A Content Based Classification of Spam Mails with Fuzzy Word Ranking

G.Santhi[1], S. Maria Wenisch[2] and Dr. P. Sengutuvan[3]

[1]CRD,
Prist University,
Thanjavur, India

[2]Department of Information Science and Technology,
Anna University,
Chennai, India.

[3]VMKV Engineering College,
Salem, India.

Abstract

The spam mails are used by spammers to steal the data of the users and organizations online. Rapid growth rate of the use of the internet has increased the spam mails. There are several methods employed for filtering spam. This approach is to classify the spam mails using spam word ranking and fuzzy rules. This work classifies the emails using word ranking database and the ranks are used based on the degree of the threat that each word possess. For this purpose the work has considered only the content of the email.

Keywords: *Email, Spam Word Ranking, Spam Classification, Fuzzy Rule, Fuzzy Inference.*

1. Introduction

Email Communication is indispensable in the present days. Spam mails are unsolicited junk mails sent by the spammers which adversely affect the email communication process. Unsolicited commercial mails are often sent by the spammer to illegally promote a service or product. Spam became an issue when the internet was opened to the public. Internet users are forced to receive spam mails in their inbox. Spammers harvest the address of internet users from various sources and serious cause inconvenience to the users. The user's inbox are flooded with enormous spam mails. U.S.A is leading spam relaying country with 18.3%. Table.1 reveals the spam rate of top twelve countries. International telecommunication union report reveals that the internet will be used by 39% of the world population by 2013.

Countries	*Percentage*
U.S.A	18.3%
China	8.2%
India	4.2%
Peru	4.0%
France	3.4%
South Korea	3.4%
Italy	3.4%
Taiwan	2.9%
Russia	2.9%
Spain	2.8%
Germany	2.7%
Iran	2.6%
Other	41.1%

Table 1: Spam Rate
(December 2012 to Feb. 2013)
(Courtesy Sophos Lab)

Spammers collect the user's personal bank details and also hang the user's systems by spreading virus. Spam mails consume the bandwidth of network, wastes memory, time of user and causes financial loss to the users and the organizations. Every internet user should spend few seconds to read and delete spam mails in their inbox. Spam mails are treated as illegitimate or black listed mails and ham as legitimate or white listed mails. The targets of the spammers are to send bulk mails to the internet users in order to receive response from few with a view to secure profit for them.

Many spam mails are being sent to undisclosed recipient address. Spammers are paying money to collect the users address. Spammers are used to send attachments with virus. The images sent by the spammers have an embedded transparent object called web bug. It tracks and gathers details about when and where each particular recipient reads email and IP address of the computer. It is difficult for the user to detect the object embedded in the image. The text of the message is stored in the image and saved as gif or jpeg image and displayed in the email. There are two ways available for filtering the illicit mails either by filtering the spam messages at inbound level or outbound level.

In rule based filters, the content based filters are used in the user inbox level to filter the spam. Like Naive Bayesian classification, Support Vector Machines for text categorization, K-NNC for classifying nearest neighbor test pattern, and other clustering methods are used for spam filtering. Spammers are increasingly employing innovative methods to send their spam mails. Some organizations and many researchers have tried to filter spam mails by applying various methods at different levels. Spam filtering is necessary to protect the internet users which is quite challenging. Spam filter is a program or software used to filter spam mails.

There are several algorithms available for filtering spam. But spammers try to break the anti spam filters by applying obfuscation and various Techniques. Christina et al (2010) proposed a study on email spam filtering techniques. They discussed about various problems aroused by spam, different filtering methods and techniques are used to filter spam. Hailong Hou et al (2008) developed a method of hyperbolic tree based algorithm for filtering spam not by matching words but matching based on the influencing factors. Gregory L. et al (2012) described about the attacking techniques used by the spammers, the challenges faced by the developers of anti-spam methods and spammers. They suggested that anti-spam developers should not only concentrate in filtering of spam but also should consider the costs associated with spam filtering.

In this work fuzzy logic is applied to classify spam. This work used a fuzzy inference system to classify spam words in an email. Words are extracted from the content of the emails which, this work compares them against a list of spam words stored in the database ranked with its values and categorizes the words in accordance to the ranking. Fuzzy inference system finally takes the input value from the above ranking and classifies the output as least dangerous or moderate or most dangerous spam mail.

The rest of this paper is organized as follows: Section 2 explains the related work. Section 3 describes the proposed work. Section 4 discusses the expected result. Finally Section 5 concludes the Paper.

