

Fake News Classification using Machine Learning

Submitted by:

HARSHVARDHAN PALAWAT

**ACKNOWLEDGMENT**

I want to extend my sincere regards to the below mentioned sources and references who helped me a lot in completion of my Project:

Team FlipRobo

Team DataTrained

scikit-learn official documentation

<https://scikit-learn.org/stable/>

geeksforgeeks

https://www.geeksforgeeks.org/

programiz

[https://www.programiz.com](https://www.programiz.com/)

Machine Learning Mastery

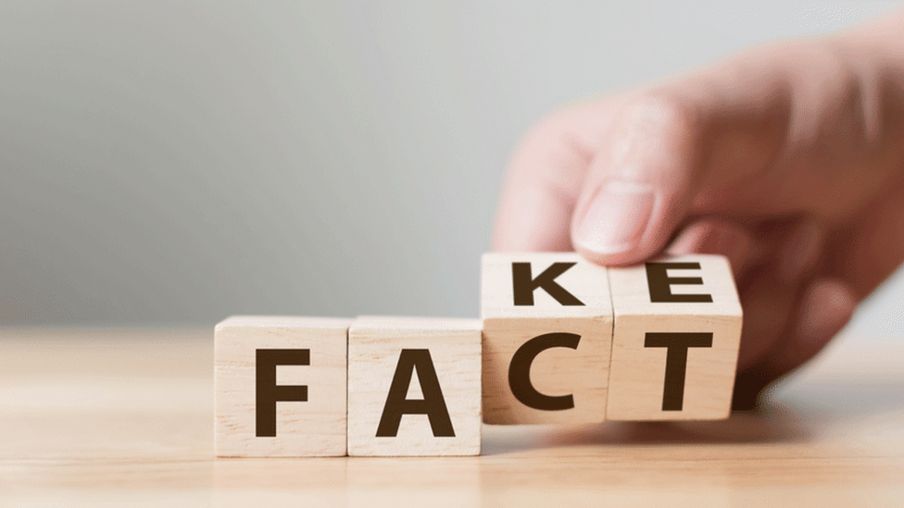
<https://machinelearningmastery.com/>

Medium

[https://www.medium.com](https://www.programiz.com/)

**INTRODUCTION**

* Business Problem Framing
* Mobile phone has a huge impact on our daily lives. It has taken over conventional methods of connectivity, entertainment as well as information distribution. A lot of people rely on their mobile phones to get the latest news. Lot of new websites have emerged which provide news in written form in real time. But amidst this, a lot of websites also run fake news for their own agendas or monetary benefits. It has become really important to identify these fake news along with their sources. In this project, we'll see how the power of Machine Learning can be leveraged to classify a news into Real or Fake.

