Author. Which sounds quite right? In this paper, the author has used linear regression algorithm to detect fake news. And to propel the model, a dataset of fake and true news has been created. Because in order to train the machine, he must have experience of both types of news. If the machine has news of both the ways, then it will be able to take the right decision. Here the author has told that he has used 65% true news and 35% false news and has tried his based pay machine. Then using this as a basis, it is divided into Trinad and Test data sets, whose ratio is 8: 2. And in this algorithm, the runn algorithm has been used for its prediction. And the author has given 91% accuracy. But after reading the whole paper it seems that the author should have described his algorithm a little more[10].

Alim Al Ayub Ahmed (2020) In this paper, the author has described the Fake News Detection very accurately and used the correct parameters to detect Fake News. In this paper, the author has set some parameters to identify fake news, in which the title, time, source, place of the news are. These parameters are very important to identify fake news. Then the author has prepared a huge dataset of 10000 news in which 8000 are true news and 2000 are fake news. In this paper, the author has also made a basis for the categories of news such as religious news, political news, social news, criminal news, and news related to rituals. Then through this dataset, Trined and Test dataset has been prepared, which has been done by the author in LSTM algorithm. In this algorithm, the author has given the accuracy of 89% which is also correct. But the author has not mentioned the current news anywhere, whereas the news is largely inspired by the current news itself[12].

III. METHODOLOGY

Proposed System

In this paper a model is made dependent on decision tree algorithm. Word counts family members to how frequently they are utilized in other artices in your dataset can help. Since this issue is a kind of text characterization, Implementing a the decision tree algorithm will be best as this is standard for textbased handling. The real objective is in fostering a model which was the content change and picking which kind of text to utilize (features versus full content). Presently the following stage is to separate the most ideal highlights for the decision tree algorithm, this is finished by utilizing a n-number of the most utilized words, as well as expressions, lower packaging or not, essentially eliminating the stop words which are normal words, for example, "the", "when", and "there" and just utilizing those words that show up in any event a given number of times in a given content dataset.

Decision Tree Algorithm

Decision Tree algorithm has a place with the group of managed learning calculations. In contrast to other administered learning calculations, the decision tree calculation can be utilized for tackling relapse and order issues as well. The objective of utilizing a Decision Tree is to make a preparation model that can use to anticipate the class or worth of the objective variable by taking in basic decision principles surmised from earlier data(training information). In Decision Trees, for anticipating a class name for a record we start from the foundation of the tree. We think about the upsides of the root trait with the records characteristic. Based on examination, we follow

the branch relating to that worth and leap to the following hub.

Decision tree algorithm steps are:

1. Read the query news in q.
2. Split the query in words w[ ] array.
3. Scraping the data using w[ ] from news sites and store in dataset[ ].
4. Read the tweets using w[ ] from tweeter and store it in tweets[ ].
5. Clean the data and create a single data set
   td[ ]=dataset[ ]+tweets[ ]
6. Extract the feature of each row
   For kx in td[ ]
   If kx.date= q.date
   If kx.text in q.text
   Collect in p[ ]=kx.text
7. Trained the dataset p[ ] and create the model m[x][y]
8. Test the query on the basis of decision tree and get classifier score.
9. if score=0 then
   Print news is fake
   Else if score>0 and score<=10
   Print news is semi true
   Else
   Print news is true

IV. RESULTS

In this part, we are using decision tree algorithm to detect fake news. This is the best algorithm to detect fake news, and out execution examination of our customary AI and neural organization based profound learning models. We present the best execution for each dataset and every lattice in strong. We compute exactness, accuracy, review, and f1- score for fake and genuine class, and track down their normal, weighted by help (the quantity of genuine cases for each class) and report a normal score of these measurements. It is observed that among the customary AI models, the decision tree algorithm, with n-gram highlights, has played out the best. Indeed, it has accomplished practically the decision tree algorithm accuracy is 97 precision on our joined corpus. We likewise find that expansion of conclusion includes alongside lexical highlights doesn't improve the exhibition fundamentally. For lexical and supposition highlights, Passive Aggressive Classifier and LR models have performed better compared to other customary AI models as proposed by the greater part of the earlier investigations. Then again, however includes produced utilizing Empath have been utilized for understanding duplicity in a survey framework, they have not shown promising execution for counterfeit news identification.

