JAUES

Journal of Al Azhar University Engineering Sector

Vol. 12, No. 45, October, 2017, 1242-1255

MAIL SPAM DETECTION USING STACKING CLASSIFICATION

Mohamed Abd El-Kareem, Ayman Elshenawy and Fawzi Elrfaey
Systems and Computers Department, Faculty of Eng., Alazhar University, Cairo, Egypt.

ABSTRACT

Spam mails are very fast growing and costly problem, that becomes a big trouble now-a days as they are very dangerous to recipients. They cause a lot of problems such as waste of storage space, reduction of communication band width and time losing for the identification and removal of their causes. In this paper a machine learning technique of two proposed stacked configuration will be applied on email data set. This data set has two types of emails, ham mails and spam mails. The preprocessing of these mails based on the analysis of all parts that constitute an email. Rather than considering only one part of an email such as content (mail body). The results of the proposed algorithm will be analyzed based on the training and testing of various performance evaluation metrics. Finally a comparative study will be applied with some of the recent models developed for this subject.

Keywords: Spam, Ham, Stacking, Classifier

1. INTRODUCTION

Electronic mail has become a powerful tool for information exchange, it affects a lot to people’s life due to its Neglible time delay during transmission, security of data being transferred, low cost and etc. they are used as a fast and inexpensive mode of sharing of both personal and business information in a convenient way. But its simplicity and ease of use has also made it exposed to scams. This leads to that a lot of users often find their inboxes full of undesirable mails or spam [1]. Spam in general is an unwanted thing that takes a lot of forms such as unsolicited bulk e-mail (UBE), unsolicited commercial e-mail (UCE), excessive multi-posting (EMP) and junk mail. A junk mail is an unwanted message published to large amount of recipients. Another form of mail spam is flooding the internet with many copies of the same mail, by sending it to a lot of users who generally would not otherwise choose to receive it [2]. Spam mail has become a big trouble for electronic users, it is a very annoying and dangerous for recipients. The severity of spam mails increases with the great development in internet in general and especially in electronic. Some email users say that they are using electronic mail less now because of spam. They fear they cannot retrieve the emails they need because of the flood of spam [2]. They are worry that their important emails to others are not being read or received. Because of the spam filtering techniques at recipients' side might detect them as a spam and mark them as junk or remove them.
Mail spams causes a lot of problems such as waste of storage space, communication bandwidth and time of users to check and delete them. Also it may contain malware as a script or dangerous attachment that can harm users' devices or networks. Another cause of mail spams is it enables spammers for stealing secret user information such as credit card informations, passwords and user's addresses. Sometimes it fills up user's mailbox and decrease the chance of reading the important mails for such user.

There are a lot of statistics of how mail spams problem becomes very dangerous. An average of spam mails is 4.7 mail per second, the max number of mail spams is 39.9 per second and the total reported mail spams is 146909593 [3]. In recent statistics, 40% of all emails are spam which about 15.4 billion email per day that cost internet users about $355 million per year [4].

To overcome the various problems results from the existence of mail spams, Spam filtering techniques or Anti-spam filtering is a must. These techniques are used to classify all the incoming mails into two general categories ham mails and spam mails. The ham mails contain legitimate mails of the user and spam mails must be refused and neglected or quarantined. Till now there is no available complete solution to spam mails, but there are several filtering techniques that have been used to solve the problem such as:

- List Based Spam Filtering: this technique classifies mails into spam or ham by categorizing senders of mails as spammers, or non spammer. All the incoming messages from the spammer sender will be blocked. Non spammer senders will be manipulated as trusted senders, and all the incoming messages from them will be allowed to be received by the recipients. The classification results depend on the analysis of the mail header especially the sender part and check it using known black and white lists, if a sender exists in the black list then this mail is classified as a spam, else it will be considered as a ham[1] [5].
- Content Based Spam Filtering: this filtering technique focuses on the body part of an email. By checking the test in the mail body it classifies a mail to ham or spam. [4] [6] [7] [8] [9] [10] [11].
- Rule Based Spam Filtering: it checks whether some rules are applied to both mail header and body, there rules as font size, font color etc. [1]

