American Journal of Applied Sciences

Original Research Paper

Features Reweighting and Similarity Coefficient Based Method for Email Spam Filtering

[1]Ahmed Osman Ali Elsiddig, [2,4]Ammar Ahmed E. Elhadi and [2,3]Ali Ahmed

[1]*University of Science and Technology - Khartoum, Sudan*
[2]*Software Engineering, Mashreq University Khartoum North, Sudan*
[3]*Faculty of Engineering, Karary University, Omdurman, Sudan*
[4]*Department of Foundation, Inaya Medical College, Riyadh*

Article history
Received: 28-02-2017
Revised: 09-08-2017
Accepted: 11-10-2017

Corresponding Author:
Ammar Ahmed E. Elhadi
Software Engineering, Mashreq University - Khartoum North, Sudan
E-mail: ammareltayeb@gmail.com

Abstract: Spam is flooding the Internet with many copies of the same message, in an attempt to force the message on people who would not otherwise choose to receive it. Anti spam by determining whether or not an incoming email is spam has become an important problem. One of the main characters or the problem of Spam filtering is its high dimension of space feature. For this reason, we need a reducing stage of dimensions. This study tried to cover this side from spam detection techniques by study the effect of re-weight of features. The works started by applying similarity coefficient in the dataset and then re-weight the features in the dataset and applying similarity coefficient in the new data set. Finally make a Comparison between the result before and after re-weight and Comparison with feature selection method. The objective of this Thesis is: Study the similarity coefficient (Cosine and Dice) and Study the effects of the important feature to other features through the re-weight process. The most important results of this study are: Reweighting process did not improve the success rate of any of the two methods (Cosine and Dice). Also, Feature selection method led to improve detection in Cosine, while reweighting method not improve detection any of (Cosine or Dice).

Keywords: Spam, Spam Filtering, Feature Selection, Similarity Coefficient

Introduction

Email spam or junk mail, or unsolicited commercial email (Michael and Mattord, 2012) is process of sending not required email messages, frequently with commercial content, in large quantities to an indiscriminate set of recipients. Spam in email started to become a problem when the Internet was opened up to the general public in the mid-1990s. It grew exponentially over the following years and today composes some 80 to 85 percent of all the e-mail in the World, by a "conservative estimate". Pressure to make email spam illegal has been successful in some jurisdictions, but less so in others. The efforts taken by governing bodies, security systems and email service providers seem to be helping to reduce the onslaught of email spam (Wikipedia, 2015a).

Spam is no more garbage but risk since it recently includes virus attachments and spyware agents which make the recipients' system ruined, therefore, there is an emerging need for spam detection (Lee *et al*., 2010).

Many spam detection techniques based on machine learning algorithms have been proposed. As the amount of spam has been increased tremendously using bulk mailing tools, spam detection techniques should deal with it. For spam detection, parameters optimization and feature selection have been proposed to reduce processing overheads with guaranteeing high detection rates (Lee *et al*., 2010).

The techniques currently used by most anti-spam software are static, meaning that it is fairly easy to evade by tweaking the message a little. To do this, spammers simply examine the latest anti-spam techniques and find ways how to dodge them (GFI, 2011).

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.

Related Work

Statistical feature selection approach combined with similarity coefficients are used to improve the accuracy and detection rate for the spam detection and filtering (Abdelrahim *et al*., 2013). At the end, the study results based on email spam datasets show that the proposed approach enhanced the detection rate, false alarm rate and the accuracy. Study was 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.

One of the main characters or the problem of Spam filtering is its high dimension of space feature. The feature space that contains words or phrases in the documents has more than ten thousand features, which is a great preventive problem for many of the machine learning algorithms (Beiranvand *et al*., 2012). For this reason, we need a reducing stage of dimensions. The previous approaches have not taken into account the importance of weights of features and there are no previous studies discuss this topic. So, in this study, we tried to cover this side from spam detection techniques by study the effect of re-weight of features.

Materials and Methods

The data used in this study, created by Mark Hopkins, Erik Reeber, George Forman, Jaap Suermondt in Hewlett-Packard Labs. It was generated in June-July 1999. The dataset is available at ftp://ftp.ics.uci.edu/pub/machine-learningdatabases/ spambase/. The number of Instances in the data set are: 4601 (1813 Spam = 39.4%, 2788 Non-Spam = 60.6%), while the number of Attributes are: 58 (57 continuous, 1 nominal class label) (Spambase Dataset).

The collection of spam e-mails came from the postmaster and individuals who had filed spam. While the collection of non-spam e-mails came from filed work and personal e-mails and hence the word 'mark' and the area code '430' are indicators of non-spam. These are useful when constructing a personalized spam filter. One would either have to blind such non-spam indicators or get a very wide collection of non-spam to generate a general purpose spam filter (Hastie *et al*., 2001).

In this study, used MATLAB software in process of experimentation and calculation results, Also, MS Excel used in order to organize Results. The experiment was implemented on ASUS laptop contains Intel core i3 with a 1.8 GHz processor running Windows 7 and a memory of 4.00 GB.

