MODEL TRAINING

Project Report

Project title: Fake News Detection

SE-Department

Submitted By:

Rafay Noor

Su92-bsdsm-f23-024

BS Data Science (3A)

3rd Semester

Submitted To:

Sir Rasikh

Fake News Detection Using Machine Learning

Abstract

In the digital age, fake news has become an issue, spreading misinformation and creating societal discord. This project aims to develop a machine learning model capable of detecting fake news by analyzing textual content. Using a dataset comprising real and fake news articles, several models, including Logistic Regression, Decision Tree, and Random Forest, were trained and evaluated. The results demonstrate the potential of these models to automate the process of fake news detection effectively.

1. Introduction

1.1 Background

Fake news, defined as fabricated information presented as legitimate news, has emerged as a significant challenge in modern times. Its rapid dissemination across social media platforms has amplified its impact, making the identification and mitigation of fake news a priority.

1.2 Problem

It's hard and time-consuming for humans to spot fake news. Using computers and machine learning can make this process faster and more accurate.

1.3 Objectives

- Process and clean the data to prepare it for analysis.
- Train machine learning models to classify news as real or fake.
- Compare the models' performances.

2. Dataset

2.1 Description

The dataset comprises two files:

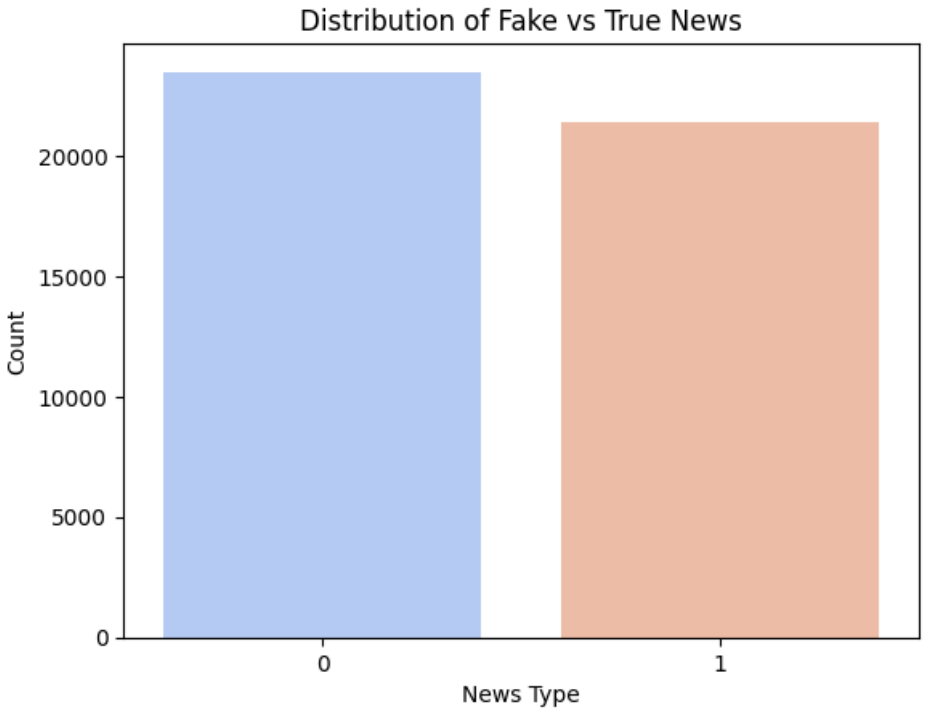
- **Fake.csv**: Contains fake news articles.
- **True.csv**: Contains true news articles.

2.2 Preprocessing

1. Text normalization:
    - Converted text to lowercase.
    - Removed URLs, special characters, and extra spaces.
2. Vectorization:
    - Used TF-IDF to convert textual data into numerical features for modeling.

2.3 Dataset Distribution

To understand the dataset composition, a count plot was created to show the distribution of fake and true news articles:



This plot demonstrates that the dataset is relatively balanced between fake and true news articles, which is important for training unbiased machine learning models.

