International Academic Institute for Science and Technology

International Academic Journal of Science and Engineering

Vol. 3, No. 2, 2016, pp. 146-160.

ISSN 2454-3896

International Academic Journal of Science and Engineering

www.iaiest.com

Filter Spamming In Computer Networks by Text Mining and Machine Learning Method

Mahsa Riahi Asl [a]*, Hasan Naderi[b]

[a] *Department of Computer, Khoramabad Branch, Islamic Azad University, Khoramabad, Iran*

[a] *University of Science and Technology*

Abstract

The goal and motivation of spam is transmitting information to the recipient, such information including an advertisement for a product (which is likely to be worthless, illegal or non-existent), prey to a fraudulent scheme or to promote a cause or destroy a computer, which often designate to influence or stealing the recipient's computer. Due to costless of sending information via an e-mail, a method is used for spammers and those who have specific targets. Only a very small part of the recipients may be as a target. This number could be one in ten thousand or less. But prevent and filtering the spam could improve the network traffic. Although in recent years the introduction of a large number of text classification algorithms have been proposed, but text mining requires a series of pre-processing, effective, efficient and fast learner algorithms. In this paper, a series of experiments are performed on spam data, and a method with high reliability and low error rates have been provided.

Keywords: spam, spam detection, machine learning, data mining

Introduction:

Spams that are as e-mail, simply just not as waste, and since including attachments that have virus and spyware agents, can be dangerous for a system and its recipients and lead to loss of information. So we need to have tools to detect spam. Spam detection through the history of data collected lead to continuous conflict between the e-mail service providers and producers of spam. A large volume of spams are distributed by robots daily around the world. Spam often annoying to many users. Because spam send to the reliability of e-mail, sometimes users affected by wasted network bandwidth, wasted time, money and

sometimes do not receive some messages. Many spam detection techniques have been proposed based on machine learning techniques (Androutsopoulos, I., et al., 2000).

Nowadays, technological and technical point of view raises a very important issue as spam, which in term of law must be avoided. In recent years, the rapid growth of computer networks has changed the world. Computer security is essential because of the attacks. Many companies in the field of anti-spam were act based on the sign, and have been successful in early detection of spam. However, a new type of anonymous spam cannot be detected by the software. Spam detection and Email classification problems are solved using machine learning systems. At first we must establish spam and non-spam detecting and then classifying e-mail using spam and non-spam to reduce the error rate (Androutsopoulos, I., et al., 2000).

Spam classification

Classification is one of the techniques that used in machine learning. Patterns belong to different classes, which limits decision borders, and then divided into two categories: learning and test, and classification were done according to the class learning, and test collection is to evaluate the performance of classes (Androutsopoulos, I., et al., 2000).

Spam is often sent as a commercial content. In spam detection, different techniques are provided to deal with these types of E-mails; the results are not good due to ongoing changes in spam behavior patterns.

Previous work in the field of spam detection:

In reference (McCallum & Nigam,1998) the authors were done systematic experimental studies to compare two different simple Bayes models: multivariate Bernoulli and MNB then come to the conclusion that the MNB in term of collection word text has higher data size than Bernoulli multivariate and acts better.

In (Yang & Liu,1999) proposed researches have confirmed the above results and in addition comparing the MNB with other classification. Their findings also suggest that the MNB significantly better than distinctive classifier, including the SVM.

To improve the performance of the MNB, (Rennie et all,2003) proposed a simple supplement Bayes (CNB). The authors note that the MNB may select poor weighting for decisions boundary when class distribution in learning data is unbalanced. . They have suggested that while learning conditional probability of a CNB class, we use a lot of information related to other classes.

Reference (Zhang & Oles,2001). checks the number of linear classification, including logistic regression, and support vector machines and MNB. The authors pointed out that numerical setting and optimization is essential for the functioning of logistic regression, and then they propose numerical optimization problem solving with a descending circular algorithms, which reduces classification error with learning parameters, one after the other. Their numerical studies suggest that their method provide compact predictive models, which can compete with the SVM class or class logistic regression border provisions in combination with feature selection in the performance. (Drucker, et all, 1999) use support vector machine in e-mail classification and classified them to spam or non-spam.

This study compared three classification algorithms such as Ripper, Rocchio, and decision trees. In this study, we have a data set, where 1000 features have been selected as the best features and another datasets that have 7000 features. The four algorithms were tested on two different datasets. When using a binary features, support vector machine is performed the best mode. For both data sets, strengthening the trees

and support vector machine have acceptable test performance in terms of accuracy and speed. However, support vector machine has less training time.

According to the above restrictions, (Carreras & Marquez,2001) used machine learning algorithms. One of the methods that are widely used is the Bayes classification.

In fact, this method works on the probability basis of whether a message is spam or not spam as well as by repeated Features performs the calculation. Spam Bayes is a striking example of open source software. (Duda et al , 2012) used the Support Vector Machine to detect and classify spam.

Known learning algorithms to increase spam detection was proposed by Salem & (Stolfo , 2010) that classified with a single vector machine algorithm, and finally has been tested with artificial neural networks (ANN).

However, these methods still incur large costs and are according to the heavy calculations and low detection rate, and that's why they do not use feature selection and use of irrelevant features. Spam is considered as an invasion of privacy.

