

Report

Fake News Detection

Jainisha Choksi

18th January, 2024

Introduction

In today's world, we get a lot of our information from the internet, especially from social media. But sometimes, not everything we read or see is true. There's a problem called "fake news", which means spreading false or misleading information on purpose.

Imagine if we could figure out which news is fake and which is real. That's where fake news detection comes in. It's like having superhero tools to spot the bad guys- except in this case, the bad guys are the fake stories.

In this report, we're going to explore how smart technology and clever strategies are helping us catch fake news. By understanding these tricks, we can make sure we're not fooled by false information and can trust what we read and see online a bit more. Let's dive into the world of fake news detection to see how it works and why it's so important.

Dataset

There 2 dataset: True and False as shown below:

True Dataset

	title	text	subject	date
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017

False Dataset

	title	text	subject	date
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn't wish all Americans ...	News	December 31, 2017
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017

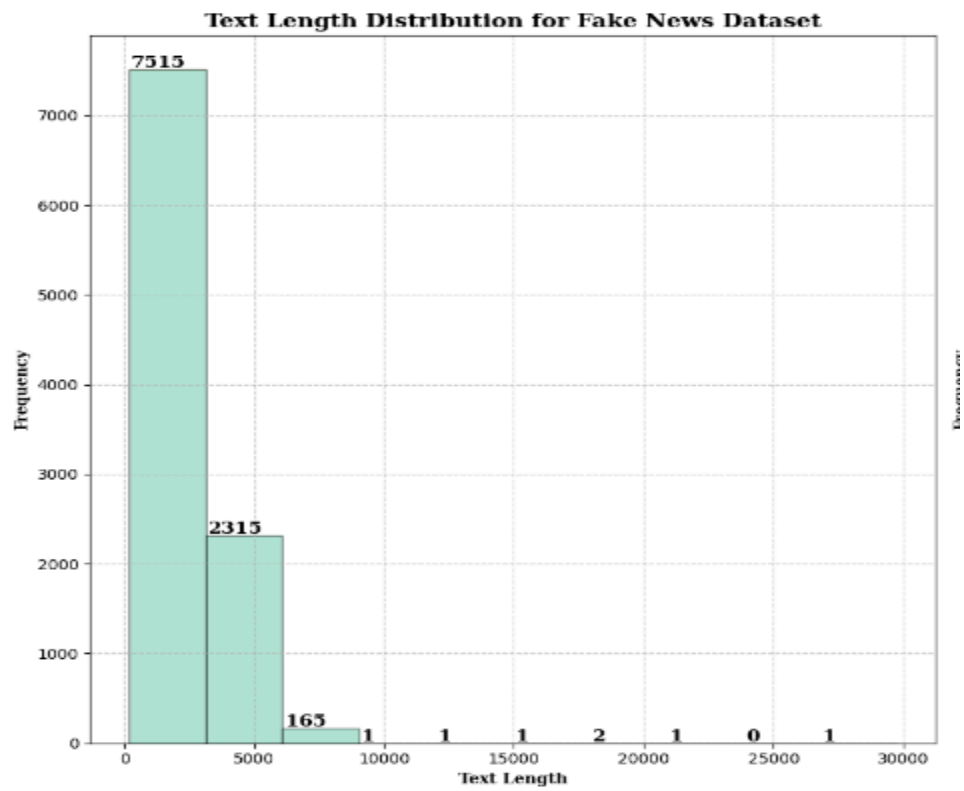
As we have two different datasets, we have to merge it to create a dataframe with labels as true -1 and false - 0.

