Fake News Detection using Logistic Regression

Tumin Sharma

Abstract

This project explores the use of machine learning algorithms for detecting fake news. The project involves preprocessing text data, applying various models, and evaluating their performance using standard metrics.

Contents

1 Introduction

1.1 Background

Fake news has become a significant challenge in today's digital age, with misinformation spreading rapidly across social media and other platforms. Identifying and mitigating fake news is essential to ensuring the reliability of information available to the public.

1.2 Motivation

With the growing influence of news on public opinion, detecting fake news using machine learning can play a crucial role in combating misinformation. This project aims to contribute to this domain by exploring effective machine learning models for fake news classification.

1.3 Objectives

- To preprocess a dataset containing real and fake news articles.
- To implement machine learning models for classifying news articles as real or fake.
- To evaluate the performance of the models using appropriate metrics.

2 Literature Review

Summarize existing research related to your topic. Discuss different methodologies previously used to tackle similar problems.

3 Dataset Description

3.1 Source

The data and the inspiration on how to use the data was taken from Kapil Singh Negi

3.2 Features

The dataset includes the following features:

- **Title**: The headline of the news article.
- **Text**: The main content of the news article.
- **Subject**: The category of the news article (e.g., politics, technology).
- **Date**: The publication date of the article.

3.3 Target Variable

The target variable is **Label**, which indicates whether the news article is:

- **1**: Fake
- **0**: Real

4 Data Preprocessing

The data was straightforward to use with. Except I needed to vectorized the words into computable matrices. And before I did that there were lot of unnecessary context free data values like stop words and symbols. So using RegEx I got read of those and uncapitalized every word for more continuity.

Also intuitively the **Date** does not have any direct correlation with the news being fake or not. So I got rid of that column totally. A lot can be at least asssumed from a **Title** alone so I kept it and the **Text** is our main input feature. Lastly for statistical theory that and particular **Subject** has more fake news or not I kept that feature too.

5 Methodology

5.1 Algorithms Used

Since there are only 2 simple labels, I used **Logistic Regression**, the best algorithm for simple binary classification.

6 Implementation

6.1 Tools and Libraries

List the tools and libraries used for implementing the solutions are

- **Python**: The primary programming language.
- **Scikit-learn**: For machine learning models and evaluation metrics.
- **Pandas**: For data manipulation.
- **Matplotlib**: For data visualization.
- **Seaborn**: For data visualization.

6.2 Parameters

The following hyperparameters were tuned for the Logistic Regression model:

- **C**: Regularization strength.
- **penalty**: 'l2' for regularization.

6.3 Training Process

The dataset was split into training (80%) and testing (20%) sets. The models were trained using 5-fold cross-validation to ensure robustness.

7 Results

The performance of the tuned Logistic Regression model is summarized below:

7.1 Classification Report

```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      4286
           1       1.00      1.00      1.00      4652

    accuracy                           1.00      8938
   macro avg       1.00      1.00      1.00      8938
weighted avg       1.00      1.00      1.00      8938
```

7.2 Confusion Matrix

The confusion matrix for the tuned model is:

```
[[4286    0]
 [   0 4652]]
```

8 Discussion

The model achieved an accuracy of 1.00 on the test set, indicating perfect classification. The confusion matrix confirms that there were no false positives or false negatives. This could suggest that the dataset is well-defined and the model is overfitted, as a real-world scenario may not always yield such perfect results.

9 Conclusion

In this project, I successfully built and evaluated machine learning models for fake news detection. The Logistic Regression model performed exceptionally well, achieving 100% accuracy. Future work could involve experimenting with more complex models, like neural networks, and evaluating the model on new datasets.

10 References

- **Data:** Kapil Signh Negi's Dataset.
- **Documentations:** Matplotlib, Pandas, Scikitlean.
- **Additional guidance:** ChatGPT, Kaggle Learn, Sentdex.