(Abu-Nimeh ,et all,2008) using Bayesian additive regression trees (BART) to classify spam, evaluate their performance against other classification methods, including logistic regression, support vector machine, classification and regression trees , neural networks, random forest, and simple Bayes. We evaluate the class estimation using three spam data sets: Ling Spam, PU1 and Spambase to determine the accuracy of the forecasts and false positive rate.

They also do not optimize machine learning algorithm parameters. The aim of parameters optimization is setting the value of several parameters in the machine learning algorithms to determine optimal values. For example, the weight and the number of hidden layer in ANN, (Breiman , 2001) and the value of the kernel function' parameter SVM, and etc.

(Liang & Winstanley ,2008) showed the parameter optimization in their algorithm, but they did not explain how to compute the optimal parameter values. In addition, they were not use feature selection methods to identify spam. The aim of selection is specifying irrelevant features including all audit data features to reduce the processing time and improve the detection rate.

All features for classifying whether an email is legitimate or spam, is not necessary and therefore irrelevant features, not only increases the computational costs (etc, time and resources), but also reduces the rate of classification.

In the reference (Liang et all, 2008) they did feature selection, but did not mention how important number of features selected, and the importance of each feature variable as a numeric value not provided. Although the authors (Liang & Winstanley, 2008) did feature selection using a ranking algorithm feature, but their detection rate is very low.

(Lee & Han, 2012) used ABPNN methods that have not BPNN algorithmic problems, such as slow learning and not being stuck in local minima. In this method, the three types of semantic similarity method are used for spam filtering. First, an atmosphere of latent semantic features (LSFS) is used to reduce the size and construction of latent semantic relationships between words, to improve the accuracy and efficiency of classification, from a set of training data with a statistical method made, then extract the relationship between words based on their occurrence in the texts and similar words together, formed a database (CBT) and finally obtained feature space will be combined with data-set made (HSS). Results are evaluated on a data set Ling Spam.

(Zhou, 2007) proposed a strategy based on a logistic regression, that the structure of each message stored in a Huffman tree and a tree is used for each class. Words and their repetition in the message stored in tree and tree feature vectors were used that encode this information in the tree.
Then the features are randomly selected among the features that have been recently used preferentially and model given training. This method has been tested on a dataset Ling Spam and has high performance, stability and speed.

Data set

Each machine learning system needs a training set to train the system. In this study, reference (Abu-Nimeh et all ,2008) was used as a standard data set, which from 2893 text messages 2412 messages (approximately 83.37%) are true and 481 message (approximately 16.63%) are spam. In this data set, all of HTML tags and headers have been removed and only the text is used. Details of ling spam data sets are shown in Table 1.

Table 1: details of the ling spam data set

Category	Number of educational documents	The total number of Category documents
Spam	481	**2893**
No spam	2412	

Preprocessing

To categorize emails after providing data collection, we go to pre-processing stage. At this stage the raw text documents must convert to a format that can be used to feature selection and learning algorithms stage. In the preprocessing stage of proposed method the Transform Case, Tokenization, Filter Stop Words, Stemming and Generate n-gram is done. Each performance is given below.

- Transform Case:. At this stage, all the characters in the text, converted to the same form. At this stage, all the characters are converted to lowercase letters.
- Tokenization :At this stage, the entire text divided as consecutive apart words
- Filter Stop Words : At this point superfluous words in the English language will be deleted
- Stemming : At this stage the roots of words in the English language, using root finding algorithms, is used to remove suffixes and prefixes of words in order to reduce the length of words and until they become their root form
- Generate n-gram: At this stage, for indexing and reduce the text size, the n-gram is used. Using n-gram can revealed text as a series of consecutive words by the length of n. This model was initially proposed for speech processing issues. But now many different versions of this model for natural language processing and classification issues have been proposed.
- After the above steps, features weighting is done. In this study, TF-IDF weighting method is used, the categories performance is evaluated using the weighting method. The figure below shows the proposed method

NAIVE BAYES algorithm

Another classification algorithms proposed in (Shipp & Kuncheva, 2002) is the simple Bayes. This is important for several reasons. The construction is simple and does not require complex iterative parameter estimation programs. This means that it can also be used for very large data sets. This algorithm is one of the fastest algorithms in classification.

Multinomial Naive Bayes algorithm

Text classification task can be considered as a Bayesian learning perspective, that the distribution of words in documents created by certain parametric model, and the parameters can be calculated through data learning. Identifying (exploring) learning data set, we can obtain all the required frequency and then calculate the corresponding conditional probability. One advantage MNB model is that perform predictive with high efficiency. So to do MNB forecasts we only need to search for the word with a circulation of zero. Traditionally, along with low duplicates, text words are used. In contrast, other classes such as network must requiring search among all words in the learning set. Since in classification, the size of words is often much larger document length. MNB algorithm can perform forecasts faster than other classes except Bayes family (Jing et all,2002).

Discriminative Multinomial Naive Bayes algorithm

However, we have introduced other distinct polynomial simple Bayes (DMNB), which uses the parameter learning method described in reference (Su et all,2008) for a polynomial simple Bayes model. DMNB goal is to keep much information while concerned to the nature of distinct classification, and so is a combination of public learning and distinguished. MNB is produced through learning algorithm documents. It is clear that DMNB has the same computational properties; this algorithm is more accurate and faster than the MNB and also supports online learning.

Complement Naive Bayes Algorithm

Simple Bayes supplementation classification (CNB), improves Simple Bayes classifier weakness with data parameters in all categories of sensitive set (Su et all, 2008).

