Fake News Detection Using Logistic Regression Method

Aayushya Kumar*, Satyam Kumar Mishra, Khushar Shukla, Smriti Srivastava
AI&ML, Dr. Akhilesh Das Gupta Institute of Professional Studies, FC-26, Shastri Park, 110053, Delhi, India

* Corresponding author

doi: https://doi.org/10.21467/proceedings.178.16

ABSTRACT

Fake News is an immense issue that is increasing worldwide. Various rumors and false information are spreading all over the internet that increase false beliefs, biases and even violence among people. As a result, it is quite challenging to differentiate between true and fake in a world where people are entangled with various false and rumored news all around their mobile phones, news articles and social media platforms. To overcome this problem, various machine learning techniques have been used so far to find and detect the authenticity of news and information. In this study, the logistic regression technique has been used under supervised learning along with text preprocessing tools and feature extraction for classifying the news as either fake or real. It evaluates the output in 0s and 1s to find the authenticity of news either fake or real respectively. To ensure the performance, the high accuracy of the trained model has also been evaluated on accuracy metrics. This study helps to overcome the spread of misinformation over the internet.

Keywords: Fake News, Logistic Regression, Vectorization, Accuracy Metric

1 Introduction

The rapid expansion of internet information has made knowledge more accessible to a wider audience, but it has also unleashed a deluge of false information, or fake news. Disinformation weakens societal cohesion, undermines institutional credibility, and impedes decisions. To contribute to the field of creation of reliable machine learning models that can find and lessen the detrimental consequences of fake news. Through examining traits and dissemination patterns of false information, it is needed to create intelligent algorithms that protect the integrity of the information environment and enable educated public discourse. However, effectively harnessing ML for this task requires a nuanced understanding of the characteristics and propagation patterns that differentiate fake news from factual content. This study delves into the potential of machine learning for fake news detection. This paper aims to find the most effective strategies for building intelligent systems that can filter out fake news and ensure a more reliable information ecosystem. The final goal is to use machine learning to empower people by authenticating the news that they are consuming.

2 Literature Review

The misinformation and rumors spreading across social media lead to immense concern in society. Studies on various aspects of fake news have been done to tackle this issue such as finding the authenticity of the information in the form of images using a Linear Support Vector Machine (LSVM) with TF IDF vectorization method that is highly efficient for analyzing fake news over social media. Numerous studies also discussed training models on Recurrent Neural Networks with LSTM to find fake and misleading news over the internet which sometimes leads to severe consequences [2]-[3]. For instance, techniques including Passive Regressive Classifier, Naive Bayes Classifier and Decision Tree are also used to detect fake news by verifying its accuracy and performance on various datasets [4]-[5]. Furthermore, fake news spreaders try to deceive user by triggering their sentiments related to a particular thing. As a result, sentiment analysis becomes more crucial to tackle these aspects of fake news [6]. Another misleading part of social media is the clickbait practices to make attractive headlines

© 2025 Copyright held by the author(s). Published by AIJR Publisher in "Proceedings of 2nd International Conference on Emerging Applications of Artificial Intelligence, Machine Learning and Cybersecurity" (ICAMC 2024). Organized by HMR Institute of Technology and Management, New Delhi, India on 16-17 May 2024.

Proceedings DOI: 10.21467/proceedings.178; Series: AIJR Proceedings; ISSN: 2582-3922; ISBN: 978-81-984081-8-1

and image previews to gain more users on social media. To tackle the clickbait detection issue, the Convolutional Neural Network (CNN) technique is used to find the clickbait video over social media [7]. However, there is still a lack in finding one robust approach to discuss all the issues of fake news collectively because of the intensity and immense size of the problem. Also, finding the authenticity of the news in real time is a rigorous puzzle to solve. Therefore, a collective approach that includes aspects like rumors, sarcasm, punctuations, misleading news and false propaganda is necessary to keep in mind for developing a highly optimized model to end the spreading of fake news all over the internet.

