Feature Selection and Similarity Coefficient Based Method for Email Spam Filtering

[1]Ali Ahmed A.Abdelrahim, [2]Ammar Ahmed E. Elhadi, [3]Hamza Ibrahim, [1]Naser Elmisbah
[1]Faculty of Engineering, Karary University, Omdurman, Sudan
alikarary@gmail.com

[2]Faculty of Electrical Engineering, Universiti Teknologi Malaysia, Skudai, Malaysiay
hamysra76@hotmail.com

[3]Almashriq College for Science and Technology, Khartoum North, Sudan
ammareltayeb@gmail.com

***Abstract*—*Many threats in the real world can be related to activities of persons on the Internet. Spam is one of the most pressing security problems online. Spam filters try to identify likely spam either manually or automatically. Most of the spam datasets used in the spam filtering area of study deal with large amounts of data containing irrelevant and/or redundant features. This redundant information has a negative impact on the accuracy and detection rate of many methods that have been used for detection and filtering. In this study, statistical feature selection approach combined with similarity coefficients are used to improve the accuracy and detection rate for the spam detection and filtering. At the end, the study results based on email spam datasets show that our proposed approach enhanced the detection rate, false alarm rate and the accuracy.*

Keywords*—*Spam; spam filtering; feature selection; similarity coefficient

I. INTRODUCTION

Spam is generally defined as "unsolicited, usually commercial, email sent to a large number of recipients" [1]. Spam is also known as unsolicited or junk email. Just like the junk email you get at home advertising everything from credit cards to local restaurants, email spam operates in the same way [2]. Spammers send out hundreds of thousands, and sometimes tens of millions, of emails to unsuspecting email recipients. These spam emails are usually trying to sell something. While most people delete these spam email messages without even reading them, a small percentage of email recipients open and read the email messages and sometimes even buy the products being sold. This is what makes it profitable for the spammers. It costs very little to send an email message. Therefore only a small percentage of people who receive spam need to make a purchase to make it profitable for the spammers.

An electronic message is spam if (A) the recipient's personal identity and context are irrelevant because the message is equally applicable to many other potential recipients, and (B) the recipient has no verifiably granted deliberate, explicit, and still-revocable permission for it to be sent [3].

The problem named spam has come to existence with the widespread usage of electronic mail (email) which not only wastes the time of the users, but also brings about other problems such as influencing bandwidth and misusing storage space [4].

Spam detection methods try to identify likely spam either manually or automatically, and then act upon this identification by either deleting the spam content or visibly marking it as such for the user.

Spam filters are email programs that attempt to organize email according to criteria that the user specifies. The ultimate goal is to filter out all unwanted email. Spam filters use a variety of techniques to determine which emails are spam. Most spam filters offer the user a variety of options for how to handle spam. Users can choose to have the emails automatically deleted, sent to a spam folder or delivered to their normal inbox marked as spam. Most spam filters will turn off all links contained in emails deemed to be spam as a protection. These links can be turned back on, however, if the user determines the email is not spam.

Most email programs come equipped with basic spam filter features. However, if users desire greater protection or control over spam, they can purchase spam filtering software.

Spam filters yield outstanding results. Laboratory testing shows that a content-based learning filter can correctly classify all but a few spam messages out of a hundred and all but a few thousand non-spam messages out of a thousand.

There is some evidence that similar results may be achieved in practice either by machine learning methods or by other methods like blacklisting, grey listing, and collaborative filtering. The controlled studies necessary to measure the effectiveness of all types of filters—and combinations of filters—have yet to be conducted. We argue that understanding and improving the effectiveness of spam filters is best achieved through a combination of laboratory and field studies, using common measures and statistical methods. According to Araúzo-Azofra and Benítez [5] and Nizamani et al. [6] by using feature selection methods one can improve the accuracy, applicability, and understand ability of the learning process.

978-1-4673-6232-0/13/$31.00 ©2013 IEEE

Feature selection is the process of finding an optimal subset of features that contribute significantly to the classification [7]. Selecting a small subset of features can decrease the cost and the running time of a classification system. It may also increase the classification accuracy because irrelevant or redundant features are removed. Our previous studies proved that feature selection was used to enhance the recall of the similarity search methods [8, 9].

Similarity coefficients are used to obtain a numeric quantification of the degree of similarity between a pair of structures (sentence, molecule, spam, etc.). There are four main types of similarity coefficients: association coefficients, distance coefficients, correlation coefficients and probabilistic coefficients [10, 11].

This paper is organized as follows. Section II describes the previous studies related to spam detection. Section III describes the material and methods used in this study. The results and discussion are illustrated in Section IV. Section V concludes the paper.