J48 Tree Algorithms

Classification of decision tree divided into two categories CART and C4.5 based on the attributes selection criteria at decision nodes. J48 decision tree is Improved C4.5 that is created pruned or not pruned C4.5tree (Zhou,2007).

Random Forest Tree Algorithm

Random Forest suggested by BREIMAN that is a set of regression and classification not pruned trees, and including the bootstrap samples of training data, which is created using random feature selection in the process of tree inference. Predicting is made by aggregating predictions of or majority vote or Voting for the classification. Its return make stronger the generalization error rate and noise. However, is like many other categories. RF tree algorithm can also cause incorrect learning and very unbalanced set of training data. As it made to minimize the overall error, it tends to more focus on accurately predicting the majority of the classes, which often include low accuracy results for a number of classes (Breiman ,2001).

Random Tree algorithm

The decision tree algorithm Random Tree is randomly makes numerous decisions trees. At the time of making each tree, the algorithm randomly selects one of the remaining properties that in each extension the nodes must be examined without any change (Duda,2012).

SVM algorithm

This method is relatively of new techniques, which in recent years show a good performance than traditional methods for classification of perceptron neural networks. The work basis of SVM classification is the linear classification data, and in linear dividing we try to choose the data that has more confident marginally. Solving the equation to find the optimal line for data by QP methods that is a known method for problem-solving is carried out in constrain method problem solving.

KNN algorithm

K nearest neighbour of a group method is including the K record of a series of educational records, which select the closest record to the trial record, and according to superiority of their category or tags, will make decision on the trial record. In simple words, this method selects the class that is in the selected neighborhood of the largest number of records assigned to that category. Therefore the class that is seen larger than other classes in the K-nearest neighbor is considered as a new record category.

Feature selection methods

IG Feature Selection Algorithm

Attribute reduction algorithm IG, do the features assessment with respect to the measurement data derived from them and given to the classes.

This discrete number of features in the first time is done using discrete-based MDL (Duda,2012). In fact, C is a set of c data samples with m distinct class. CI training data set includes examples of class I. Expected and required information to classify the sample is calculated by the following formula:

$$I = (C_1, C_2, \dots, C_m) = -\sum_{i=1}^{m} \frac{c_i}{c} \log_2 \left(\frac{c_i}{c}\right) \quad (1)$$

Where the probability custom samples are belongs to the class Ci. Feature F include V separated and specified value of {F1 ،F2 ،... ،FV} that can divide training class to V subclass {C2…., Cv,C1,} where F is the CI subclass for each feature and CJ include classes Cij of class I. F features Entropy is as follows

$$(F) = \sum_{i=1}^{v} \frac{c_{ij+\dots+c_{mj}}}{c} \times I\left(c_{ij+\dots+c_{mj}}\right) \quad (2)$$

Finally Information Gain (IG) is calculated for F by following formula:

$$Gain\,(F) = I\,(C_1, C_2, \dots, C_m) - E(F) \quad (3)$$

Feature Selection with Cfs Subseteval Methods

It is a filter that evaluated the subset of features based on their predict ability with a degree of redundancy. As a result, features that has highly correlated with the target variable and while the correlation between features is low is preferred (Jing,2002).

Combination Classification

Bagging algorithm

Bagging algorithm is only one method of learning. With different sampling along with replacement from the training set, several new training set of the same size is achieved. In practice it has been shown that Bagging has very good results, especially when the size of the data set is small (Shipp,2002).

Boosting method

In Boosting the classification algorithm which is called weak learning method (Polikar ,2006) several times, and each time with different training data that selecting with regard to the previously implementation, implemented and finally the most frequently answers is selected. However this method is time-consuming, but its answers will be sure. In this method, each sample of training data set is assigned a weight. The weight of all samples is the same, and in the next iterations with respect to the category performance, the weight of each sample will be reformed, so that the weight of samples were classified as wrong is reduced. ADABOOST is the most popular Boosting method.

