Email Spam Classification via Machine Learning and Natural Language Processing

Saraswathi Morthala #1, Ms R Madhuri Devi #2
#1 M.Tech., Scholar, #2 Head of the Department,
Computer Science and Engineering,
Priyadarshini Institute of Technology and Management

ABSTRACT

Today, spamming mails is one of the biggest issues faced by everyone in the world of the Internet. In such a world, email is mostly shared by everyone to share the information and files because of their easy way of communication and for their low cost. But such emails are mostly affecting the professionals as well as individuals by the way of sending spam emails. Every day, the rate of spam emails and spam messages is increasing. Such spam emails are mostly sent by people to earn income or for any advertisement for their benefit. This increasing amount of spam mail causes traffic congestion and waste of time for those who are receiving that spam mail. The real cost of spam emails is very much higher than one can imagine. Sometimes, the spam emails also have some links which have malware. And also, some people will get irritated once they see their inbox which is having more spam mails. Sometimes, the users easily get trapped into financial fraud actions, by seeing the spam mails such as job alert mails and commercial mails and offer emails. It may also cause the person to have some mental stress. To reduce all these risks, the system has proposed a machine learning model which will detect spam mail and non-spam emails, and also this system will optimize the data by removing the unwanted mails which contain the advertisement mails and also some useless emails and also some fraud mails. This proposed system will detect the spam mails and ham emails with the dataset consisting of spam mails and after identifying spam mails this system will remove that spam emails and this proposed system will calculate the amount of storage before and after the removal of spam mails.

INTRODUCTION

The major issues faced by all the email users are spam mails which contain unwanted information and data and some fake data to spoil the life of the people and also some mails which cause harmful effects. Today, the job issues are faced by fifty percent of the people by both educated and uneducated people. In such a case, these people will get emails about advertisement mails which are completely fake. But by seeing that mail, this people will get interested or have a thought to communicate through the mail for what they are looking into it. More people are affected by this spam mails in similar cases. To reduce this risk and to save the people from this danger of spam mails, we are proposing this system to remove the spam mails. For filtering the spam mails, in this system we are using two filtering model. Namely, Opinion Rank and NLP based n-grams model. By using these two models we will filter the spam mails and non-spam mails. And this system will optimize the data by removing the spam mails and also it calculates the storage of the mails. The finding of trust rank of the mail and classifying

those mails as spam and ham mails based on their content. And detection of advertisement mails in those mails. After detecting the advertisement mails, optimizing the storage by deleting those advertisement mails. By deleting mails, the proposed system will optimize the data. And also, the system will identify the fake mails which look similar to real mails that people can believe. The main objective of the project is to detect the spam mails and to optimize the data storage. This detection of spam mails in this proposed system is done through the two filtering models. One is Opinion Rank which is based on the trustworthiness of the mail id and this rank uses the two algorithms namely, high page rank and inverse page rank. By combining these results, and by calculating the mean of this results, the Opinion Rank will perform. And the data optimization is done by removing the advertisement mails with the help of Latent Dirichlet Allocation which is a probabilistic topic modelling to classify the contents or documents based on the topics. The proposed system of the project will effectively detect the spam mails and the system will extract the spam mails by using some machine learning algorithms and it gives the result with greater accuracy and with good performance. Also, this proposed system will optimize the data storage by blocking and deleting the spam mails. And with the help of the Opinion Rank model, this proposed system will find trustworthiness of the mail and it will carry the filtering of spam messages. This proposed system will save the user's time and it destroys the risk of spam mails.

LITERATURE SURVEY

In [2], an integrated approach involving all the three processes has more accuracy than any of the processes having a standalone approach (URL Analysis, NLP, ML). [3] further highlights the system of URL Classification and the Decision Tree algorithm is used, and the model is trained using a data-set from Phishtank. In [4], the URL is analyzed based on parameters like the number of special characters and dots. Random tree and KNN have the same numerical value for the parameters but KNN requires more time to build the model than random tree, hence random tree is the most efficient machine learning algorithm used to classify the emails. In [5], Random forest, K nearest neighbour, decision making tree algorithm and support vector machines algorithms were implemented; Enron Spam Project data-set was used to train the model. Furthermore, an add-on to outlook was created in C in a visual studio. Random Forest was the machine learning algorithm which was the most accurate. Initially, the model is trained for text classification from the data-sets available on Enron Email Corpus and CMU Corpus. In [6] A total of 32 parameters are taken into consideration for file classification. Support Vector Machine algorithm is the most accurate for both the stages i.e. text classification and file classification. In [7], the drawbacks of models based on term frequency were considered which leads to huge computational load and slow training speed due to the size of huge feature vector space. Implementation of a model based on semantic similarity which extracts semantic meanings layer by layer using effective information retrieval techniques. In [8], performance of different ML algorithms by using features of keyword extraction, uni-grams, big-rams and n-grams for classification of spam and fake online reviews. In [9],the measure of cosine similarity function

