# Fake News Detection Project Report

March 2020 – 2021

### Introduction

"Fake news" is "fabricated information that mimics news media content in form but not in organizational process or intent. It has been existing for a long time and becomes a common problem we face every day since the advent of internet. The phenomenon has becomes more evident when social media are increasingly popular.

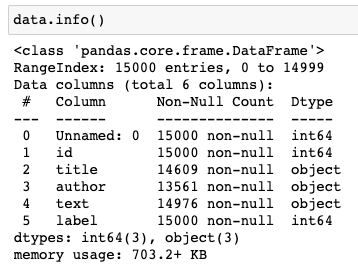
This case study is to build models to detect fake news from a given dataset, to evaluate the accuracy of different models, and to perform a prediction.

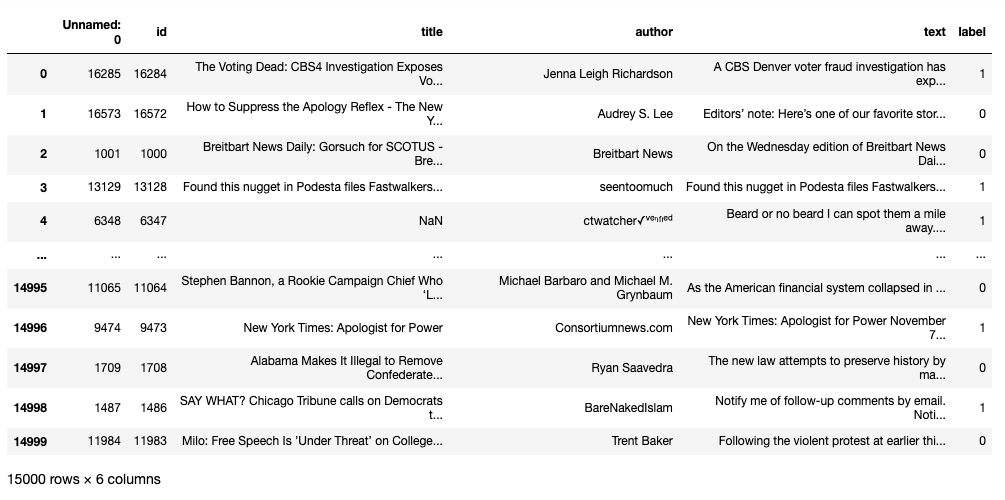
### Methodology

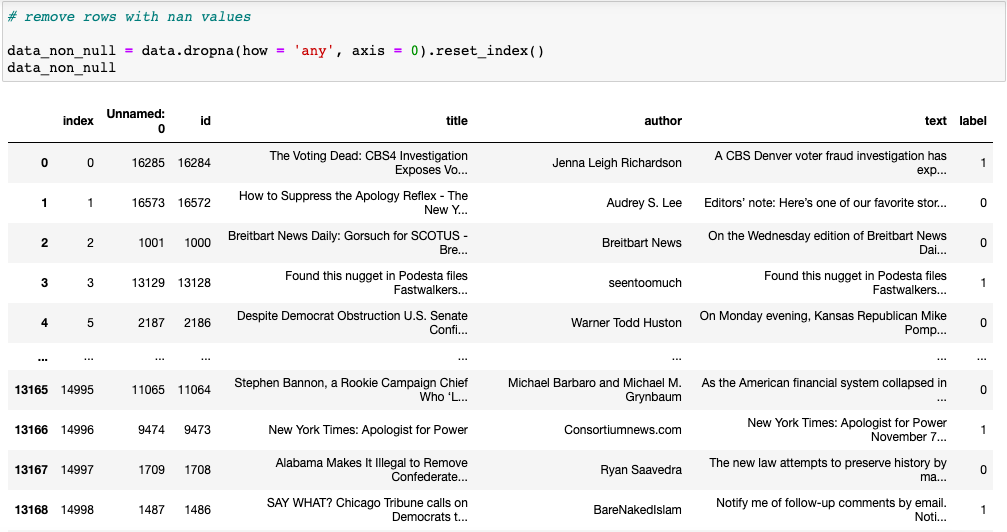
2 datasets are given for this case study: training and testing datasets. Since they are labelled and time independent, I use supervised machine learning to study the problem. I inspect, visualize, and clean the training dataset to see their patterns and characteristics. Then I deploy various models to find out which one is the most suitable for the study case. Finally I test the best model with the testing dataset to see how the prediction is.

