Natural Language Processing

Analysing Fake News

Projects Presented

By

Kodangada Ketan Bopanna

Subramani Praveen

Iyaju Funso Stephen

Scholl Liam

Table of Content

1.0 INTRODUCTION

Fake news datasets are problematic for governments, individuals and companies. In this project, we will focus on text-based news and try to build a model that will help us to identify if a piece of given news is fake or real.

Simple content related classification n-gram and part of speech (POS) tagging have proven insufficient in fake news context. Fake news detection through classification is not sufficient since it missed the important context of the information, however a deep analysis of the content shows that context-free grammar (CFG) produced good results with the combination of the n-gram in deception related classification. The accuracy achieved 85%-91% when applied on news article datasets through classification. In this project, we build a text classifier system that use machine learning approach to categorise news text. This process of classification is not restricted to text alone. It is also used in other domains including science, healthcare, weather forecasting, and technology. The main objective of this project is to develop a model based on the count vectorization and TF-IDF.

1.1 PROJECT

In this project, we build a text classifier system accurately classified collection of news as real or fake using machine learning model. The system was developed in python using the TF-IDF library. We first create the model, then we initialize the classifier, transform and fit the model using our training data set split. We check the accuracy of the model using our test data split to calculate the performance of the model using the appropriate performance metrics. We do a comparative analysis of the model to check the model that has the best performance.

1.2 DATA ANALYSIS

In this project, our dataset contains 4594 rows and 4 columns. The columns are:

Title - this is the title of the news.

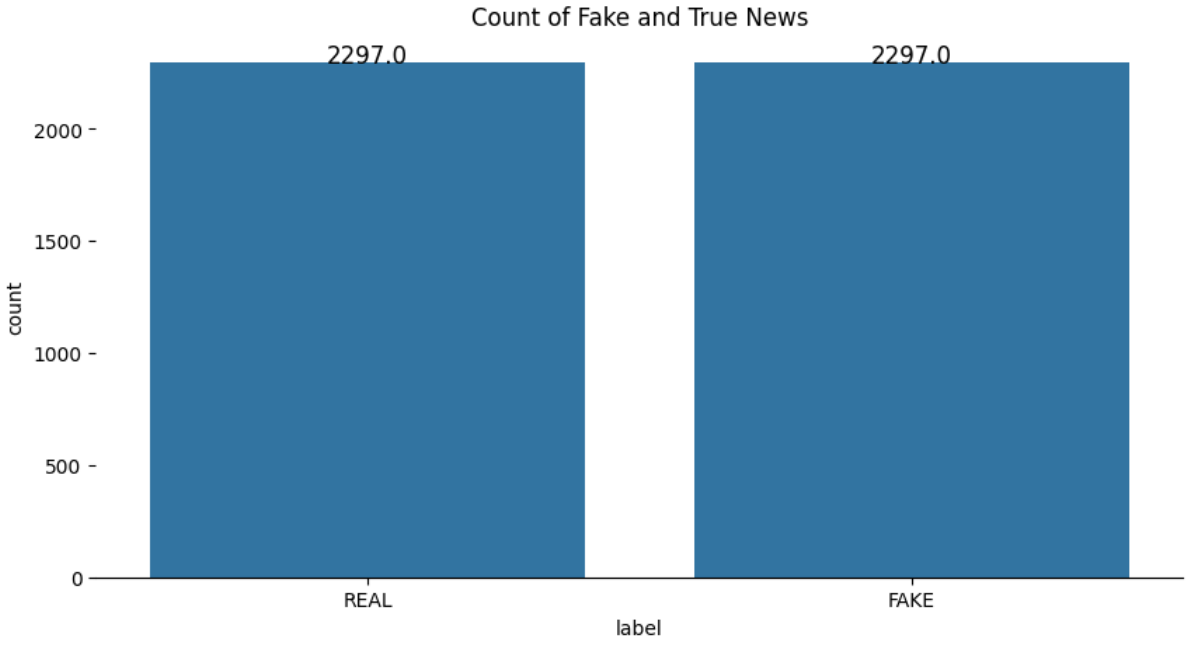
Text - this is text of the news itself.