II. RELATED WORK

The traditional methods of filtering spam based on signature; a signature should be able to identify any spam exhibiting the malicious behavior specified by the signature. Most anti-spam scanners are signature based. Signature-based spam detection technique cannot meet its security challenges [12], since it can only detect a small number of generics or extremely broad signatures and thus is poor in detecting new spam threats.

Sang Min Lee [1] proposed a spam detection model based on Random Forests (RF) using parameters optimization and feature selection simultaneously. Liang et al. [13] performed the feature selection using feature ranking algorithm but the detection rates are very low. Selvakuberan et al. [14] have applied filtered feature selection on web page classification; according to their results, the evaluator CfsSubsetEval yields better performance with search methods BestFirst, Ranker search, and Forward selection. Pineda-Bautista et al. [15] proposed a method for selecting the subset of features for each class in multi-class classification tasks. Bursteinas et al. [16] and Zhu [17] have performed the feature selection but they have not mentioned how they decided the number of important features and provided the variable importance of each feature as a numerical value.

The problem of spam filtering is the high dimensionality of feature space. The feature space that contains words or phrases in the documents has more than ten thousands features, which is a great preventive problem for many of the similarity methods. For this reason, there is a strong need for reducing the dimensions.

III. MATERIAL AND METHODS

This study has compared three evaluation metrics, which are detection rate, false alarm rate and accuracy, and observed the effects of the feature selection process for the seven association coefficients shown in Table I. The dataset consists of 1813 spam and 2788 non-spam samples created by Hopkins et al. [18]. The results were obtained two times, first based on all features of the spam messages, and second, after applying the features selection process.

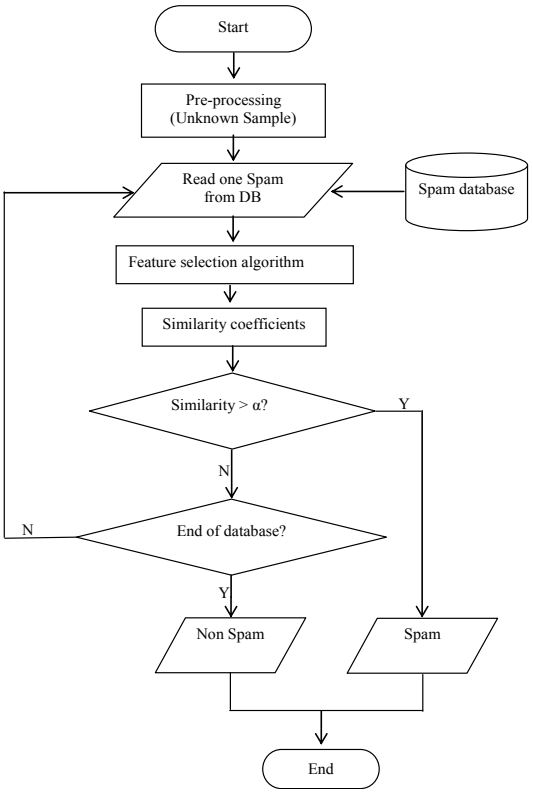

Fig. 1. proposed method flowchart.

TABLE I. ASSOCIATION (SIMILARITY) COEFFICIENTS

Coefficient	Formula	Range
Cosine	$\frac{a}{\sqrt{(a+b)(a+c)}}$	0 to 1
Dice	$\frac{2a}{2a+b+c}$	0 to 1
Rao	$\frac{a}{n}$	0 to 1
Sokal	$\frac{a}{a+2b+2c}$	0 to 1
Simple Matching	$\frac{a+d}{n}$	0 to 1
Hamann	$\frac{a+d-b-c}{n}$	-1 to 1
Tanimoto/Jaccard	$\frac{a}{a+b+c}$	0 to 1

Statistical analysis method for feature selection was performed using SPSS Clementine 11.1. Class identifiers were set as output variables and the all other spam features as input

variables. All features were classified as continuous. Finally, the features that contributed to a class identifier were selected.

A feature selection algorithm was applied to find important features which showed a strong correlation with the class identifier. The algorithm considered one attribute at a time to see how well each predictor alone predicted the target variable. The importance value of each variable was then calculated as (1-p) where p is the association strength between the candidate predictor and the target variable. Since the target values were continuous, p-values based on the F-statistic. The general flowchart of the proposed method is outlined in Figure. 1.

For each similarity coefficient described in the above table, n is the total number of feature positions in the strings representing the two spams compared, b is the number of feature positions set in only one of the two spams, while c is the number of feature positions set in only the other spam. Finally, d is the number of n features not set in either one of the spams, and a is the number of features set in both spams. Thus, n = a + b + c + d.

A. Simulation Setup

In this study, a set of experiments were conducted on real spam samples to evaluate the detection rate and the accuracy of the different similarity methods; two sets of experiments were designed to investigate the impact of using feature selection over the different similarity methods and show its detection rate and accuracy. In the first set of experiments, the similarities between spams to spam were presented to evaluate detecting spam and compare the different similarity methods before feature selection and after it. In the second set of experiments we evaluate the detection rate, false alarm rate and accuracy for the different similarity methods before feature selection and after it.

