Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, January 2023

Spam/Ham Email Classification Using Machine Learning

Rajeev Yadav
Professor
Computer Science Engineering
Arya Institute of Engineering and Technology, Jaipur, Rajasthan

Ramakant Gzautam
Assistant Professor
Computer Science Engineering
Arya Institute of Engineering and Technology, Jaipur, Rajasthan

Dimpy Arora
Asst. Professor
Engineering Mathematics
Arya College of Engineering and Technology, Jaipur

Payal Rathore
Science Student
Sant Meera Convent Senior Secondary School, Pratapgarh, Rajasthan

Akshat Pareek
Science Student
A's Steward Morris School, Bhilwara, Rajasthan

Abstract

Email stands for Electronic mail which means it is the way to distribute messages by electronic means from a computer source to one or more by means of network. It is one of the fastest method of distributing messages between sender's and receiver's computer system via Internet. In these days, email has become the most frequent communication system and a significant portion of the population depends on accessible email or texts from strangers. One

http://doi.org/10.36893/IEJ.2023.V52.195-201

of the biggest issues email consumers confront is the increase of spam communications. The tools that determine if an email is spam or not are known as spam filters. Currently there are many spam filter tools available on internet. Identifying these spammers is one of the hot topic of research and arduous tasks. In this paper we will look into the spam filter techniques using machine learning. Logistic regression algorithm and the spam emails will be categorized using the Natural Language Toolkit. The dataset utilized comes from the public datasets of Apache Spam Assassin and includes samples of spam and ham. The accuracy is defined on the basis of score of cross validation, precession and recall.

Keywords: Spam, Machine learning, Logistic regression algorithm, Natural Language Toolkit, Apache Spam Assassin's public datasets.

Introduction

One of the most efficient and well-organized ways to share or exchange information with others is via email [9]. Spam email influx is currently a big issue for web users and web services [8]. Unwanted or unsolicited emails that a recipient gets without prior notification of the sender are referred to as spam [6]. Spam emails are used for email spoofing, virus distribution, financial fraud, and marketing. Therefore, it can lead to serious issues for email users, such as excessive network traffic, compromised data, and time wastage. To stop unsolicited mailings, a prompt and efficient mail screening strategy is required [8].

Filtering is a method of organizing emails that involves removing viruses and removing spam. As a result, it can lead to serious issues for email users, such as data breaches, time wastage, and network traffic. A timely and well-organized mail filtering approach is necessary to prevent unsolicited mailings [8].

Filtering is the process of organizing emails such that just ham mail can pass through and viruses are eliminated [7]. Ham emails are another name for non-spam emails. Currently, categorization is used in spam filtering techniques. Data mining uses the classification technique. Identifying important knowledge from a large dataset is known as data mining. Finding a model that represents and distinguishes between several classes or generalizations of data is the process of classification. A model is the portion of the machine learning system that is capable of learning and making predictions. Analysis serves as the basis for training the models,a collection of class objects with known class labels that belong to various classes.

http://doi.org/10.36893/IEJ.2023.V52.195-201

Important data class concepts are described by models that are extracted through classification. There are two main processes in the classification process. The learning phase comes first, during which a categorization model must be created. Using the classification algorithm, the training data is examined in this stage. The second stage is classification, in which the model is used to forecast class labels based on the data. Data is needed in this phase to estimate the categorization rules' accuracy [1]

Objects belonging to distinct classes are classified using machine learning methods. These algorithms do a pretty good job of identifying spam from ham emails. The classification in this study is done using the Logistic Regression technique. The experiment's findings demonstrate the accuracy of the trained model, where accuracy is determined by using cross-validation and precession-recall[6].

Methodology

There are many methods that are proved to filter spam emails. In this paper, Logistic Regression algorithm is used for classification [7]. The dataset is taken from the examples of Apache Spam Assassin's public datasets. Firstly, download the examples of spam and ham from Apache Spam Assassin's public datasets using libraries: we get there are 2500 ham emails in the file where as 500 spam emails To parse these emails, we may utilize Python's email module, which takes care of encoding and headers. To get a sense of what the data looks like, let's look at some samples of both spam and ham emails. Some emails are truly multipart, with photos and attachments (which may have their own attachments). After that, have a look at the many kinds of structures we have. It appears that spam emails contain a significant amount of HTML, whereas ham emails are often plain text [6]. Furthermore, very few spam emails are signed using PGP, but a lot of ham emails are. To sum up, it appears that knowing the email structure is helpful. Divide the dataset into a training and a test set before you learn too much about it [7]. Using the train test split() function from the learn.model selection package, divide the data. We will now require a method to convert HTML to plain text. Using the excellent Beautiful Soup package is arguably the best way to accomplish this.. Then write a function that takes an email as input and returns its content as plain text. Then use nltk (Natural Language Toolkit) to throw in some stemming [6]. All of this is ready to be assembled into a transformer so that emails can be converted to word counters. Take note that the split () method in Python, which employs whitespaces as word borders, is used to divide

