Assignment 1: Fake News Detection via Text Classification

CS525: Natural Language Processing
Thabsheer Jafer Machingal

Dataset

For this assignment, we are using the Fake and real news dataset on kaggle [1]. The dataset contains two set ‘Fake.csv’ and ‘True.csv’. The dataset has 21417 data points for true news and 23481 for fake news. It has four different categories,namely title, text(news content), subject and date of release.

Task 1 - Explore Essential Information from Text Data Preprocessing

The dataset is fairly balanced between true and fake news, a pie chart is given below which shows the percentage of each true and fake news in the datset

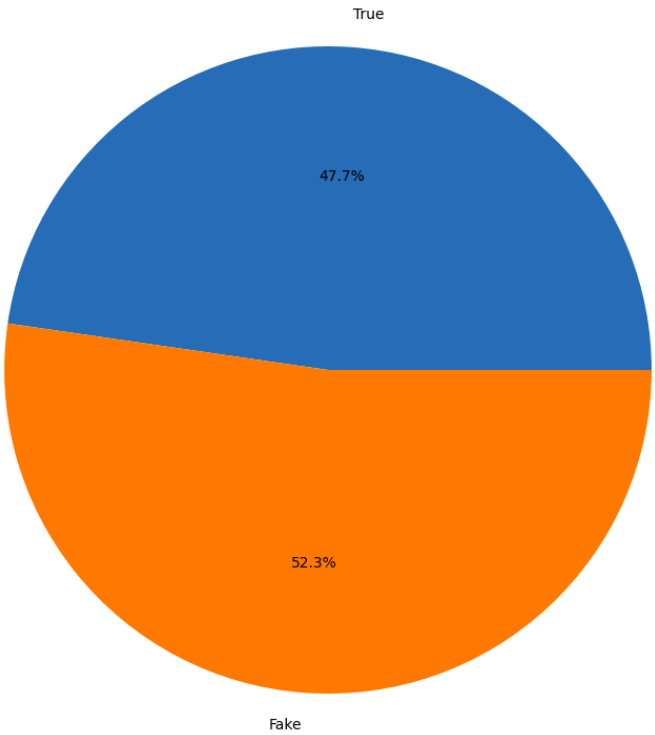

Figure 1: The distribution of fake (orange color) and real news (blue color).

The dataset has 17 duplicated data points (combined real and fake news), which is about 0.04% of the total. Hence we dropped these data points and updated the dataset.

A distribution of the dataset between different subjects is given below. Political and world news contributed more to the dataset.

