Fake News Detection Using Machine Learning

Naive Bayes Algorithm

Rushikesh Joshi

29/08/2020

What Is Fake News?

- Fake News is a news designed to spread hoaxes, propaganda and misinformation
- Fake News is different from satirical sites like “The Onion”
- Fake News Stories Usually Spread through social media like Facebook, Twitter etc.
- Often,Fake News will Mimic real headlines and twist the stories

Defining Fake News

There are articals that :-

- Are blatantly False
- Provide Truthful events with False Interpretation
- Pseudoscientific
- Opinion piece disguised as news
- Are satirical
- Some Articles that were marked as ‘Fake’ sometimes had truthful article.
- There is No way to distinguish them without doing a *sanity* check.

Why Important ?

Every 60 seconds :-

98,000 Tweets Are shared on Twitter

600 Videos are uploaded to You Tube

600,000 Status updated and posts are created

……nobody but a machine could keep up with it all

How to Recognize ?

Check who is reporting ? . If its a mainstream source chances are it is True. If it's a site you never heard of, be skeptical.

- Watch for Headlines and content typos
- Watch for excessive punctuation
- Watch for biased vocabulary
- Example: " Immigrants Vs Illigle"

Detecting Fake News using Machine Learning

- As a human being when we read paragraph, we can interpret the words with the hole document and understand the context.

- Given todays volume of news , it is possible to teach to a computer how to read and understand the differences between fake news and real news using Naive Bayes Machine Learning Algorithm. The building blocks are data set and Machine Learning Algorithm