A prototype using Matlab programming language has been implemented. All experiments were conducted on AMD Phenom™ II X4 with a 3.25 GHz processor running Windows 7 and a memory of 4.00 GB. The dataset was downloaded from ftp://ftp.ics.uci.edu/pub/machine-learningdatabases/spambase/.

B. Evaluation Measures

To evaluate the proposed mechanism three metrics are used, first, detection rate is defined as the percentage of correctly identified spam samples as spam, as in

$$Detection\ Rate = TP/(TP + FN) \quad (1)$$

Second, false alarm rate is the percentage of non-spam labeled as spam samples, which the number of non-spam samples classified as spam divided by the total number of non-spam samples, as illustrated in

$$False\ Alarm\ Rate = FP/(FP + TN) \quad (2)$$

Last, accuracy, which is the overall accuracy of the system to detect spam and non-spam files, as shown in

$$Accuracy = (TP+TN)/(TP + TN+FP+ FN) \quad (3)$$

IV. Results and Discussion

In the first set of experiments, the similarity was calculated for all spam to spam samples using the different similarity methods. Table II shows statistics for the different similarity methods before feature selection. In Table II, Hamann and Matching succeed to detect all spams with maximum similarity 1, but Matching method is better than Hamann because of a better minimum similarity of 0.8. Cosine and Dice methods come after that with 1807 spam samples successfully detected, Sokal and Tanimoto detected 1792 samples, whereas Rao method had the worst detection rate.

Table III shows the same statistics after feature selection; the main observation after applying feature selection is that the similarity value was increased in the Cosine, Hamann, Matching and Rao methods. As for the other methods, it is not clear if their similarity value was increased, but their rate of success to detect was, which proved that the similarity value of some samples was increased (Rao, Tan). After applying feature selection, Dice and Sokal methods were affected negatively in their similarity values.

TABLE II. STATISTICAL RESULTS BEFORE FEATURE SELECTION

Similarity method	Max similarity	Min similarity	Success	Fail
Cosine	1	0.45	1807	6
Dice	1	0	1807	6
Hamann	1	0.59	1813	0
Matching	1	0.8	1813	0
Roa	0.56	0	54	1759
Sokal	1	0	1792	21
Tanimoto	1	0	1792	21

TABLE III. STATISTICAL RESULTS AFTER FEATURE SELECTION

Similarity method	Max similarity	Min similarity	Success	Fail
Cosine	1	0.47	1812	1
Dice	1	0	1800	13
Hamann	1	0.62	1813	0
Matching	1	0.81	1813	0
Roa	0.59	0	116	1697
Sokal	1	0	1727	86
Tanimoto	1	0	1795	18

In the second set of experiments, the similarity was calculated for all spam to non-spam samples using the similarity methods, and the results of the previous experiment were included to generate the following comparison. The results of this experiment are illustrated in Figure 2, Figure3 and Figure 4. The figures illustrated the detection rate, false alarm rate and accuracy for all compared methods before and after feature selection.

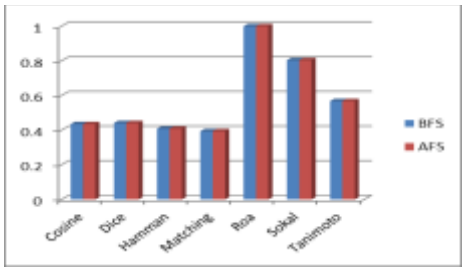

Fig. 2. Detection Rate before and after feature selection.

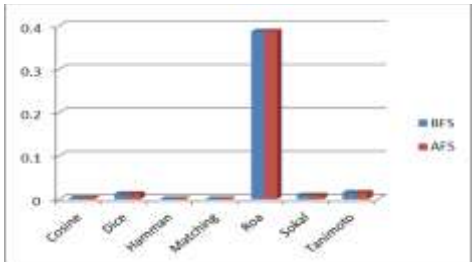

Fig. 3. False alarm Rate before and after feature selection

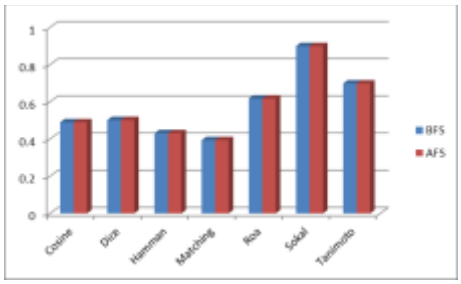

Fig. 4. Accuracy before and after feature selection

According to the result shown in the previous figures, before applying features selection, the highest value of accuracy and detection rate with the lowest false positive rate was achieved by the Sokal method. Tanimoto and Rao methods have acceptable results compared with all other methods that have low accuracy and detection rates. After applying feature selection, the main observation is that the accuracy and detection rates were improved in the Tanimoto, Sokal, Rao, Dice and Cosine methods, while they were not affected for the rest of the methods.

