

Fake News Detection

Submitted by:

NITIN SINGH TATRARI

1825

**ACKNOWLEDGMENT**

I have studied articles on fake news and its impact through Google. I have also studies similar model on kaggle. I have studied on how we can stop its spread and what role we plan in minimizing the damage caused by fake news.

**INTRODUCTION**

* Business Problem Framing

With the advancement of technology, digital news is more widely exposed to users globally and contributes to the increment of spreading [hoaxes](https://en.wikipedia.org/wiki/List_of_hoaxes) and disinformation online. Fake news can be found through popular platforms such as social media and the Internet and even print media. The rise of fake news affects not only individuals, narratives, and the platforms that publish them, but also businesses and jobs across all manner of industries, as well. Thus this false news can adversely impact the businesses, as they can alienate their portion of target market due to false fact spread to them.

* Conceptual Background of the Domain Problem

Fake news is very harmful for businesses as it tarnishes the brand image. Even it impacts the business financially. If a business is subject to fake damning reviews, it can negatively impact consumer confidence. If less people feel confidently about the product or service, then less people are likely to spend money on it. Fake news is incredibly divisive and exacerbates polarisation. It creates unrest among employee and creates confusion, thus dividing employees.

Fake news can also be use for someone’s politically gain. These political influences can spread hatred and even violence in some instances. This situation could effects the working of business. Many businesses may inadvertently find themselves taking a political position by simply acting on the fact as they are known.

* Review of Literature

We today world, we should not believe on forwarded message at facebook, whatsapp or other social site. We should try to investigate it, try to find its sources, verify it from different websites, and then only act upon it. As this fake article, can be to just for someone personal gain like payperclick , or even political gain or business rivalries. We would also check the author for reassurance. Sometimes even sponsored content can help us. This fake news are mostly planned with a motive, for example, influencing people during election time or creating riots like situation in the country. Fake news can mislead people regarding some diseases, spreading fear or misconception among people. False news on a product, can effect it perception in public, thus leading to the loss of revenue of company and till the time company clarify it, the loss was already done. Thus, fake news can destroy the image of a brand which it has built over years just like that. Thus, this fake news has to be filtered and remove, before it makes any damages emotionally, socially and financially.

* Motivation for the Problem Undertaken

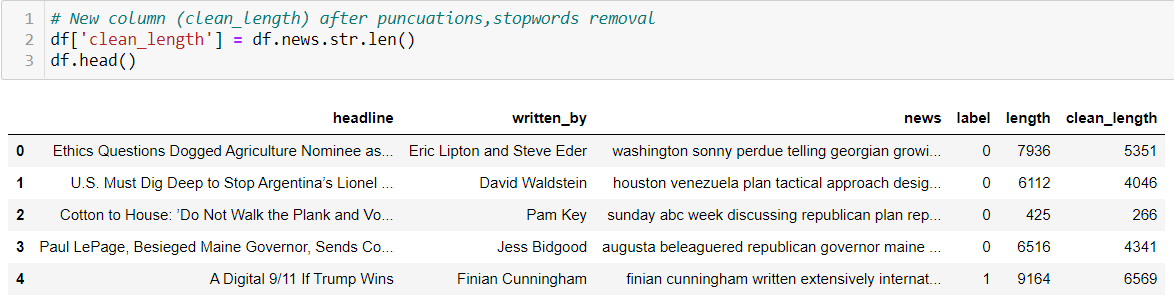
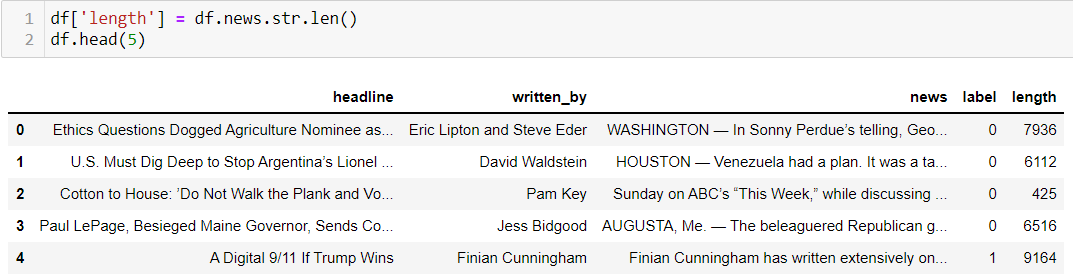
Since fake news spread confusion, hatred, losses of business, revenues, it is really concern issue. Nowadays, people devote much of their time online, for them news comes mostly from social site, not by reading newspaper, thus any fake news will effects their judgement of people and business.