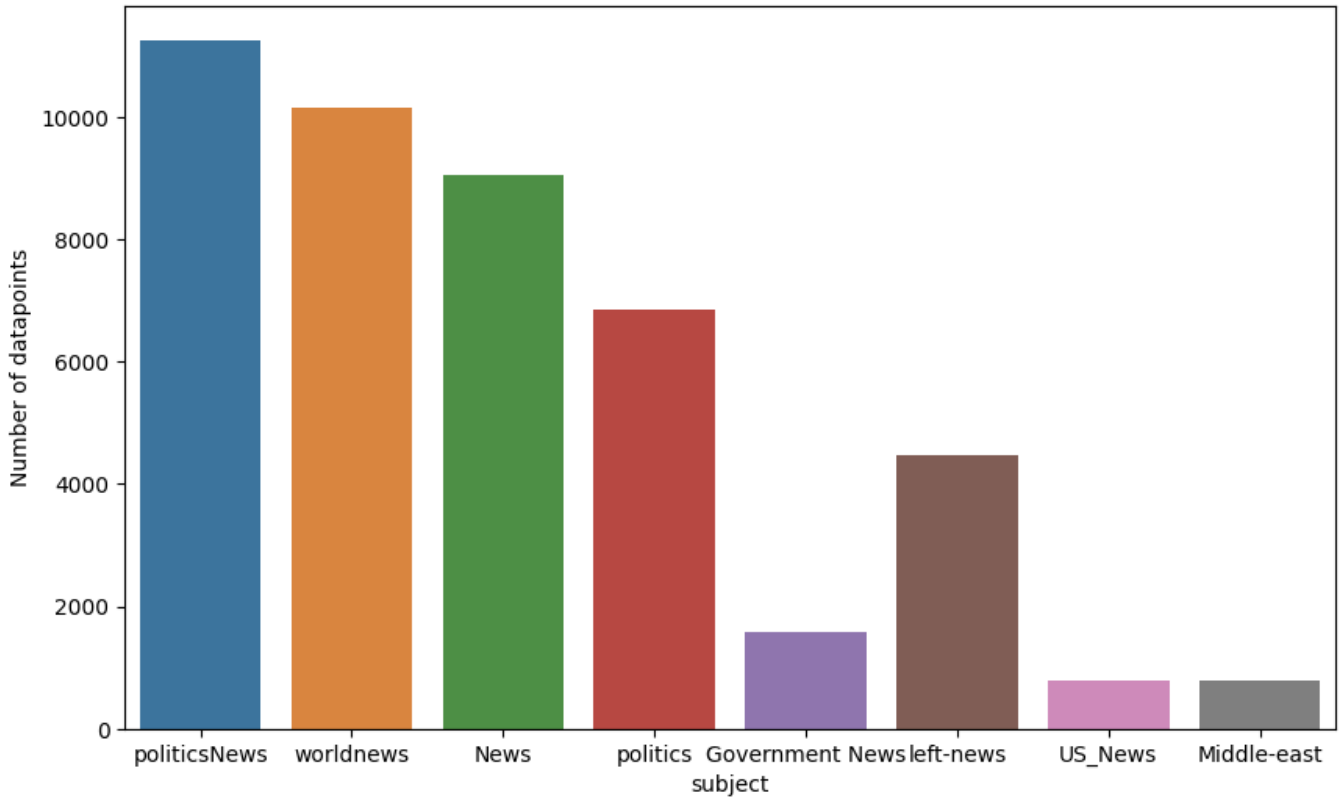

Figure2: Distribution of data among different subjects

1. *What are the most common;y used words in fake and real news?*

Number of common words in true and fake news are given below as word clouds for easy understanding. I extracted 100 common words in true and fake news separately.