N.S. Kumar, D.P. Rana and R.G.Mehta used content based technique on "Ling Spam" dataset. Each mail in this data set is converted into document which is represented as a vector using Vector Space Model (VSM). A cluster of similar mails were formed using k-means and the spam cluster is selected. And a spam words list was created from it. A various combinations in which spam words occur in a set of mail is obtained, then the association rules is generated using Apriori algorithm and applied to detect if the mail is spam or not. The accuracy achieved by this method is 89.31% with precision of 60.1% and recall of 71.6% [6].

Prajakta Ozarkar, and Dr. Manasi Patwardhan proposed a model that classifies a set of mails using appropriate feature selection. A multiple feature selection methods are applied on different datasets using multiple machine learning algorithms. Finally the results were compared for each classifier to find the best one. Two datasets were used spam assassin and enron. The traditional preprocessing were performed using chisquare, information gain, gain ratio, relief, one R, correlation and symmetrical uncertainty feature selection methods. The Partial Decision Tree classifier and Random Forest classifier were applied [10]. Table 1 summarizes the results of this work.

Table 1: accuracy results of random forest and partial decision tree on Assassin and Enron data sets for different percentages of used features.

	Random forest		Partial decision tree	
	All features	87% of feature	All features	87% of feature
Assassin data set	94.35%	94.14%	92.29%	93.16%
Enron dataset	93.62%	93.43%	91.78%	90.73%

Masoumeh Zareapoor and Seeja K. R proposed a model depend on Feature Extraction or Feature Selection for Text Classification. This technique is used to classify mails into phishing and legitimate using feature extraction or feature selection method. They have used spam assassin data set to check their model. A traditional preprocessing on the data set was applied to build the Term Document Frequency (TDF) vector. The TDF vector is then simplified using the Principal Components Analysis (PCA) when feature extraction is used. Information Gain or Chisquare is used when the feature extraction method. And the classification process is applied by training the generated short vector using different classifiers. The achieved results depend primarily on the number of used features, that is 90%, 92.4% and 93.6% when the number of features is 10, 20, 500 and using the information gain method [11].

All the above researchers have developed a mail spam classification methods by manipulating only on part of an email at time [6] [10]. This manipulation leads to inaccurate results since they ignore other parts of an email that may include spam or annoying the recipients. Ignoring some mail parts in the classification process may affect the result. Some mails may contain spam data on the headers and have a clean body and vice versa [11].

In this paper a mail spam detection using stacking classification will be developed. This technique manipulates more than one part of an email concurrently. It can detect spam mails with more accuracy and efficient classification. This paper is organized as follow: section 1 include an introduction, in section 2 e-mail classification using Stacking method will be described. The experiment results are discussed in section 3. Finally our work will be concluded and the future work will be suggested.

2. E-MAIL CLASSIFICATION USING STACKING METHOD

In this section a model for mail spam classification using stacking will be presented. And a multiple parts of an email (Header, Content and Links) will be analyzed concurrently. The flow chart in figure (1) describes the steps of mail spam classification.

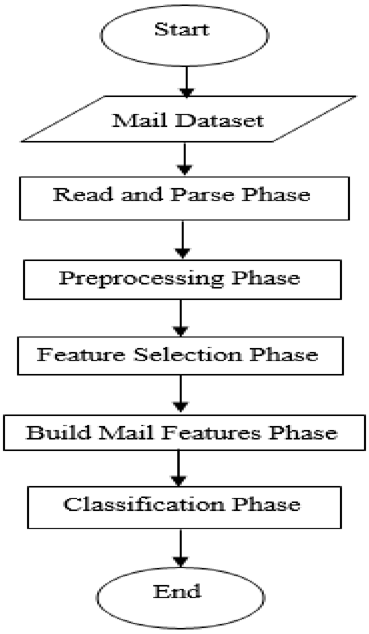


Figure 1: Cascaded Phases of Proposed Model

2.1 READING AND PARSING

In this phase, a set of mails will be entered to the system as a plain text. And MBX2EML tool will be used to generate the EML format (Microsoft Outlook Express file extension). The output EML file is accessed and parsed. The parsing process distinguishes between headers, content, links, URL's and HTML tags.

