ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

Survey Paper on Effective Email Classification into Spam and Non-Spam Mails

Ankita Mohite[1], Rupali Lokam[2], Simeen Wasta[3], Pooja Chikane[4], Snehal Mangale[5]

Students, Department of IT, RMCET, Devrukh, India [1,2,3,4]

Lecturer, Department of IT, RMCET, Devrukh, India[5]

ABSTRACT: Emails are used by number of users for educational purpose or professional purpose. But the spam mails causes serious problem for email users likes wasting of user's energy and wasting of searching time of users. This paper present as survey paper based on some popular classification technique to identify whether an email is spam and non-spam. For representing spam mails ,we use vector space model(VSM). Since there are so many different word in emails, and all classifier can not be handle such a high dimension ,only few powerful classification terms should be used. Other reason is that some of the terms may not have any standard meaning which may create confusion for classifier.

KEYWORDS: Spam and non-spam email filtering, K-Nearest Neighbour, Naïve Bayes, Stemming, Stop word removal, Vector Space model**.**

I. INTRODUCTION

Today ,emails are used for communication purpose by many users. Emails are broadly classified as spam mails and non-spam mails.First we will try to explain what is spam mails and non-spam mails and how affects on email user Spam is defined as bad emails and unwanted emails sent with the purpose of spreading viruses, for fraud in business and causing harm to email users. Non-spam mails are nothing but our regular emails which is useful for email users. According to the survey, today email users received spam emails than non-spam emails**.** In 1997 ,corporate networks received 10% emails are spam.The objective of email classification is to decide spam emails and not let them be delivered to email users. In document classification technique, document can be categorized into different predefined categories, have been applied to email classification with satisfactory result. In document classification, the document can be represented by vector space model(VSM)[1]. Each email can be represented into the vector space model ,i.e. each email is considered as a vector of word term. Since there are so many different word in emails, and all classifier can't be handle such a high dimension ,only few powerful classification terms should be used.

II. LITERATURE SURVEY

Tak-lam wong , Kai-on chow, Franzwong **(**19-22 August 2007**)**"[7] Incorporting keyword-based filtering to document classification for email spamming" contributed to "the research of email filtering to soften the hard clustering decision and also acheived the result of cost evaluation of ham to spam is better than spam to ham.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

Ren wang (IEEE CCECE/CCGEI, Ottawa, May 2006)"[10]On Some Feature Selection Strategies for Spam Filter Design "concluded that use of optimization techniques as feature selection strategies reduce the dimension of email as well as improve the performance of the classification filter.

III. METHODOLOGIES

A. Document Preprocessing:
Document pre-processing is the process of absorbing a new text document into text classification system.
Document preprocessing can be used for following purposes:

- Represent the document efficiently by removing useless keywords.
- Improve retrieval performance.

Document pre-processing consist of following stages:
a. Lexical analysis
b. Stop word elimination
c. Stemming

a. Lexical analysis
Lexical analyzer extracts keywords from text document by using tokenizer. It determines words from text documents[4]. Lexical analysis separates the input alphabet into characters (the letters a- z) and separators (space, newline, tab).
Lexical analysis removes digits, punctuation marks because these are useless for making decision in text classification.
b. Stop Word Elimination
In the context of text classification stop words referred as useless symbols. So it is important to remove these stop words from text document in order to improve the performance of text classifier. Stop words include articles, prepositions, conjunctions, pronouns and possibly some verbs, nouns, adverbs. Stop word elimination improves the size of the indexing structures.
c. Stemming
In information retrieval system morphological variants of words have similar semantic interpretations and can be considered as equivalent. For this purpose number of stemming Algorithms have been designed, which reduce a word to its root form. Thus, document is represented by stems rather than by the original words which helps to reduce dictionary size. The meaning of "writing", "written", "write" and "write" is same in context of information retrieval system. A stemming algorithm reduces the words "writing", "written", and "write" to the root word, "write".