2. Related Works

MD. Rafiqul Islam et al (2005) discussed about different machine learning algorithms for spam filtering and presented a comparative study of spam filters. Their research includes a study of automated filtering and machine learning techniques like rule based, content based, personalized, collaborative, support vector machine and kernel based algorithms for filtering spam. They presented a comparative analysis on different filtering techniques and its advantages.

Ni Zhang et al (2006) developed a method for filtering spam mails from the Internet service providers in its heavy traffic. Finger print method is used to detect the similar earlier mails and sets a parameter for the email category. Mail database and finger print database are used to store information. By simply adding the entry in the MD and delete the unimportant mails. They explained about the three advantages of BMTC. They are automatic hand-free deployment and online update mechanism, high accuracy in identifying emails, and handling a large amount of data with small memory and reasonable CPU time.

Seongwook Youn et al (2007) proposed a comparative study for email classification. Neural Network, SVM, Naive Bayesian and J48 classifiers are used to filter spam from the datasets of emails. J48 is a decision tree creates a binary tree used for classification of legitimate and spam. They suggested J48 and NB classifiers obtained a better result and accuracy than SVM and NN classifiers.

Enrico Blanzieri et al (2008) proposed a survey on learning based techniques of spam filtering. This paper discussed about the learning based methods of spam filtering like keyword filtering, image based filtering, language based filtering, filters based on non-content features, collaborative filtering and hybrid approaches. They presented the evaluation and comparison of the results obtained from the various filtering methods.

Ali Cıltık et al (2008) proposed a method of spam email filtering methods with high accuracies and low time complexities. They took Turkish mails for their research. They used PC-KIMMO system, a morphological analyzer to extract root forms of words as input and produce parse of words as output. This method is based on the n-gram approach and a heuristics. They developed two models, a class general model and an e-mail specific model. The general model classifies the mail as spam or legitimate by using bayes rule. The second model determines the correct class of a message by comparing it with the similar previous message for matching. The third model is a combined perception refined model. It is a combination of above two models. Free word order is used for ordering the word in fixed order for n gram model. This spam filtering method is based on classifying text contents and raw contents of emails obtaining results from the categorization of data sets. They faced the increase of time complexity problem when handling the larger number of words. Adaboost ensemble algorithm is used to compare with its previous work. They performed extensive tests on various number datasets sizes and initial words. They have obtained a result of high success rates in both Turkish language and English.

A.G. López-Herrera et al (2008) developed a multiobjective evolutionary algorithm for filtering spam. They evaluated the concepts of dominance and pareto–set. SPAM-NSGA-II-GP is used for filtering spam mails. MOEA is used to learn a set of queries with good precision and recall. PUI datasets are used for spam filtering. SPAM-NSGA-II-GP with very strong filtering rules are (high recall and low precision) used to block all the legitimate emails and labeled as spam. They used the weak filtering rules (high precision and low recall) for labeling a minimum portion of spam emails.

Liu Pei-yu et al (2009) suggested the method of improved bayesian algorithm for filtering spam. KNN algorithm, SVM, decision tree, and improved bayesian algorithm are used for classifying texts. KNN algorithm is a simple and accurate method for spam filtering by using the k nearest neighbor. SVM is also used for filtering spam and finds hyper plane to classify the legitimate and spam mails. It works with smaller training set. Decision tree is used for faster and simple classification which gives higher accuracy of judgment. Bayesian algorithm is a base and simple classification method classifies the mail as C_{legal} and spam C_{rubbish}. In the bayesian method one feature is treated as independent of other. Improved naive bayesian algorithm is a combination of bayesian algorithm with boosting method, developed to reduce the rate of misjudgment and improve the accuracy of classification. Boosting is a universal learning algorithm. They treated the naive bayesian algorithm as weaker learning algorithm and made it stronger by boosting it with boosting algorithm. They obtained better result by applying this boosted naive bayesian algorithm for filtering spam.

Alaa El-Halees et al (2009) developed to filter spam messages in mixed Arabic and English. Six classifiers are used for filtering spam messages and compared the results obtained from these classifiers. Maximum entropy, decision trees, artificial neural sets, naive bayes, support system machines and k-nearest neighbor are used for spam filtering. Recall and precision are the two ways for presenting the system performance. SVM is used as the best classifier for English and ME performed better than NB in Arabic messages. They suggested increasing the parameter will improve the performance.