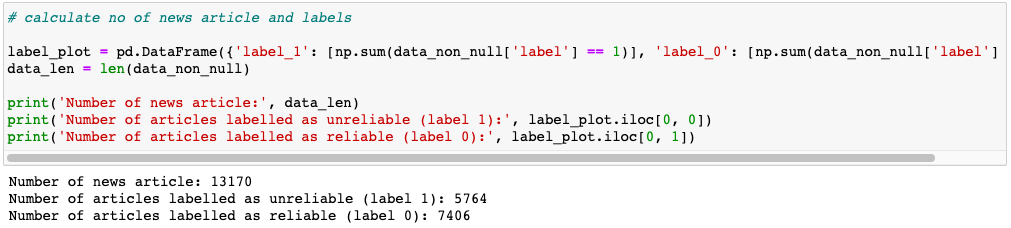
### Data Description

The training dataset contains 15000 rows and 6 columns. The 6 variables are:

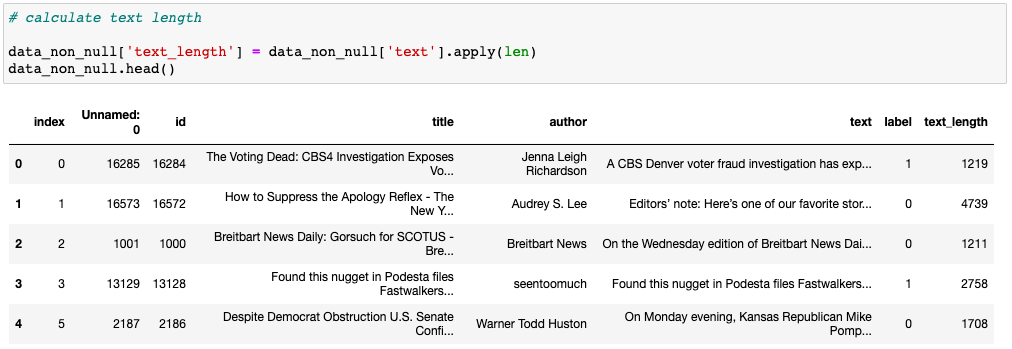
* *id*: unique id for a news article
* *title*: the title of a news article
* *author*: author of the news article
* *text*: the text of the article; could be incomplete
* *label*: a label that marks the article as potentially unreliable
  + 1: unreliable
  + 0: reliable

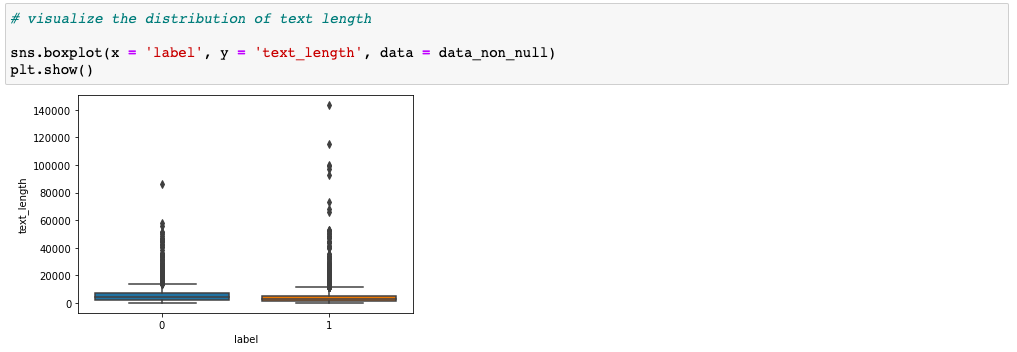
From the details of the dataset, we can see it contains a number of rows of *nan* values in the columns of title, author, and text. Other columns have integers only. Hence I remove those *nan* rows in order to let models successfully read the data for training and prediction.

After removing the *nan* values, we can now see how many rows are left and how many data belong to label 1 and 0 respectively. There are only 13170 rows of data left. The number of label 1 data is 5764 and that of label 0 data is 7406.



Next we calculate the text length of each text to see how they distribute by using the function *apply* which carries out the function of *len* for every row of the dataframe.



Using boxplot we can observe that there are some texts that are relatively long, much higher than the maximum of boxplot. The maximum length of label 0 is more than 80000 and that of label 1 is more than 140000.

