International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 05 | May - 2023 SJIF 2023: 8.176 ISSN: 2582-3930

Email Spam Detection Using Machine Learning

Prof. Abhimanyu Dutonde[1] , Prashik .P. Ramteke[2], Akanksha Burade[3] , Rajani Bansod[4], Pratham Telang[5]

Project Guide, Department of Computer Science Engineering, Tulsiramji Gaikwad-Patil College of Engineering & Technology, Nagpur.

Students[234567], Department of Mechanical Engineering, Tulsiramji Gaikwad-Patil College of Engineering & Technology , Nagpur.

Abstract -

Nowadays, Email spam has become a big problem, with the fast growth of internet users, email spams are also increasing. People are using them for phishing, illegal and unethical practices and frauds. Sending malicious links through spam emails that can harm for our system and may also get into your system. It is very simple for spammers to create a fake profile and email account, they show like a real person in their spam emails, these spammers simply target people who are not aware of these frauds. then there is a need to identify those spam mails which are frauds, this project will identifies those spams using techniques of machine learning, this paper will discuss machine learning algorithm's and apply all these algorithm's to our dataset. it select the best algorithm, for this project algorithm will be chosen based on the best accuracy and precision in email spam detecting.

Keywords— Email Spam Detection, Machine Learning, Neural Networks, Support Vector Classifier, Logistic Regression etc.

1. Introduction

Technology has become a vital part of life in today's time. With each passing day, the use of the internet increases exponentially, and with it, the use of email for the purpose of exchanging information and communicating has also increased, it has become second nature to most people. While e-mails are necessary for everyone, they also come with unnecessary, undesirable bulk mails, which are also
called Spam Mails [29]. Anyone with access to the internetcan receive spam on their devices. Most spam emails divert people's attention away from genuine and important emails and direct them towards detrimental situations. Spam emails are capable of filling up inboxes or storage capacities, deteriorating the speed of the internet to a great extent. These emails have the capability of corrupting one's system by smuggling viruses into it, or steal useful information and scam gullible people. The identification of spam emails is a very tedious task and can get frustrating sometimes.

While spam detection can be done manually, filtering out a large number of spam emails can take very long and waste a lot of time. Hence, the need for spam detection softwares has become the need of the hour. To solve this problem, various spam detection techniques are used now. The most common technique for spam detection is the utilization of Naive Bayesian [5] method and feature sets that assess the presence of spam keywords. The main purpose is to demonstrate an alternative scheme, with the use of Neural Network (NN) [4] classification system that utilises a collection of emails sent by several users, is one of the objectives of this research. One other purpose is the development of spam detection with the help of Artificial Neural Networks, resulting in almost 98.8% accuracy.

2. Problem Identification

Email or electronic mail spam refers to the "using of email to send unsolicited emails or advertising emails to a group of recipients. Unsolicited emails mean the recipient has not granted permission for receiving those emails. "The popularity of using spam emails is increasing since last decade. Spam has become a big misfortune on the internet. Spam is a waste of storage, time and message speed. Automatic email filtering may be the most effective method of detecting spam but nowadays spammers can easily bypass all these spam filtering applications easily. Several years ago, mos t of the spam can be blocked manually coming from certain email addresses. Machine learning approach will be used for spam detection. Major approaches adopted closer to junk mail filtering encompass "text analysis, white and blacklists of domain names, and community-primarily based techniques". Text assessment of contents of mails is an extensively used method to the spams. Many answers deployable on server and purchaser aspects are available.

3. Objectives

There are four objectives that need to be achieved in this project:

- To study on how to use machine learning techniques for spam detection.
- To modify machine learning algorithm in computer system settings.
- To leverage modified machine learning algorithm in knowledge analysis software.
- To test the machine learning algorithm real data from machine learning data repository.

4. Literature Review

Most research has been conducted into detecting and filtering spam email using a variety of techniques.

