

Email Spam Classifier

Submitted by:

AMIT PURI GOSWAMI

**ACKNOWLEDGMENT**

The project is done under the kind guidance of Mr. Kasif Khan. Some learnings are made from different informative websites to come out of the some difficult situations. The material which is freely available for any aspirants are used for better understanding of concepts of different techniques. Few of them to name are [www.google.com](http://www.google.com) etc.

**INTRODUCTION**

* Business Problem Framing

There are few mail which come with some malicious intentions and those may be dangerous for the user. These mails may harm the user and his system. These mails are classified as spam and are stored in a different folder called spam folder. The user need to check these mails sender properly and if found suitable then only open such mails or delete these mails without opening these as it may harm the computer system or the user by any way it can.

With the help of data available we need to build a model which can predict that whether the mail falls under spam category or not.

* Conceptual Background of the Domain Problem

All the suspicious mails are classified as spam and other are called as ham. Spam mails are stored in separate folder.

Review of Literature

It is researched that the problem is very important and spam mail may cause harm to the users so the classification is very important for removal of inappropriate or spam mails to avoid any unforeseen event due to these mails.

* Motivation for the Problem Undertaken

Just to keep email users safe is the motivation for the problem.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

Ham mails are labelled as 0 and spam mails are labelled as 1 for the model.

Data Sources and their formats

We received Data spam.csv file it came from one of the email platform and which contains the data for using in the classification of mails.

* Data Preprocessing Done
* punctuation removal
* splitting the words by space
* applying stemmer and lemmatizer
* recombining the words again for further processing
* Data Inputs- Logic- Output Relationships

Input is comment field which is highly useful to decide the comment is of which category or target label.

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

Previous history of classification is used for further classification

* Hardware and Software Requirements and Tools Used

Basic Hardware are : i5 processor, 8GB RAM, other standard h/w

Softwares used : Jupyter, MS Excel, MS Word

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

Standard process of data cleansing, feature selection, EDA, Outliers Removal, Skewness Removal, Handling Multi collinearity , dividing data in train and test, Standardization, finding best random state, model selection, finding best cv, Parameter finetuning, Finalization of the model and predicting.

* Testing of Identified Approaches (Algorithms)

Algorithms used are

1. MultinomialNB

* Run and Evaluate selected models

1. After running and evaluating MultinomialNB

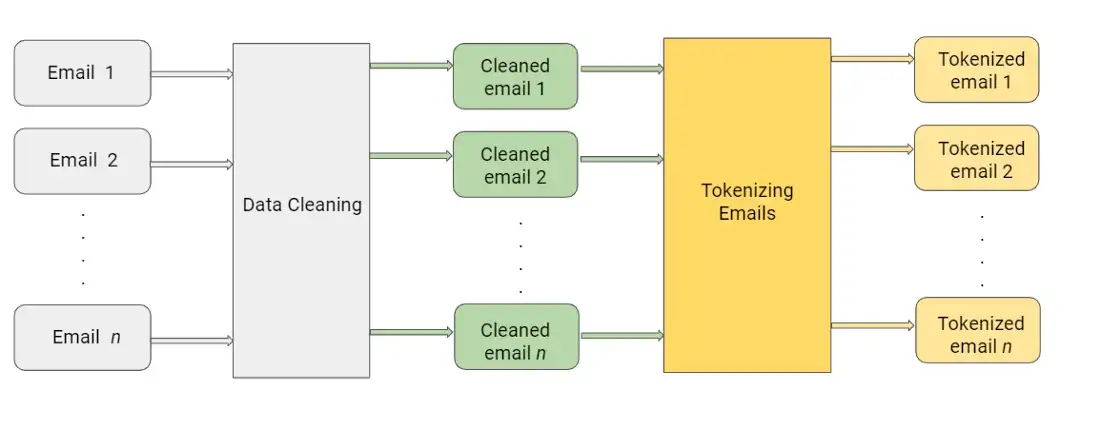
is found to be the best model.

* Key Metrics for success in solving problem under consideration

Best feature selection

* Visualizations : following can be used

1. histplotplot
2. ScatterPlot



* Interpretation of the Results

Spam mails are classified.

**CONCLUSION**

* Key Findings and Conclusions of the Study

All comments are classified.

* Learning Outcomes of the Study in respect of Data Science

Visualization helps us a lot in understanding the problem and identifying the work which will be suitable for model building.

* Limitations of this work and Scope for Future Work

No. of Target variable is 1 but after tokenization it is converted in multiple (huge number) of columns and it is difficult for a standard computer to work on huge number of rows and columns.