Fake News Detection using NLP and ML

Pari Chhoriya, Komal Bahadurge

May 22, 2023

Contents

1 Introduction

The fake news has been rapidly increasing in numbers. It is not a new problem but recently it has been on a great rise. According to Wikipedia, Fake news is false or misleading information presented as news. Detecting the fake news has been a challenging and a complex task. It is observed that humans have a tendency to believe the misleading information which makes the spreading of fake news even easier.

Fake news is dangerous as it can deceive people easily and create a state of confusion among a community. This can further affect the society badly .The spread of fake news creates rumors circulating around and the victims could be badly impacted. Fake news might be created by people or groups who are acting in their own interests or those of third parties. The creation of misinformation is usually motivated by personal, political, or economic agendas. Therefore it is very important to detect if the news is fake or not.

Since a lot of time is spent by users on social media and people prefer online means of information it has become difficult to know about the authenticity of the news. People acquire most of the information by these means as it is free and can be accessed from anywhere irrespective of place and time. Since this data can be put out by anyone there is lack of accountability in it which makes it less trustable unlike the traditional methods of gaining information like newspaper or some trusted source.

In this paper, we deal with such fake news detection issue. We have used the techniques of NLP and ML to build the model .We have also compared text vectorization methods and obtained the one which gives a better output.

1.1 Objective

The objective of this project is to examine the problems and possible significances related with the spread of fake news. We will appl machine learning algorithms to train the data and test it to find which news is the real news or which one is the fake news. By using the artificial intelligence and the machine learning, the problem can be solved as we will be able to mine the patterns from the data to maximize well defined objectives.

2 Methodology

2.1 Datasets

The datasets in this project are from kaggel.com. The data is in CSV format and each column of the dataset represents one movie information.

There are now 5 features that are used in the project. They are **id, title, author, text,** and **label**.

2.2 Libraries

In this project we are using different packages and to load and read the data set we are using pandas. For training and testing the data, we use **Scikit Learn(sklearn)** library and supervised learning model **Logistic Regression**. For data preprocessing, we use **nltk** library and **TF-IDF Vectorizer** method. Next step is, by using this data, getting the visual reports, which we will get by using the **matplotlib** and **seaborn**.

2.2.1 Pandas

Pandas provides highly efficient and easy-to-use data structures, such as DataFrame and Series, along with a wide range of functions for data cleaning, transformation, exploration, and visualization. We can use "pandas" for following:

1. Data Handling
2. Data Cleaning and Preparation
3. Data Exploration and Analysis
4. Data Visualization
5. Data Integration
6. Data Transformation and Feature Engineering
7. Data Transformation and Feature Engineering
8. Data Export and Integration

2.2.2 Re

Regular expressions (often abbreviated as "regex" or "RE") are a powerful tool for pattern matching and text manipulation. In Python, the "re" module provides support for working with regular expressions. We can use "re" for following:

1. Pattern Matching
2. String manipulation

3. Data cleaning and validation
4. Data information
5. Data analysis

2.2.3 NLTK

Natural Language Processing (NLP) is the process of getting a computer to understand natural language. The data can be in the form of a text document, image, audio, or video. The Natural Language Toolkit (NLTK) is an open-source toolkit for natural language processing. This toolkit is one of the most powerful NLP libraries which contains packages to make machines understand human languages and respond in an appropriate manner. Using NLTK we can perform variety of tasks such as data cleaning, tokenizing, visualization, and vectorization that will help us in classifying our text.

2.2.4 Scikit-Learn (sklearn)

Scikit-learn, often referred to as sklearn, is a popular open-source machine learning library in Python. It is built on top of other scientific computing libraries such as NumPy, SciPy, and matplotlib. Following are advantages of sklearn:

1. Ease of Use
2. Comprehensive Machine Learning Algorithms
3. Efficient Data Preprocessing
4. Integration with Other Libraries
5. Large and Active Community
6. Integration with Ecosystem

2.2.5 Matplotlib and Seaborn

Matplotlib and Seaborn are two popular data visualization libraries in Python. They are widely used for creating visualizations and plots to analyze and communicate data. Following are advantages of matplotlib:

1. Flexible Customizable Plots
2. PublicationQuality Visualizations

3. Wide Compatibility and Integration
4. Extensive Community and Documentation

Following are advantages of seaborn:

1. Statistical Data Visualization
2. Beautiful Default Styles
3. Integration with Pandas
4. Advanced Statistical Visualizations

2.3 Data Preprocessing

Data preprocessing is a data mining technique that involves transforming raw data into understandable form. In natural language processing, text preprocessing is the practice of cleaning and preparing text data. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Data preprocessing methods such as tokenization, stemming, stop word removal and lowercasing.