System Architecture

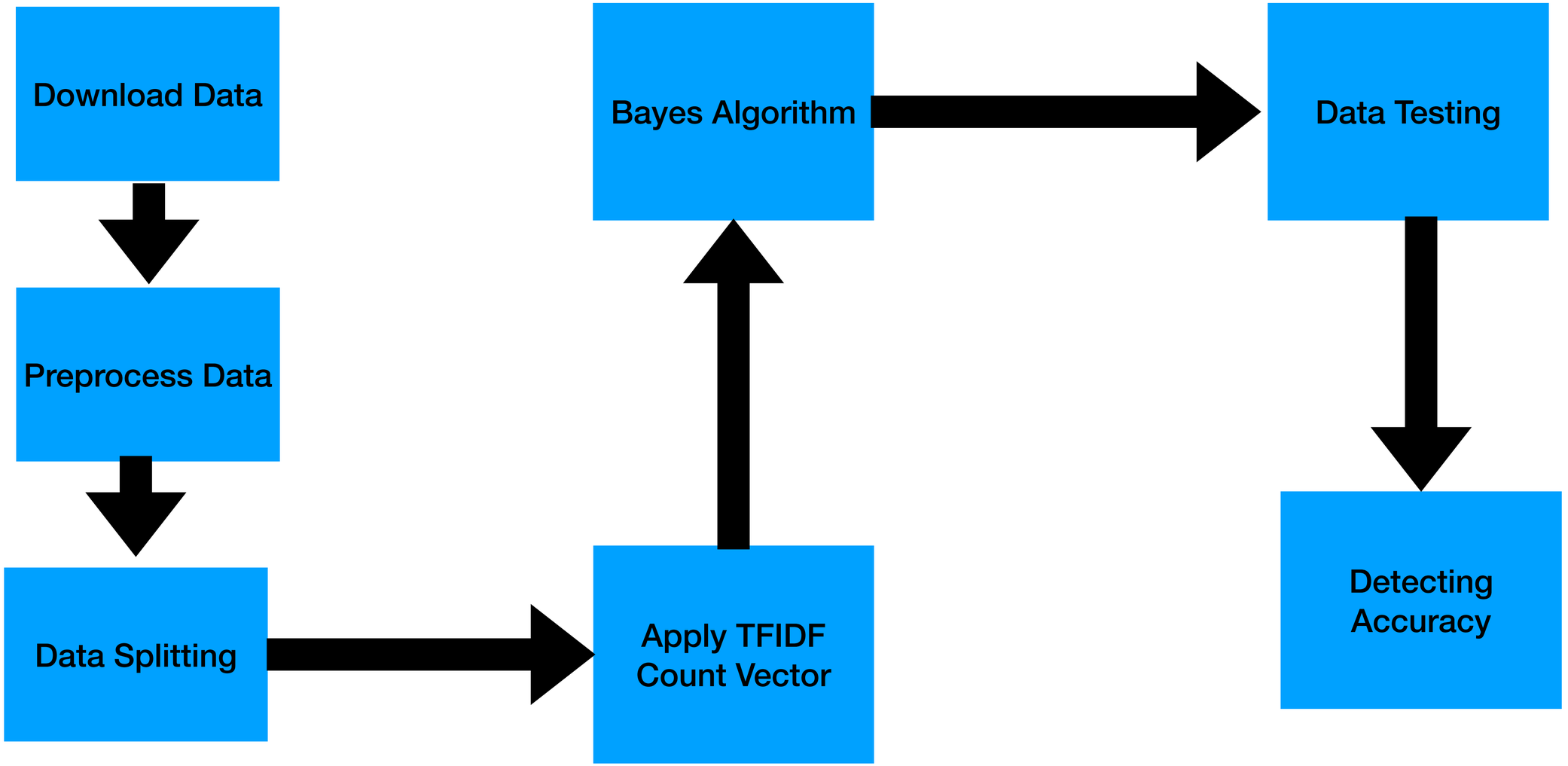


Fig:- System Architecture of Machine Learning

System Architecture

- The First Step in detection of Fake News is to download Data , in this case we downloaded data from Kaggle (source:- https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset)
- Data set is split into two parts that is train and test , we used scikit learn library to split the data
- Now Multinomial Naive Bayes Algorithm is used to classify the train data in groups of similar entities.
- Naive Bayes Algorithm is applied to the test data set
- Finally we determine the accuracy of the model and weather the given news is Fake or real

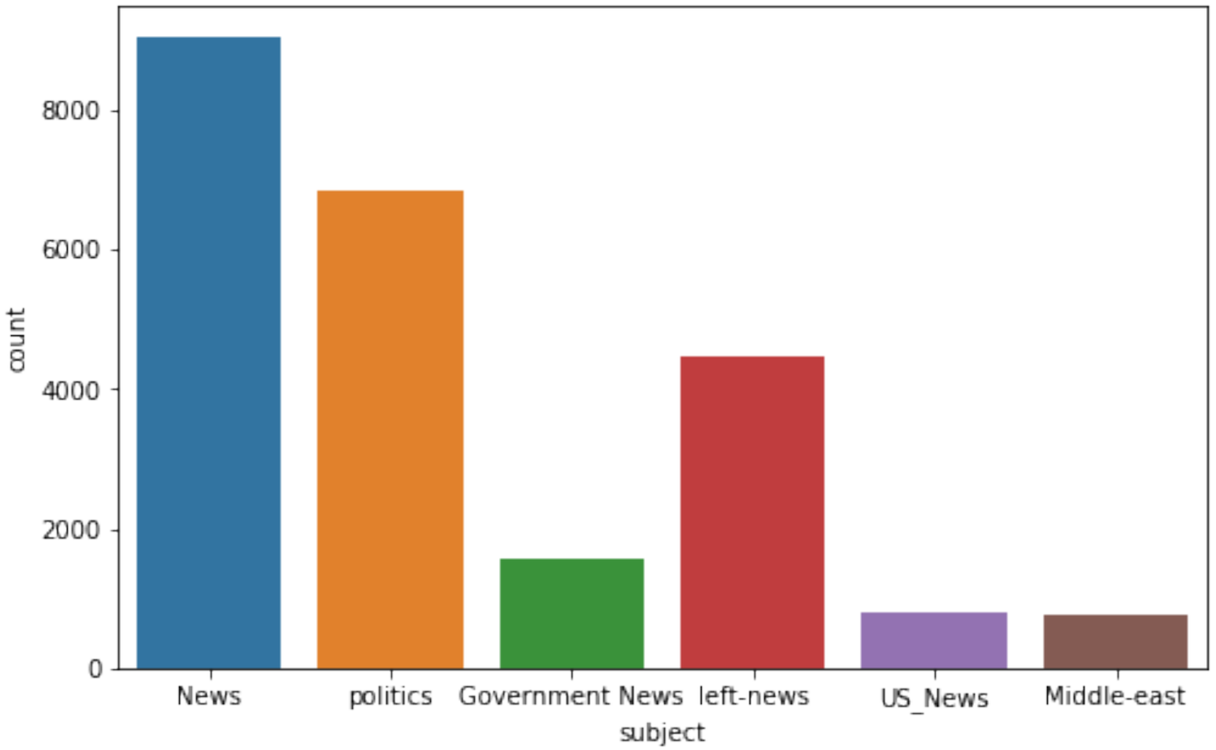
EDA

- Exploratory Data Analysis is one of the most important step in any Machine Learning Project
- While exploring data we observed that real news seems to have source of publication whereas fake news don’t
- Most of the text contains reuters information such as “WASHINGTON (Reuters)”
- Some texts are tweets from twitter
- Few text do not contain any publication information.