Table 1: Showing the classifier accuracy

Subjects	Politics	Sports	Social Issues
Algorithm	Logistic Regression	Naive Bayes	Decision Tree + My App
Accuracy	56	80	96
	75	78	92
	89	87	97

V. CONCLUSION & FUTURE SCOPE

In this research, we were successful in developing such a fake news detection tool for social media platforms, so that any fake news can be detected in time, for this we have used a dataset of both new and old news, so that the accuracy of the results is maintained. Remain and any prediction has a big role of Trined and test dataset that is why we have also taken care that any fake news has its own identity such as its timing, its source, its title, its location and it is new. Tell which category does it belong to, that's why the accuracy of our algorithm is 97%, which is a big thing in itself and somewhere its accuracy is also 100%.

But now more work is needed in this because in this we have analyzed only text data whereas fake news can also be spread through images and wrong videos and I would like to work on this in my next research through which I can make my tool stronger and more reliable and detect fake news before it is posted.

REFERENCES

[1] Khan, W. & Haroon, M. (2022). An efficient framework for anomaly detection in attributed social networks. *International Journal of Information Technology*, *14*(6), 3069-3076.

[2] Khan, W. (2021). An exhaustive review on state-of-the-art techniques for anomaly detection on attributed networks. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *12*(10), 6707-6722.

[3] Khan, W., Haroon, M., Khan, A. N., Hasan, M. K., Khan, A., Mokhtar, U. A. & Islam, S. (2022). DVAEGMM: Dual variational autoencoder with gaussian mixture model for anomaly detection on attributed networks. *IEEE Access*, *10*, 91160-91176.

[4] Nida Khan, M. H. (2022). Comparative study of various crowd detection and classification methods for safety control system. *International Journal of Engineering and Management Research*, 124-130.

[5] Ethar Qawasmeh, Mais Tawalbeh & Malak Abdullah. (2019). Automatic identification of fake news using deep learning. *Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 978-1-7281-2946-4/19/$31.00 ©2019 IEEE.

[6] William Yang Wang. (2017). Liar, liar pants on fire": a new benchmark dataset for fake news detection. *arXiv:1705.00648v1 [cs.CL]*.

[7] Costin BUSIOC, Stefan RUSETI & Mihai DASCALU. (2020). A literature review of nlp approaches to fake news detection and their applicability to Romanian- language news analysis. *Romanian Ministry of Education and Research, CNCS - UEFISCDI, project number PN-III-P1-1.1-TE-2019-1794, within PNCDI III.*

[8] Alim Al Ayub Ahmed, Ayman Aljarbouh & Praveen Kumar Donepudi. (2020). Detecting fake news using machine learning: A systematic literature review. *IEEE Conference*.

[9] Razan Masood & Ahmet Aker. (2018). The fake news challenge: Stance detection using traditional machine learning approaches. In: *Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (KMIS 2018)*, pp. 128-135.

[10] Sohan De Sarkar & Fan Yang. (2018). Attending sentences to detect satirical fake news. *Proceedings of the 27th International Conference on Computational Linguistics, pages 3371–3380 Santa Fe, New Mexico, USA.*

[11] Abdullah-All-Tanvir, Ehesas Mia Mahir & Saima Akhter. (2019). Detecting fake news using machine learning and deep learning algorithms. *7th International Conference on Smart Computing & Communications (ICSCC).*

[12] Hadeer Ahmed, Issa Traore & Sherif Saad. (2017). Detection of online fake news using n-gram analysis and machine learning techniques. In: *International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*, pp. 127–138. Springer.

[13] Hunt Allcott & Matthew Gentzkow. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives, 31*(2), 211–36.