This is also one of the most popular data science projects. Thus building this model will surely give me a good exposure.

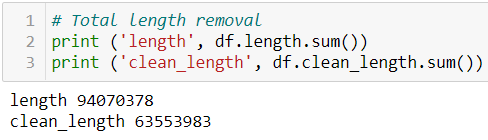
**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

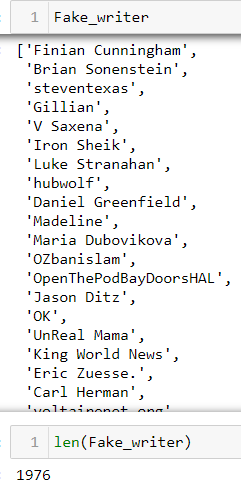
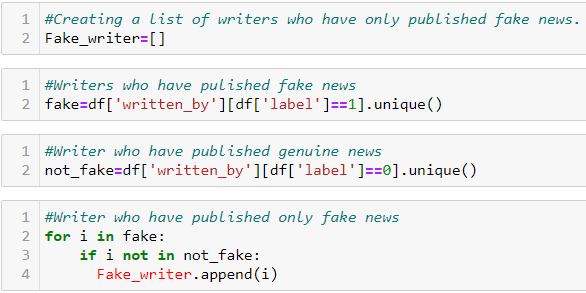
1. We have added two columns describing length of comments before and after cleaning.



We can study the length is reduced significantly after cleaning.



1. We have tried to make a list of authors who have only posted fake news.



There are total 1976 out of 4202 authors.

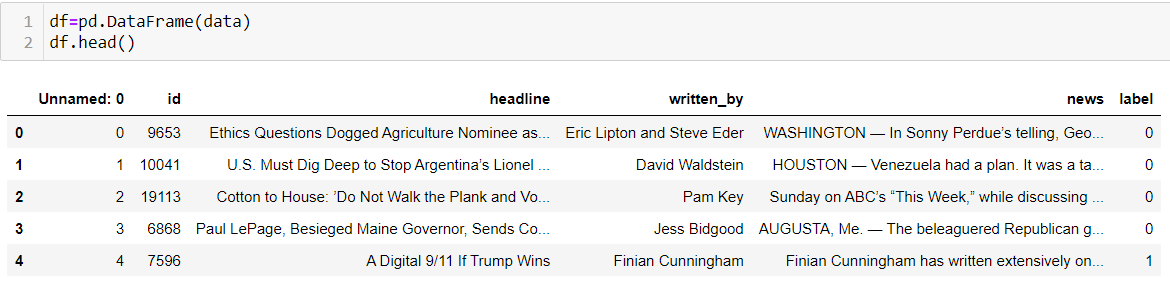
* Data Sources and their formats

What are the data sources, their origins, their formats and other details that you find necessary? They can be described here. Provide a proper data description. You can also add a snapshot of the data.

1. Uploading the data

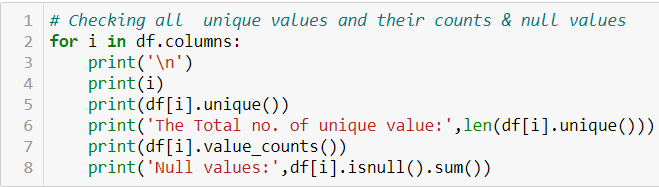
C:\Users\Nitin Singh Tatrari\Desktop\2.5.png

1. Creating Data Frame



There are 20800 rows and 6 columns in the data frame.