Jan Gobel et al (2009) proposed a method to filter spam mail in a proactive way which intercepts the communication held between spambot and the intended server and redirects its communication with local mail server at the gateway. They collect spam messages at the gateway and obtain the current spam messages sent by spambotnet. They clean the machine system using a software based restoring system to execute next spambot. They have collected the spam messages by resetting the honeypot. The next process is filtering the message in a proactive way. Longest common string algorithm is used for the extraction of emails and single raw template. They generate templates taking subject, xmailer and complete body of the message for consideration. Longer emails are processed first as they took the first email from the list sorted based on text length and considered it as first raw template named α. They took second email from the list and merge it with α to form a second raw template named β which was more specific than previous one. Then they compared both α and β and determine the amount of text that was replaced by the placeholders. If the removed text percentage is below a predefined threshold Ø then they are treated β as their new α. The email used to form β was removed from the list and they continued with third step. If the changed text percentage is above Ø the current β is too generic and is therefore discarded. They used template generation process for detecting the spam rate.

M. Basavaraju et al (2010) proposed the text based clustering method for spam detection. Pre processing of data, methodology of classification, vector space model, and data reduction are the methodologies used for spam filtering. The Porters stemming and stopping algorithm are used for preprocessing of data. Hierarchical and partition clustering algorithms are used for partitioning and clustering. They used BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) method to clustering the documents. NNC and K-NNC are the two classifiers used to classify the neighbors. K-NNC classifier is used to classify the patterns. The vector space model is used to calculate the inverse document frequency of each word i.e. tf-idf test patterns. After clustering of training patterns the non-spam data are stored in the centroids. Test patterns and centroids are passed in to the classification module for spam and non-spam detection.

Alireza Nemaney Pour et al (2012) proposed for Minimizing the time of spam detection by relocating the filter to the sender messages. They used DSPAM and TREC anti-spam software for filtering spam. They used four steps to detect spam. The first two steps are used to check the IP validity of both the sender and receiver. The Sender IP validity is checked by the mail server and receiver IP validity is checked by the DNS server respectively. The third step checks the category of mail belongs to white list or black list or grey list. In the Final step they applied rule and content based filters for detection of spam. This research helps them to preserve the network resources such as bandwidth, time and memory and also minimize the time.

Dhananjay et al (2012) developed an adaptive neural fuzzy inference system classifier which includes both the neural networking and fuzzy logic concept to detect the spam message on social networking websites. ANFIS classifier is used to identify the spam from input vector. They identified five input vector parameters. Number of associated user pages, number of times marked as spam, text priority, presence of URL or Hyperlink and the number of common timestamps are the parameters used to classify the spam. They developed the fuzzy inference system with three parts. They are input member function, output member function and the rule set linking the two member functions. Input member function has five parameters like the number of associated user pages, number of times marked as spam by user, presence of Hyperlink or URL, the number of instances of common timestamps and the priority of text in the message. Seven fuzzy rules are used and the output member functions produce a result which equals the rules in the fuzzy rule set. The researcher suggested increasing the parameters would decrease the false positives and improve the detection rate of spam. This system is not designed for a particular specialized social networking website.

Sudhakar. P (2011) et al developed fuzzy logic concept for spam detection. They applied five fuzzy rules on five fuzzy parameters. The 5 fuzzy parameters are sender address, sender IP, subject words, content words, and attachments. All the five parameters are compared against the black list and white list. If match was found they considered the parameter as spam or ham. This approach consumes large amount of time to identify spam words.

Subhodini gupta (2012) et al suggested a fuzzy filtration module for spam detection. They developed two modules. The first module applied stemming, stop-word elimination and tokenization process on the extracted email words. Fuzzy rules are applied on the document set to verify it for spam or ham. In this method five fuzzy parameters are used. They are sender address, sender IP, subject words, content words and attachments. These extracted parameters are passed through the fuzzy rules for detecting spam. This method applied only for plain text used in subject and body content.

Dr. Sonia et al (2010) developed a vector space model to classify the mail. It converts the mail into matrix and inverse frequency is calculated. They calculated the similarity coefficient by using term frequency and inverse message frequency. The fuzzy decision maker used to take the sc as input for fuzzification. Fuzzification classifies the input as legitimate or spam mail.

Jitendra Nath Shrivastava et al (2012) discussed about the trends, issues and challenges concerning the spam. They present the role of botnets in spreading spam and predicts the statistical figures of spam mails from 2011 - 2015. It covers the area of the anti spam filtering methods, its classification and consequences on stopping the spam mails.