2.2 PREPROCESSING

In this phase the tokenization process will be applied on each mail file. First, each mail will be divided into a set of tokens. A token may be a word, stop word, stop character or special character. Second, the contents of each mail will be enhanced by removing stop words, stop characters, special characters. Finally, the word stemming process will be applied to find the root of all words. The stemming process decreases the number of tokens. Stemming is a process for removing the commoner morphological and in-flexional endings from words in English [10]. Some algorithms are available for stemming such as Suffix-stripping algorithms, Lemmatization, Stochastic algorithms and Affix stemmers. [14].

2.3 FEATURE SELECTION

In this phase, the stemmed tokens of each mail will be used to construct a long vector term of contents. This vector can be expressed as table, with number of columns equal to the number of the stemmed tokens without repetition. Each token here will be considered as one feature. The number of rows in the table equal to the number of mails constitutes the data set. The cells of the table contain a value that represents the occurrence of this feature in that mail. The value takes a various forms such as:

- Binary value that represent the presence or the absence of a specific feature in a specific mail.
- Term Document Frequency (TDF) to represent the frequency (number of occurrence) of a specific feature in a specific mail.
- Term Document Frequency (TDF) * Inverse Document Frequency (IDF) to represent the multiplication of TDF and IDF. The IDF equal to log (N/dfi) where N is number of all mails in data set and dfi is number of mails that contain this feature.

The features/mail table is always a huge matrix and contains some ineffective data which has no or less affect in classification process. This ineffective data must be removed and the result of this process is a short vector. This conversion can be performed using the following processes [9].

1. Dimensionality Reduction Techniques

This method used to transform a huge table of features into a shorter one that is more compact, predictive, and easier to handle. Feature selection is considered as one of dimensionality reduction techniques. In a feature selection technique, a subset of original features is selected to be used in the training and testing. This can be performed using Chi-Square, Information gain, gain ratio, one R or symmetrical uncertainty [10].

2. Extract Most Important Features of Header and Links

This step is used to extract the features exist in the header and links parts of an email to enhance the process of spam identification. For the header part, most important features can be extracted such as empty To part, From part = To part, Recipient list > 10, Recipient in BCC only and invalid mail address. For the links part, most important features can be extracted such as domains count, text link difference, dots count, image URL’s and IP URLs [16].

The header and links features are structured in a table. The table columns represent header and link features and the rows represents the mails. Each cell in the table has a binary value [0, 1] that represent that a specific feature is achieved in a specific mail or not.

2.4 FEATURE BUILDING AND PREPARATION

In this step all the mails features list will be prepared and formatted in a readable form for any classifier. This can be achieved by combining the short vector term with header and links' features for all mails in CSV File.