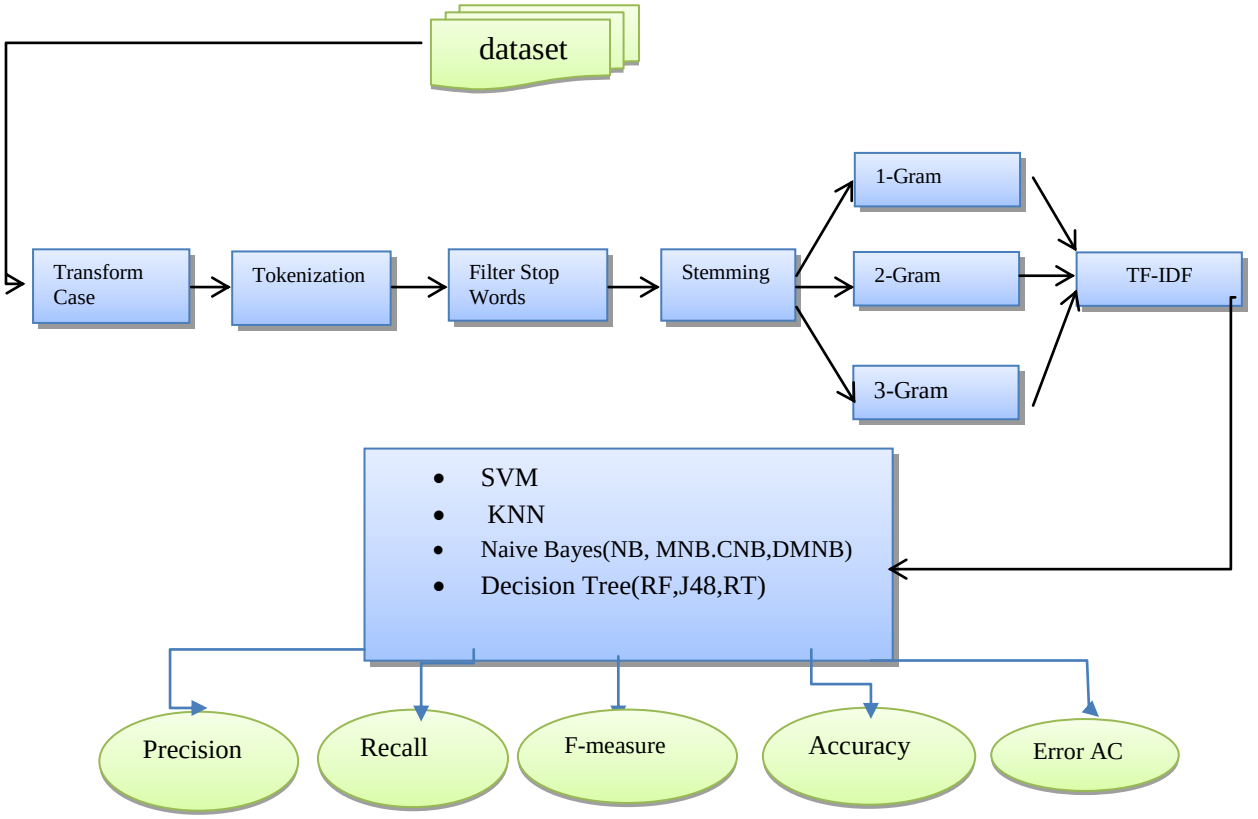


Figure 1: Simple Methods

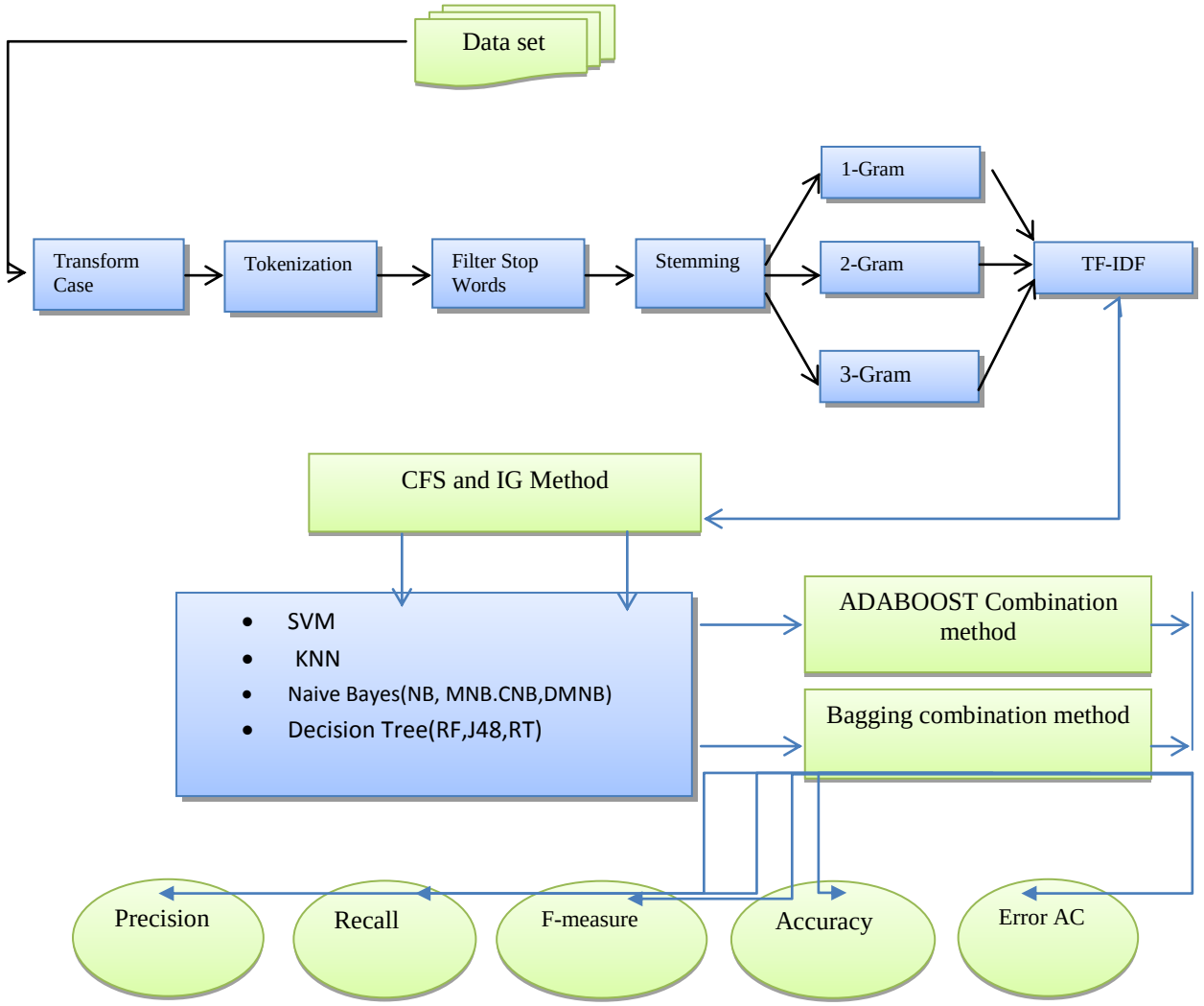


Figure 2: Combined Methods

Performance evaluation of proposed method

WEKA tool includes a collection of machine learning algorithms for data mining tasks. Algorithms can be directly applied to the data set or be called by your Java code (Patterson et all,2008) and (Hall et all,2009) .WEKA machine learning tool also includes tools for primary processing, classification, regression, clustering, association rules, and visualization. This tool can be useful for developing new machine learning schemes. WEKA contains tools such as investigative tools, testing, data flow and knowledge, a simple command line interface and a Java interface (weka softwere).

Evaluation Criteria

To evaluate the suggested methods, we used data collection test. Below are the criteria by which we evaluate the proposed method. They are:

- Accuracy
- Recall
- Precision
- F-Measure

Several parameters are used to evaluate the effectiveness of the algorithms. Following shows the performance categories algorithm according to the input data set based on types of issue category.

TN (True negative): Percentage of valid records that are classified correctly

TP (True positive): The percentage of record that are incorrectly classified

FP (False positive): The percentage of records that were mistakenly put in wrong class, while have valid activities.

Percent of the records that were mistakenly considered as properly function, while they are no correctly.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (4)$$

