AN EVALUATION OF TIME COMPLEXITIES OF BAYESIAN BASED AND HYBRIDIZED WORD STEMMING TECHNIQUE FOR FILTERING ADVANCED FEE FRAUD EMAILS

Okunade Oluwasogo Adekunle

Department of Computer Science, Faculty of Sciences, National Open University of Nigeria,

Abuja, Nigeria.

ABSTRACT

Challenges day in out occupied electronic mail classification processes, as a result of dynamism in spam attacks. Different techniques has being implemented to combat this attacks to the extent of using combined techniques, called hybridized spam filter techniques. However, time execution variance of hybridized filtering techniques against an individual technique need to be carefully examine in spam filtering processes. To avert the further manipulation and reoccurrence of scammers and implementation of their enterprises, to prevent future reoccurrence. Time execution of content based spam filter is being described using the Bayesian statistical algorithm versus Bayesian statistical algorithm incorporated with a word stemmer algorithm. The execution time interval for the two algorithms implementing the two techniques were evaluated by subjecting the filters to manipulated and non-manipulated spam and ham mails. Result of the tested time variance of the two algorithm signify that ordinary Bayesian statistical technique (single filter) took one quarter (¼) of the entire time used by Bayesian statistical integrated with word stemmer classifier algorithm (hybridized spam filter techniques). The implication is that when a word stemmer is incorporated with other Bayesian statistical classifiers, email classification is optimized and improved in performance, but with significant increased in execution time.

Keywords: Time execution, Classification, Spam, Ham, Suspicious terms, Word stemmer.

INTRODUCTION

Electronic mail also known as Email is the cheap and fast means of communication, among the individual and cooperate organizations. It is efficient, simple and accessible means of communication at the availability of internet (Ahmed & Hani, 2017). Email availability, simplicity and cheapness are prone to a lot of threats among which is spam (Cormack, Smucker & Clarke, 2011 in Ahmed & Hani, 2017 and Zahra & Seyyed, 2017). Email has a peculiar challenges that turns its expediency burdensome among the users, called spam. Spam is unsolicited massive number of commercial bulk and harmful e-mail sent to multiple recipients, that are irrelevant to the specified recipients. Spam also called Jung mail is one of the major internet challenges, that contributed to today's internet step back in various ways, such as financial lost to individuals and cooperate organizations and annoyance to individual users. Managing these emails becomes a significant challenge to individuals and cooperate organizations, since most of the traffic comprises of unsolicited bulk email according to (Karthika & Visalakshi, 2015). Spam filtering technique methods are categories into two: methods that avoid spam distribution at the origin and methods that avoid spam at destination point (Saadat, 2011).

Various attempt have being taken to prevent, reduce and even eliminate spam existence, but as they are being prevented in one way, they are coming up in several ways. Several techniques have being applied to address this challenges among which are Naïve Bayes Classifier, K-Nearest Neighbour classifier, Support Vector Machine, Decision Tree, Fuzzy logic and so on (Upasana, 2010). This methods had been applied and re-applied in several ways due to their manipulations by scammers in order to implement their enterprises. Among several re-use is the combination of Word Stemmer with Bayesian classifier to form a stronger hybridized classifier. This is a type of classifier that will firstly exposed the real mail content to its actual original, by eliminating the manipulations within the keywords used to thwart the filters by the scammers. And latter apply the proper Bayesian based classifier for proper and accurate filtering and mail classification. However, hybridized classifier algorithm complexity need to explore for advancement. According to Lasmedi and Retantyo (2017) complexity of the algorithm is divided into complexity of time and space. This paper focused on time complexity of an improved algorithm (hybridized Bayesian based and word stemming filtering techniques algorithm (Spam filter with additional function) against the existing filtering techniques (ordinary Bayesian based filtering techniques algorithm). Measuring time complexity of an algorithm is computing the number of stages required to run the algorithm as a function of a number of data n (size of the input), against the execution time of the tested data. Execution time needs to explore to test for variances in execution time of Bayesian based filtering technique algorithm against the hybridized filtering technique algorithm. This prevent execution time variance manipulation by scammers that may further used for advance spam attack.

