IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Fake News Detection Using Logistic Regression And Decision Tree Algorithm

[1]Mr. R. Arihara Suthan, [2]Dr. R. Sri Devi
[1]Student, [2]Assistant Professor
[1]Department of Computer Applications (PG),
[1]Hindusthan College of Arts and Science, Coimbatore, India

Abstract: In today's digital age, where the internet is ubiquitous, people increasingly rely on online resources for news. With the prevalent use of social media platforms like Facebook and Twitter, information spreads rapidly among millions of users within seconds. However, this rapid dissemination also facilitates the extend of fake news, which can have serious consequences ranging from shaping biased opinions to influencing election outcomes in Favor of certain candidates. Additionally, spammers exploit sensational headlines as clickbait to generate ad revenue. The aim of the presented work to empower users with a tool to determine the authenticity of news articles and verify the credibility of the sources publishing them. In this paper, we propose a binary classification approach such as Logistic Regression and Decision Tree Classifier to analyse online fake new detection. The comparative result shown that decision tree classifier is more accurate to detect fake news than logistic regression.

Index Terms **- Classification techniques, Decision Tree Algorithm, Fake News Detection, Logistic Regression, Natural Language Toolkit (NLTK).**

I. INTRODUCTION

In recent years, social media has become a dominant part of everyday life, serving as a primary source of news for many individuals. However, it has also become a major channel for the spread of fake news, posing significant threats to politics, finance, education, democracy, and business. While misinformation has always existed, the growing reliance on social media has made it easier for people to believe and spread fake news without verification. Distinguishing between true and false information has become increasingly challenging, leading to widespread confusion and misunderstandings. Manually identifying fake news is difficult and often requires extensive knowledge of the topic. Meanwhile, advancements in computer science have made it easier to create and distribute misinformation, but detecting and verifying its authenticity remains a complex task. Fake news can impact businesses by damaging reputations and misleading consumers, while in politics, it can influence public opinion and even affect careers.

Jency Jacob, Managing Director at the Mumbai-based fact-checking website BOOM, highlighted the increasing challenge of misinformation, stating, “2019 has been a unique year where fact-checkers continuously kept moving from one event to the other, and this has been the busiest year for us so far.”

To address this issue, we compare different supervised classification techniques, including Logistic Regression and Decision Tree Classifier, for detecting fake news. The dataset containing both real and fake news articles is taken from Kaggle and implemented using python. The proposed approach achieves optimal results in distinguishing misinformation from credible sources.

The second part contains literature review, third one is proposed work, the next methodology has been discussed, result and discussion, and last conclusion.

II. LITERATURE REVIEW

Mykhailo Granik *et al.* [1] proposed a simple approach for fake news detection using a Naïve Bayes classifier. Their method was implemented as a software system and tested on a dataset of Facebook news posts, collected from three large Facebook pages each from the right and left political spectrums, along with three mainstream political news pages (Politico, CNN, ABC News). Their model achieved a classification accuracy of approximately 74%, though the accuracy for fake news detection was slightly lower, likely due to dataset skewness, with only 4.9% of the data consisting of fake news.

Marco L. Della Vedova *et al.* [2] introduced a novel machine learning-based fake news detection method that combines both news content and social context features. Their approach outperformed existing methods, achieving an accuracy of 78.8%. Additionally, they implemented their method within a Facebook Messenger chatbot, validating its effectiveness in a real-world setting and achieving a fake news detection accuracy of 81.7%.

Veronica Perez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea [3] developed computational models and resources for automated fake news detection. They used two datasets—one collected directly from the internet and another created through a combination of manual curation and online sources. Their analysis identified key linguistic properties associated with fake news content. Using a fake news detection model based on linguistic features, they achieved an accuracy of up to 78%.

Hadeer Ahmed *et al.* [4] proposed a machine learning model for fake news detection, incorporating n-gram analysis. Their best results were obtained using the Term Frequency-Inverted Document Frequency (TF-IDF) technique in combination with a Linear Support Vector Machine (LSVM) classifier, achieving an impressive accuracy of 92%.

Additionally, Mykhailo Granik [5] applied a Naïve Bayes classifier for fake news detection, implementing it as software and testing it on a dataset of Facebook news posts. Their model achieved a classification accuracy of approximately 74%.

III. PROPOSED WORK

Text data requires preprocessing to be transformed into a suitable format for data modelling. Various techniques are widely used for this purpose, and in our approach, we utilized Natural Language Processing (NLP) techniques, specifically the Natural Language Toolkit (NLTK). The preprocessing steps applied to news headlines and articles include stop-word removal, punctuation removal, and stemming. These steps help reduce the data size by eliminating superfluous information while retaining essential content.