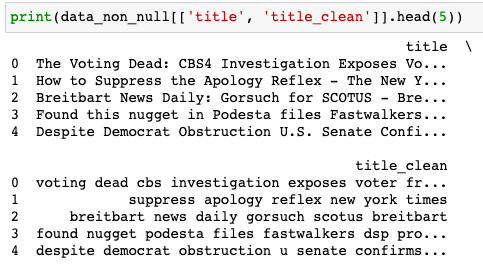
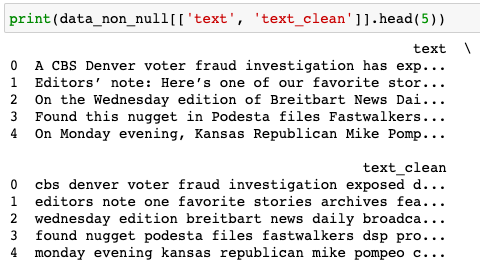
We can plot another graph to visualize the number of labels to see how they differ by a simple bar plot. It is obvious that there are more label 0 than label 1 in the dataset. The size of label 1 is slightly less than 6000 while that of label 0 is more than 7000, meaning there are more reliable articles than unreliable ones in the dataset.

### Data Preparation and Visualization

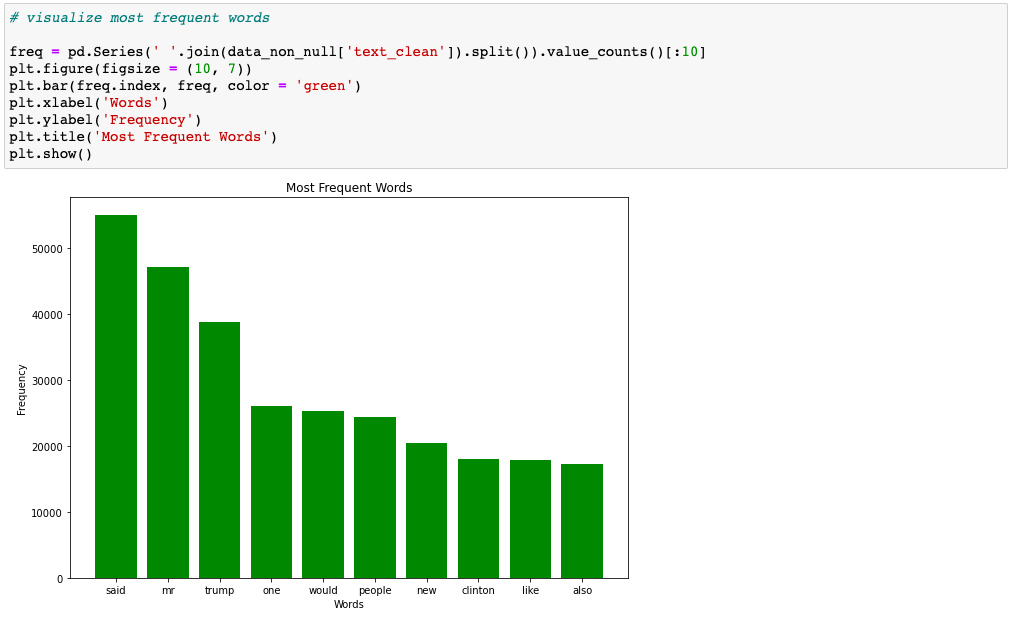
Before managing the dataset, we have to first define a function *cleaning* to remove unnecessary elements to lower the workload of the program. The stopword list is from the library Natural Language Toolkit which deals with classification, tokenization, stemming, tagging, parsing, and semantic reasoning for natural language processing.

*cleaning* removes extra space, punctuations, numbers, and capitalization with functions from built-in *str* library. We have to tokenize the text, i.e. breaking down a sentence into words in order to remove stopwords with the help of the stopword list. We combine the words again to make sentences for further actions.



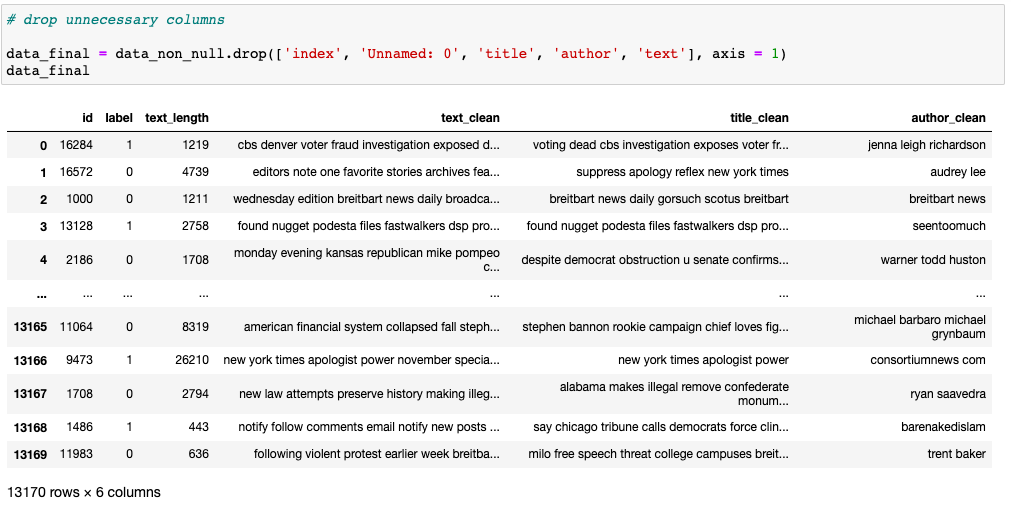
Then we can compare the results before and after cleaning for the columns text, title, and author. Evidently all the unwanted elements are removed so that we can perform analysis and prediction more easily.