http://doi.org/10.36893/IEJ.2023.V52.195-201

sentences into words. Not all written languages can use this, but many can. For instance, Vietnamese script frequently employs gaps even between syllables, whereas Chinese and Japanese scripts typically do not. Since the dataset in this experiment is (largely) in English, it's OK. The word counts are now available, and these must be converted to vectors [6]. To do this, we'll construct an additional transformer, the fit () method of which will generate the vocabulary (an ordered list of the most frequently used terms), and whose transform () function will turn word counts into vectors by using the vocabulary. A sparse matrix is the result [7] For instance, the 99 in the first column of the second row indicates that 99 of the terms in the second email are not in the lexicon. The number 11 next to it indicates that this email contains 11 instances of the vocabulary term.

Using the *.vocabulary_() method, you may view the vocabulary to see the words we are discussing. "The" is the first word, "of" is the second, etc. We may now begin training our initial spam classifier! Transform the entire dataset now [6]. Utilizing logistic regression, train the spam classifier. A typical method for determining the likelihood that an instance belongs to a specific class is logistic regression. The model predicts that the instance belongs to that class (referred to as the positive class, labeled "1") if the estimated probability is higher than a predetermined threshold, which is usually 50% ,if not, it anticipates that it won't (i.e., it falls into the negative class, designated "0"). It is a binary classifier as a result. As opposed to producing the result immediately, as the linear regression model does, a logistic regression model produces the logistic of the weighted sum of the input characteristics (plus a bias component) [6].

The logistic—noted σ (·)—is a sigmoid function (i.e., S-shaped) that outputs a number between 0 and 1.

Equation 1.1 Logistic function

$$\sigma(t) = \frac{1}{1 + e^{-t}}$$

http://doi.org/10.36893/IEJ.2023.V52.195-201

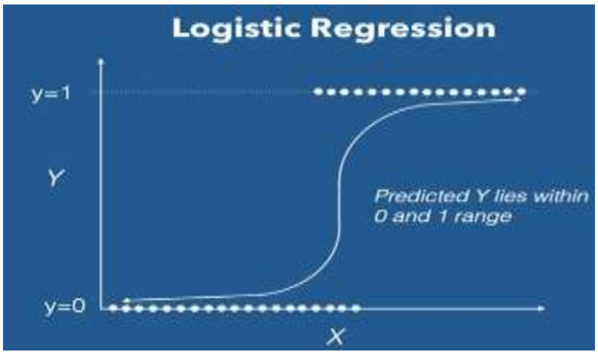

Figure 1. Logistic Regression

The scoring function for the cross-validation score is really the opposite of the RMSE because Scikit-Learn's cross-validation features anticipate a utility function (larger is better) rather than a cost function (lower is better). Since the value is negative, you must change the output's sign in order to obtain the RMSE scores[7]. Verify the trained model's recall and precession as well. Recall is a metric that measures the quantity of items that a model successfully classifies, whereas precession of the classifier refers to the accuracy of the positive predictions.TABLE 1. COMPARISON TO PRIOR RESEARCH

Method	Result
Cost-sensitive three-way decision[2]	89.88%
NSA-DE[3]	83.06%
NSA-PSO[4]	91.22%
NSA-PSO[5]	83.20%
Bayes' theorem and NaïveBayes'Classifier[9]	96.06%
Naïve Bayes, J48 Decision Tree[8]	96.13%
Proposed method[LR]	98.50%

Result

In this paper of spam classification, Logistic regression is used. Email textual data, including the subject line and message body, is examined.

