

Fake News Detection

Submitted by:

Amit Gawande

**ACKNOWLEDGMENT**

I have referred below resources that helped and guided me in completion of this project as below:-

[www.w3resource.com](http://www.w3resource.com)

[www.towardsdatascience.com](http://www.towardsdatascience.com)

[www.stackoverflow.com](http://www.stackoverflow.com)

**INTRODUCTION**

* **Business Problem Framing**

The authenticity of Information has become a longstanding issue affecting businesses and society, both for printed and digital media. On social networks, the reach and effects of information spread occur at such a fast pace and so amplified that distorted, inaccurate, or false information acquires a tremendous potential to cause real-world impacts, within minutes, for millions of users. Recently, several public concerns about this problem and some approaches to mitigate the problem were expressed.

* **Review of Literature**

The relevant content used in the project are

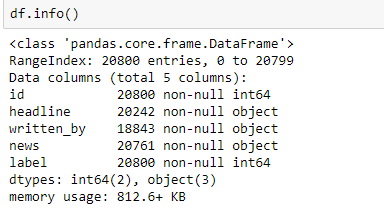
* NumPy
* Pandas
* Sklearn
* Matplotlib
* Nltk
* Seaborn
* Wordcloud
* Tfidvectorizer
* Logistic Regression
* MultinomialNB
* Pasive Aggressive Classifier
* **Motivation for the Problem Undertaken**

A type of yellow journalism, fake news encapsulates pieces of news that may be hoaxes and is generally spread through social media and other online media. This is often done to further or impose certain ideas and is often achieved with political agendas. Such news items may contain false and/or exaggerated claims, and may end up being vitally used by algorithms, and users may end up in a filter bubble.

**Analytical Problem Framing**

* **Mathematical/ Analytical Modeling of the Problem**

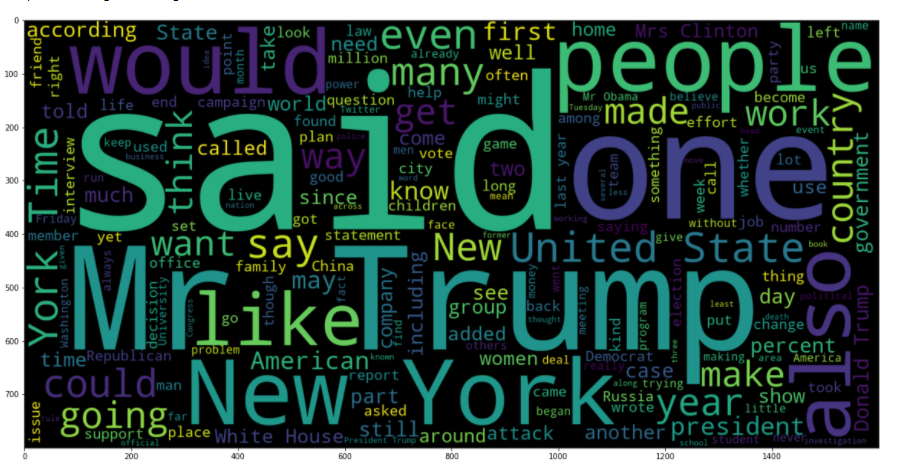
Dataset Info:

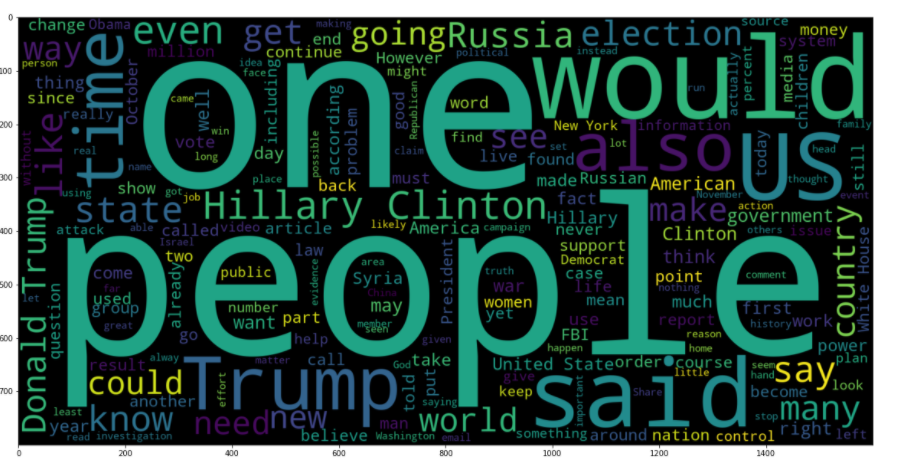


* **Data Sources and their formats**
* There are 6 columns in the dataset provided to you. The description of each of the column is given below:
* “id”: Unique id of each news article
* “headline”: It is the title of the news.
* “news”: It contains the full text of the news article
* “Unnamed:0”: It is a serial number
* “written\_by”: It represents the author of the news article
* “label”: It tells whether the news is fake (1) or not fake (0).
* **Data Preprocessing Done**

The data was cleaned by removing punctuations ,stop words.Stemming and lemmatiztaion are used to reduce word to its root form.