We would like to see which words are the most frequent in the text. With a simple wordcloud and bar plot, we can visualize them. From the graphs, we can know that “said”, “trump”, “one”, “would”, “people” etc. are example of common words. They have each appeared for more than 20000 times.

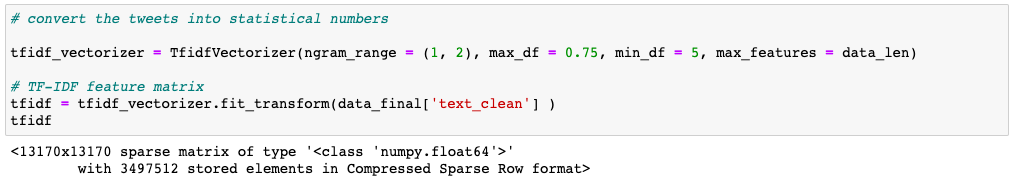
Before heading to modelling, we analyze the sentiment of the text. With the help of the powerful library TextBlob, a convenient tool for processing textual data, we obtain a general sentiment of 0.0109, which is slightly positive. TextBlob defines 1 as the most positive text while -1 as the most negative.

### Modelling

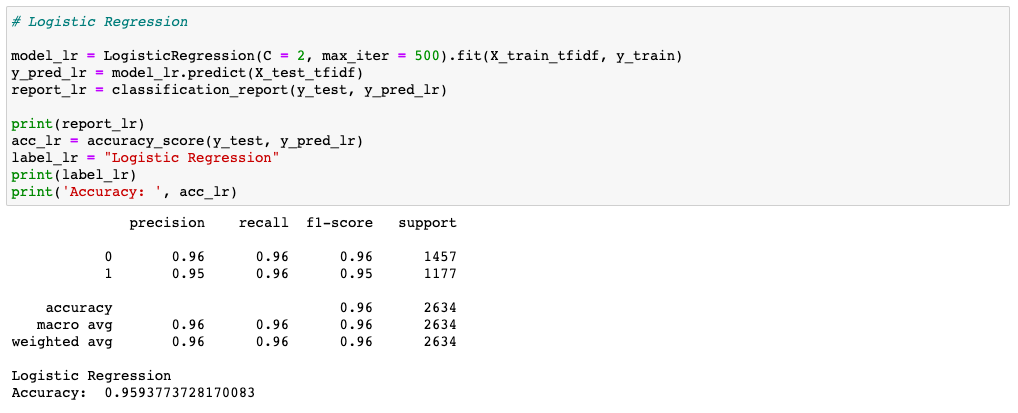
Before splitting the data, we need to drop redundant columns such as ‘index’, ‘Unnamed: 0’, ‘title’, ‘author’, and ‘text’ since we already have the cleaned data ‘text\_clean’, ‘title\_clean’, and ‘author\_clean’.



From the library sklearn, there is a function called TfidfVectorizer to convert a collection of raw documents to a matrix of TF-IDF features. TF-IDF is a statistical measure that evaluates how relevant a word is to a document in a collection of documents. We apply the vectorizer on the text to transform it into an array of numerical data for model training and testing.

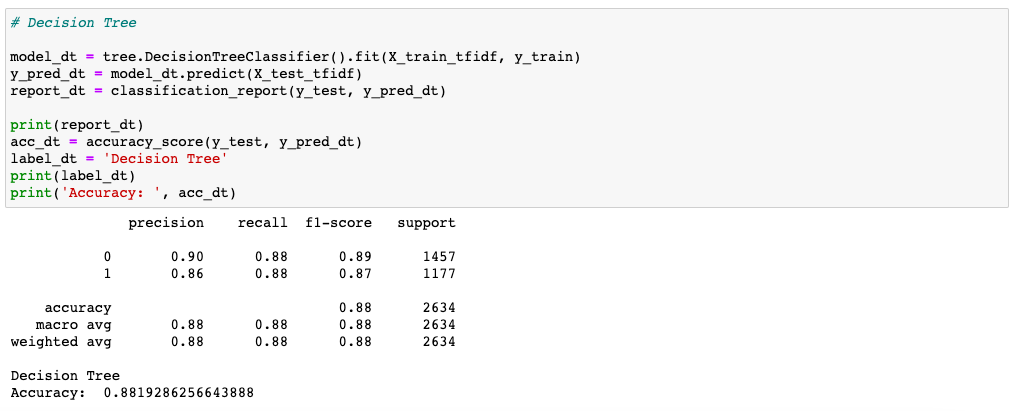
*i. Logistic Regression*

The result of logistic regression is excellent for both labels, with an accuracy of 95.94%.



