Effective Fake News Detection Using Logistic Regression: A High-Accuracy Approach

Abstract

The proliferation of fake news on digital platforms has become a significant challenge, necessitating the development of automated systems to detect and mitigate its impact. This study presents a Logistic Regression-based approach for fake news detection, leveraging a dataset of news articles. The model achieved a high accuracy of 98.95% on the training data and 97.76% on the test data, demonstrating its effectiveness in distinguishing between genuine and fake news. The simplicity and interpretability of Logistic Regression make it a strong candidate for real-world applications where transparency and quick decision-making are crucial.

1 Introduction

The digital age has led to an unprecedented increase in the dissemination of information. However, this has also given rise to the spread of misinformation or fake news, which can have severe societal impacts. The ability to automatically identify and filter out fake news is critical to maintaining the integrity of information. This paper explores the use of Logistic Regression, a statistical method for binary classification, to address the problem of fake news detection. By analyzing various features of news articles, the model predicts whether an article is likely to be fake or genuine.

2 Related Works

The issue of fake news detection has garnered significant attention in recent years. Researchers have employed various machine learning techniques, including Support Vector Machines (SVM), Naive Bayes, and deep learning models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), to tackle this problem. While deep learning models often yield high accuracy, they require large amounts of data and computational resources. On the other hand, traditional machine learning models like Logistic Regression, although simpler, have proven to be effective in scenarios where interpretability and lower computational costs are desired.

3 Algorithm

Logistic Regression is a statistical method that models the probability of a binary outcome based on one or more predictor variables. The model assumes a linear relationship between the input features and the log-odds of the outcome. The logistic function, or sigmoid function, is used to map the predicted values to probabilities between 0 and 1.

Mathematically, the logistic function is defined as:

$$P(y=1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n)}}$$

Where:

P(y=1|X) is the probability that the target variable y y equals 1 given the input features X.

β_0, β_1, $\dots\beta_n$ are the coefficients of the model, learned during training.

The model is trained by maximizing the likelihood function, which measures how well the model fits the data. The cost function, known as the log-loss or binary cross-entropy, is minimized during the training process.

4 Methodology

The methodology involves several key steps:

Data Collection: The dataset used in this study consists of news articles labeled as fake or genuine. The data is split into training and test sets, with 80% used for training and 20% for testing.

Data Preprocessing: The articles undergo preprocessing, including the removal of stop words, stemming, and vectorization. The vectorization process converts the text into numerical features using methods such as TF-IDF (Term Frequency-Inverse Document Frequency).

Model Training: Logistic Regression is employed to train the model on the processed dataset. The training process involves optimizing the model coefficients to minimize the log-loss function.

Model Evaluation: The trained model is evaluated on the test set to assess its performance. Accuracy, precision, recall, and F1-score are calculated to provide a comprehensive evaluation of the model.

Experimental Work

The experiments were conducted using a standard dataset of news articles. The dataset was preprocessed to remove noise and irrelevant features. Logistic Regression was implemented using the scikit-learn library in Python. The model was trained on 80% of the data and tested on the remaining 20%. The results showed that the model performed exceptionally well, with an accuracy of 98.95% on the training data and 97.76% on the test data.

5 Results

The Logistic Regression model demonstrated high accuracy in predicting fake news. The accuracy on the training set was 98.95%, indicating that the model learned the underlying patterns in the data effectively. The test accuracy was slightly lower at 97.76%, suggesting that the model generalizes well to unseen data. These results validate the effectiveness of Logistic Regression as a tool for fake news detection, especially in scenarios where interpretability and quick decision-making are required.

6 Conclusion

This study presents a Logistic Regression-based approach for fake news detection, achieving high accuracy on both training and test datasets. The results indicate that Logistic Regression is a viable option for real-world fake news detection systems, offering a balance between simplicity, interpretability, and performance. Future work could explore the integration of more complex features and hybrid models to further enhance detection accuracy.

7 References

- Graves, S. Fernández, and J. Schmidhuber, "Bidirectional LSTM networks for improved phoneme classification and recognition," in Artificial Neural Networks: Formal Models and Their Applications – ICANN 2005, 2005, pp. 799–804.

- R. Johnson and T. Zhang, "Supervised and Semi-supervised Text Categorization using LSTM for Region Embeddings," in Proceedings of the 34th International Conference on Machine Learning - Volume 70, Sydney, Australia, 2017, pp. 526–534.
- J. B. Polson and N. S. Scott, "Data Science and Fake News Detection," Applied Stochastic Models in Business and Industry, vol. 35, no. 1, pp. 77–86, 2019.
- P. Ferrara, H. Harman, J. J. Jiang, and H. Zheng, "Linguistic Features for Fake News Detection," International Journal of Advanced Computer Science and Applications, vol. 10, no. 8, pp. 6–15, 2019.