* **Data Inputs- Logic- Output Relationships**





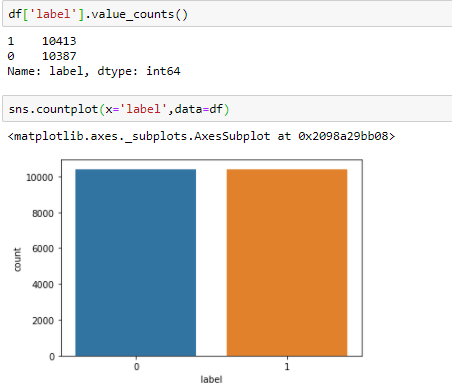
* **State the set of assumptions (if any) related to the problem under consideration**

The model is trained on a concatenation of the headline, written by and the news, the model would be more generalized because adding more words to the input might increase the reliability of the model.

* **Hardware and Software Requirements and Tools Used**
* **Hardware**-64bit, 4GB RAM.
* **Software-**Excel, Anaconda,jupyter notebook

**Model/s Development and Evaluation**

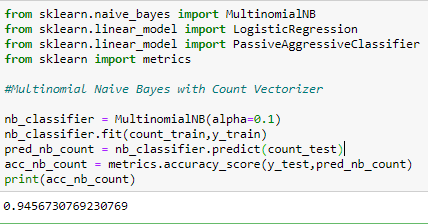
* **Identification of possible problem-solving approaches (methods)**

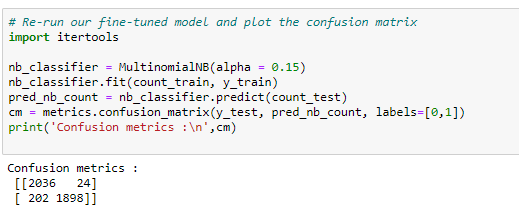


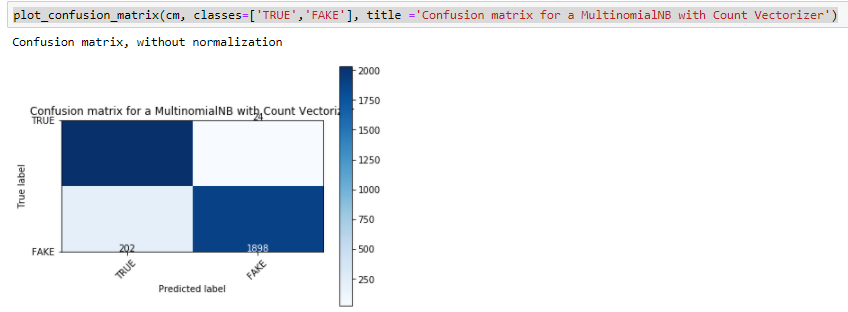
* **Testing of Identified Approaches (Algorithms)**

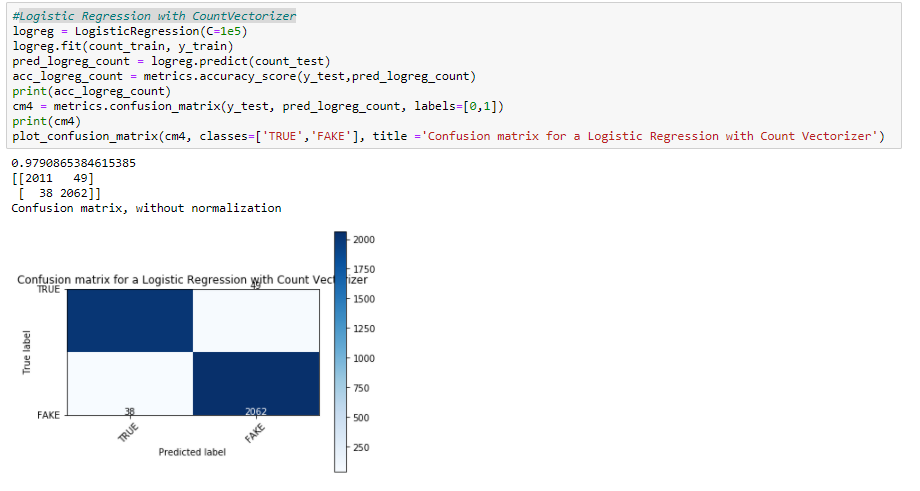
Below are classification algorithms used for the training and testing this dataset.

* MultinomialNB Classifier
* LogisticRegression
* PassiveAggressiveClassifier
* Decision Tree Classifier
* Gaussian NB
* Run and Evaluate selected models









* **Interpretation of the Results**

Selecting Logistic Regression with CountVectorizer as final model , since it has less fake news labeled as true news and also it maximizes accuracy.

**CONCLUSION**

* **Key Findings and Conclusions of the Study**

Every features of dataset plays an important role in identifying the results accurately.

* **Learning Outcomes of the Study in respect of Data Science**
* Visualize the data using univariant / multi-variant analysis.
* Check the prediction score using accuracy score & get ROC-AUC score.
* Train data using classification models to get the best score & finalise best score giver model for this dataset.
* Save file using pickle/joblib library.