3. Methodology

3.1 Machine Learning Models

The following models were trained and evaluated:

1. **Logistic Regression**: A linear model used for binary classification.
2. **Decision Tree Classifier**: A tree-based model that splits data based on features.
3. **Random Forest Classifier**: An ensemble of decision trees to improve accuracy and reduce overfitting.

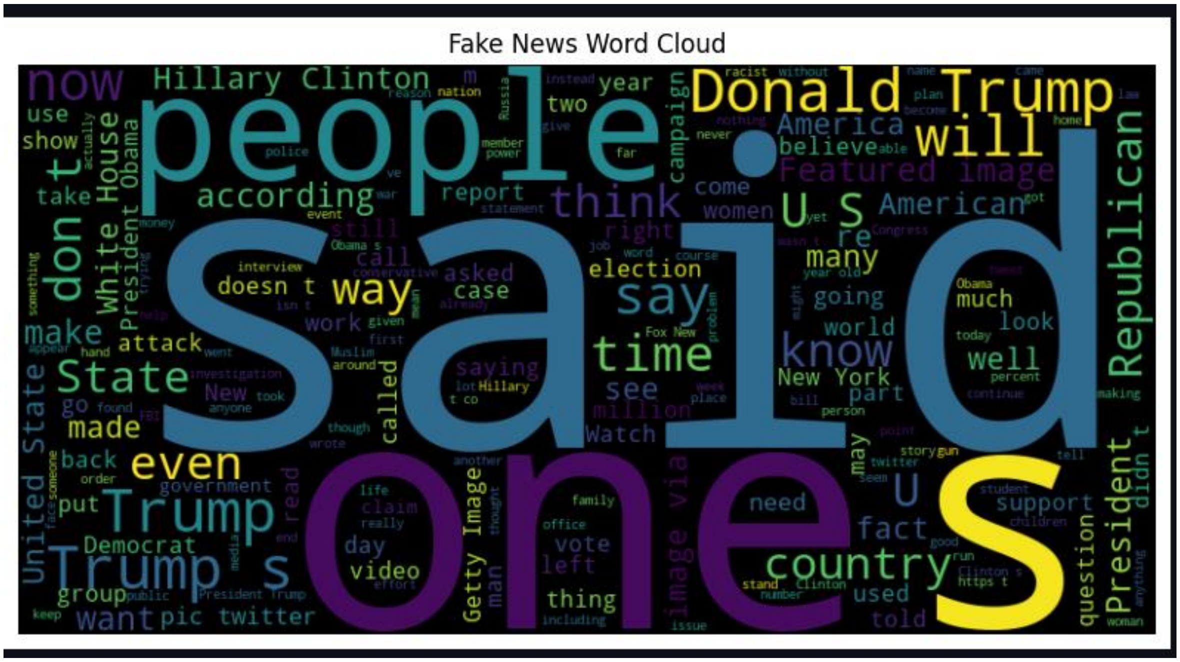
3.2 Workflow

1. **Data Splitting**: Split the data into training (70%) and testing (30%) sets.
2. **Feature Extraction**: Applied TF-IDF vectorization to transform text data into numerical format.
3. **Model Training**: Trained the models on the processed training data.
4. **Evaluation**: Assessed the models using metrics like accuracy, precision, recall, and F1-score.