M. Muztaba Fuad al (2004) proposed a method of trainable fuzzy filters for filtering spam mails automatically. A trainable fuzzy classification module consists of a set of fuzzy rules and fuzzy inference system used for the classification of spam.   The messages from the corpus parsed and features are extracted. It extracts the features from, to, cc, subject, and header fields. In the fuzzification five fuzzy sets are used in which two sets are used for feature extraction from the header part and others for the body features. In the fuzzification it determines the degree of input that belongs to which fuzzy set. Then the rule antecedents are evaluated by fuzzy AND operation and the consequences are combined by OR operation and passed the output for defuzzification process. It will produce a crisp output and it is compared with threshold value and predicts the output value as a spam or ham. They suggested that this method can eliminate a large amount of spam from inbox of the user.

Mehdi Samiei yeganeh et al (2012) developed a model for fuzzy logic based machine learning approach for filtering spam. They discussed the methods of automatic spam filters like naive bayes classifier, artificial immune classifier and fuzzy logic. They have enhanced the functionality of the model and also enhanced the feature identification of emails and deletion of spam mails on its own. They suggested that the fuzzy logic is adaptable for spammer tactics.

Begol et al (2011) proposed a fuzzy system method to detect the edges of image. Two types of pixel vicinities used for edge detection. They are four and eight pixel vicinities. In an image grey pixels are treated as noise. The noises are omitted from the image. Canny and sobel method did not detect the edges properly. But fuzzy technique is used to detect the original image edges correctly and eliminates the noises in the image.

3. Proposed Frame work

The paper's related work consist of methods like naive bayesian, K-NNC, SVM, J48, and fuzzy logic filtering methods to detect the emails as spam or ham. The proposed work relies on content based classification which includes fuzzy inference system with fuzzy rules for classifying spam mails. This work focused on classifying the spam words by applying fuzzy rule. The spam words are assigned different values by using fuzzy rules. Collection of spam words list from the paper's related works, spam words list available in the website and spam mails in the inbox. This value helps us to rank the spam words as shown in fig.1 and finally fuzzy inference system classifies the input rank values and produces the output.

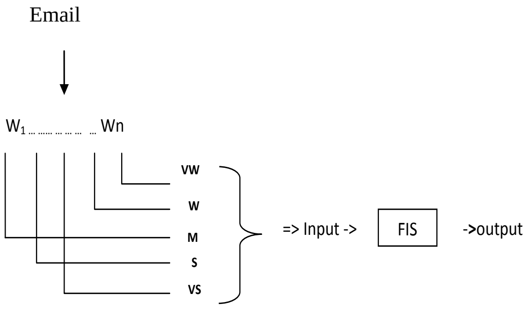

Fig. 1 Concept of classification of spam words

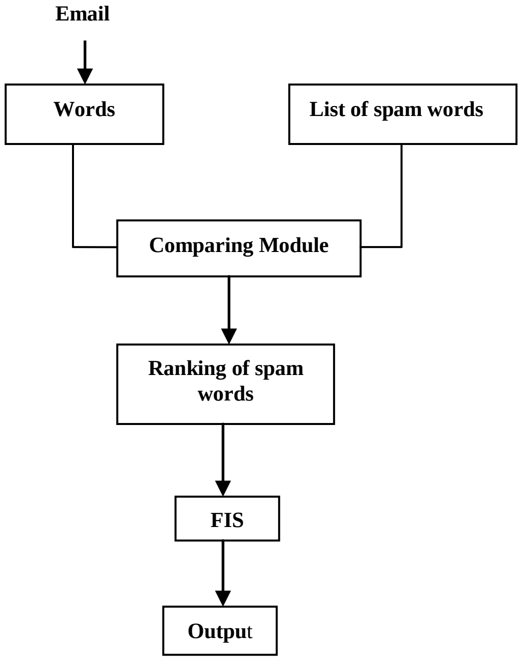

Fig. 2 Spam classification model

3.1 Fuzzy Classification Module

The fuzzy classification contains fuzzy inference system and a set of rules. The proposed fuzzy rule based classification shown in fig.3. This work has a list spam words in the database with its ranked value. The spam words are extracted from the content of email. The spam words are assigned a value and categorized in to five linguistic variables i.e. weak (W), very weak (VW), moderate (M), strong (S), and very strong (VS). Email contains many spam words. This work extracted spam words from 100 mails. Winner, dollar, award, cash prize, top job opportunities, earn more, beneficiary, good news, claim, high salary and payment are the few most attracted words used by the spammers to cheat the users. The internet user gets attracted by these words in the mail and contacts the sender immediately and gets deceived. So these words are treated as very strong spam words. The list of spam words in a database is available. The database contains ID, spam words and rank value fields. The actual words are extracted from the user inbox and compared against the spam words list in the database. If the actual word is equal to the spam word then corresponding rank value and the ID is assigned to the actual word. In order to classify the spam fuzzy inference system has been designed to take the ranked input value and produce the output. Fig.2. explains the classification of spam procedure. The output data are classified in to three linguistic variables i.e. Least dangerous, Moderate dangerous and Most dangerous.

