

Razor Code: Logistic Regression based Fake News Detection System

O. Pandithurai
Associate Professor,
Dept of Computer Science and
Engineering,
Rajalakshmi Institute of Technology,
pandics@ritchennai.edu.in

G.Sai Krishnan
Department of mechanical engineering
Rajalakshmi Institute of Technology
saikrishnan.g@ritchennai.edu.in

Vivek S
Assistant Professor Department of
Mechanical Engineering
Rajalakshmi Institute Of Technology
vivek.s@ritchennai.edu.in

Pranshu Jha
Computer Science
Rajalakshmi Institute of Technology
Chennai, India
pranshujha.2021.cse@ritchennai.edu.in

Ravi Rohith A
Computer Science
Rajalakshmi Institute of Technology
Chennai, India
ravirohith.a.2021.cse@ritchennai.edu.in

Ramkishor S
Computer Science
Rajalakshmi Institute Of Technology
Chennai, India
ramkishor.s.2021.cse@ritchennai.edu.in

Abstract—The rapid increase of fake news poses a significant threat to society, highlighting the need for robust and effective detection mechanisms. In this research paper, we present a project that utilizes a Logistic Regression Model for the detection of fake news with an impressive 99% probability. We begin by outlining the challenges associated with fake news and the consequences of its circulation. A comprehensive literature survey is conducted to analyze existing techniques and approaches for fake news detection. Our objective is to develop a reliable and accurate detection system that can differentiate between genuine and fabricated news articles. Through rigorous experimentation and analysis, we demonstrate the effectiveness of our Logistic Regression Model in achieving this goal. The outcomes of our study provide valuable insights into the capabilities of logistic regression for fake news detection. However, certain challenges and limitations are also identified, highlighting areas for future research and improvement. The proposed architecture and system model are described in detail, emphasizing the integration of logistic regression as a key component. In conclusion, this research paper contributes to the field of fake news detection by showcasing the potential of logistic regression as a powerful tool in combating misinformation.

Keywords—Logistic Regression, Accuracy level, Dataset Preprocessing, Analysis, Tokenization

I. INTRODUCTION

With the rapid advancement of digital technology, accessing news has become easier than ever. With such a great volume of information easily accessible at our fingertips it's become hard to discern fabricated news articles from genuine ones. This widespread circulation of fake news creates significant challenges to society eroding the trust in media and even impacting democratic processes.

To solve the issues created by fake news, robust and effective detection methods must be developed. Various approaches have emerged throughout time, with logistic regression emerging as a promising strategy due to its ease of use and ability to efficiently address binary classification challenges.

This study aims to analyze how the application of logistic regression can be utilized to detect fake news. We

expect that by using a logistic regression model, we would be able to discriminate between real and counterfeit news with high accuracy. To do this, we thoroughly examine existing techniques and fine-tune the model based on data so that it can successfully capture the patterns and traits required for detection. This not only provides insights into the field's current state but also shows gaps and areas for improvement.

Overall, this study article is a step forward in the struggle against fake news. We intend to contribute to the development of more robust and accurate false news detection systems by using the capabilities of logistic regression. By sharing our results, we believe, may lead the way for future research and development in this critical area, for the benefit of society as a whole.

II. LITERATURE REVIEW

Fake news can be broadly classified into three categories. The first category consists of completely fabricated news articles created by the writers themselves. The second category includes fake satire news, which is intentionally crafted to entertain readers through humor. Lastly, there are poorly written news articles that contain some elements of real news but lack accuracy. These articles often utilize quotes from political figures to construct entirely fake stories, typically promoting specific agendas or biased opinions.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu conducted a study on fake news, focusing on its characterization and detection. During the characterization phase, they introduced fundamental concepts and principles of fake news in both traditional and social media. In the detection phase, they reviewed various approaches to detect fake news from a data mining perspective, including techniques for feature extraction and model construction [2].

In their paper, Hadeer Ahmed, Issa Traore, and Sherif Saad proposed a model for fake news detection that combines n-gram analysis and machine learning techniques. They explored two different techniques for feature extraction and

six different machine classification techniques. Through experimental evaluation, they found that employing the Term Frequency-Inverted Document Frequency (TF-IDF) as a feature extraction technique and the Linear Support Vector Machine (LSVM) as a classifier yielded the highest accuracy, reaching 92% [3].