Idd - this is the internal id for uniquely representing each new items

Label - this is a text column representing if the news is real or fake

1.3 DATA EXPLORATION

In this project, we carried out an exploratory analysis of the distribution of our target labels. From the graph below, we can see that the distribution of our target labels is balanced



1.4 DATA PREPROCESSING

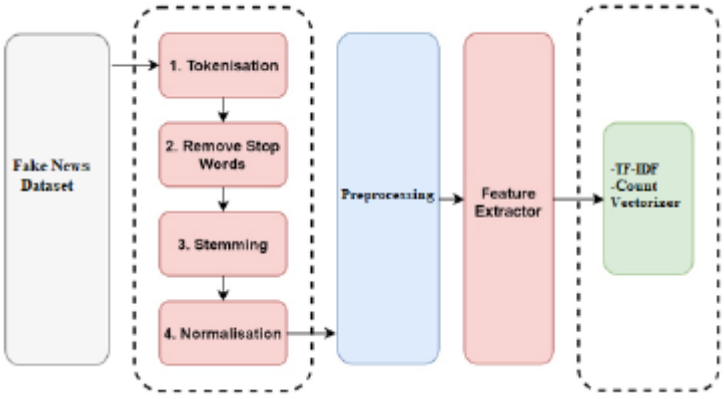
In data pre-processing phase, we check the title column for null value. We replace any null title column with a default title and concatenate the title column with the text column to replace the text column. We focus on the text column to prepare data for modelling.

1.4.1 TEXT NORMALIZATION

In this project we created some functions to handle text normalization. Our text normalization techniques involves the removal of special characters and symbols, text cleaning, removing the English stopwords, converting text to lower case and text standardization through stemming and lemmatization

1.4.2 TOKENIZATION

We apply tokenization step on the text column to break into tokens of words. The diagram below shows the flow of steps carried out the text columns



1.5 FEATURE ENGINEERING

Usually extracted features are fed into ML algorithms for learning patterns that can be applied on future new data points for getting insights. These algorithms usually expect features in the form of numeric vectors because each algorithm is at heart a mathematical operation of optimization and minimizing loss and error when it tries to learn patterns from data points and observations.

To make the text column relatable for machine learning, they must first be converted into some numerical structure or numerical vector space. This process is a mathematical and algebraic model for transforming and representing text documents as numeric vectors of specific terms that form the vector dimensions.

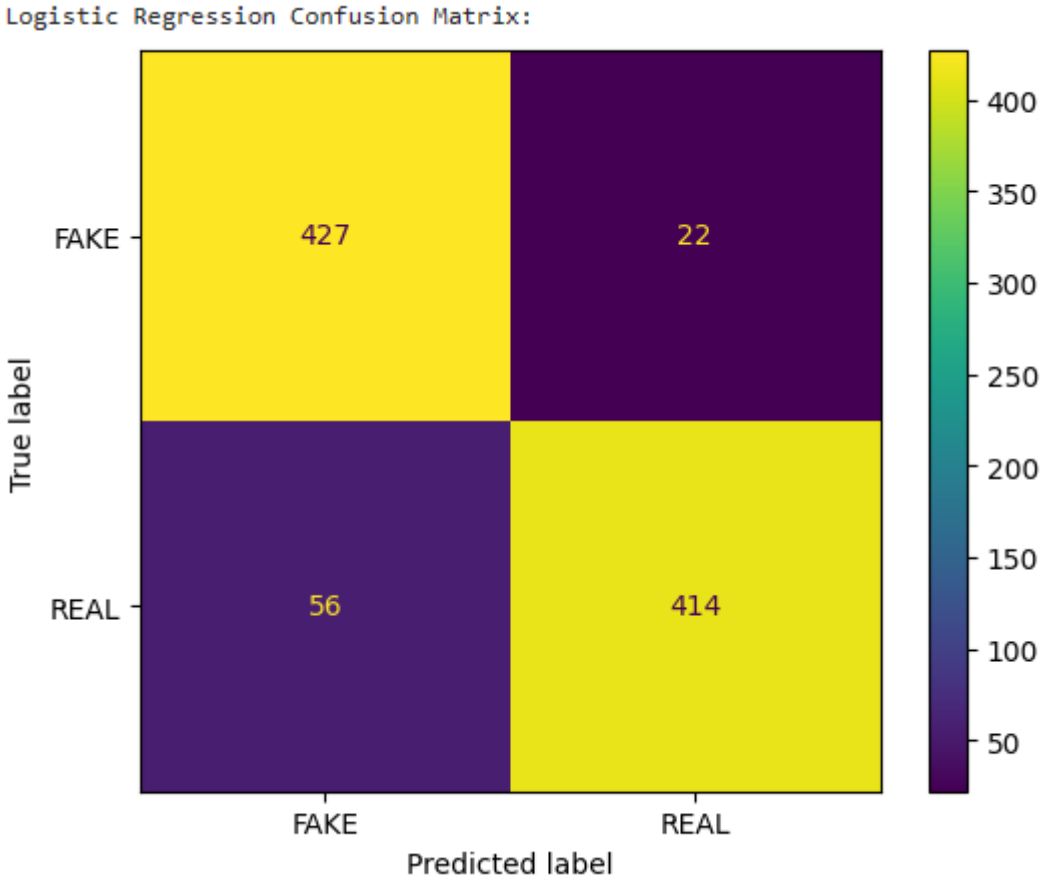
There are few techniques that are used to achieve this but in this project, we are using TF-IDF model. Sci-kit learn provide 'Tfidfvectorizer' library that enable us to carry out TF-IDF Vectorization. TF-IDF stands for "term frequency-inverse document frequency", meaning the weight assigned to each token not only depends on its frequency in a document but also how recurrent that term is in the entire text corpora. The process involves computing the word count and then computing the Inverse Document Frequency

1.6 MODELING

After the vectorization of the text, we split the data into train and test data set. We then fit three machine learning models to the training data set. We experimented with classification machine learning models that have proven to be effective and give good results in text classification. We used Logistic Regression, Multinomial Naive-Bayes and Support Vector machine Classifier.