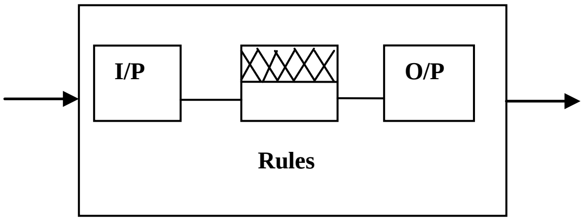

Fig. 3 Fuzzy rule based classification.

Step1: It is used to read the email from the inbox of user
Step2: The words are extracted from content of email
Step3: It is used to count the number of spam words
Step4: The words are compared against a list of ranked spam words in the database.
Step5: The spam words are classified according to the rank
Step6: Input the rank values to the input and produce the result.

The Mamdani fuzzy inference model is followed in the implementation. This model's simplicity helps anyone to understand the concept easily. Fig.4 shows the process of ranking and classification of spam words.

```
Ranking of Words
If spam word <=0.9 And > =0.7
It is a Very Strong spam word
Elseif spam word < 0.7 And >=0.5
It is a Strong spam word
Elseif spam word < 0.5 And >=0.4
It is a Moderate spam word
Elseif spam word <0.4 And >=0.2
It is a Weak spam word
Else Very Weak spam word
```

```
Classification of Spam emails
If spam words <=0.9And >=0.5
It is most dangerous spam mail
If spam words <0.5And >=0.4
It is moderate dangerous spam mail
If spam words <0.4And >=0
It is least dangerous spam mail
```

4. Results and discussion

This work used 100 emails for experiment. This method extracted the words and applied fuzzy inference system with fuzzy rules for classification of spam mails. It is a novel approach used for classification of spam. Table.2 presents the ranked spam words with its value shown in the appendix.

This approach helps the end-users to identify the spam mails by using the linguistic terms i.e., least dangerous, moderate dangerous, and most dangerous. The user can easily distinguish the spam mails and delete the spam mails in the inbox level.

5. Conclusion and Future work

This presented paper is a content based classification of spam mails with fuzzy word ranking. There are many classifiers and filters available for classifying and filtering spam mails. This study analyzed the previous related works. The proposed work used two sets of linguistic terms for ranking and classifying spam mails. This method has extracted only the features from the content of an email instead of extracting all the features from the mail.

The actual words are extracted from the inbox of an email are compared with a list of spam words in the database and the words are categorized according to its rank value. This input value is passed to the fuzzy inference system. FIS classifies the spam and produces the output. This work obtains a better result from ranking and classifying of spam words. The future work aims at classification of spam words in the subject and html also.