was used and applied only on specific parts of speech (POS) and by employing lemmatization algorithms and effects of different pre-processing algorithms on classification accuracy. In [10], a semantic feature space was created from training data using statistical methods and solving the problems of conventional neural networks using BP algorithm. Keyword based spam filtering model efficient feature selection method using adaptive learning rate is used. In [11], modern day Spam Statistics and different types of Spam attacks (Email-phishing, spear phishing and spoofing) have been studies, existing software's and scope of techniques for spam classification have been analyzed, performance metrics of various Supervised Learning algorithms are compared and extraction of email routing information from a spam source. In [12], spam management for SMS services using text mining applications such as Rapid Miner for use in classification and clustering algorithms have been focused upon, they have justified the use of SVMs and Naive Bayesian models as they have high performance even in the absence of large data and feature modelling and engineering. In [13], the main focus is on comparing four different machine learning algorithms for content based spam filtering technique, experimental results show that Neural Network classifier is more sensitive to the training set size and unsuitable for using alone as a spam rejection tool, generally, the performances of the SVM and RVM classifiers are less influenced by data sets and feature sizes, and obviously superior to the Naive Bayesian classifier. In [14] features for spam classification and spamming behavior of mails are discussed, the paper have presented a rule-based method for instantiating behavior-based features into discrete values, the paper presents the design and implementation of back-propagation neural networks for spam classification using behavior-based features. In [15] Spam classifier which is based on RF algorithm is used, they have used parameter optimization and feature selection , optimization of two parameters of Random Forest to maximize the Spam detection rates is done, and also provided the importance of individual feature selection. It can detect spam with low processing resources with high accuracy. In [16] research work on several State-of-the-art approaches used for various spam filtering methods. Different spam filtering formulas implemented by Gmail, Yahoo and Outlook; various performance evaluation measures are discussed on basis of which the performances are measured. All the machine learning algorithms used for filtering are discussed with their comparative analysis, the paper also includes various open research problems faced by existing techniques. [24] Highlights an algorithm where SVM and KNN both were implemented for Chinese web page classification, it improved the classifying predictability and gave a better feedback.

SYSTEM ANALYSIS

In an integrated approach involving all the three processeshas more accuracy than any of the processes havinga standalone approach (URL Analysis, NLP, ML). Furtherhighlights the system of URL Classification and the DecisionTree algorithm is used, and the model is trained using a data-setfrom Phishtank. In the URL is analyzed based onparameters like the number of special characters and dots.Random tree and KNN have the same numerical value for theparameters but

KNN requires more time to build the modelthan random tree, hence random tree is the most efficientmachine learning algorithm used to classify the emails.

The project presents the design and implementationof back-propagation neural networks for spam classificationusing behaviour-based features. In Spam classifier whichis based on ML algorithm is used, they have used parameteroptimization and feature selection, optimization of two parametersof Random Forest to maximize the Spam detectionrates is done, and also provided the importance of individualfeature selection. It can detect spam with low processingresources with high accuracy.

METHODOLOGY

Spam Detection Using N-Grams Model

Spam emails occupy a lot of storage and hence it is detected using the N-Grams Model in which it detects the spam by analyzing the n set of words. When the n set of words occur it is detected as spam or ham using the frequency and probability.

Data-set

5000 input mails are taken from kaggle and tested for spam using the NLP N-Grams Model. Also, it is tested with the personal mail. The figure 4.1 shows the data-set for spam detection.

Data Pre-processing

Data pre-processing is an important step in this. The data that is given to the model will affect the performance of the model. Therefore the data is processed before proceeding. In this model the punctuation's will be removed for better prediction. The data which will be given to this model will be of string data type. This will be fed to the process which will identify each of the characters and the punctuation's will be removed. This refers to the processes of identifying and correcting the errors that are present in the data-set that may negatively impact the model. In this process the stop words will be removed. Generally the stop-words will be removed. A stop-word is a word which is usually the most used words in the natural language.