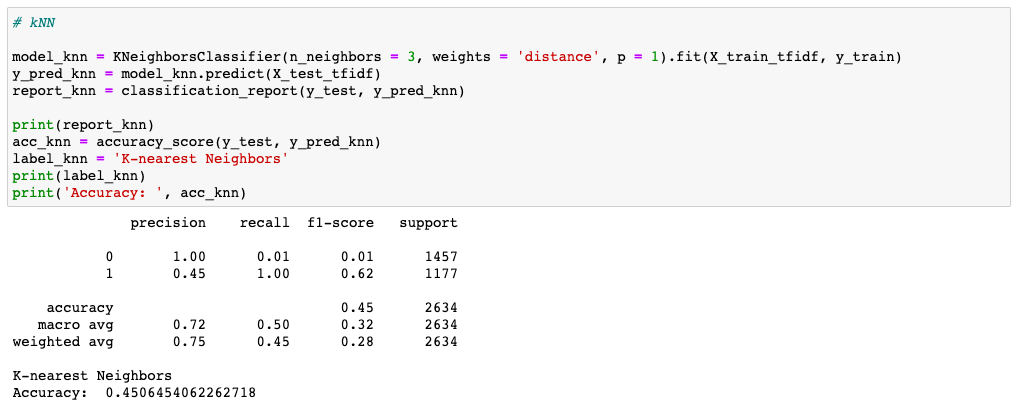
*ii. Decision Tree*

Decision tree is slightly behind logistic regression with an accuracy of 88.19% but still the result is very satisfactory.



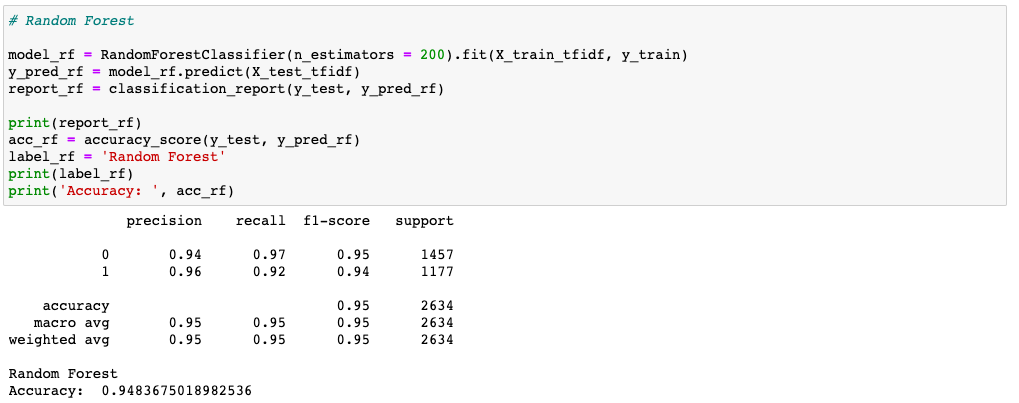
*iii. K-nearest Neighbors*

The outcome of kNN is disappointing, with an accuracy of 45.06%. It may be due to the huge size of the dataset. It is also relatively slower than the other algorithms.



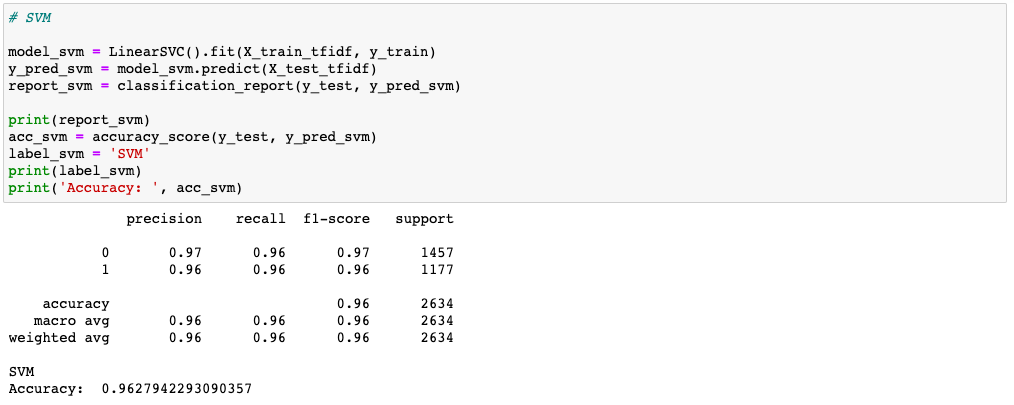
*iv. Random Forest*

Random forest also gives an outstanding result, with an accuracy of 94.84%, close to that of logistic regression.