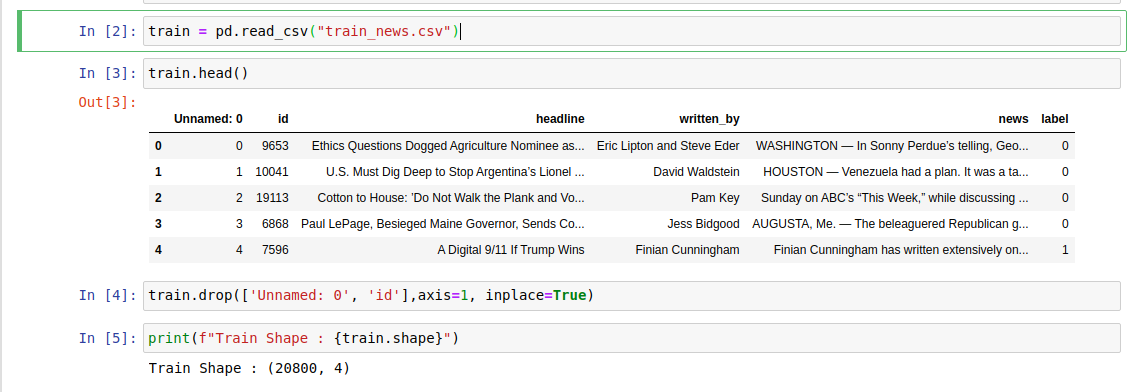


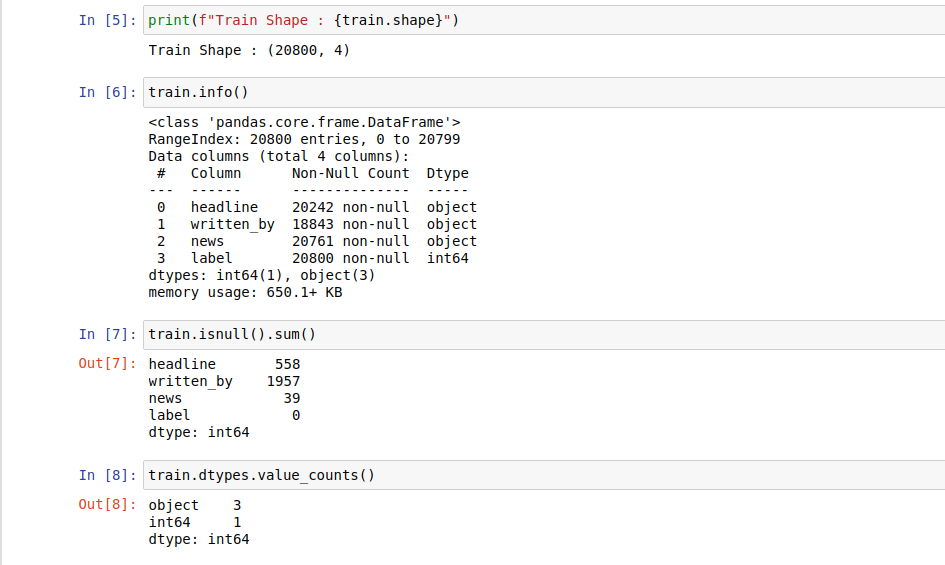
* Conceptual Background of the Domain Problem
* The domain of this project is Natural Language Processing. We want to leverage the power of Machine Learning to classify a news as real or fake out of pool of news available on internet. Here, we are given with the labelled data classified as fake or real. We are given with news headlines along with the full text, but we can not use text features in model building. Therefore, it is very necessary to have a deep understanding of how we can vectorize the data using concepts like Bag of Words, TF-IDF, Word2Vec etc.
* Review of Literature
* As the problem at hand is related to Natural Language Processing, it is very important to study how can human language be converted into mathematical form that machines could understand. The word embeddings is the most important part of this project. We did a lot of research on methods of tokenization such as Bag of Words, TF-IDF, Word2Vec and the comparison between listed methods. Also, a deep research on how state of the art techniques (like LSTMs) can leverage the sequence information present in text was conducted.
* Motivation for the Problem Undertaken

Information and facts have the power to change the whole world. It is really important to remain informed of what is going on around you. It is a natural tendency of human beings to know what is happening and the best source for this is News. Some people have converted this into business and they run Fake news either for TRP or other monetary gains. With the help of this project, we want to make a system that can identify such fake news and provide the readers with authentic information.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem
* First, we used mathematical techniques like Bag of Words, TF-IDF and Word2Vec to convert news text and headlines into numerical vectors because we can not leverage text features directly in machine learning. Machine learning models like Multinomial Naive Bayes, Logistic Regression and Passive aggressive classifier were used because they are good in dealing with high-dimensional text data. Also, state-of-the-art LSTMs were used so that we can leverage the sequence information present in news text.
* Data Sources and their formats
* The client provided the data in CSV format. The data contains total of five columns along with the target column. Below snapshots give a brief insight into what data looks like and what are the data types.

