# Fake News Detection

Aarush Mathur

## INTRODUCTION

The objective of this project is to detect fake news using Natural Language Processing technique. We use a corpus of labeled real and fake new articles to build a classifier that can make decisions about information based on the content from the corpus. We use a text classification approach, using four different classification models, and analyze the results. The model focuses on identifying fake news sources, based on multiple articles originating from a source. Once a source is labeled as a producer of fake news, we can predict with high confidence that any future articles from that source will also be fake news. We use a text classification approach, using four different classification models, and analyze the results. The intended application of the project is for use in applying visibility weights in social media. Using weights produced by this model, social networks can make stories which are highly likely to be fake news less visible.

## METHODS

It is a supervised learning problem where we have a dataset of 44898 rows × 6 columns where we will be predicting whether some news is FAKE or REAL, represented as Target 0 and 1, based on these 6 features and observing and comparing the accuracy of different algorithms.

We are using 4 algorithms-

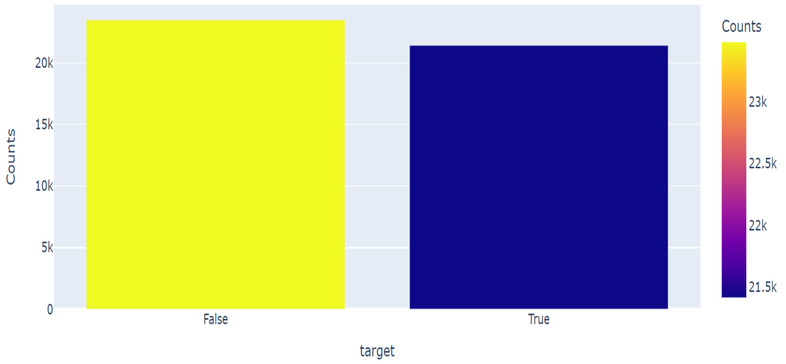
1. Passive Aggressive Classifier
2. Logistic Regression
3. Naïve Bayes
4. SVM

For the purpose of smooth analysis, we removed some components which are not helpful for our analysis, like removing unnecessary words. We compared the amount of fake and real news(Target 0 and 1).

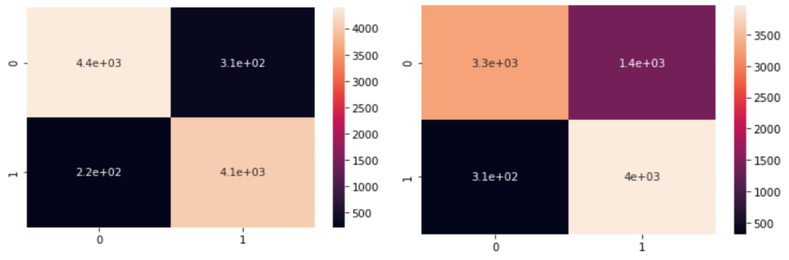
## RESULTS

I have 6 features in the dataset. I took one variable as the target dataset. The target dataset is in binary “Real” and “Fake”. My dataset has 44898 rows. For training I used “text”, “title” features and for target I used “label” feature.

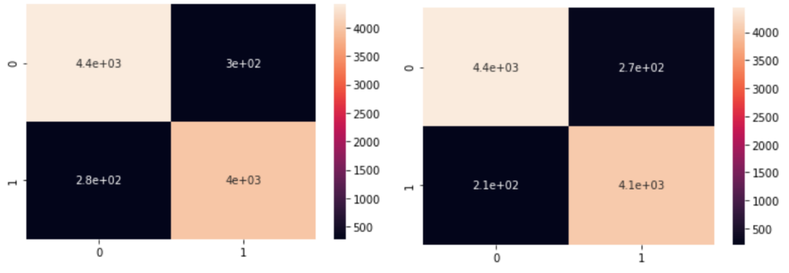
### Exploratory Data Analysis



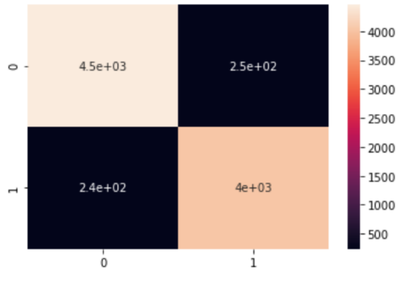

### Confusion Matrices




Logistic Regression | Naïve Bayes




Passive Aggressive Classifier | SVM




Grid Search (Logistic Regression)

## DISCUSSION

| Model Name | Accuracy % | Wrong Predictions |
|---|---|---|
| Passive Aggressive Classifier | 93.57 | 580 |
| Logistic Regression | 94.05 | 530 |
| Naïve Bayes | 80.82 | 450 |
| SVM | 94.6 | 480 |
| Grid Search CV(Logistic Regression) | 94.62 | 490 |

I used four classifiers Passive Aggressive Classifier, Logistic Regression, Naïve Bayes and SVM. When I compared them, Passive Aggressive Classifier gave me accuracy of 93.57%, Logistic Regression gave me accuracy of 94.05%, Naïve Bayes gave me accuracy of 80.82%, SVM gave me accuracy of 94.60%. While as expected Naïve Bayes gave the most wrong predictions and SVM has the least wrong predictions, Then I applied Grid Search CV on Logistic Regression which resulted in highest accuracy of 94.62%. I created confusion matrix to visualize the result of these classifiers.

## CONCLUSIONS

My goal was to detect fake news using NLP. From the given dataset I used a text classification approach, using four different classification models, and analyzed the results. The best one in terms of accuracy was SVM with 92.6% accuracy rate.

I also performed Hyperparameter Tuning on Logistic regression using grid search to get accuracy of 94.62%.

I will try more algorithm in future for this project and continue my search for best algorithm for the given problem.

## REFERENCES

[1]https://thecleverprogrammer.com/2021/07/09/end-to-end-fake-news-detection-with-python/