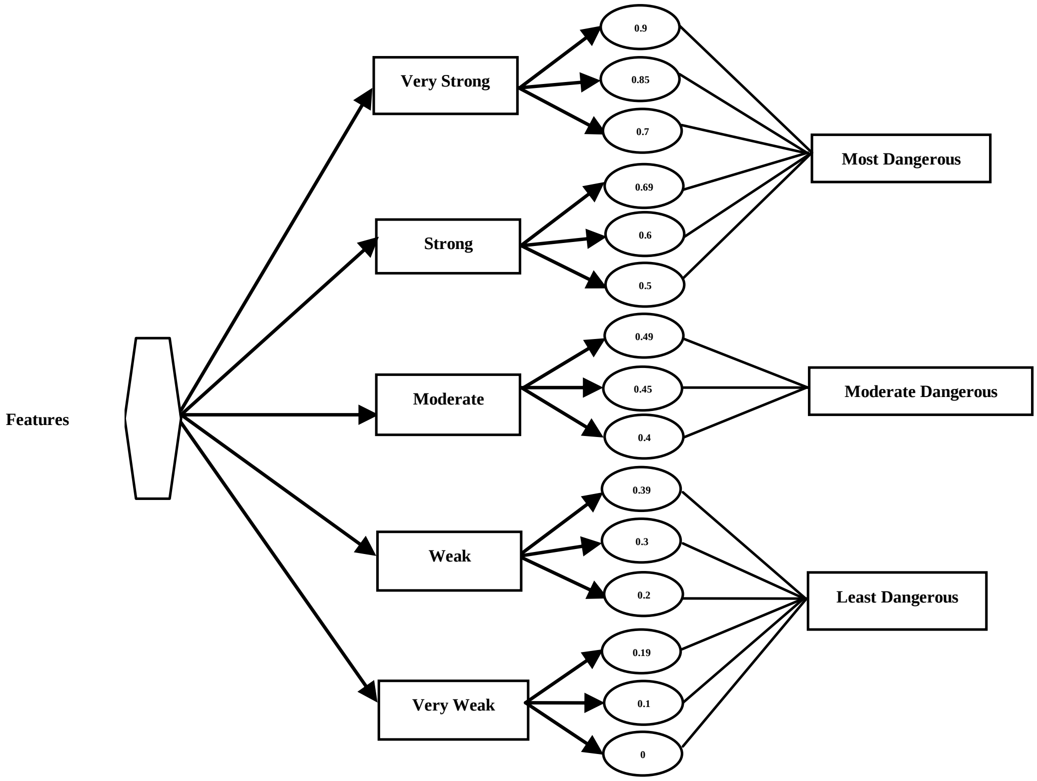

Fig.4 Ranking and classification of spam

Appendix

Very strong (0.7 -0.9)

congratulations	cash prize	unclaimed funds	Urgent loan
lottery	donated	Attention	registration
details	confidential	Click here	information
$Dollars	selected	Amount	attachment
prize	lotto	claim your prize	Email ID
High salary	receive	winning	Bank name
pounds	grant	hearing from you	Urgent response
awarded	funds	Re-Confirm their details	crores
Dear Winner	notification	draw	Apply now
60sec Approval (loan)	immigration	account number	Annual promotions
View our mail immediately	Remember	announce	claim
Luckily winner	won	$1500	Immediately
Euro	money	Beneficiary	earn more
positive response	information	reply	Notice
loan	transfer	attach	Lakhs

Strong (0.7 -0.5)

Payment	directly	Bank branch	partnership
compensation	online service	immediate employment	Mobile number
identification	offer loans	online pharmacy	Undelivered.
Secret pin code	Batch number	requesting	Investment
free installation	Ref Number	Abandoned Money	greetings
successfully	Provide	Offer for few days	property
visit our website	Unpaid contract funds	business proposal	approved
for more details	File	transaction	Assistance
quick deposit	Sum	Kindly contact me	Promo
check your account	top job opportunities	offer	occupation
Needed	open	mail	info
contact	Further instructions	national	support

Moderate (0.5 – 0.4)

part time jobs	Get	BBC	Form
Home jobs	Trust worthy	mobile phone	Fill
job opportunity	online jobs	lotto information	waiting
confirm	Good news	e-approval	interested
online survey	Vacancies	Skill requirement	For further clarification
access great jobs	Job description	Consultancy	Invite
post your resume	Send your cv	Experienced	Email

Moderate (0.5 – 0.4)

interview	please	Complement	Pay cash
job location	Good news	Kindly	Recruitment
Now	More	waiting	Office Address
receive	Looking for	Read	Mobile phone
Kindly find the attachment	Send	Grab	security deposit
Office Tel	Enable us	Acknowledge	resume
New openings	Job Site	Feel free to contact	online degree

Weak (0.4 – 0.2)

Best sellers	initial amount	seek	Lack of fund
has	collaborate	kindly	Free IPod
resident	suggestion	regards	Correctly
addresses	reasonable	Brand	Have been
Dear sir	government	magazine	madam
country	Your mail	Family	purchase
short listed	staff	company	recommend
Your	pleased	Age	Free tickets
Phone no	University diplomas	Free cell phone	valid Id proof
name	Online education	country	free installation

Very Weak (0.2 – 0)

real estate	jewel	hello	Weight loss
hand bag	free	your	Sale
watch	only $19.99	because	No experience
sun classes	Automobiles	like	Good health
Cosmetics	Best price	dealers	Hair loss
Hand bags	sex	You	randomly
thank you	face to face	Package	Free
Free trial	meeting	electronic	Permit
Shoes	permitted	Kind Regards	visit
buying	Free shipping	gift	product

Table.2 Ranked words

X→ spam word

0<= X<0.2 → Very weak spam word

0.2<=X<0.4 →Weak spam word

0.4<=X<0.5 → Moderate spam word

0.5<=X<0.7 → Strong spam word

0.7<=X<=0.9 → Very strong spam word

Email-1		
Actual Words	***Ranks***	ID
urgent response	0.9	1
open	0.5	2
the	not spam	
attachment	0.8	1
in	not spam	
your	0.2	4
mail	0.5	2
and	not spam	
fill	0.4	3
form	0.4	3
correctly	0.3	4
and	not spam	
contact	0.6	2
immediately	0.9	1