****

****

* Data Preprocessing Done
* There are null values in the data and we filled all the null values with ' '. We also curated two new features where we combined the headlines with text in one and created a length based feature in another. We used Bag of Words and TF-IDF to convert the comment text into vector.
* Data Inputs- Logic- Output Relationships
* This a classification task having 2 labels. The data is balanced. We have used Logistic Regression, Multinomial Naive bayes and Passive Aggressive Classifier to logic the input-output relationships.
* State the set of assumptions (if any) related to the problem under consideration
* Here, we have assumed that the data is missing completely at random. We have also used the multinomial naive bayes for modelling and the assumptions associated with it are the independence of features and multinomial distribution of the data.
* Hardware and Software Requirements and Tools Used
* The size of data is very small, therefore any system running on Windows 7 or higher, Mac or Linux based operating systems with 4 GB of RAM is more than sufficient for the given task. We can use any Python IDE or Jupyter notebooks or Google Colab for modelling.

Below is the list of tools used for the task:

sklearn for model building,

pandas for reading and manipulation of data,

numpy for numerical operations,

matplotlib and seaborn for data visualization

NLTK (TF-IDF) and gensim (Word2Vec) for tokenization and text pre-processing

tensorflow and keras for modelling the LSTM

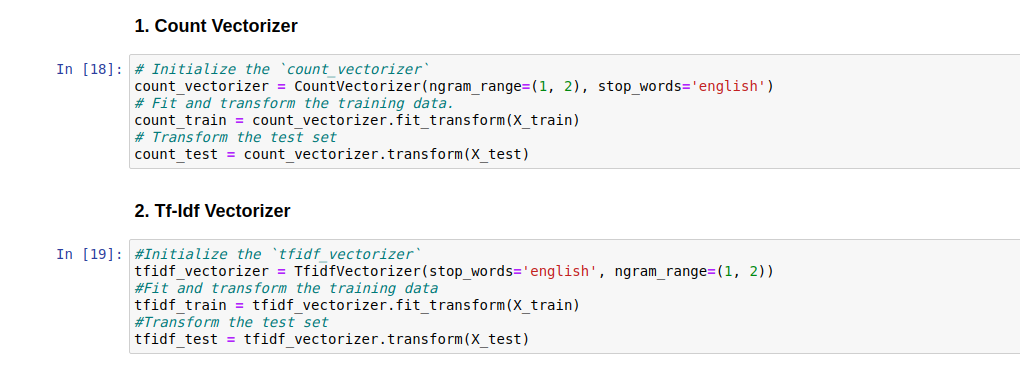
joblib for saving the model

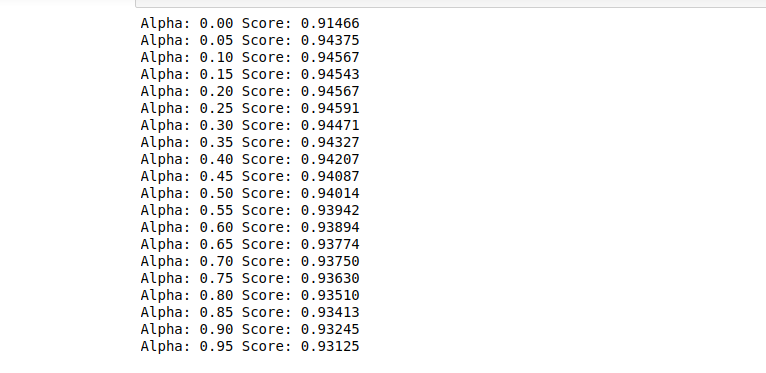
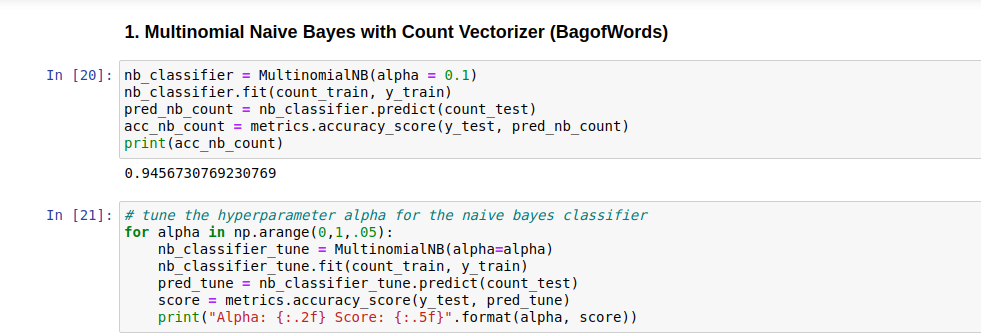
**Model/s Development and Evaluation**