Tokenization And Lemmatization

The next process in this is the tokenization process. In tokenization larger text is into smaller words. That is the will be split individually and will be put into appropriate data type. These tokenized words will be then used for the next purpose. Lemmatization isthe process of converting a word into its natural base form. This will aim to remove the inflectional ending and return the base word.

Visualization of spam messages

Email messages are separated into spam and ham from the data-set and it is visualized using word cloud. Using this user can easily visualize the words that commonly occurred in spam messages and in ham messages.

Vectorization using TF-IDF

TF-IDF refers to Term Frequency and Inverse Document Frequency. Term Frequency refers to how frequent a word appears in a document divided by total number of words in the document.

TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document)

Inverse Document Frequency refers to the importance of a term in a document which is calculated by taking the logarithm of the number of documents in a corpus which is divided by how many times the specific pattern appears.

IDF(t) = log_e(Total number of documents / Number of documents with term t in it)

Detection of spam using Bi-gram

Bi-gram is the two-word sequence of n-grams where n refers to 2 here. Bi-gram is the model which predicts the following word using the occurrence of the previous word in a document. Here, using the occurrence of words in the mail data-set bi-gram predicts whether the mail is spam or ham.

Detection of spam or ham

The system displays whether the email is spam or ham by taking the set of n-grams and analyzing whether the set of words occurs in the spam message or in a ham message. Like any other Machine Learning or Artificial Intelligence models, the model should be trained with a huge corpus of data. After training this N-grams model, the system will have an idea on the probability of the spam or ham message by analyzing the sequence of words in that message.

To measure the probability of spam or ham in a bi-gram model, the occurrence of every two words has been analyzed by the model and then determines whether that is a spam or ham. The predictions will get improved when we give a bigger corpus. In a bi-gram the probability of spam and ham is classified by using the following formula.

- Probability of spam = count(word2 spam) / count(word 2)
- Probability of ham = count(word2 ham) / count(word 2)

Probability calculation is done by calculating the probability of the word spam or ham occurring after the word w2 which is how many times the word occurs in the required sequence and it is divided by number of times the word before the expected word occurs in the corpus. For

example, if the given sentence is cash offer, then the following probabilities are calculated, P(null, cash), P(cash, offer), P(offer, spam). P(offer, spam) is calculated by,

- Probability of spam=count(number of times the word offer occurs in the given sentence * number of times the word spam occurs i.e 1 if mentioned as spam or 0 if mentioned as ham) / number of times the word offer occurs in the entire corpus which is mentioned as spam.

CONCLUSION

A comprehensive and efficient spam classification system has been created which follows a two step methodology to completely ensure that the mail received is spam or not. Initially, text classification takes place which is followed by URL analysis and filtering in order to determine if any link present in the mail is malicious or not. For text classification, machine learning algorithms were studied and analyzed. Various data-sets have been referred to for a list of spam trigger words and a list of blacklisted URLs. This model was hosted as an API which was then called by the java-script code in the google apps script in order to classify mails in real time in gmail.

REFERENCES

[1] Statista, accessed 3 November2020, https://www.statista.com/statistics/255080/number-of-e-mail-usersworldwide/

[2] E. Markova, T. Bajto ´ s, P. Sokol and T. M ˇ eze ´ sov ˇ a, “Classification of ´ malicious emails”, 2019 IEEE 15th International Scientific Conference on Informatics, Poprad, Slovakia, 2019, pp. 000279-000284, doi: 10.1109/Informatics47936.2019.9119329.

[3] M. S. Swetha and G. Sarraf, “Spam Email and Malware Elimination employing various Classification Techniques”, 2019 4th International Conference on Recent Trends on Electronics, Information, Communication Technology (RTEICT), Bangalore, India, 2019, pp. 140-145, doi: 10.1109/RTEICT46194.2019.9016964.

[4] S. Nandhini and D. J. Marseline.K.S, “Performance Evaluation of Machine Learning Algorithms for Email Spam Detection”, 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), Vellore, India, 2020, pp. 1-4, doi: 10.1109/icETITE47903.2020.312.

[5] K. Kandasamy and P. Koroth, “An integrated approach to spam classification on Twitter using URL analysis, natural language processing and machine learning techniques”, 2014 IEEE Students’ Conference on Electrical, Electronics and Computer Science, Bhopal, 2014, pp. 1-5, doi: 10.1109/SCEECS.2014.6804508.

