

Fake News Detection Project

Submitted by:

Abhishek Murarka

**ACKNOWLEDGMENT**

I would like to express my sincere gratitude to FlipRobo Technologies for supporting me throughout the internship and giving me the opportunity to explore the depth of Data Science by providing multiple projects like this, there are multiple people, organizations, youtubers, who guided me in this wonderful journey and few journals which helped me develop my models in this project. I would like to thank following people for the inspiration and help,

* FlipRobo Technologies
* DataTrained Team
* Victor Lavrenko
* Jason Schvach

Research papers and youtube videos that helped me in this projects is as follows:

* <https://data-flair.training/blogs/advanced-python-project-detecting-fake-news/>
* <https://www.sciencedirect.com/science/article/pii/S1877050918318210>
* https://www.youtube.com/watch?v=TJU8NfDdqNQ&ab\_channel=VictorLavrenko

**INTRODUCTION**

* Business Problem Framing

The project is concerned with identifying a solution that could be used to detect and ﬁlter out sites containing fake news for purpose of helping users to avoid being lured by clickbait. It is imperative that such solutions are identiﬁed as they will prove to be useful to both readers and tech companies involved in the issue.

* Conceptual Background of the Domain Problem

The idea of fake news is not a novel concept. Notably, the idea has been in existence even before the emergence of the Internet as publishers used false and misleading information to further their interests. Following the advent of the web, more and more consumers began forsaking the traditional media channels used to disseminate information for online platforms. Not only does the latter alternative allow users to access a variety of publications in one sitting, but it is also more convenience and faster. The development, however, came with a redeﬁned concept of fake news as content publishers began using what has come to be commonly referred to as a clickbait

* Review of Literature

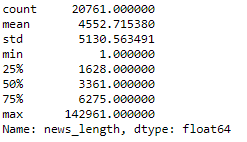
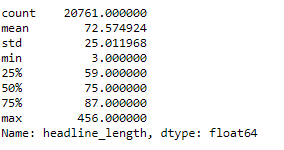
If we look at some scholar work shows the issue that the fake news has been major concerned amongst scholar from various background. For instance, some authors have observed that fake news is no longer a preserve of the marketing and public relations departments. Instead there is a increasing risk of IT security, therefore, IT department is premised on the idea that it would help avert the various risks associated with the problem. So, if we good deeply into it we could find that the hackers use clickbait with the help of fake news and make some professional of the organization downloads their malicious exploits in their system or leak sensitive information, albeit in an indirect manner. The user may, for instance, be tricked into believing that they are helping to disseminate the news further when, in the actual sense, they are providing the perpetrators with access to their emails, and we can also see that the fake news are worked extensively as they are using videos with original massage and uses their facial structure to replace the massage with false massage they want us to believe, these fake news issues is bigger day by day and we need to implement more our research and extensive knowledge to solve the problem.

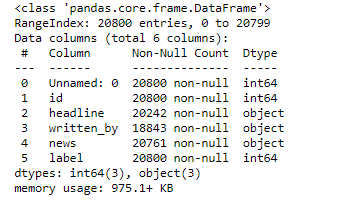
* Motivation for the Problem Undertaken

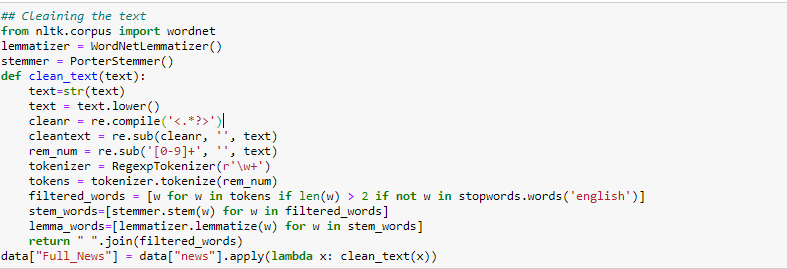
This project was highly motivated project as it includes the real time problem of fake news which if we see are getting bigger, as there various concern as people do good things work hard to build a reputation, and only one false news is enough to ruin it all, it also have inverse effect on the financial market as if we observe there will a good amount of fluctuation on stock markets based on news.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

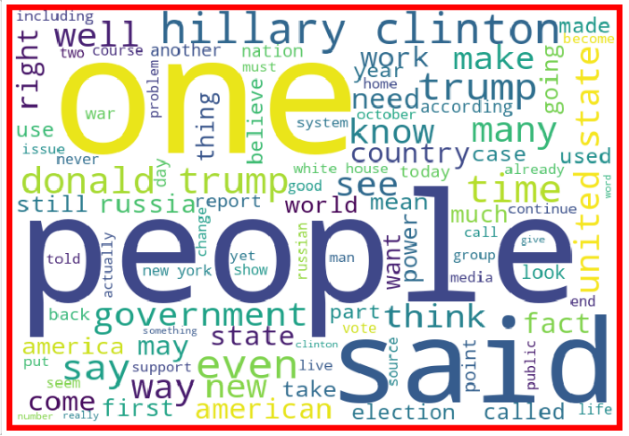
 From the above we can see that length of headlines and news and their statistical summary, so left side table is for headline and we can see that the average length of headlines is 72 and minimum length is 3 and max length is 456, and on right side table is for news, and we can see that the average is 4552, and minimum length is 1 and maximum length is 142961, so some news can be genuine, so we can’t consider them outliers.