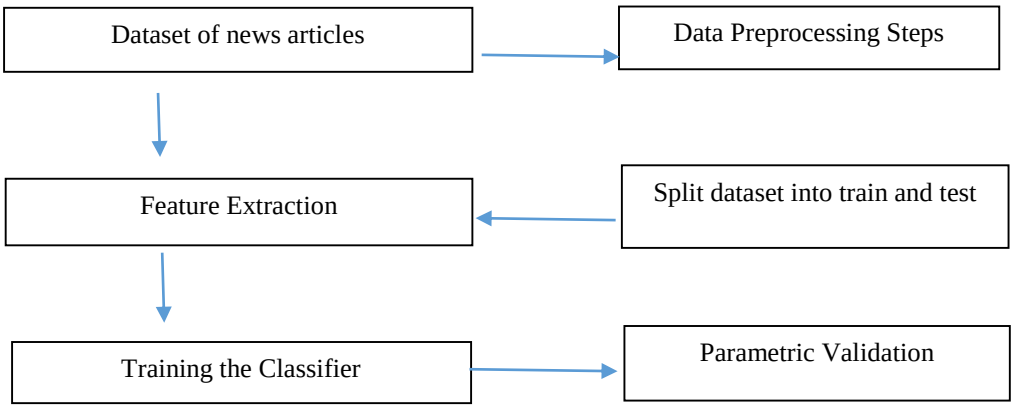

fig.1. flowchart for fake news detection

Dataset of News Articles

The dataset consists of news articles labelled as fake or real to train the machine learning model for spam news detection. It contains text data along with metadata such as titles, subjects, and dates.

Data Preprocessing Steps

Data preprocessing involves removing special characters, punctuation, stop words, and converting text to lowercase. Additionally, URLs, HTML tags, and numerical values are eliminated to clean the dataset for better model performance.

Split Dataset into Train and Test

The dataset is divided into training (75%) and testing (25%) sets to train the classifier and evaluate its performance. This ensures that the model generalizes well to unseen data.

Feature Extraction

TF-IDF (Term Frequency-Inverse Document Frequency) is used to convert the text data into numerical vectors, enabling the machine learning model to understand and process textual information effectively.

Training the Classifier

Different machine learning models such as Logistic Regression and Decision Tree are trained using the training dataset to classify news as fake or real.

Parametric Validation

The trained models are evaluated using precision, recall, F1-score, and accuracy to measure their effectiveness. The best-performing model is selected based on these performance metrics to ensure reliable spam news detection.

We explored two feature selection methods: Term Frequency (TF) and Term Frequency-Inverted Document Frequency (TF-IDF). TF measures the importance of a word based on its occurrence within a document, representing the document as a collection of words. In contrast, Inverse Document Frequency (IDF) determines the rarity of a word across multiple documents. TF-IDF combines both metrics to assess the significance of a word within a document while adjusting for its frequency across the entire corpus. Words with higher TF-IDF scores are considered more important for distinguishing content.

Our preprocessing pipeline began with cleaning the dataset by removing unnecessary words and characters. Next, we performed feature extraction using the TF-IDF method. To evaluate the effectiveness of our approach, we implemented different classification algorithms: Logistic Regression and Decision Tree Classifier. These classifiers were developed using Python's Natural Language Toolkit (NLTK). We then split the dataset into training and testing sets, allocating 80% for training and 20% for testing to assess model performance.

IV. METHODOLOGY

The dataset is collected from Kaggle of 11,220 samples, comprising both fake and real news articles, is loaded and labelled appropriately—assigning a class of 0 to fake news and 1 to real news. To facilitate manual testing, the last ten rows from each dataset are set aside, while the remaining data is merged and shuffled to ensure randomness. The implementation is done by python to predict Fake News.

Next, unnecessary columns such as title, subject, and date are removed. Text preprocessing techniques are applied to clean the text by removing punctuation, URLs, special characters, and converting all text to lowercase. The dataset is then split into training and testing sets, followed by vectorization using the TfidfVectorizer, which converts textual data into numerical features for machine learning models.

Several machine learning models, including Logistic Regression and Decision Tree Classifier, are trained and evaluated using accuracy scores and classification reports. Additionally, a manual testing function is implemented, allowing users to input news text, which is then classified as either "Fake News" or "Real News" using the trained Logistic Regression model. The script concludes by prompting the user to input a news article and returning a prediction based on the model’s classification.

