A Real Time Spam Classification of Twitter Data with Comparative Analysis of Classifiers

Sandeep Kumar Rawat  
M. Tech Student  
Department of Computer Science & Engineering  
Sri Sai University, Palampur, India

Saurabh Sharma  
Assistant Professor  
Department of Computer Science & Engineering  
Sri Sai University, Palampur, India

Abstract

In today’s scenario, online networking is fast becoming popular among Internet users. The rise in the use of social networking sites such as Twitter are gaining much importance because it plays a double role of online social networking and micro blogging. These sites have a constraint to them, i.e., the spammers. Twitter is becoming a popular site in the face of micro-blogging service for the users to share short messages, called Tweet. The spamming fact widely spread-out in the tweets is now affecting micro blogs and exploits specific mechanisms of the messaging process. Spammers try to attack the trending topics over the twitter to spoil the useful content. Social spamming is more successful as compared to the email spamming as it uses social relationship between the users. Spam detection in real time is very important because Twitter is largely used for the commercial advertisements. The spammers attack the private information of the user. Spammers can be detected by using the content and user based attributes. In this paper we will try to classify the spam from real-time Twitter data using Twitter Streaming API, text mining and use the classifiers for spam classification.

Keywords: Twitter Spam Detection, API streaming, Text mining, Pre-processing, Classification, machine learning approaches BPNN classifier, Naïve baye’s classifier

I. INTRODUCTION

Online social networking sites are becoming more popular each day, such as Facebook, Twitter, and LinkedIn. Among all these sites, Twitter is the fastest growing site than any other social networking site. Twitter has recently emerged as a popular social site where the users share and discuss about everything, including news, jokes, their take about events, and even their mood. With a simple interface where only 140 character messages can be posted, Twitter is fast becoming a system for obtaining real-time information. When a user posts a tweet, it is immediately delivered to all those persons who are following the person who has posted the tweet, allowing them to spread the received information even more. Twitter’s 140-character limit on a message serves modern day busy people’s trend of acquiring information in a short and quick way. It does not take much time to read through such short tweets. People need to spend only 5 to 10 minutes on Twitter to find out what is happening around in the world. Twitter was started in 2006 by Jack Dorsey as an online social networking and micro blogging service for the users to send and receive short messages (called tweets) of up to 140 characters. Only text and HTTP links can be included in the tweets. Such tweet exchanges allow friends/colleagues to communicate and stay connected thereby making twitter as the fastest growing social networking site. Spammers use Twitter as a tool to send unsolicited messages to legitimate users, post malicious links, and hijack the trending topics. Spam is becoming a major problem on Twitter as well as on other online social networking sites. Spammers could be phishers, Malware Propagators, Marketers, and Adult Content Propagators. These all can spread malicious URLs and links, advertisements and adult contents. We need some tools that can automatically identify these spammers. In addition, we need more accurate but efficient spam detection methods to avoid causing inconvenience to legitimate users. In this paper, we focus more on spammers in Twitter. Some spammers are record clients who spread malignant URLs through their tweets. Unaware of these URLs the clients click on these connections and end up in entering wrong sites. This prompts the spammers to hack passwords, charge card data etc. Malware propagators are the clients who tweet malevolent connections, which on clicking prompts the downloading of malwares. Malwares are malicious programming which can end up being unsafe to real clients. Marketers are the spammers who focus on spreading ads. They attempt to drift distinctive items. Advertisers are typically safe in light of the fact that, the just thing they do is, promoting their items. Sometimes these clients can deceive the true blue clients. Adult Content Propagators are the spammers who spread the adult content through tweets.

In this paper we are proposing an application which can classify a Twitter user into spam or legitimate. To achieve this we use text mining, and Machine Learning techniques and certain classifiers are also used. The rest of the paper is organized as follows. Section II talks about the related works done in this field. Section III describes the system design and methodology. Section IV explains the experiments done and results thereof. Finally, the paper is concluded in the section V.
Investigation into micro blogs for spam detection is still in its infancy and not much literature related to the subject exists. Machine learning techniques have been used by Benevenuto, et al., to identify spammers. They identified two sets of each tweet, number of hash tags on each tweet etc.