The rest of this paper is structured as follows: Literature review, discusses and analyse other researchers write up on ham and spam mails, combined and individual filtering techniques, and current research on content-based spam filter execution time variance. While methodology and material discusses the method of approach applied in measuring the execution time variance of ordinary Bayesian based filtering technique algorithm against the hybridized word stemming plus Bayesian based filtering techniques algorithm. Result show the outcome of the experiment and discussion displayed and analyze the outcome of the research and paper then concludes in the conclusion.

LITERATURE REVIEW

Hybridization, combined and process two categories of spam filtering techniques, with the aims of integrating their both advantages over their disadvantages, to come up with a new hybridized idea with better and improved performance. Any of the two combined filtering techniques may apply to a particular segment of the hybridized technique operational process to perform a certain designated function. And the other technique to perform another designated function in another segment that may later combined the result, each of the two designated functions to get an expected end result.

Otherwise combined and use the two techniques side by side to generate the expected end result (Abdullah, Abdul, Azuraliza & Mohd, 2015). According to Subhana and Pramod (2016) each algorithm is only suitable for filtering specific spam. It is not reliable and inefficiency to use a single algorithm to separate spam out rightly, in this case hybridized filtering techniques is highly appropriate and recommended for effective filtering collaboration. The author proposed hybridization of two different algorithms, Bayesian classifier and back propagation neural network, and tested variants of the hybridized algorithm on numbers of different data sets against the individual traditional filtering algorithm. This show that combined algorithm performed and achieved very good and accurate results, with poor time consumption. The time complexity of the proposed data model is on the high side, despite of it accurate performance.

Mail classification Techniques

Mail classification techniques are techniques presented as a means to identify, differentiate and separate between the legitimate (Ham) and spam mails. It help to avert the scammers from successful achievement of their enterprises. Email classifications were being used to present spam to the recipient as a spam and ham as ham within the set of received buck mails. It is technical ideologies that prevent one from individual physical rigorous of identification and separation of ham from spam within the set of buck mail received. There are various types of mail classification techniques, among are:

Support Vector Machine

Support Vector Machine (SVM) is one of the most commonly used filtering algorithms for spam detection (Subhana, Nadir, Othman & Waheeb, 2014). Support Vector Machine (SVM) is a statistical learning method for pattern recognition. It applied Kernel function method that does not increase the computational complexity. According to Priyanka, Rajesh and Sanyam (2010) in their experiment, stated that spam: ham is ratio 1:3 given as their appropriate result that the classification is appropriate for more legitimate mails compare to that of spam mail. Then concluded that SVM is a good classifier compared to Decision Tree classifier, that have large memory requirement, because of its poor data format. However, Subhana, Nadir, Othman & Waheeb (2014) further stated that it was shown in many cases that it takes a long processing time and at the same time provides a less accurate rate for classification due to the content volume (size).

Decision Tree mail classification

Decision Tree is a common data mining classification. The principle idea of a decision tree is to split data recursively into subsets so that each subset contains more or less homogeneous states of targeted variable. All input attributes are evaluated for their impact on the predictable attribute, at each split in the tree, and then for a decision tree having completed the recursive process (Akhilesh, & Rahul, 2015). Priyatharsini and Chandrasekar (2017) implement different various decision tree classifiers for evaluation, and stated that decision tree filters are easy to implement and understand. It provides an overall satisfactory performance as far as spam mail detection is concerned. However, the algorithms takes more time to execute than other algorithms, despite its advantages, it short coming relied on time complexity.

Naive Bayesian Based Filter

Naive Bayesian based filter is the application of Bayesian statistical formula for mail classification with assumption of strong independence (Tang, Kay & He, 2016 in Priti & Uma, 2018). Nearly all the statistic based spam filters uses Bayesian probability calculation for classification of mail, according to Heckerman & Wellman (1995) in Kang & Zhenyu (2006). Similarly, Bayesian probability combination has been widely used successfully in various message classifications. Bayesian filter should be trained to work effectively, since every word has certain probability of occurring in either spam or ham email, in the given database, that further used to determine the total words probabilities. if exceeds a certain limit, the entire mail is classified as spam otherwise as ham (Awad & ELseuofi, 2011).