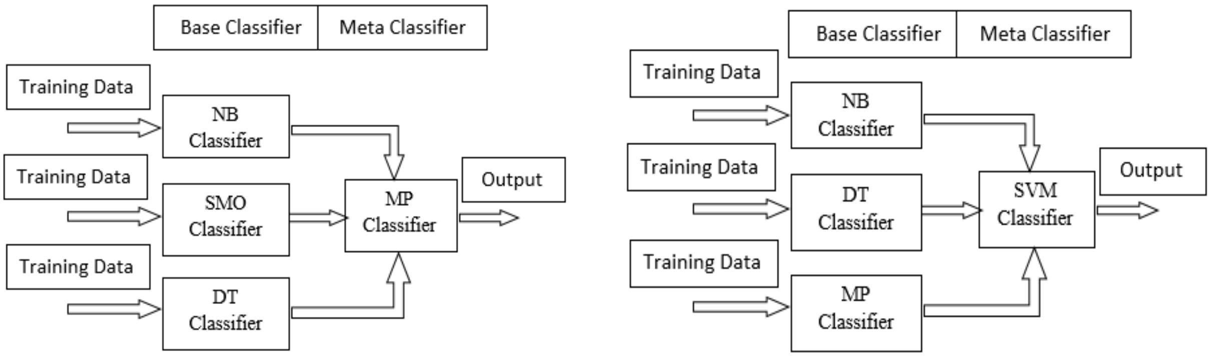
2.5 Stacking Classification

Stacking used to combine different classifier models for two levels base learner (level-0 model) and Meta learner (level-1 model). The predictions of base learners are feed as input to the Meta learner. The procedure of Stacking is as follow:

- Split training set into two disjoint sets.
- Train several base learners on the first part.
- Test base learners on the second part [17].

Using predictions from previous step as inputs and correct responses as outputs, train a higher level learner as shown in Figure 2. In this paper we will use two models:

1. Stacking NB-SMO-DT (MP) Model: in this model NB-SMO-DT will be used in parallel as base learner. And MP will be used as Meta learner as shown in figure (3-a).
2. Stacking NB-DT-MP (SVM) model: in this model NB-DT-MP will be used in parallel as base learner. And SVM will be used as Meta learner as shown in figure (3-b).



(a) NB-SMO-DT (MP) model **(b) NB-DT-MP (SVM) model**

Figure 2: used stacking classification model

3. EXPERIMENTAL RESULTS

This section presents the results of our proposed model. Stacking classification method has been used for the testing process. The previous described stacking models in figure 2 will be applied to the features extracted from the dataset.

The experiments were executed using PC with processor core i7, 32 GB RAM and using WEKA data mining software [15], using 10-fold cross validation technique which uses 0.9 of training data set for training the classifier and remaining 0.1 for testing the classifier and repeat this procedure 10 times with changing training and testing each time.

The following metrics are used to evaluate all the configured models: Accuracy, True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN), Precision (P), F-measure, Recall (R) and Receive Operating Characteristics Curve (ROC) [9]. Table 2 summarizes the evaluation of the used models related to the metrics listed above by analyzing only the body part of each mail. The results show that stacking with (NB - SVM – DT) (MP) achieves accuracy of 91.23% and accuracy of 91.40% when the stacking using (NB-DT-MP) (SVM) is applied.

Figure 3 and Figure 4 show the comparison between the used stacking models.

Table 2: Performance Evaluation Using Content (Body) Only

	Evaluation Metrics						
Classifier	A	TP	FP	P	R	F-Measure	ROC Area
Stacking (NB - SVM – DT) (MP)	91.23%	0.912	0.104	0.914	0.912	0.912	0.962
Stacking (NB-DT-MP) (SVM)	91.40%	0.914	0.103	0.916	0.914	0.913	0.905