Wang [7] used a directed social graph model to explore friend and follower relationships among Twitter. Twitter’s spam policy that, if a person has a small number of followers compared to the amount of he is following the account is considered as spam was used to compute three features namely the number of friends, the number of followers, and the reputation of a user. Other features such as hyperlinks, replies and mentions and trending topics are also used as attributes. Several methods of classification like decision tree, neural network, support vector machines, and k-nearest neighbors are compared.

C. Grier et al. [6] show that it takes a few weeks for URLs posted in Twitter to be on its blacklist. In addition to the fact that Twitter itself does to 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. Currently, much research is going on to find a method to detect Twitter spamming in an efficient and automated way.

Previous study at U.C. Berkeley shows that 45% of users on a social network site readily click on URLs without doubt. Grier et al. collected over 400 public tweets and reported that 8% of 25 million unique URLs posted to twitter point to phishing, malware, and scam.

Grier proposed a schema based on URL blacklisting.

Evaluation of the context-aware spam that could result from information that is shared on the social networks is dealt in [4]. The mitigation techniques are also discussed here. The authors have done analysis on Facebook. The authors concluded that context-aware e-mail attacks have a high rate of success. The paper also mentions the defense strategies taken by other social networks like LinkedIn and MySpace.

Harvested Twitter dataset and links are examined in [5]. Here the authors have found features using which content polluters can be easily identified. The authors proposed a long term study of protecting social networks using honeypots. Almost 60 honeypots were deployed for seven months which resulted in the harvesting of more than 30000 spam data. The spam classification was done using machine learning algorithms.

Wang proposed a new approach that employed features from the social graph from Twitter users to detect and filter spam. Although these approaches are able to deal with the problem, a new spammer account may emerge to substitute the filtered accounts. Hence, these blacklisting systems, similarly as happens with e-mails, should be complemented with content-based approaches commonly used in e-mail spam filtering.

### III. SYSTEM DESIGN AND METHODOLOGY

We have come up with an application which can classify real time tweet using API and classify them correctly. From the extensive literature survey, it is concluded that early classification of spam is very critical for maintaining the security. Spam classification helps user to report the spam in account. In case of larger number of spam tweets, it takes lot of time to classify spam in twitter. Therefore, there is need for designing a system which would automatically classify the spam based on their textual description.

Although, many attempts has been made at the problem and several machine learning and text mining algorithms have been applied for the same, a novel idea of using dictionary for the prediction of spam in real time test set is proposed in this paper. BPNN and Naïve Bayes are being exploited as classifiers with information gain and Chi square.

In this paper our goal is to identify useful features that can be used in traditional machine learning schemes to correctly classification of twitter data major contributions of this paper are as follows:

1. Real time tweets extraction using API
2. Preprocessing of the tweet set by using text mining approach.
3. Use TF-IDF score for creation of a (Term-document matrix) TDM matrix
4. Use information gain and chi-square for dimensionality reduction.
5. Classification of the spam by training the dataset using BPNN and Naïve Bayes Classifier and testing the efficacy of the proposed algorithm in terms of accuracy, and precision.

Textual description and Machine learning algorithms on dictionary for classifying twitter spam. A novel idea of utilizing a tf-idf score of real time tweets data set and feature selection method to create dictionary of critical terms and using this dictionary for prediction of spam is proposed in this thesis. For this purpose we intend to use text mining algorithms for pre-processing steps. Pre-processing of tweets is done using steps such as tokenization, stop word removal and stemming. Further TF-IDF score is calculated for each term. Then two feature selection methods are used for creating dictionary of terms. Therefore two ML algorithms such as Naïve Bayes algorithms and Back Propagation Neural Network are used as classifiers. The approach is validated using real time tweets that are extracted using API and performance is analyzed using precision and accuracy.
Natural Language Processing (NLP) is a technique which enables a machine to process a natural language (like English) and do all the things that a human can do.