The works will start by applying similarity coefficient in the dataset and then re-weight the features in the dataset and applying similarity coefficient in the new data set. Finally make a Comparison between the result before and after re-weight and Comparison with feature selection method.

The Experiment Phases

Figure 1 represents the steps that have been taken to achieve the comparison between Reweighting and features selection.

Phase (1): Similarity

In this phase the work started by applying the similarity coefficient (cosine and dice) on the spam database, the objective of this phase is to calculate the accuracy of the spam detection when apply the cosine and dice only.

Cosine Similarity

Measuring of similarity between two vectors of an inner product space that measures the cosine of the angle between them is called Cosine similarity. The cosine of 0° is 1 and it is less than 1 for any other angle. It is thus a judgment of orientation and not magnitude: Two vectors with the same orientation have a Cosine similarity of 1, two vectors at 90° have a similarity of 0 and two vectors diametrically opposed have a similarity of -1, independent of their magnitude. Cosine similarity is particularly used in positive space, where the outcome is neatly bounded in [0, 1] (Wikipedia, 2015c).

Dice Coefficient

Using statistic for comparing the similarity of two samples is defined Dice coefficient. It's mainly useful for ecological community data, As compared to Euclidean distance, Dice distance retains sensitivity in more heterogeneous data sets and gives less weight to outliers. The outcome in the range [0, 1] (Wikipedia, 2015c).

Phase (2): Reweighting

Contains two levels as follows:

- Level (1): In this level, the spam database was re-weighed, by dividing the largest value in every feature on all other features in the same column
- Level (2): During this level, similarity coefficient (cosine and dice) was applied again, the objective of this level is to calculate the accuracy of spam detection after reweighting the spam database and apply cosine and dice

Phase (3): Results

In this phase Comparison was done between reweighting and feature selection in order to get the best method between them.

Experiment Steps

The objective of the experiment is to calculate the accuracy of the spam detection from the (spam database

and Reweighted spam data base), by using similarity coefficient (Cosine and Dice), so in order to clarify the processes that has done during the experimental work, we will use a Mini sample of the spam database.

Step (1)

Selected Mini sample includes 10 spam message and 10 non spam messages (labeled with the letters A to J in the tables). Note that the main data base contains 4601 message (1813 Spam = 39.4%, 2788 Non-Spam = 60.6%).

In this sample rows represents the spam, while columns represent features. Also rows form 1-10 is spam while other rows are non-spam.

Step (2)

In this step the similarity coefficient (Cosine and Dice) was apply in Mini sample As follows:

Cosine Formula (Wikipedia, 2015b):

$$\frac{a}{\sqrt{(a+b)(a+c)}} \quad (1)$$

Dice Formula (Wikipedia, 2015c):

$$\frac{2a}{2a+b+c} \quad (2)$$

where, $a + b + c$ is the total number of feature positions in the strings, a is the number of features set in both spams, 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.

Now, through the use of MATLAB, the output was as follows:

- Based on the results described on Table 1-5, it is clear that 60% is the best similarity rate, where it is in the case of Cosine was discovered 7 spams out of 10 spams
- Based on the results described on Table 6-10, it is clear that 50% is the best similarity rate, where it is in the case of Dice was discovered 7 spams out of 10 spams

When comparing the Cosine and Dice it is clear that Cosine similarity (detect 7 spams with similarity 60%) is best than Dice similarity (detect 7 spams with similarity 50%).

Step (3)

At this stage, start the process of re-weight spam database, through a process of division on the largest value in the column (largest Feature) as follows:

$$Re-weight\ Features = \frac{F(ij)}{LF(j)} \quad (3)$$

where, as $F(ij)$ *is* feature in row (i) and column (j), while $LF(j)$ is largest feature in the column (j).

Table 1. Minimum sample of the results after applying Cosine algorithm on the spam database

A	B	C	D	E	F	G	H	I	J
0.612372	0.612372	0.408248	0.258199	0.258199	0.408248	0.408248	0.408248	0.471405	0.654654
0.408248	0.875000	0.875000	0.632456	0.632456	0.500000	0.500000	0.500000	0.721688	0.801784
0.258199	0.632456	0.790569	0.790569	0.790569	0.500000	0.500000	0.500000	0.866025	0.801784
0.258199	0.632456	0.790569	1.000000	1.000000	0.632456	0.632456	0.632456	0.730297	0.507093
0.408248	0.500000	0.500000	0.632456	0.632456	0.632456	0.632456	0.632456	0.730297	0.507093
0.408248	0.500000	0.500000	0.632456	0.632456	0.500000	0.500000	1.000000	0.288675	0.267261
0.408248	0.500000	0.500000	0.632456	0.632456	1.000000	0.500000	0.500000	0.577350	0.267261
0.471405	0.721688	0.866025	0.730297	0.730297	0.288675	0.577350	0.288675	0.288675	0.267261
0.654654	0.801784	0.801784	0.507093	0.507093	0.267261	0.267261	0.267261	0.771517	0.771517

Table 2. Maximum value of the features after applying Cosine algorithm on the spam database

A	B	C	D	E	F	G	H	I	J
0.654654	0.875	0.875	1	1	1	0.632456	1	0.866025	0.801784

Table 3. Minimum value of the features after applying Cosine algorithm on the spam database

A	B	C	D	E	F	G	H	I	J
0.258199	0.5	0.408248	0.258199	0.258199	0.267261	0.267261	0.267261	0.288675	0.267261

Table 4. Average value of the features after applying Cosine algorithm on the spam database

	B	C	D	E	F	G	H	I	J
0.43198	0.641751	0.670244	0.64622	0.64622	0.525455	0.501975	0.525455	0.605103	0.538412

Table 5. Result of spam detection after applying Cosine algorithm on the spam database

Similarity	50%	60%	70%	80%	90%
	0	0	0	0	0
	1	1	1	1	0
	1	1	1	1	0
	1	1	1	0	0
	1	1	1	0	0
	1	1	0	0	0
	1	1	0	0	0
	1	0	0	0	0
	1	1	1	1	0
Detection	8	7	5	3	0

Table 6. Minimum sample of the results after applying Dice algorithm on the spam database

A	B	C	D	E	F	G	H	I	J
0.545455	0.545455	0.363636	0.25	0.25	0.4	0.4	0.4	0.444444	0.6
0.363636	0.875	0.875	0.615385	0.615385	0.4	0.4	0.4	0.714286	0.8
0.25	0.615385	0.769231	0.769231	0.769231	0.4	0.4	0.4	0.857143	0.8
0.25	0.615385	0.769231	1	1	0.571429	0.571429	0.571429	0.727273	0.5
0.4	0.4	0.4	0.571429	0.571429	0.571429	0.571429	0.571429	0.727273	0.5
0.4	0.4	0.4	0.571429	0.571429	0.5	0.5	1	0.25	0.222222
0.4	0.4	0.4	0.571429	0.571429	1	0.5	0.5	0.5	0.222222
0.444444	0.714286	0.857143	0.727273	0.727273	0.25	0.5	0.25	0.25	0.222222
0.6	0.8	0.8	0.5	0.5	0.222222	0.222222	0.222222	0.769231	0.769231

Table 7. Maximum value of the features after applying Dice algorithm on the spam database

A	B	C	D	E	F	G	H	I	J
0.6	0.875	0.875	1	1	1	0.571429	1	0.857143	0.8

Table 8. Minimum value of the features after applying Dice algorithm on the spam database

A	B	C	D	E	F	G	H	I	J
0.6	0.875	0.875	1	1	1	0.571429	1	0.857143	0.8

Table 9. Average value of the features after applying Dice algorithm on the spam database

A	B	C	D	E	F	G	H	I	J
0.405948	0.596168	0.626027	0.619575	0.619575	0.479453	0.451675	0.479453	0.582183	0.5151

Table 10. Result of spam detection after applying Dice algorithm on the spam database

Similarity	50%	60%	70%	80%	90%
	1	1	0	0	0
	0	0	0	0	0
	0	0	0	0	0
	1	1	1	0	0
	1	0	0	0	0
	1	0	0	0	0
	1	0	0	0	0
	1	0	0	0	0
	1	1	1	1	0
Detection	7	3	2	1	0

Now, through the use of MATLAB, The output was as follows:

- Based on the results described on Table 11-16, it is clear that 60% is the best similarity rate, where it is in the case of Cosine was discovered 7 spams out of 10 spams
- Based on the results described on Table 17-21, it is clear that 50% is the best similarity rate, where it is in the case of Cosine was discovered 7 spams out of 10 spams

Table 11. Re-weight: Represents a mini sample of the result after applying Re-weight equation into the spam database

	A	B	C	D	E	F	G	H	I	J
1	0.000000	1.000000	0.831169	0	0.156863	0.000000	0.000000	0.000000	0.000000	0.000000
2	1.000000	0.437500	0.649351	0	0.068627	0.875000	0.552632	0.037234	0.000000	1.000000
3	0.285714	0.000000	0.922078	0	0.602941	0.593750	0.500000	0.063830	0.695652	0.265957
4	0.000000	0.000000	0.000000	0	0.308824	0.000000	0.815789	0.335106	0.336957	0.670213
5	0.000000	0.000000	0.000000	0	0.308824	0.000000	0.815789	0.335106	0.336957	0.670213
6	0.000000	0.000000	0.000000	0	0.906863	0.000000	0.000000	0.984043	0.000000	0.000000
7	0.000000	0.000000	0.000000	0	0.941176	0.000000	0.000000	0.000000	0.000000	0.680851
8	0.000000	0.000000	0.000000	0	0.921569	0.000000	0.000000	1.000000	0.000000	0.000000
9	0.714286	0.000000	0.597403	0	0.29902	0.000000	0.789474	0.000000	1.000000	0.808511
10	0.285714	0.187500	1.000000	0	0.093137	1.000000	1.000000	0.000000	0.065217	0.000000
11	0.000000	0.000000	0.000000	0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
12	0.000000	0.000000	0.000000	0	0.000000	0.000000	0.000000	0.000000	0.000000	0.904255
13	0.000000	0.000000	0.000000	0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
14	0.000000	0.000000	0.000000	0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
15	0.000000	0.000000	0.000000	0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
16	0.000000	0.000000	0.000000	0	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000
17	0.000000	0.000000	0.000000	0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
18	0.000000	0.000000	0.000000	0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
19	0.000000	0.000000	0.000000	0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
20	0.000000	0.000000	0.000000	0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