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

Accuracy is the most important criteria for determining the efficiency of a standard clustering algorithm. This measure calculates the accuracy of a classifier. This criterion indicates that how percentage of the entire data set is correctly classified. Of note is that, according to many criteria Recall shows classifier efficiency according to the number of category i accordance, while the Precision criteria is basically based on the accuracy of predicted category, and indicates that how much we can trust the output classifier. F-measure criteria is achieved from the composition of Precision and Recall criteria and used in some cases, which cannot consider great importance for non-Precision and Recall criterion relative to each other. Equation (7) shows how to calculate this measure.

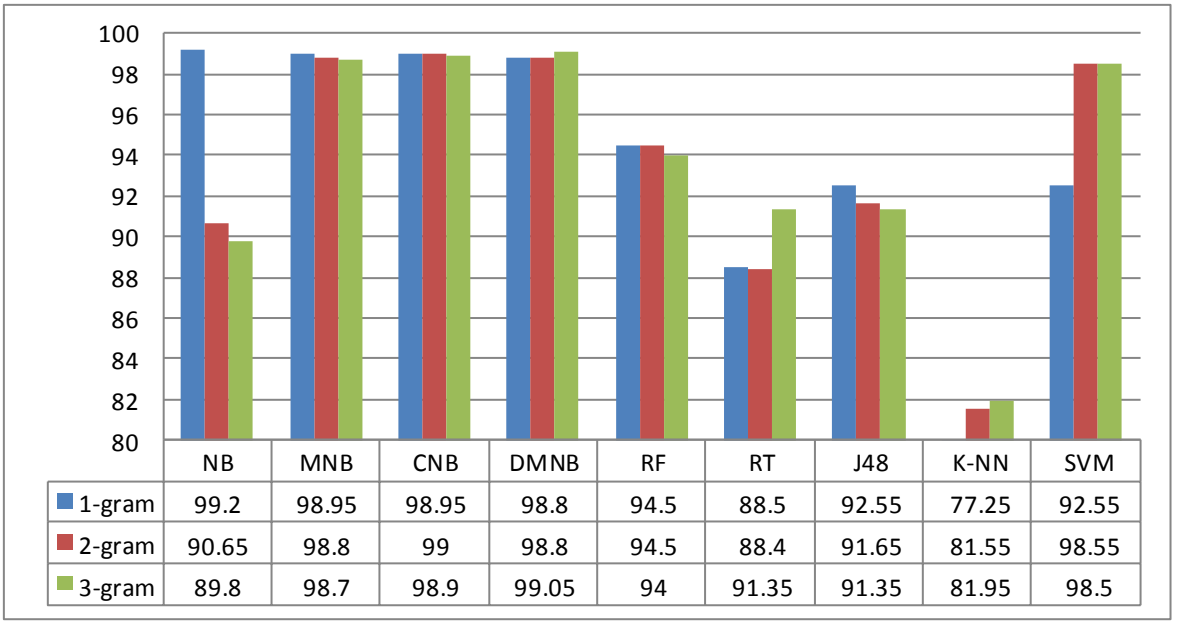
$$F - Measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (7)$$

Results

Figure (3) as you can see, is the highest performance of the two algorithms of the Bayes family namely CNB and the MNB with one gram. The performance of both algorithms is 98.95% and categories error is 1.05%. The DMNB algorithm also has high efficiency among other algorithms and has little difference with the two algorithms CNB and MNB.

The algorithm can also be proposed as useful and efficient. The highest performance in two gram belongs to an algorithm of Bayes family named CNB.

