Spam Detection using KNN

Sunidhi Bansal¹, Kanwal Garg²
¹Scholar, M.Tech, CSE,  
²Supervisor, Assistant Professor,  
Department of Computer Science and Application, Kurukshetra University, Kurukshetra

Abstract- Social networking sites have become part of life for most of the people today. Among all OSNs twitter is one of the most used and powerful way of communication and news source. With twitter growth spamming activities in it has also increased. There is a need for more accurate but efficient spam detection methods to avoid causing inconvenience to legitimate users. This paper presents the implementation of KNN algorithm for spam detection marking tweets as spam or non-spam and experiment is done with different training percentages. Performance evaluation of KNN is also done using different standards like execution time, accuracy, sensitivity, specificity, precision, recall, f-measure, g-mean. The results show that KNN provides good accuracy even when less percent data is trained and it increases more when training percentage is taken high.

Index Terms- Spam Detection, KNN, Feature selection.

I. INTRODUCTION

Social networking sites have become part of life for most of the people today. Online social networking sites (OCNs) have grown both in size as well as in popularity in recent years. People spend a significant amount of time on social networking sites interacting with other people. Among all OSNs twitter is one of the most used and powerful way of communication and news source. Twitter is a micro-blogging site that allows registered members to send and read short messages of about 140 characters called tweets. With twitter growth spamming activities in it has also increased. Spam is irrelevant or unsolicited messages sent over the Internet, typically to large numbers of users, for the purposes of advertising, phishing, spreading malware, etc. Unfortunately, social networking sites do not provide strong authentication mechanisms, and it is easy to impersonate a user and sneak into a person’s network of trust [2]. Twitter itself does not prevent spamming, Twitter relies on users to report spam. Once a report is filed, Twitter investigates it to decide to suspend an account or not [1]. The accounts that are found suspicious are deleted by twitter but some users complain that their account is mistakenly deleted.

Detection of spam has become a challenger task for researchers as well as for Twitter itself. Twitter spam detection consist both the varieties of detecting spammers and detecting spam links which is posted by the users. We need some tools that can automatically identify spammers. In addition, we need more accurate but efficient spam detection methods to avoid causing inconvenience to legitimate users. To classify tweets as spam or non-spam machine learning methods can be applied. Classification algorithms perform classification in two phases training and testing. In first step, a classifier is built describing a predefined set of data classes or concepts. It is a leaning step. Second step is of using the classifier for classification. Here the test data is used to estimate the accuracy of classification rules. The classification rules can be applied to the new data tuples if the accuracy is considered acceptable.

II. RELATED WORK

Spam detection has been studied for a long time. The previous work mainly focuses on email spam detection and Web spam detection. M. Klasssen [1] explored attribute reduction and data pre-processing such as data normalization and discretization from the aspect of Twitter spam detection using various machine learning algorithms. Results show that when top 24 attributes were selected and used for these classifiers, the overall classification rates obtained were very close in range 84.30% and 89%. G. Stringhini et al. [2] developed techniques to detect spammers in social networks, and aggregated their messages in large spam campaigns. Results show that
it is possible to automatically identify the accounts used by spammers, and analysis was used for take-down efforts in a real-world social network. More precisely, during this study, author collaborated with Twitter and correctly detected and deleted 15,857 spam profiles. Z. Chu et al. [3] proposed a classification system which detects spam campaigns that manipulate multiple accounts to spread spam on Twitter. Complementary to conventional detection methods, proposed system brings efficiency and robustness.

M. McCord and M. Chuah[4] compared the various classifiers for detecting spams. Results show that among the four classifiers evaluated, the Random Forest classifier produces the best results. Results based on the 100 most recent tweets also show that spam detection based on author’s suggested features achieved 95.7% precision and 95.7% f-measure using the Random Forest classifier. A. Kumar et al. [5] describe a hierarchy of spamming techniques, defense approaches and evasion tactics adopted by the spammers to evade detectors.

III. MACHINE LEARNING APPROACH (KNN)

KNN is among the simplest of all machine learning algorithms. KNN a non-parametric lazy algorithm delay the process of modeling training data until they need it to classify new data. A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function. If K = 1, then the case is simply assigned to the class of its nearest neighbor. A Euclidean Distance measure is used to calculate how close each member of the training set is to the test class that is being examined [2]. say, X=(x1,x2,……,xn) and Y=(y1,y2,……,yn), is

Detection of spam has become a challenger task for researchers as well as for Twitter itself. In this paper accuracy of KNN algorithm is evaluated in classifying tweets as spam or non-spam. Machine learning algorithms cannot be directly applied to tweets. Several steps need to be followed before applying any machine learning algorithms. These steps are described in next section.

IV. SPAM DETECTION USING KNN

Steps for spam detection include data reading and preprocessing, feature extraction, binary classification, evolution of results obtained. Figure shows the process involved in spam detection.

**Reading/Preprocessing:** The data (tweets) that are gathered are in unstructured form. There is much irrelevant and redundant information or noisy and unreliable data present which do not contribute to detection of spam. Data must be preprocessed in order to perform better data mining functionality.