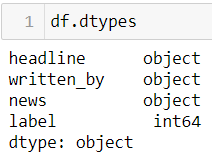
1. Checking unique values and its count & Null values



Since column ‘Unnammed: 0’ is just index column and column ‘Id’ have all unique values, we can drop them.

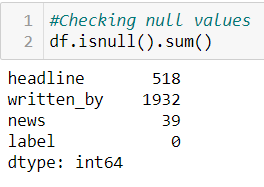


1. Checking Data type



All the variables have string data type expect target variable ‘label’ which is integer type.

1. Checking Null values



We drop all rows which have null values in ‘news’ variable.

C:\Users\Nitin Singh Tatrari\Desktop\9.png

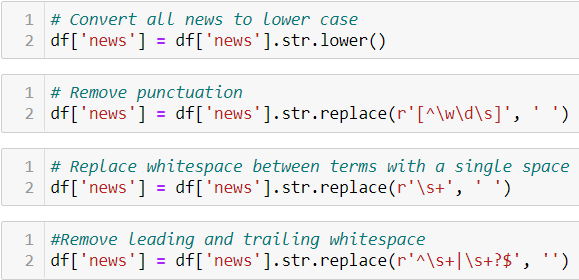
We replace all null values in ‘written\_by’ column by ‘No\_data’.



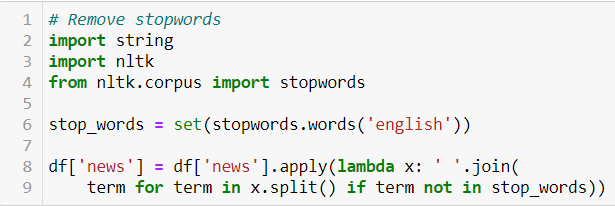
* Data Preprocessing Done

What were the steps followed for the cleaning of the data? What were the assumptions done and what were the next actions steps over that?

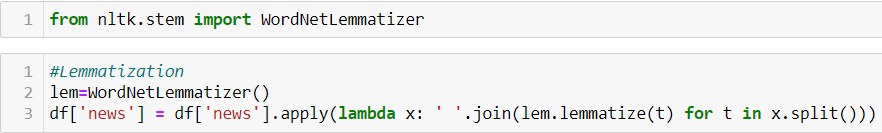
1. Converting data in ‘news’ column into lower case and removing punctuations and replacing whitespaces.



1. Removing stopwords



1. Applying lemmatization



* Data Inputs- Logic- Output Relationships

1. Term Frequency Inverse Document Frequency Vectorizer(TFIDF)

We have converted the text in ‘news’ column into meaningful representation of numbers which is used to fit machine algorithm for prediction.

* State the set of assumptions (if any) related to the problem under consideration

No assumption were taken.

* Hardware and Software Requirements and Tools Used

The libraries used are: pandas, numpy, matplotlib.pyplot, seaborn, and scikit\_learn. The laptop used is with Intel I5 10th generation, 4GB RAM, 4GB GPU.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

1. I have cleaned the data by removing punctuation, whitespace, and stop words.
2. I have lemmatized the text in ‘new’ column.
3. Then I have changed the text in news column into vector form using TF-DIF vectorizer.
4. I have trained and tested my data in the following algorithms and metrics.

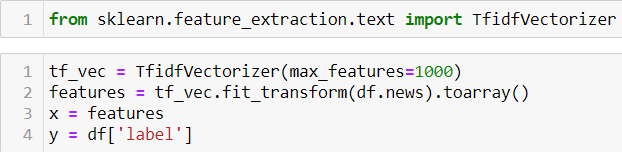
* Testing of Identified Approaches (Algorithms)

The algorithms used for testing are as follows:-

1. Term Frequency Inverse Document Frequency Vectorizer(TF-IDF)
2. Multinomial Naïve Bayes
3. Gaussian Naïve Bayes
4. Support Vector classifier
5. Logistic Regressor
6. Decision Tree classifier
7. Random Forest classifier
8. Ada Boost Classifier

