



Social Networking Spam Detection using R Package and K-Nearest Neighbor Classification

Dr. Sanjeev Dhawan

Faculty of Computer Science & Engineering
University Institute of Engineering and Technology
Kurukshetra University, Kurukshetra, Haryana, INDIA

Jyotika Verma

Student of M.Tech. Software Engineering
University Institute of Engineering and Technology
Kurukshetra University, Kurukshetra, Haryana, INDIA

Abstract: Spam is one of the problems which are faced on Social Networking Websites. Everyone today uses the internet very much and hence, everyone is familiar with the spam too. Spam is nothing but the mails or more appropriately it is defined as junk mails which create hindrance in doing the work. Various techniques like Decision Tree, Naïve Bayesian etc are available to deal with this problem. Here, in this paper a spam detection system has been introduced involving the use of the Clustered KNN approach because of which great accuracy has been achieved and that too in no time.

Keywords: Spam; Social Networks; Preprocessing; KNN; R Package; Feature Extraction

I. Introduction

Social Networking Sites are very much commonly used in today's life. If there are some advantages of these sites along with them there are some disadvantages too. Social Networking Sites is one of the services which are provided by the Internet. But there are certain issues also which are related to this like Phishing, Cracking, Spoofing and the most common is spamming. Spam is the problem which is faced by every single person who is using the Internet. Spam is an unsolicited bulk email messages which creates disturbance in the network [8]. So, detection of this type of problem was so much important and hence, various techniques like KNN, Naïve Bayesian, SVM, GA, AIS etc are used for the detection of the spam. Some of the techniques are good in providing the results with great accuracy but their work takes long time to complete these kinds of problems which have been encountered in most of the techniques. So, here in this paper the system which is introduced does its work in less time and with great accuracy. The classifier which is used to perform this work is KNN because it is considered as one of the most used classification process which gives the accurate result but it is a time consuming process so, to avoid this problem a clustered approach is used.

II. Literature Review

Internet is referred to as one of the mostly used ways for communicating with the other person whether that person is living in your city, state or at any other location around the globe. But, everything on this earth comes with the advantages and with several disadvantages also. Same is the case with Internet, if Internet have provided us the services like email, blogs etc. they also come with several problems like spam. Spam is the problem which is faced by every individual working on the Internet. The present literature survey concentrates on the work done by a number of researchers Worldwide in the field of spam detection both in e-mail systems and social networking Websites. This literature survey is carried out in chronological order and is presented from year 1998 to 2014 as under. Several work has been performed by *Naïve Bayes* for filtering the spam out of ham which was basically used for the text classification [1]. The first program which has used the Bayes classifier was of Jason Rennie's *ifile* program which was introduced in 1996[1]. The very first publication about the Bayesian work was done by Sahami *et al.* in 1998 [2]. Their method made it possible that the spam filters automatically learn to separate the junk email out of the legitimate emails and they have also improved the efficiency of the filters by providing the domain specific feature they the filters didn't have to read the whole text but with the help of certain words the mail could be classified as spam or ham. Spam has caused very much problem in our day to day life as it doesn't allow us to do our work efficiently. The spam always fills up the user's account inbox and because of this the legitimate emails doesn't get the chance to get addressed because sometimes with the purpose of getting this spam email out of the inbox we by chance delete the legitimate mails too. And the second problem related to the spam is sometimes there is a case when we encounter a large amount of spam in our email account and hence it takes a lot of time to get rid of those spam messages. There also chances that spam messages may contain the malicious content which may crash the user's system. It is

experienced that the situation in China and America is worse than 50% of the mail which have been received or more than that is spam [3]. Federal Trade Commission is U.S commission which works for enhancing the security among the users and various economies. It address the various privacy issues including spam, social networking, peer to peer, mobile etc and all these matters include over 130 spam and spyware cases [4]. The mechanism of the spam detection system which is based on the intelligent cognition of the email user interest within life and work is discussed under the guidance of the cognitive pattern recognition [5]. Later on, Kunal Jain *et al.* [6] have then proposed the technique for the classification of the spam which was based on the combination of the local concentration based feature extraction method with k-means clustering. They have performed the experiments on the PU Corpora which was in a controlled environment and have concluded that there proposed system was better than the existing one. They have also provided the future scope that the performed work can be made better by applying the concept of the multiclass classification.

