Communication in Physical Sciences, 2021, 7(1):40-44

Evaluation of Time Complexities of Bayesian Vs Hybedrized Word Stemming Techniques for Advanced Fee Fraud Emails Filtering

Okunade Oluwasogo Adekunle, Afolorunso Adenrele and Adebayo Adegboyega
Received 18 December 2020/Accepted 20 March 2021/Published Online 29 March 2021

Abstract: *Time execution of content-based spam filter was investigated using the Bayesian statistical algorithm against Bayesian statistical algorithm incorporated with a word stemmer. The execution time intervals for the algorithm's implementation of the two techniques were evaluated by subjecting the filters to manipulated and non-manipulated spam mails. Tests conducted with ham mails (mails without suspicious terms) took the Bayesian statistical method integrated with a word stemmer ¼ of the time used by ordinary Bayesian statistical algorithm. The implication is that when a stemmer is incorporated with other email classifiers, classification is optimized and the performance of the algorithm does not degrade in terms of execution time.*

Key Words: *Time execution, Classification, Spam, Mail, Word stemmer.*

Okunade, Oluwasogo Adekunle*
Department of Computer Science, Faculty of Sciences, National Open University of Nigeria, Cadastral Zone, Nnamdi Azikiwe Expressway, Jabi, Abuja, Nigeria
Email: aokunade@noun.edu.ng
Orcid id: 0000-0002-1625-8749

Afolorunso, Adenrele Abolanle
Department of Computer Science, Faculty of Sciences, National Open University of Nigeria, Cadastral Zone, Nnamdi Azikiwe Expressway, Jabi, Abuja, Nigeria
Email: aafolorunsho@noun.edu.ng
Orcid id: 0000-0002-1194-7799

Adebayo, Adegboyega
Department of Computer Science, Faculty of Sciences, National Open University of Nigeria, Cadastral Zone, Nnamdi Azikiwe Expressway, Jabi, Abuja, Nigeria
Email: aadebayo@noun.edu.ng
Orcid id: 0000-0002-2886-5933

1.0 Introduction

The Internet has become an important and fastest means of communication. It makes use of electronic mail (eMail) for communication, which is is one of the most personal and professional ubiquitous communication methods. Consequently, spam mail tends to dominate and compete with real mail in a manner that seems to be explosive (Sanjay, 2015). Despite expanding roles and relevance of the internet and communication via email, reported challenges associated with email have been confirmed to be significant (Ali and Tunga, 2007). More is known of the assurance of the mail one sends than the safety of unsolicited mails that have been received. Spam or junk mails are unsolicited email messages sent in bulk (multiple recipients) by spamming and may have some fraudulent benefit to the sender if their mission is not detected by the (Tian, 2020). It is currently regarded as one of the major problems on the internet that is yet to be completely neutralized (Garacia *et al*., 2004). ~~s~~pam message volumes have doubled over the past year and now account for about 80% of the total messages on the Internet. (Zhe, *et al., 2007*). Spam is waste of time, storage space and communication bandwidth and can be a source of virus attack on the internet, which may be potent in destroying users' information or revealing identity or data. Emails are used by several users to communicate around the world. Along with the growth of the internet and email, there has been dramatic growth in spam in recent years. Spam can originate from any location across the globe, where internet access is available (Savita and Santoshkumar, 2014).

Most emails circulating on the Internet are unsolicited bulk emails called Spam (Albercht, 2006). According to The United States, Federal Trade Commission in Alireza, Raheleh and Soheil (2012) 66% of spam have false information somewhere in the message and 18% of spam advertise "Adult" material. Several years ago most of the spam could be reliably dealt with

by blocking the address of such e-mails or filtering out messages with certain subject lines (Fight Cybercrime, 2008; Hall, 1996; Monthy, 1989). However, in recent times, spammers have stepped beyond to the extent of escaping mechanisms that could trap their messages through filtering–(Awad and. ELseuofi, 2011). Consequently, global research efforts are concentrated on the development of varied spam filtering techniques because current spam filters is prone to collapse if the spam keywords are manipulated or avoided in the email system (Almomani *et al*., 2015). Like other types of filtering programs, a spam filter looks for certain criteria on which it bases its' judgments (Hall, 1996; Rekha and Sandeep, 2014).