* Run and Evaluate selected models

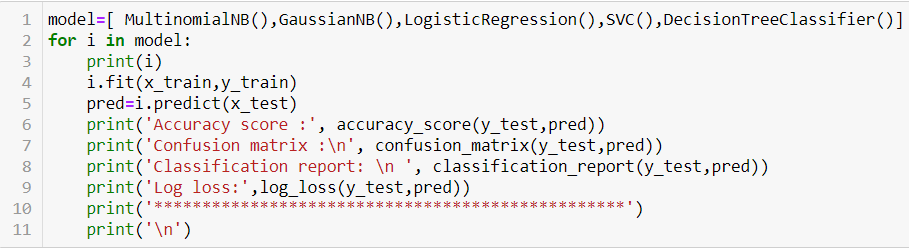
1. We convert the text in comment\_text into vectors using TF-IDF vectorizer.



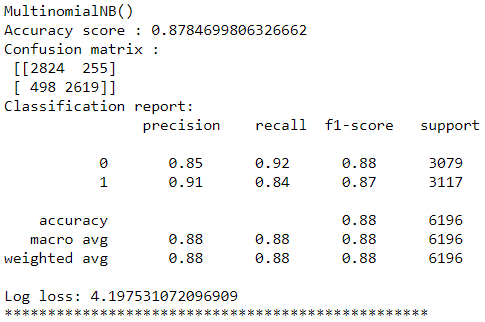
1. We split the data into train and test.



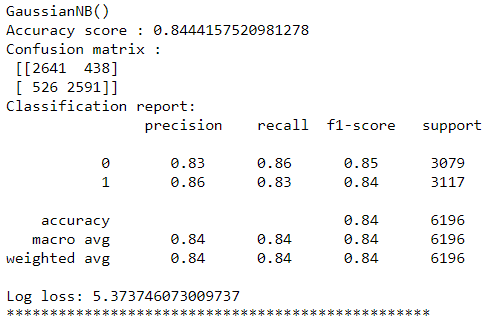
1. We have train and test the data in few algorithms and found out its accuracy score, log loss, classification report and confusion metrics.



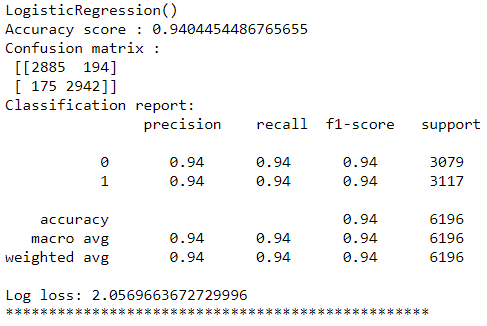
1. MultinomialNB()



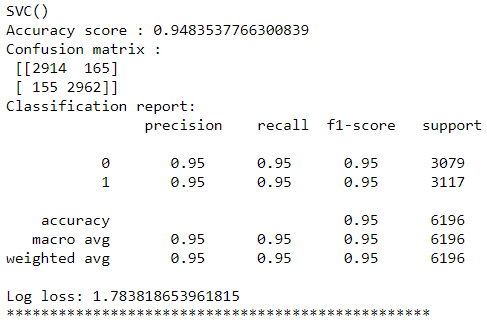
1. GaussianNB()



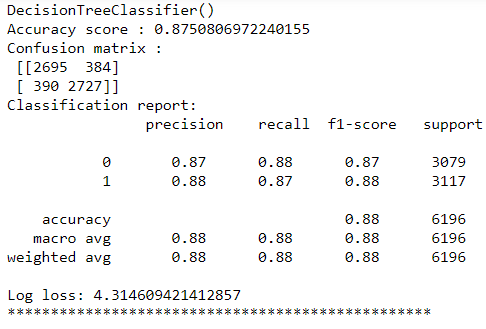
1. LogisticRegression()



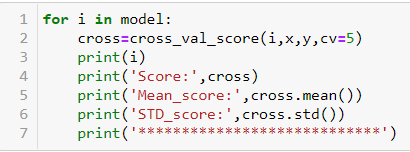
1. SupportVectorClassifier()