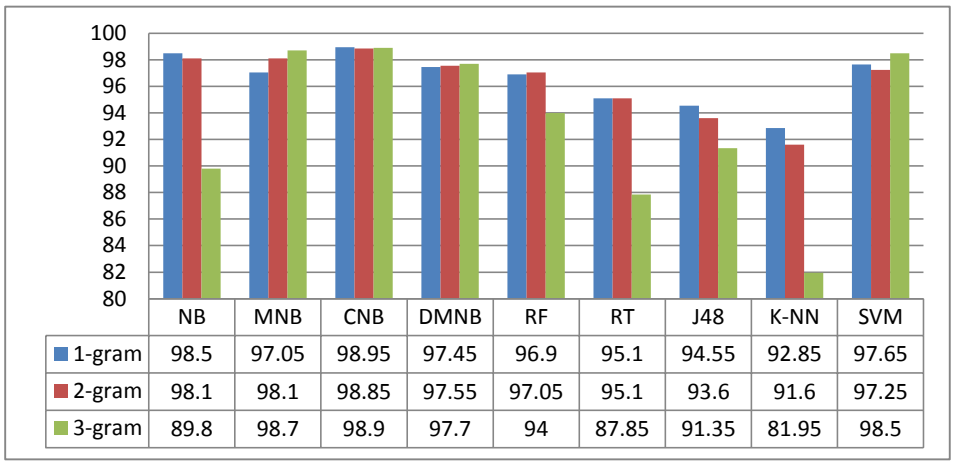
Functionality of algorithm is 99.00% and the categories error is 1.00%. And finally the highest performance of the algorithm belongs to DMNB of Bayes families with three grams. Functionality of algorithm is 99.05% and categories error is 0.95% respectively.



	NB	MNB	CNB	DMNB	RF	RT	J48	K-NN	SVM
1-gram	99.2	98.95	98.95	98.8	94.5	88.5	92.55	77.25	92.55
2-gram	90.65	98.8	99	98.8	94.5	88.4	91.65	81.55	98.55
3-gram	89.8	98.7	98.9	99.05	94	91.35	91.35	81.95	98.5

Figure 3: Comparison of 1 to 3 grams with simple algorithms

Figure 3 shows the performance of the three parameters 1 gram to3 gram. As you can see Bayes algorithms have shown good performance and have a smaller margin of error. The best algorithm is DMNB with 3 grams and 99.05 performance and margin of error equal to 0.95 percent. The following are methods to reduce the BEST CFS features with N-Gram.

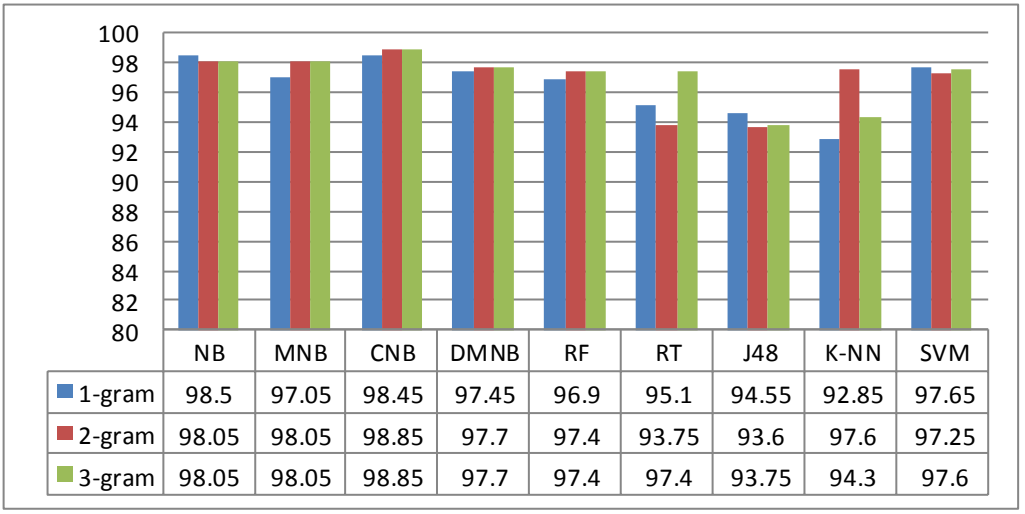


	NB	MNB	CNB	DMNB	RF	RT	J48	K-NN	SVM
1-gram	98.5	97.05	98.95	97.45	96.9	95.1	94.55	92.85	97.65
2-gram	98.1	98.1	98.85	97.55	97.05	95.1	93.6	91.6	97.25
3-gram	89.8	98.7	98.9	97.7	94	87.85	91.35	81.95	98.5

Figure 4: Comparison of 1 to 3 grams with CFS BEST method

Figure 4 shows the highest performance of Bayes algorithms with all the methods 1 gram to3 gram and reduction. As you can see Bayes algorithms have shown good performance and have a smaller margin of