The efficiency of ML algorithm Logistic regression are tested for randomly splits k-folds using cross validation. And the cross validation score of this trained model is 0.985 which means the score is over 98.5% .And the precession is:96.88% & the recall is:97.89%.

http://doi.org/10.36893/IEJ.2023.V52.195-201

Conclusion

These days, email is the most important form of communication because it can be sent anywhere in the globe thanks to internet connectivity. Every day, about 270 billion emails are exchanged, with spam accounting for roughly 57% of them. Spam emails are unsolicited, malicious emails that breach personal information, compromise banks, or do anything else that could harm a person, a business, or a group of people. These emails could include links to phishing hosting websites, which are designed to collect personal data, in addition to advertisements.

Spam is a major problem that poses a risk to users' security, causes financial harm, and annoyance. As a result, this system is built to identify and stop spam and unsolicited emails, which will assist reduce the amount of unsolicited messages and be very advantageous to both the company and the people using it.Different algorithms may be used to construct this system in the future, and the current system may receive additional functionality.

References

1. Proceedings of The International Conference on Innovations in Intelligent Systems and Computing Technologies, Philippines, “Text Mining Approach to Detect Spam in Emails,” no. February, 2016.
2. B. Zhou, Y. Yao, and J. Luo, “Cost-sensitive three way email spam filtering,” J. Intell. Inf. Syst., vol. 42, no. 1, pp. 19–45, 2014.
3. Idris, A. Selamat, and S. Omatu, “Hybrid email spam detection model with negative selection algorithm and differential evolution,” Eng. Appl. Artif. Intell., vol. 28, pp. 97–110, Feb. 2014.
4. Idris and A. Selamat, “Improved email spam detection model with negative selection algorithm and particle swarm optimization,” Appl. Soft Comput., vol. 22, pp. 11–27, 2014.
5. Idris, A. Selamat, N. Thanh Nguyen, S. Omatu, O. Krejcar, K. Kuca, and M. Penhaker, “A combined negative selection algorithm-particle swarm optimization for an email spam detection system,” Eng. Appl. Artif. Intell., vol. 39, pp. 33–44, 2015.
6. Aurelien Geron, “Hands on Machine Learning with Scikit Learn, Keras and TensorFlow”

http://doi.org/10.36893/IEJ.2023.V52.195-201

7. Adi Wijaya, Achmad Bisri, "Hybrid Decision Tree and Logistic Regression Classifier for Email Spam Detection",2016
8. Thashina Sultana, K A Sapnaz, Fathima Sana, Mrs. Jamedar Najath, "Email based Spam Detection", Vol. 9 Issue 06, June-2018
9. Anju Radhakrishnan , Vaidhehi V, "Email Classification Using Machine Learning Algorithms", Vol 9 No 2 Apr-May 2017
10. Shukor Bin Abd Razak, Ahmad Fahrulrazie Bin Mohamad "Identification of Spam Email Based on Information from Email Header" 13th International Conference on Intelligent Systems Design and Applications (ISDA), 2013.
11. R. Kaushik, O. P. Mahela, P. K. Bhatt, B. Khan, S. Padmanaban and F. Blaabjerg, "A Hybrid Algorithm for Recognition of Power Quality Disturbances," in *IEEE Access*, vol. 8, pp. 229184-229200, 2020.
12. Kaushik, R. K. "Pragati. Analysis and Case Study of Power Transmission and Distribution." *J Adv Res Power Electro Power Sys* 7.2 (2020): 1-3.
13. R. Kaushik, O. P. Mahela and P. K. Bhatt, "Hybrid Algorithm for Detection of Events and Power Quality Disturbances Associated with Distribution Network in the Presence of Wind Energy," *2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, Greater Noida, India, 2021, pp. 415-420.
14. P. K. Bhatt and R. Kaushik, "Intelligent Transformer Tap Controller for Harmonic Elimination in Hybrid Distribution Network," *2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, India, 2021, pp. 219-225
15. Kumar, G., Kaushik, M. and Purohit, R. (2018) "Reliability analysis of software with three types of errors and imperfect debugging using Markov model," International journal of computer applications in technology, 58(3), p. 241. doi: 10.1504/ijcat.2018.095763.
16. Sharma, R. and Kumar, G. (2017) "Availability improvement for the successive K-out-of-N machining system using standby with multiple working vacations," International journal of reliability and safety, 11(3/4), p. 256. doi: 10.1504/ijrs.2017.089710.

http://doi.org/10.36893/IEJ.2023.V52.195-201