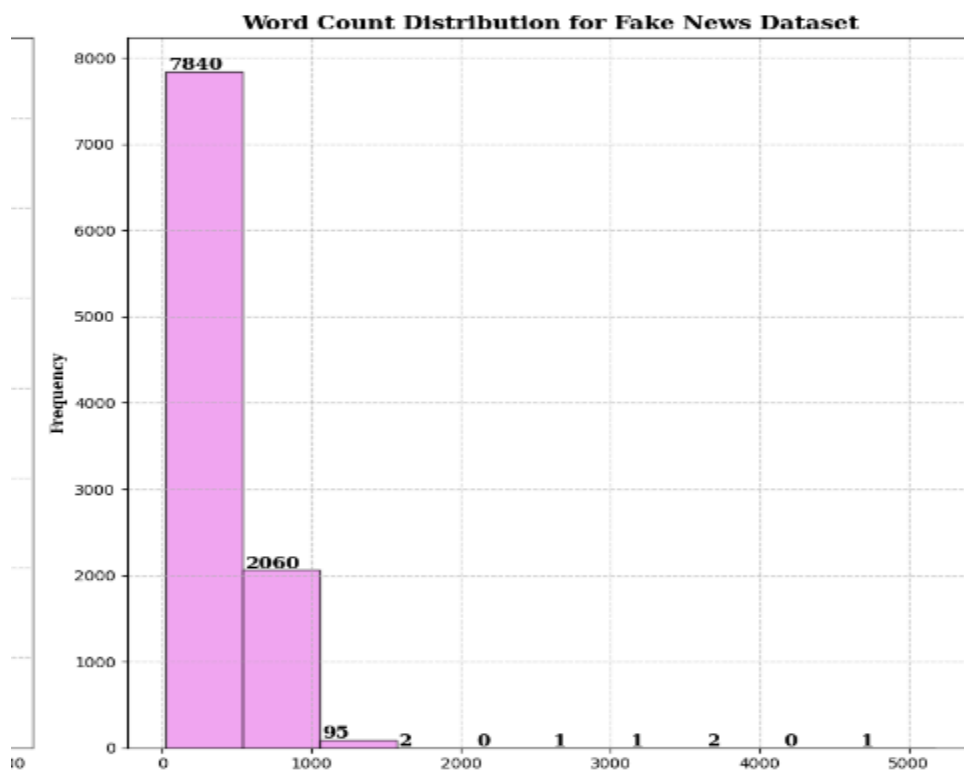
	index	title	text	subject	date	label
0	0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017	1
1	1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017	1
2	2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017	1
3	3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017	1
4	4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017	1

Exploratory Data Analysis

In the Exploratory Data Analysis phase, we first examined the distribution of text length and word count in the dataset. This step allowed us to understand the basic characteristics of the text corpus we are working with. Subsequently, we visualized the importance of words in the corpus by creating a word cloud.


The text length distribution and word count analysis provide insights into the overall structure and variability of the text data. This information is crucial for understanding the nature of the content we are dealing with.





The word cloud, on the other hand, is a visually impactful representation of the most frequently occurring words in the dataset. Larger words in the cloud indicate higher frequency, highlighting the key terms that might carry significance in the context of fake news detection.





This combination of analysis allows for a comprehensive view of the textual data, aiding in the identification of patterns and potential features that can be leveraged for fake news detection. The visual appeal of the word cloud also makes it an effective tool for communicating the weightage of words in the text corpus.

Data Preprocessing

In the data preprocessing, we implemented a series of essential Natural Language Processing tasks to enhance the quality of the text data for effective fake news detection. Here is a brief summary of the preprocessing steps undertaken:

1. Text Cleaning:

Removal of HTML tags, if any, to ensure clean text.

Handling special characters and punctuation removal to maintain consistency in the data.

2. Tokenization:

Breaking down the text into individual words or tokens. This step facilitates further analysis on a word-level basis.


3. Lowercasing:

Converting all text to lowercase. This ensures uniformity and prevents the model from treating words with different cases as distinct.

4. Stopword Removal:

Elimination of common words that do not contribute much to the meaning of the text. This step reduces noise in the data.

5. Stemming:



Reducing words to their root/base form. This step helps in standardizing variations of words, improving feature extraction and reducing dimensionality.

6. TF-IDF Vectorization:

Transforming the preprocessed text into numerical vectors using TermFrequency-Inverse Document Frequency vectorization.

