IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 4, April 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.488

9940 572 462 | 6381 907 438 | ijircce@gmail.com | www.ijircce.com

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 4, April 2021 ||

| DOI: 10.15680/IJIRCCE.2021. 0904055 |

An Collaborative and Early Detection of Email Spam Using Multitask Learning

Jospin Jeya J[1] , Rajan L[2], Ram Kumar T[3]

Associate Professor, Department of Computer Science, Jeppiaar engineering College, Chennai, India[1]

UG Student, Department of Computer Science, Jeppiaar engineering College, Chennai, India[2,3]

ABSTRACT: The problem of Email spam has grown significantly over the past few years. It is not just a nuisance for users but also it is damaging for those who fall for scams and other attacks. This is due to the complexity intensification of Email spamming techniques which are advancing from traditional spamming (direct spamming) techniques to a more scalable, elusive and indirect approach of botnets for distributing Email spam messages. This paper proposes a hybrid solution of spam email classifier using context based email classification model as main algorithm complimented by information gain calculation to increase spam classification accuracy. Proposed solution consists of three stages email pre-processing,feature extraction and email classification.

KEYWORDS: Dataset Pre-processing, Analysis Of Emails, Feature Selection, Spam Prediction.

I. INTRODUCTION

E-MAIL becomes a necessary means of communication because of its convenience and high efficiency. But the number of spam is increasing since it can make big profits with a small spending by spreading advertisement or other disgust news to mail users. Some lawbreakers even send computer virus with an e-mail which results in a huge threat of computer. Spam, usually considered as unsolicited bulk e-mail or unsolicited commercial e-mail, has brought many troubles to our normal communication by e-mail. Ferris Research Group indicated that the number of spam was so large that a majority of network bandwidth and mailbox server's storage are unable to be used in other important applications. The huge amount of spam also brought much interference to users and had very severe influences for people to work effectively. Moreover, the spam always had threats once it carrying malicious codes secretly which would affected the safety of computer and personal information. It can be seen from the Symantec Internet Security Threat Report 2015 that there are nearly 60% of e-mails are spam in 2014 and the report of Cyren Internet Threats Trend revealed a more serious statistical result with the spam rate more than 68% in the third quarter of 2014. In a word, spam detection is still a severe challenge.

II.LITERATURESURVEY

[1]"Ghada Al-Rawashdeh" and "RabieiMamat" and "Noor Hafhizah Binti Abd Rahim**"," Hybrid Water Cycle Optimization Algorithm With Simulated Annealing for Spam E-mail Detection"**, IEEE Journal Article (IEEE Volume 7) 2019.
The aim of this research is to improve the accuracy of feature selection by applying hybrid Water Cycle and Simulated Annealing to optimize results and to evaluate the proposed Spam Detection. The methodology used in this study which consists of groundwork, induction, improvement, evaluation and comparison quality. The cross-validation was used for training and validation dataset and seven datasets were employed in testing the spam classification proposed. The results demonstrate that the meta-heuristic namely water cycle feature selection (WCFS) was employed and three ways of hybridization with Simulated Annealing as a feature selection employed.

[2] **"**EnaitzEzpeleta" and "UrkoZurutuza" and "José María Gómez Hidalgo**" –** "A Study of the Personalization of Spam Content using Facebook public information"**,**IEEELogic Journal of the IGPL Year: 2017 | Volume: 25, Issue: 1.

In this task we considered two options: thefirstone,obtainingtheemailaddresses, where they get e-mail addresses using various combinations of public information from OSN users. The second, usingpubliclyavailableapplicationsthatautomaticallyharvestemail addresses from simple search queries over known

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 4, April 2021 ||

| DOI: 10.15680/IJIRCCE.2021. 0904055 |

search engines. The one used by the authors, generates a query for the search engine using a given keyword, and extracts email address patterns from the search result.

[3]" Ghulam Mujtaba, LiyanaShuib","Ram Gopal Raj"," Nahdia Majeed" and "Mohammed Ali Al-Garadi" – **"Email Classification Research Trends: Review and Open Issues",**IEEE Access Journal Article,Year: 2017 | Volume: 5 |.