3 Methodology

To tackle the issue of fake news worldwide many studies and techniques have been used [8]. Numerous studies have been done to find the source and root cause of the spread of this fake news. In aspects of machine learning, lots of methods like Support Vector Machine, Naive Bayes, K Means, etc. are used to find false news along with verifying the accuracy and classification report of these models. In this study, the Logistic Regression model is used to detect and find the authenticity of news by classifying them as fake or real based on training data. This technique provides results in 0 or 1 i.e., binary classification for finding news either true or false. This method includes cleaning and preprocessing of the labeled dataset such as removing stop words, punctuation, and outliers. Methods like lemmatization are also used for data preprocessing that can convert words into their base forms. Also, after preprocessing the TF IDF vectorization method is used to vectorize the dataset into numerical values. Then the dataset is divided into two parts i.e. training and testing data. Afterward, this logistic regression model will be fitted to the training data. Logistic regression calculates the sum of the weighted inputs and then passes this value into a sigmoid function that converts it into a probability between 0 and 1. This probability is then compared with a threshold value which is used to make the final classification. If the probability value is less than the threshold then it will be classified as class 0 otherwise class 1. Thus, the model will predict the output in the form of class 0 or 1 where class 0 depicts the false news and class 1 depicts the true news of the input text. To evaluate the performance of this model, accuracy metrics will be used to find its efficiency and accuracy on both training and testing data. Hence, with the help of the right tools and algorithms, finding the news whether fake or real and ensuring the authenticity of the news articles has been done. Figure 1 shows the pictorial and systematic representation of the method consisting of all the preprocessing aspects and tools used.