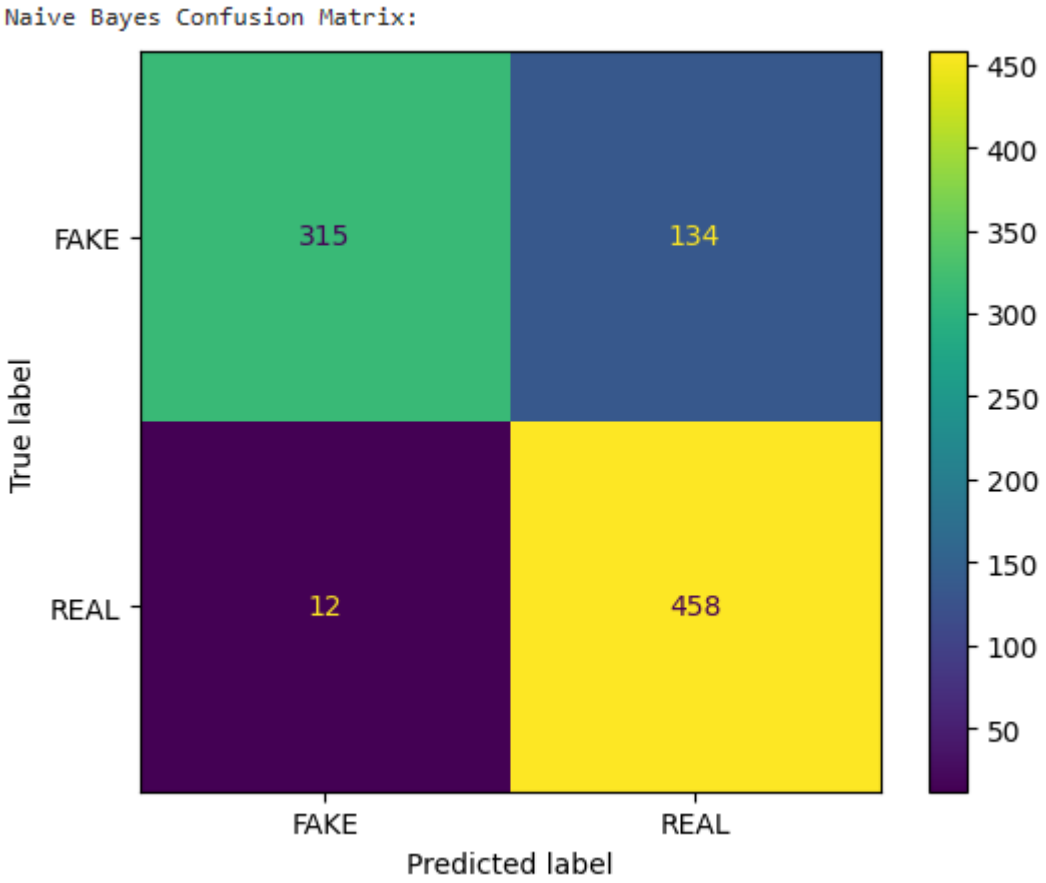
1.6.1 LOGISTIC REGRESSION

It is used to estimate the relationship between variables after using statistical methods. It performs well in binary classification problems because it deals with classes and requires a large sample size for initial classification. The confusion matrix below helps us to gain insight into the number of false and true negatives and positives obtained after fitting and using the model