K-Nearest Neighbour classifier

The K-Nearest Neighbour (*K*-NN) classifier is a classifier that search for the most similar documents (neighbours), if a enough large proportion of the document

have been assigned to a certain classification category, similar thing may likewise apply to the new document. If not it categorize otherwise (Awad & ELseuofi, 2011).

Fuzzy logic

The concept of Fuzzy Logic was first proposed in (Zadeh, 1965). It is a flexible approach of mail classification, that give room for partial membership in a particular set of given mail. It stated that, it does not require precise, numerical information input, to get the expected output. If feedback controllers could be programmed to accept noisy, imprecise input, they would be much more effective in classification (Yeganeh, Bin and Babu, 2012).

METHODOLOGY

Bayesian Spam Filtering Technique

It employs the principle of Mathematical Probability formula to classified email messages to be ham (legitimate) or spam (unwanted). It identifies the suspicious terms within the email content, and pick from the database already assigned numerical values for the identified suspicious term. To calculate the email chances of becoming a spam or ham mail. The final calculated result is compared against the particular set threshold, if greater than the threshold value (the entire mail concluded spam and classified as spam) otherwise lesser (the entire mail concluded as spam and classified as ham). The threshold value (that could be any of 0.3, 0.4 or 0.5) figure 1.

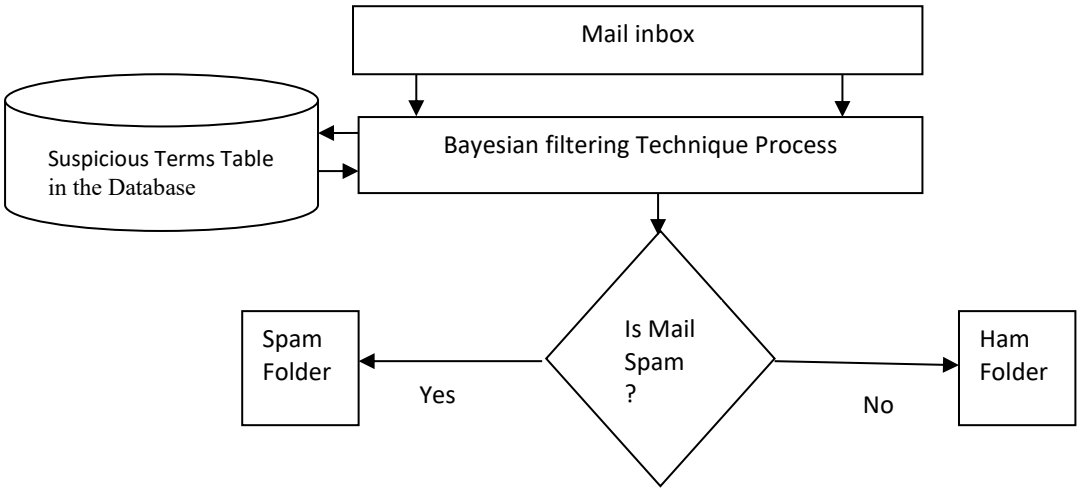

Figure 1: Pure Bayesian Spam filtering Technique Experimental setup

Hybridized Bayesian Filtering with Word Stemming Technique

Word Stemming removed all unwanted prefixes, affixes and suffixes within and around the suspicious terms to generate the suspicious terms actual root, placed by scammers to thwart the filters, in order to successful implement their enterprises. Having done this using the word stemmer, Bayesian filtering technique is then applied to actually filter the real mail content (Okunade, n. d).