This study comprehensively reviews articles on email classification published in 2006–2016 by exploiting the methodological decision analysis in five aspects, namely, email classification application areas,datasetsusedineachapplicationarea,featurespaceutilized in each application area, email classification techniques, and useof performance measure . To achieve the objective of the study, a comprehensive review and analysis is conducted to explore the various areas where email classification was applied.

[4]" Wazir Zada Khan","Muhammad Khurram Khan","Fahad Bin Muhaya","Muhammad Y Aalsalem" and "HanChieh Chao" **– "A Comprehensive Study of Email Spam Botnet Detection"** - IEEE Communications Surveys &Tutorials,Year: 2015 | Volume: 17, Issue: 4 | Journal Article | .

In this paper, they first discuss the sources and architectures used by the spamming botnets for sending massive amount of email spam. Then they present detailed chronicles of spamming botnets which systematically describes the timeline of events and notable occurrences in the advancement of these spamming botnets. This paper also aims to represent a comprehensive analysis of particular email spamming botnet detection techniques proposed in the literature.They attempt to categorize them according to both their nature of defense and method of detection, also revealing and comparing their advantages and disadvantages extensively. They also present a qualitative analysis of these techniques.

[5] "Haiying Shen" and "Ze Li" – **"Leveraging Social Networks f**or Effective Spam Filtering**"**.IEEE Transactions on Computers,Year: 2014 | Volume: 63, Issue: 11 | Journal Article |.

In order to develop an accurate and user-friendly spam filter, they propose a SOcial network Aided Personalized and effective spam filter (SOAP) in this paper. In SOAP, each node connects to its social friends; i.e., nodes form a distributed overlay by directly using social network links as overlay links. Each node uses SOAP to collect information and check spam autonomously in a distributed manner. Unlike previous spam filters that focus on parsing keywords (e.g., Bayesian filters) or building blacklists, SOAP exploits the social relationships among email correspondents and their (dis)interests to detect spam adaptively and automatically.

III.PROPOSED SYSTEM

The proposed system first discuss the sources and architectures used by the spamming botnets for sending massive amount of email spam. Then we present detailed chronicles of spamming botnets which systematically describes the timeline of events and notable occurrences in the advancement of these spamming botnets. This paper also aims to represent a comprehensive analysis of particular Email spamming botnet detection techniques proposed in the literature. We attempt to categorize them according to both their nature of defense and method of detection, also revealing and comparing their advantages and disadvantagesextensively.Wealsopresentaqualitativeanalysisofthesetechniques.Finally we summarize the future trends and challenges in detecting email spammingbotnets.
Graph-mining approaches to email classification take advantage of semantic features and structureinemailsbyconvertingemailsintographsandmatchingtemplategraphswithgraphs made from each emails. Typical graph mining algorithm converts emails into graphs. Substructures of graphs are then extracted from graphs. Parameters prune substructures. Representative substructures remain. Substructures are ranked just so that in case an email graph matches more than two representative substructures, emails go into a folder which the matched representative with higherrank.

In the proposed system, Two important techniques of neural network are Dropout and Activation. hyperparameter tuning can also be done based on these techniques, but are not used here. Dropout technique is used to improve the generalization error of large neural networks. In this method the noise zeros, or drops out a fixed fraction of the activation of the neurons in a given layer. Rectified linear unit (ReLU) uses the activation function max(0; x). ReLUs are incorporate into a standard feed-forward neural net, to maintain the probabilistic model with the max(0; x). GloVe (Global vectors) used here is one of the approach where each word is mapped to 100-dimension vector. These vectors can be used to learn the

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 4, April 2021 ||

| DOI: 10.15680/IJIRCCE.2021. 0904055 |

semanticsofwordslikeManisWomanasKingistoQueen.OrMan+Female=Woman.This embeddingplaysanimportantroleinmanyapplications.Itiskindofatransferlearningwhere word embedding are learnt from large corpus of data and then can be used on smaller datasets. The vectors are generated by an unsupervised learning algorithm (PCA). Training isperformedonaggregatedglobalword-wordco-occurrencestatisticsfromacorpus,andthe resulting representations showcase linear substructures of the word vector space. The main intuition underlying the model is the simple observation that the ratio of word-word co- occurrence probabilities have the potential for encoding some form ofmeaning.