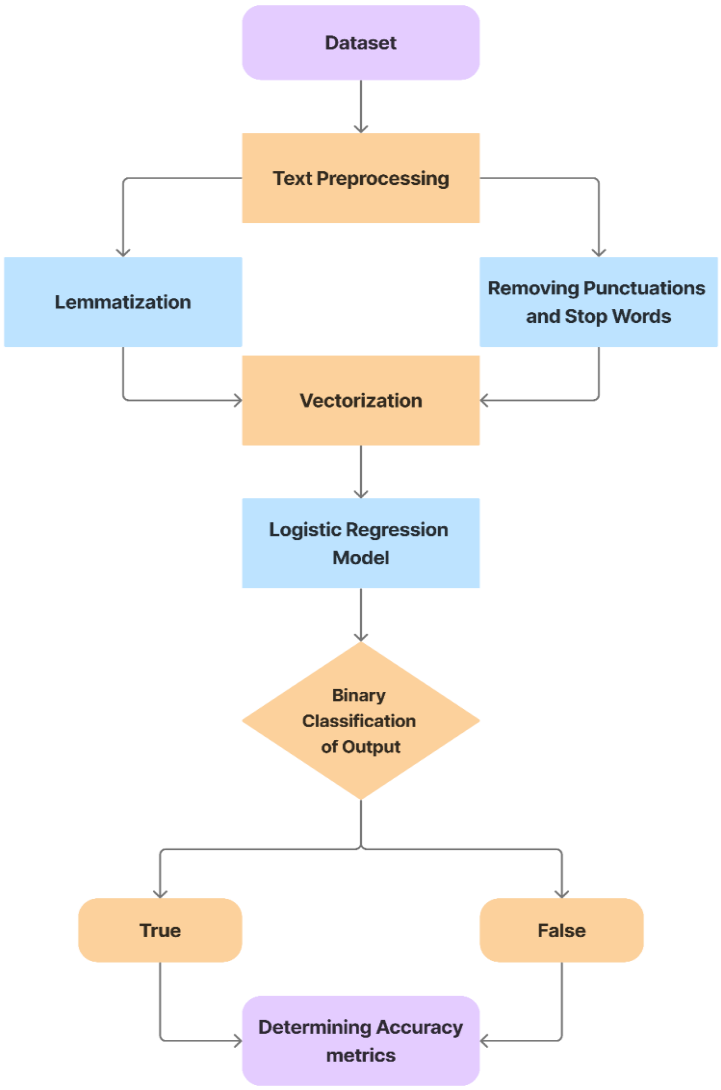

Figure 1: Pictorial representation of working of logistic regression model

4 Data Preprocessing

Data processing is the crucial aspect of refining the labeled dataset for further model training. It shapes the quality of the model as it ensures which data points are needed to train it. Firstly, stop words, unnecessary punctuation and outliers are removed from the datasets to reduce the noise. This is the most fundamental step in data cleaning as it is always required while training machine learning models. Secondly, the data is reduced to its base form using lemmatization as it cuts irrelevant texts and increases the consistency of the dataset. It leads to taking meaningful data only and dropping the irrelevant ones. Additionally, the TF IDF vectorization technique is used to vectorize the data points into numerical values allowing the logistic regression model to train efficiently. It is highly efficient in text classification and allows machine learning models to differentiate between key terms in a document. Also, finding useful features and properties in the dataset ensures efficient training and accuracy of the model. Thus, data preprocessing plays an immense role in developing and training a model.

Proceedings DOI: 10.21467/proceedings.178
ISBN: 978-81-984081-8-1

Series: AIJR Proceedings
ISSN: 2582-3922

4.1 Lemmatization

Lemmatization is a text preprocessing technique in natural language processing that concise words to their root forms. Unlike stemming, lemmatization considers the grammatical context for correct root word identification. It benefits machine learning models in several ways. Firstly, it reduces vocabulary size, enhancing model efficiency and mitigating overfitting. Secondly, it ensures consistent representation of words with the same meaning, bolstering feature representation. Lastly, by focusing on core word meanings, lemmatization aids in generalization, making models less susceptible to minor variations and better equipped to handle unseen data.

4.2 Eliminating Punctuations and Stop Words

To perform text preprocessing, removing punctuations and stop words is needed as they have no significance while training the model to classify fake news. Besides, most of the punctuation should be eliminated but there are still some punctuations like question marks and exclamation marks which affect the meaning of the news. E.g., People are dying due to starvation. People could die due to starvation. Similarly, stop words such as the, is, are, has etc. must be removed for better text preprocessing as they do not provide any significance while processing. E.g., Delhi is a polluted city. Required Words for text preprocessing are Delhi, polluted and city.

4.3 Missing Variables

People now use social media extensively in their daily lives for content creation and communication. Likes are an essential part of the several ways one can respond to material on social media. They express preferences, which propel current markets or open the door to the emergence of fresh ones. However, the circumstances prevent the target universe of the respondents from having a dimension, which calls for vigilance in the handling of the missing numbers. The handling of missing data offers a pertinent issue in terms of statistical analysis.

4.4 Useful Feature Extraction and Vectorization

There is always a requirement for feature extraction to find relevant features. These features lead to the development of a robust machine learning model. The most popular and practical approach to data representation for tasks involving regression and classification still features vector definition. The size of the data table decides which feature extraction technique is most often used. The size of data tables is growing along with the efficiency of data storage. The secret for conducting the experimental study is to extract useful information from the text while avoiding needless data processing. Text information holds a variety of data sizes. Structured data is a crucial consideration when trying to extract features from the text. In most cases, unprocessed raw data is transformed into structured data in the text. This conversion enables the model to interpret the data efficiently [9].

5 Model Description

Logistic regression works in stages, combining concepts from linear regression with a special function to handle probabilities for classification. It takes features (x axis) and predicts a continuous value on the y axis. Logistic regression builds on this idea, but instead of predicting a value anywhere on the number line, it wants to predict the probability of an event happening (coded as 1) or not happening (coded as 0). The sigmoid function, which is also referred to as the logistic function, takes output of the linear equation and puts it between 0 and 1. A value near to 1 refers to a higher chance of the occurrence of an event i.e., class 1, and a value near to 0 refers to a lower chance i.e., class 0. Each feature holds a weight, and the sigmoid function works on a linear combination of these weights and the feature values. Then the model adjusts these weights during training to minimize the error in predicting probabilities

for the training examples. The sigmoid function gives a probability, but one certain answer is needed for classification i.e., class 1 or class 0. A common approach is to set a threshold, usually at 0.5 [10]. Figure 2 illustrates that if the predicted probability from the sigmoid function is greater than 0.5, the model classifies the data point as belonging to class 1. Conversely, if the prediction is lower than 0.5 then it is classified as class 0.