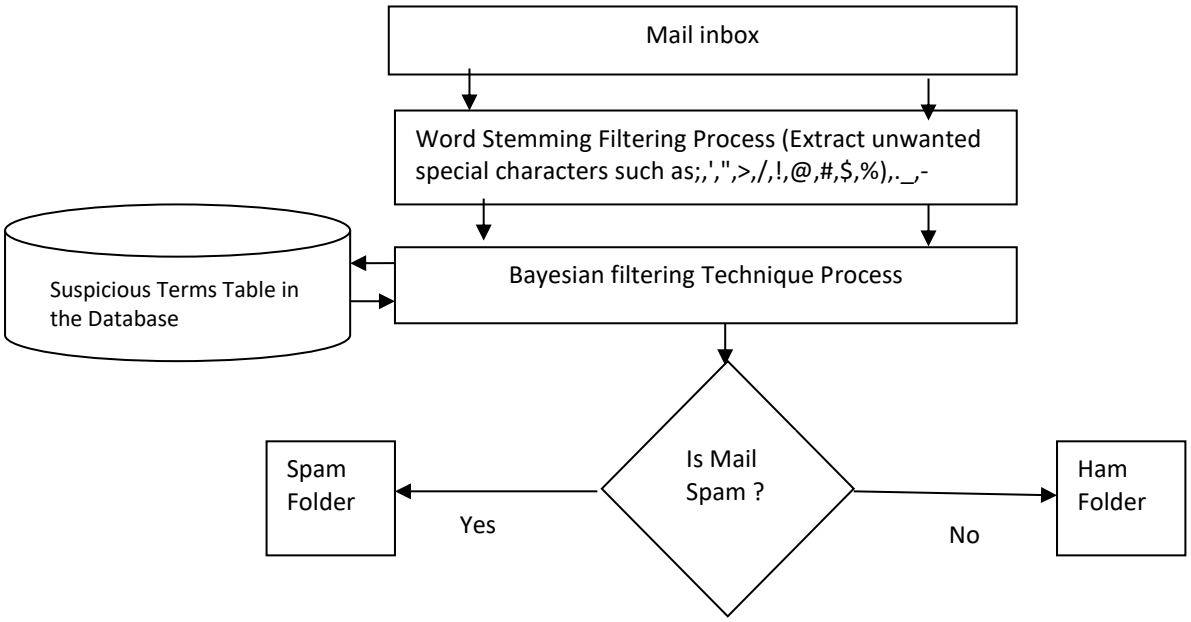

Figure.2: The Hybridized Algorithm (Bayesian Spam filter Technique Incorporated with Word Stemmer Technique) Experimental Setup data flow.

The experimental setup show in figure 1 and 2; Figure 1 is the execution process of pure Bayesian Statistical filtering technique process while the figure 2, is the execution process of Bayesian Statistical filtering technique incorporated with the Word Stemmer technique processes

RESULT

Chart 1 show the execution time variance of the experimental result of figure1 and 2 conducted. The x-axis signify mail size (volume) measured by words count make up the mail content. The y-axis measure time, it signify time taken an algorithm to complete circle of a particular mail execution, measure per seconds. In chart 1 x-axis, two bars contained same values (same numerical value of word count), first of the same two values, in each set of bars (1^{st}: 173, 199, ..., 448) in blue colour is for hybridized Bayesian filtering combined with Word Stemming techniques examined result. While the second same values in red colour is for ordinary Bayesian filtering technique examined result (2^{nd}: 173, 199, …,448). The first mail value (1^{st} 173) is the spam mail executed without manipulating the suspicious terms while the 2^{nd} mail with 173 numbers of words is the spam mail with manipulated suspicious terms, the third spam mail (1^{st} 199) is the spam mail without manipulated suspicious terms while the forth mail (2^{nd} 199) is the spam mail with manipulated suspicious terms and so on, up to the second to the last mail (1^{st} 448) is spam mail without manipulate suspicious terms and the last mail (2^{nd} 448) is spam mail with manipulated suspicious terms.

The execution time of each mail on y-axis is the time interval taken the algorithm to complete the execution. Time measured in blue colour (Bayesian algorithm with word Stemmer) taken larger time for the completion of the execution, against the time measured in red colour (ordinary Bayesian algorithm without word Stemmer) that take lesser time for the completion of the execution. Also the chart indicate that the experiment take lesser time in executing mails with suspicious terms manipulated using any of the two algorithm rather than executing the mail with suspicious terms not manipulated except the mail contained 404 words where the execution of the manipulated offensive words is a little bit increased by 0.03 seconds)