III. Spam Detection Techniques

There are various spam detection techniques but broadly they are classified in two categories and these are supervised and unsupervised spam detection techniques. These two techniques differ from each other in many ways but the one important difference between them is that supervised techniques work with the trained dataset whereas the unsupervised techniques always work with the untrained dataset [9]. Supervised techniques are used for the two purposes i.e., classification and regression. KNN, SVM, Naïve Bayesian all comes in this supervised category. An unsupervised spam detection technique involves Clustering (e.g., k-means, mixture models, hierarchical clustering). The approach which is used in this paper is Clustered KNN approach. However, Clustering is an unsupervised learning algorithm but as it is used with the KNN approach so, it became a supervised training algorithm which is used in the implementation of the spam detection system [10].

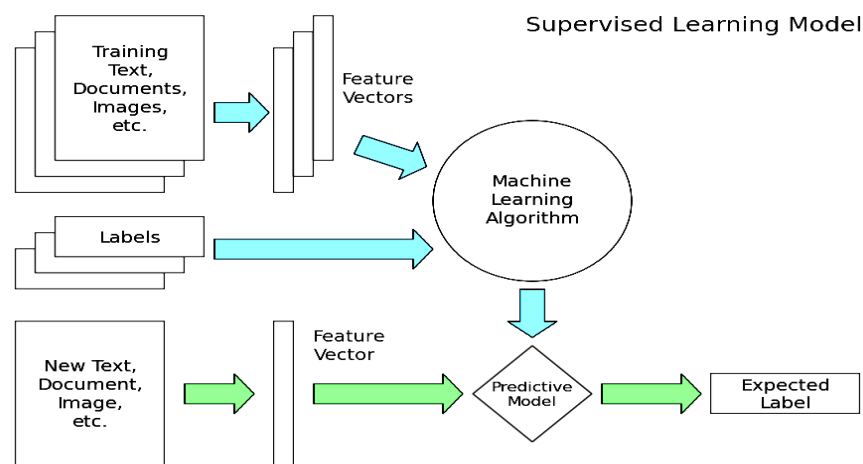


Figure 1 Supervised Learning Model [9]

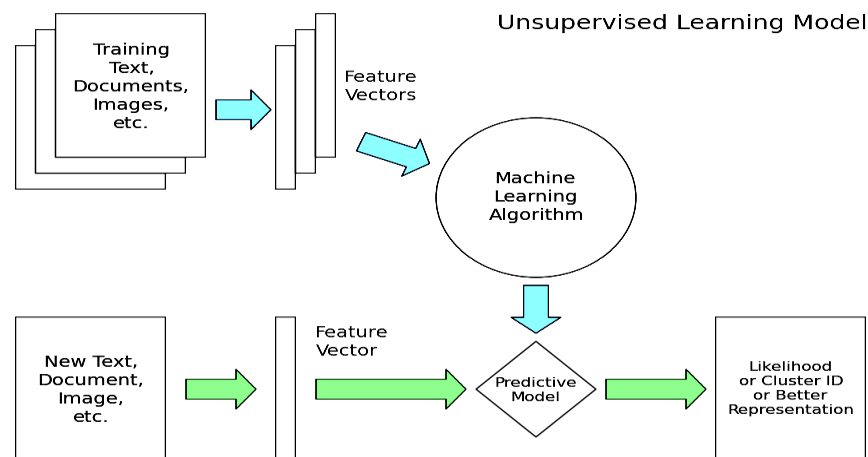


Figure 2 Unsupervised Learning Model [10]

IV. R package

R package is an environment which is used for data manipulation, graphical display and calculation. It is also defined as the integrated suite of facilities [11]. Data Analysis, Data Handling and Storage Facility can be done effectively in this package. It is a coherent systematic and planned system. It is a simple and effective Programming Language called “S” Language. It was developed at the Bell Laboratories by Robert Becker, John Chamber and Allan Wilks [11]. It is also referred to as one of the Mining tool. It can easily work with any Operating System like Windows, Linux, Mac OS. There are various functions which can be performed with the R package like Classification, Clustering, Save and Load R data, Text Mining. Different functions are used for each of the process for example if one want to use the classification process in Decision Trees then `rpart` and `party` functions are used , to use the classification process in Random Forest `randomForest` and `Party` functions are used [12]. For performing the clustering with R in DBSCAN `fpc` function is used and in BIRCH `birch` function is used. The data in R can be saved as `.Rdatafiles` with the `save()` function and then the loading of that data is performed with the help of `load()` function. In this research work R package has been used with the purpose of extracting the tweets from the Twitter as R package provides interface with the Twitter Web API.