Email-2		
Actual Words	***Ranks***	ID
please	0.4	3
open	0.5	2
the	not spam	
attach	0.7	1
file	0.5	2
fill	0.4	3
the	not spam	
form	0.4	3

Email-3		
Actual Words	***Ranks***	ID
winning	0.8	1
award	0.7	1
read	0.4	3
attach	0.7	1

Email-4		
Actual Words	***Ranks***	ID
cash	0.9	1
loan	0.7	1
$1,500	0.8	1
click here	0.9	1
get	not spam	
your	0.2	4
money	0.9	1
now	0.45	3

Email-5		
Actual Words	***Ranks***	ID
hello	0.1	5
your	0.2	4
partnership	0.5	2
needed	0.6	2
reply	0.7	1
for	not spam	
more	0.43	3
info	0.5	2

Email-6		
Actual Words	***Ranks***	ID
your	0.2	4
email ID	0.7	1
has	0.2	4
won	0.9	1
750000	not spam	
gbp	not spam	
pounds	0.9	1
in	not spam	
BBC	0.4	3
one	not spam	
draw	0.8	1

Email-6		
Actual Words	***Ranks***	ID
send	0.45	3
name	0.35	4
address	0.3	4
emails	0.4	3
phone no	0.38	4
country	0.36	4
for	not spam	
more	0.43	3
information	0.7	1
contact	0.6	2
Mr.	not spam	

Email-7		
Actual words	***Rank***	ID
I	not spam	
need	0.6	2
your	0.2	4
support	0.5	2
for	not spam	
the	not spam	
transfer	0.8	1
of	not spam	
money	0.9	1
from	not spam	
my	not spam	
country	0.36	4
to	not spam	
your	0.2	4
country	0.36	4
you	0.1	5
will	not spam	
get	0.4	3
more	0.43	3
info	0.5	2
when	not spam	

Email-7

Actual Words	*Ranks*	ID
I	not spam	
receive	0.7	1
reply	0.7	1
from	not spam	
you	0.1	5

Email-8

Actual words	*Rank*	ID
your	0.2	4
email ID	0.7	2
has	0.2	4
successfully	0.6	2
selected	0.7	1
in	not spam	
ongoing	not spam	
2013	not spam	
draw	0.8	1
and	not spam	
you	0.1	5
have been	0.3	4
awarded	0.8	1
a	not spam	
whooping	not spam	
sum	0.6	2
of	not spam	
four	not spam	
crores	0.9	1
eighty	not spam	
lakhs	0.8	1
in	not spam	
punjab	not spam	
lottery	0.9	1

Email-8

Actual Words	*Rank*	ID
draw	0.8	1
please	0.4	3
provide	0.5	2
name	0.3	4
address	0.3	4
mobile number	0.6	2
age	0.2	4
sex	0.1	5
email	0.4	3

Email-9

Actual Words	*Rank*	ID
I	not spam	
need	0.6	2
your	0.2	4
support	0.5	2
for	not spam	
the	not spam	
transfer	0.8	1
of	not spam	
money	0.9	1
from	not spam	
my	not spam	
country	0.36	4
to	not spam	
your	0.2	4
country	0.36	4
you	0.1	5
will	not spam	
get	0.4	3
more	0.43	3
info	0.5	2
when	not spam	

Email-9

Actual Words	*Rank*	ID
I	not spam	
receive	0.7	1
reply	0.7	1
from	not spam	
you	0.1	5

Email-10

Actual words	*Rank*	ID
your	0.2	4
email	0.4	3
address	0.3	4
won	0.9	1
70000000gbp	not spam	
and	not spam	
blackberry	not spam	
mobile phone	0.4	3
in	not spam	
blackberry	not spam	
promo	0.5	2
2013	not spam	
to	not spam	
receive	0.7	1
award	0.7	1
send	0.45	3
your	0.2	4
name	0.35	4
age	0.2	4
occupation	0.6	2
address	0.3	4
mobile number	0.6	2
to	not spam	
email	0.4	3

References

[1] MD.Rafiqul Islam and Morshed U. Chowdhury, “Spam Filtering Using ML Algorithms”, IADIS International Conference on WWW/Internet 2005, pp. 419-426.