- Thiago S. Guzella et. Al (2009) has conducted "A Review of Machine Learning Approaches to Spam Filtering". In their paper, they found that Bayesian Filters that are used to filter spam required a long training period for it to learn before it can completely well function.
- S. Ananthi (2009) has conducted a research on "Spam Filtering using K-NN". In this paper, she used KNN as it is one of the simplest algorithm. An object is classified by a majority vote of its neighbours where the class is typically small.
- Anjali Sharma et. Al (2014) has conducted "A Survey on Spam Detection Techniques". In this paper, they found that Artificial Neural Network (ANN) must be trained first to categorize emails into spam or non-spam starting from the particular data sets.
- Simon Tong and Daphne Koller (2001) has conducted a research on "Support Vector Machine Active Learning with Applications to Text Classification". In this paper, they presented new algorithm for active learning with SVM induction and transduction. It is used to reduce version space as much as it can at every query. They found out that the existing dataset only differ by one instance from the original labelled data set.
- Minoru Sasaki and Hiroyuki Shinnou (2005) has conducted a research on "Spam Detection Using Text Clustering". They used text clustering based on vector space model to construct a new spam detection technique. This new spam detection model can find spam more efficiently even with various kinds of mail.
- Aigars Mahinovs and Ashutosh Tiwari (2007) has conducted a research on "Text Classification Method Review". They test the process of text classification using different classifier which is natural language processing, statistical classification, functional classification and neural classification. They found that all the classifier works well but need more improvement especially to the feature preparation and classification engine itself in order to optimise the classification performance.
- Email system is one of the most common and popular communication systems. Organisations from all over the world are making their efforts in order to identify the spam mails. The work of authors to identify the ham and spam emails is discussed here. Table 1 illustrates the comparative work of authors by stating the classification techniques, dataset, feature extraction approaches and drawbacks.

5. Proposed System

5.1. Process Model

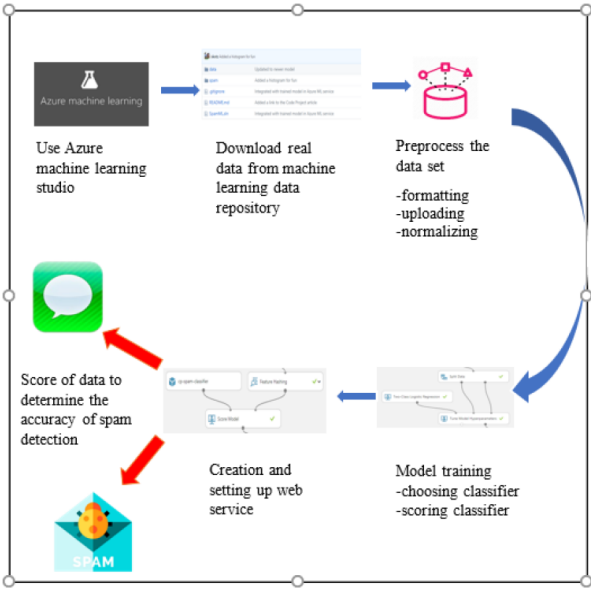

Fig. 1. Process framework

Process model is a series of steps, concise description and decisions involved in order to complete the project implementation. In order to finish the project within the time given, the flows of project need to be followed. The framework below shows how the overall flow of this project in order to separate between a spam and ham message.

5.2. Data Model

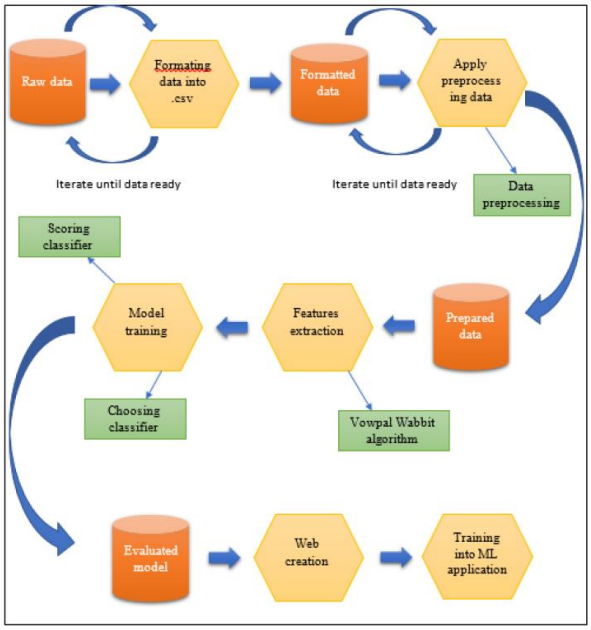

Figure. 1 Data model flow

The data model flow is essential to this project to show the structured of the project on how it should be built and how the process is related to each other. It helps to make organize the process of project smoothly and clearly.

Based on the framework, Azure ML Studio is used as the platform to develop the project. First, study and discovered all the functionality on Azure ML to make sure the project can achieve its objectives. After that, make sure to download and used the real existing dataset from machine learning data repository as training and testing data.