ADVANTAGES OF PROPOSED SYSTEM

- ✓ Fake emails maybe recognized with high accuracy andprecision.
- ✓ Higher efficiency andpreciseness.
- ✓ Much accurate and higher than any of the existingtechniques
- ✓ It runs efficiently on largedatasets.
- ✓ It can handle thousands of input variables without variabledeletion.

IV.RESULTS AND DISCUSSIONS

PREPROCESSING:

The first and foremost thing in machine learning is the dataset collection and preprocessing. Here for our Project we have collected the dataset form Kaggle dataset.Once the dataset is collected we preprocessed them ie(cleaning the unwanted things).

ANALYSIS OF EMAIL:

Our dataset consist of 2 columns .First column represent whether the message is spam or ham. Second column represent the context of the message. So using these datas we are going to do analysis. First we are analyzing the total number of messages and frequency of the total message and finding total number of Spam and ham message and finding the frequency of spam and ham etc…

For our model to find whether the message is spam or ham we don't want the full content of the message ,we just need the main keywords to find the prediction ie(Spam or ham) so lot of unwanted and grammatical words are present in the context of the message so we need to remove those things using NLTK(Natural Language Toolkit). After cleaning up the context using NLTK we just have main keywords,so we are going to do analysis on those main keywords ie(What are the top words that are consider as spam and ham).

We use plots and graphs to represent the analysis visually and in addition we use wordcloud to visually represent the main keywords.

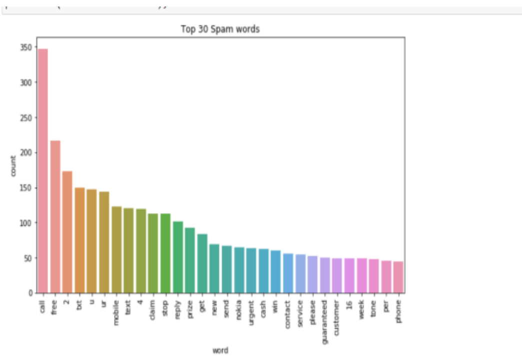

Fig.1.Shows Top 30 Spam words

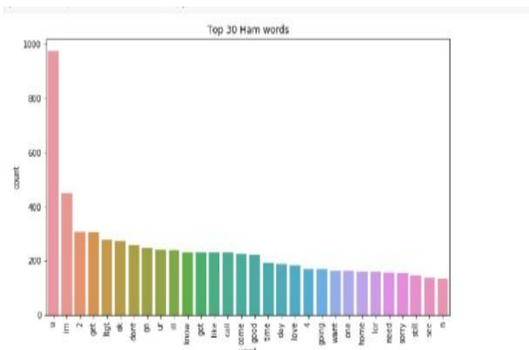

Fig.2.Shows Top 30 Ham Words

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 4, April 2021 ||

| DOI: 10.15680/IJIRCCE.2021. 0904055 |

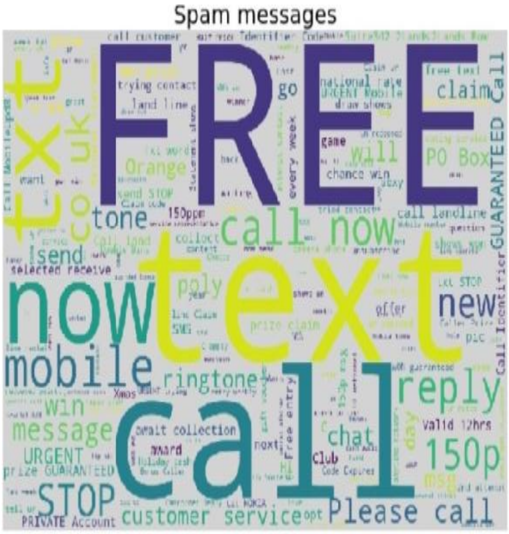

Fig.3.Spam Messages

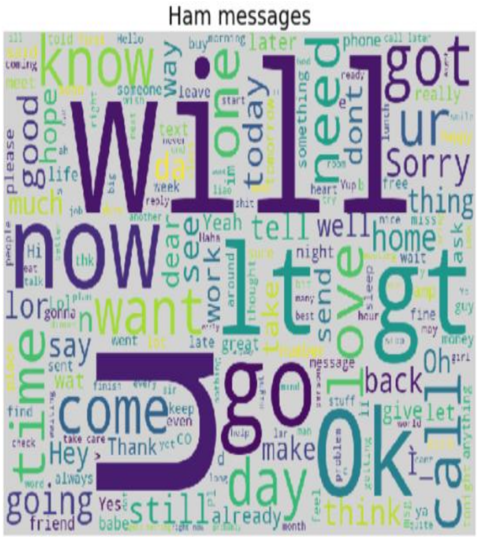

Fig.4.Ham Messages

BUILDING MODEL:

For every machine learning projects we first need a build a model to train our datasets using some algorithms.so here we build a model using DEEP LEARNING library KERAS and all the necessary things that are needed to build a model are also imported.So we are going to train our model now,but before training we need to notice one thing (ie) our context of the message contain main keywords only and we are going to train our model by splitting the sentence into separate words so for this need we used a tokenizer class from keras.The purpose of this class is split the whole sentence into separate words. Once splitting is done we are ready to train our dataset to the model we built using the DEEP NEURAL NETWORK (DNN) algorithm. The purpose of using DNN is already discussed before.

PREDICTION OUTPUT :

DNN consist of 3 layers ie (input,hidden,output) We are building these layers using keras and we are able to find that our model have learned manythings from these dataset and lost something, so we plot how much that the model has learned and lost in the graph.Once the first learning is finished we have dropoutsomehiddenlayer to increase the efficiency of the model so after dropping some hidden layers we again train our model and we have seen that our model have learned more efficiently and got good accuracy when compared to the previous training.

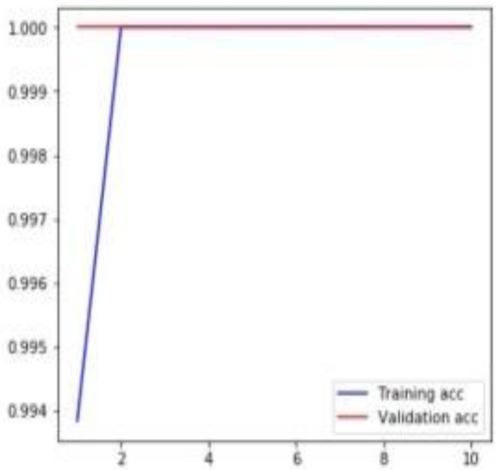

Fig.5. Shows Accuracy Level

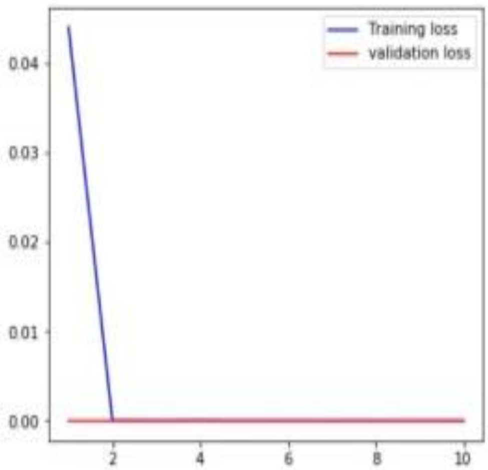

Fig.6.Shows how much algorithm learnt

Here in our project we just built a model and trained using DNN algorithm and got more accuracy of learning when compared to the other models. So the ultimate goal of our project is how much that the model is learned in the training phase.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 4, April 2021 ||

| DOI: 10.15680/IJIRCCE.2021. 0904055 |

We got nearly 99% of accuracy in our model so in future we need to implement this model as a real time application or as a software to find the spam messages.

V.CONCLUSION

Based on this finding, it can be concluded that the content classification performance will be improved with enhancements as a feature selection. The second finding is that the use ofthe interleaved hybridization generated better optimal features for the classifier than using all the features From this observation, it can be stated that content classification can be better performed using all the optimal features generated by the interleavedhybridization.

REFERENCES

[1] S. Wasi, S. Jami and Z. Shaikh, "Context-based email classification model", Expert Systems, vol. 33, no. 2, pp. 129-144, 2015.

[2]Alsmadi and I. Alhami, "Clustering and classification of email contents", Journal of King Saud University - Computer and Information Sciences, vol. 27, no. 1, pp. 46-57, 2015.