Data Preprocessing involves the following tasks:

- **Stop words removal:** Tweets contain prepositions, articles, and pro-nouns which are of no use in detection of spam. These words are treated as stop words. These words need to be eliminated. Any group of words can be chosen as the stop words for a given purpose. This process also reduces the text data and improves the system performance. Example are ‘the’, ‘in’, ‘a’, ‘an’, ‘with’ etc.

- **Special characters removal:** Special characters like (), {}, [], / needs to be removed for better performance of classifier.

- **Stemming:** Stemming or lemmatization is a technique for the reduction of words into their root. Many words in the English language can be reduced to their base form.
or stem e.g. agreed, agreeing, disagree, agreement and disagreement belong to agree. In this we applied standard Porter Stemming Algorithm for find the root words in the document.

- Upper/lowercase conversion: Case sensitive systems could have problems when making a comparison between a word in capital letters and another with the same meaning in lower case. So all the words which are in uppercase are converted to lower case.

**Feature reduction:** Next step in spam detection is feature extraction. The features extracted are used to facilitate spam detection. A feature matrix is created by using these features. Features that are used are given below.

- Spam words: There are certain words which are mostly used by the spammers to attract twitter users by advertising about some product or offering them with some scheme which make them click the links provided with those words which take the uses to some untrusted or malicious site. Example of some spam words are ‘free’, ‘credit’, ‘deals’, ‘diet’ etc. So a list of spam words is created that are often found in spammer’s tweets. The tweets containing these words are considered as spam.

- Word count: Tweets containing one or two words give no information so they are often considered as spam.

- Number of URL’s: Since Twitter only allows a message with a maximum length of 140 characters, many URLs included in tweets are shortened URLs. Spammers often include shortened URLs in their tweets to entice legitimate users to access them. Twitter filters out the URLs linked to known malicious sites. However, shortened URLs can hide the source URLs and obscure the malicious sites behind them. While Twitter does not check these shorten URLs for malware, any user’s updates that consist mainly of links are considered spam according to Twitter’s policy [4]. In our work, we use the number of HTTP links that are contained in tweets.

- Number of url’s/words: If in a tweet ratio of number of url’s to the words is more than that tweet is treated as spam.

- Hashtags: Hash tagged words that become very popular are often Trending Topics. For example #symbol followed by a term describing or naming the topics, to a tweet. If there are many tweets containing the same term, the term will become a trending topic. Spammers often post many unrelated tweets that contain the trending topics to lure legitimate users to read their tweets. If a user posts multiple unrelated updates to a topic using the # symbol” then that tweet is considered as spam.

**Binary classifier:** After feature matrix creation KNN is applied which classify tweets as spam or non-spam. Classification is done in two phases training and testing. In this experiment accuracy of KNN is analyzed with different training percentages of data and 100% data is used for testing.

**Evaluation:** Lastly the results of classifier are analyzed. Performance of KNN is analyzed using different parameters like execution time, accuracy, sensitivity, specificity, precision, recall, f-measure, g-mean.

**Experimental results**

This section present the result of classification performed. For experiment, twitter data was extracted using R tool. Steps followed are as given in [4]. We have used 150 tweets for our experiment. Performance evaluation was performed using different percentages of training data and taking 100% data as testing data. The experiment is conducted several times and then the average value of these results is taken as final results. Table 1 shows the final results obtained with different percentages of training data.

The results of classification using KNN are shown in following figures. The results indicate the successful implementation of algorithm in spam detection whose accuracy increases with the increase in training percentages. When training and testing is done using 100% data then according to figure 3 classifier accuracy is maximum. Figure 2 shows the changes in execution time when training percentages are changed.
Table 1: Performance of KNN in spam detection with different percentages of training data.

<table>
<thead>
<tr>
<th>TrainPer(%)</th>
<th>Accuracy</th>
<th>Specificity/TNR</th>
<th>Precision</th>
<th>Recall/Sensitivity</th>
<th>F_measure</th>
<th>gmean</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>93</td>
<td>79</td>
<td>90</td>
<td>100</td>
<td>95</td>
<td>89</td>
<td>2.6</td>
</tr>
<tr>
<td>60</td>
<td>95</td>
<td>87</td>
<td>93</td>
<td>100</td>
<td>97</td>
<td>93</td>
<td>1.5</td>
</tr>
<tr>
<td>80</td>
<td>99</td>
<td>98</td>
<td>99</td>
<td>99</td>
<td>99</td>
<td>99</td>
<td>1.7</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>2</td>
</tr>
</tbody>
</table>

Figure 2: Time graph of KNN with different training percentages

Figure 3: Performance graph of KNN with Percentages training percentages

V. CONCLUSION

In this paper, performance of KNN a machine learning algorithm is analyzed in detecting spam from twitter data. Several experiments are conducted on twitter data to detect spam and results obtained show that KNN give very promising results in terms of execution time, accuracy, sensitivity, specificity, precision, recall, f-measure, g-mean. KNN provides accuracy of more than ninety percent even when less percent of data is used for training and accuracy is increased when training percentage is increased.

REFERENCES


