FLIP ROBO

Email Spam Classifier Project

Submitted by:

PRINCE KUMAR

INTERNSHIP - 31

ACKNOWLEDGMENT

[1] E.M. Bahgat, S. Rady, W. Gad An e-mail filtering approach using classification techniques the 1st International Conference on Advanced Intelligent System and Informatics (AISI2015), November 28-30, 2015, Springer International Publishing, BeniSuef, Egypt (2016), pp. 321-331.

[2] N. Bouguila, O. Amayri, a discrete mixture-based kernel for SVMs: application to spam and image categorization Inf. Process. Manag., 45 (6) (2009), pp. 631-642

[3] Y. Cao, X. Liao, Y. Li An e-mail filtering approach using neural network International Symposium on Neural Networks, Springer Berlin Heidelberg (2004), pp. 688-694

[4] F. Fdez-Riverola, E.L. Iglesias, F. Diaz, J.R. Méndez, J.M. Corchado SpamHunting: an instance-based reasoning system for spam labelling and filtering.

[5] S. Mason New Law Designed to Limit Amount of Spam in E-Mail (2003) http://www.wral.com/technolog Google Scholar

INTRODUCTION

- ## Business Problem Framing

  The majority of us view spam emails as irritating and frequently utilised for product promotion and advertising. We continue to block these email addresses, however, it is pointless because spam emails are still widely used. There are several significant types of spam emails that pose a serious threat to security, including fake emails, identity theft through hacking, viruses, and malware. Building a strong real-time email spam classifier that can effectively and correctly mark incoming mail as spam if it is a spam message or appears to be one is necessary to deal with spam emails. Additionally, the latter will aid in creating an anti-spam filter.

  While Google and other email providers are useful for identifying spam emails, they are still in their early stages and require continuous user feedback. The end-user receives basic services for free from well-known email providers like Gmail, Yandex, Yahoo Mail, etc., with a EULA of course. As private companies run their own email servers and desire them to be more secure due to the secret data, there is a large potential for developing email spam classifiers. In such circumstances, email spam classifier solutions can be supplied to such companies.

- ## Conceptual Background of the Domain Problem

  Although there are other email spam filtering techniques, the most recent ones are described below.

  - Content Based Filtering Technique: Automatic filtering rules and email categorization utilising machine learning techniques like Naive Bayesian classification, Support Vector Machine, K Nearest Neighbor, and Neural Networks are typically created using content-based filtering. In order to filter incoming email spam, this technology often analyses terms, the incidence, and distribution of words and phrases in email content.

- Case Base Spam Filtering Method: One of the widely used spam filtering techniques is case base or sample base filtering. First, using a collection approach, all emails—spam and non-spam—are collected from each user's email. Then, utilising the client interface, feature extraction, selection, grouping email data, and process evaluation, pre-processing stages are carried out to change the email. After that, the information is divided into two vector sets. In order to determine whether incoming emails are spam or not, machine learning algorithms are employed to train datasets and test them.
- Heuristic or Rule Based Spam Filtering Technique: This method compares a large number of patterns, most of which are regular expressions, against a selected message using pre-made rules or heuristics. A message's grade is raised when there are several related patterns. If any of the patterns didn't match, it subtracts from the score. Any communication that receives a score beyond a certain level is classified as spam; otherwise, it is considered to be authentic. While certain ranking criteria do not vary over time, others need to be updated often in order to successfully combat the threat of spammers who constantly add new spam messages that can easily evade detection by email filters. Spam Assassin is a good illustration of a rule-based spam filter.
- Previous Likeness Based Spam Filtering Technique: This method classifies incoming emails based on how closely they resemble stored examples using memory-based, or instance-based, machine learning techniques (e.g. training emails). A multi-dimensional space vector is created using the email's properties, and new instances are plotted as points using this vector. The most well-liked class of the K-closest training instances is then given the new instances. This method filters spam emails using the k-nearest neighbour (kNN) algorithm.

- Adaptive Spam Filtering Technique: The technique classifies spam into distinct categories in order to detect and filter it. It separates an email corpus into different groups, each of which has a distinctive text. Each incoming email is compared to each group, and a percentage of similarity is calculated to determine the most likely group to which it belongs.

• Review of Literature

Different email spam classification approaches have been proposed by numerous researchers and academics, and they have been utilised successfully to divide data into groups. These techniques consist of probabilistic, decision-tree, artificial immune system, support vector machine (SVM, artificial neural networks (ANN), case-based technique, and artificial immune system (AI). The application of these classification techniques for spam mail filtering has been demonstrated in the literature by utilising a content-based filtering methodology that will identify specific traits (normally keywords frequently utilised in spam emails). The frequency with which these characteristics exist in emails determines the probabilities for each attribute, which are then compared to the threshold value. Spam is defined as email messages that exceed the threshold value.

• Motivation for the Problem Undertaken

One of the most well-liked and effective ways to share data or messages online is through email. Given the importance and widespread use of emails, the quantity of spam emails has also significantly expanded. Spam emails are unwelcome emails with a variety of contents, including offers, adverts, harmful links, malware, trojans, etc. Spammers send junk mail with the purpose to perpetrate email fraud, so it's critical to separate spam from legitimate communications. The goal of this project is to develop machine learning-based email spam detection models that can accurately distinguish between spam and valid emails.

Analytical Problem Framing

- Mathematical/ Analytical Modelling of the Problem

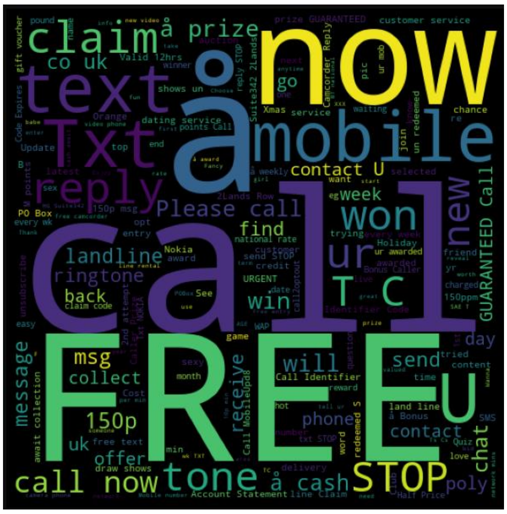
  We use accuracy score, classification report, and confusion matrix as our evaluation metrics because the case study calls for the classification of Labels "1" and "0," where "1" denotes a spam message and 0 denotes a genuine message

- Data Sources and their formats

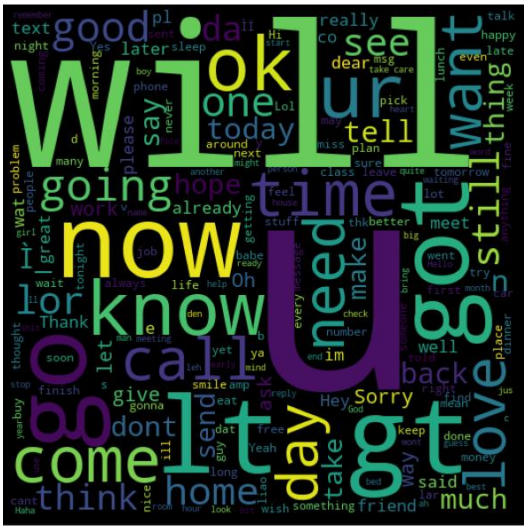
  A group of SMS-tagged messages known as the SMS Spam Collection have been gathered for SMS Spam study. It includes a single batch of 5,574 English SMS messages that have been classified as either spam or ham (legal). Each line in the files holds one message. Two columns make up each line: v1 carries the label (such as "ham" or "spam") and v2 contains the actual text. This corpus was compiled using free or open-access Internet resources. From the Grumble text website, a collection of 5573 rows of SMS spam messages was manually collected. In this UK forum, cell phone users complain about SMS spam messages in the open, the majority of them without ever reporting the spam message they received. It took a lot of time and effort to thoroughly review hundreds of web sites in order to identify the spam messages contained in the claims. 3,375 SMS messages from the NUS SMS Corpus (NSC), a corpus of roughly 10,000 genuine messages gathered for research at the National University of Singapore's Department of Computer Science, were randomly selected as ham messages. The majority of the messages are sent by Singaporeans who attend the University. These comments were gathered from volunteers who had been informed that their contributions would be made public.

- Data Inputs- Logic- Output Relationships

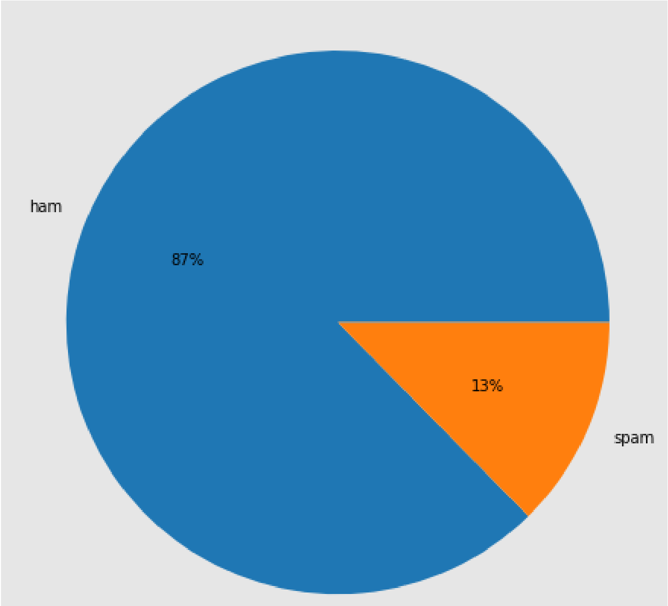
After visualising the most repeated words in the spam messages, the below words are obtained.



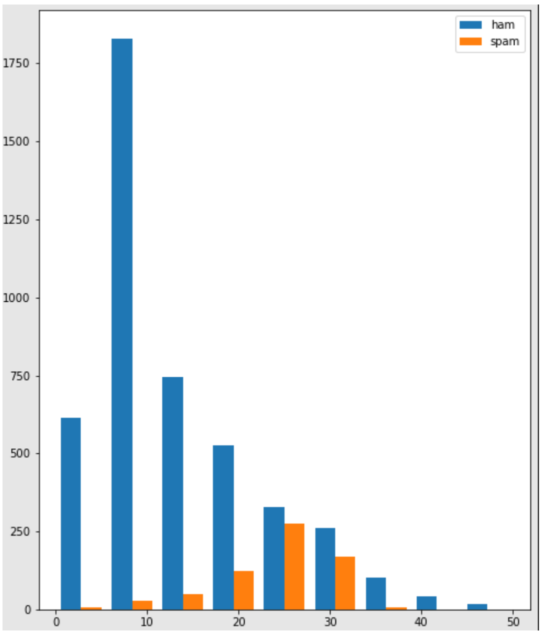
The most repeated words for ham messages are shown below.



On plotting a pie chart of the messaged, 13% of the messages are spam messaged and 87% are genuine messages.



When histogram is plotted for the messages with respect to the number of words for both ham and spam messages, it is observed that most of the ham messages contain 0 to 10 words and majority of spam messages are longer and contain between 20 to 30 words



- Data Pre-processing

From the given data, the unnecessary columns are removed and after removing the duplicates the percentage loss in data was found to be 7%. A function was defined to clean the text and return the tokens. The cleaning of the text was done by first removing punctuation and then removing the useless words also known as stop words. The target column which denotes whether the message is spam or not is converted to numerical feature. Further, the message text is converted into vectors using TFid Vectorizer.

- Hardware and Software Requirements and Tools Used

Anaconda Navigator is a desktop graphical user interface (GUI) included in Anaconda® Distribution that allows you to launch applications and manage conda packages, environments, and channels without using command line interface (CLI) commands. Navigator can search for packages on Anaconda.org or in a local Anaconda Repository. It is available for Windows, macOS, and Linux.

The Jupyter Notebook application allows you to create and edit documents that display the input and output of a Python language script.

Model/s Development and Evaluation

- Testing of Identified Approaches (Algorithms)

The features and target are split into train and test sections using the train_test_split function for random state-0 and test-size-0.25. The algorithms used for model training and testing are-

1. Logistic Regression
2. Random Forest Classifier
3. Decision Tree Classifier
4. Gaussian NB Classifier
5. K Neighbors Classifier
6. Support Vector Machine
7. Gaussian Bayes Classifier

- ## Run and evaluate selected models

  When the train and test data were used for developing models for the above algorithms the following observations were made:

  - Linear Regression models has very less difference in accuracy score of training and testing data which is beneficial but the F1 score is only 75%.
  - Random Forest Classifier had difference of 0.02 between accuracy score of training and test data and F1 score of 90%.
  - Decision Tree Classifier had difference of 0.03 between accuracy score of training and test data.
  - Gaussian NB Classifier had difference of 0.06 between accuracy score of training and test data.
  - KNeighbors Classifier had a difference of 0.04 between accuracy score of training and test data.
  - Support Vector Classifier had a difference of 0.03 between accuracy score of training and test data.

- ## Key Metrics for success in solving problem under consideration

  To forecast the target class of the data sample in classification issues, classification models are used. The likelihood that each occurrence belongs to a certain class is predicted by the categorization model. To effectively employ classifications models in production for resolving practical issues, it is critical to assess their effectiveness. Machine learning classification models' performance metrics are used to evaluate how well they perform in a specific situation. Accuracy, precision, recall, and F1-score are some of these performance indicators. Model performance is vital to machine learning since it enables us to comprehend the advantages and disadvantages of these models while making predictions in novel circumstances.

CONCLUSION

Random Forest Model is the effective model for the case study since it has a good training score and the difference between training and testing score is less compared to other models.

The model was further analysed using the hyperparameter tuning and from the results the False Negative reduced from 39 to 38.