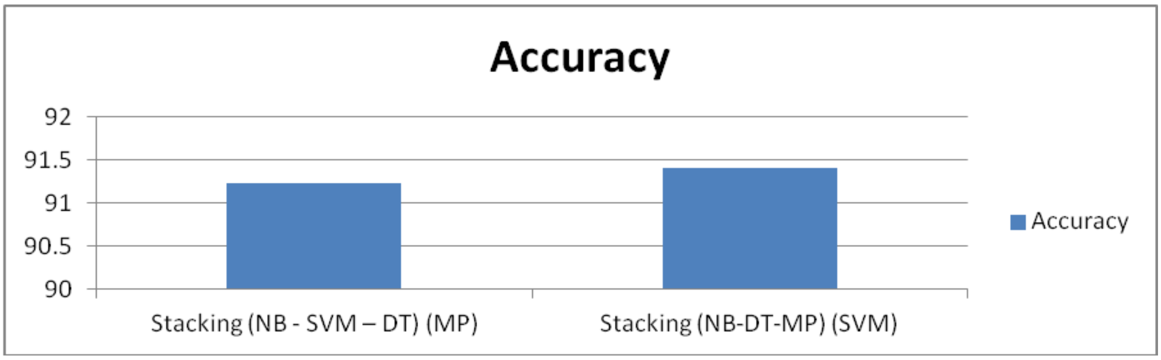


Figure 3: Accuracy of Proposed Model Using Content (Body) Only

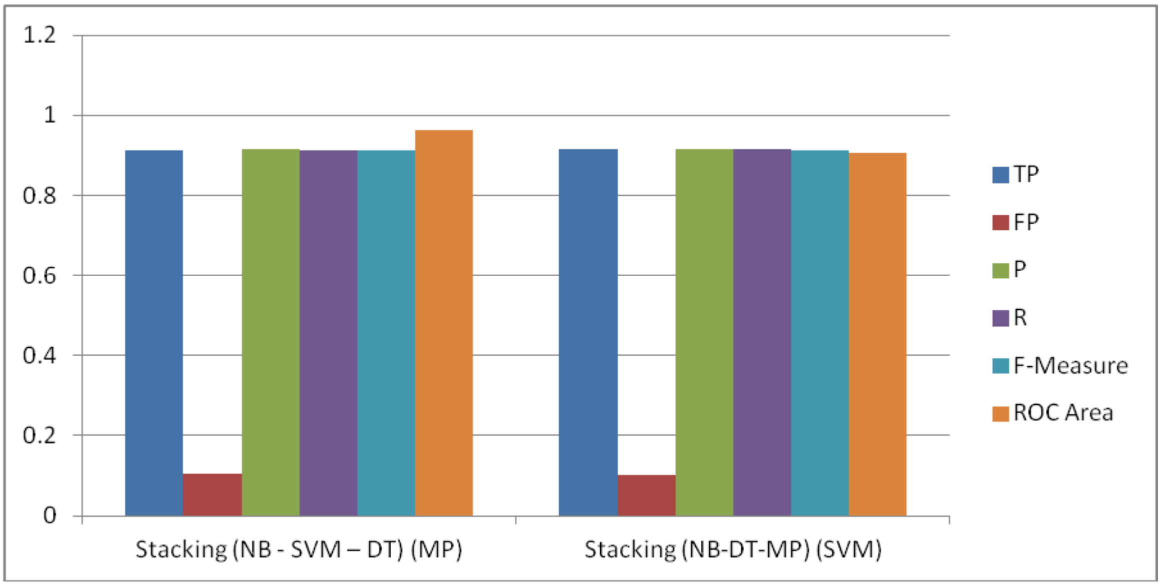


Figure 4: All Metrics of Proposed Model Using Content (Body) Only

Table 3 summarizes the evaluation of the used models related to the metrics listed above by analyzing all parts of an email, Body, headers and links. The results show that stacking with (NB - SVM – DT) (MP) achieves accuracy of 95.67% and accuracy of 95.66% when the stacking using (NB-DT-MP) (SVM) is applied. Figure 5 and Figure 6 show the comparison between the used stacking models.

Table 3: Performance Evaluation Using All Parts of Mail (Content, Header and Links)

Classifier	Evaluation Metrics						
	A	TP	FP	P	R	F-Measure	ROC Area
Stacking (NB - SVM – DT) (MP)	95.67%	0.957	0.051	0.957	0.957	0.957	0.985
Stacking (NB-DT-MP) (SVM)	95.66%	0.957	0.051	0.957	0.957	0.957	0.953