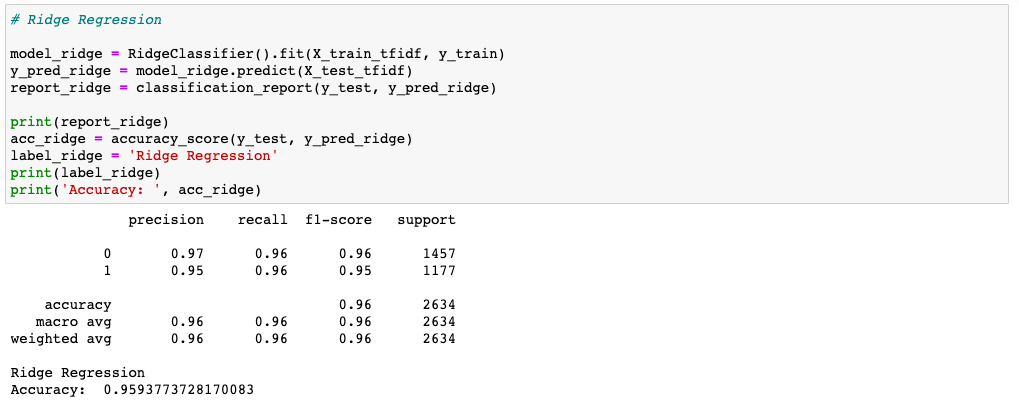


*v. Support Vector Machine*

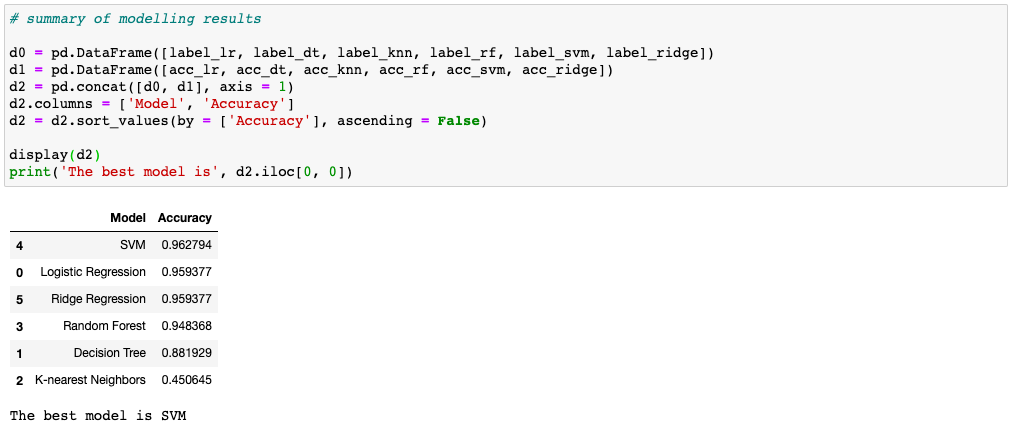
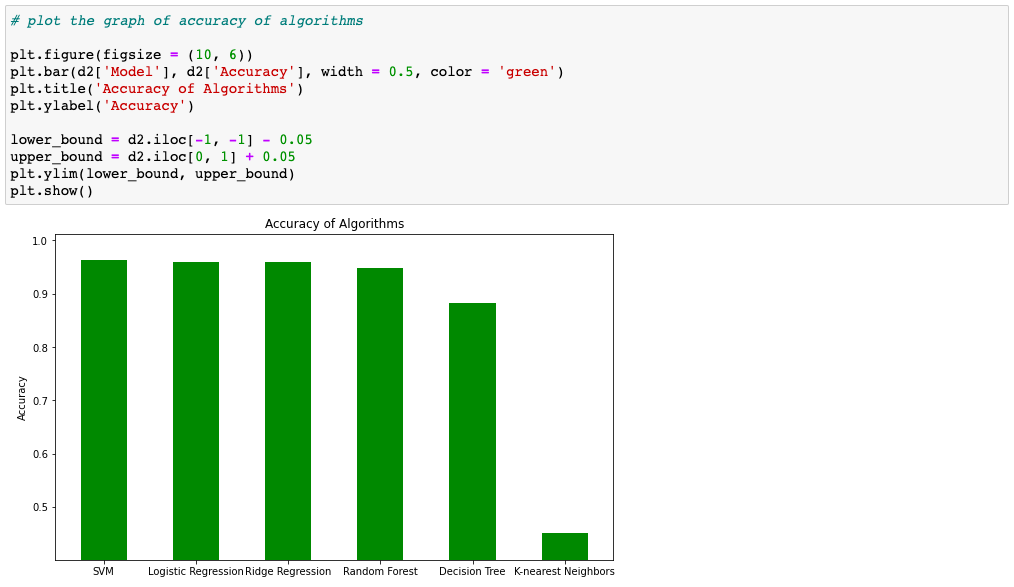
SVM is the best algorithm so far, with an accuracy of 96.28%.