2.4 Algorithms

2.4.1 Porter stemmer

Stemming is a natural language processing technique that is used to reduce words to their base form, also known as the root form. The process of stemming is used to normalize text and make it easier to process. It is an important step in text pre-processing, and it is commonly used in information retrieval and text mining applications. A stemming algorithm reduces the words “chocolates”, “chocolatey”, “choco” to the root word, “chocolate” and “retrieval”, “retrieved”, “retrieves” reduce to the stem “retrieve”. There are several different algorithms for stemming.

The **Porter stemmer** is the most widely used algorithm for stemming. It is based on the idea that the suffixes in the English language are made up of a combination of smaller and simpler suffixes. This stemmer is known for its speed and simplicity. The main applications of Porter Stemmer include data mining and Information retrieval.

2.4.2 TF-IDF

TF-IDF stands for term frequency-inverse document frequency. It highlights a specific issue which might not be too frequent but holds great importance. The TF–IDF value increases proportionally to the number of times a word appears in the document and decreases with the number of documents in the corpus that contain the word. TF-IDF(Term Frequency/Inverse Document Frequency) is one of the most popular IR(Information Retrieval) techniques to analyze how important a word is in a document. TF-IDF is the product of TF and IDF. A high TF-IDF score is obtained by a term that has a high frequency in a document, and low document frequency in the corpus. For a word that appears in almost all documents the IDF value approaches 0, making the tfidf also come closer to 0. TF-IDF value is high when both IDF and TF values are high i.e the word is rare in the whole document but frequent in a document.

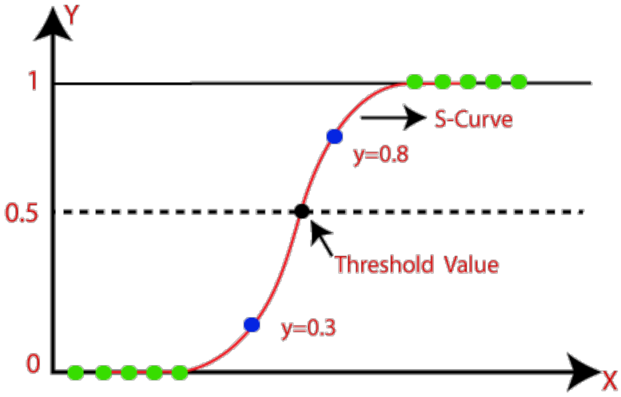
2.5 Models

We used 2 models in this project: **Logistic Regression** and **Decision Tree Classifier**.

2.5.1 Logistic Regression

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. It predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1. It is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems. In Logistic regression, instead of fitting a regression line, we fit an ”S” shaped logistic function, which predicts two maximum values (0 or 1). The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc. Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. The below image is showing the logistic function:

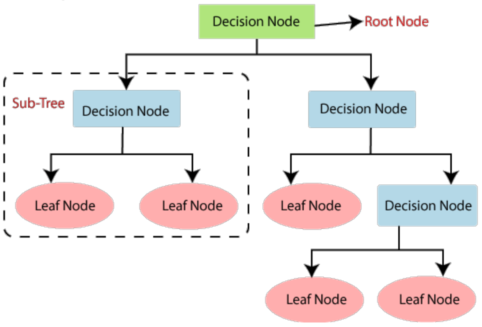
Figure 1: Sample Logistic Regression



2.5.2 Decision Tree Classifier

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees. Below diagram explains the general structure of a decision tree:

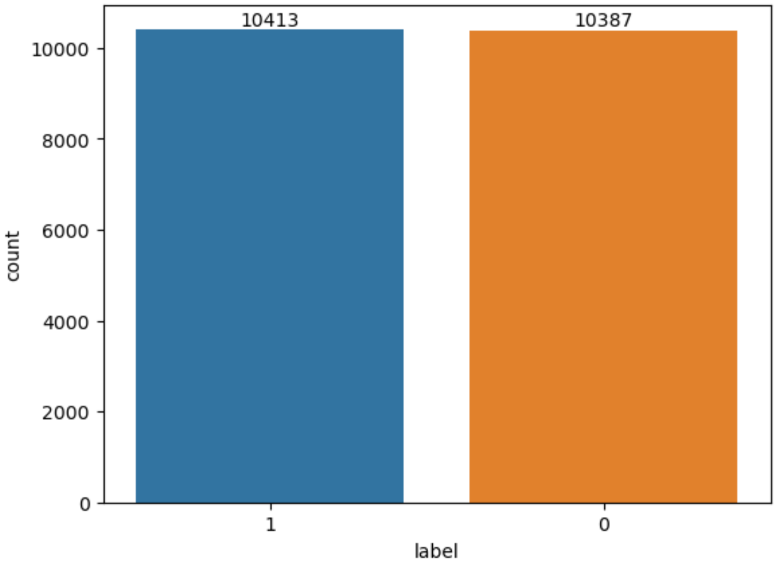
Figure 2: Sample Decision Tree



3 Results

'Figure 3' shows the number of fake news and real news in the dataset.

Figure 3: Number of fake news and real news in the dataset



We have used word clouds to check which are the words which appear frequently in the fake and real news. **'Figure 4'** shows the Word Cloud for real news and **'Figure 5'** shows the Word Cloud for fake news.

Figure 4: Word Cloud for Real news

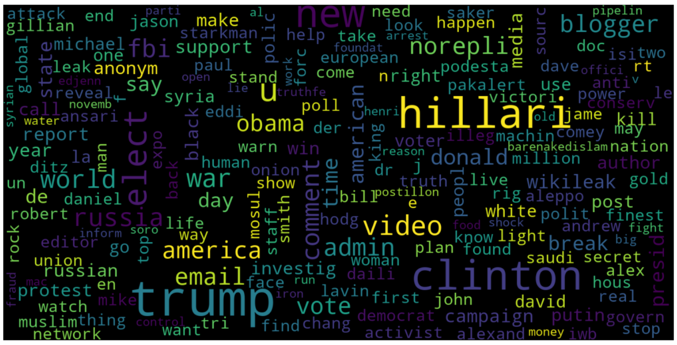
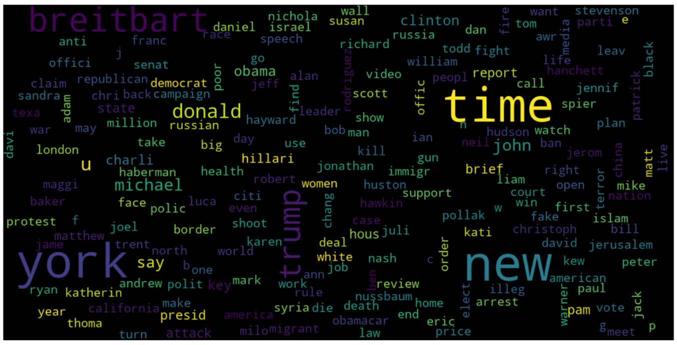


Figure 5: Word Cloud for Fake news



After the models were trained we calculated the performance metrics accuracy and precision. ‘Table.I’ shows the performance metrics of the models.

Table 1: Performance metrics of the models

Metric	**Logistic Regression**	**Decision Tree Classification**
Accuracy for train dataset	98%	100%
Accuracy for test dataset	98%	99%
Precision for train dataset	98%	100%
Precision for test dataset	96%	99%

4 Conclusion

Our findings indicate that our fake detection system achieved a high level of accuracy in distinguishing between genuine and fake content. We trained the model on a diverse dataset consisting of labeled examples of both real and fake content, enabling it to learn patterns and characteristics that differentiate between the two. In conclusion, our model successfully developed a robust fake detection model capable of accurately identifying fake content across various domains. This model can serve as a valuable tool for combating misinformation and aiding in the promotion of credible information online.