Table 12. Minimum sample of the results after applying Cosine algorithm on the Re-weight spam database

A	B	C	D	E	F	G	H	I	J
0.612372	0.612372	0.408248	0.258199	0.258199	0.408248	0.408248	0.408248	0.471405	0.654654
0.408248	0.875000	0.875	0.632456	0.632456	0.500000	0.500000	0.500000	0.721688	0.801784
0.258199	0.632456	0.790569	0.790569	0.790569	0.500000	0.500000	0.500000	0.866025	0.801784
0.258199	0.632456	0.790569	1.000000	1.000000	0.632456	0.632456	0.632456	0.730297	0.507093
0.408248	0.500000	0.500000	0.632456	0.632456	0.632456	0.632456	0.632456	0.730297	0.507093
0.408248	0.500000	0.500000	0.632456	0.632456	0.500000	0.500000	1.000000	0.288675	0.267261
0.408248	0.500000	0.500000	0.632456	0.632456	1.000000	0.500000	0.500000	0.577350	0.267261
0.471405	0.721688	0.866025	0.730297	0.730297	0.288675	0.57735	0.288675	0.288675	0.267261
0.654654	0.801784	0.801784	0.507093	0.507093	0.267261	0.267261	0.267261	0.771517	0.771517

Table 13. Maximum value of the features after applying Cosine algorithm on the Re-weight spam database

A	B	C	D	E	F	G	H	I	J
0.654654	0.875	0.875	1	1	1	0.632456	1	0.866025	0.801784

Table 14. Minimum value of the features after applying Cosine algorithm on the Re-weight spam database

A	B	C	D	E	F	G	H	I	J
0.258199	0.5	0.408248	0.258199	0.258199	0.267261	0.267261	0.267261	0.288675	0.267261

Table 15. Average value of the features after applying Cosine algorithm on the Re-weight spam database

A	B	C	D	E	F	G	H	I	J
0.43198	0.641751	0.670244	0.64622	0.64622	0.525455	0.501975	0.525455	0.605103	0.538412

Table 16. Result of spam detection after applying Cosine algorithm on the Re-weight spam database

Similarity	50%	60%	70%	80%	90%
	0	0	0	0	0
	1	1	1	1	0
	1	1	1	1	0
	1	1	1	0	0
	1	1	1	0	0
	1	1	0	0	0
	1	1	0	0	0
	1	0	0	0	0
	1	1	1	1	0
Detection	8	7	5	3	0

Table 17. Minimum sample of the results after applying Dice algorithm on the Re-weight spam database

A	B	C	D	E	F	G	H	I	J
0.545455	0.545455	0.363636	0.250000	0.250000	0.400000	0.400000	0.400000	0.444444	0.600000
0.363636	0.875000	0.875000	0.615385	0.615385	0.400000	0.400000	0.400000	0.714286	0.800000
0.250000	0.615385	0.769231	0.769231	0.769231	0.400000	0.400000	0.400000	0.857143	0.800000
0.250000	0.615385	0.769231	1.000000	1.000000	0.571429	0.571429	0.571429	0.727273	0.500000
0.400000	0.400000	0.400000	0.571429	0.571429	0.571429	0.571429	0.571429	0.727273	0.500000
0.400000	0.400000	0.400000	0.571429	0.571429	0.500000	0.500000	1.000000	0.250000	0.222222
0.400000	0.400000	0.400000	0.571429	0.571429	1.000000	0.500000	0.500000	0.500000	0.222222
0.444444	0.714286	0.857143	0.727273	0.727273	0.250000	0.500000	0.250000	0.250000	0.222222
0.600000	0.800000	0.800000	0.500000	0.500000	0.222222	0.222222	0.222222	0.769231	0.769231

Table 18. Maximum value of the features after applying Dice algorithm on the Re-weight spam database

A	B	C	D	E	F	G	H	I	J
0.6	0.875	0.875	1	1	1	0.571429	1	0.857143	0.8

Table 19. Minimum value of the features after applying Dice algorithm on the Re-weight spam database

A	B	C	D	E	F	G	H	I	J
0.25	0.4	0.363636	0.25	0.25	0.222222	0.222222	0.222222	0.25	0.222222

Table 20. Average value of the features after applying Dice algorithm on the Re-weight spam database

A	B	C	D	E	F	G	H	I	J
0.626027	0.619575	0.619575	0.479453	0.451675	0.479453	0.582183	0.5151	0.626027	0.619575

Table 21 Result of spam detection after applying Dice algorithm on the Re-weight spam database

Similarity	50%	60%	70%	80%	90%
	1	1	0	0	0
	0	0	0	0	0
	0	0	0	0	0
	1	1	1	0	0
	1	0	0	0	0
	1	0	0	0	0
	1	0	0	0	0
	1	0	0	0	0
	1	1	1	1	0
Detection	7	3	2	1	0

When comparing the Cosine and Dice after Re-weight it is clear that Cosine similarity (detect 7 spams with similarity 60%) is best than Dice similarity (detect 7 spams with similarity 50%).