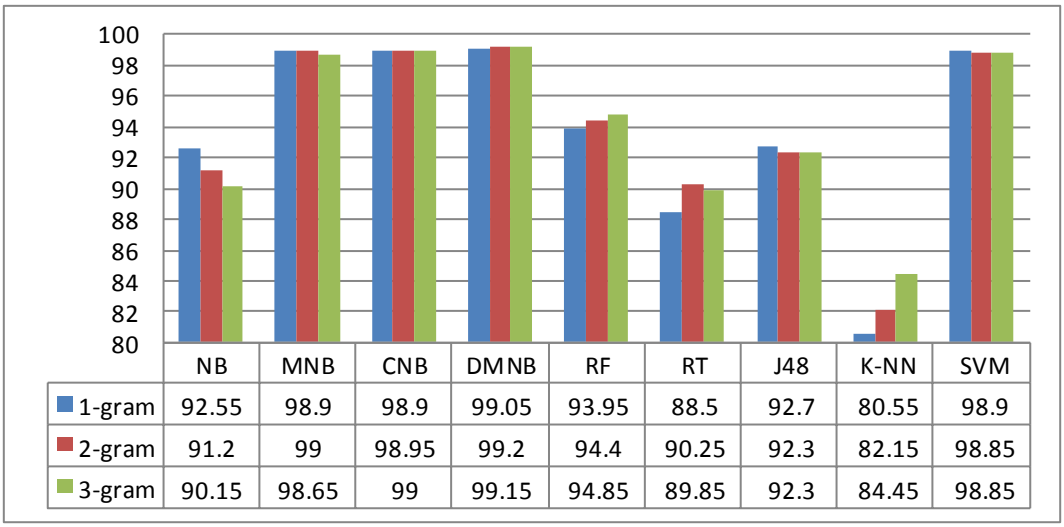
error CFS BEST feature. The performance of both algorithms is 98.95%, 98.85% and 98.90%, respectively. Performance of 1gram to 3 gram with three parameters was compared. As you can see, Bayes algorithms have shown good performance and have a lower error rate. The best algorithm is the CNB with 1 gram and 98.95 performance and margin of error is 1.05%. Next we investigate methods of reducing features with N-Gram CFS greedy.



	NB	MNB	CNB	DMNB	RF	RT	J48	K-NN	SVM
1-gram	98.5	97.05	98.45	97.45	96.9	95.1	94.55	92.85	97.65
2-gram	98.05	98.05	98.85	97.7	97.4	93.75	93.6	97.6	97.25
3-gram	98.05	98.05	98.85	97.7	97.4	97.4	93.75	94.3	97.6

Figure 5: Comparison of 1 to 3 grams with CFS greedy method

In figure 5 the highest performance belongs to Bayes family algorithms namely Naïve Bayes and CNB with methods 1 gram feature reduction CFS greedy. Performance of both algorithms is 98.50% and 98.45%. CNB is the highest performance in the two gram. Functionality of algorithm is 98.85% and categories error is 1.15% respectively. Finally the highest performance belongs to Bayes family algorithms CNB with 3 gram. Functionality of algorithms is 98.85% and categories error is 1.15% and there is no difference, but because of the complexity in 3 gram, two gram methods are recommended.

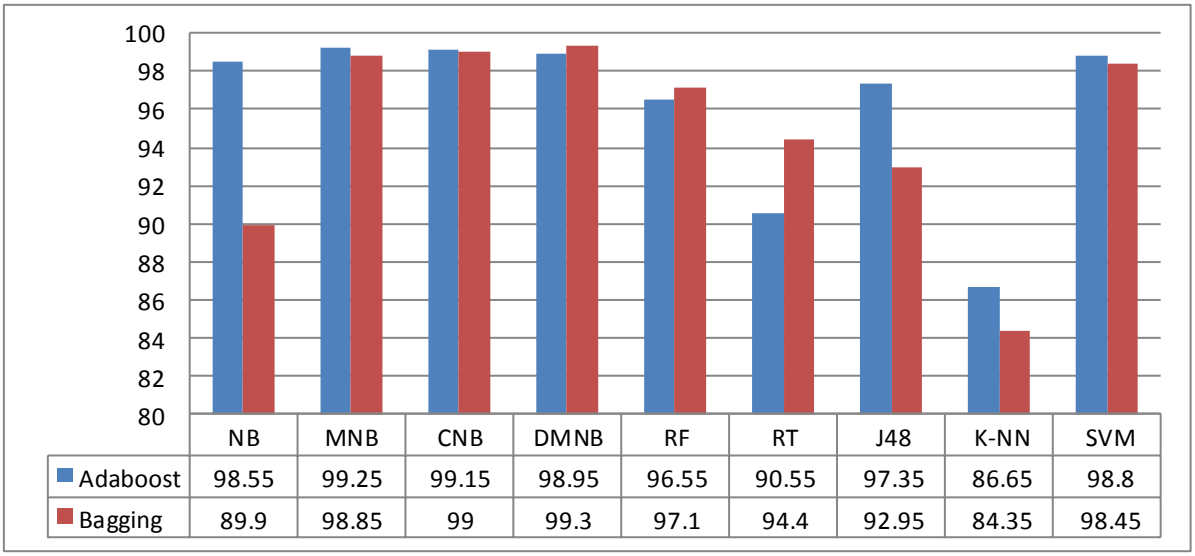


	NB	MNB	CNB	DMNB	RF	RT	J48	K-NN	SVM
1-gram	92.55	98.9	98.9	99.05	93.95	88.5	92.7	80.55	98.9
2-gram	91.2	99	98.95	99.2	94.4	90.25	92.3	82.15	98.85
3-gram	90.15	98.65	99	99.15	94.85	89.85	92.3	84.45	98.85

Figure 6: Comparison of 1 to 3 grams with IG method

In figure (6) highest performance of the algorithm belongs to DMNB method using 1 g and features reduced IG. The algorithm efficiency is 99.05% and the error is 0.95%. Compared with CFS methods we have higher efficiency with 1 gram. In two grams the highest performance is belong to DMNB. The algorithm efficiency is 99.20% and the error is 0.80%. Finally the highest performance belongs to DMNB method using 3 g. The algorithm efficiency is 99.15% and the error is 0.75%.Finally, by comparing the ways to reduce features it can be realized that IG method performance is better than CFS. In addition to this, two gram has the best results.