[6] S. B. Rathod and T. M. Pattewar, “A comparative performance evaluation of content based spam and malicious URL detection in E-mail”, 2015 IEEE International Conference on Computer Graphics, Vision and Information Security (CGVIS), Bhubaneswar, 2015, pp. 49-54, doi: 10.1109/CGVIS.2015.7449891.

[7] Wei Hu, Jinglong Du, and Yongkang Xing, “Spam Filtering by Semantics-based Text Classification”, 8th International Conference on Advanced Computational Intelligence Chiang Mai, Thailand; February 14-16, 2016

[8] Crawford, M., Khoshgoftaar, T.M., Prusa, J.D. et al. ,“Survey of review spam detection using machine learning techniques”, Journal of Big Data 2, 23 (2015). https://doi.org/10.1186/s40537-015-0029-9

[9] Vlad Sandulescu, Martin Ester “Detecting Singleton Review Spammers Using Semantic Similarity”, WWW ’15 Companion Proceedings of the 24th International Conference on World Wide Web, 2015, p.971-976 10.1145/2740908.2742570

[10] Cheng Hua Li, Jimmy Xiangji Huang “Spam filtering using semantic similarity approach and adaptive BPNN”, Neurocomputing Journal, Elsevier, https://doi.org/10.1016/j.neucom.2011.09.036

[11] Krishnan Kannoorpatti, Asif Karim , Sami Azam, BharanidharanSanmugam, ”on A Comprehensive Survey for Intelligent Spam Email Detection,” IEEE Journal of Computational Intelligence, 2015.

[12] Zainal K, Sulaiman NF, Jali MZ, “An Analysis of Various Algorithms For Text Spam Classification and Clustering Using RapidMiner and Weka”, (IJCSIS) International Journal of Computer Science and Information Security, Vol. 13, No. 3, March 2015

[13] B. Yu, Z. Xu, “A comparative study for content-based dynamic spam classification”, Knowl. Based Syst. , China, 2008, doi:10.1016/j.knosys.2008.01.001 [14] C.H.Wu, “Behavior based spam detection using a hybrid method of rule based techniques and neural networks”, Expert Systems with Applications, Kaohsiung, Taiwan, 2009, doi:10.1016/j.eswa.2008.03.002

[15] S.M.Lee, D.S.Kim, J.H.Kim, J.S.Park, “Spam Detection Using Feature Selection and Parameter Optimization”, 2010 International Conference on Complex, Intelligent and Software Intensive Systems, DOI 10.1109/CISIS.2010.116

[16] E.G.Dada, J.S.Bassi, H.Chiroma, S.M.Abdulhamid, A.O.Adetunmbi, O.E.Ajibuwa, “ Machine learning for email spam filtering: review, approaches and open research problems”, Heliyon (2019) DOI:doi.org/10.1016/j.heliyon.2019.e01802

[17] 455 Spam Trigger Words to Avoid in 2019, accessed 3 November 2020, https://prospect.io/blog/455-email-spam-trigger-words-avoid-2018/

[18] PhishTank, accessed 3 November 2020, https://www.phishtank.com/

[19] Word2vec skip gram and cbow, accessed 3 November 2020, https://towardsdatascience.com/nlp-101-word2vec-skip-gram-andcbow-93512ee24314 [20] Asif Karim, Sami Azam, BharanidharanShanmugam, Krishnan Kannoorpatti, and MamounAlazab - "A Comprehensive Survey for Intelligent Spam Email Detection", College of Engineering, IT and Environment, Charles Darwin University, Casuarina, NT 0810, Australia.

[21] Two Simple Adaptations of Word2Vec for Syntax Problems - Scientific Figure on ResearchGate, accessed 3 November 2020, https://www.researchgate.net/figure/Illustration-of-the-Skip-gram-andContinuous-Bag-of-Word-CBOW-modelsfig1281812760

[22] Enron Spam data set accessed on 3 November 2020, http://nlp.cs.aueb.gr/softwareanddatasets/Enron-Spam/index.html

[23] Kaggle data set accessed on 3 November 2020, https://www.kaggle.com/uciml/sms-spam-collection-dataset

[24] Y. Lin and J. Wang, "Research on text classification based on SVMKNN," 2014 IEEE 5th International Conference on Software Engineering and Service Science, Beijing, 2014, pp. 842-844, doi: 10.1109/ICSESS.2014.6933697.