1) **Tokenization**: The purpose of tokenization is to remove all the punctuation marks like commas, full stop, hyphen and brackets. It divides the whole text into separate tokens to explore the words in document.

2) **Stop word removal**: The purpose of this process is used to eliminate conjunction, prepositions, articles and other frequent words such as adverbs, verbs and adjectives from textual data. The reason to remove these words is that these words are very common and rarely contain any important information. Thus it reduces textual data and system performance is increased.

3) **Stemming**: Stemming is used to reduce the words to their root words e.g. words like “computing”, “computed” and “computerize” has its root word “compute”.

4) **Weighting Factor**: Features are extracted from overloaded large datasets. TF-IDF (Term frequency-Inverse document frequency) score is generally is used to give weight to each term. TF-IDF is multiplication of term frequency and inverse document frequency.

\[
TF-IDF = n_d^w \cdot \log \left( \frac{N}{n_w^d} \right)
\]

Where \( n_d^w \) = frequency of word \( w \) in document \( d \)
\( N \) = total document and \( N_w^d \) = document containing word \( w \).

5) **Term - document matrix**: After initial steps of preprocessing text in documents is converted into term-document matrix. Rows in matrix represents document in which word appears and columns represent the words that are extracted from documents. The cell of matrix is filled with TF-IDF score.

6) **Dimensionality Reduction**: After preprocessing steps, dimensionality reduction is performed. Here original TDM (term document matrix) is replaced with smaller matrix by using a SVD (singular value decomposition technique). This technique discards unimportant word and relevant and important word are filtered out. The new matrix is generated of terms and documents. Mining the reduced data with traditional data mining techniques- Classification, clustering and predictive methods are applied to the reduced datasets using data mining techniques to analyze the pattern and trends within data.

B. **Keywords/Word Weight**:

Since we observe that the contents in spammers’ tweets contain similar words, we define two metrics to help identify spammers. First, we created a list of spam words that are often found in spammer’s tweets and the associated probabilities of these words and a list of popular words in legitimate tweets and the associated probabilities of these words. Comparison with a set of already identified expressions the next task is to identify a set of expressions or words in a tweet which can prove that the tweet is a spam...

The presence of even one of these expressions can conclude that tweet classified as spam. If the user is not classified as spam in this step, or if he hasn’t tweeted any term related to it, then it is classified as a non-spam.

C. **Machine Learning Techniques**:

Machine learning, is a branch of artificial intelligence, which is concerned with the construction and study of systems that can learn from data. For example, a machine learning system could be trained on email messages to learn to distinguish between spam and non-spam messages. After learning, it can then be used to classify new email messages into spam and non-spam folders. The various machine learning algorithms used in our thesis work are:

1) BPNN classifier.

2) Naïve-Bayes classifier

In this paper we have used Naïve-Bayes which is a supervised machine learning algorithm. The dataset used here were first tested using two algorithms: Naïve-Bayes and BPNN. Out of this Naïve-Bayes was found to be more accurate.
D. The Back-Propagation Neural Network:

BPNN is solution to the problem of training multi-layer perceptions. The fundamental advances represented by the BPNN were the inclusion of a differentiable transfer function at each node of the network and the use of error back-propagation to modify the internal network weights after each training approach. Theory of artificial neural networks starts and developed in the line with elementary principle of operation of neural system. Since, a awfully kind of networks have been created. All square measure are composed of units, and connections between them, that along confirm the behavior of network. The selection of network depends on matter to be solved; rear propagation gradient network is most often used. This network includes 3 or lot of somatic cell layers: one is input layer, one output layer and a minimum hidden layer. In most of the cases, network with hidden layer is used the limit calculation time, specifically once results are attained the square measure satisfactory. All somatic cells of layer square measure connect by the nerve fiber to neuron of the successive layer a pair of 1. Signal propagation is the input layer that includes n neurons code for n items of signaling of Network. The quantity of the neurons of hidden layer is selected with empirical observation by user. Finally, output layer consist k neurons for k categories. Every association between the 2 neurons is related to weight issue; this weight can be changed by the ordered iterations through the coaching of network in line with the input and output information. With the input layer, the state of somatic cell is set by input variable; the opposite neurons judge the state of signal from previous layer.

