AMERICAN JOURNAL OF INNOVATION IN SCIENCE AND ENGINEERING (AJISE)

ISSN: 2158-7205 (ONLINE)

VOLUME 1 ISSUE 1 (2022)

Indexed in

PUBLISHED BY: E-PALLI, DELAWARE, USA

American Journal of Innovation in Science and Engineering (AJISE)

Volume 1 Issue 1, Year 2022
ISSN: 2158-7205 (Online)
DOI: https://doi.org/10.54536/ajise.v1i1.996
https://journals.e-palli.com/home/index.php/ajise

Detection of Spam Email

Manish Panwar[1*], Jayesh Rajesh Jogi[1], Mahesh Vijay Mankar[1], Mohamed Alhassan[1], Shreyas Kulkarni[1]

Article Information

Received: November 23, 2022

Accepted: December 21, 2022

Published: December 30, 2022

Keywords

Spam Email, Classification, Dataset, Performance Metrics

ABSTRACT

Spam, often known as unsolicited email, has grown to be a major worry for every email user. Nowadays, it is quite challenging to filter spam emails since they are made, created, or written in such a unique way that anti-spam filters cannot recognize them. In order to predict or categorize emails as spam, this paper compares and reviews the performance metrics of a few categories of supervised machine learning techniques, including Svm (Support Vector Machine), Random Forest, Decision Tree, Cnn, (Convolutional Neural Network), Knn(K Nearest Neighbor), Mlp(Multi-Layer Perceptron), Adaboost (AdaptiveBoosting), and Nave Bayes algorithm. Thegoal of this study is to analyze the specificsor content of the emails, discover a limited dataset, and create a classification model that can predict or categorize whether spam is present in an email. Transformers' Bidirectional Encoder Representations) has been optimized to perform the duty of separating spam emails from legitimate emails (Ham). To put the text's context into perspective, Bert uses attention layers. Results are contrasted with a baseline Dnn (deep neural network) modelthat consists of two stacked Dense layers and a Bilstm (bidirectional Long Short-Term Memory) layer. Results are also contrasted with a group of traditional classifiers, including k- Nn (k-nearest neighbours) and Nb (Naive Bayes). The model is tested for robustness andpersistence using two open-source data sets, one of which is utilized to train the model.

INTRODUCTION

The internet has progressively assimilated into daily life. Email users are growingdaily as a result of increased internet usage. Because of the rise in email usage, there are issues brought on by spam, or unsolicited bulk email. Spam emails are produced as a result of the email is one of the best platforms for advertising nowadays. Spam emails are those that the recipient has requested not to receive. To many email recipients, many copies of the same message are sent. Giving away our email address on an unlawful or dishonest website frequently results in spam. Spam has a wide range of negative impacts. a large number of absurd emails fill our inbox. greatly reduces the speed of our Internet. stealsvital information from your contact list,such as our data. on any computer software, modifies the search results. Spam is a major waste of time for everyone and may get annoying very fast if you get a lot of it.

It takes a lot of work to locate these spammers and the spam content. Despite the large number of studies that have been conducted, they have not yet been able to discriminate between legitimate surveys and spam, and none of them have shown the advantages of each eliminated component. Spam messages are utilized for someattacks despite increasing network communication and using up a lot of RAM. Unsolicited commercial or malicious emails sent to a single person, a business, or a group of individuals are referred to as spam emails or non-self. In addition to advertisements, they could have connections to websites harbouring malware or phishing attacks that have been known to steal personal data. Many spam filtering methods are employed to address this issue. Our mailbox is protected against spam emails by spam filtering algorithms.

LITERATURE REVIEW

The authors of the paper have identified some email header attributes that can be effectively used to identify and categorize spam communications. These features were chosen based on how well they detected spam messages. This study also compares the properties of Hotmail, Gmail, and Yahoo Mail so that a universal spam message detection method can be suggested for all significant email providers.

In the paper, a novel method based on the principle of word repetition frequency was applied. Before determining the grammatical functions of the full words in the phrase, the essential sentences—those with the keywords—of the incoming emails must first be tagged.