*vi. Ridge Regression*

Ridge regression also yields a stunning accuracy of 95.94%.

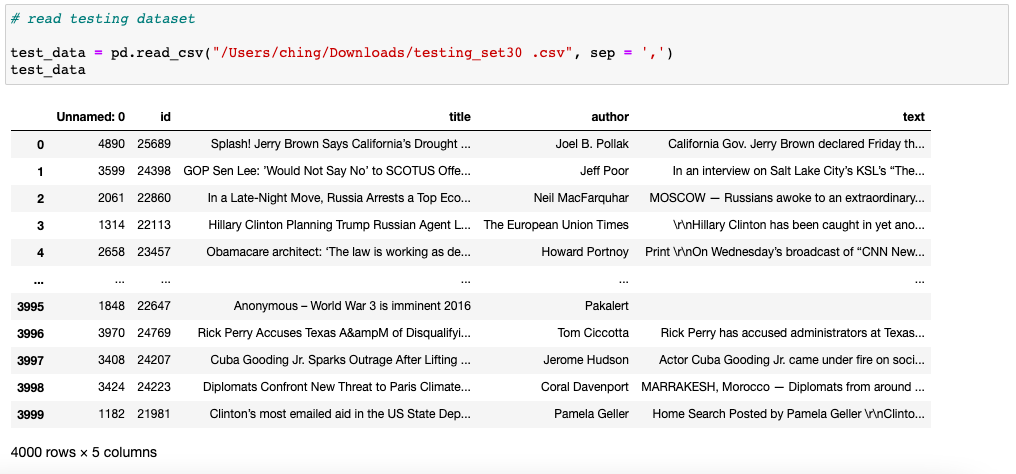
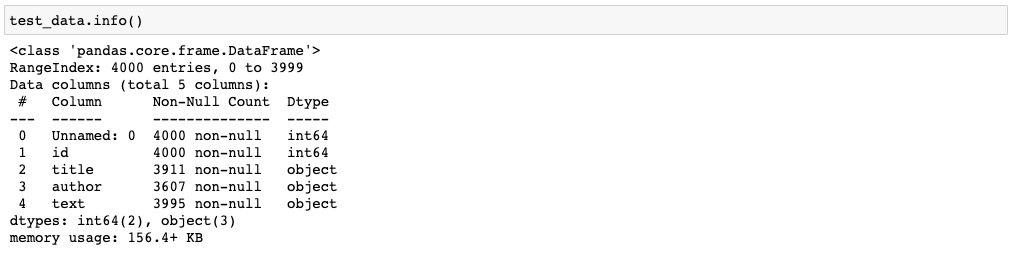
### Result