B. Weighting Scheme:

tf–idf, short for term frequency–inverse document frequency used as weighting factor in text mining. It reflects how word is important to a document. The value of tf-idf rising as the number of times a word appears in the document.
term frequency–inverse document frequency is combination of two terms:

- Term Frequency

The term frequency is a concept which can be defined as number of occurrences of the term t_i within particular documents d_j.

$$tf_{i,j} = n_{i,j}$$

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

Where $n_{i,j}$ is number of occurrence of term t_i in document d_j.
To prevent bias for larger documents, term frequency often normalized equation(1) as shown below,

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (1)$$

- Inverse Document Frequency

The inverse document frequency is defined as total number of documents divide by number of documents containing the term t_i and taking logarithm of quotient as shown in equation(2) below,

$$idf_{i=\log} \frac{|D|}{|\{ d : t_i \in d \}|} \quad (2)$$

To avoid divide by zero error, we can use $1+|\{ d : t_i \in d \}|$.
C. Data set:

It is very difficult to collect non-spam emails because of protecting personal privacy. Therefore we collect dataset which is available on www.csmining.org/index.php/pu1-and-pu123a-datasets.html pu1corpus [1].
This corpus consists of total 1099 emails ,out of which 481 are spam emails and 618 are non-spam emails. The emails in pu1 corpus only have subject name and email body text ,header fields and HTML tags removed.

IV. CLASSIFICATION ALGORITHM

A. Naïve Bayesian Classification:

Naïve Bayesian classifier is defined by a set C of classes and a set A of attributes. A generic class belonging to C is denoted by c_j and a generic attribute belonging to A as A_i.
Consider a database D with a set of attribute values and the class label of the case.The training of the Bayesian Classifier consists of the estimation of the conditional probability distribution of each attribute, given the class[5].
Let $n(a_{ik}|c_j)$ be the number of cases in which A_i appears with value a_{ik} and the class is c_j.
Then $p(a_{ik}|c_j) = n(a_{ik}|c_j)/\Sigma n(a_{ik}|c_j)$ Also $p(c_j) = n(c_j)/n$.
This is only an estimate based on frequency. To incorporate our prior belief about $p(a_{ik}|c_j)$ we add α_j imaginary cases with class c_j of which α_{jk} is the number of imaginary cases in which A_i appears with value a_{ik} and the class is c_j.
Thus $p(a_{ik}|c_j) = (\alpha_{jk} + n(a_{ik}|c_j))/(\alpha_j + n(c_j))$
Also $p(c_j) = (\alpha_j + n(c_j))/(\alpha + n)$ where α is the prior global precision.
Once the training (estimation of the conditional probability distribution of each attribute, given the class) is complete we can classify new cases.
To find $p(c_j|e_k)$ we begin by calculating
$p(c_j|a_{1k}) = p(a_{1k}|c_j)p(c_j)/\Sigma p(a_{1k}|c_h)\ p(c_h)$
$p(c_j|a_{1k}, a_{2k}) = p(a_{2k}|c_j)p(c_j|a_{1k})/\Sigma p(a_{2k}|c_h)\ p(c_h|a_{1k})$ and so on.

B. Decision Trees:

Decision trees also known as classification trees. It learns from set of independent instances by applying 'Divide and conquer" approach. Decision trees are designed its node contain attribute test conditions to classify instances which have different characteristics[6]. Decision trees branches leads to those classification and leaf node represent respective class. Constructing optimal decision trees is an NP complete problem; heuristics are used for constructing optimal trees. It select those features that best divide the training data to partition the records into smaller subsets. The important step is how to determine which feature to split on. There are different feature

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

evaluations techniques are available to find for optimal splitting feature. These techniques are derived from information theory and most of them are based on Shannon's entropy.

Information gain is one of technique which can also be applied as feature ranking method. It is one of most widely used splitting criteria. By expanding tree nodes that contribute to largest gain in average maximize the global mutual information. Rules derived from distance measures calculate the separability and discrimination between classes. Gini index of diversity one of the popular distance measures, a measure of the inequality of a distribution, the Kolmogorov-Smirnov distance.

The decision tree algorithm works as below: First, it selects an attribute to create root node and create one branch for each value of this attribute. This divides the training set into subsets, one for every value of the attribute. Then, it repeats the process recursively for each branch. If at any particular time all records at a node have the same classification then stop developing that part of the tree.

C. K-Nearest Neighbor Algorithm:

The k-nearest neighbor (KNN) algorithm belongs to category of instance-based learners which is simple and one of important machine learning algorithms. Instance-based learners are also called lazy learner algorithm because it delays actual generalization process until classification is performed. There is no model building process. Instance-based learners do not abstract any information from the training data during the learning phase. Learning is merely a question of encapsulating the training data.

KNN is works based on principle that instances within dataset generally exist in close proximity to other instances within dataset that have similar properties. If the objects are tagged with a classification label then objects are classified by a taking majority vote of their neighbors and it assigns to most common class amongst its k-nearest neighbors[2]. K is small odd positive number and correct classification is prior known. The objects can be n-dimensional points within n-dimensional instance space where each point corresponds to one of the n features which describe objects. The distance of object is calculated by using distance metric, for example the Manhattan distance or the Euclidean distance[7] .KNN is highly susceptible to noise in the training data due to high degree of local sensitivity. Thus the value of K influences the performance of KNN algorithm. The optimal choice of k is a problematic issue, but cross validation can be used to reveal optimal value of k for objects within training set.

V. **CONCLUSION**

In the proposed system, we consider the requirements of improving the efficiency, accuracy of data mining classification technique like Naïve Bayesian, Decision Tree, K-Nearest neighbor, which is good data mining algorithms.

VI. **ACKNOWLEDGEMENT**

It is an opportunity of immense pleasure for us to present the paper " Evaluation Of Classification Techniques for Email Classification As Spam and non-Spam "expressing our heart left gratitude to all those who have generously offered their valuable suggestions towards the completion of the paper. The credit goes to our Prof. Mangale S.R. (RMCET, Ambav, Ratnagiri) whose positive attitude; moral support and encouragement lead to the success of the paper.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

REFERENCES

[1] RenWangI, Amr M. Youssef , Ahmed K. Elhakee "On Some Feature Selection Strategies forSpam Filter Design" 1- 4244-0038-4 2006 IEEE CCECE/CCGEI Ottawa, May 2006
[2] T.M. Cover, P. E. Hart, "Nearest Neighbor Pattern Classification". Knowledge Based Systems, 1995.
[3] C. Lai and M. Tsai, "An empirical performance comparison of machine learning methods for spam email categorization," Proceeding of the 4th international conference on hybrid intelligent systems (HIS'04), 2004.
[4] J. F. Pang, D. Bu and S. Bai , "Research and Implementation of Text Categorization System Based on VSM,"Application Research of Computers, 2001.
[5] D. Vira, P. Raja, and S. Gada,"An Approach to Email Classification using Bayesian Theorem"‖ GJCST. (USA),vol. 12, Issue 13, ver. 1.0, 2012.
[6] J. Han, M. Kamber, and J. Pei, (2011,July 6) "Data Mining Concepts and Techniques" (3rd ed.). The Morgan Kaufmann Series in Data Management Systems.
[7] TAK-LAM WONG , KAI-ON CHOW, FRANZWONG" Incorporting keyword-based filtering to document classification for email spamming" Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, Hong Kong, 19-22 August 2007.

BIOGRAPHY

Prof. Mangale S.R. I have completed Bachelor of Computer Engineering. I am interested in Software Project Management and Data mining.

Ms.Ankita Mohite.I am is pursuing my B.E. in Information Technology, RMCET, Mumbai University. I am a member of ISTE. My area of interest is networking and Security.

Ms.Rupali Lokam.I am is pursuing my B.E. in Information Technology, RMCET, Mumbai University. I am a member of ISTE. My area of interest is networking and Security.

Ms.Simeen Wasta .I am is pursuing my B.E. in Information Technology, RMCET, Mumbai University. I am a member of ISTE. My area of interest is networking and Security.

Ms.Pooja Chikane.I am is pursuing my B.E. in Information Technology, RMCET, Mumbai University. I am a member of ISTE. My area of interest is networking and Security.