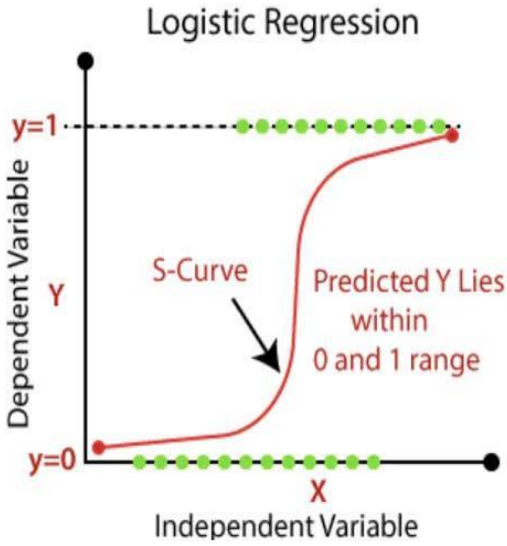

Figure 2: Prediction of the dependent variable using Logistic Regression

6 Accuracy Metric

Various evaluation metrics can be used to check the performance of the machine learning model such as f1 score and confusion matrix on multiple parameters [11]. The accuracy score metric from the scikit learn library evaluates the performance of the logistic regression model in machine learning. It calculates the percentage of right predictions of the model on a reference dataset. This parameter verifies the performance and accuracy of the trained model. The function takes two arguments consisting true label and prediction label in which the true label describes the ground truth labels or actual labels of the data while prediction labels are predicted by the model. The accuracy of the model is evaluated by dividing the correctly predicted samples by the total number of samples. As a result, the model shows an accuracy score of 0.98 on training data and 0.97 on texting data as shown in Figure 3.

Proceedings DOI: 10.21467/proceedings.178
ISBN: 978-81-984081-8-1

Series: AIJR Proceedings
ISSN: 2582-3922

```
In [81]: # accuracy score with training data
         X_train_prediction = model.predict(X_train)
         training_data_accuracy = accuracy_score(X_train_prediction, Y_train)

In [86]: print('Accuracy score of training data : ', training_data_accuracy)

         Accuracy score of training data :  0.9866586538461538

In [87]: print(training_data_accuracy)

         0.9866586538461538

In [88]: # accuracy score with testing data
         X_test_prediction = model.predict(X_test)
         test_data_accuracy = accuracy_score(X_test_prediction, Y_test)

In [89]: print('Accuracy score of testing data : ', test_data_accuracy)

         Accuracy score of testing data :  0.9790865384615385
```

Figure 3: Accuracy score of logistic regression model on training and testing data

7 Result

To resolve the issue of fake news, the logistic regression model is discussed in this study. The logistic regression model works on the binary classification of the data and gives answers in classes 0 and 1 Useful text preprocessing and vectorization technique such as TF IDF vectorization in the labeled datasets allows the model to train on significant data points by recognizing the relations among them. The output of this method gives a probability, if it is higher than the defined threshold value then it will be classified as true or real news. Similarly, if the probability of the model falls below the threshold value, then it will be classified as fake news. This approach helps to distinguish between fake and real news with remarkable accuracy and efficiency. The accuracy scores of the logistic regression model are 0.98 and 0.97 on training and testing data. Other parameters of performance metrics such as F1 score, and confusion matrix are also evaluated with high results showing the compatibility and high performance of the model. As the performance scores of the model are high, predicting the news either fake or real would be done efficiently. The model is trained on labeled datasets and works well while predicting output on similar data and news statements. Also, the model predicts correct results on unseen data that are related to the training and testing dataset. Thus, the ability to predict output with high accuracy allows the model to prevent misinformation and rumors among the vast network of social media.