Results and Discussion

After implementation the experiment on the mini sample and from the results, it is clear that Reweight operation does not have any positive impact on the detection process.

Phase (1)

Cosine and Dice Similarity

In this phase the experiment started by calculate the value of cosine and Dice, As shown in Table 22 and 23 Best ratio of similarity is at 60%, at this ratio Cosine method succeeded in identifying 1793 spam and Failure to identify 20 spam, while Dice method at the same ratio (60%) succeeded in identifying 1796 spam and Failure to identify 17 Spam. As noted earlier, the original data base containing 1813 spam.

Results

At this phase Experiment proved that Dice similarity is best than Cosine similarity, As shown in Table 23, Dice fail in identify 17 spam while Cosine Fail in identify 20 spam.

Phase (2)

Reweighting Spam Database

During this phase reweighting operation was apply and then applied Cosine and Dice methods in the new database.

Cosine and Dice Similarity after Reweighting

As shown in Table 24 and 25, Best ratio of similarity is at 60%, at this ratio Cosine method succeeded in identifying 1790 spam and Failure to

identify 23 Spam, while Dice method at the same ratio (60%) succeeded in identifying 1792 Spam and Failure to identify 21 Spam.

Results after Reweighting

As shown in Tables 22 to 25, Experiment proved that Dice similarity is best than Cosine similarity, Dice fail in identify 21 spam while Cosine Fail in identify 23 spam.

Results of the Experiment

From the results before and after applying reweighting clear to us the following:

- Dice similarity Give the best results in both cases (before and after reweighting)
- Reweighting process did not improve the success rate of any of the two methods (Cosine and Dice)

Figure 2. Shows the comparison between Cosine and Dice after applying features selection similarity, while Fig. 3. Shows the comparison between Cosine and Dice after applying Re-weight Process. The results of Reweighting shown in Fig. 4 and Fig. 6. clarify the comparison between Cosine and Dice before and after applying feature selection. Figure 7 shows the results obtained by (Abdelrahim *et al*., 2013) in their study after using similarity algorithms to detect spam.

Phase (3)

Reweighting VS Feature Selection

From the results obtained in Tables (22-25), it is clear that the process of re-weight did not have a positive impact in improving the spams detection Fig 5.

Results after Features Selection

The results from the study (Abdelrahim *et al*., 2013) Table 26 were as follows:

- The paper proved that features selection process had a positive effect in improving the accuracy of the spams detection
- Features selection process led to the improvement of detection in the Cosine, while it has a negative effect in the case of Dice

Based on the experiment and the results from the study (Abdelrahim *et al*., 2013), following results were Obtained Feature selection method is best than Reweighting:

- Feature selection method led to improve detection in Cosine, while reweighting method not improve detection any of (Cosine or Dice)

Table 22. Cosine and dice similarity

Similarity	50%	60%	70%	80%	90%
Cosine	1800	1793	1719	1155	239
Dice	1803	1796	1705	1130	238

Table 23. Results of cosine and dice similarity

Similarity method	Max similarity	Min similarity	Success	Fail
Cosine	1	0.19	1793	20
Dice	1	0.17	1796	17

Table 24. Cosine and dice after reweighting

Similarity	50%	60%	70%	80%	90%
Cosine	1797	1790	1716	1154	239
Dice	1799	1792	1701	1129	238

Table 25. Results of cosine and dice after reweighting

Similarity method	Max similarity	Min similarity	Success	Fail
Cosine	1	0.19	1790	23
Dice	1	0.17	1792	21

Table 26. Represents the results obtained by (Abdelrahim *et al.*, 2013) in their paper after applying spam detection algorithms in the same spam database

Similarity method	Max similarity	Min similarity	Success	Fail
Cosine	1	0.47	1812	1
Dice	1	0	1800	13