To determine the degree of similarity between received emails, they will finally be combined into a vector. The received e- mail is categorized using the K-Mean method. The process used to decide which category the email belongs to is called vector determination.

The authors of the report discussed cyberattacks. Email services are commonly used by phishers and malicious attackers to deliver fake communications that can cause target users to lose money and their social standing. These lead to the theft of private information, including passwords, credit card numbers, and other private information. The authors of this paper used Bayesian classifiers. Take into account each message. continuously adapts to new spam forms.

In the study, a proposed system attempts to identify a

[1] Department of MCA, Vishwakarma Institute of Technology Pune, India
[*] Corresponding author's e-mail: manish.panwar21@vit.edu

pattern of recurring terms that are considered spam using machine learning techniques. The system also suggests categorizing emails based on additional variables found in their structure, such as the domain, header, and Cc/Bcc fields. When applied to the machine learning method, each parameter would be viewed as a feature. It will be possible to discern between a proper output and an ambiguous output using the machine learning model's feedback mechanism, which has been pre-trained. This approach offers an alternate architecture for the implementation of a spam filter. The email body with its widely used terms and punctuation is also taken into account in this paper

Problem Statement

The spammer is the one who is in charge of sending the spam messages. This individual collects email addresses from various websites, chat forums, etc. The massive amount of spam emails streaming across computer networks harms user time, CPU power, communication bandwidth, and email server memory. Overall, the current approach is ineffective in detecting spam emails. As a result, many users also suffer immeasurable financial losses. It results in poor test and prediction accuracy, decreased security, and data loss.

Related Work

Researchers have already becomeinterested in the topic of spam email detection. Spam e-mail detection has been the subject of several important works. This section discusses earlier related efforts that concentrate on classifying spam using ML and deep learning methods. The effectiveness of word embedding in deep learning for email spam detection is discussed by Srinivasan *et al.* The proposed method outperformed other traditional methods of email representation. To identify phishing emails, Soni introduced THEMIS, a profound learning model that employs an upgraded RCNN and displays the email body and header at both the character and word levels.

According to test results, THEMES' accuracy was 99.84 percent, which is higher than that of LSTM and CNN's experiment. Instead of employing rule- based techniques, Hassanpur *et al.* represented emails to vectors using the word2vec package. A NN, the learning model, receives input in the form of vector representations. Compared to the conventional machine learning techniques, their method achieves an accuracy of over 96 percent. By processing the email samples' content and extracting features concentrated on word counts, stopword counts, punctuation counts, and uniqueness factors, Egozi *et al.* attempted to validate the efficacy of using NLP approaches to detect phishing emails.

Over 80% of phishing emails and 95% of spam emails could be successfully identified using the 26 extracted characteristics that were utilized to train an ensemble learning model based on linear kernel SVM. To categorize emails as spam or ham, Seth *et al.* suggest a hybrid CNN model that examines both the email'stextual and visual content. Their model reaches 98.87 percent accuracy, which is very good. By including a polarity score feature that represents the semantic of email content, Ezpeleta *et al.* increase the accuracy and dataset used for spam classification using Bayesian filteringclassifiers up to 99.21%, concluding that sentiment analysis of the emails may aid in the detection of spam emails. Comparison analysis for earlier spam filtering systems' accuracy is proposed by Bibi *et al.*

METHODOLOGY

The general strategy and techniques employed to carry out the task of spam email detection are detailed in detail in this section. Data collection, data pre-processing, feature extraction, model training, and model evaluation are the five essential stages of any NLP activity. As a result, feature extraction in this study will be carried out automatically throughout the deep learning model training.

Data Collection

In this study, two open-source data setswith two columns each—one containing theemail body text and the other the classification of spam or ham—were employed. The first data set comes from the UCI machine learning repository and is theopen-source Spambase data set. It contains5569 emails, of which 745 are spam. The second set of data consists of 5728 emails, 1368 of which are spam, and is an open- source spam filter data set from Kaggle. Both data sets are unbalanced when thedistributions for the SPAM and HAM classes are examined; the SPAM class is therare class.