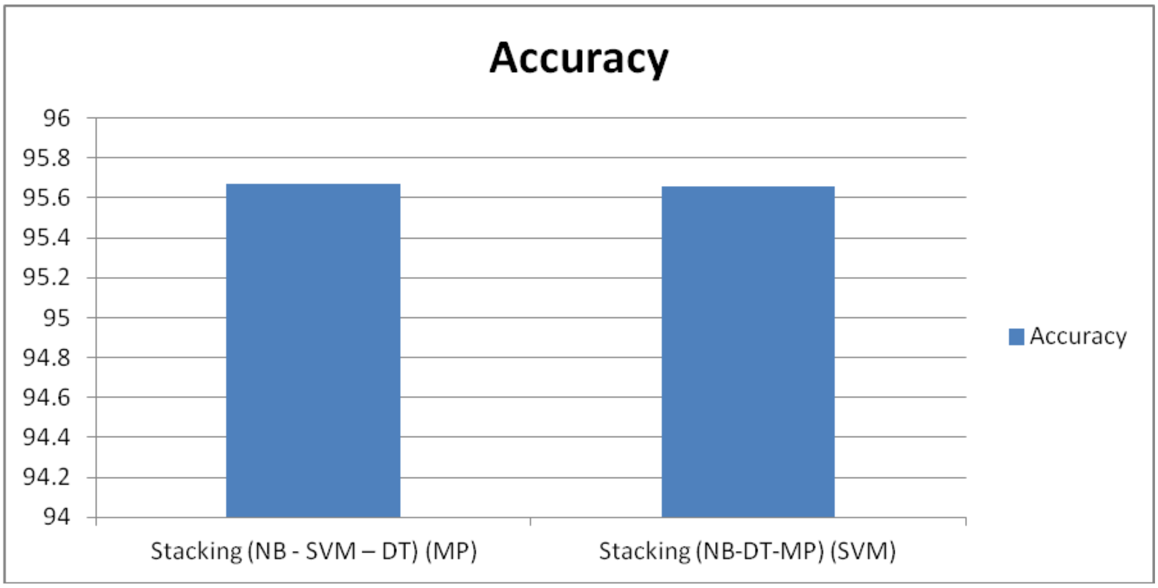


Figure 5: Accuracy of Proposed Model Using All Parts Of Mail

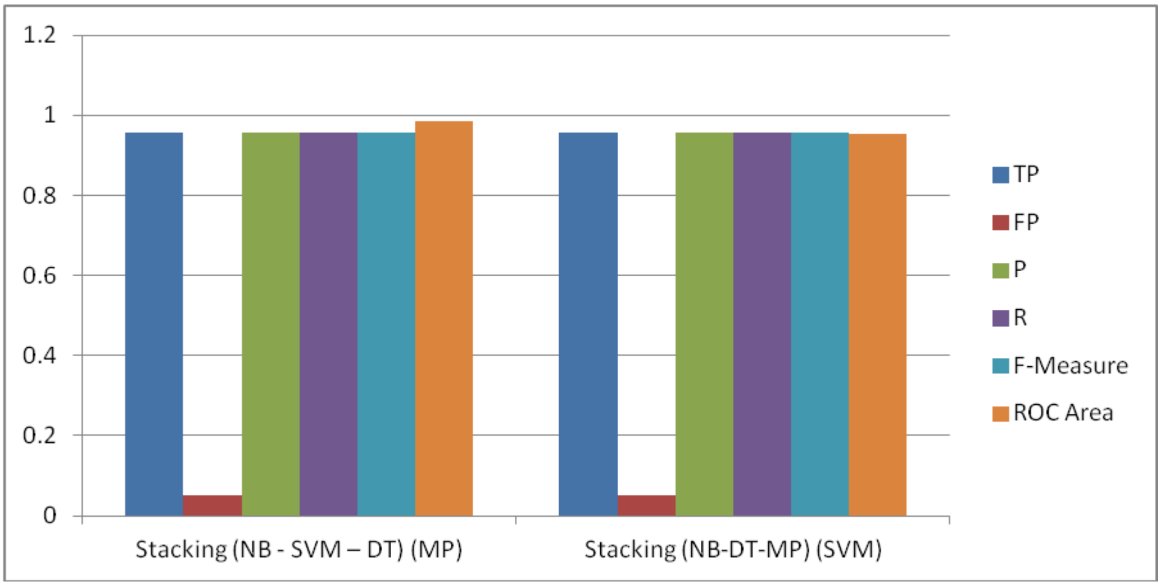


Figure 6: All Metrics of Proposed Model Using All Parts of Mail

Figure 7 shows a comparison between the used two stacking models when only the body of an email is used in the preprocessing and when all the items of an email are used. As the figure shows when using all mail parts in the analysis increase the accuracy by 4% for the two models.

