

**SPAM EMAIL DETECTION**



**Submitted by:**

**ASHISH KUMAR RATHOR**

**ACKNOWLEDGMENT**

I would like to express sincere gratitude for presenting this report on the “SPAM EMAIL DETECTION ” Project. Working on this project was a good experience to give me a basic knowledge of the Machine Learning Model on NLP. This project helps me research whether I know many new things related to this.

At the commencement of this project report, I would like to thank my SME “MR.MOHD. KHASIF”. Without her guidance, decision, and valuable comments and corrections on my all submitted project, I try to do better compared to last time. It would not possible to reach that point without his direction.

I would like to say thanks to Flip Robo Technologies and Data Trained for providing me with a suitable environment and guidance to complete my project work.

**REFERENCE:**

I have also used a few external resources that helped me to complete this project successfully. Below some resources are available that I am used to completing my project.

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

**TABLE OF CONTAINS: -**

## 1. Introduction

1. Business Problem Framing
2. Conceptual Background of the Domain Problem
3. Review of Literature
4. Motivation for the Problem Undertaken

## 2. Analytical Problem Framing

1. Mathematical/ Analytical Modelling of the Problem
2. Data Sources and their formats
3. Data Pre-processing Done
4. Data Inputs- Logic- Output Relationships
5. Hardware & Software Requirements & Tools Used

## 3. Model/s Development and Evaluation

* + - 1. Identification of possible Problem-solving approaches

1. (Methods
   * 1. Visualizations
     2. Testing of Identified Approaches (Algorithms)
     3. Run and Evaluate Selected Models
2. Key Metrics for success in solving the problem under consideration
   * 1. Interpretation of the Results

## 4. Conclusion

* + 1. Key Findings and Conclusions of the Study
       1. Learning Outcomes of the Study in respect of Data Science
    2. Limitations of this work and Scope for Future Work

**INTRODUCTION**

A spam Detector is used to detect unwanted, malicious, and virus-infected texts and helps to separate them from no-spam texts. It uses a binary type of classification containing the labels such as ‘ham’ (no spam) and spam. Application of this can be seen in Google Mail (GMAIL) where it segregates spam emails to prevent them from getting into the user’s inbox.

The files contain one message per line. Each line is composed of two columns: v1 contains the label (ham or spam) and v2 contains the raw text.

This corpus has been collected from free or free research sources on the Internet:

-> A collection of 5573 rows of SMS spam messages was manually extracted from the Grumble text Web site. This is a UK forum in which cell phone users make public claims about SMS spam messages, most of them without reporting the very spam message received. Identifying the text of spam messages in the claims is a very hard and time-consuming task, and it involves carefully scanning hundreds of web pages.

-> A subset of 3,375 randomly chosen ham messages of the NUS SMS Corpus (NSC), a dataset of about 10,000 legitimate messages collected for research at the Department of Computer Science at the National University of Singapore. The messages largely originate from Singaporeans and mostly from students attending the University. These messages were collected from volunteers who were made aware that their contributions were going to be made publicly available.

Below are some of the most popular machine-learning methods:

a)**Naïve Bayes classifier**: It is a supervised machine learning algorithm where word probabilities play the main role. If some words often occur in spam but not ham, this incoming e-mail is probably spam. The naïve Bayes classifier technique has become a very popular method in mail filtering software. The Bayesian filter should be trained to work effectively. Every word has a certain probability of occurring in spam or ham email in its database. If the total of word probabilities exceeds a certain limit, the filter will mark the e-mail to either category.

b)**Artificial Neural Networks classifier**: An artificial neural network (ANN), also called simply a "Neural Network" (NN), is a computational model based on biological neural networks. It consists of an interconnected collection of artificial neurons. An artificial neural network is an adaptive system that changes its structure based on information that flows through the artificial network during a learning phase.

**BUSINESS PROBLEM FARMING**

Junk emails or unsolicited bulk emails sent to a large list of email users through the email system are referred to as email spam. Typically, they are misleading ads that promote low-quality services and, in some instances, include images with content that is inappropriate for children. Whether commercial or not, many of them are dangerous since they may contain links that appear to be legitimate and recognizable, but they lead to [phishing](https://heimdalsecurity.com/blog/phishing-attack/) websites that host [malware](https://heimdalsecurity.com/blog/what-is-malware-as-a-service-maas/) or include malware in the form of file attachments.

Typically, spammers obtain recipients’ email addresses from publicly available sources and use them to advertise and promote their businesses; they may also use them to collect sensitive information from the victim’s machine. These collected email addresses are sometimes also sold to other spammers.

**REVIEW OF LITERATURE**

Spam text is a complex phenomenon, and different knowledge fields try to study and tackle this problem. In this study, serval related literature is used to express different types of spam text. This research found a few dedicated works that address the effect of incorporating different text transformations on the model accuracy for text classification. In this work, we performed a systematic review of state of art in spam text detection classification using the machine learning method with NLP text processing. In our analysis of every primary study, we investigated the data set used, evaluation metrics, used machine learning methods.