V. Implementation

The main aim of this implementation is to build a system which is time efficient, less complex and gives high accuracy. The process which is followed to implement this system is to first extract the data with the help of the R package, then performing the preprocessing which involves further processes named stop word removal and stemming, then feature extraction is performed and then the classifier is applied to the trained clustered data [7]. R package is used for maintaining the dataset on which the further work is performed and the dataset is a collection of the tweets which are extracted from the Twitter with the help of R as it provides an interface with Twitter web API. Preprocessing of the extracted data is performed involving the steps stop word removal and stemming. Stop word removal is referred to as the process of removing the unnecessary words which are increasing the complexity of the document nothing else [13] and later on stemming is performed. Stemming is referred to as the process of finding out the root or the stem of the word so that the words with the similar root or stem can be identified easily. These both the steps are basically information retrieval systems. Feature Extraction is the next step which is to be performed and the various factors like `scount` (no. of spam words), `wordcount` (total number of words), `url/wordcount` (number of url per wordcount), `url` (total number of url), `rt` (retweets) are calculated. After this when the data gets trained then that trained data is provided as input to the clustering process. It is the process of selecting the objects and arranging them in a group in such a way that they are more similar to each other rather than with other groups. After all this work, KNN classifier is applied whose work is to give the accurate result and the problem related to the KNN that it is a time consuming process when the data set is large is now resolved by using the Clustering approach and hence, it is said as Clustered KNN Approach.

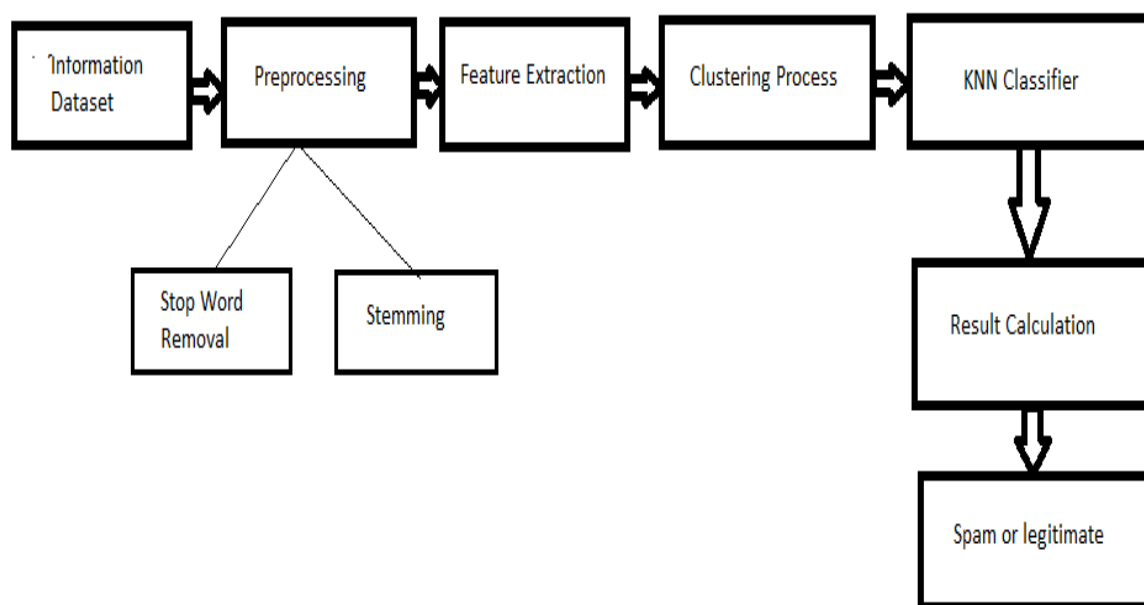
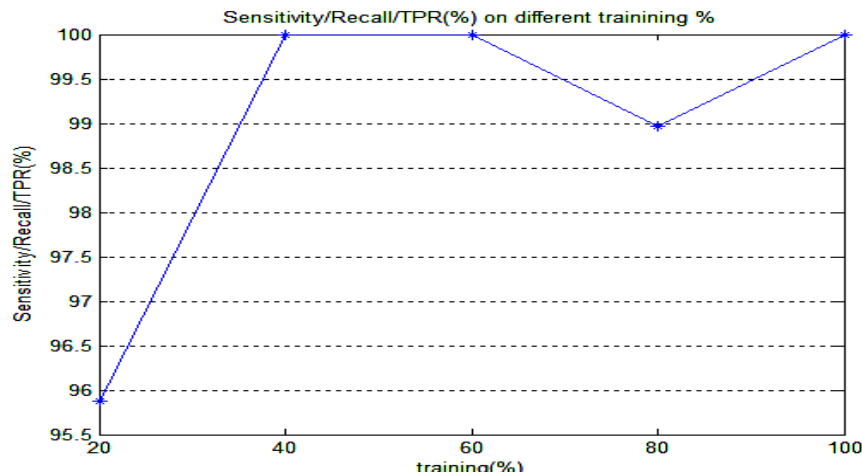


Figure 3 An Overview of the Proposed System.