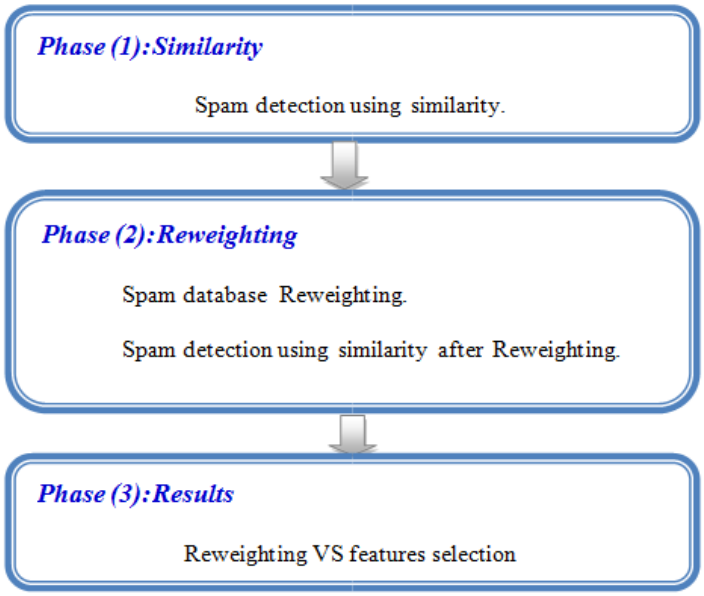

Fig. 1. Experiment phases

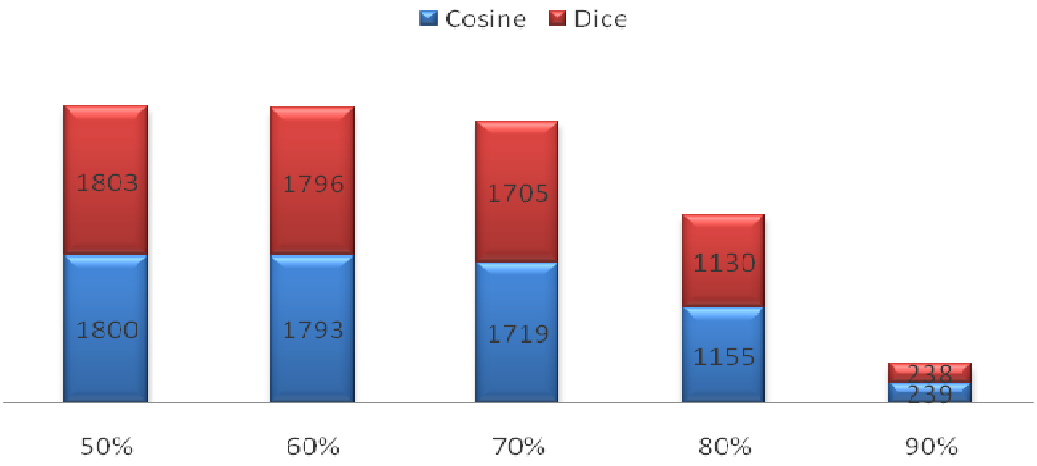

Fig. 2. Cosine and dice similarity

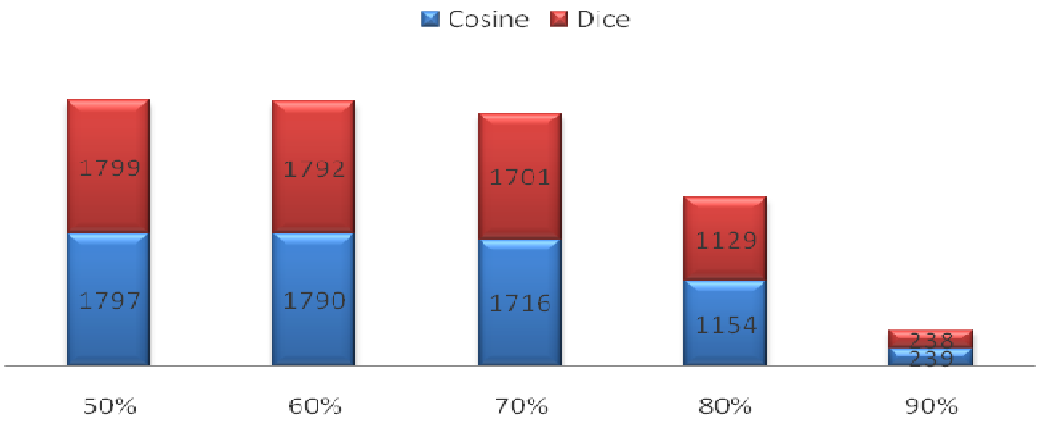

Fig. 3. Cosine and dice similarity after reweighting

DOI: 10.3844/ajassp.2017.983.993

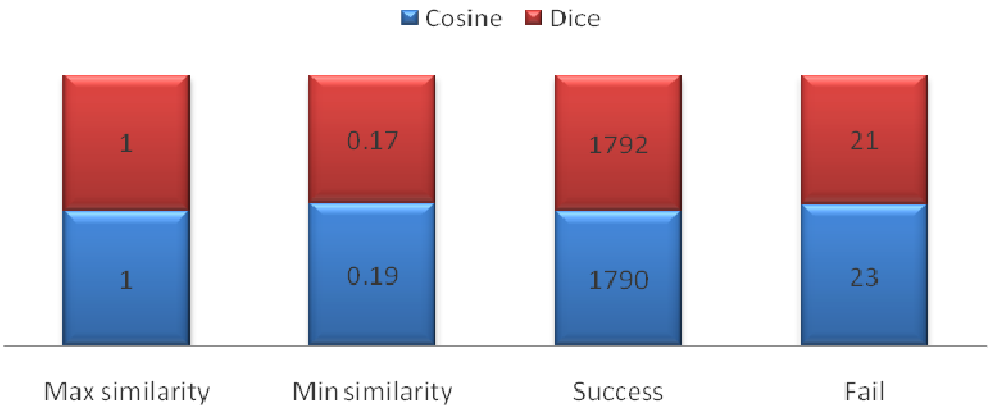

Fig. 4. Results of cosine and dice after reweighting

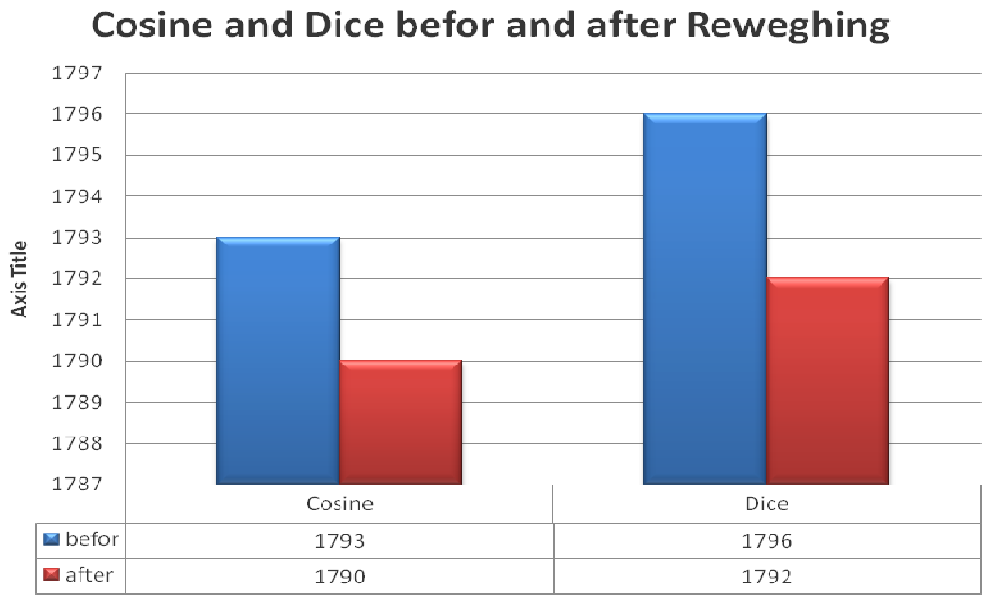

Fig. 5. Results of cosine and dice after reweighting

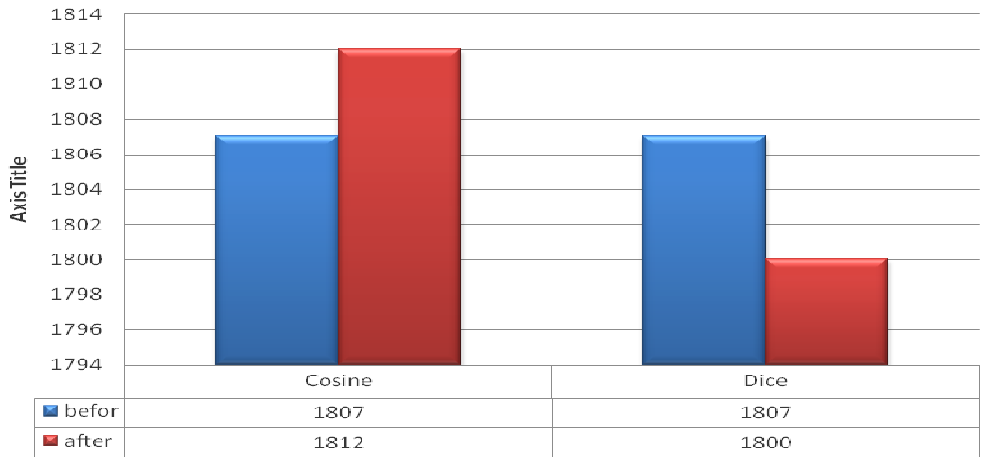

Fig. 6. Cosine and dice before and after feature selection

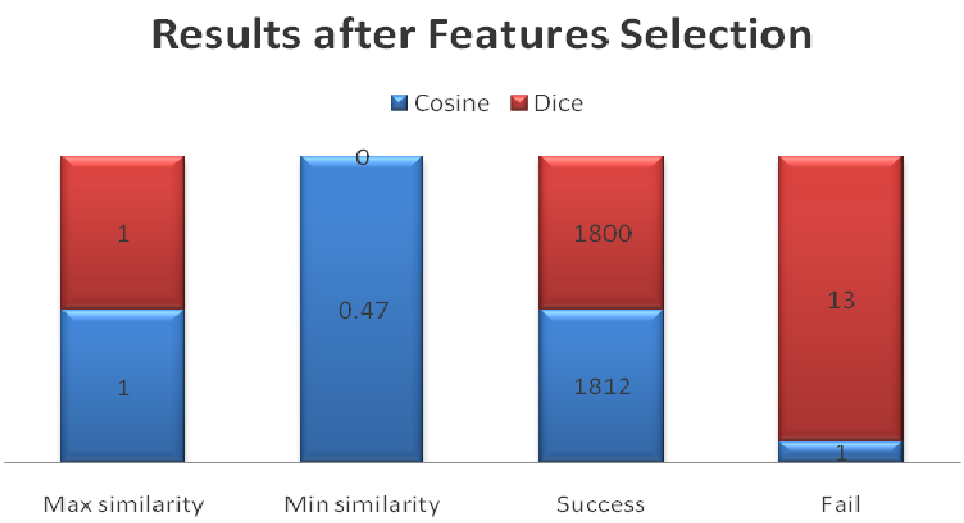

Fig. 7. Results of cosine and dice after feature selection

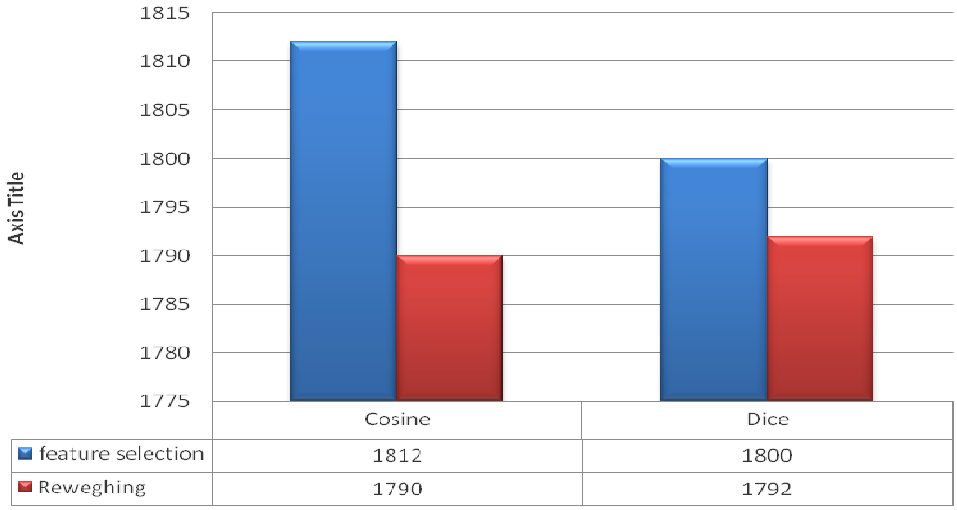

Fig. 8. Feature selection VS reweighting

Recommendations and Future Research

Based on the findings of this study, the following recommendations are offered for consideration

- Despite the findings in this research, we cannot say that the re-weight process does not have a positive impact in all cases, so it's recommended to looking for other ways to improve the process of re-weight. Because the logic indicates that there are some features have higher weight and its presence indicates a large margin that the message is spam