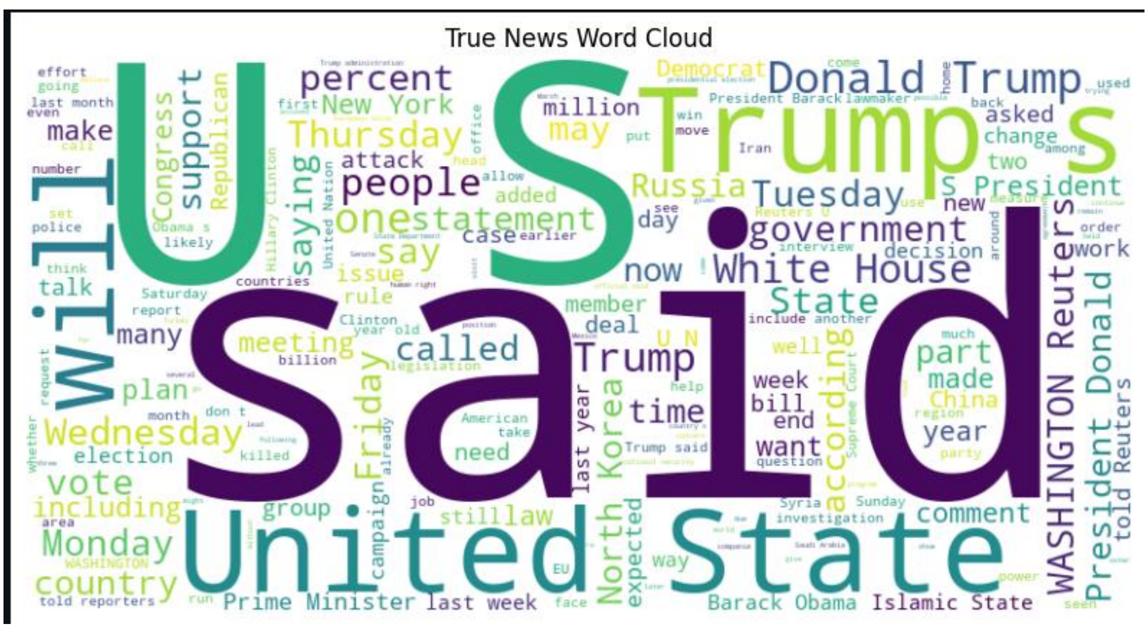
3.3 Word Cloud Visualization

Word clouds were generated to highlight the most frequently occurring words in the datasets. These visualizations provided insight into the common themes and patterns in fake and true news articles.

FAKE:

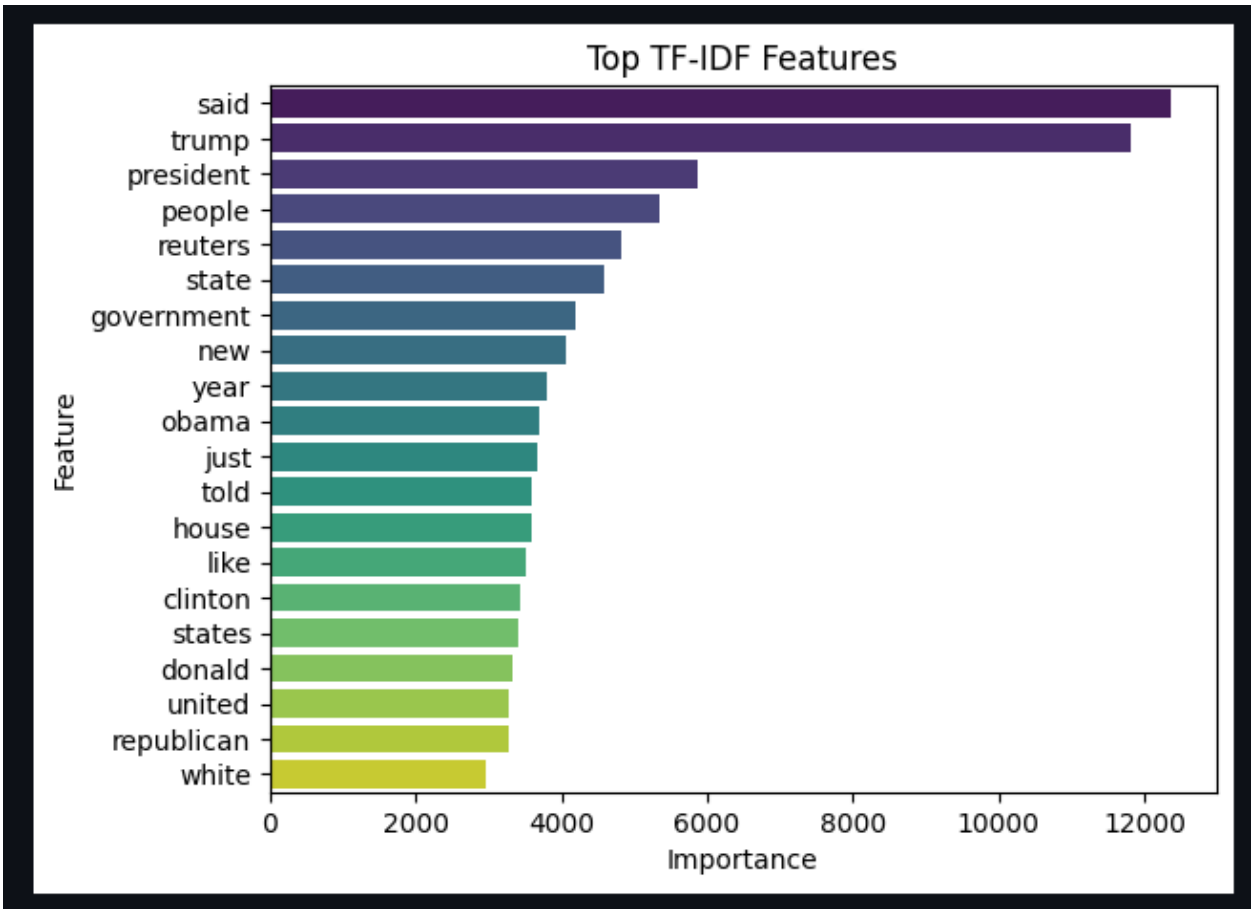

Fake News Word Cloud
said
one
people
Donald Trump
will
Hillary Clinton
Republican
Featured image
American
know
time
say
State
Trump
Trump s
country
way
according
think
even
White House
President Obama
United State
President
Getty Image
image via
New York
election
many
Democrat
video

TRUE:


True News Word Cloud
said
U
S
Trump s
Donald Trump
United State
will
percent
New York
Thursday
people
onestatement
government
White House
Tuesday
Wednesday
Monday
Friday
WASHINGTON Reuters
President Donald
Russia
State
North Korea
Prime Minister
last week
Barack Obama
Islamic State
Congress
support
Republican

3.2 TF-IDF Feature Analysis

The TF-IDF (Term Frequency-Inverse Document Frequency) method was used to identify the most important words in the dataset. The top 20 features with the highest importance scores were visualized using a bar chart, helping to identify key distinguishing terms.



4. Results

The performance of the models was evaluated on the test dataset. Below are the key metrics for each model:

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.99	0.99	0.99	0.99
Decision Tree	1.00	1.00	0.99	1.00
Random Forest	0.99	0.99	0.99	0.99

Key Observations:

1. **Logistic Regression** achieved an accuracy of **99%**, showing its effectiveness as a baseline model.
2. **Decision Tree** provided the best overall performance, achieving an accuracy of **100%** with near-perfect precision, recall, and F1-score.
3. **Random Forest** matched Logistic Regression in performance, with an accuracy of **99%** and balanced precision, recall, and F1-score.

These results indicate that all three models perform exceptionally well on the dataset, with the Decision Tree achieving the highest overall accuracy and F1-score.

5. Discussion

5.1 Observations

1. The **word clouds** provided a quick overview of common terms in fake and true news articles, showing distinct patterns in word usage.
2. The **TF-IDF bar chart** highlighted the most important words in the dataset, providing valuable insights into the differences between fake and true news.
3. Among the models, **Random Forest** achieved the highest accuracy and F1-score, demonstrating its ability to handle the complexity of the dataset.
4. **Logistic Regression** showed competitive performance and was computationally efficient.

5.2Challenges

1. Balancing the dataset to avoid bias in model training.
2. Selecting the optimal number of features for the TF-IDF vectorizer to improve model accuracy.

6. Conclusion

This project successfully demonstrated the use of machine learning to detect fake news. Visualizations like word clouds and TF-IDF feature analysis helped understand the data better. The Random Forest model emerged as the most effective in classifying news articles. Future work could include:

1. Exploring advanced techniques like deep learning.
2. Using larger and more diverse datasets to improve generalization.

7. References

1. Scikit-learn Documentation: https://scikit-learn.org.
2. Python Libraries: Pandas, NumPy, Matplotlib, Seaborn.