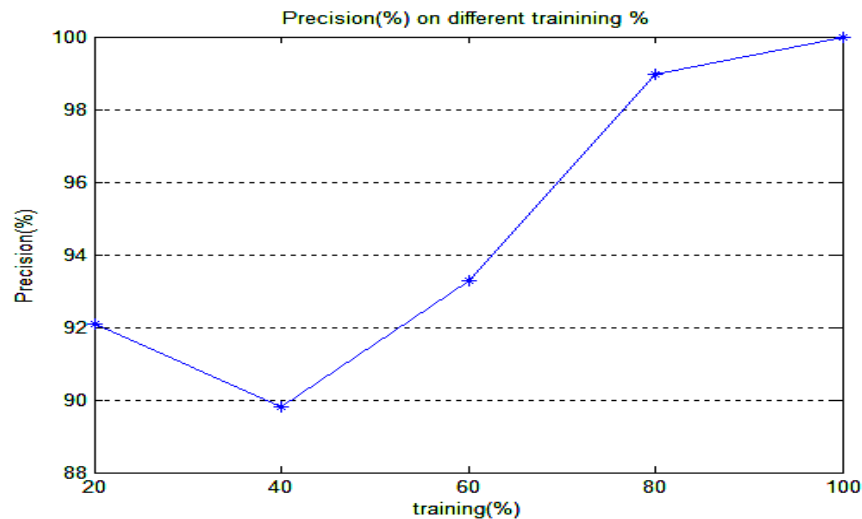
VI. Results and Analysis

Based on the proposed model above implementation is performed and on the basis of which various results are calculated and all these metrics are calculated on 20% 40% 60% 80% trained data.

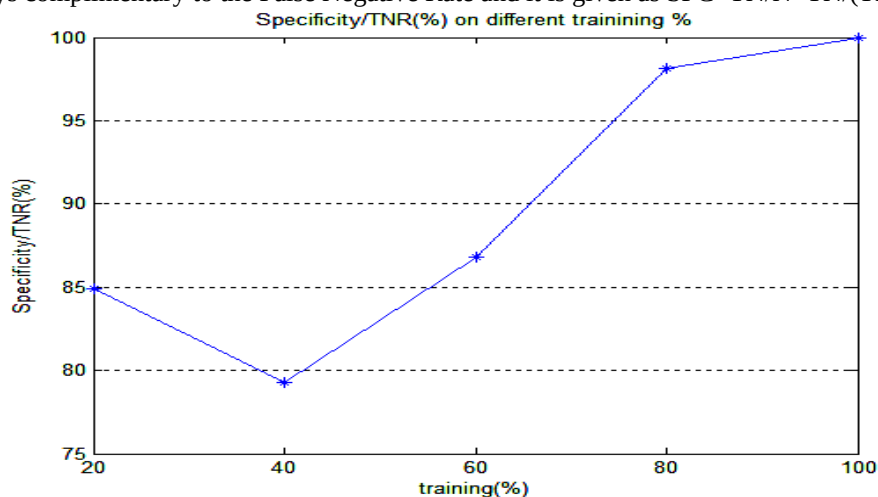
- Sensitivity/Recall: It is a fraction of the documents that are relevant to the query that are successfully retrieved. In binary classification it is called as sensitivity and formula is $TPR = TP/P = TP/(TP+FN)$ [14].



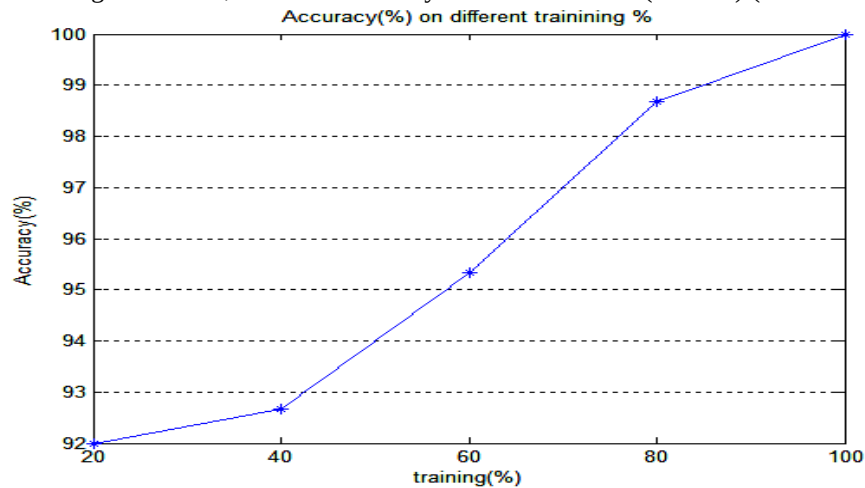
- Precision: It is defined as list of the documents available on the Internet related to a certain topic and it is calculated as $PPV = TP/(TP+FP)$ [15].



