

EMAIL SPAM CLASSIFIER PROJECT

Submitted by:

REBANTA DUTTA

**ACKNOWLEDGMENT**

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

* https://www.javatpoint.com/nlp
* https://www.educative.io/answers/preprocessing-steps-in-natural-language-processing-nlp
* https://www.youtube.com/watch?v=5ctbvkAMQO4
* https://www.youtube.com/watch?v=X2vAabgKiuM

**INTRODUCTION**

* Business Problem Framing

The cell phone users in the UK make public claims of receiving SMS Spam messages, but most of them do not report about the very spam messages received. It is a very time consuming task of identifying the spam messages in mails because it involves careful scanning hundreds of webpages. Spam Detector is used to detect unwanted, malicious and virus infected texts and helps to separate them from the nonspam texts

* Conceptual Background of the Domain Problem

People were facing the issue of spam emails and therefore a spam detector was necessary for detections of mails that are malicious, unwanted and virus infected and they need to be separated from non spam or general text messages or emails. Messages have been collected from a website and they were stored in two columns- one column containing the message and the other one containing the label.

* Review of Literature

Spam is a waste of time to the user since they have to sort the unwanted junk mail and it consumed storage space and communication bandwidth. Spam filters detect unsolicited, unwanted, and virus-infested email (called spam) and stop it from getting into email inboxes. Internet Service Providers (ISPs) use spam filters to make sure they are not distributing spam. This is called Spam Detection, and it is a binary classification problem. The reason to do this is simple: by detecting unsolicited and unwanted emails, we can prevent spam messages from creeping into the user's inbox, thereby improving user experience.

* Motivation for the Problem Undertaken

A business email system without spam filtering is highly vulnerable, if not unusable. It is important to stop as much spam as you can, to protect your network from the many possible risks: viruses, phishing attacks, compromised web links and other malicious content.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

Data was collected from a website called Grumbletext and it was stored into 2 columns- one containing the label and the other containing the text. The dataset was imported in the Jupyter notebook and any presence of null values was checked. Further, using various NLP techniques like stemming, and countVectorizer(), stopwords were identified and the dataset was prepared for modelling.

* Data Sources and their formats

The dataset contains 5573 rows and 2 columns.

Label: object

Message:object

* 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)

Multinomial NB

* Run and Evaluate selected models
* Key Metrics for success in solving problem under consideration

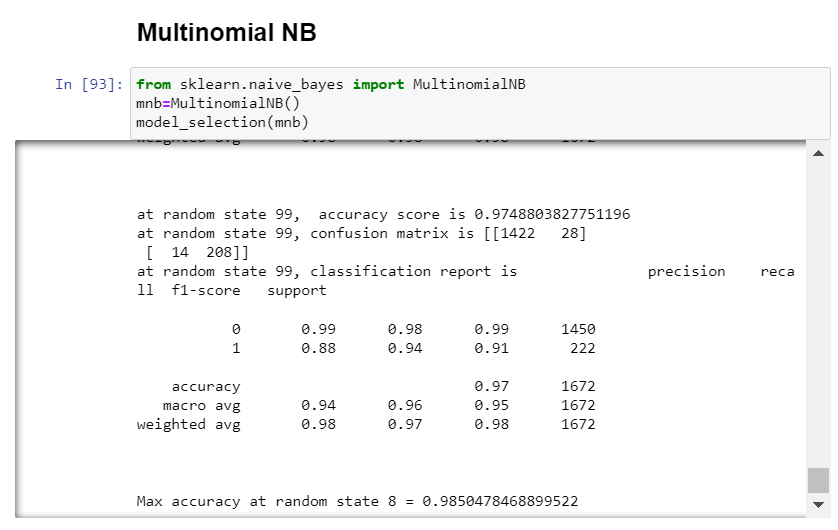
The key metrics that have been used are r2score() and mean absolute error.

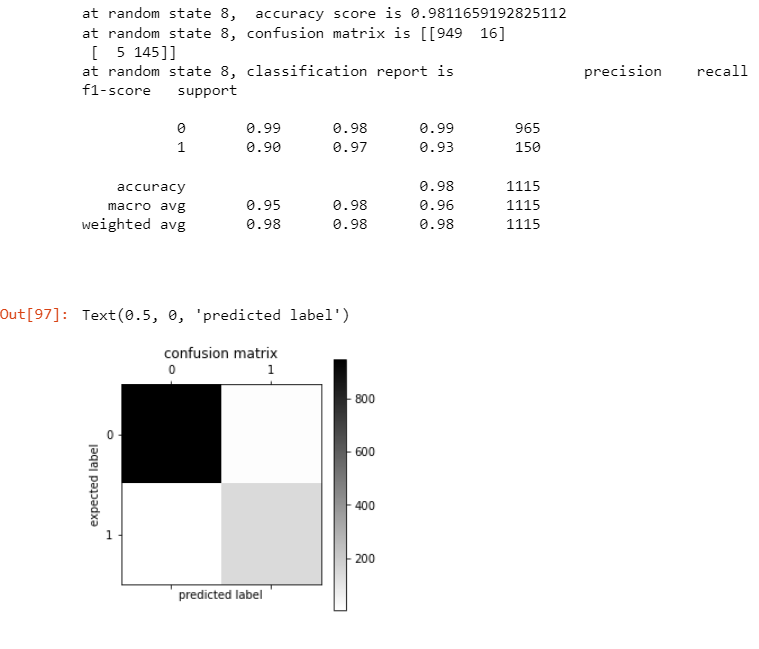
* Visualizations

***Barplots***

* Interpretation of the Results

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





**CONCLUSION**

* Key Findings and Conclusions of the Study

For building the model to detect spam and ham emails, text messages have been collected from a source and from the messages unnecessary words have been removed.

* Learning Outcomes of the Study in respect of Data Science

To detect whether a mail is spam or ham, Multinomial NB algorithm was used.