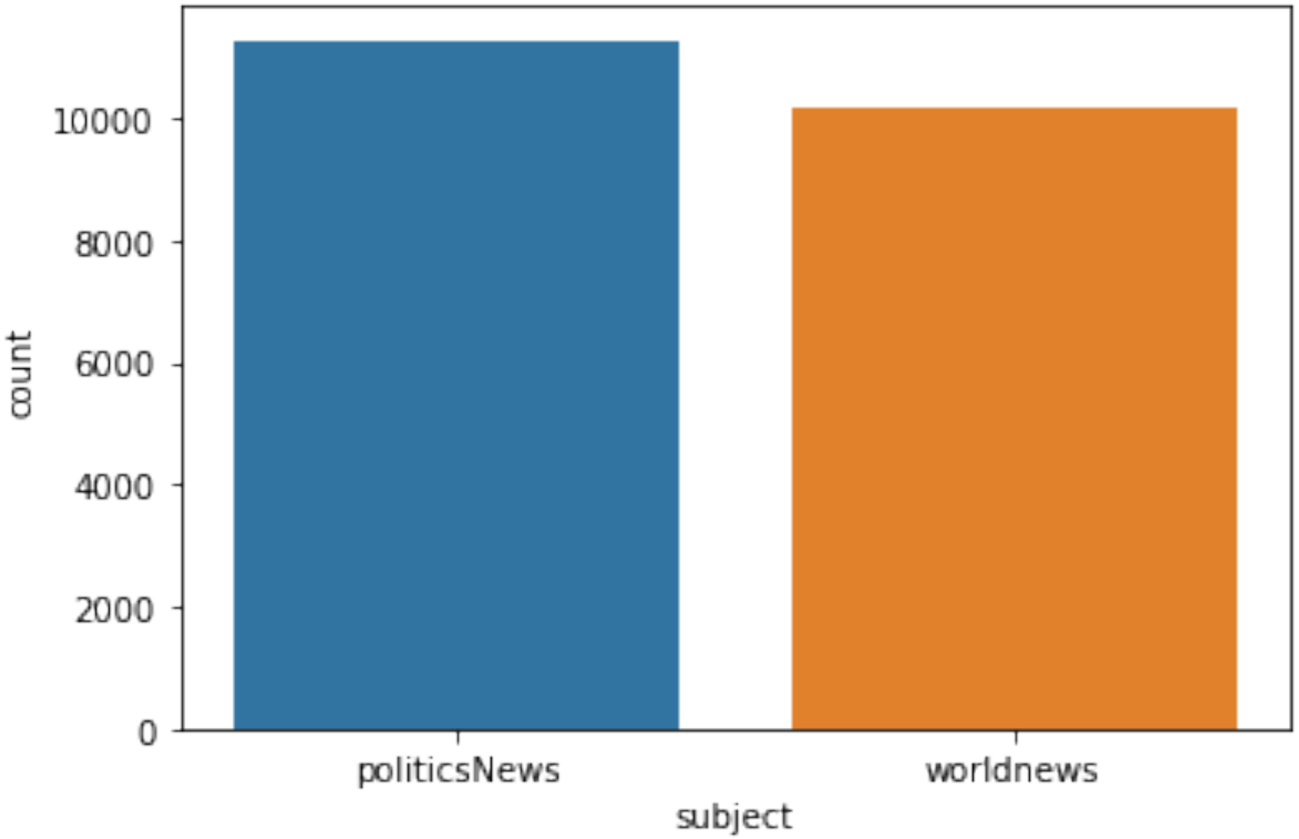
Data Pre-processing

- This process Contain all the data which must be checked thoroughly and preprocessed
- Check data if it has missing values
- Counting subject wise news spread and it seems that we have most of the news from politics
- While doing EDA we found some interesting facts we will discuss it in Data Cleaning