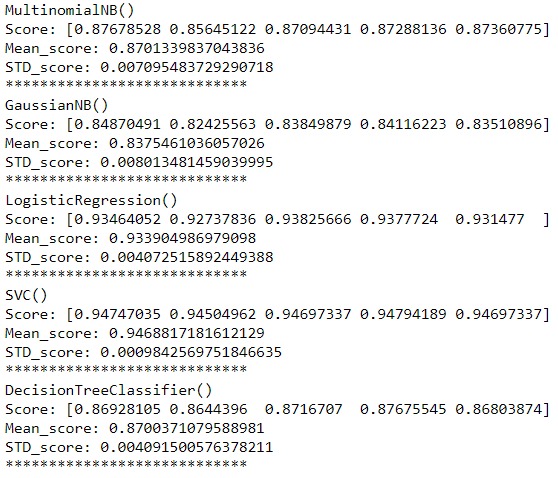
1. DecisionTreeClassifier()



1. Cross Validation

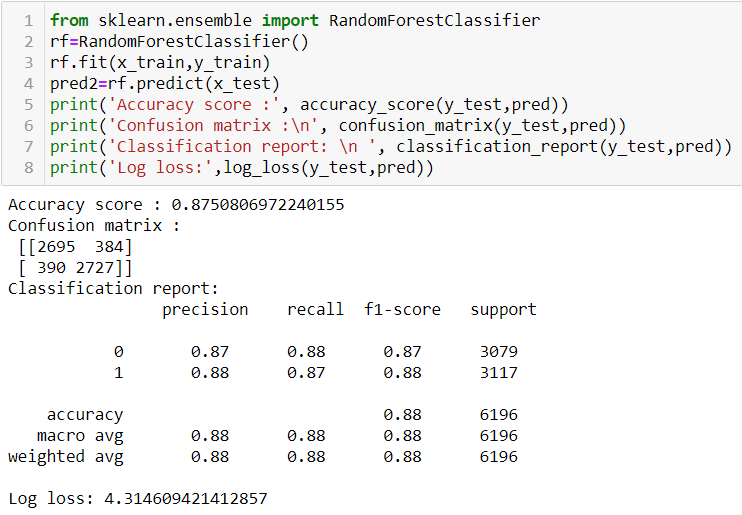


We check cross validate the accuracy of all algorithm.

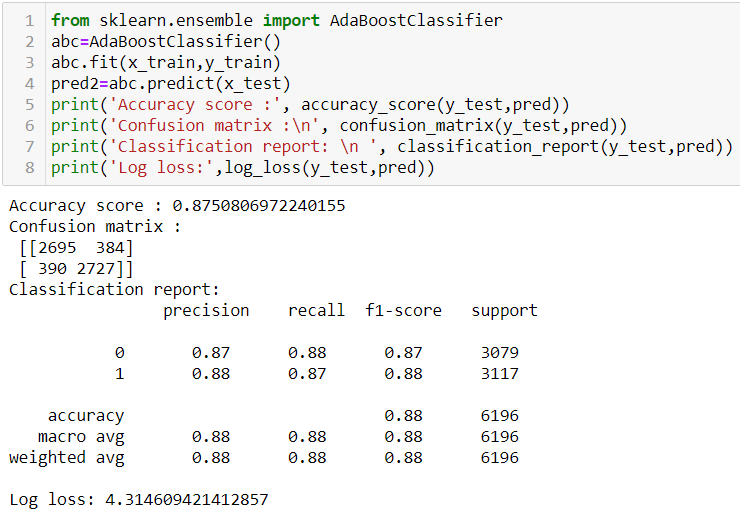


Standard deviation of all algorithms is under acceptable range.

