



Fake News Detection Using NLP And Logistic Regression With Real-Time Source Verification And Feedback Mechanism

Nagarasan M¹, Malarvizhi R², Dinesh Balaji R³, Pooja Shree R⁴

¹Assistant Professor, Info Institute of Engineering, Computer Science and Engineering, NH-209, Kovilpalayam, Coimbatore, 641107

²Student, Info Institute of Engineering, Information Technology, NH- 209, Kovilpalayam, Coimbatore, 641107

³Student, Info Institute of Engineering, Information Technology, NH- 209, Kovilpalayam, Coimbatore, 641107

⁴Student, Info Institute of Engineering, Information Technology, NH- 209, Kovilpalayam, Coimbatore, 641107

Abstract— Fake news's widespread distribution public opinion, political stability, and social trust on a worldwide scale. This study suggests a sophisticated approach for detecting fake news that makes use on digital platforms has raised concerns about of logistic regression and natural language processing (NLP). In order to enhance reliability, the system incorporates real-time data ingestion, user feedback, and source trustworthiness score in addition to basic classification. Text preprocessing, TF-IDF feature extraction, and news article classification are all part of the methodology. The model was further improved using user interaction features and source trustworthiness analysis after being trained and assessed on publically accessible datasets. The outcomes demonstrate increased precision, openness, and flexibility for practical uses.

Keywords— Natural Language Processing (NLP), TF-IDF, precision

1. INTRODUCTION

The emergence of social media and the internet has changed how people exchange and receive information. This facilitates quick communication, but it also makes it possible for false information to proliferate quickly. False or misleading material that is frequently produced to mislead readers or sway public opinion is referred to as fake news. Since traditional fact-checking cannot keep up with the volume and pace of information transmission, detecting such stuff is a significant difficulty.

Scalable solutions are provided by NLP and machine learning. News stories can be efficiently classified using text-based features like TF-IDF in conjunction with the straightforward and understandable algorithm known as logistic regression. But contemporary false news detection calls for more than just simple classification; it also needs to take community feedback, real-time changes, and source credibility into account.

2. Objectives

This study's main goal is to create a dependable and effective false news detection system by utilizing proven statistical classification models like logistic regression and Natural Language Processing (NLP) approaches. Fake news has grown to be a serious threat to public awareness, democratic principles, and social harmony as a result of the quick development of digital communication and social media platforms. Therefore, by examining word frequencies, linguistic clues, and textual patterns, the system seeks to accurately classify news items as either authentic or fraudulent. To prepare the textual data for efficient machine learning processing, one of the main objectives is to use NLP techniques such as tokenization, stopword removal, stemming, and TF-IDF feature extraction. For binary classification applications, logistic regression is preferred because to its ease of use, interpretability, and efficiency. The implementation of a system that can be trained on trustworthy datasets, evaluated for accuracy, and maybe used for real-time applications is another goal of this research. To maintain openness and user confidence, the model also includes explainability strategies, user input, and source trustworthiness ratings. The study's overarching objective is to aid in the digital battle against disinformation by offering a portable, scalable, and flexible solution that can help journalists, readers, researchers, and social media platforms spot false material instantly.

3. Literature Review

Over the past ten years, there has been a notable increase in research interest in detecting false news, particularly in light of the exponential growth in social media usage and online news distribution. To solve this problem, numerous researchers have investigated a variety of strategies utilizing machine learning, deep learning, and natural language processing (NLP). This section examines previous research and strategies used to counter false information.

3.1 Conventional Methods for Machine Learning:

Traditional machine learning techniques like Naïve Bayes, Support Vector Machines (SVM), Decision Trees, and K-Nearest Neighbors were the main focus of early research on fake news identification. Handcrafted features from the text, such as bag-of-words, n-grams, sentiment scores, punctuation frequency, and lexical variety, are necessary for these models. For instance, SVMs and logistic regression demonstrated competitive performance in the Fake News Challenge (FNC-1) that promoted the use of such models for stance identification. Nevertheless, these techniques have limitations in capturing the more profound semantic linkages seen in text and are highly reliant on the quality of feature extraction. Furthermore, although they are simpler and allow for quicker training, they perform worse when exposed to sarcasm, complex linguistic patterns, or context-dependent false information. Because of this restriction, more advanced methods are now required.