Fake Vs. Real

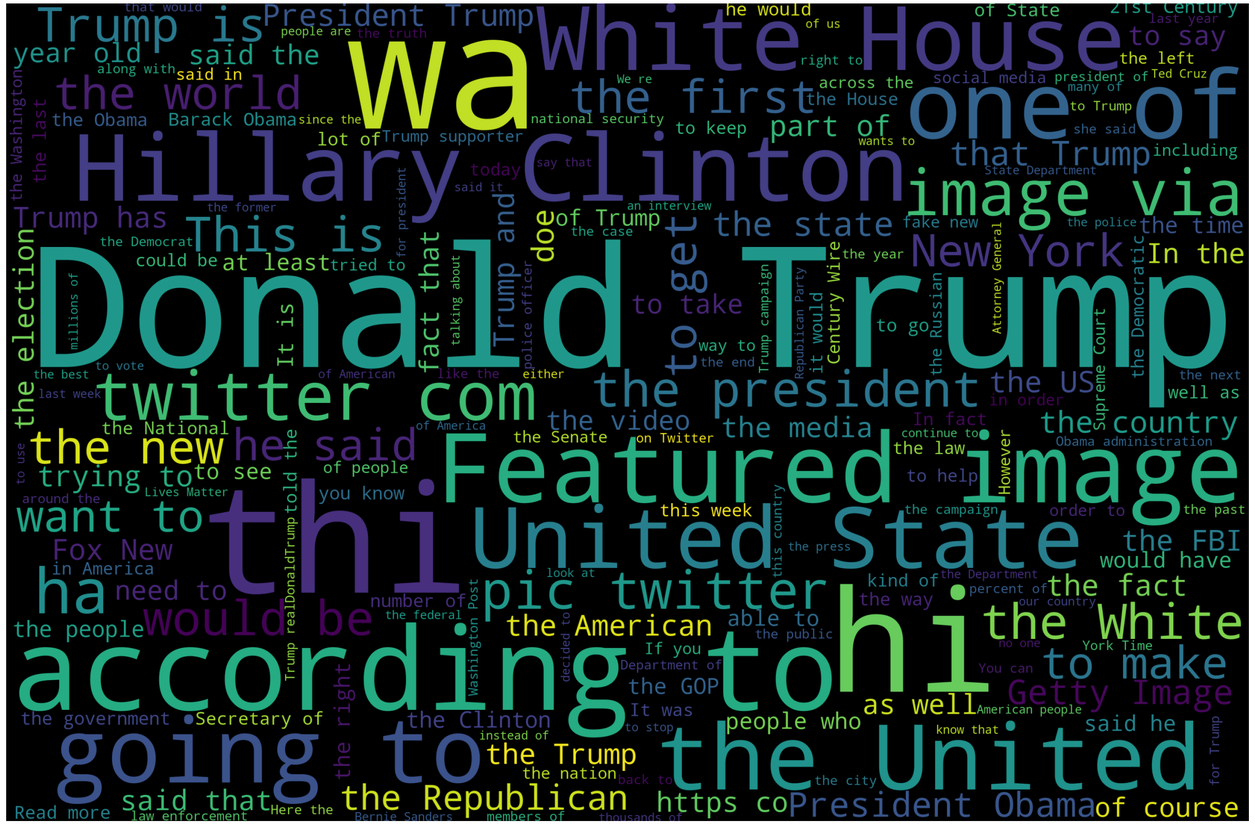


Fake News Subjects

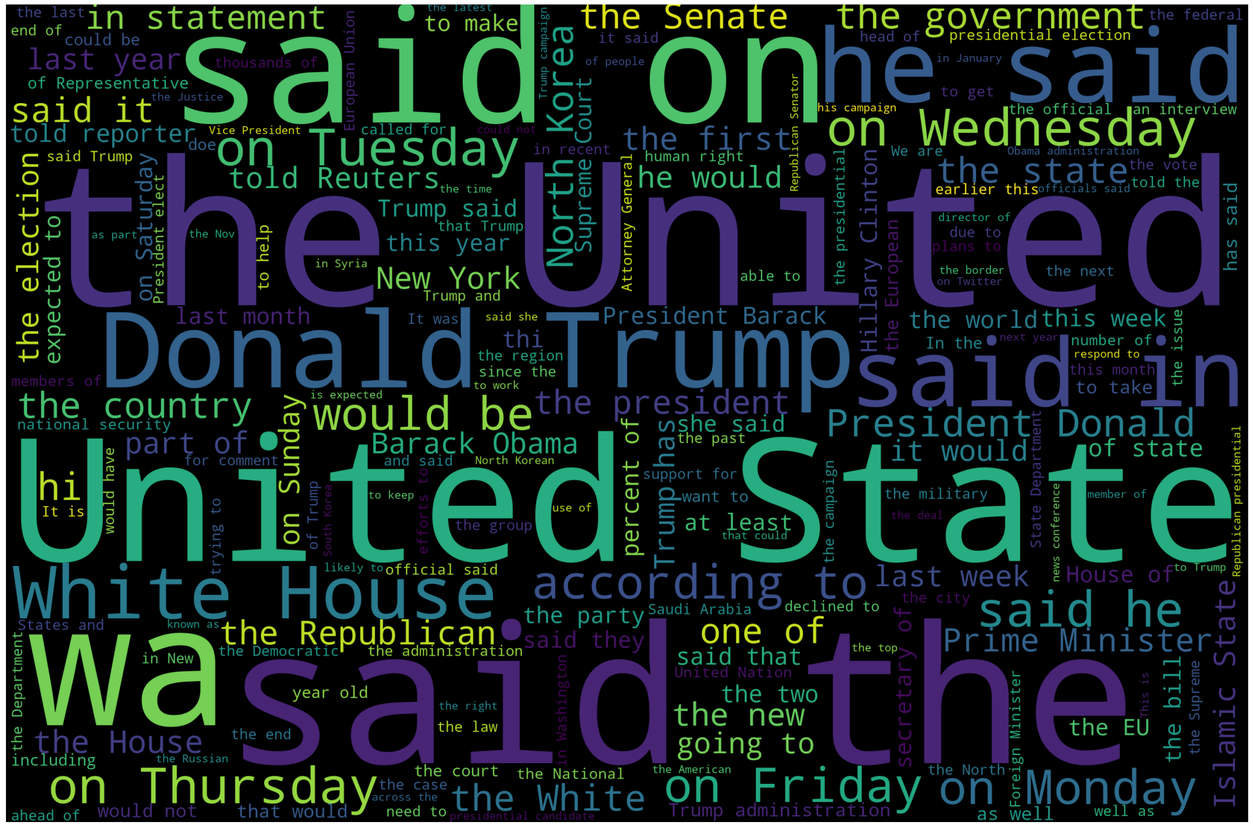


Real News Subjects

Fake Vs. Real Word Cloud



Fake News



Real News

Data Cleaning

- While Cleaning Data we have to check null values and empty text
- Fake news has only one empty text whereas real news has 630 empty text
- Removing reuters or twitter tweet information from the text
- Separate text part and publication part , as we already know fake news data do not have publication part
- If we do not get the text part this means publication details wasn’t given for the record.

Data Treatment

- Adding Class information , we are giving class 1 for real news and class 0 for fake news.
- Combining ‘text’ and ‘title’ in both data set
- Subject is different for real and fake thus dropping it
- Dropping ‘subject’ ,’date’, ‘publisher’ and ‘title’ real data
- Also dropping ‘subject’ ,’date’ and ‘title’ in fake news data
- Combining ‘fake’ and ‘real’ news data and creating new data set having only two columns ‘text’ and ‘class’
- We removed all stopwords and punctuation marks in text data by using ‘nltk’ library

Model

- Naive Bayes Classifier is used here to classify fake and real news using Multinominal NB and Pipelining concept
- There are number of algorithms that focus on common principle and its not the only algorithm for training such classifier
- It is kind of algorithm which is used in text classification , the use of token is correlated with the news that may be fake or not fake in the Naive Bayes classifier and then the accuracy is find out using Bayes Theorem
- Naive Bayes classification uses the probability of previous event and compares it with existing event , this process is repeated for each event
- At last overall probability of the news as compared to dataset is calculated
- Calculating overall probability we can get the approximate value and can detect whether the news is fake or real

Results

	precision	recall	f1-score	support
0	0.93	0.95	0.94	4668
1	0.95	0.92	0.93	4312
accuracy			0.94	8980
macro avg	0.94	0.93	0.94	8980
weighted avg	0.94	0.94	0.94	8980

Summary of Model

Feature Selection Technique

- In this project we have done feature extraction and selection from scikit learn library
- To perform feature selection we used method ‘CountVectorizer’ and ‘TF-ID Vectorizer’
- Also pipelining has been used to ease the code

Working Code

Conclusion

Therefore by using Naive Bayes theorem we can conclude that any news from large or small dataset can be classified as fake or real news by matching it with the previous dataset values in less time which intern helps the users to believe in a particular news.

Thank you……