**MOTIVATION FOR THE PROBLEM UNDERTAKEN**

The main objective of this study is to investigate which method from a chosen set of machine learning techniques performs the best. So far, we have a range of publicly available models served through the Perspective API, including spam comments. But the current models still make errors and don’t allow users to select which type of toxicity they are interested in finding.

The project which is given by Flip ROBO as a part of the internship program gives insight to identify major factors that lead to spam comments. The exposure to real-world data and the opportunity to deploy my skillset in solving a real-time problem has been the primary objective. However, the motivation for taking on this project was that it is relatively a new field of research. Here we have many options but less concrete solutions. The main motivation was to classify the news to bring awareness and reduce unwanted chaos and make a good model to help us know such kinds of miscreants. Our goal is to build a prototype of an online text classifier that can be used to classify spam and not spam comments.

**ANALYTIC PROBLEM FRAMING**

**MATHEMATICAL/ANALYTICAL MODELLING OF THE PROBLEM**

The SMS Spam Collection is a set of SMS-tagged messages that have been collected for SMS Spam research. It contains one set of SMS messages in English of 5,574 messages, tagged according to being ham (legitimate) or spam

A spam Detector is used to detect unwanted, malicious, and virus-infected texts and helps to separate them from non-spam texts. It uses a binary type of classification containing the labels such as ‘ham’ (non-spam) and spam. Application of this can be seen in Google Mail (GMAIL) where it segregates spam emails to prevent them from getting into the user’s inbox.

The files contain one message per line. Each line is composed of two columns: v1 contains the label (ham or spam) and v2 contains the raw text.

This corpus has been collected from free or free research sources on the Internet:

**DATA SOURCES AND THEIR FORMATS**

A collection of 5573 rows of SMS spam messages was manually extracted from the Grumble text Web site. This is a UK forum in which cell phone users make public claims about SMS spam messages, most of them without reporting the very spam message received. The identification of the text of spam messages in the claims is a very hard and time-consuming task, and it involved carefully scanning hundreds of web pages.

-> A subset of 3,375 randomly chosen ham messages of the NUS SMS Corpus (NSC), a dataset of about 10,000 legitimate messages collected for research at the Department of Computer Science at the National University of Singapore. The messages largely originate from Singaporeans and mostly from students attending the University. These messages were collected from volunteers who were made aware that their contributions were going to be made publicly available.

**DATA PRE-PROCESSING DONE**

Data pre-processing is the process of converting raw data into a well-readable format to be used by a machine learning model. Data pre-processing is an integral step in machine learning as the quality of the data and the useful information that can be derived from it directly affects the ability of our model to learn, therefore we must pre-process our data before feeding it into our model. I have used the following pre-processing steps:-

* Importing necessary libraries and loading datasets into in data frame.
* After loading the data checked some statistical information like the shape of the dataset, the number of unique values, info about the dataset, null values, values count of the target, duplicated values, etc.
* Checked null values and find in 3 columns out of 5 near 99%, so we drop these columns.
* After this, we change both columns’ names and make two labels and text to replace v1 and v2.
* Visualized each feature using seaborn and matplotlib libraries by plotting pie charts and count plots.
* Done text pre-processing techniques like Removing Punctuations and other special characters, Splitting the comments into individual words, Removing Stop Words, Stemming, and Lemmatization.

After getting cleaned data used the TF-IDF vectorizer. It’ll help to transform the text data to a feature vector which can be used as input in our

**DATA INPUT- LOGIC OUTPUT RELATIONSHIP**

The dataset contains two variable types and one feature The features are independent and the label is dependent as the values of our independent variables changes as our label varies

To know about the scene of a loud and the most common word in each label we use word cloud which gives the words frequented in the labels

I have also checked the correlation between the label and the length of the text using the heatmap

**HARDWARE & SOFTWARE REQUIREMENTS & TOOLS USED**

To build the machine learning project it is important to have the following hardware and software requirements and tools

|  |  |
| --- | --- |
| Hardware | Processor: intel core i5  RAM: 8GB  SSD: 512GB |
| Software | Distribution: Anaconda Navigator  Programming language: Python  Browser-based language shell: Jupyter Notebook |

**MODELS DEVELOPMENT AND EVALUATION**

**IDENTIFICATION OF POSSIBLE SOLVING APPROACHES(METHOD)**

In this project, there was 1 feature that defines the type of text like spam.

We created two columns of name labels and text from v1 and v2 for better understanding. which is combined all the above features and contains the labeled data in the format of 0 and 1, where 0 represents “NO SPAM” and 1 represents “SPAM YES”. In this NLNLP-based project, we need to predict the multiple labels which are binary. I have converted text into features vector using TF-IDF vectorizer and separated our features and label. Also, before building the model, I made sure that the input data is cleaned and scaled before it was fed into the machine-learning models.

**TESTING OF IDENTIFIED APPROACHES(ALGORITHMS)**

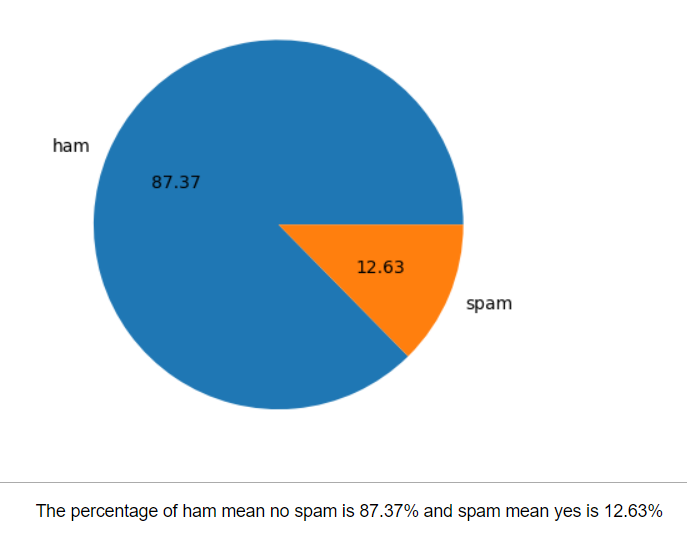
Since the target variable is categorical in nature, from this I can conclude that it is a classification-type problem hence I have used the following classification algorithms. After the pre-processing and data cleaning, I have two columns one is the target and 2nd is text-nlp after applying to pre-process of NLP.

The algorithms used for training the data are as follows:

1. LOGISTIC REGRESSION
2. MULTINOMIALNB
3. GAUSSIANNB
4. BERNOULLINB
5. DECISION TREE CLASSIFIER
6. KNEIGHBORS CLASSIFIER
7. ADA BOOST CLASSIFIER
8. GRADIENT BOOSTING CLASSIFIER
9. RANDOM FOREST CLASSIFIER
10. VOTING CLASSIFIER

**VISUALIZATIONS**

I used pandas profiling to get the overviewed visualization of the pre-processed data. Pandas is an open-source python module that can do an Exploratory data analysis to get a detailed description of the features and it helps in visualizing and understanding the distribution of each variable. I have used word cloud to get the sense of loud words in the labels.



**Observation🡪**

From the pie chart, we can notice approximately 87.37% of the text are not spam, and 12.63% of the text is spam contains the text. The count of the not-spam text is very high compared to spam text. Here unbalancing problem so we solve this problem by SMOTE method.

**INTERPRETATION OF THE RESULTS**

**VISUALIZATION:**- From the pie chart, we can notice approximately 87.37% of the text are not spam, and 12.63% of the text is spam contains the text. The count of the not-spam text is very high compared to spam text. Here unbalancing problem so we solve this problem by SMOTE method.

Pre-Processing:- The dataset should be cleaned and scaled be cleaned and scaled to build the Machine Learning Models to get a good prediction. I have performed a few NLP text processing steps which I have already mentioned in the pre-processing steps where all the important features are present in the dataset and ready for model building.

**Model Building:-** After cleaning and processing data, I performed a train test split to build the model. I have built multiple models. I have multiple classification models to get the accurate accuracy score and evaluation metrics like precision score, confusion matrix, classification matrix, etc. I got a Random Forest classifier as the best model. When I performed the Hyperparameter tuning on the Random Forest classifier then I got the highest accuracy score of 98.89% and the precision score of 99.32% and an f1 score is 99%. I saved my final model and got good prediction results for the test dataset.

**CONCLUSION**

**Key finding and Conclusion of the Study:-**

From the above analysis, the below-mentioned results were achieved which depict the chances and condition of a text being a spam comment or a normal comment. With increasing fraud in society, people spread it on a large level and it increasing day by day. It has strong negative impacts on individual users and broader society.

The conclusion for our Study:

* In our dataset text we have only 12% spam-mentioned text out of the whole dataset text.
* In the 12% of text comments used spam words
* After using the word cloud, we find that there are so many negative words. While in not spam text no negative and spam including word.
* Some of the text is very long while some are short.

**Learning Outcomes of the Study in respect of Data Science**

**Limitations:** This project was amazing to work on, it creates new ideas to think about but there were some limitations in this project like an unbalanced dataset. Every effort has been put into it for perfection but nothing is perfect and this project is no exception. Certain areas can be enhanced.

Future work: In future work, we can focus on performance and error analysis of the model as lots of text is misclassified into the spam category. Previous work has achieved success using various algorithms on data in the English language but in the future, we can consider having data in regional languages. We can also work on after work of the detection of Spam texts like automatic blocking of the user, and auto-deletion of harmful text on mail, and message platforms. Spam detection is an emerging research area with few public datasets. So, a lot of work need to be done in this field