Preprocessing data start with reformat the dataset into 2 separate files which is training.csv and testing.csv format. Then, upload the formatted dataset into Azure ML under dataset function/menu and drag them onto the workspace to visualize the data. Choose any desired filters to clean the raw data such as “remove numbers” filter.

6. Implementation

All the implementation process of the spam detection by using machine learning based binary classifier project will be presented.

6.1. Deployment

After spam detection ML model has been trained, an API key will be generated by the server in order to deploy the web service.

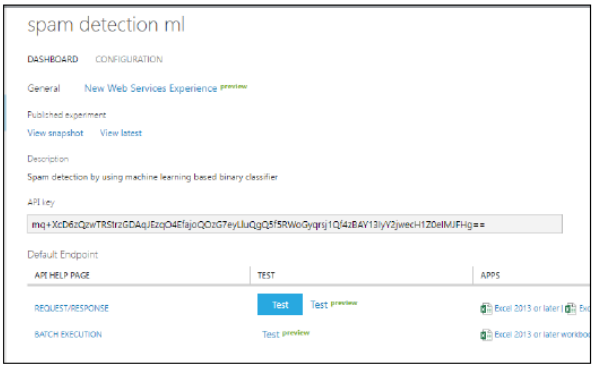

Figure.3. API generated by ML studio server

Then, the API will be entered into API key form from the VW algorithm using Visual Studio

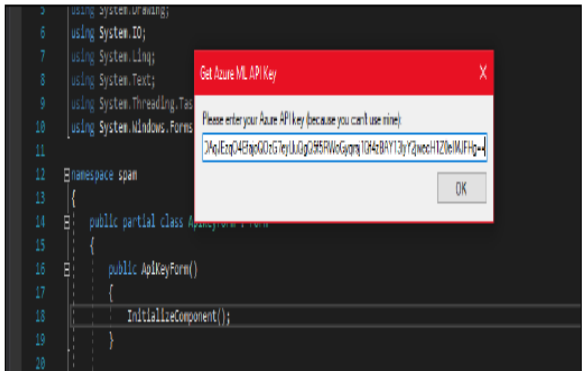

Figure 4 Insert API key into form in VS 28

After the API key has been confirmed by server, the spam ML web service will be shown as below.

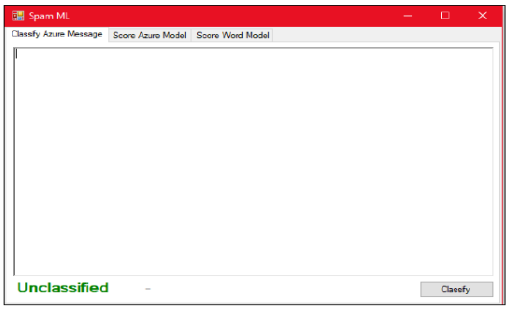

Figure 5 Spam ML web service

WordSearchClassifier then created to test the accuracy of detection by using VW algorithm which works by dividing messages into bigrams.

7. Results & Discussion

the result for spam probability, time elapsed and comparison of spam detection using different malware detection is presented. This section presented the results based on experiments and study of this project. The entire graph above focused on the comparison of detection using different malware detection tools which is Joe Sandbox cloud, hybrid analysis (Falcon sandbox) and visual studio.

A. Classification And Probability

The probability of classification is measured by counting the number of spam flags. According to Dr. Matt Peters, google has examined a great number of potentials factors that predicted that a site might be penalized or banned due to spam [9]. Each flag has its own warning sign that indicates the message as spam. So, to calculate this probability, spam score will records the quantity of flags that triggers the data. Hence, the graph below shows the relationship that numbers of flags effect the probability of classification type. The overall likelihood of spam increases as the number of flags increases.

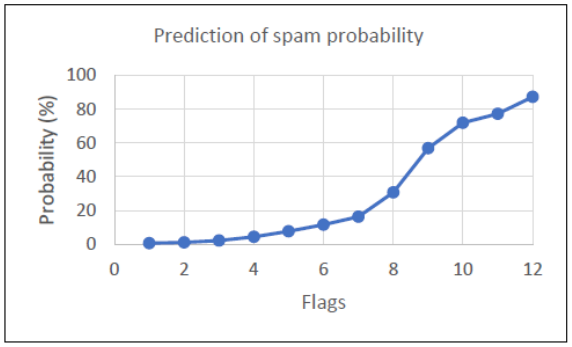

Figure 6 Relationship between number of flags and classification type

b. Elapsed Time And Message Count

Elapsed time is time different or amount of time between the beginning and the end of execution process. In simplest terms, elapsed time is the processing time of a process or event. In this project, both elapsed time and message count are taken into consideration in order to score the accuracy. This is to ensure the efficiency of model by decreasing the processing time even when the messages counts are large. As shown in the graph below, it shows the comparison of elapsed time using same messages between four different tools.