1. Ensemble Technique
2. RandomForestClassifier()

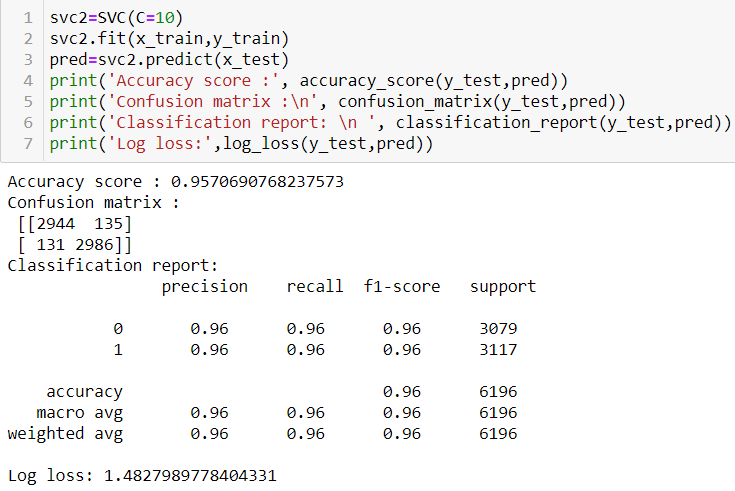
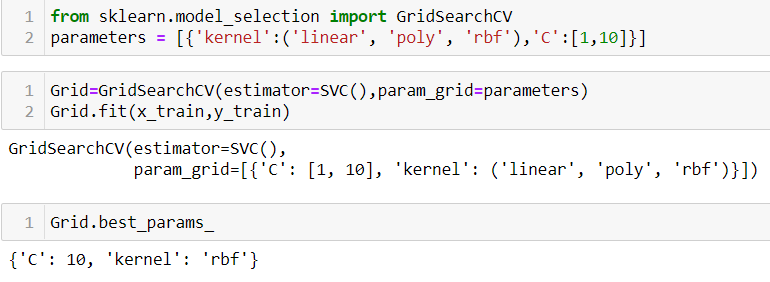


1. AdaBoostClassifier()



The best accuracy score of 94.83% is given by SVC() and log loss of 1.78.

1. Hyper Parameter Tuning



The best accuracy score of 95.7% is given by SVC() and log loss of 1.48 with parameter C=10.

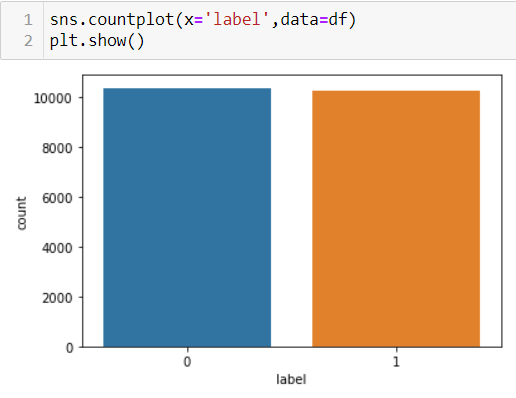
* Key Metrics for success in solving problem under consideration

The metrics used are accuracy\_score, classification\_report, confusion metrix and Log\_loss.

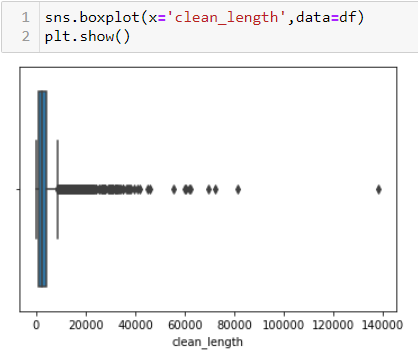
Above it is shown.

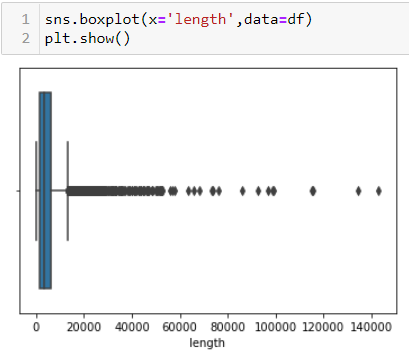
* Visualizations

1. Plotting graphs

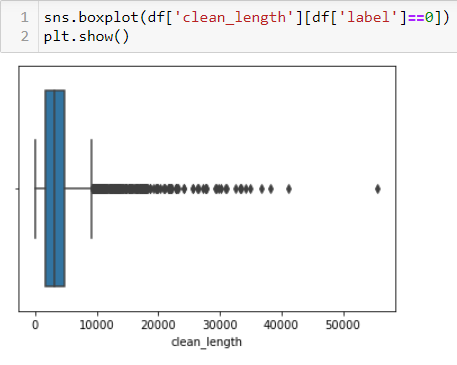


We can see data is well balanced as target variable have nearly equal ratio of 0(not fake news) & 1(Fake news)