[3] S. Sayed, "Three-Phase Tournament-Based Method for Better Email Classification", International Journal of Artificial Intelligence & Applications, vol. 3, no. 6, pp. 49-56, 2012.

[4] D. Patil and Y. Dongre, “A Clustering Technique for Email Content Mining,” Int. J. Comput. Sci. Inf. Technol., vol. 7, no. 3, pp. 73–79, 2015.

[5] H. He, A. Tiwari, J. Mehnen, T. Watson, C. Maple, Y. Jin, and B. Gabrys, “Incremental information gain analysis of input attribute impact on rbf-kernel svm spam detection,” in Evolutionary Computation (CEC), 2016 IEEE Congress on, pp. 1022–1029, IEEE, 2016.

[6] A.-Z. Ala’M, H. Paris, et al., “Spam profile detection in social networks based on public features,” in Information and Communication Systems (ICICS), 2017 8th International Conference on, pp. 130–135, IEEE, 2017.

[7] A.-Z. AlaM, H. Faris, M. A. Hassonah, et al., “Evolving support vector machines using whale optimization algorithm for spam profiles detection on online social networks in different lingual contexts,” Knowledge-Based Systems, vol. 153, pp. 91– 104, 2018.

[8]Barushka and P. Hájek, “Spam filtering using regularized neural networks with rectified linear units,” in Proc. AI*IA Adv. Artif. Intell. 15th Int. Conf. Italian Assoc. Artif. Intell., Genoa, Italy, Nov./Dec. 2016, pp. 65–75.

[9] H. Faris, I. Aljarah, and B. Al-Shboul, “A hybrid approach based on particle swarm optimization and random forests for e-mail spam filtering,” in International Conference on Computational Collective Intelligence, pp. 498–508, Springer, 2016.

[10] F. Wang, T. Xu, T. Tang, M. Zhou, and H. Wang, ‘‘Bilevel feature extraction-based text mining for fault diagnosis of railway systems,’’ IEEE Trans. Intell. Transp. Syst., vol. 18, no. 1, pp. 49–58, Jan. 2017.

[11] S. Gupta, A. Khattar, A. Gogia, P. Kumaraguru, and T. Chakraborty, ‘‘Collective classification of spam campaigners on Twitter: A hierarchical meta-path based approach,’’ in Proc. World Wide Web Conf., Apr. 2018, pp. 529–538.

[12] M. Chakraborty, S. Pal, R. Pramanik, and C. R. Chowdary, “Recent developments in social spam detection and combating techniques: A survey,” Inf. Process. Manag., vol. 52, no. 6, pp. 1053–1073, 2016.

[13] H. Fu, X. Xie, and Y. Rui, “Leveraging careful microblog users for spammer detection,” in Proc. 24th Int. Conf. World Wide Web, 2015, pp. 419–429.

[14] S. K. Trivedi and P. K. Panigrahi, ‘‘Spam classification: A comparative analysis of different boosted decision tree approaches,’’ J. Syst. Inf. Technol., vol. 20, no. 3, pp. 105–298, Aug. 2018.

[15] Y. Ren and Y. Zhang, “Deceptive opinion spam detection using neural network,” in Proc. COLING 26th Int. Conf. Comput. Linguist. Conf. Tech. Papers, Dec. 2016, Osaka, Japan, 2016, pp. 140–150. [Online].

[16] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” CoRR, vol. abs/1409.0473, Sep. 2014. [Online]. Available: http://arxiv.org/abs/1409.0473

[17] S. Zhao, Z. Xu, L. Liu, and M. Guo, “Towards accurate deceptive opinion spam detection based on word order-preserving CNN,” CoRR, vol. abs/1711.09181, 2017. [Online]. Available: http://arxiv.org/abs/1711.09181

[18] G. Jain, M. Sharma, and B. Agarwal, ‘‘Spam detection on social media using semantic convolutional neural network,’’ Int. J. Knowl. Discovery Bioinf., vol. 8, no. 1, pp. 12–26, Jan. 2018.

[19] S. Saha, S. DasGupta, and S. K. Das, ‘‘Spam mail detection using data mining: A comparative analysis,’’ in Smart Intelligent Computing and Applications. Singapore: Springer, 2019, pp. 571–580.

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING