Performance Analysis of Email Classifiers for Detection of Spam

Zahra Masood, Javeria Khadim, Syed Khaldoon Khurshid
zm.masood@gmail.com , javeriakhadim@gmail.com , khaldoon@uet.edu.pk
Computer Science and Engineering Department, UET, Lahore

Abstract

Growing usage of email has also increased size of email data, this data involves important as well as undesirable emails. Amount of unwanted emails(spam) has increased enormously. Blocking spam sources doesn't works well in this era. For saving resources its vital to separate spam and essential emails(ham). Email servers are prepared to tackle this situation. Problem is handled by different algorithms that automate the system instead of manually separating emails. Our work addresses the selection of algorithm, whose outcome will precisely allocate labels to emails and will be efficient enough to give results in adequate time. So, that emails can be classified correctly into inbox and spam folders in adequate time by email server. Three different machine learning classifiers are analyzed over a dataset, providing a criterion that will categorize them according to their time, precision, recall and accuracy.

Keywords: Email, classifier, time, spam, ham, precision, recall, machine learning

Introduction

Irrelevant information in any form is never useful for anybody. Instead it's a burden that everyone desires to get rid of. As electronic e-mail is now used extensively for communication purposes whether official or unofficial, email data is getting huge and has a large variety now. For managing such massive amount of email, handling irrelevant and uninvited emails is essential. Now a day's email is being used as a source for accessing consumers for marketing purposes. Such emails are generated massively and can consume large amount of space in inbox. Preserving one's resources from such unwanted data is crucial. For this purpose, automated methods are introduced that separate such emails before reaching our inbox. This will save time and memory resources. Many techniques are introduced to distinguish between email that is relevant known as ham [1], from email that is irrelevant or annoying known as spam.

For establishing a difference between spam and ham, they should be distinguished in proper terms. And some rules should be defined that can classify emails into either of two categories. This rule defining approach is Knowledge Based Approach. It requires constant updating of rules to fulfill needs of new data. It's manual approach and requires frequent maintenance. Another better approach is the use of Machine Learning Techniques, which will practice data to train an algorithm for filtering and then use that trained algorithm to predict new data that will arrive. Machine learning approach is more efficient then prior one.

In our work, we considered three supervised learning classifiers that are Support Vector Machine (SVM), Naïve Bayes (NB) and Multinomial Naïve Bayes (MNB). Super Vector Machine (SVM) is off the shelf supervised machine learning classifier algorithm and it has very high-quality performance as it can classify the high dimensional data using kernels. It takes less time and give maximum accuracy [2]. SVM is the most extensively used technique that has many applications. It can be used to classify the hypertext, text and images. SVM cannot only do the linear classification but can also do non-linear classification competently. SVM construct hyperplanes in high dimensional space to perform classification. Hyperplane is used to achieve the separation in nearest training data of particular class can be known as functional margin [2].

Variables (feature vectors or training data) are placed in this high dimensional space and are separated nonlinearly in that space. The feature space (hyperplane) separates the positive and negative samples with greatest margin. To maintain the computational weight reasonable SVM scheme use mapping that are planned to make sure the easy computation of dot products of the variables in original space using the kernel function that suits the problem.

The representation of SVM model is like points in space and is mapped as separate categories having obvious gap that is as broad as possible. New values are placed in that space and then predicted that to which side category it belongs to, so it falls in that side of gap. SVM in this way is used to classify the ham and spam emails.

Naive Bayes is available in several versions. It is the most popular open source and commercial spam filter. There is the data set which is comparable it can be called the bag of words. This bag of words can be used to separate the messages into two categories spam and ham [3]. Naive Bayes is simple, can be implemented easily, its complexity and accuracy can be comparable to other algorithms. Naive Bayes is based on Bayesian algorithm it is used where input comes from different dimensions. In naive bayes classifiers different class labels are defined and certain features are mapped to each class. These features are independent from each other and can be represented in the form of vectors. Probability of each feature in particular class is calculated and then compared it with fixed value to separate the ham and spam emails.

In Naive bayes stemming words like preposition, conjunction, helping verbs etc. are ignored and features or attributes for the categorization of email are selected from message content. Triversing of bag of words occur for the feature at examined node . Probability of the feature in email is calculated to declare email as spam and ham.

Multinomial Naive Bayes (MNB) is the version of naive bayes that is specifically for the text documents classification. Simple naïve bayes classify the document on

https://doi.org/10.24949/njes.v12i2.262

the bases of presence and absence of any word while MNB perform the word count and make calculations to classify document. The count of any term in the document is called frequency of term [5]. Document is the sequence of words that belongs to vocabulary. In MNB document length is class independent and probability of every word in document is independent of its position in document and context of document [6].

MNB is the classification approach that builds the model ignoring the dependence of the features that's why no sequence, history and order is used in this model [7]. MNB performance can be improved by the effect of weighted normalization and complement based classification [8]. MNB is very efficient and accurate due to which it can be used as baseline in classification of text and research analysis.

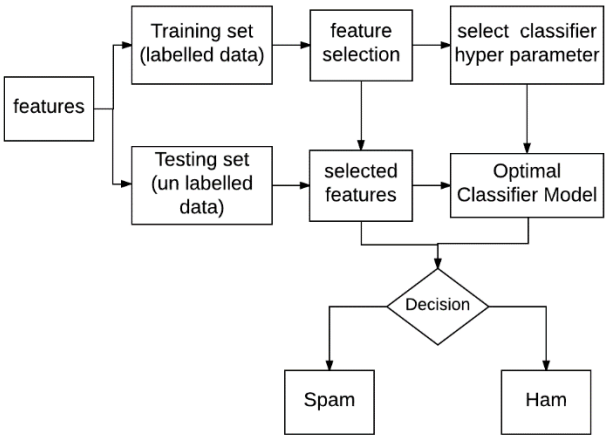

Fig 1: Data Flow Diagram for Classifier

Each Classifier works in sequence of some steps before it classifies documents into their classes according to their category. Figure 1 shows that a classifier uses labeled data for training and unlabeled data for testing, wether results match with actual labels or not.

This paper is divided in different sections, section II contains previous work done for classifying emails , section III explians dataset, that has been used in our work, section IV is about evaluation, calculations and results, section V concludes our work and proposed future work and last section has all references related to our research.

Literature Review

In 2014 the hybrid spam identification system was introduced in which the features of Artificial Immune System and Bayesian filters are combined to solve the problem faced by email users regarding spam emails. Artificial immune system follows the more flexible algorithm as compare to Bayesian filters, it has the ability to detect spam drift and can be trained easily in less time while on the other hand Bayesian filters are weak in spam drift detection and take more time to train. This hybrid system decomposes the email into two parts 1) header that contain IP address and mail address of sender along with number of receivers 2) body of email and filter them separately. This hybrid algorithm takes time to deliver emails and have much processing time. [9]

Many spam filtering techniques create clusters and labels for email classification then Lingo Algorithm was introduced in 2014 that is efficient for automatic clusters and labels generation. This proposed algorithm clusters the emails on the bases of its content such that the users can easily search their required email and make user friendly environment. [10]

Raghvendra and Johan proposed the Kernel Spectrum Clustering (KSC) as another clustering technique. The KSC model is built on subset of data with training, model selection and test phase. This clustering model is formed by the dual solution of problem. This algorithm is evaluated on the bases of precision and recall. Results are compared with neural gas and k-mean. KSC is good technique as it forms small size and homogeneous clusters. [11]

In 2013 Karthika and Visalakshi concluded that Support Vector Machine (SVM) is a good technique to classify the spam emails but it does not give much efficient results. To enhance the accuracy of SVM Latent Semantic Indexing (LSI) is used for feature extraction so the suitable feature selection will be done for email classification. Hybrid model of LSI and SVM can give more accurate and effective result. Term Frequency and Inverse Document Frequency features are extracted and LSI feature reduction techniques are used for feature selection them SVM classifiers are used for classifying the email as ham and spam [12].

LSI is effective for synonymy problem in email but LSI does not consider class label for training document due to which classification tasks are not very effective. To overcome this, sprinkling is used. Sprinkling is LSI extension based on features that are used for encoding class knowledge. Adaptive sprinkling is used for classifying the classes that are difficult to separate. Hybrid of LSI and Vector Space Model (VSM) enhance the performance of spam filtering. Combining the best characteristics of LSI like selection of word and high order co-occurrence with knowledge of relationships among classes the result is adapted for vector space representation. [13]

Izzat Alsmadi (2015) published the evaluation of clustering techniques in which large size dataset is collected and arranged for classification as spam email, subject and different folders classification. Different data mining techniques are utilized for clustering like parsing, stemming etc. The important challenge is to overcome large number of emails, different words used in emails and different formats used by different spammers. The challenge, faced by all clustering techniques is the use of many different terms and large amount of emails. Such large number of emails requires different folders to separate the emails [14].

In 2016 an international journal published efficiency evaluation of hybrid systems for email categorization. Several famous email classification methods are reviewed in this paper to evaluate the performance of machine learning classification systems. This paper suggests that the problems in Naive Bayes can be overcome by hybrid the other machine learning techniques with it. Hybrid of Naive bayes and Rough set has satisfactory performance [15].

With increased use of email as official announcements, appointments and business promotion. It is essential to classify email for quick and easy retrieval of required information. For shaping vast number of objects into concise meaningful groups clustering technique is used. Utilizing same technique for email categorization [refer] proposed SMTP that considers feature that is part of any two emails, feature that exists in only one email and the feature that is part of no email. Using Similarity Measure for Text Processing with k-means algorithm gives better result than other methods used for same purpose [16].

In 2016 L. Jiang and S. Wang proposed the Structure extended multinomial naïve bayes (SEMNB). This method removes the assumption of feature independence in MNB. SEMNB have simple algorithm without structure and searching. This algorithm has very good and accurate performance [17].

A.M. Kibriya, E. Frank, B. Pfahringer and G. Holmes performed the empirical comparison of MNB with SVM. They concluded that if data set is significantly large and time taken for classification do not matters then SVM is more acceptable as compare to MNB. [5]

In 2014 L. Jiang, C. Li and S. Wang proposed the correlation feature selection based feature weighted approach for naïve bayes classifiers. Their proposed solution reduces the assumption of feature independence by finding frequency of correlated features [18].

Dataset Specifications

Dataset utilized in our analysis is publicly available for experimental purposes. Federal Energy Regulatory Commission has dispatched it on web. It comprises of emails of employees that contain both ham and spam emails. This dataset is utilized widely in many forms. We have used classified form of set. In which each email is labeled as one of two labels. Thus, a standard is available against which each algorithm performance is evaluated.

Detail description of dataset is such that, there are six different folders having emails in .txt format, in a count of 5k to 6k. Each text file has email body and subject in it. No two folders contain same email in them. Every folder has emails from a different source. Thus, have different context. Each of six folders comprises of two labels of email. One is ham which has emails that have space in user's inbox. Second is spam type of email which comprises unwanted promotions or other objectionable emails, that one doesn't want to see in inbox. For experimental purpose, each folder is divided in two weightages. One has sixty percent of emails and other has forty percent. The first portion is used for training the classifier, the second portion is for testing it. For instance, a dataset containing 6000 emails, 3600 among them will be used to train the classifier and remaining 2400 will be used to test performance of test.

Methodology and Results

Each folder is passed from three classifiers. In first step text files are converted into `.arff ` format, that is attribute-relation format. Its output is one file having all emails in form of list, across each email is its assigned label, that is ham or spam. At this stage, there are two attributes which are labels of class (ham, spam). Each email is considered as string. Next step is to separate each word as a separate attribute. Such that frequency of each word in corpus can be accumulated. Fig 2 shows steps followed in our methodology in order to discover performance of each classifiers. For this purpose, file is passed through an unsupervised attribute function that will separate each word as a feature and maintain its count number, mean, standard deviation and minimum, maximum values. In this case as we are considering words it will have 0 as minimum which means doesn't exist and 1 as maximum that implies inexistence of word. No in between value will exist as shown in fig 3.

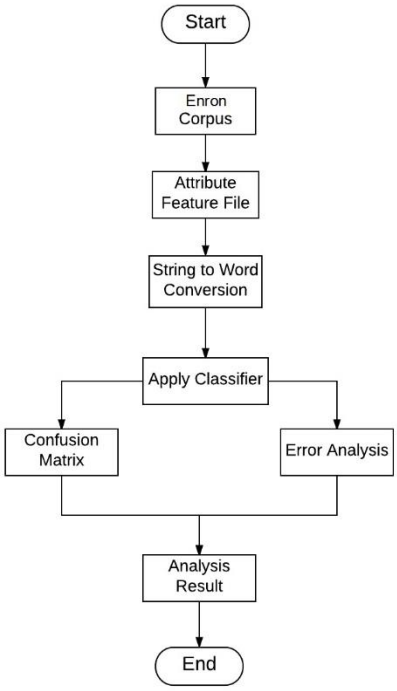

Fig 2: Classifier Evaluation Data Flow Diagram

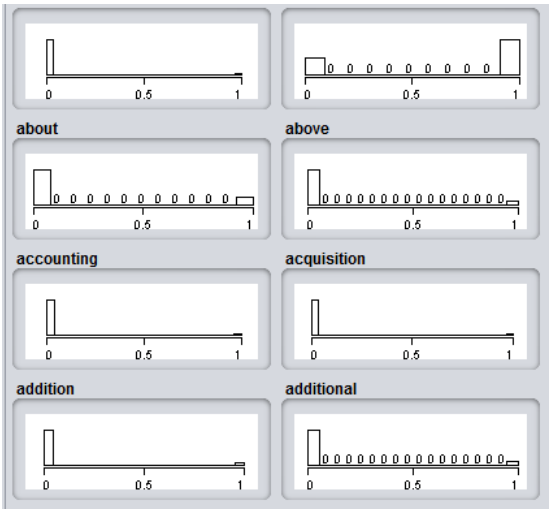

Fig 3: Standard Deviation, Minimum, Maximum Values shown for some words

Next step is to pass each corpus through a classifier. Results of each classifier will compute following values

- Precision
- Recall
- Time taken to build model
- Incorrectly Classified Instances
- Mean absolute error
- Root mean squared error
- Relative absolute error
- Root relative squared error

Precision and recall are elementary procedures for estimating performance of classifiers. Inverse relationship exists between these two measures.

Precision is defined as number of emails detected as spam are really spam. It is calculated as

$$\text{Precision} = \frac{N_{spam \rightarrow spam}}{N_{spam \rightarrow spam} + N_{ham \rightarrow spam}}$$

Recall is measure of correctly classified spam emails, computed as

$$\text{Recall} = \frac{N_{spam \rightarrow spam}}{N_{spam \rightarrow spam} + N_{spam \rightarrow ham}}$$

In these formulas, N is number of emails and their subscript show type of emails, spam->spam show emails that are correctly classified as spam, ham->ham are emails correctly classified as ham, spam->ham, ham->spam show incorrect classification of emails to other label instead of actual label specified according to ground truth [1]. Measuring Recall and Precision of each model is done through its confusion matrix. A confusion matrix marks result each sample of dataset's label from classifier across its original label known already from standard. From this matrix, these two parameters can be easily computed from their formulas.

In figure, each classifier is shown as a unique color, and each graph point shows a particular dataset. As recall graph shows that naïve bayes has values that range from 0.85 to 0.975. Which are lower than SVM and Multinomial NB. Analyzing upper plots show that both classifiers have almost equivalent values for each dataset.

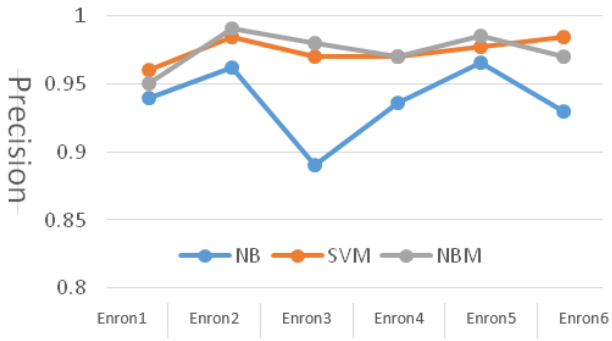

Fig 4: Precision Graph Plot for three Classifiers

In case of precision three classifiers perform in same manner as in recall. Naïve Bayes has lower values than other two classifiers, as it considers features to be independent of each other, while in real world it is not possible. Multinomial NB and SVM do not encounter this problem.

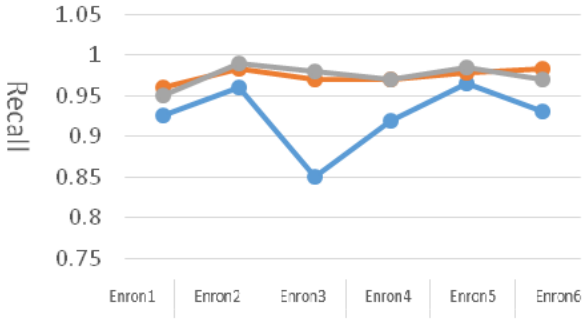

Fig 5: Recall Graph Plot for three Classifier

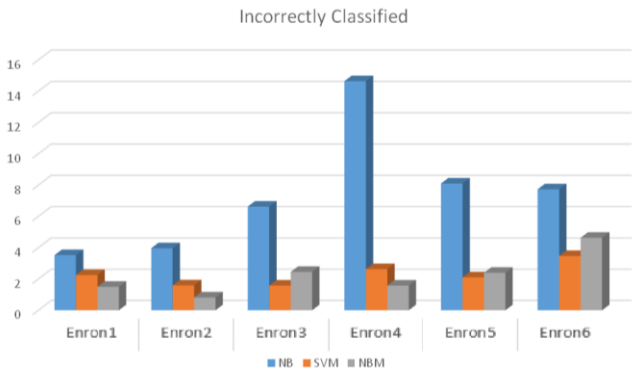

Fig 6: Graph for incorrectly classified emails by each classifier

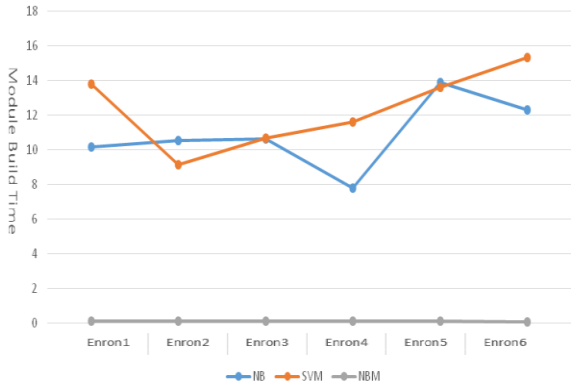

Fig 7: Graph shows time taken by each classifier to build model on training set.

Measuring time build for each model has shown diverse results as compared to previous calculations. Graph in Fig 7 show that Multinomial Naive Bayes takes least time to train itself for dataset. While SVM has almost equivalent performance results with Multinomial but time taken by it is comparatively higher.

Conclusion and Future Work

Here we conclude that, Support Vector Machine and Multinomial Naïve Bayes have almost equivalent results in classifying correctly. But time taken by Support Vector Machine is comparatively higher from Multinomial Naïve Bayes. It shows that NBM can be a better choice over SVM in order to increase throughput in classification of email as ham or spam. Analyzing each classifier across six corpuses give results about their performance. Results deduced from precision, recall and time taken for each build show that Multinomial Naive Bayes can be more advantageous than other two classifiers. In our analysis, it has occurred that for a dataset precision and recall has fallen in an abrupt manner. In extensive version of this work this aspect analysis will be helpful to judge that which features of corpus have caused this downfall of classifier.

REFERENCES

1. W. A. Awad and S. M. ELseuofi, "MACHINE LEARNING METHODS FOR SPAM E-MAIL," *International Journal of Computer Science & Information Technology (IJCSIT)*, vol. 3, no. 1, 2011.
2. V. P. Karthika Renuka D, "Latent Semantics Indexing Based SVM Model for Email Spam Classification," *Journal of Science and Industrial Research*, vol. 73, pp. 437-442, 2014.
3. T. M. N. Mirza, "An Evaluation on the Efficiency of Hybrid Feature Selection in Spam Email Classification," *International Journal of Advance Scientific Research and Engineering Trends*, vol. 1, no. 4, 2016.
4. K.Renuka , Visalakshi, "Latent Semantics Indexing Based SVM Model for Email Spam Classification", *NISCAIR-CSIR, JSIR* vol.73, no 07, 2014.
5. A. M. Kibriya, E Frank, B. Pfahringer, and G. Holmes, "Multinomial Naive Bayes for Text Categorization Revisited", *Australasian Joint Conference on Artificial Intelligence*, 2005
6. S. Raschka, "Naive Bayes and Text Classification", *Cornell University Libraray* 2014
7. S. Wanger, M. Zimmermann, E. Ntoutsi, M. Spiliopoulou, "Ageing-Based Multinomial Naive Bayes Classifiers Over Opinionated Data Streams", 2016.
8. L. Jiang, C. Li, "A CFS-Based Feature Weighting Approach to Naive Bayes Text Classifiers",2014
9. Smera, Rockey, Rekha, T. Sunny, "A Hybrid Spam Filtering Technique Using Bayesian Spam Filters and Artificial Immunity Spam Filters", *International Journal of Engineering Research & Technology (IJERT)*, Vol. 3 No. 5, 2014.
10. S. Maharana, M. Mohite and P. Wadekar, " Email Clustering Using Lingo Algorithm ", *International Journal of Computer Science Trends and Technology*, Vol. 2, No.6, 2014.
11. R. Mall, K. Arenberg, J. A.K. Suykens , "Kernel Spectral Document Clustering Using Unsupervised Precision-Recall Metrics.", *International Joint Conference on Neural Networks (IJCNN)*, 2015
12. K.Renuka , Visalakshi, "Latent Semantics Indexing Based SVM Model for Email Spam Classification", *NISCAIR-CSIR*, 2014
13. I. Alsmadi, I. Alhami, "Clustering and classification of email ", *Journal of King Saud University*, 2015.
14. V. Metsis, I. Androutsopoulos, G. Paliouras "Spam Filtering With Naïve Bayes" K.P.Murphy, "Naïve Bayes Classifiers" , 2006.

15. A. Bhowmick, M. Hazarika, "Machine Learning for E-mail Spam Filtering: Review, Techniques and Trends", 2016.
16. S.Pundalik Teli, S. Biradar, "Effective Email Classification for Spam and Non-spam", *International Journal of Advanced Research in Computer Science and Software Engineering*, vol 4, Issue 6, 2014
17. S. Wanger, M. Zimmermann, E. Ntoutsi, M. Spiliopoulou, "Ageing-Based Multinomial Naive Bayes Classifiers Over Opinionated Data Streams",2016.
18. L. Jiang, C. Li, "A CFS-Based Feature Weighting Approach to Naive Bayes Text Classifiers", 2014