

EMAIL SPAM CLASSIFIER PROJECT

Submitted by:

REBANTA DUTTA

**ACKNOWLEDGMENT**

I would like to acknowledge some of the websites from where I have taken help:

1-https://www.javatpoint.com/nlp

2-https://www.educative.io/answers/preprocessing-steps-in-natural-language-processing-nlp

3-https://www.youtube.com/watch?v=5ctbvkAMQO4

4-https://www.youtube.com/watch?v=X2vAabgKiuM

**INTRODUCTION**

* Business Problem Framing

In present times, fake news has become one of the biggest problems. It negatively affects social media and online based outlets because 58% of the people have a trust on this fake news. Moreover, fake news are also dangerously affecting societal values and redefining the truths, facts and beliefs. Therefore, fake news detection model is ML technology used for detection of whether a news is fake or true.

* Conceptual Background of the Domain Problem

Fake news are misleading and exposure of people to these news can have a serious negative impact on the society. A large number of people use social media platforms and trusts the fake news that are mostly circulated in these platforms. Overall, this news are affecting the societal values and threatening lives of people. Therefore, the fake news detection models helps to determine whether a news is fake or not.

* Review of Literature

Fake news (FN) on social media (SM) rose to prominence in 2016 during the United States of America presidential election, leading people to question science, true news (TN), and societal norms. Fake news are having an impact on societal values and changing opinions on issues. FN content, which is divided into individual opinions and scientific consensus on trending issues such as COVID-19, evolution, and climate change.

* Motivation for the Problem Undertaken

With increasing use of social media platforms, the opinions, beliefs and actions of majority of the people are getting influenced by the fake news. Thereby, it is one of the biggest motivations to develop a ML model that would detect whether a news is fake or not.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

Data was collected from a website called Kaggle and there are two datasets- one is for fake news and the another one for true news. The dataset was imported in the Jupyter notebook and any presence of null values was checked. Further, using various NLP techniques like lemmatization, and countVectorizer(), stopwords were identified and the words were converted into vectors for modelling.

* Data Sources and their formats

The dataset contains 44898 rows and 5 columns.

Title: object

Text: object

Subject: object

Date: object

Label: int

* Data Preprocessing Done

After loading the dataset, data preprocessing was done by removing extra characters, and unnecessary words. Additionally, all the letter were also converted to lower case. Following this, the Bag of Words model was developed using the countVectorizer(). After that, label encoding was done to convert the labels into int32 type and the dataset was prepared for modelling.

* Hardware and Software Requirements and Tools Used

The Jupyter notebook was used for analysis and building of the machine learning model. In addition to this, different libraries were used like the matplotlib, seaborn, pandas and numpy.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)
* After identification of object type data, I have used Label encoding technique to convert it into numeric type
* Label encoding technique was used to convert object type data into numeric type.
* Testing of Identified Approaches (Algorithms) like Multinomial NB and passive aggressive classifier
* Run and Evaluate selected models
* Key Metrics for success in solving problem under consideration

The key metrics that have been used are accuracy\_score() and confusion matrix.

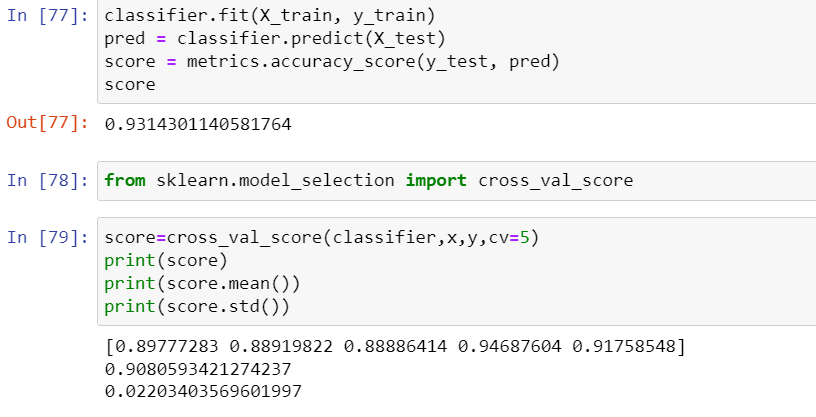
* Visualizations

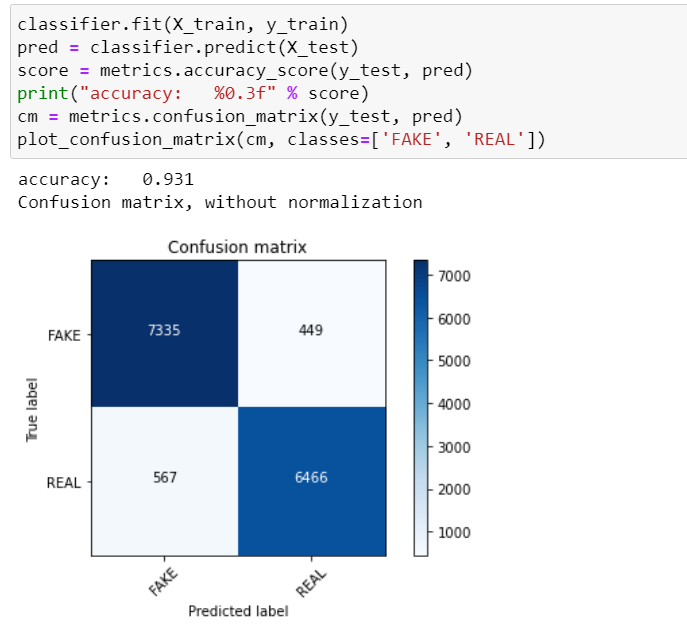
***Barplots***

* Interpretation of the Results

From the visualisation, it could be interpreted that the model is performing with nearly 98% accuracy.

**Multinomial NB**





**CONCLUSION**

* Key Findings and Conclusions of the Study

For building the model to detect fake and true news, it can be concluded that fake news classification to supress the impact of fake news on public and society.

* Learning Outcomes of the Study in respect of Data Science

To detect whether a news is fake or true, Multinomial NB algorithm was used.