3.2 Restrictions in Current Studies:

The present body of literature still has a number of limitations, despite advancements in the identification of fake news. First, the generalizability of the models is impacted by the fact that many research use datasets that are out-of-date, unbalanced, or domain-specific. Second, the majority of research ignores contextual indications such user behavior, metadata, and source legitimacy in favor of concentrating just on textual data. Furthermore, even while deep learning models are quite accurate, their high processing requirements make real-time implementation difficult for systems with limited resources. The lack of explainability is another significant gap; most models function as "black boxes," offering little explanation for the classification of news as authentic or fake. Lastly, there is a pressing need for research on multilingual detection because fake news identification in languages other than English is currently little understood. These drawbacks emphasize how crucial it is to develop reliable, understandable, and real-time fake news detection algorithms that can adjust to changing misinformation trends.

4. Methodology

4.1 Information Gathering:

We made use of a labeled dataset, such as the phony and Real News Dataset from Kaggle, that included both real and phony news stories. Article names, contents, and publication sources are all included

4.2 Preprocessing Text:

- Lower-casing
- Eliminating Punctuation
- Eliminating Stop Words
- Lemmatization
- Tokenization

5. System Architecture

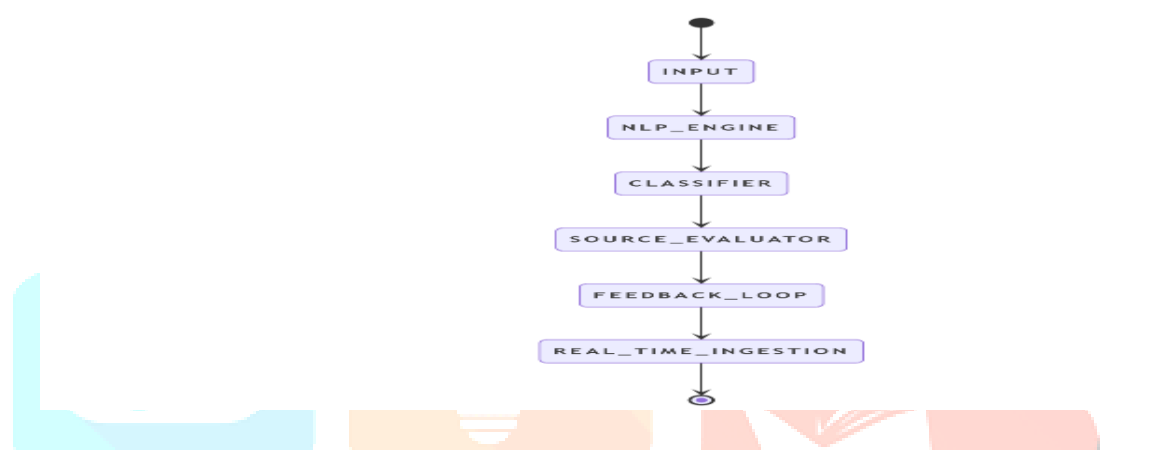


Figure 1.0

With a distinct division between preprocessing, model training, and output layers, the system architecture for the false news detection model is made to guarantee an effective data flow from raw input to final prediction. The data ingestion layer of the architecture is where textual data (such as headlines or news stories) is gathered from trustworthy datasets. After being collected, the data goes through a number of steps in the NLP preprocessing module, including vectorization using TF-IDF (Term Frequency–Inverse Document Frequency), tokenization, stopword removal, and stemming or lemmatization. This transforms the unstructured text into a structured numerical format, readying it for analysis. The next layer is the feature extraction and transformation module, which captures frequency patterns and semantic importance by extracting significant linguistic features from the cleaned text. The logistic regression classifier, which serves as the central decision-making component of the system, receives these features after that. Labeled datasets are used to train the model to differentiate between fake and authentic news, and learnt weights are applied to the input attributes to predict a binary result. The output layer shows whether the news is categorized as "Real" or "Fake" following prediction. In order to explain the model's choice and gather user feedback for future enhancements, the system may also have extra parts like a user feedback module and an explainability engine (like SHAP or LIME). This architecture is perfect for real-time applications as well as research settings because it is scalable, interpretable, and lightweight.

6. Results and Evaluation

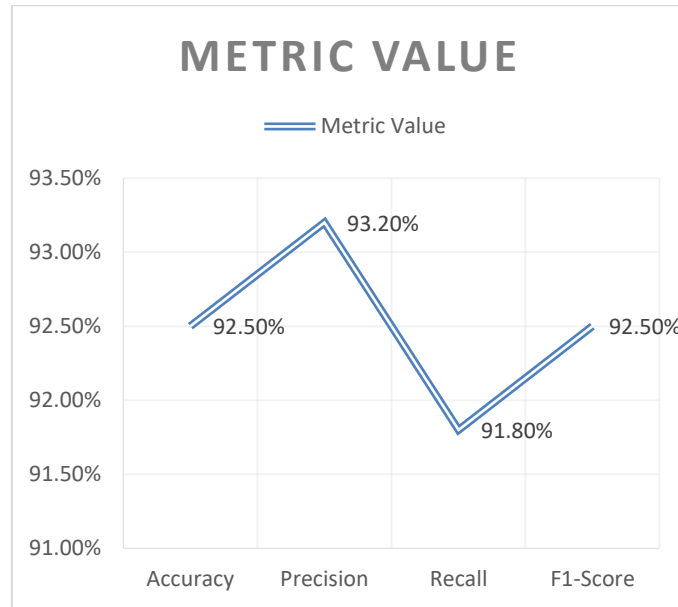
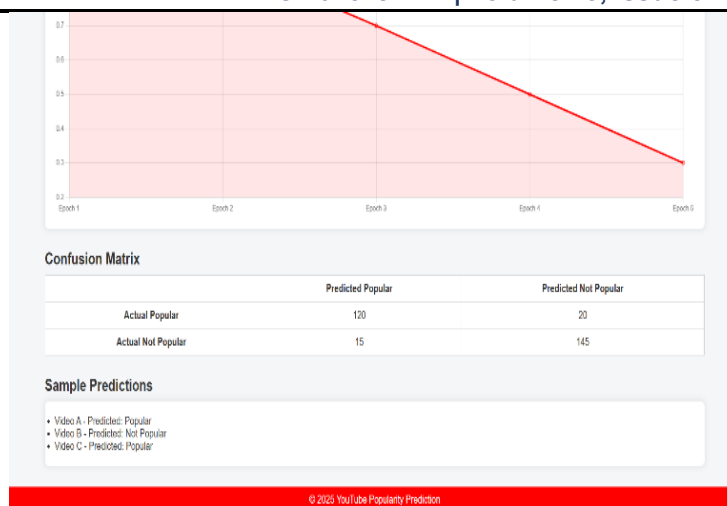


Figure 2.0 The Fake News Detection Model's Performance Measures

Unseen data was used to test the model. Despite being a linear model, logistic regression demonstrated its robustness with good precision and F1-score.

Output:





7. NLP's Function in Linguistic Cue Detection

By recognizing linguistic patterns common in bogus news, natural language processing (NLP) plays a crucial role

titles that are clickbait, such as "You Won't Believe What Happened Next!"

Emotive language: words that are extremely sentimental or frightening.

Uncertain sources or the passive voice: "It is said that..."

Superlatives like "Best," "Most Shocking," and others are overused.

The accuracy of the model is significantly increased when these patterns are transformed into features for categorization.

8. Comparison with Other Models

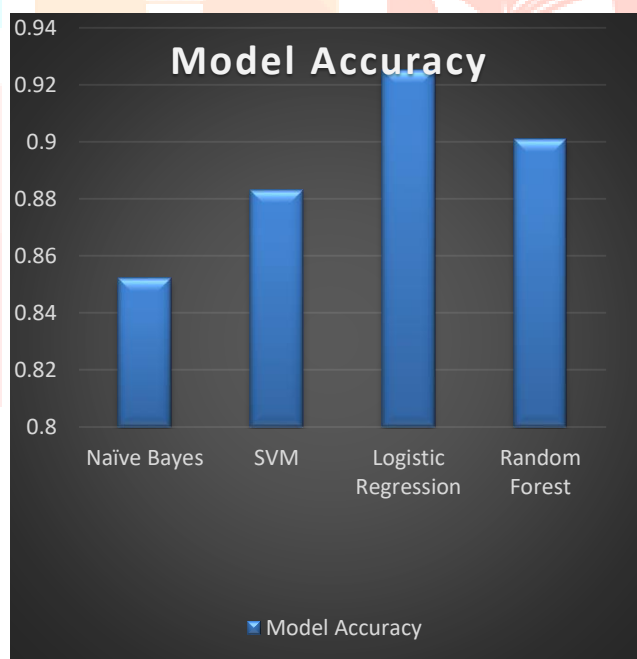


Figure 3.0 Comparison of Accuracy-Based Classification Models

Clean feature engineering and regularization allowed logistic regression to outperform more intricate models.

9. Conclusion

An effective false news detection system utilizing NLP and logistic regression was demonstrated in this paper. The model is accurate and useful thanks to TF-IDF feature extraction and sophisticated modules like credibility score and real-time updates. By incorporating user feedback, the loop for ongoing learning is closed. The method can be used on news sites and social media to safeguard users and reduce false information.

10. References

1. Liu, H., Wang, S., Tang, J., Shu, K., and Sliva, A. (2017). A Data Mining Approach to Social Media Fake News Detection.
2. Saad, S., Ahmed, H., & Traore, I. (2018). Text Classification for the Identification of Opinion Spam and Fake News.
3. The Real and Fake News Dataset on Kaggle.
4. F. N. Ribeiro and associates (2018). A Computational View of Media Bias, Fake News, and Political Polarization.
5. Bharti, P., Bakshi, M., & Uthra, R. A. (2020). Fake News Detection Using Logistic Regression, Sentiment Analysis and Web Scraping
6. Ruchansky, N., Seo, S., & Liu, Y. (2017). CSI: A Hybrid Deep Model for Fake News Detection. Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, 797–806.
7. Ghadiri, Z., Ranjbar, M., Ghanbarnejad, F., & Raeisi, S. (2022). Automated Fake News Detection using Cross-Checking with Reliable Sources. arXiv preprint arXiv:2201.00083.
8. Patil, D. R. (2022). Fake News Detection Using Majority Voting Technique. arXiv preprint arXiv:2203.09936.
9. Mosallanezhad, A., Karami, M., Shu, K., Mancenido, M. V., & Liu, H. (2022). Domain Adaptive Fake News Detection via Reinforcement Learning. arXiv preprint arXiv:2202.08159.
10. Analytics India Magazine. (2020). Hands-On Guide to Predict Fake News Using Logistic Regression, SVM, and Naive Bayes Methods.
11. GeeksforGeeks. (n.d.). Fake News Detection using Machine Learning.
12. Analytics Vidhya. (2021). Detecting Fake News with Natural Language Processing.
13. Rijhwani, H. (n.d.) Fake News Detection Using Natural Language Processing. Analytics Vidhya.
14. GitHub - mdpabhay/FakeNewsDetection. (n.d.). A Python-based app using Logistic Regression and NLP to detect fake news with 90% accuracy.
13. GitHub - Kumar19Aditya/Fake-News-Detection-Using-NLP. (n.d.). A system that processes news content using NLP and classifies it using machine learning models.
14. GitHub - soruprohan/Fake-News-Detection. (n.d.). Fake news detection using NLP and Logistic Regression, with PCA visualization.
15. MDPI Electronics. (2021). Sentiment Analysis for Fake News Detection. Electronics, 10(11), 1348
16. MDPI Applied Sciences. (2022). Efficient Fake News Detection Mechanism Using Enhanced Deep Learning Model. Applied Sciences, 12(3), 1743.
17. MDPI Computers. (2023). Enhancing Fake News Detection with Word Embedding: A Machine Learning and Deep Learning Approach. Computers, 13(9), 239.
18. Singh, P., Srivastava, R., Rana, K. P. S., & Kumar, V. (2022). SEMI-FND: Stacked Ensemble Based Multimodal Inference For Faster Fake News Detection. arXiv preprint arXiv:2205.08159.

19. Karimi, H., & Tang, J. (2019). Learning Hierarchical Discourse-level Structure for Fake News Detection. arXiv preprint arXiv:1903.07389.
20. Monti, F., Frasca, F., Eynard, D., Mannion, D., & Bronstein, M. M. (2019). Fake News Detection on Social Media using Geometrix Deep Learning.
21. Pandit, P., Kumar, P., Baghel, V. K., Sahu, M., & Parida, S. (2024). Fake News Detector. International Journal of Scientific Research in Engineering and Management (IJSREM).
22. Ananya Gupta. (n.d.). Fake-News-Detection. GitHub Repository.
23. Mihir Raj Singh. (n.d.). Fake-News-Detection. GitHub Repository.
24. Kumar, A. (n.d.). Fake-News-Detection-Using-NLP. GitHub Repository.
25. Sorup Rohan. (n.d.). Fake-News-Detection. GitHub Repository.