- Specificity: The proportions of negatives which are correctively identified get measured here and it is always complimentary to the False Negative Rate and it is given as $SPC = TN/N = TN/(TN+FP)$ [16].



- Accuracy: It is a statistical measure of how well a classification test correctly identifies or excludes a condition. If the value of the accuracy is 100% then that means the measured values are exactly the same as of the given values, it is determined by the formula $ACC = (TP+TN)/(TP+FP+TN+FN)$ [17].



VII. Conclusion

The implemented work results in the improvement of the accuracy and time of the classification process and hence, the work of spam detection can be done easily but the features which are identified here are just related to the spam data like no. of URL in the tweet or number of spam words etc. The future researcher can do the work of spam detection by including the features like user's friend, frequent links etc or can introduce an entire new methodology by using any other classification technique.

References

- [1] Wikipedia, "Naïve Bayes spam filtering," 2014, http://en.wikipedia.org/wiki/Naive_Bayes_spam_filtering. Accessed on 29-January-2015
- [2] Meheran Sahami, Dumai, S., Heckman, D., Horvitz, E. "A Bayesian Approach to Filtering Junk E-Mail," AAAI'98 Workshop on Learning for Text Categorization, Madison, Wisconsin, July 1998.
- [3] Nello Cristianini, John Shawe – Taylor "An Introduction to Support Vector Machines and Other Kernel Based Learning Methods", Cambridge University, 2001, Vol. 22, pp. 103-104.
- [4] Federal Trade Commission, USA: False Claims in Spam, April 30,
- [5] Tiancai Liang, Younguo Pi "On Spam Detection based on Cognitive Pattern Recognition," IEEE International Conference on Computational Intelligence and Security Workshops, Harbin, 15-19 December 2007, pp. 136-139. M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.
- [6] Kunal Jain, Sanjay Aggarwal "A Hybrid Approach for Spam Filtering using Local concentration based K-means Clustering", IEEE 5th International Conference on Confluence The Next Generation Information Technology Summit (Confluence), Noida, 25-26 September 2014, pp. 194-199
- [7] Jyotika Verma, Dr. Sanjeev Dhawan, "Detection of Spam in Social Networks using Clustered K-Nearest Neighbor approach", International Journal of Advanced Research in Computer Science and Software Engineering, 2015, Vol. 5, Issue 3, March 2015, pp. 1120-1123.
- [8] Oxford Dictionary, "Definition of spam", <http://www.oxforddictionaries.com/definition/english/spam> . Accessed on 29-January-2015
- [9] Wikipedia, "Supervised Learning Model", 2015, https://en.wikipedia.org/wiki/Supervised_learning. Accessed on 7-August-2015.
- [10] Wikipedia, "Unsupervised Learning Model", 2015, http://en.wikipedia.org/wiki/Unsupervised_Learning. Accessed on 7-August-2015.
- [11] Rdatamining, "Regression and Classification with r", 2015, <http://rdatamining.com/docs/regression-and-classification-with-r>. Accessed on 7-August-2015
- [12] CRAN, "R environment", 2015, <https://cran.r-project.org/doc/mauals/r-release/R-intro.html#The-R-environment> . Accessed on 7-August-2015.
- [13] Wikipedia, "Stop Words", 2015, https://en.wikipedia.org/wiki/Stop_words . Accessed on 7-August-2015.
- [14] Specificity, https://en.wikipedia.org/wiki/Sensitivity_and_specificity. Accessed on 8-July-2015.
- [15] Precision, https://en.wikipedia.org/wiki/Precision_and_recall. Accessed on 8-July-2015.
- [16] Sensitivity, https://en.wikipedia.org/wiki/Sensitivity_and_specificity. Accessed on 8-July-2015
- [17] Accuracy, https://en.wikipedia.org/wiki/Accuracy_and_precision. Accessed on 8-July-2015