- Because this study did not get a positive effect after applying Cosine and Dice similarity, (Fig. 8) the study recommended re-applying this experience to other similarity coefficients

Conclusion

The result, which reached in this study is that the re-weight process did not have a positive impact on the Spam detection process, this result can be explained as follows: There are other algorithms for the process of re-weight can be tested in the future, also previous studies have shown that features selection had a positive impact in improving Spam detection, so can Combine the process of features selection with re-weight to get the best results.

Acknowledgement

The authors would like to thank our colleagues and staff at University of Science and Technology and Mashreq University for their contribution and comments.

Author's Contributions

Ahmed Osman Ali Elsiddig: Has given a significant contribution in the preparation of this article. Produced the initial draft of the article, develop and carry out this manuscript.

Ammar Ahmed E. Elhadi and Ali Ahmed: Are the supervisors, who oversaw the overall research article, reviewed and finalized the draft of the article before it is being submitted.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

Abdelrahim, A.A.A., A. Ahmed, E. Elhadi, H. Ibrahim and N. Elmisbah *et al*., 2013. Feature selection and similarity coefficient based method for email spam filtering. Proceedings of the International Conference on Computing, Electrical and Electronics Engineering, Aug. 26-28, IEEE Xplore Press, Khartoum, Sudan, pp: 26-28. DOI: 10.1109/ICCEEE.2013.6634013

Beiranvand, A., A. Osareh and B. Shadgar, 2012. Spam filtering by using a compound method of feature selection. J. Acad. Applied Studies, 2: 25-31.

GFI, 2011. Why Bayesian filtering is the most effective anti-spam technology. Software.

Hastie, T., R. Tibshirani and J.H. Friedman, 2001. The Elements of Statistical Learning. 1st End., Springer, Science and Business Media.

Lee, S.M., D.S. Kim, J.H. Kim and J.S. Park, 2010. Spam detection using feature selection and parameters optimization. Proceedings of the International Conference on Complex, Intelligent and Software Intensive Systems, Feb. 15-18, IEEE Xplore Press, Krakow, Poland, pp: 15-18. DOI: 10.1109/CISIS.2010.116

Michael, E.W. and H.J. Mattord, 2012. Principles of Information Security. 4th Ed, Cengage Learning, USA.

Spambase Dataset, http://ftp.ics.uci.edu/pub/machine-learning-databases/spambase/

Wikipedia, 2015a. Cosine Similarity.

Wikipedia, 2015b. Dice Coefficient.

Wikipedia, 2015c. Spamming. /Spamming