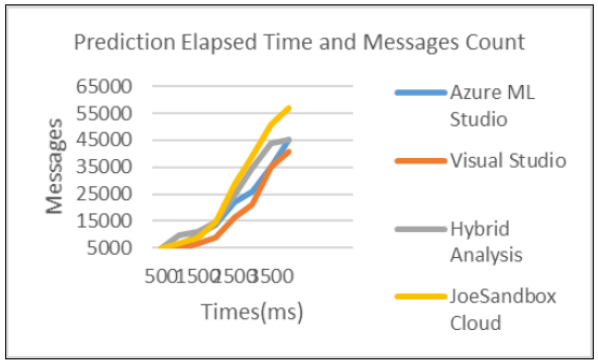

Figure 7 The relationship between message count and elapsed time

c. Accuracy and Message Count

Based on research, the message count or frequency of words is calculated in order to get the most accurate percentage of accuracy. This is because, the messages are the important element to test spam detection. Figure below shows that all the tools used verified that accuracy of detection affected by the messages count.

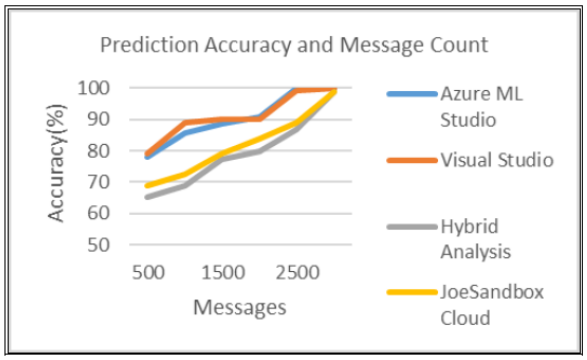

Figure 7 Relationship between accuracy and message count

d. Accuracy and Elapsed Time

The relationship between elapsed time and accuracy also take into consideration. Sometime, shorter time does not mean more accurate. The time affects the accuracy by processing as much as possible data.

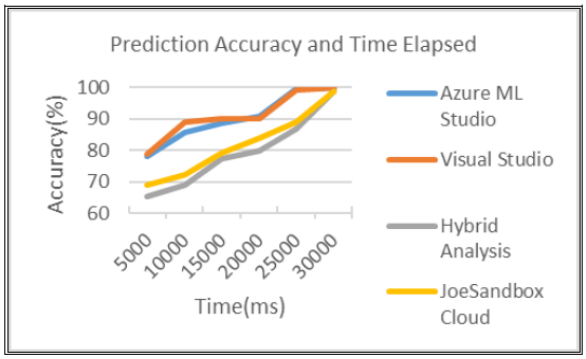

Figure 8 Relationship between accuracy and elapsed time

Researchers have become increasingly interested in spam detection and filtering over the last two decades. Several studies have been conducted in this area because of its substantial impact on a variety of areas, such as consumer behavior or fake reviews. In the study, lessons learned from each machine learning category are compared with previous approaches. Additionally, spam filters find it challenging to evaluate features from multiple angles, including temporal, writing style, semantic and statistical ones. Models are trained primarily on balanced datasets, while self-learning models are not feasible. Deep fake is another challenge facing spam detection systems. According to the findings of this study, most proposed spam email detection techniques are based on supervised machine learning techniques. This project provides an in-depth analysis of these Logistic Regression algorithm and some future directions for searching and detecting spam email.

8. Conclusion

The performance of a classification technique is affected by the quality of data source. Irrelevant and redundant features of data not only increase the elapse time, but also may reduce the accuracy of detection. Each algorithm has its own advantages and disadvantages as stated in Chapter 2. As state before, supervised ML is able to separate messages and classified the correct categories efficiently. It also able to score the model and weight them successfully. For instances, Gmail's interface is using the algorithm based on machine learning program to keep their users' inbox free of spam messages.

During the implementation, only text (messages) can be classified and score instead of domain name and email address. This project only focus on filtering, analysing and classifying message and do not blocking them. Hence, the proposed methodology may be adopted to overcome the flaws of the existing spam detection.

