

FAKE NEWS DETECTION - NLP

Submitted by:

Narendran Suresh

**ACKNOWLEDGMENT**

I would like to take the opportunity to thank the organisation ‘FlipRobo Technologies”, “Datatrained “and my mentor Ms. Astha Mishra for their immense support and guidance for making this feat achievable. I would also like to thank my dad Mr. Suresh Kumar S who is a statistics subject expert for guiding me along the right path and techniques to be used.

FlipRobo Technologies who are specialised in making ML / AI models provided me the Data set as a part of my internship.

References were made to several articles among Medium, KdNuggets, towardsdatascience, realpython, machinelearningmastery, python and sklearn documentation for the successful completion of the project.

**INTRODUCTION**

* Business Problem Framing

To develop a natural language processing model to distinguish between fake news and actual news. With the current developments in technology and automation techniques, generation of system generated news is a common problem.

* Conceptual Background of the Domain Problem

With the current developments in technology and automation techniques, generation of system generated news is a common problem. Fake news can be very misleading and has a potential to create a lot of chaos between the public. Hence, it is necessary to develop a model to distinguish between them.

* Review of Literature

TF IDF Vectorization and Passive aggressive Classifier literature work was done to get the achieved result.

* Motivation for the Problem Undertaken

A classification model which can classify between fake news and actual news is a real necessity as the amount of fake news is increasing with the advancement in technology.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

Language processing techniques were used to clean the data and remove the stop words and common and unnecessary data.

Vectorization and transformation of text to machine understandable form was done using TF IDF vectorizer and transformer.

* Data Sources and their formats

Data for the scope of the project was provided by Flip Robo. The intend is to create a model to distinguish between the news.

* Data Preprocessing Done

Null data was treated and removed.

The unnecessary columns were removed namely written\_by, unnamed:0 and id.

Trailing and leading spaces and unnecessary spaces between words were removed.

Stop words were removed from the news and headline column, and both were merged to form a single feature.

* Data Inputs- Logic- Output Relationships

The headline along with the main news merged into a single feature was the only data where the processing and learning occurs.

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

N/A

* Hardware and Software Requirements and Tools Used
* Hardware:
* Inter Core (i7) – 5500U, clock speed at 2.40GHz
* RAM – 12.0 GB
* Software:
* Jupyter Notebook (Anaconda 3) – Python 3.7.6
* Microsoft Excel

Libraries & Packages used – Pandas, numpy, sklearn, matplotlib, seaborn, sklearn, scipy, imblearn, tensorflow, keras, Image Data Generator, Maxpool2D

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

Cleaning the text and removing the unnecessary stop words, vectorization and training the models for performance.

* Testing of Identified Approaches (Algorithms)
* Run and Evaluate selected models

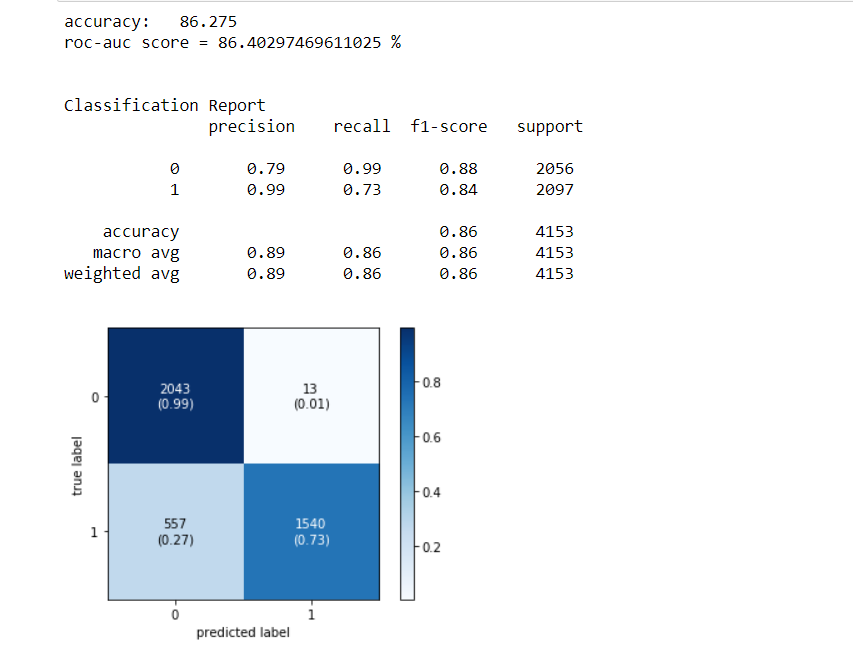
The below models were checked for performance