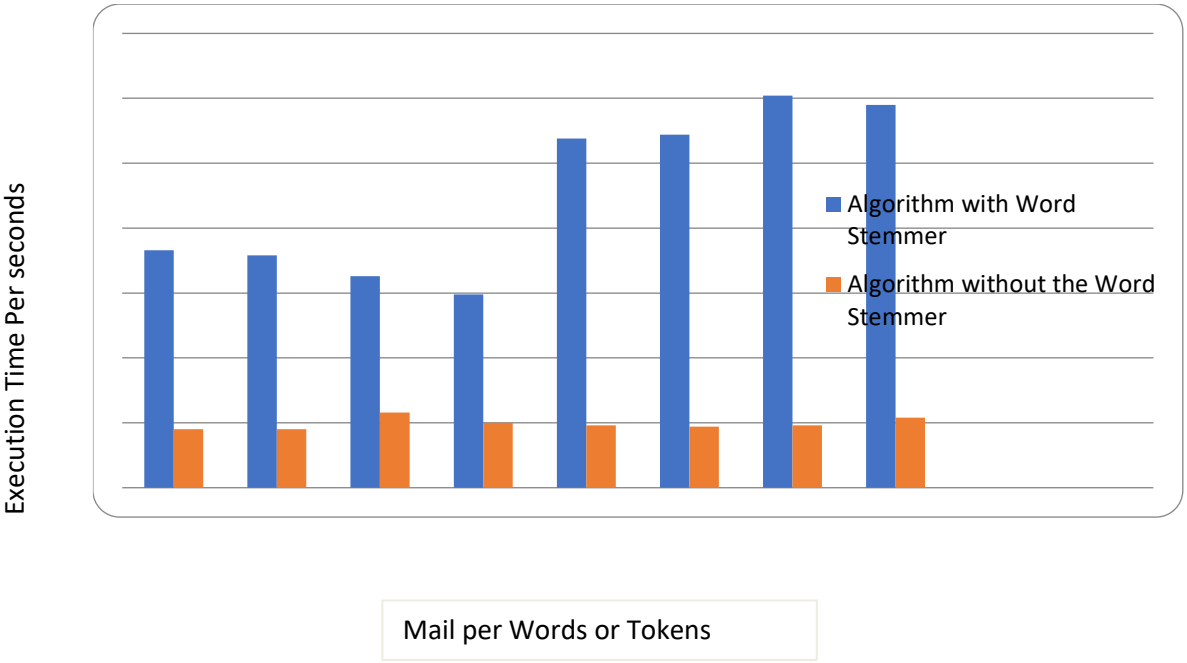

Chart 1: The result of The Execution Time comparison of Bayesian Statistical Spam Filter Against The Bayesian Statistical Incorporated with Word Stemming Spam Filter.

DISCUSSION

Result of execution time comparism of the two algorithm experiments show that the execution time of the Bayesian incorporated with Word Stemmer was far larger/higher compared to that of ordinary Bayesian mail classification. Result of the tested time variance of the two algorithm signify that ordinary Bayesian statistical technique took one quarter (¼) of the entire time used by hybridized Bayesian statistical integrated with word stemmer classifier algorithm. Also, spam with suspicious terms manipulated takes lesser time in execution compared with those without manipulated suspicious terms. Similarly, the work of Subhana and Pramod (2016) sited in the literature review, stated that, hybridization of Bayesian classifier and back propagation neural network show that combined algorithm performed accurately better than single traditional filtering technique. But with poor time consumption, high time complexity compared to single traditional filtering technique.

CONCLUSION

Experiment show that the execution of mail classifier using the Word Stemmer incorporated with the Bayesian mail filter takes larger time in execution compared to that of ordinary Bayesian mail classifier. However, hybridized filtering techniques performed more accurately better than ordinary single filtering technique, but having higher time complexity compared with ordinary single traditional filtering technique.