* Data Sources and their formats
* Data Preprocessing Done



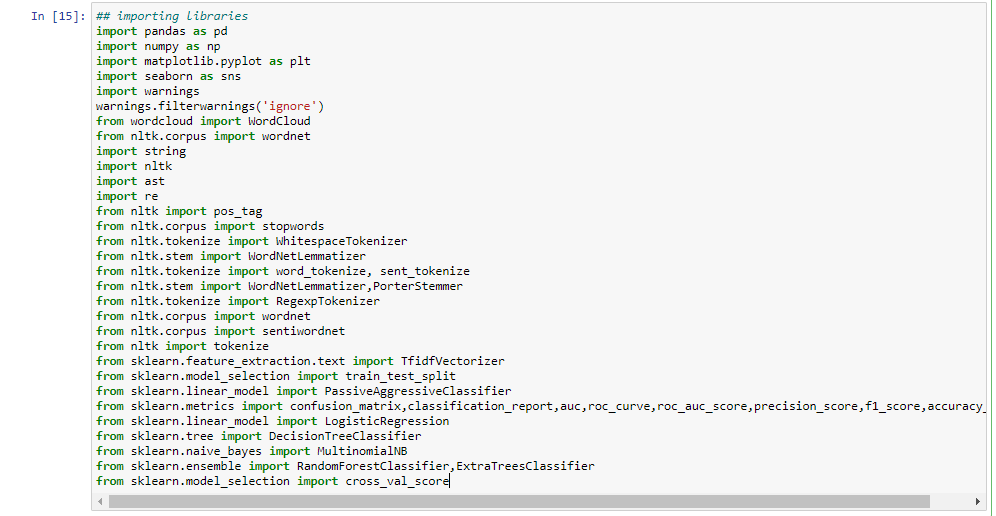
For Data pre-processing we did some data cleaning, where we used wordNet lemmatizer and porterStemmer to clean the words and removed special characters using Regexp Tokenizer and filter the words by removing stop words and then used lemmatizers and joined and return the filtered words.

Used TFIDF vectorizer to convert those text into vectors, and split the data and into test and train and trained various Machine learning algorithms.

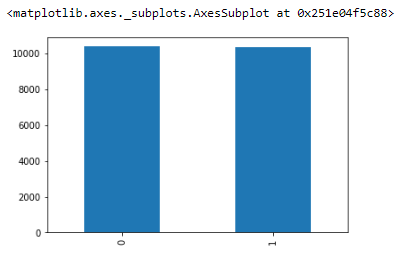
* Data Inputs- Logic- Output Relationships

  
From the above we can see that most frequent words on both labels and we can observe the words which are leading to fake new and real news, so from the left side we can see the wordcloud of fake news, and on the right hand side we can see that the wordcloud of real news.

* Hardware and Software Requirements and Tools Used
* Hardware: 8GB RAM, 64-bit, i7 processor.
* Software: Excel, Jupyter Notebook, python 3.6.
* Libraries used:



**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (

From the above we can see that the dataset is balanced which is good as it will help our model to classify more accurately, so we should expect good accuracy score, and as the volume of data was also good, so was able to implement Passive Aggressive Algorithms.

* Testing of Identified Approaches (Algorithms)
* LR=LogisticRegression()
* DT=DecisionTreeClassifier()
* GNB=MultinomialNB()
* RFC=RandomForestClassifier()
* ETC=ExtraTreesClassifier()
* PAC=PassiveAggressiveClassifier()
* Run and Evaluate selected models

Describe all the algorithms used along with the snapshot of their code and what were the results observed over different evaluation metrics.

* Key Metrics for success in solving problem under consideration

What were the key metrics used along with justification for using it? You may also include statistical metrics used if any.

* Visualizations

Mention all the plots made along with their pictures and what were the inferences and observations obtained from those. Describe them in detail.

If different platforms were used, mention that as well.

* Interpretation of the Results

Give a summary of what results were interpreted from the visualizations, preprocessing and modelling.

**CONCLUSION**

* Key Findings and Conclusions of the Study

Describe the key findings, inferences, observations from the whole problem.

* Learning Outcomes of the Study in respect of Data Science

List down your learnings obtained about the power of visualization, data cleaning and various algorithms used. You can describe which algorithm works best in which situation and what challenges you faced while working on this project and how did you overcome that.

* Limitations of this work and Scope for Future Work

What are the limitations of this solution provided, the future scope? What all steps/techniques can be followed to further extend this study and improve the results.