[2] Ni Zhang, Yu Jiang, Binxing Fang, Xueqi Cheng and Li Guo, “Traffic Classification-Based Spam Filter”, IEEE International Conference on Communications, 2006, Vol.5, pp. 2130 – 2135.

[3] Youn, Seongwook, and Dennis McLeod, “A Comparative Study for Email Classification”, Editor. Khaled Elleithy, Advances and Innovations in Systems, Computing Sciences and Software Engineering, Print ISBN 978-1-4020-6264-6 pp.387-391, 2007.

[4] Enrico Blanzieri and Anton Bryl, “A Survey of Learning Based Techniques of Email Spam Filtering”, Artificial Intelligence Review, Vol.29,No.1, 2008, pp.63- 92.

[5] Çıltık, Ali, and Tunga Güngör, “Time-Efficient Spam E-mail Filtering Using n-gram Models”, Pattern Recognition Letters, Volume 29, No.1, 2008, pp.19–33.

[6] López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F., “A Multiobjective Evolutionary Algorithm for Spam E-Mail Filtering”, Proceedings of 2008 3rd International Conference on Intelligent System and Knowledge Engineering, 2008, Vol.1, pp. 366-371.

[7] Liu Pei-yu, Zhang Li-wei and Zhu Zhen-fang, “Research on Email Filtering Based on Improved Bayesian”, Journal of Computers, Vol. 4, No. 3, March 2009, pp. 271-275.

[8] El-Halees, Alaa, “Filtering Spam E-Mail from Mixed Arabic and English Messages: A Comparison of Machine Learning Techniques”, The International Arab Journal of Information Technology Vol. 6, No. 1, 2009, pp.52-59.

[9] Göbel, Jan, Thorsten Holz, and Philipp Trinius, “Towards Proactive Spam Filtering”, Editors. Ulrich Flegel and Danilo Bruschi, in Detection of Intrusions and Malware and Vulnerability Assessment, 6th International Conference, DIMVA, proceedings, Italy, Print ISBN 978- 3-642-02917-2, pp.38-47, 2009.

[10] M. Basavaraju and Dr. R. Prabhakar, “A Novel Method of Spam Mail Detection using Text Based Clustering Approach”, International Journal of Computer Applications, 2010, Vol. 5, No.4, pp.15-25.

[11] Alireza Nemaney Pour, Raheleh Kholghi and Soheil Roudsar, “Minimizing the Time of Spam Mail Detection by Relocating Filtering System to the Sender Mail Server”, International Journal of Network Security & Its Applications, 2012, Vol.4, No.2, pp.53-62.

[12] Dhananjay Kalbande, Harsh Panchal, Nisha Swaminathan and Preeti Ramaraj, “ANFIS based Spam Filtering Model for Social Networking Websites”, IJCA, 2012, Vol. 44, No.11, pp. 32-36.

[13] Sudhakar, P., G. Poonkuzhali, K. Thiagarajan, R. Kripa Keshav, and K. Sarukesi, “Fuzzy Logic for E-mail Spam Deduction”, In Proceedings of the 10th WSEAS International Conference on Applied Computer and Applied Computational Science, 2011, pp. 83-88.

[14] Subhodini gupta, Parekh .B.S and Jaimine N.Undavia, “A Fuzzy Approach for Spam Mail Detection Integrated with Wordnet Hypernyms Key term Extraction”, IJERT, 2012, Vol. 1, No.5, pp.1-5.

[15] Dr. Sonia, “Spam Filter: VSM based Intelligent Fuzzy Decision Maker”, International Journal of Computer Science and Technology, 2010, Vol.1, No.1, pp.48-52.

[16] Jitendra Nath Shrivastava and Maringanti Hima Bindu, Trends, Issues and Challenges Concerning Spam Mails”, I.J. Information Technology and Computer Science, 2012, Vol.4, No.8, pp.10-21.

[17] Muztaba Fuad, Debzani Deb and M. Shahriar Hossain, “A Trainable Fuzzy Spam Detection System”, In Proc. of the 7th International Conference on Computer and Information Technology, 2004.

[18] Mehdi Samiei yeganeh, Li Bin and G. Praveen Babu, “A Model for Fuzzy Logic Based Machine Learning Approach for Spam Filtering”, IOSR Journal of Computer Engineering 2012, ISSN: 2278-0661 Vol.4, No.5, pp. 07-10.

[19] Begol, Moslem, Maghooli and Keivan, “Improving Digital Image Edge Detection by Fuzzy Systems”, World Academy of Science, Engineering and Technology, 57, 2011, pp.76-79.