A fresh balanced training data set is constructed by merging spam samples that are randomly selected from the two data sets Spambase and Spam filter while ensuring that there are no duplicate records to avoid biassing for the main class, which is the HAM class. 2000 samples of SPAM and 3000 samples of HAM can be found in the new data set. Additionally, a portion of the Spambase data set was held back to evaluate the model's durability against unknown samples. To determine if the text is labeled as SPAM 1 or HAM 0, the task is a binary classification problem. In both the training and testing data sets, Table 2 displays the distribution for the classes HAM and SPAM

Data Cleaning and Pre- Processing

The distribution of the word number in each record is examined after going through the class distribution for each target. In general, SPAM texts are longer than HAM texts; in our case study, the input sequence for our model is set to 300 tokens or words. The next step in the data cleaning process involved extracting and removing stop words using the Sklearn library because stop words offer little to no distinctive information that can be used for classification. Punctuation was then extracted and removed because it has an impact on the text encoding, particularly when it is attached to a word, as in the case of [users, user's] having a different encoding. Text retained

in case because it could be a sign of spam particularly spam or other forms of promotion fraud. The words were divided into tokens depending on space using the Keras tokenization program. For instance, ["We went to Amman."] will be tokenized as ["We went to Amman."], "went to," and "to." Using Keras' TfidfVectorizer, each token was converted to a vector for use with traditional classifiers. The binary encoding of the labeled target variable HAM 0 and SPAM 1

Base Line Model

The modern BiLSTM model is used as the baseline model. It uses an embedding layer's input and feeds a dense layer with its output vectors. To enable speedy convergence, the Dense layer uses the Relu activation function. To prevent overfitting, the next layer is a Dropout layer of 0.1. The output is normalized by the Dense layer, which is the final layer, using the Sigmoid activation function. Many traditional classifiers, including KNN with n-neighbors equal to 3 and MultinomialNB, are trained to compare results in addition to the BiLSTM model.

Transformer Model

A pre-trained model called BERT transformer was made public by Google AI language. The contextual relationship between the words in a phrase is learned using attention models. It is primarily made up of two components: an encoder that encrypts the input text and a decoder that decodes the output data following the job. The Simple Transformers library an API built on top of the Hugging face library, was used to create the best- base-cased model. The Bert cased model has been trained on the 2500 million words of the English Wikipedia and the 800 million words of BookCorpus. Hyper- parameters such as 300 sequence length, three epochs, 32 batch size, 4e-5 learning rate, and AdamW as the optimizer was used to train the model.

System Architecture

The system architecture demonstrates the process used to classify mail as spam or not (ham). The architecture consists of several processes, including the uploading of a dataset, preparing the data by dividing it into training and testing sets, training the model by using the appropriate classification algorithm, and lastly determining whether or not the supplied email is spam.

Technology Used in Project

User Interface

This system's user interface was developed using the python Tkinter module and provides a user-friendly graphical user interface.

Hardware Interfaces

Python features are used to enable interaction between the user and the console.

Software Interfaces

The required software is python.

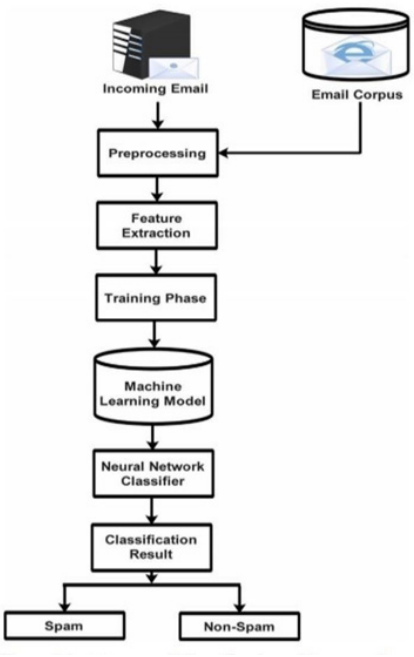

Figure 1: Architecture of Classification

Windows 10 Operating Environment

Hardware Requirements

a) Processor - Pentium –IV
b) Speed - 1.1 Ghz
c) RAM - 256 MB(min)
d) Hard Disk - 6 GB and above
e) Key Board - Standard Windows Keyboard
f) Mouse - Two or Three Button Mouse
g) Monitor - SVGA