V. CONCLUSION

This study has proved that feature selection has a positive impact on the similarity methods used for spam filtering. Feature selection increased spams filter accuracy and detection rate. Also, the degree of similarity between spam to spam samples was increased.

ACKNOWLEDGMENT

The authors would like to thank our colleagues and staff at the Faculty of Engineering at Karary University, Almashriq College for Science and Technology and the Faculty of Electrical Engineering at the Universiti Teknologi Malaysia for their contribution and comments.

REFERENCES

[1] S. M. Lee, D. S. Kim, J. H. Kim, and J. S. Park, "Spam detection using feature selection and parameters optimization," in Complex, Intelligent and Software Intensive Systems (CISIS), 2010 International Conference on, 2010, pp. 883-888.

[2] T. Bogers and A. Van den Bosch, "Using language models for spam detection in social bookmarking," in Proceedings of 2008 ECML/PKDD Discovery Challenge Workshop, 2008, pp. 1-12.

[3] T. Subramaniam, H. A. Jalab, and A. Y. Taqa, "Overview of textual anti-spam filtering techniques," Int. J. Phys. Sci, vol. 5, pp. 1869-1882, 2010.

[4] A. Beiranvand, A. Osareh, and B. Shadgar, "Spam Filtering By Using a Compound Method of Feature Selection," Journal of Academic and Applied Studies, vol. 2, pp. 25-31, 2012.

[5] A. Araúzo-Azofra and J. M. Benítez, "Empirical study of feature selection methods in classification," in Hybrid Intelligent Systems, 2008. HIS'08. Eighth International Conference on, 2008, pp. 584-589.

[6] S. Nizamani, Memon, N., Wiil, U.K., Karampelas, P., "Modeling suspicious email detection using enhanced feature selection," Int. J. Modeling and Optimization, vol. 2, pp. 371-377, 2012.

[7] A. R. Behjat, A. Mustapha, H. Nezamabadi-pour, M. Sulaiman, and N. Mustapha, "GA-based feature subset selection in a spam/non-spam detection system," in Computer and Communication Engineering (ICCCE), 2012 International Conference on, 2012, pp. 675-679.

[8] A. Ahmed, A. Abdo, and N. Salim, "An Enhancement of Bayesian Inference Network for Ligand-Based Virtual Screening using Features Selection," American Journal of Applied Sciences, vol. 8, pp. 368-373, 2011.

[9] A. Ahmed, A. Abdo, and N. Salim, "Ligand-based Virtual screening using Fuzzy Correlation Coefficient," International Journal of Computer Applications, vol. 19, pp. 38-43, 2011.

[10] D. Ellis, J. Furner-Hines, and P. Willett, "Measuring the degree of similarity between objects in text retrieval systems," Perspectives in Information Management, vol. 3, pp. 128-149, 1993.

[11] J. Willett, Similarity and clustering in chemical information systems: John Wiley & Sons, Inc., 1987.

[12] A. A. E. Elhadi, M. A. Maarof, and A. H. Osman, "Malware detection based on hybrid signature behavior application programming interface call graph," American Journal of Applied Sciences, vol. 9, pp. 283-288, 2012.

[13] J. Liang, S. Yang, and A. Winstanley, "Invariant optimal feature selection: A distance discriminant and feature ranking based solution," Pattern Recognition, vol. 41, pp. 1429-1439, 2008.

[14] K. Selvakuberan, M. Indradevi, and R. Rajaram, "Combined feature selection and classification—a novel approach for the categorization of web pages," Journal of Information and Computing Science, vol. 3, pp. 083-089, 2008.

[15] B. B. Pineda-Bautista, J. A. Carrasco-Ochoa, and J. F. Martínez-Trinidad, "Taking advantage of class-specific feature selection," in Intelligent Data Engineering and Automated Learning-IDEAL 2009, ed: Springer, 2009, pp. 1-8.

[16] B. Bursteinas and J. Long, "Transforming supervised classifiers for feature extraction," in Tools with Artificial Intelligence, 2000. ICTAI 2000. Proceedings. 12th IEEE International Conference on, 2000, pp. 274-280.

[17] Z. Zhu, "An Email Classification Model Based on Rough Set and Support Vector Machine," in Fuzzy Systems and Knowledge Discovery, 2008. FSKD'08. Fifth International Conference on, 2008, pp. 236-240.

[18] Spambase Dataset, ftp://ftp.ics.uci.edu/pub/machine-learningdatabases/spambase/