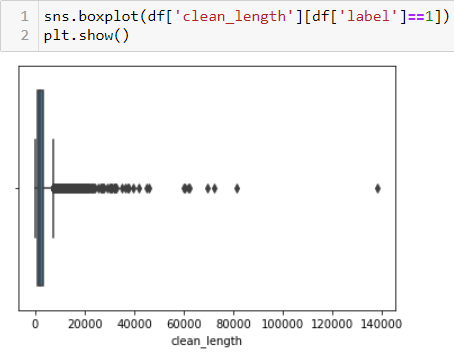




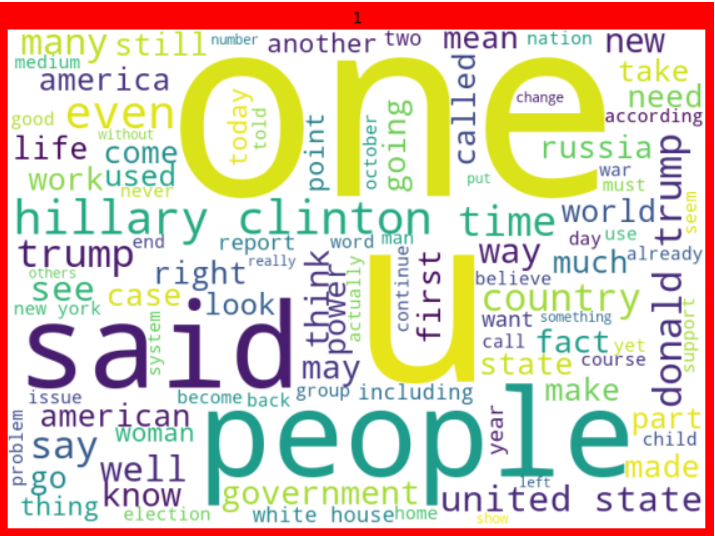
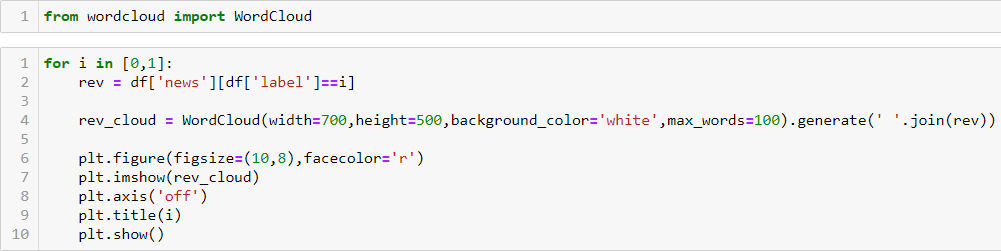
We can see the length is reduced significantly after cleaning.



We can see news which is not fake have length not more than 55000. The mean length is around 3000.



We can see fake news could have length more than 55000 lengths. The mean length is around 1000.



* Interpretation of the Results

1. The length of text was reduced considerably after applying all pre-processing steps.
2. Column ‘news’ which is not fakes have length not more than 55000.
3. We can see fake news could have length more than 55000 lengths.

**CONCLUSION**

* Key Findings and Conclusions of the Study

1. There were punctuations, stop words and white spaces which were removed to decrease the length.
2. Column ‘news’ which is not fakes have length not more than 55000.
3. We can see fake news could have length more than 55000 lengths.
4. Best algorithm is given by support vector classifier

* Learning Outcomes of the Study in respect of Data Science

Firstly the data was cleaned to removes all punctuation, whitespaces and stop words, which have no impact on classification. Then I have lemmatized the text.

Then I have converted the text into vector form for machine learning.

I have train the model with different algorithm. Support vector classifier gave the best accuracy and also gave the least log loss.

* Limitations of this work and Scope for Future Work

I would like to train the model with neural network for future work.