1.6.2 MULTINOMIAL NAÏVE BAYES

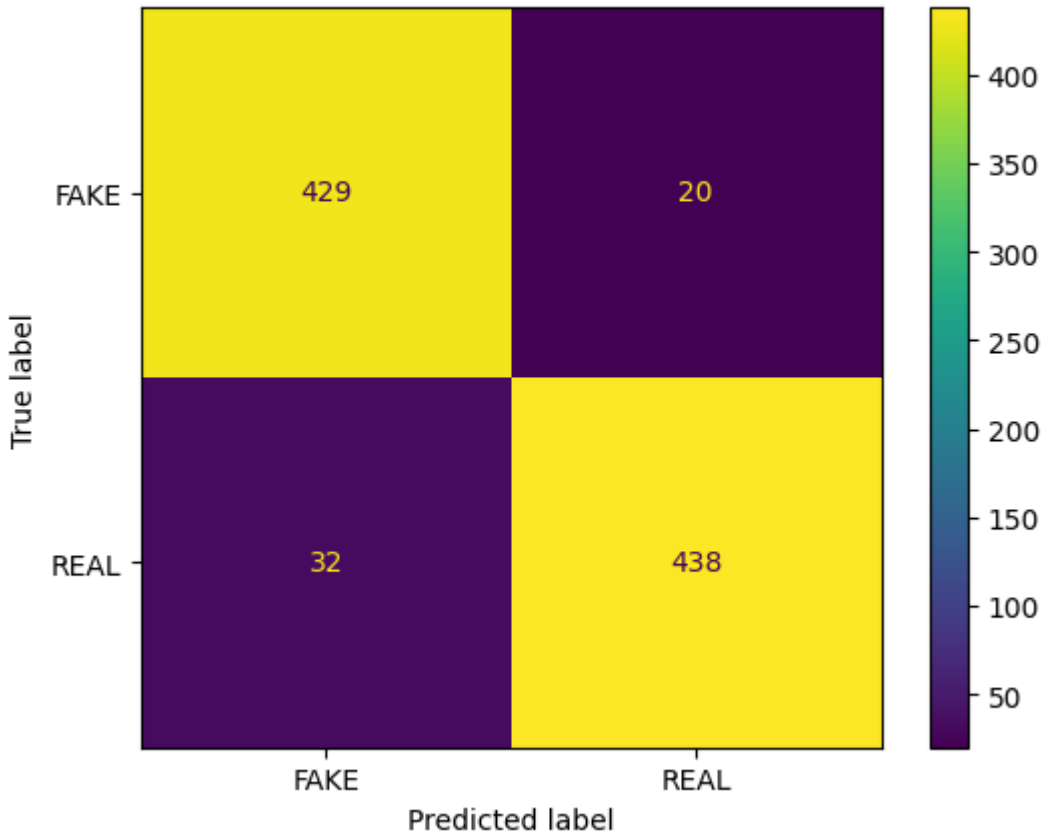
It is a powerful classification model that performs well when we have a small dataset and it requires less storage space. It does not produce good results if words are co-related between each other. The confusion matrix below helps us to gain insight into the number of false and true negatives and positives obtained after fitting and using the model



1.6.3 SUPPORT VECTOR MACHINE

It performs supervised learning on data for regression and classification. The SVM computes the data and converts it into different categories. The advantages of Support Vector Machine are learning speed, accuracy, classification and tolerance to irrelevant features. Support Vector Machine is one of the most researched classifiers nowadays and it performs well in the fake news detection problem. The confusion matrix below helps us to gain insight into the number of false and true negatives and positives obtained after fitting and using the model

Support Vector Machine Confusion Matrix:



1.7 RESULT

We apply the model for prediction on the test set from the TfidfVectorizer and calculated the accuracy with accuracy_score() from sklearn.Metrics and other metrics. The result and comparison of the performance of the machine learning model is shown below

model	accuracy	precision	recall	f1
Naive Bayes	0.8411316648531012	0.866308814629686	0.8411316648531012	0.8377685091473562
Logistic Regression	0.9151251360174102	0.9175478048479129	0.9151251360174102	0.9150807079675467
Support Vector Machine	0.9434167573449401	0.9437529426911697	0.9434167573449401	0.9434239937391791

From the result of the performance comparison, the support vector machine classifier performed the best here and gave an accuracy of 94.34%.

1.8 CONCLUSION

In conclusion, while fake news remains a major challenge in the digital age, Natural Language Processing (NLP) provides us with a promising solution for its detection and mitigation. NLP models can analyse given data using techniques like sentiment analysis, semantic analysis to detect patterns and inconsistencies in the content. By combining the strengths of NLP with human expertise, we can detect fake news and further avoid sharing misleading information.

1.9 REFERENCES

https://medium.com/analytics-vidhya/fake-news-detection-using-nlp-techniques-c2dc4be05f99

https://www.analyticsvidhya.com/blog/2021/07/detecting-fake-news-with-natural-language-processing/

https://arxiv.org/abs/2201.07489

https://iopscience.iop.org/article/10.1088/1757-899X/1084/1/012018

https://www.analyticsinsight.net/how-to-detect-fake-news-with-natural-language-processing/

https://baobabsoluciones.es/en/blog/2022/10/07/nlp-as-a-means-to-detect-fake-news/

https://arxiv.org/pdf/1901.09657

https://medium.com/@mariomg85/resolving-the-issue-of-fake-news-detection-through-natural-language-processing-109e6b0f78ec

[Bishop, 2006] Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer, New York.

[VanderPlas, 2014] VanderPlas, J. (2014). Frequentism and Bayesianism: A Python driven Primer.