8 Conclusion

This is a comprehensive study on the vital issue of fake news over the internet. The humongous situation of rumors and misleading news leads to an immense matter of concern among people. To address this problem, the logistic regression technique is used to figure out the authenticity of the news and find whether it is fake or not. Firstly, data processing is done by converting the words into their base form using lemmatization, removing outliers, stop words, and punctuation to refine the dataset. Secondly,

the TF IDF vectorization method is used to assign weight to each text for further classification. The logistic regression model is then trained and tested on the labeled dataset. Also, the performance of the model is verified using accuracy metrics, which are 0.98 and 0.97 on training and testing data. The analysis and prediction capability of the model makes it compatible with predicting fake news spreading over the internet. Although the accuracy of the model is high, it cannot classify results in the case of complex data as the algorithm works upon the linear relationship between the variables only. Furthermore, the model is trained on a limited dataset and cannot predict the output upon news spreading over the internet in the present time. However, integrating the model with deep learning techniques and advanced natural language preprocessing methods would allow the model to make the right predictions on more complex data points. Thus, the process of increasing the scalability of the model and working upon the real time fake news detection are crucial aspects for future consideration.

9 Declarations

9.1 Competing Interests

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

9.2 Publisher's Note

AIJR remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

How to Cite

Aayushya Kumar, Satyam Kumar Mishra, Khushar Shukla, Smriti Srivastava (2025). Fake News Detection Using Logistic Regression Method. *AIJR Proceedings*, 132-138. https://doi.org/10.21467/proceedings.178.16

References

[1] J. Shaikh and R. Patil, "Fake News Detection using Machine Learning," *IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC)*, pp. 1–5, Dec. 2020. https://doi.org/10.1109/iSSSC50941.2020.9358890

[2] P. Bahad, P. Saxena, and R. Kamal, "Fake News Detection using Bi-directional LSTM-Recurrent Neural Network," *Procedia Computer Science*, vol. 165, pp. 74-82, 2019. https://doi.org/10.1016/j.procs.2020.01.072

[3] M. F. Mridha, A. J. Keya, M. A. Hamid, M. M. Monowar, and M. S. Rahman, "A Comprehensive Review on Fake News Detection With Deep Learning," in *IEEE Access*, vol. 9, pp. 156151-156170, 2021. doi: 10.1109/ACCESS.2021.3129329

[4] A. Pal, Pranav, and M. Pradhan, "Survey of fake news detection using machine intelligence approach," *Data & Knowledge Engineering*, vol. 144, 2023, 102118. https://doi.org/10.1016/j.datak.2022.102118

[5] V. Agarwal, H. P. Sultana, S. Malhotra, and A. Sarkar, "Analysis of Classifiers for Fake News Detection," *Procedia Computer Science*, vol. 165, pp. 377-383, 2019. https://doi.org/10.1016/j.procs.2020.01.035

[6] M. A. Alonso, D. Vilares, C. Gómez-Rodríguez, and J. Vilares, "Sentiment Analysis for Fake News Detection," *Electronics*, vol. 10, no. 11, 2021. https://doi.org/10.3390/electronics10111348

[7] S. Rastogi and D. Bansal, "A review on fake news detection 3T's: typology, time of detection, taxonomies," *International Journal of Information Security*, vol. 22, pp. 177–212, 2023. https://doi.org/10.1007/s10207-022-00625-3

[8] S. Pal, T. Kumar, and S. Pal, "Applying Machine Learning to Detect Fake News," *Indian Journal of Computer Science*, vol. 4, pp. 7, 2019. https://doi.org/10.17010/ijcs/2019/v4/i1/142411

[9] S. Gilda, "Notice of Violation of IEEE Publication Principles: Evaluating machine learning algorithms for fake news detection," *IEEE 15th Student Conference on Research and Development (SCOReD)*, pp. 110–115, 2017. https://doi.org/10.1109/SCORED.2017.8305411

[10] M. Mittlböck and M. Schemper, "Explained variation for logistic regression," *Statistics in Medicine*, vol. 15, no. 19, pp. 1987–1997, Oct. 1996. https://doi.org/10.1002/(SICI)1097-0258(19961015)15:19<1987::AID-SIM318>3.0.CO;2-9

[11] F. Özbay and B. Alatas, "Fake news detection within online social media using supervised artificial intelligence algorithms," *Physica A: Statistical Mechanics and its Applications*, vol. 540, 123174, 2019. https://doi.org/10.1016/j.physa.2019.123174