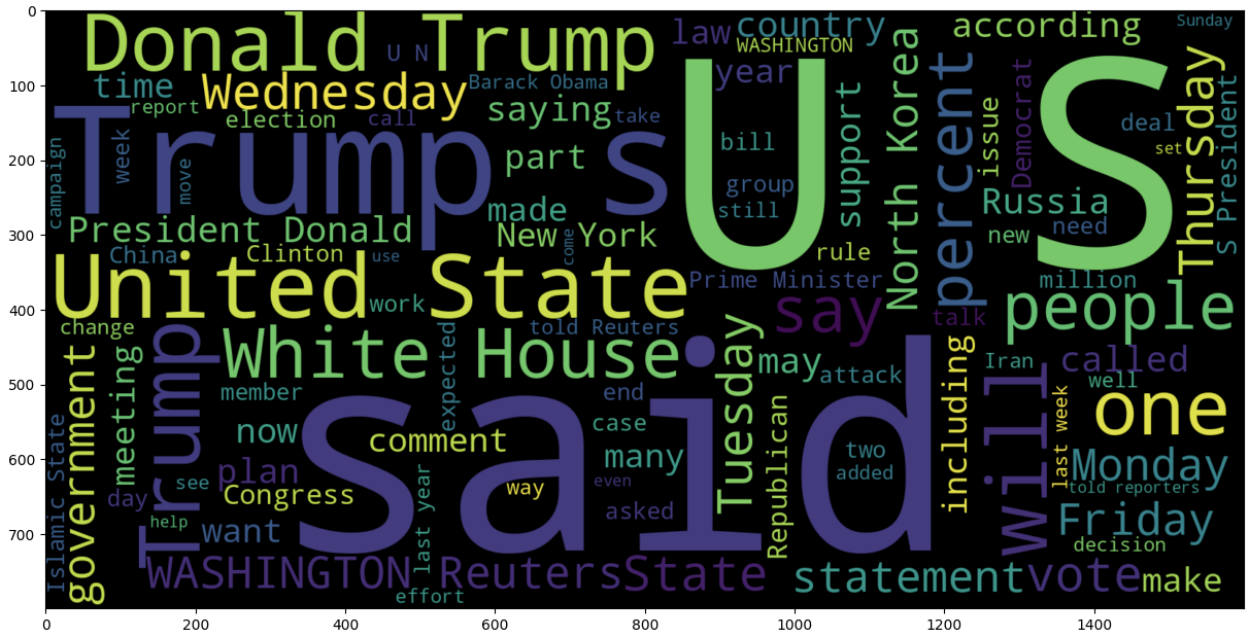

Figure3: Showing the 100 most common words in the true dataset

The most common words in the true data are US, Said, Trump e.t.c.. If the word in the word cloud is bigger, which occurs more frequently in the dataset.

One interesting fact to note that, US and United States are among the two common words, even though it means the same, this could be one way to improve feature extraction. Considering synonyms as just one word could improve the efficiency by reducing the size of the test set, it intuitively make sense.

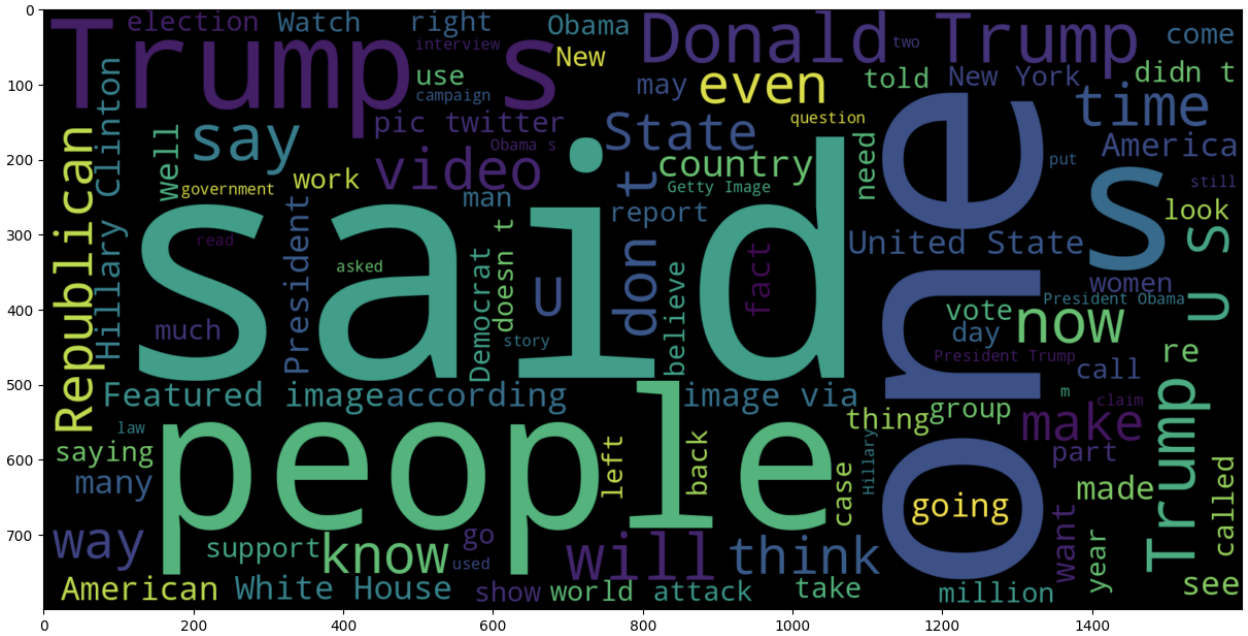

Figure4: Showing the 100 most common words in the fake dataset

The common words in fake dataset are said, one, people, Donald Trump e.t.c..

2. *By looking at the data Can you tell the difference between real and fake news?*

By looking at the word cloud, I cannot actually differentiate fake or real news. For humans these are commonly used words and their frequency might not make sense, but our model can identify these frequencies as a feature, either TF, IDF or TF-IDF. There is no way we can predict if the news is real or fake by looking at the data above.

3. *What does the strongest feature set (for machine learning) look like*?

For machine learning, we can use features like term frequency, TF-IDF, bag of words e.t.c..

Task 2 - Build Machine Learning Model

I implemented two machine learning models, **Multinomial Naive Bayes** and **Logistic regression.**

Two different feature sets were used, Bag of words (frequency of words in the collection) and TF-IDF (Term Frequency- Inverse Document Frequency).

The following results were observed:

ML Model	Feature	Precision	Recall	Accuracy
Multinomial NB	Bag of Words	0.94	0.94	0.94
Multinomial NB	TF-IDF	0.93	0.92	0.92
Log Regression	Bag of Words	0.93	0.93	0.93
Log Regression	TF-IDF	0.83	0.75	0.75

Table 1: Table showing the performance on test set

CONFUSION MATRIX -Multinomial NB
Bag of Words

Confusion matrix for Multinomial Naive Bayes using ‘bag of words’ as feature set is given below.

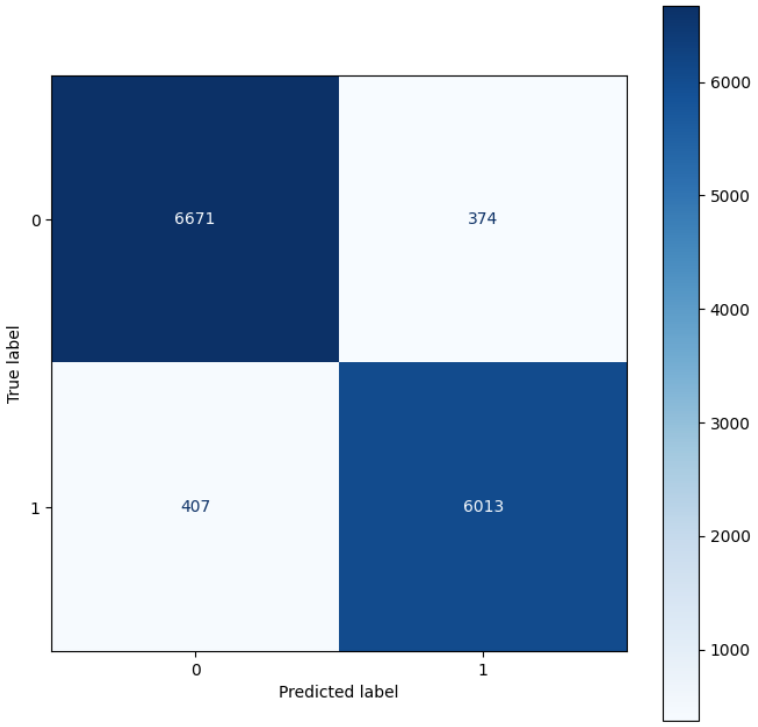

Figure 6: Confusion matrix of the Multinomial NB on Bag of words

TF-IDF

Confusion matrix for Multinomial Naive Bayes using tf-idf as feature set is given below.

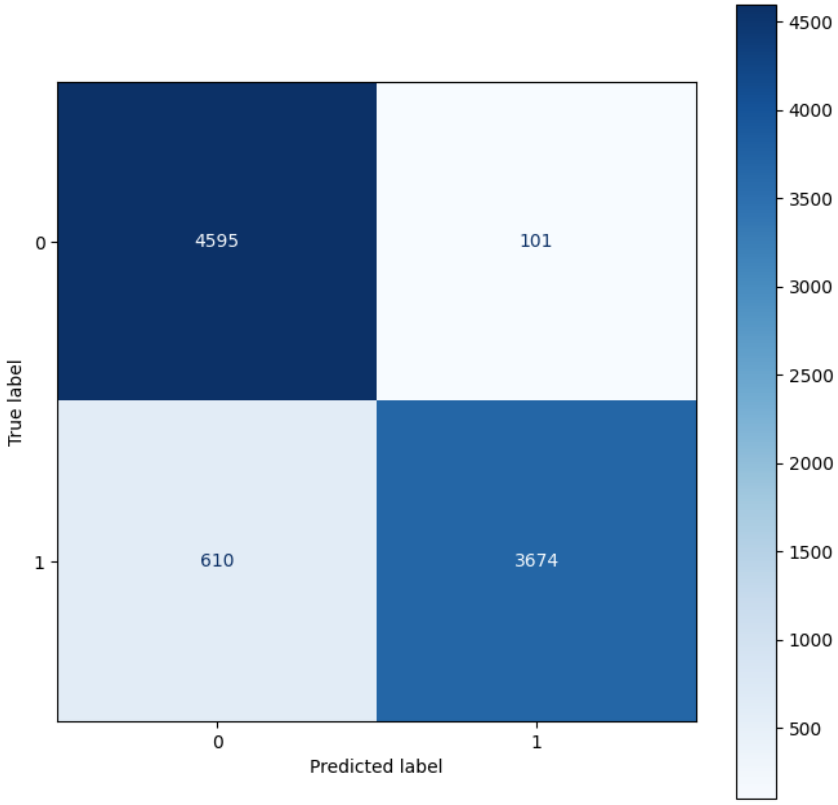

Figure 7: Confusion matrix of the Multinomial NB on Bag of words

Task 3 - Enhanced NLP Features

For this part of the assignment I modified the dataset, used POS tagging and removed words other than nouns in the text.
The new dataset has words that are only nouns.

ML Model	Feature	Filter	Precision	Recall	Accuracy
Multinomial NB	Bag of Words	Noun	0.92	0.92	0.92
Multinomial NB	TF-IDF	Noun	0.90	0.80	0.89
Log Regression	Bag of Words	Noun	0.90	0.90	0.90
Log Regression	TF-IDF	Noun	0.80	0.70	0.70

Table 2: Table showing the performance on test set

As a second filter I kept both Nouns and adjectives

ML Model	Feature	Filter	Precision	Recall	Accuracy
Multinomial NB	Bag of Words	Noun+Adjective	0.92	0.92	0.92
Multinomial NB	TF-IDF	Noun+Adjective	0.90	0.89	0.89
Log Regression	Bag of Words	Noun+Adjective	0.91	0.91	0.91
Log Regression	TF-IDF	Noun+Adjective	0.81	0.71	0.71

Table 3: Table showing the performance on test set

Evidently filtering the words and using the frequency as a feature reduces the accuracy and performance on the unseen test set. This could be because of loss of information.

Task 4- Future Work

I believe we can improve efficiency and performance not just in terms of accuracy but also the quality of the dataset.

1. Better Machine learning models
   ofcourse, using better ML models or testing out different ML and exploring options can actually leads to good results. People have done work on this dataset using Support Vector Machine(SVM)[2], Linear SVM and obtained better results (in terms of accuracy).
2. It would be interesting if we define word embeddings as a feature set. We can use language models and train the machine to try and understand why fake news is fake. Fake news can be different kinds, they could be purposefully created, created for humor or due to poor quality of writing. This can be addressed by sentiment Analysis[3]. Understanding the data better can be helpful to improve efficiency and precision. Using machine learning with semantic features have proved to yield better results [4],[5].

References

[1] “Fake and Real News Dataset,” accessed October 5, 2022, https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset.

[2] Rubin., Victoria, L., et al.: Fake news or truth? Using satirical cues to detect potentially misleading news. In: Proceedings of NAACL-HLT (2016)

[3]S. Volkova, K. Shaffer, J. Jang Yea and N. Hodas, "Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter", *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, pp. 647-653, 2017.

[4] N. J. Conroy, V. L. Rubin and Y. Chen, "Automatic deception detection: Methods for finding fake news", *Proc. Annu. Meeting Assoc. Inf. Sci. Technol.*, pp. 1-4, 2015.

[5] W. Y. Wang, "Liar liar pants on fire: A new benchmark dataset for fake news detection", *Proc. Annu. Meeting Assoc. Comput. Linguistics*, pp. 422-426, 2017.