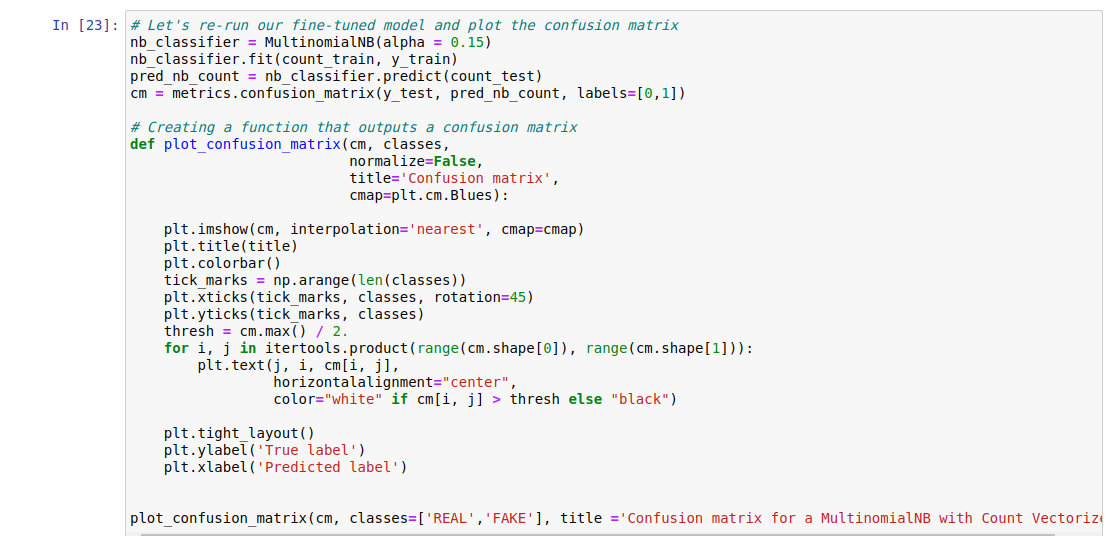
* Identification of possible problem-solving approaches (methods)
* We have used Bag of Words, TF-IDF and Word2Vec for tokenization, Logistic Regression, Multinomial Naive Bayes, Passive Aggressive classifier and Bi-directional LSTMs for modeling.

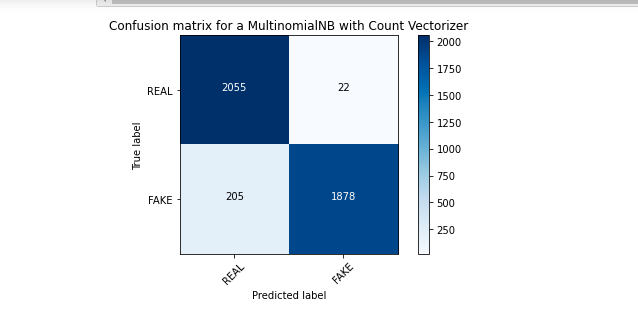
* Testing of Identified Approaches (Algorithms)
* The following algorithms were used for training and testing:
* Logistic Regression,
* Multinomial Naive Bayes,
* Passive Aggressive classifier and
* Bi-directional LSTMs
* Run and Evaluate selected models