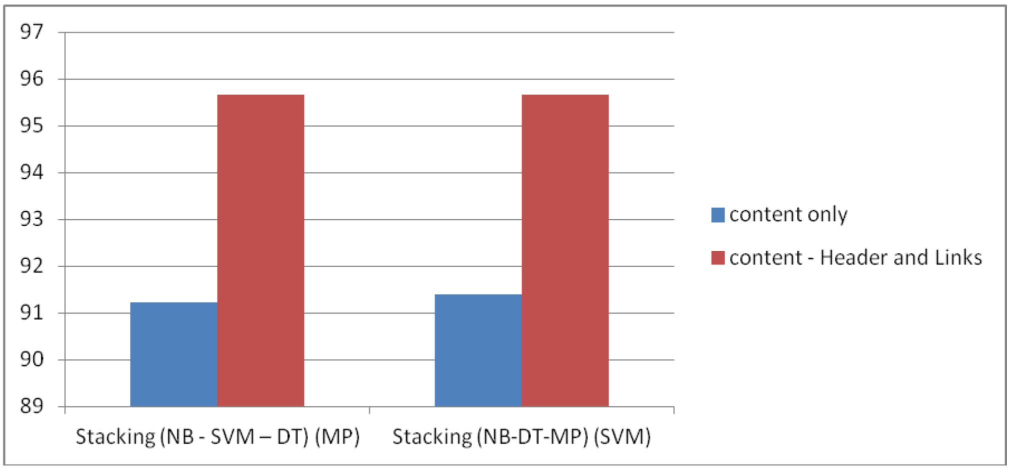


Figure 7: Comparison of Accuracy between Using Content Only and Using All Parts

In table 4 and figure 8, a comparison was performed between some of the selected previous work and the stacking models. This comparison indicates that mail spam with the stacking models gives better results than the other models. The accuracy is increased by nearly 1% from the highest one.

Table 4: Comparison between Accuracy of Our Model and Other Models

Accuracy	Used data set	Used Technique	Paper
89.31%	Ling Spam	Content Based	[6]
94.63%	Spam assassin	Content Based with feature selection method	[10]
93.6%	Spam assassin	Content Based with feature extraction method	[11]
95.67%	Spam assassin	Content Based with feature selection method and stacking classification method	Our model

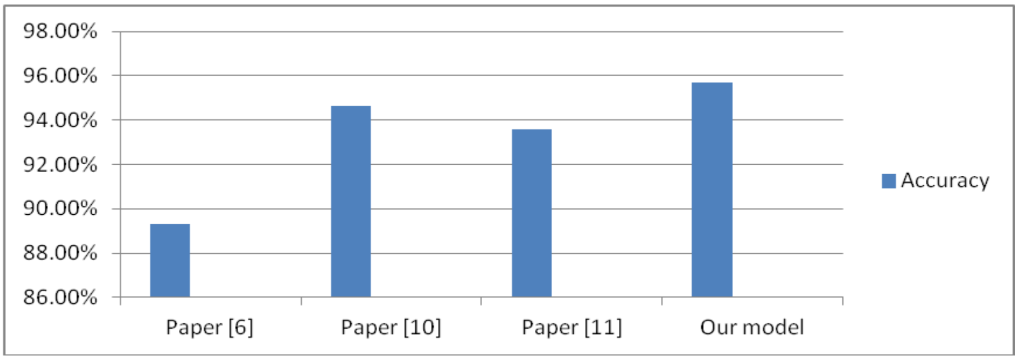


Figure 8: Comparison between Accuracy of Our Model and Other Models

4. CONCLUSION AND FUTURE WORK

This paper proposed a model for mail classification into ham and spam by applying machine learning techniques, firstly we collected suitable labeled data set and used intelligent preprocessing including word stemming steps to prepare data then we used Dimensionality reduction techniques to select most important features of all parts of mail (Header, Body and Links). The model was experimented using different classifiers as SVM, Navie Bayes, Bagging, Voting and Stacking, in each time we used 10-fold cross validation technique in training and testing processes to overcome overfitting problem.