2.0 Bayesian Spam Filtering Method

This is a content-based spam filtering method that contains the word probability database that checks for the matches of any of the contents of the mail against the suspicious terms in the database table named (Offensive) (Okunade *et al*., 2009). Content-based spam filtering is a promising filtering approach capable of executing automatic identification of spam and legitimate email messages (Andrej, *et al*., 2006). It employs the laws of mathematical probability to determine which messages are legitimate (ham) and those that are spam. The word probabilities (also known as likelihood functions) are used to compute the probability that an email with a particular set of words in it belongs to either of the categories. This contribution is called the posterior probability and is computed using Bayes' theorem (Christina *et al*., 2010). It searches for the keywords in the mail, that is it scans through the mail content for suspicious related terms. This is a simple language analysis, which operates by matching match specific terms or phrases. This method makes use of the Bayesian Statistical Probability formula (Process, 2010). The probability formula enhances each term be checked, compare and contrast for the similarity/equality with the terms enlisted in the content of the Offensive table in the database where the entire mail can then be classified to be Mail/Spam, depending on the result values of the calculation of the suspicious terms ~~calculated~~. If the result value calculated (that is the Spamicity value) or Probability value calculated is less than or equal to (<=) 0.5 (set threshold), the entire mail will be classified as Ham and will be sent to the Ham folder of the recipient inbox but if otherwise (that is the Spamicity value) greater than (>) 0.5 (set threshold), then the entire mail can be classified as Spam. The Spamicity value of 0.5 is neutral, meaning that it does not affect the decision as to whether a message is a spam or not. (See Fig. 1).

2.2 Bayesian Spam Filtering Method Incorporated with Word Stemming Technique

Stemming is the removal of all unwanted prefixes, affixes and suffixes from a term to generate its actual value/root. When an incoming mail is received through the Mail Transfer Agent (MTA), it will pass through the word Stemmer where the term stemming processing activities will be implemented through checking and extraction, when it comes across any of unwanted special characters used to modified the suspicious terms to deceit the Bayesian filters. Also, the word stemmer will equivalent any identified modified terms to their original value if any of the characters of the suspicious terms is been rearranged/modified to foil the Bayesian filter. Having done the above Word Stemming process activities on the mail content, mail can then be transferred to the Spam filter using next to the Bayesian Statistical Probability formula as used above in the Bayesian Spam filtering process. See figure 2.

3.0 Materials and Methods

The experimental setup shown in Fig. 1 is the executing process of the pure Bayesian Statistical filtering process while Fig. 2 is the executing process of the Bayesian Statistical filter incorporated with the Word Stemmer (Word Stemming process), and the experimental result is shown in Fig. 1.

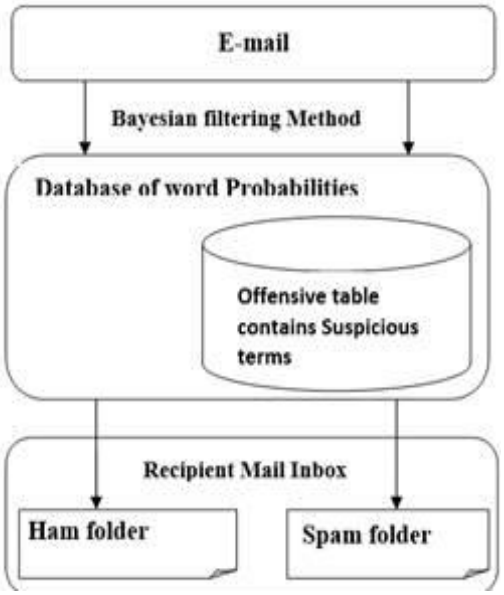

Fig. 1: Bayesian Spam filtering Method Experimental setup

Bayesian filtering Method Experimental Algorithm

Step 1

Let $Productsum = 1, $Differentialsum=1

Step 2

If Term > Total Mail terms

End

Step 3

Check the input terms against the Suspicious terms database

If fund/matches?, then

Calculate values

End

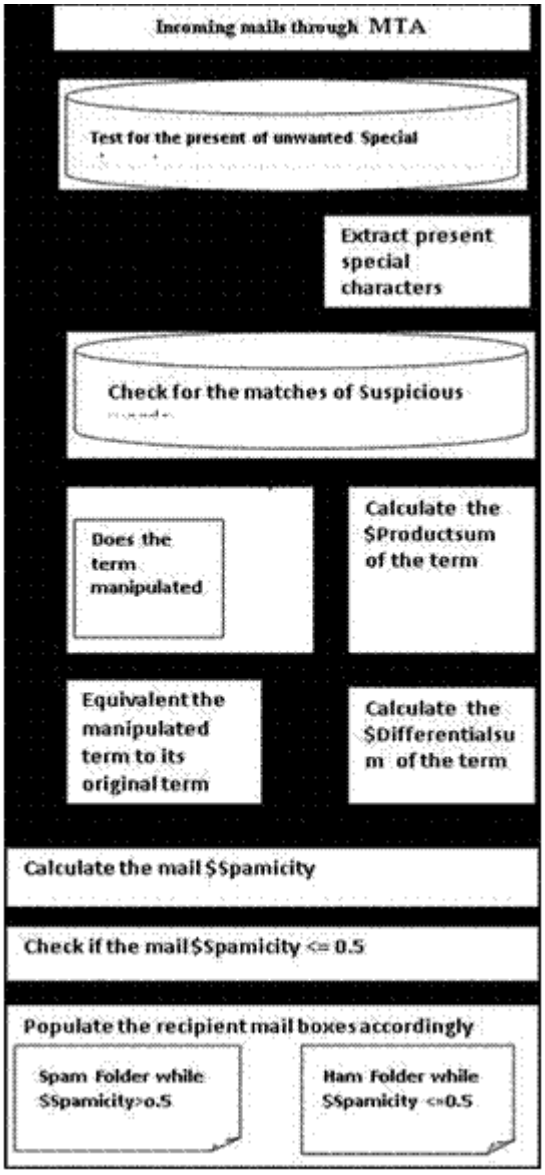

Fig. .2: Bayesian Spam filter Incorporated with Word Stemmer Experimental flow process

Step 4

Calculate Values:

*$Productsum *= $Spamicitydb*

*$Differentsum *= (1 - $Spamicitydb)*

Step 5

$Spamicity = $Productsum / ($Productsum + $Differentialsum)

If $Spamicity <=0.5 Then

Populate the recipient mail box Ham folder

Else, Populate the recipient mailbox Spam/Junk folder

End

Bayesian Spam filter Incorporated with Word Stemmer Experimental Algorithm

Step 1

Let $Productsum = 1, $Differentialsum=1

Step 2

If Term > Total Mail terms

End

Step 3

Check input term against unwanted Special characters used as prefix, infix and suffix

If any fund?, then

Extract them all

End

Step 4

Check the input terms against the Suspicious terms database

If fund/matches?, then

Calculate values

End

Step 5

Check input term for any modified/manipulated suspicious terms If fund, then Equivalent it to the actual suspicious term

End

Step 6

Calculate Values:

*$Productsum *= $Spamicitydb*

*$Differentsum *= (1 - $Spamicitydb)*

Step 7

$Spamicity = $Productsum / ($Productsum + $Differentialsum)

If $Spamicity <=0.5 Then

Populate the recipient mail box Ham folder

Else, Populate the recipient mailbox Spam/Junk folder

End

3.0 Results and Discussion

Chart 1 shows the result of the Experiment of execution time interval (shown in Figs. 1 and 2) conducted in the previous section. X-axis signifies spam mail content measured per number of words that make up the spam mail content (such as 173,199,..,…, and 448) and the y-axis signifies the time it takes an Algorithm to

complete each execution, measured per second. As a group of numbers of words made up of a complete mail, from chart 1 bellow appeared each number values in twos that is: 173,173,199,199,…,448,448, the first mail value (1st "173") is the spam mail executed without manipulating the Suspicious terms while the 2nd mail with "173" numbers of words/terms is the spam mail with manipulated Suspicious terms, the third spam mail (1st "199") is the spam mail without manipulated Suspicious terms while the forth mail (2nd "199")is the spam mail with manipulated Suspicious terms and so on up to the second to the last mail (1st "448") is spam mail without manipulate Suspicious terms and the last mail (2nd "448") is spam mail with manipulated Suspicious terms.

They appeared to be a longer execution time per second for each mail. The y-axis represents the execution time interval of the algorithm with word stemmers (which appeared in blue) while the shorter execution value per second on the y-axis is the execution time of the Algorithm without the word Stemmer (which is indicated with brown colour).

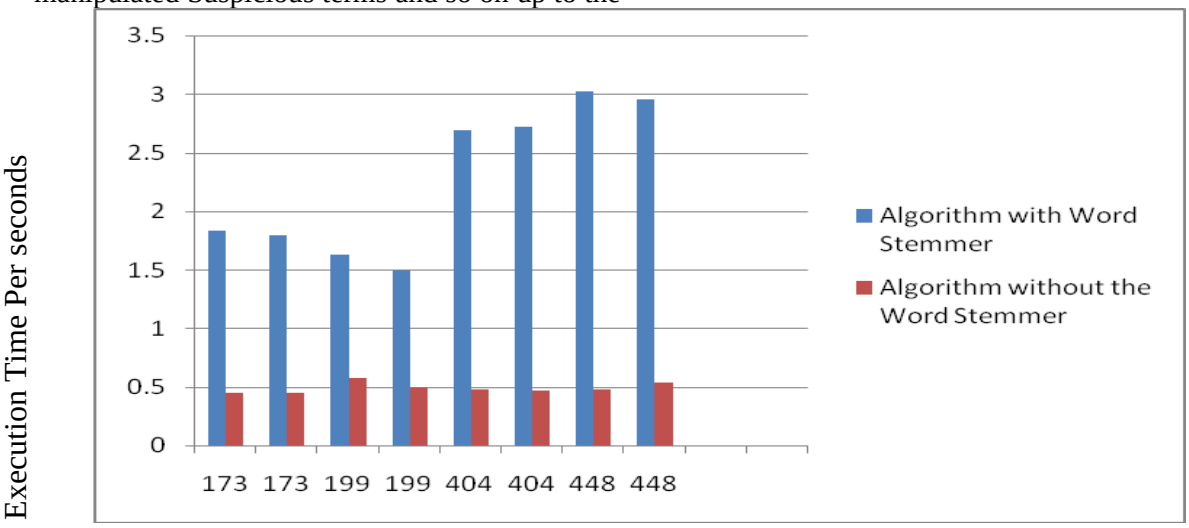

Fig. 3: Mail per Words or Tokens

Fig. 1: The result of The Execution Time comparison of the Bayesian Statistical Spam filter and Bayesian Statistical Incorporated with the Word Stemming Spam filter.

The result of execution time comparison of the two Algorithm experiments indicated that the execution time of Bayesian incorporated with the Word Stemmer is significantly longer compared to that of ordinary Bayesian mail classification. Also, that suspicious terms manipulation has no or less effect on the execution time of both algorithms.

4.0 Conclusion

Our experiment showed that the execution of a mail classifier using the Word Stemmer incorporated with the Bayesian mail filter takes a lot of time in execution compares to that of an ordinary Bayesian mail classifier. Also with this, we can easily know the amount of time the Algorithm actually spent in executing the manipulated terms only before the real execution of the Bayesian aspect of the Algorithm since the second Algorithm figure 2 can give us the time spent in executing the Bayesian filer only by way of subtracting the result gotten from ordinary Bayesian classifier/filter from the result of the word stemmer incorporated with the Bayesian filter/classifier.

5.0 References

Albercht. K, (2006). *Mastering Spam: A Multifaceted Approach with the Spamato Spam Filter System* DSS. ETH NO. 16839

Ali, C. and Tunga, G. (2007). *Time-efficient spam e-mail filtering using n-gram models*. Department of Computer Engineering, Bogazic University, Istanbul 34342, Turkey

Alireza, N. P., Raheleh, K. & Soheil, B. R. (2012). Minimizing The Time of Spam Mail Detection by Relocating Filtering System to the Sender Mail Server. International *Journal of Network Security & Its Applications* (IJNSA), 4, 2. doi:10.5121/ijnsa.2012.4204 53

Almomani, A., Obeidat1,A., Alsaedi, A., Obaida, M.A. & Al-Betar, M. (2015). Spam E-mail Filtering using ECOS Algorithms. *Indian Journal of Science and Technology*,8, S9, pp. 260-272

Andrej, B., Gordon, V. C., Bogdan, F. C., Thomas, R. L. & Blaz, Z. (2006). Spam

Filtering Using Statistical Data Compression Models. *Journal of Machine Learning Research*, 6, pp. 2673-2698.

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, DOI: 10.5121/ijcsit.2011.3112 173.

Christina, V., Karpagavalli, S. & Suganya, G. (2010). A Study on Email Spam Filtering Techniques. *International Journal of Computer Applications*, 12, 1, DOI: 10.5120/1645-2213

Fight Cybercrime (November, 2008) *Anti-phishing Techniques* SpamAlert.org

Garacia, F.. D, Hoepman. J and Nieuwenhuizen, J. (2004) Twente, Netherlands

Hall, R. J, (1996). *Channels: Avoiding unwanted electronic mail*. In Proceeding .DIMACS Symposium on Network Threats. DIMACS.

Monthy, P. (1989). *Flying Circus. Just the word*. Volume 2, Chapter 25, pg 27-28. Menthuem Publishing Ltd.

Okunade O. A, Robert A.B.C, Longe O.B & Onifade O.F.W (2009): *Word-Stemming*

Algorithms to Improve Bayesian Classification of Electronic Spam Mails. International Conference of the Nigerian Computer Society. Volume 20th Mathglo 2009 pg 215 - 220, Abuja FCT, Nigeria. www.ncs.org.ng

Process (2010). *Bayesian Filtering Example Using Bays' formula to Keep Spam Out of Your Inbox.* http://www.process.com/

Rekha & Sandeep N. (2014). A Review on Different Spam Detection Approaches.

International Journal of Engineering Trends and Technology (IJETT), 11, pp.315-319.

Sanjay, K. N. (2015). Spam Filtering using the Social Anthropology and Data Mining Technique. International Journal of Computer Science and Mobile Computing. IJCSMC, , 44, pp. 234-237.

Savita, T. & Santoshkumar, B. (2014). *Effective Spam Detection Method for Email. IOSR*

Journal of Computer Science (IOSR-JCE). e-ISSN: 2278-0661, p-ISSN: 2278-8727 PP 68-72 *www.iosrjournals.org*

Tian, X. (2020). *Constant Time Complexity Spam Detection Algorithm for Boosting Throughput on Rule based Filtering Systems*. IEEE Access.

Zhe, W., William, J., Qin, L., Moses, C. and Kai, L.(2007). *Filtering Image Spam with Near-Duplicate Detection*. Computer Science Department, Princeton University, USA

Conflict of Interest

The author declared no conflict of interest