References

1) Anitha, PU & Rao, Chakunta & , T.Sireesha. (2013). A Survey On: E-mail Spam Messages and Bayesian Approach for Spam Filtering. International Journal of Advanced Engineering and Global Technology (IJAEGT). 1. 124-136.

2) Attenberg, J., Weinberger, K., Dasgupta, A., Smola, A., & Zinkevich, M. (2009, July). Collaborative email-spam filtering with the hashing trick. In *Proceedings of the Sixth Conference on Email and Anti-Spam*.

3) Awad, W. A., & ELseuofi, S. M. (2011). Machine learning methods for spam e-mail classification. *International Journal of Computer Science & Information Technology (IJCSIT)*, *3*(1), 173-184.

4) Barnes, J. (2015). Azure Machine Learning. *Microsoft Azure Essentials. 1st ed, Microsoft*.

5) Chang, M. W., Yih, W. T., & Meek, C. (2008, August). Partitioned logistic regression for spam filtering. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 97-105). ACM.

6) Çıltık, A., & Güngör, T. (2008). Time-efficient spam e-mail filtering using n-gram models. *Pattern Recognition Letters*, *29*(1), 19-33.

7) Crawford, M., Khoshgoftaar, T. M., Prusa, J. D., Richter, A. N., & Al Najada, H. (2015). Survey of review spam detection using machine learning techniques. Journal of Big Data, 2(1), 23.

8) Dai, W., Xue, G. R., Yang, Q., & Yu, Y. (2007, July). Transferring naïve bayes classifiers for text classification. In *AAAI* (Vol. 7, pp. 540-545).

9) Fishkin, R. (2015, November 06). Spam Score: Moz's New Metric to Measure Penalization Risk. Retrieved from https://moz.com/blog/spam-score-mozsnew- metric-to-measure-penalization-risk

10) Guzella, T. S., & Caminhas, W. M. (2009). A review of machine learning approaches to spam filtering. Expert Systems with Applications, 36(7), 10206- 10222.

11) Ikonomakis, M., Kotsiantis, S., & Tampakas, V. (2005). Text classification using machine learning techniques. WSEAS transactions on computers, 4(8), 966-974.

12) Introduction | ML Universal Guides | Google Developers. (n.d.). Retrieved from https://developers.google.com/machine-learning/guides/textclassification/

13) Jindal, N., & Liu, B. (2007, May). Review spam detection. In Proceedings of the 16th international conference on World Wide Web (pp. 1189-1190). ACM.

14) Joachims, T. (2002). Learning to classify text using support vector machines: Methods, theory and algorithms (Vol. 186). Norwell: Kluwer Academic Publishers.

15) Khan, A., Baharudin, B., Lee, L. H., & Khan, K. (2010). A review of machine learning algorithms for text-documents classification. *Journal of advances in information technology*, *1*(1), 4-20.

16) Kolari, P., Java, A., Finin, T., Oates, T., & Joshi, A. (2006, July). Detecting spam blogs: A machine learning approach. In Proceedings of the national conference on artificial intelligence(Vol. 21, No. 2, p. 1351). Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

17) Korde, V., & Mahender, C. N. (2012). Text classification and classifiers: A survey. *International Journal of Artificial Intelligence & Applications*, *3*(2), 85.
18) Lewis, D. D., & Gale, W. A. (1994). A sequential algorithm for training text \classifiers. In *SIGIR'94* (pp. 3-12). Springer, London.
19) Lewis, D. D., & Ringuette, M. (1994, April). A comparison of two learning algorithms for text categorization. In *Third annual symposium on document analysis and information retrieval*(Vol. 33, pp. 81-93).
20) Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., & Watkins, C. (2002). Text classification using string kernels. Journal of Machine Learning Research, 2(Feb), 419-444.
21) Mahinovs, A., Tiwari, A., Roy, R., & Baxter, D. (2007). Text classification method review.
22) Mund, S. (2015). *Microsoft azure machine learning*. Packt Publishing Ltd.
23) Rogati, M., & Yang, Y. (2002, November). High-performing feature selection for text classification. In *Proceedings of the eleventh international conference on Information and knowledge management* (pp. 659-661). ACM.
24) Sasaki, M., & Shinnou, H. (2005, November). Spam detection using text clustering.In 2005 International Conference on Cyberworlds (CW'05) (pp. 4- pp). IEEE.
25) Sasaki, M., & Shinnou, H. (2005, November). Spam detection using text clustering. In null (pp. 316-319). IEEE.

.