Software Requirements

h) Operating System - Windows 10
i) Programming Language-Python 3.8(Pycharm Editor)

CONCLUSION

Today, email is the most significant form of communication because it allows for the delivery of any message anywhere on the globe thanks to internet connectivity. Every day, more than 270 billion emails are sent and received, of which roughly 57% are spam. Spam emails also referred to as "non- self," are unwanted commercial or harmful emails that damage or compromise personal information, such as bank account information, financial information, or anything else that harms a single person, a business, or a group of people. In addition to advertisements, they could have connections to websites hosting phishing or malware created to steal personal data. Spam is a severe problem that ends consumers find bothersome but is also financially harmful and a security risk. Therefore, this system is created so that it can identify undesired and unsolicited emails and stop them, aiding in the decrease of spam messages, which would be extremely beneficial to both individuals and the business. In the future, this system can be developed using various algorithms, and it can also get new features added to it.

REFERENCES

Shukor Bin Abd Razak, Ahmad Fahrulrazie Bin Mohamad (*2013).* Identification of Spam Email Based on Information from Email Header. 13th *International Conference on Intelligent Systems Design and Applications (ISDA).*

Mohammed Reza Parsei, Mohammed Salehi (2015). E-Mail Spam Detection Based on Part of Speech Tagging. *2 nd International Conference on Knowledge Based Engineering and Innovation (KBEI).*

Sunil B. Rathod, Tareek M. Pattewar (2015). Content Based Spam Detection in Email using Bayesian Classifier, presented at the *IEEE ICCSP conference.*

Aakash Atul Alurkar, Sourabh Bharat Ranade, Shreeya Vijay Joshi, Siddhesh Sanjay Ranade, Piyush A. Sonewa, Parikshit N. Mahalle, Arvind V. Deshpande (2017). A Proposed Data Science Approach for Email Spam Classification using Machine Learning Techniques.

Kriti Agarwal, Tarun Kumar (2018). Email Spam Detection using integrated approach of Naïve Bayes and Particle Swarm Optimization, *Proceedings of the Second International Conference on Intelligent Computing and Control Systems (ICICCS).*

Cihan Varol, Hezha M.Tareq Abdulhadi (2018). Comparison of StringMatching Algorithmson Spam Email Detection, International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism Dec.

Duan, Lixin, Dong Xu, and Ivor Wai-Hung Tsang. (2012). Domain adaptation from multiple sources: A domaindependent regularization approach. *IEEE Transactions on Neural Networks and Learning Systems 23.3.*

Mujtaba, Ghulam, *et al.* (2017). Email classification research trends: Review and open issues. *IEEE Access 5*

Trivedi, Shrawan Kumar (2016). A study of machine learning classifiers for spam detection Computational and Business Intelligence (ISCBI), *4th International Symposium on. IEEE.*

You, Wanqing, *et al.* (2015). Web Service-Enabled Spam Filtering with Naïve Bayes Classification. *IEEE First International Conference on Big Data Computing Service and Applications (BigDataService). IEEE.*

Rathod, Sunil B., and Tareek M. Pattewar. (2015). Content based spam detection in email using Bayesian classifier. *International Conference on. IEEE.*

Sahın, Esra, Murat Aydos, and Fatih Orhan. (2018). Spam/ham e-mail classification using machine learning methods based on bag of words technique. 26th *Signal Processing and Communications Applications Conference (SIU). IEEE, 2018.*

Kuldeep Vayadande, Aditya Bodhankar, Ajinkya Mahajan, Diksha Prasad, Shivani Mahajan, Aishwarya Pujari and Riya Dhakalkar (2022). Classification of Depression on social media using Distant Supervision, *ITM Web Conf. 50.*

Kuldeep Vayadande, Rahebar Shaikh, Suraj Rothe, Sangam Patil, Tanuj Baware and Sameer Naik, (2022). Blockchain-Based Land Record Syste M, *ITM Web Conf. 50.*

Samruddhi Mumbare, Kunal Shivam, Priyanka Lokhande, Samruddhi Zaware, VaradDeshpande and Kuldeep Vayadande, (2022). Software Controller using Hand Gestures, *ITM Web Conf. 50.*