V. RESULT AND DISCUSSION

The comparative result between logistic regression and decision tree algorithm are explained below.

Logistic Regression

Performs well on linearly separable data. Efficient and interpretable but struggles with complex relationships. The classification report shows that the spam news detection model performs exceptionally well with 98.56% accuracy. It achieves high precision (0.99) and recall (0.98-0.99), ensuring minimal misclassification of fake and real news. The F1-score (0.99) confirms a strong balance between precision and recall. With a test dataset of 11,220 samples, the model effectively distinguishes between real and fake news, making it highly reliable for spam news detection.

Decision Tree Classifier

Easy to interpret but prone to overfitting. The spam news detection model achieves near-perfect accuracy (99.53%), with precision, recall, and F1-scores close to 1.00 for both fake and real news. With 11,220 test samples, it effectively classifies news with minimal errors, making it highly reliable for spam detection.

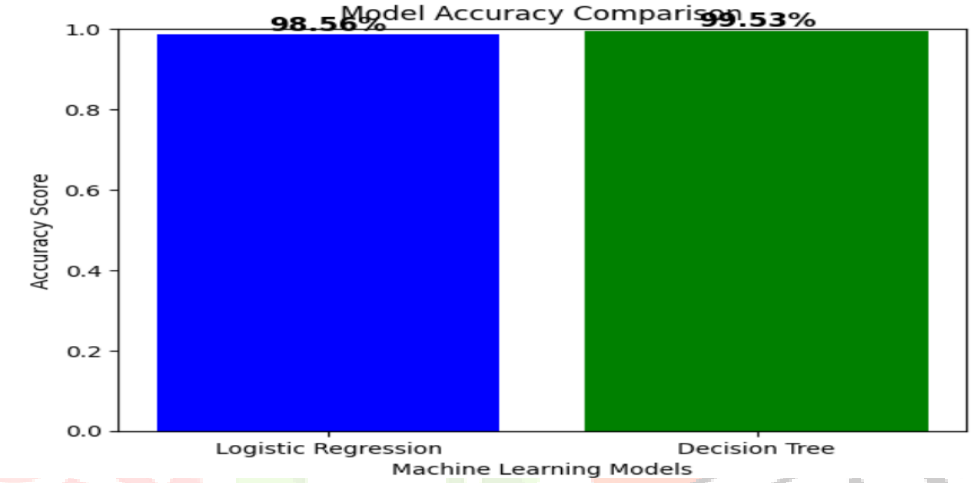

fig.2 comparison between logistic regression and decision tree algorithm for detecting fake news

VI. CONCLUSION

Our Fake News Detection System classifies news as true or fake using NLP and ML techniques. Trained on a relevant dataset, it identifies linguistic and contextual patterns in misinformation. By detecting fake news early, the system helps mitigate its impact. To enhance accuracy, integrating diverse datasets and traditional fact-checking methods is essential. The classification report of logistic regression shows that the spam news detection model performs exceptionally well with 98.56% accuracy. The classification report of Decision Tree shows that the spam news detection model performs exceptionally well with 99.53% accuracy.

REFERENCES

[1] M. Granik and V. Mesyura, "Fake news detection using naive Bayes classifier," 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), Kiev, 2017, pp. 900-903.
[2] M. L. Della Vedova, E. Tacchini, S. Moret, G. Ballarin, M. DiPierro and L. de Alfaro, "Automatic Online Fake News Detection Combining Content and Social Signals," 2018 22nd Conference of Open Innovations Association (FRUCT), Jyvaskyla, 2018, pp. 272-279.
[3] Pérez-Rosas, Verónica & Kleinberg, Bennett & Lefevre, Alexandra & Mihalcea, Rada. (2017). Automatic Detection of Fake News.
[4] Traore, Issa & Saad, Sherif. (2017). Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. 127- 138. 10.1007/978-3-319-69155-8_9.

[5] Mykhailo Granik, Volodymyr Mesyura, "Fake News Detection Using Naive Bayes Classifier", 2017 IEEEFirst Ukraine Conference on Electrical and Computer Engineering (UKRCON).

[6] Akshay Jain and AmeyKasbe. "Fake News Detection." 2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS). Bhopal, India: IEEE. 2018.

[7] W. Y. Wang, "Liar, Liar Pants on Fire: A New Benchmark Dataset for Fake News Detection," *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2017.

[8] H. Allcott and M. Gentzkow, "Social Media and Fake News in the 2016 Election," *Technical Report National Bureau of Economic Research*, 2017.

[9] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake News Detection on Social Media," *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, pp. 22–36, 2017.