Detection of fake news using machine learning

September 2024

Introduction

In today's digital world, the spread of fake news has become a growing concern, often leading to misinformation and confusion. Distinguishing between fact and fiction has become more difficult, especially with AI advancements that generate highly realistic but misleading content. This accelerates the spread of fake news, influencing people's opinions, emotions, and decisions.

With this project our aim is to build a machine learning model that can detect fake news articles by applying Natural Language Processing (NLP) techniques where we aim to classify news as either true or false based on its content. Our goal is to create a system that helps people trust the information they read and reduce the impact of misinformation. By training our model on a large dataset of labeled news articles, we hope to develop a reliable tool that can differentiate between fake and real news.

This report will cover the problem formulation, including the dataset used and key features of news articles. Then, we will discuss the methods applied, including preprocessing steps and the machine learning models used for classification. Finally, the performance of the models will be evaluated, and the results will be compared to determine the most effective approach for detecting fake news.

Problem Formulation

This project addresses fake news detection as a binary classification problem, where the goal is to predict whether a news article is real or fake. The dataset contains labeled news articles, each instance consisting of four main attributes: the article's **title**, **text**, **subject**, and **date** of publication.

Data Points:
Each data point corresponds to a unique news article. The **title** provides a brief description of the content, the **text** contains the full body of the article, the **subject** indicates the general category (such as article or news), and the **date** refers to when the article was published. These textual data points are essential for building a model that captures linguistic patterns associated with fake news.

Features:
The features of this dataset are primarily based on the content of the articles. Textual features like word frequency, sentence length, and linguistic markers will be extracted using Natural Language Processing (NLP) techniques. Additionally, categorical data such as the subject of

the article (e.g., politics or world news) will provide contextual features that help the model differentiate between real and fake news. The text of each article will be preprocessed, including cleaning, tokenization, and stop-word removal, to prepare the data for model training.

Labels:
The labels for each data point are binary, with **1** representing fake news and **0** representing real news. These labels provide the ground truth for training the machine learning model, allowing it to learn from both fake and real examples to make predictions.

Source of the Dataset:
The dataset is sourced from a publicly available repository on Kaggle, specifically designed for fake news detection. The dataset has been preprocessed to include labeled examples of news articles, helping to ensure the reliability of the machine learning model and its ability to generalize to unseen data.

Methods

Dataset and Preprocessing:

The dataset consists of 17,903 fake and 20,826 true news articles. Each article contains four attributes: **title**, **text**, **subject**, and publication **date**. The text data is preprocessed by removing special characters, stop words, and converting the text to lowercase. Tokenization is performed to convert the text into word units, and term frequency-inverse document frequency (TF-IDF) is applied to represent the text numerically for the model.

Feature Selection:

Key features include word frequency, sentence structure, and contextual markers like article subject. We selected these features as they provide insight into the writing patterns and language use that can distinguish real news from fake news. We chose TF-IDF to prioritize important terms within each article, ensuring the model focuses on distinguishing features rather than common words.

Model choice:

Since both logistic regression and decision trees work well for binary classification applications, we evaluated both of them. The decision trees are helpful in identifying non-linear relationships in the data, but the interpretability and simplicity of logistic regression made it the preferred method. Because they may represent intricate linkages, neural networks were also taken into consideration for investigating more complex patterns in the dataset.

Loss Function:

We used binary cross-entropy as the loss function for our binary classification task. This loss function evaluates the divergence between the actual labels (false or real) and the anticipated

probabilities, making it ideal for assessing probabilistic outputs. It guarantees that as training goes on, the model gains greater assurance in its ability to distinguish between false and authentic news.

Model Validation

15 percent of the dataset was set aside for testing, 15 percent for validation, and 70 percent for training. With this distribution, the model may validate its performance and fine-tune its hyperparameters using the validation set, all while learning from a significant portion of the data. To make sure the model applies well to previously unseen data, the test set will be put aside for the final assessment. In order to evaluate model performance on various data splits and prevent overfitting even more, we additionally employed cross-validation.

Results

In this project, Logistic Regression and Decision Tree models were evaluated for classifying fake and real news. Logistic Regression achieved an accuracy of 98.5%, while the Decision Tree performed slightly better with an accuracy of 99.5%. Both models demonstrated strong precision and recall, but the Decision Tree was chosen as the final model due to its superior validation accuracy and fewer misclassifications. On the test set, the Decision Tree maintained its high performance with a test accuracy of 99.5%, confirming its reliability.

Conclusion

The project successfully applied machine learning to differentiate between fake and real news, with the Decision Tree model emerging as the best performer. The high accuracy across both validation and test sets indicates the problem was addressed effectively. While the model performed exceptionally well, there is potential for future improvements, particularly in managing overfitting and handling more complex cases. Advanced techniques could enhance the model's robustness, but the current solution provides a reliable method for this classification task. Overall, the results suggest that the model can accurately and consistently classify news articles with minimal errors.

REFERENCES

1. Paramartha Sengupta 2023, Fake News Detection, Available at: https://www.kaggle.com/code/paramarthasengupta/fake-news-detector-eda-prediction-99

2. A. Jung, "Machine Learning: The Basics," Springer, Singapore, 2022.

3. CS-C3240 - Machine Learning Lecture Slides and videos

4. Logistic regression

   https://scikit-learn.org/1.5/modules/generated/sklearn.linear_model.LogisticRegression.html

5. Decision tree classifier
   https://scikit-learn.org/stable/modules/tree.html

6. Bubble chart
   https://www.geeksforgeeks.org/bubble-chart-using-plotly-in-python/

7. Wordcloud
   https://medium.com/@m3redithw/wordclouds-with-python-c287887acc8b

8. Decision tree
   https://scikit-learn.org/stable/modules/tree.html