Fig. 3: Structure of a neural network used in the Experiments

Fig. 4: Detail of one Neuron
E. Naïve-Bayes:

Naïve-Bayes is a probabilistic classifier which uses the Bayes Theorem. Each feature is independent of each other. Consider a Test Set T with attributes (features) a1, a2 . . . an.

\[ T = \{a_1, a_2 . . . a_n\} \] and a set of labels \( L = \{\text{Spam, Legitimate}\} \). Then, \( P(L | a_1, a_2 . . . a_n) = p(L) \cdot (a_i | L) \) where \( i = 1 \) to \( n \)

Whichever label has the higher probability is the label of that particular test set T.

Requirements for the implementation are a training set and a test set. The most important thing for developing an efficient classifier is to construct a good training set. The success of the classifier lies in the efficiency of the training set. Inefficient training set will lead to a classifier with low accuracy. The results obtained after using BPNN and Naïve-Bayes are given.

IV. EXPERIMENT RESULTS

The application works for all kinds of tweets. For checking the accuracy, a set of 70 tweets were used in this application.

By Naïve Bay’s classifier, Out of the 70, 60 tweets were correctly classified and 10 were spam. Whereas BPNN classifier, only 52 tweets were correctly classified. This result also shows that the NB performs better than BPNN.

A. Overview of Performance Measures:

We perform a binary classification as to whether a given tweets is spam or not. In case of a highly imbalanced data accuracy alone is not sufficient measure to estimate the performance of classifier. For example, if had only 1% of positive instances in data we can achieve an accuracy of 99% but simply classifying all instances as false. In such situations the following measure are more useful and informative for evaluating the performance of such binary classifiers. Accuracy: Accuracy is calculated as fraction of sum of correct classification to total number of classification. It is defined as:

\[ \text{Recall} = \frac{TP}{TP+FN} \]

In simple terms this means the number of correctly classified positive instances out of the total instances which are positive. For example, if we had 10 fruits of which 5 are apples then recall will be how many of the 5 apples are correctly classified over number of apples. Recall is the same as sensitivity.

\[ \text{Precision} = \frac{TP}{TP + FP} \]

From the above fruits example, it would mean ratio of how many correctly classified apples over total number of fruits classified as apples. These measures are computed for all the classification tests in this work. In general we can say that if the above measures have higher values the classifier performs well and accuracy is higher. However a high accuracy but lower values for above measures may indicate a poor classifier performance.

<table>
<thead>
<tr>
<th>Component</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>84.28%</td>
<td>85.77%</td>
<td>42.85%</td>
</tr>
<tr>
<td>BPNN</td>
<td>74.25%</td>
<td>72.18%</td>
<td>37.14%</td>
</tr>
</tbody>
</table>

V. CONCLUSION AND FUTURE SCOPE

Spammers are the problem in any online social networking sites. Once a spammer is detected it is easy to suspend his/her account or block their IP address. This research deal with the study of spam classification techniques in twitter. Twitter API is developed to collect real data set from Twitter public available information. The real time tweet dataset was obtained and the document was pre-processed. Various special characters, stop words etc have been removed and finally the TF-IDF scores of each word are calculated. A Naïve Bayes algorithm was utilized for classifying the tweets dataset as spam and non-spam. The results are found to be quite satisfactory in terms of accuracy, precision and recall. The results have also been compared to other algorithms which have been implemented namely Back Propagation Neural Network. It was found that while the NB performs better than BPNN,
the results of NB are the best. In future, other algorithms can be implemented for real time tweets comparison purpose and for the better results than these classifiers.

REFERENCES