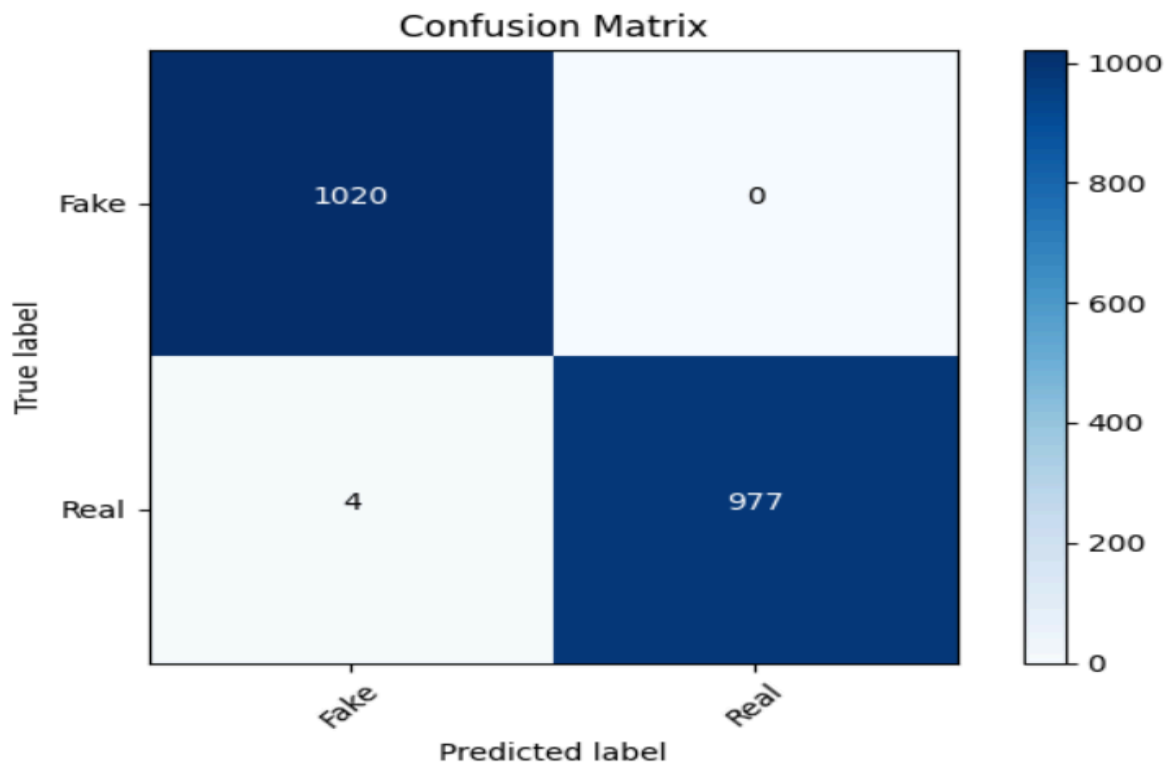
TF-IDF captures the importance of words in a document relative to the entire corpus, assigning higher weights to terms that are more discriminative.

By applying these preprocessing tasks and employing a TF-IDF vectorization, we have effectively transformed the raw textual data into a format that is suitable for machine learning models. This processed data serves as the input for subsequent stages in our fake news detection pipeline, contributing to the creation of a robust and efficient model.

Machine Learning Models

We used a machine learning model called the Passive Aggressive Classifier. After training and evaluating the model on our dataset, we achieved an impressive accuracy of 99.8%.

This means that the model successfully classified news articles as either fake or genuine with an accuracy rate of 99.8%. In simple terms out of every 100 articles, the model correctly predicted the authenticity of 99 of them. Such a high accuracy indicates the effectiveness of the Passive Aggressive Classifier in distinguishing between fake and genuine news based on the features and patterns it learned during training.



Conclusion

Detecting and addressing fake news demands a holistic approach. By leveraging tech tools, promoting media literacy, and fostering partnerships, we can build a robust defense against misinformation.

Future Scope

Use of Advance Machine Learning Algorithm

Natural Language Processing Enhancements

Cross-Platform Collaboration

Ethical AI Practices

Online Machine Learning



Public Awareness Campaigns