The following two gram method, feature reduction methods and combination methods were tested.



	NB	MNB	CNB	DMNB	RF	RT	J48	K-NN	SVM
Adaboost	98.55	99.25	99.15	98.95	96.55	90.55	97.35	86.65	98.8
Bagging	89.9	98.85	99	99.3	97.1	94.4	92.95	84.35	98.45

Figure 7: Two gram, reduction features IG and two hybrid approach

Figure 7 shows the performance with the three parameters: 1 gram to 3gram. As you can see the Bayes algorithm has the best efficiency and have a lower error rate. Best algorithm is DMNB with 2 gram and a Bagging combined method, with a 99.30% efficiency and margin of error is 0.70 percent, which is 0.10% performance better than the simple IG.

Conclusions:

Nowadays, there are many ways to remove designed spam. This method uses different techniques for the analysis legitimate email of illicit (spam).Of all the methods of spam filtering, machine learning methods can have a better performance with the best and highest performance in the classification of spam. This method does not require any specific rules. In this paper with standards and parameters of Data Recovery, we discuss the spam filtering function. Tests show that a learning content-based filter can function properly.

But there may be some messages incorrectly classified as spam, but these statistics are so less. Studies for more effective spam filtering and sorting can be done from a variety of filters and combinations of filters. This paper presents a method to classify spam based on machine learning. In this study, we aim to a spam detecting with high speed and high performance. This detection method can be efficient in information retrieval and search engine results.

References

Androutsopoulos, I., et al., 2000, an experimental comparison of naive bayesian and keyword-based antispam filtering with personal e-mail messages. Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, NewYork, NY, USA, pp 160-167.

Drucker, H., Wu, S., &Vapnik, V. N. (1999). Support vector machines for spam categorization. Neural Networks, IEEE Transactions on, 10(5), 1048-1054.

Carreras, X., & Marquez, L. (2001). Boosting trees for anti-spam email filtering.arXiv preprint cs/0109015.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32

Zhou, Y, (2007). Adaptive spam filtering using dynamic feature spaces. International Journal of Artificial Intelligence Tools, 16(4), 627–646.

Spam Abuse Corporation, <http: //spam.abuse.net/overview/whatisspam.shtml>, Visited in 2013.

Duda, R. O., Hart, P. E., & Stork, D. G. (2012). Pattern classification. John Wiley & Sons.

Liang, J., Yang, S., &Winstanley, A. (2008). Invariant optimal feature selection: A distance discriminant and feature ranking based solution. Pattern Recognition, 41(5), 1429-1439.

Cheng Hua Li, Jimmy Xiangji Huang, (2012), Spam filtering using semantic similarity approach and adaptive BPNN, Neurocomputing 92 , 88–97.

Salem, M. B., & Stolfo, S. J. (2010). Detecting masqueraders: A comparison of one-class bag-of-words user behavior modeling techniques. Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, 1(1), 3-13.

Abu-Nimeh, S., Nappa, D., Wang, X., & Nair, S. (2008, March). Bayesian additive regression trees-based spam detection for enhanced email privacy. InAvailability, Reliability and Security, 2008. ARES 08. Third International Conference on (pp. 1044-1051). IEEE.

Tretyakov, K., 2004, Machine Learning Techniques in Spam Filtering. Data Mining Problem-Oriented Seminar, pp 62-79.

McCallum, A., & Nigam, K. (1998, July). A comparison of event models for naive bayes text classification. In AAAI-98 workshop on learning for text categorization (Vol. 752, pp. 41-48).

Rennie, J. D., Shih, L., Teevan, J., & Karger, D. R. (2003, August). Tackling the poor assumptions of naive bayes text classifiers. In ICML (Vol. 3, pp. 616-623).

Yang, Y., & Liu, X. (1999, August). A re-examination of text categorization methods. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (pp. 42-49). ACM.

Zhang, T., & Oles, F. J. (2001). Text categorization based on regularized linear classification methods. Information retrieval, 4(1), 5-31.

D. Patterson, F. Liu, D. Turner, A. Concepcion, and R. Lynch.(2008) Performance Comparison of the Data Reduction System. Proceedings of the SPIE Symposium on Defense and Security, Orlando, FL.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I.H. Witten,(2009) "The WEKA Data Mining Software: An Update", ACM SIGKDD Explorations Newsletter, Vol(11), pp. 10-18.

Weka: Data Mining Software in Java http://www.cs.waikato.ac.nz/ml/weka

[Su, J., Zhang, H., Ling, C. X., & Matwin, S. (2008, July). Discriminative parameter learning for Bayesian networks. In Proceedings of the 25th international conference on Machine learning (pp. 1016-1023). ACM.

Jing, L. P., Huang, H. K., & Shi, H. B. (2002). Improved feature selection approach TFIDF in text mining. In Machine Learning and Cybernetics, 2002. Proceedings. 2002 International Conference on (Vol. 2, pp. 944-946). IEEE.

Shipp, C. A., & Kuncheva, L. I. (2002, January). An investigation into how adaboost affects classifier diversity. In Proceedings of 9th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (pp. 203-208).

Polikar, R. (2006). Ensemble based systems in decision making. Circuits and Systems Magazine, IEEE, 6(3), 21-45.