Achieved results are competitive and enhance classification accuracy as mentioned in results and discussion section and show that the best results was achieved when using all parts of mail (not Content only) with stacking method due to multi levels of classification which are executed in it.

As future work, our proposed model can be enhanced by using images and attachments which are contained in mail, by using image processing techniques and by examining attachments we can deal with images and attachments respectively and we can extract important features from them which can help us in classification and surely this will enhance accuracy.

REFERENCES

[1] Sahil Puri1, Dishant Gosain2, Mehak Ahuja3, Ishita Kathuira4, Nishtha Jatana5, "COMPARISON AND ANALYSIS OF SPAM DETECTION ALGORITHMS", International Journal of Application or Information in Engineering & Management (IJAIEM), Volume 2, Issue 4, April 2013. ISSN 2319 – 4847.

[2] Mr. Santosh A. Shinde, Dr. R.K. Kamat, "Synopsis Report on a Semi Custom ASIC for Circumventing Spam", July 2007.

[3] https://www.spamcop.net/spamgraph.shtml?spamyear accessed January 2017.

[4] S.Divya1, T.Kumaresan2, "Email Spam Classification Using Machine Learning Algorithm", International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297:2007 Certified Organization), Vol.2, Special Issue 1, March 2014.

[5] Alaa H. Ahmed, Mohammed Mikki", "Improved Spam Detection using DBSCAN and Advanced Digest Algorithm", International Journal of Computer Applications (0975 – 8887) Volume 69– No.25, May 2013.

[6] N.S.Kumar1, D.P.Rana2, R.G.Mehta3, "Detecting E-mail Spam Using Spam Word Associations", International Journal of Engineering Technology and Advanced Engineering", (ISSN 2250-2459, Volume 2, Issue 4, April 2012).

[7] Ms.D.Karthika Renuka, Dr.T.Hamsapriya, "Email Classification for Spam Detection Using Word Stemming", ©2010 International Journal of Computer Applications (0975 – 8887) Volume 1– No. 5.

[8] Mumtaz M. Al-Mukhtar, Yasmine M. Tabra, "An effective spam filter based on a combined support vector machine approach", Int. J.Internet Technology and Secured Transactions, Vol. 4, No. 1, 2012.

[9] Adwan Yasin and Abdelmunem Abuhasan, "AN INTELLIGENT CLASSIFICATION MODEL FOR PHISHING EMAIL DETECTION", International Journal of Network Security & Its Applications (IJNSA) Vol.8, No.4, July 2016.

[10] Prajakta Ozarkar, & Dr. Manasi Patwardhan,” Efficient Spam Classification by Appropriate Feature Selection”, Global Journal, of Computer Science and Technology Software & Data Engineering Volume 13 Issue 5 Version 1.0 Year 2013, Online ISSN: 0975 –4172 & Print ISSN: 0975-4350.

[11] Masoumeh Zareapoor, Seeja K. R, "Feature Extraction or Feature Selection for Text Classification: A Case Study on Phishing Email Detection", I.J. Information Engineering and Electronic Business, 2015, 2, 60-65.

[12] http://spamassassin.apache.org/publiccorpus/ accessed December 2016.

[13] https://monkey.org/~jose/phishing/ accessed December 2016.

[14] Porter, M.F. (1980), “An algorithm for suffix stripping”, Program, Vol. 14 No.3, pp. 130-137.

[15] ark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

[16] Mike Spykerman – CEO Red Earth Software, "Typical spam characteristics, How to effectively block spam and junk mail", Red Earth Software.

[17] Mi ZhiWei, Manmeet Mahinderjit Singh, Zarul Fitri Zaaba, "EMAIL SPAM DETECTION: A METHOD OF METACLASSIFIERS STACKING", A method of meta classifiers stacking in Zulikha, J. & N. H. Zakaria (Eds.), Proceedings of the 6th International Conference on Computing & Informatics (pp 750-757). Sintok: School of Computing.