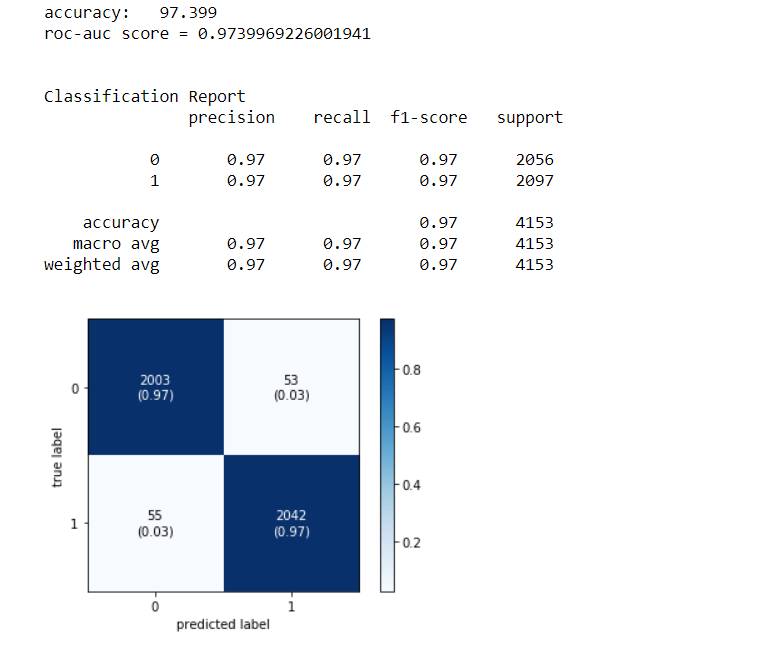
Multinomial NB

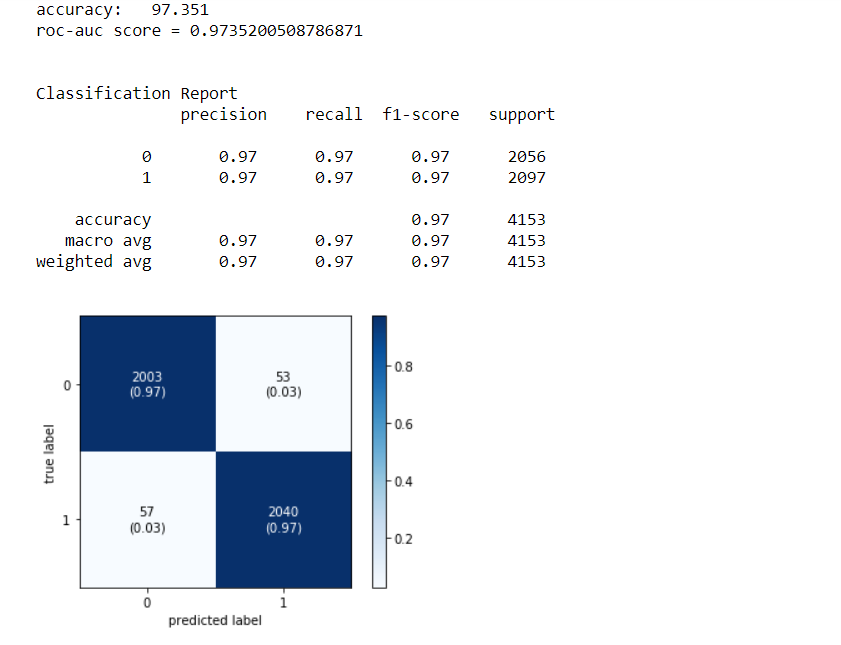
Linear SVC

Passive Aggressive Classifier

The performance metrics values are as mentioned below:





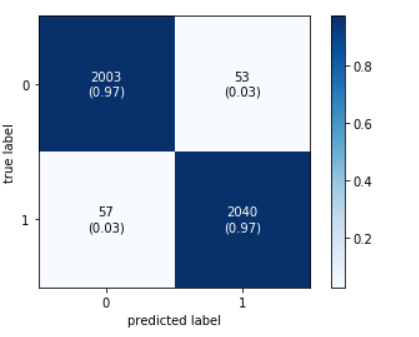


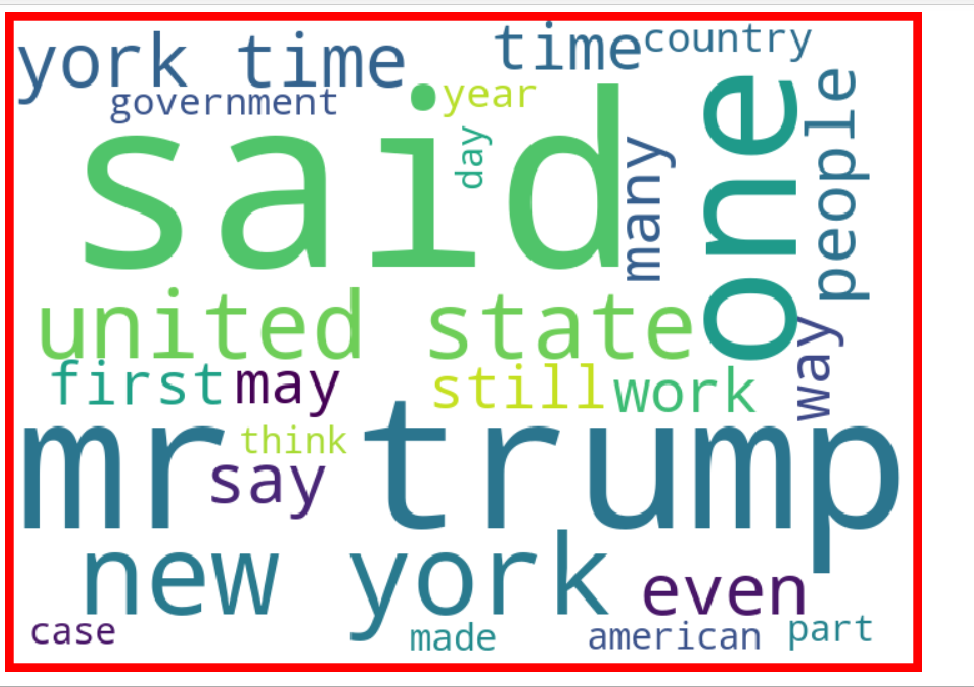
* Key Metrics for success in solving problem under consideration

Key metrics were accuracy, confusion matrix, precision, recall and F1-score

* Visualizations

The confusion matrix of the model shows that the model in hand is a very dependable one and can be deployed for solution to the problem.





* Interpretation of the Results

The model does a very good job in classifying the news categories and can be used for predictions.

**CONCLUSION**

* Key Findings and Conclusions of the Study

There were a lot of fake news about trump, new York, USA and so was the case with original news, probably due to the dataset.

* Learning Outcomes of the Study in respect of Data Science

Tfidf transformer and data pipelines were new topics learnt in order to attain completion of the project.

* Limitations of this work and Scope for Future Work

The data for some categories were limited and hence model may not perform as well if more data is added.