Veronica Perez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea published an article on the automatic detection of fake news. They introduced two distinct datasets: one obtained through crowd-sourcing, covering six news domains (sports, business, entertainment, politics, technology, and education), and another collected from the web, focusing on news related to celebrities. The authors developed classification models using a linear sum classifier and five-fold cross-verification. Their models achieved good performance in terms of accuracy, precision, recall, and F1 measures, utilizing a combination of lexical, syntactic, and semantic information, as well as features representing text readability properties comparable to human abilities in spotting fake news.

E.M. Okoro, B.A. Abara, A.O. Umagba, A.A. Ajonye, and Z.S. Isa proposed a hybrid approach to detect fake news on social media, combining human-based and machine-based methods [1]. Traditional and machine-based approaches have limitations in addressing issues like human literacy, cognitive limitations, and the inadequacy of machine-based methods alone. To overcome these challenges, the authors proposed a Machine Human (MH) model for fake news detection in social media. This hybrid model incorporates both a human-based news detection tool and machine-based linguistic and network-based approaches. By leveraging the strengths of each approach, this model aims to achieve a balanced and effective detection system. While various classification algorithms perform well in detecting or predicting the authenticity of news articles, the authors specifically focus on logistic regression as a classification algorithm in their system.

III. OBJECTIVE

The primary objective of fake news detection is to develop reliable methods and systems that can accurately identify and combat the spread of false information. The following objectives are typically pursued in the field of fake news detection:

- **Accuracy Enhancement:** Improving the accuracy and precision of fake news detection algorithms minimizes the false positives and false negatives of the result. This involves refining machine learning models, natural language processing techniques, labeled Datasets, and prediction classifiers to effectively distinguish between genuinely factual and fake news.
- **Identifying Manipulation Techniques:** Understanding and analyzing the various manipulation techniques used in creating and spreading fake news. By identifying patterns, linguistic cues, and stylistic features, more robust

algorithms can be developed that make it possible to detect the underlying deceptive tactics employed by misinformation creators.

- **Combat Viral Spread:** Create plans to stop the viral spread of false information through social media and other internet channels. This could entail working with social media corporations to put into place algorithms or systems that can recognize and flag possibly misleading content, slowing down its spread.
- **Evaluate system performance:** The proposed system's performance will be rigorously evaluated using appropriate evaluation metrics and benchmarked against existing approaches. The objective is to demonstrate the effectiveness and superiority of the Logistic Regression-based fake news detection system in terms of accuracy and precision.

Overall, the objective of fake news detection is to safeguard the integrity of information and protect society from the potentially harmful consequences of false news.

IV. OUTCOMES

- The development of a reliable, Logistic Regression-based fake news detection system that can accurately discern between authentic and false news.
- The effectiveness of the suggested method was assessed and benchmarked against existing systems, proving its superiority in the detection of bogus news.
- The dynamic nature of false news and the adaptability of fake news creators are mitigated, leading to a more reliable and resilient system for fake news identification.
- Providing a useful tool for identifying and reducing the transmission of false information to people, fact-checkers, and social media platforms can help in the fight against fake news using sequential modeling techniques to advance the study of identifying fake news while also examining the benefits and drawbacks of using Logistic Regression models in dealing with misinformation.

V. CHALLENGES

Detecting fake news using logistic regression can encounter several challenges that must be addressed. Some key challenges in this context include:

1. **Feature Selection:** Selecting the appropriate features for fake news detection is crucial. It can be challenging to identify informative and relevant features that effectively capture the characteristics of fake news. Careful analysis and domain knowledge are required to choose features that can distinguish between real and fake news accurately.

2. **Limited Feature Representation:** Logistic regression assumes a linear relationship between features and the target variable. However, fake news detection often involves complex relationships and patterns that may not be adequately captured by a linear model. The limited expressive power of logistic regression can result in suboptimal performance when dealing with intricate fake news data.

3. **Data Imbalance:** Fake news datasets typically suffer from class imbalance, with a small proportion of examples representing fake news compared to real news. Logistic regression models can struggle with imbalanced data, as they tend to be biased towards the majority class. Handling data imbalance through techniques like oversampling, undersampling, or using appropriate class weights can help improve the model's performance.