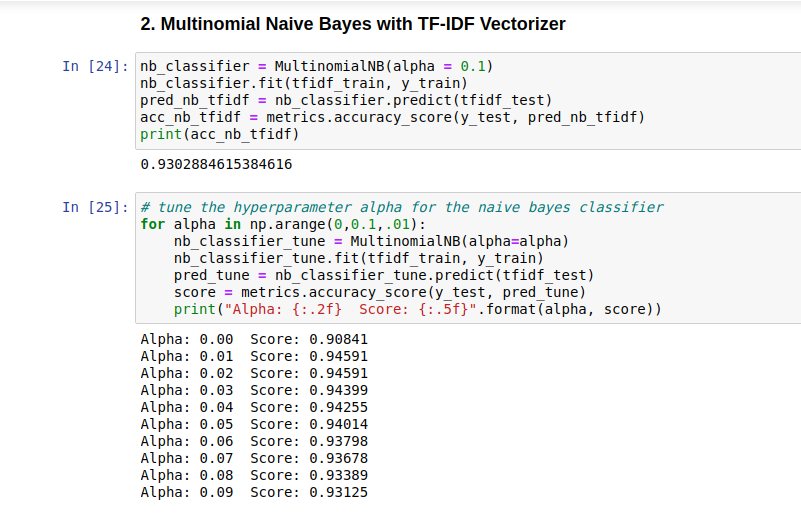
Below are the snapshots of the code used for modeling along with the results:

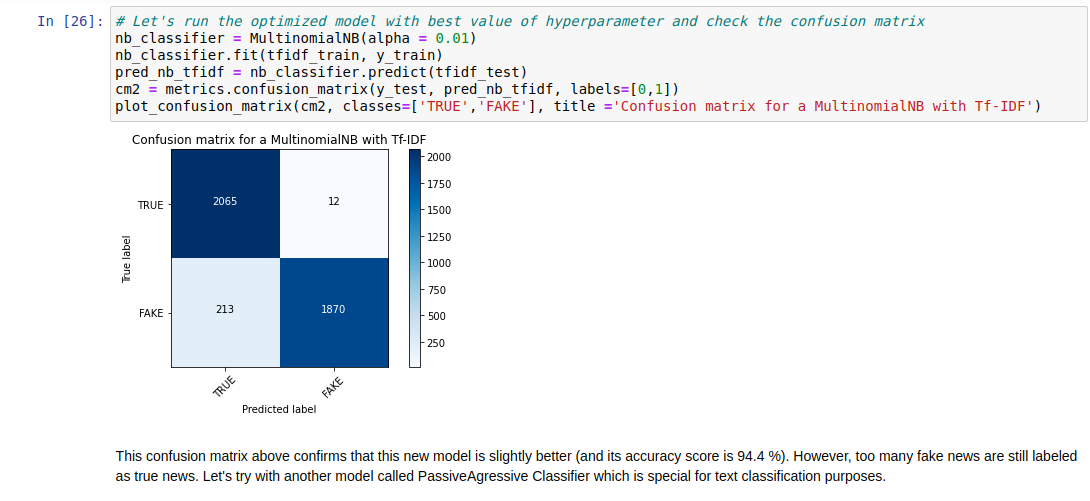


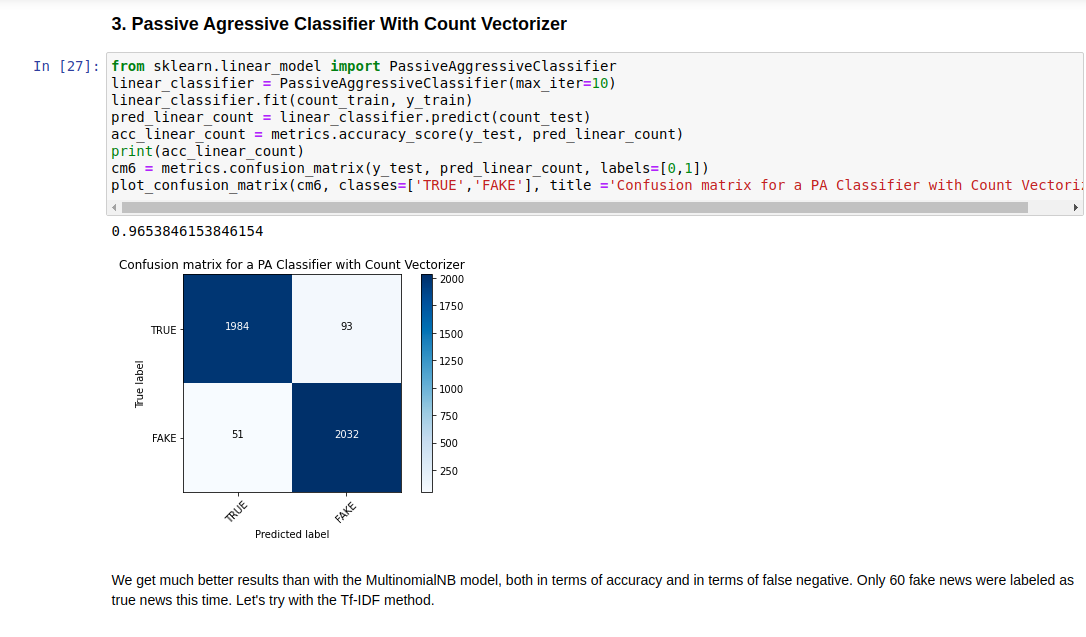


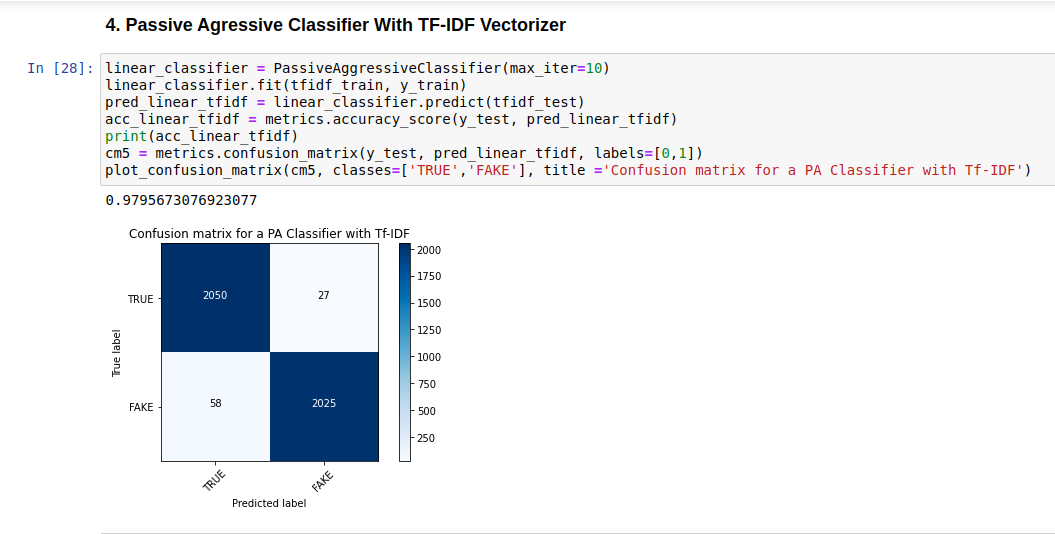


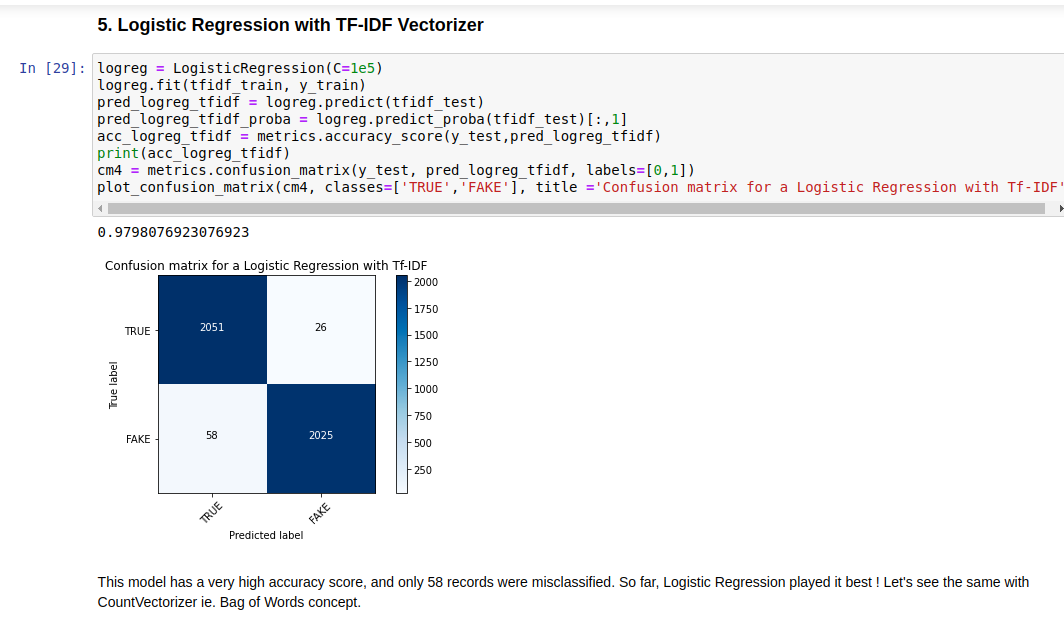


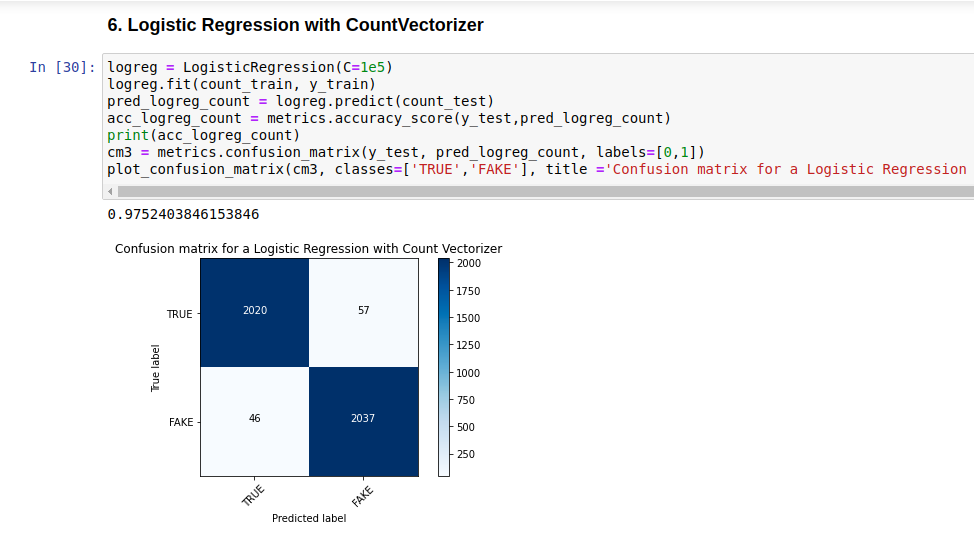


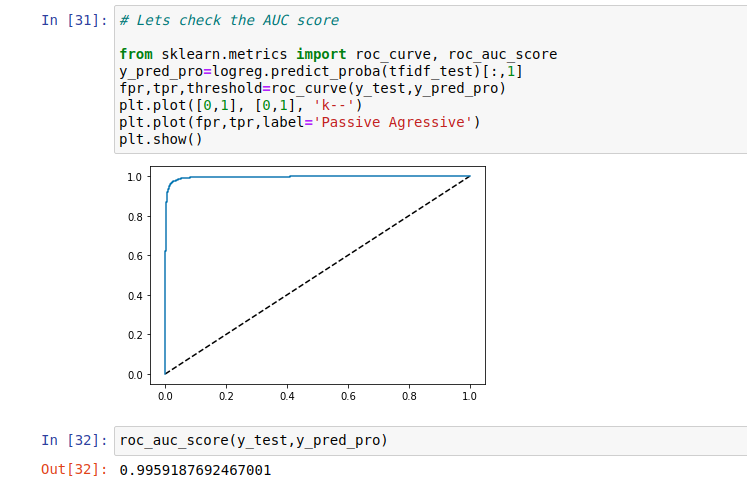




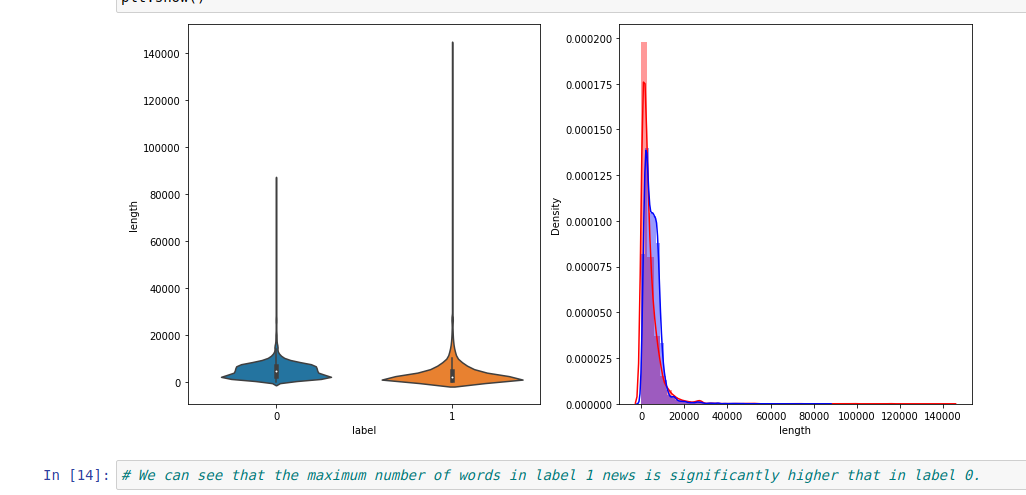








* Key Metrics for success in solving problem under consideration
* As the data is balanced, we've used accuracy as our key metrics. Though, we measured the effectiveness of our model based on multiple metrics like Precision, Recall and AUC-ROC score.
* Visualizations
* Below are the sanpshots of all the plots made:



* Interpretation of the Results
* We have drawn the distribution of length of the news articles along with their headline label wise. We can see that the maximum number of words in label 1 of news is significantly higher that in label 0.

**CONCLUSION**

* Key Findings and Conclusions of the Study
* We have observed that Passive Aggressive Classifer works better than other algorithms even with simpler vectorization techniques such as Bag of Words.
* Learning Outcomes of the Study in respect of Data Science
* We have learnt that the maximum number of words in label 1 of news is significantly higher that in label 0. We also saw a comparison between two vectorization strategies with various models like Passive Aggressive classifier, Logistic Regression and Multinomial Naive Bayes. Due to lack of data, we could not train a word2vec model ourselves, so we tried using a pre-trained word2vec model. Also, tuning the hyper parameters of various algorithms is quite a challenge, specially for deep learning based models.
* Limitations of this work and Scope for Future Work
* We can use state of the art algorithms like Word2Vec or BERT to vectorize the data in future to improve the results. Also, we can use LSTMs and RNN based algorithms for modeling to obtain an even better classifier. Also, we can engineer new features to add more value to the existing data.