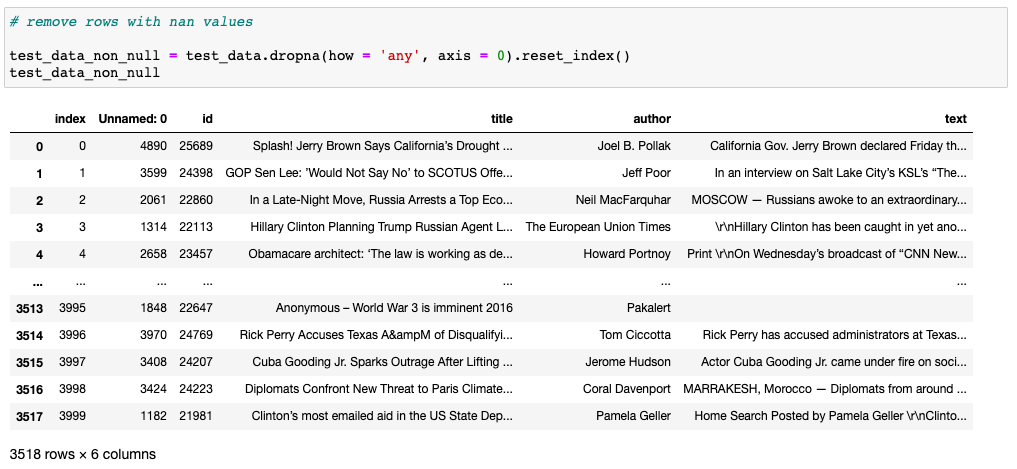
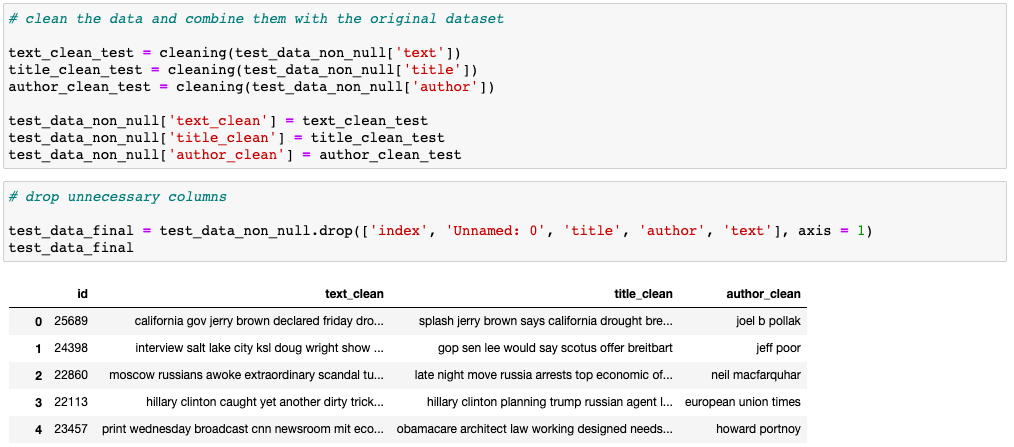
All the results are summarized in the below table and bar plot. The best model is SVM which I am going to use to predict. Other than kNN, all the other algorithms show similar accuracy, which is around 90% or more.

I have also generated a confusion matrix of SVM to take a look of how the algorithm performs. You can see SVM is a successful model to obtain a good result.

### Prediction

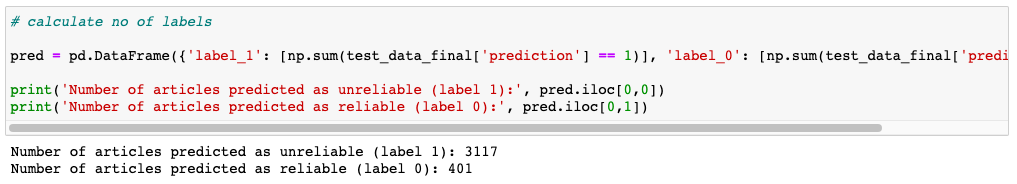
We read the testing dataset to predict. It has 4000 rows and 4 variables, namely ‘id’, ‘title’, ‘author’, and ‘text’, similar to the training dataset.

Again with a simple function we can immediately discover that some rows have *nan* values. We remove them and perform data cleaning just like before as the model can only take the same data structure.

Fit the testing dataset into the SVM model and we can immediately obtain the prediction:



The number of predicted unreliable article is 3117 and that of reliable article is 401. In contrast to the training dataset, the testing dataset has a much higher proportion of fake news as you can see from the bar plot. Since the model is highly accurate, we can conclude that the testing dataset has more unreliable news than reliable news, and higher proportion of unreliable news than that of training dataset.





### Recommendation

We all have to be careful when we read the news online. Since there is an explosive number of news media and social media, we are exposed to a sea of overwhelming sources of news. We have to always pay attention to the news source we read. Pick the credible ones such as AFP, AP, and Reuters. These are wire service companies which provide the most objective news you can find.

Other than that, we can also read news from renowned public agencies such as BBC, NPR, France 24, and DW which also offer news of quality. They are regulated by the government and that’s why they have more limitations on the news and messages they can deliver. Advertisement and propaganda are strictly banned among these outlets.

Avoid unknown online news media. You will never know how they come up with the information or whether the news is correct or not. On the other hand, we do have some impressive and well-known online news media such as Vice News and Vox but they tend to be more subjective than the aforementioned ones. Be more careful when you read news from these outlets.

### Final Thoughts

It is hard to completely avoid fake news in these days and age as we are already part of the internet. We have to always understand the news source first before absorbing the information. Fact checking can also help but can we really check every piece of news? It seems that more regulations may be a solution to this problem.