REFERENCES

Abdullah, S. G., Abdul, R. H., Azuraliza, A. B. and Mohd, R. Y. (2015). Hybrid Statistical Rule-Based Classifier for Arabic Text Mining. Journal of Theoretical and Applied Information Technology 71(2): 194-204. ISSN: 1992-8645. E-ISSN: 1817-3195. *www.jatit.org*

Ahmed, H. and Hani, M. A. (2017). Feature Weight Optimization Mechanism for Email Spam Detection based on Two-Step Clustering Algorithm and Logistic Regression Method. International Journal of Advanced Computer Science and Applications (IJACSA). 8(10): 420-429 *www.ijacsa.thesai.org*

Akhilesh, K. S. and Rahul, H. (2015). Decision Tree Model for Classification of E-mail Data with Feature Selection. International Journal of Research Studies in Computer Science and Engineering (IJRSCSE). pp15-19. ISSN 2349-4840 (Print) & ISSN 2349-4859 (Online). *www.arcjournals.org*

Awad, W. A. and ELseuofi, S.M. (2011). Machine Learning Methods For Spam E-Mail Classification. International Journal of Computer Science & Information Technology (IJCSIT), 3(1): 173-184. DOI : *10.5121/ijcsit.2011.3112*

Cormack, G. V., Smucker, M. D. and Clarke, C.L. (2011). Efficient and effective spam filtering and re-ranking for large web datasets. Information retrieval. 14(5): 441-465.

Heckerman, D. and Wellman, M. P. (1995). Bayesian networks. In Communications of the ACM, 3: 27-30.

Kang, L. and Zhenyu, Z. (2006). Fast Statistical Spam Filter by Approximate Classifications SIGMetrics/Performance. ACM 1595933204/06/0006.

Karthika, R.D. And Visalakshi, P. (2015). A Hybrid ACO Based Feature Selection Method for Email Spam Classification. WSEAS Transactions on computers. 14. 171-177. E-ISSN: 2224-2872

Lasmedi, A. and Retantyo, W. (2017). Calculation Algorithm Complexity of Porter‟s Algorithm in Information Retrieval. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET). 6(7). pp. 1010- 1012. *www.ijarcet.org*

Okunade, O. A.(unpublished) Improved Electronic Mail Classification Using Hybridized Root Word Extractions

Priti, S. and Uma, B. (2018). Machine Learning based Spam E-Mail Detection. International Journal of Intelligent Engineering and Systems, 11(3): 1-10.
DOI:10.22266/ijies2018.0630.01 *www.inass.org*

Priyanka, C., Rajesh, W. and Sanyam, S. (2010). Spam Filtering using Support Vector Machine. IJCCT for International Conference [ACCTA-2010]. (1): 2, 3, 4. 166-171

Priyatharsini, P. and Chandrasekar, C. (2017). Email Spam Filtering using Classifiers in Data Mining. International Journal of Engineering Science and Computing. IJESC 2(11): 15474-15478. *http://ijesc.org/.*

Saadat, N. (2011). Survey on Spam Filtering Techniques. Communications and Network. 3, 153-160. doi:10.4236/cn.2011.33019 http://www.SciRP.org/journal/cn

Subhana, K., Nadir, O. F., Othman, I. and Waheeb, A. (2014). An Improved Of Spam E-Mail Classification Mechanism Using K-Means Clustering. Journal of Theoretical and Applied Information Technology. 60(3): 568-580. ISSN: 1992-8645 E-ISSN: 1817- 3195. *www.jatit.org*

Subhana, K. and Pramod, S. N. (2016). A Hybrid E-Mail Spam Filtering Technique using Data Mining Approach. International Journal of Latest trends in Engineering and Technology (IJLTET). 6(3):118-195

Tang, B., Kay, S. and He, H. (2016). Toward optimal feature selection in naive Bayes for text categorization. IEEE Transactions on Knowledge and Data Engineering, 28(9): 2508-2521

Upasana, S. C. (2010).A Survey of Text Classification Techniques for E-mail Filtering. Second International Conference on Machine Learning and Computing. IEEE. DOI 10.1109/ICMLC.2010.61. p.32-36

Yeganeh, M. S., Bin, L. and Babu, G. P. (2012). A Model for Fuzzy Logic Based Machine Learning Approach for Spam Filtering. IOSR Journal of Computer Engineering (IOSRJCE). ISSN: 2278-0661 4(5): 07-10. *www.iosrjournals.org.www.iosrjournals.org*

Zadeh, L. A. (1965). Fuzzy sets. Information and Control

Zahra, R. and Seyyed, A. A. (2017). Providing an Improved Feature Extraction Method for Spam Detection Based on Genetic Algorithm in an Immune System. Journal of Knowledge-Based Engineering and Innovation. 3(8): 569-605