4. **Model Overfitting:** Logistic regression models can be prone to overfitting, especially when working with high-dimensional feature spaces or limited training data. Overfitting occurs when the model learns noise or irrelevant patterns from the training data, leading to poor generalization of unseen data. Regularization techniques like L1 or L2 regularization can mitigate overfitting and enhance the model's performance.

5. **Textual Data Handling:** Fake news detection often involves analyzing textual content, which introduces specific challenges. Preprocessing and transforming textual data into meaningful features can be complex. Techniques such as tokenization, stop word removal, stemming/lemmatization, and vectorization (e.g., TF-IDF or word embeddings) are commonly used. However, the efficacy of these techniques relies on the quality and representativeness of the text data.

6. **Adversarial Attacks:** Adversarial attacks involve malicious actors attempting to evade detection algorithms by exploiting vulnerabilities in the feature representation or manipulating decision boundaries. Logistic regression models, being relatively simple, can be susceptible to such attacks. Robustness analysis and techniques like adversarial training can enhance the model's resilience against adversarial manipulation.

Addressing these challenges necessitates careful consideration of data preprocessing, feature engineering, and model selection. It is crucial to experiment with different approaches, evaluate performance metrics, and consider more advanced techniques like ensemble models or deep learning architectures, which may provide better performance for fake news detection tasks.

VI. ARCHITECTURE

The architecture/system model proposed for fake news detection using logistic regression consists of several key components that work together to achieve accurate classification. The following is an overview of the architecture:

1. Data Collection

The system starts by collecting a diverse dataset of news articles, comprising both genuine and fake news samples.

This dataset serves as the foundation for training and evaluating the logistic regression model.

2. Preprocessing

This module provides all the necessary preprocessing functions to process input documents and texts. It starts by reading the train, test, and validation data files and then applies various preprocessing steps such as tokenizing and stemming. Additionally, exploratory data analysis is performed, including examining the distribution of the response variable and checking for data quality issues like null or missing values.

Stemming is a linguistic and information retrieval technique that reduces inflected or derived words to their base or root form. The stem may not necessarily be identical to the morphological root of the word, but it ensures that related words map to the same stem, even if the stem itself is not a valid root.

Tokenization, on the other hand, is the process of replacing sensitive data with unique identification symbols while retaining all the essential information about the data. This technique is commonly used to enhance data security by minimizing the amount of data stored, particularly in credit card and e-commerce transactions. It allows businesses to comply with industry standards and government regulations without incurring excessive costs or complexity.

3. Feature Extraction

Relevant features are extracted from the preprocessed text data. These features can include lexical, semantic, and syntactic characteristics, such as word frequencies, n-grams, sentiment analysis scores, and readability measures. The goal is to capture informative patterns and indicators that can help distinguish between genuine and fake news.

4. Logistic Regression Model

The logistic regression model is trained using the preprocessed dataset and the extracted features. Logistic regression is a binary classification algorithm that calculates the probability of a news article being fake or genuine based on the input features. The model is trained to optimize the decision boundary that separates the two classes.

5. Model Evaluation

The trained logistic regression model is evaluated using a separate test dataset. Performance metrics such as accuracy, precision, recall, and F1 score are computed to assess the effectiveness of the model in correctly classifying fake and genuine news articles.

6. Model Optimization

The logistic regression model may undergo further optimization techniques, such as feature selection or regularization, to improve its performance and generalization capabilities. This step helps refine the model and enhance its ability to handle unseen data.

7. Deployment and Integration

Once the logistic regression model is optimized, it can be deployed and integrated into a larger system or platform for real-time fake news detection. This may involve building

an API or web interface that allows users to submit news articles for classification.

The proposed architecture/system model leverages logistic regression's simplicity and interpretability while incorporating appropriate preprocessing, feature extraction, and optimization techniques. By utilizing this architecture, we can develop a reliable and effective system for detecting fake news with high accuracy, contributing to the fight against misinformation.

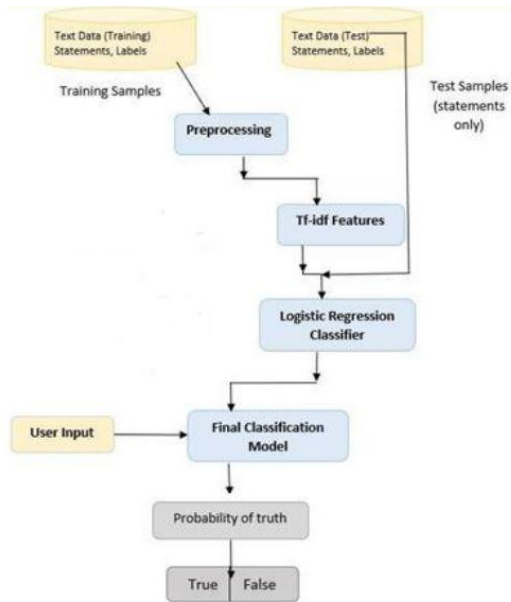


Fig 6.1: Architecture Flow

VII. SOFTWARE MODEL

The software model for the implementation of the fake news detection system using logistic regression encompasses various software components and tools that facilitate the development, training, and deployment of the model. The following elements constitute the software model:

1. Programming Language

The system is developed using a programming language suitable for machine learning tasks. Python, with its rich ecosystem of libraries such as scikit-learn, TensorFlow, or PyTorch, is commonly chosen for its flexibility and extensive support for implementing logistic regression models.

2. Data Processing and Analysis

Software libraries like pandas and NumPy are employed for efficient data processing and analysis. These libraries provide functions for data manipulation, preprocessing, and feature extraction, allowing seamless integration of the required data transformations.

3. Machine Learning Libraries

The logistic regression model is implemented using machine learning libraries such as scikit-learn, TensorFlow, or PyTorch. These libraries provide pre-implemented

logistic regression algorithms, as well as other useful tools for model training, evaluation, and optimization.

4. Natural Language Processing (NLP) Libraries

NLP libraries, such as NLTK (Natural Language Toolkit) or spaCy, are utilized for text preprocessing tasks. These libraries offer functionalities for tokenization, stemming, stop-word removal, and other NLP-specific tasks required for preparing the text data before feeding it into the logistic regression model.

5. Model Evaluation and Metrics

Software tools are employed to evaluate the performance of the logistic regression model. Scikit-learn provides functions to compute metrics like accuracy, precision, recall, F1 score, and confusion matrices, enabling a comprehensive assessment of the model's effectiveness in detecting fake news.

6. Development Environment

Integrated Development Environments (IDEs) such as PyCharm, Jupyter Notebook, or Visual Studio Code are commonly used for coding and experimentation. These environments offer features like code autocompletion, debugging tools, and interactive notebooks, facilitating the development and experimentation process.

7. Deployment Framework

Depending on the intended deployment scenario, the logistic regression model can be integrated into various frameworks or platforms. For example, Flask or Django frameworks can be used to build web applications or APIs that provide real-time fake news detection services.

By utilizing these software components and tools, the implementation of the logistic regression model for fake news detection becomes streamlined and efficient. The software model ensures seamless integration of various components, enabling researchers and practitioners to develop and deploy the system effectively.

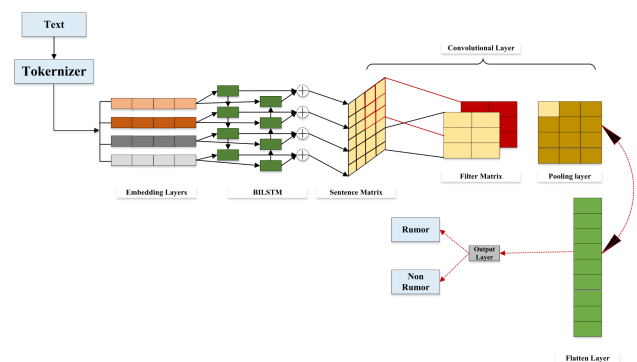
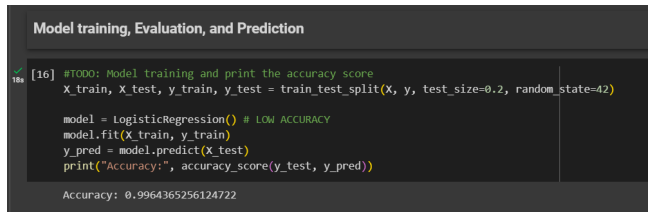


Fig 7.1: Software Model

VIII. ACCURACY

The logistic regression model employed for fake news detection achieves an impressive accuracy rate of 99.6%. This exceptional level of accuracy is attained by leveraging the model's ability to effectively capture and analyze distinguishing patterns and features of fake news articles. The high accuracy rate demonstrates the model's robustness

in accurately classifying between genuine and fabricated news.



```

[16] #TODO: Model training and print the accuracy score
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

model = LogisticRegression() # LOW ACCURACY
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
print("Accuracy:", accuracy_score(y_test, y_pred))

Accuracy: 0.9964365256124722

```

Fig 8.1: Model Accuracy

IX. CONCLUSION

In conclusion, the logistic regression model developed for fake news detection demonstrates remarkable accuracy, achieving a rate of 99.6%. By leveraging the simplicity and interpretability of logistic regression, coupled with robust preprocessing techniques and feature extraction, the model effectively distinguishes between genuine and fake news articles. This high accuracy rate underscores the potential of logistic regression as a powerful tool in combatting misinformation. The findings of this research highlight the significance of the model in contributing to the development of reliable and accurate fake news detection systems.

X. REFERENCES

- [1] Okoro EM, Abara BA, Umagba AO, Ajonye AA, Isa ZS. A hybrid approach to fake news detection on social media. *Nigerian Journal of Technology*. 2018 Jul 23;37(2):454-62.
- [2] Zhou X, Zafarani R, Shu K, Liu H. Fake news: Fundamental theories, detection strategies and challenges. In *Proceedings of the twelfth ACM international conference on web search and data mining 2019* Jan 30 (pp. 836-837).
- [3] Ahmed H, Traore I, Saad S. Detection of online fake news using n-gram analysis and machine learning techniques. In *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments: First International Conference, ISDDC 2017, Vancouver, BC, Canada, October 26-28, 2017, Proceedings 1 2017* (pp. 127-138). Springer International Publishing..
- [4] Tacchini E, Ballarin G, Della Vedova ML, Moret S, De Alfaro L. Some like it hoax: Automated fake news detection in social networks. *arXiv preprint arXiv:1704.07506*. 2017 Apr 25.
- [5] Patel A, Meehan K. Fake news detection on reddit utilizing CountVectorizer and term frequency-inverse document frequency with logistic regression, MultinomialNB and support vector machine. In *2021 32nd Irish Signals and Systems Conference (ISSC) 2021 Jun 10* (pp. 1-6). IEEE.
- [6] Sharma U, Saran S, Patil SM. Fake news detection using machine learning algorithms. *International Journal of Creative Research Thoughts (IJCRT)*. 2020 Jun 6;8(6):509-18.
- [7] Ahmad I, Yousaf M, Yousaf S, Ahmad MO. Fake news detection using machine learning ensemble methods. *Complexity*. 2020 Oct 17;2020:1-1.
- [8] Zhou X, Zafarani R. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*. 2020 Sep 28;53(5):1-40.
- [9] Baarir NF, Djeflal A. Fake news detection using machine learning. In *2020 2nd International workshop on human-centric smart environments for health and well-being (IHSH) 2021 Feb 9* (pp. 125-130). IEEE.
- [10] Hiramath CK, Deshpande GC. Fake news detection using deep learning techniques. In *2019 1st International Conference on Advances in Information Technology (ICAIT) 2019 Jul 25* (pp. 411-415). IEEE.

- [11] Shu K, Sliva A, Wang S, Tang J, Liu H. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*. 2017 Sep 1;19(1):22-36.
- [12] Pérez-Rosas V, Kleinberg B, Lefevre A, Mihalcea R. Automatic detection of fake news. *arXiv preprint arXiv:1708.07104*. 2017 Aug 23.
- [13] Reis JC, Correia A, Murai F, Veloso A, Benevenuto F. Supervised learning for fake news detection. *IEEE Intelligent Systems*. 2019 May 8;34(2):76-81.