Improving Spam Detection in Online Social Networks

Arushi Gupta
B.Tech.(6th Sem. IT), Department of Information Technology
Indira Gandhi Delhi Technical University for Women
Kashmere Gate, Delhi
arushigupta12@gmail.com

Rishabh Kaushal
Asst. Prof., Department of Information Technology
Indira Gandhi Delhi Technical University for Women
Kashmere Gate, Delhi
rishabh.kaushal@gmail.com

ABSTRACT

Online Social Networks (OSNs) are becoming one of the most popular platforms for people to interact and share information. While it is true that OSNs have become a new medium for dissemination of information, at the same time, they are also fast becoming a playground for the spread of misinformation. Consequently, we can say that an OSN platform comprises of two kinds of users namely, Spammers and Non-Spammers. Spammers, out of malicious intent, post either unwanted (or irrelevant) information or spread misinformation on OSN platforms. As part of our work, we propose mechanisms to detect such users (Spammers) in Twitter social network (a popular OSN). In our work, we have applied three learning algorithms namely Naive Bayes, Clustering and Decision trees. Furthermore, to improve detection of Spammers, a novel integrated approach is proposed which "combines" the advantages of the three learning algorithms mentioned above. Results, thus obtained, show that our novel integrated approach that combines all algorithms outperforms other classical approaches in terms of overall accuracy and detect Non-Spammers with 99% accuracy with an overall accuracy of 87.9%.

1. MOTIVATION

Online Social Networks (OSNs) are a platform where people with common interests and beliefs, interacts and connect. However, at the same time, some of the users, called Spammers, are misusing these OSN platforms, thereby spreading misinformation, unsolicited messages, etc. Sometimes, this spamming is done with the intent of advertising and other commercial purposes. Such activities disturb Non- Spammers and also decrease the reputation of OSN platforms. Therefore, there is a need to devise mechanisms to detect Spammers so that corrective actions can be taken thereafter. We have worked in this area to detect spam accounts in one of the most popular OSN, Twitter. Twitter was selected as an OSN. In our work, an algorithm is proposed to successfully detect spam accounts with 86.5% accuracy. Finally, this algorithm was compared with various other classification algorithms and the results show better performance.

2. METHODOLOGY

An overview of the complete process of spam detection is shown in the diagram in Figure 1, each of whose steps are explained in this section.

2.1 Data Set Description and Feature Identification

We have used the dataset obtained from Fabricio Benevenuto *et al.* [2] which consists of labeled record of 1064 Twitter users. Dataset comprises of 62 features containing user specific and tweet specific information. They have used SVM based machine learning approach as opposed to our work in which we have used other learning approaches namely Naive Bayes, Clustering, Decision Trees and finally combined all of them together to achieve a higher spam detection accuracy.

Since, spammers behave differently from non-spammers; therefore we can identify some features or characteristics in which both these categories differ. In this preliminary work of ours, we have used very basic features to help us in detection of spam accounts, namely: number of followers and followees, number of URLs, usage of spam Words, fraction of hashtags, etc.

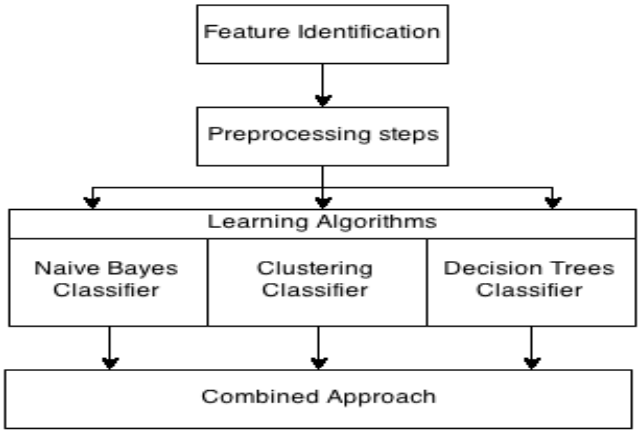

Figure 1: Proposed Spam Detection Approach

2.2 Preprocessor

In preprocessing step, all the continuous features were converted into discrete using a procedure adopted to select the intervals for a particular feature from the work of *Alex Hai Wang* [4] according to which all user accounts are arranged in increasing order of their feature values. Processing begins from the first account, if we encounter an account whose category is different from the category of the next account, and then an interval is created as a mean of both the feature values.

2.3 Learning Algorithms

We have combined Naive Bayes, Clustering and Decision trees into an integrated algorithm in order to increase the accuracy. As

outlined in Figure 1, the preprocessed data is first classified using different learning algorithms to predict the class {"Spammers" or "Non-Spammers"} of all Twitter user accounts. In *Naive Bayes approach,* accounts were classified by calculating the probability of the given account to be Spammer/Non-Spammer, given the feature values of that account. Bayes theorem was used to calculate this probability. On the basis of similar feature values, *Clustering algorithm* could classify the entire set of accounts into two classes (Spammers and Non-Spammers). In *Decision trees* learning method, a tree structure was prepared and the decisions, on the basis of feature values, were made at every level of the tree.

2.4 Combined Approach

As part of combined approach, we compare the classification results of any two learning algorithms, if both the learning algorithms predict the same result, then we finalize the class of the Twitter account under investigation. Otherwise, if the predicted class of both the classification techniques differ, then we use the prediction of third algorithm as the final class.

3. RESULTS

In order to check the correctness of our proposed algorithm, the results obtained from the algorithm were compared with the labels (Spammers/Non-Spammers) in the dataset [2]. It is evident in Figure 2, that the proposed algorithm was able to successfully identify an account as Spammer or Non-Spammer with 87.9% accuracy. The algorithm's accuracy of detection of non-spammers was higher (99.1%) as compared to the accuracy of detection of spammers (68.4%). This integrated algorithm was then compared with each of the learning algorithm, Naive Bayes, Clustering and Decision Trees. The results showed that Clustering algorithm performs better in detection of non-spam accounts but was very poor in detecting spam accounts. Our algorithm was able to maintain the high accuracy of Clustering algorithm in detecting non-spam and at the same time, retain the accuracy of Naive Bayes in detecting Spammers accounts thereby, increasing the overall accuracy.

4. IMPLICATION OF RESULTS

Spamming is an undesirable activity in OSNs and effective mechanisms need to be developed to detect it and thereafter take remedial steps. Our work is a small first step in that direction in which we make preliminary investigations in applying various machine learning algorithms in detection of spam. Though, we have been able to correctly detect most of non-spammers (over 99%), however detection of spammers is not satisfactory (at 68%). Use of a very basic feature set could be a potential cause of this outcome. In our future work, we plan to address this issue.

5. REFERENCES

[1] Fabricio Benevenuto, Gabriel Magno, Tiago Rodrigues, and Virgilio Almeida. "Detecting Spammers on Twitter". In Proceedings of the Conference on Email and Anti-Spam (CEAS), 2010.

[2] De Wang, DaneshIrani, and Calton Pu. "A Social-Spam Detection Framework". In proceeding of Conference on Email and Anti-Spam (CEAS), 2011

[3] Dewan Md. Farid1, Nouria Harbi1, and Mohammad Zahidur Rahman. "Combining Naive Bayes And Decision Tree For Adaptive Intrusion Detection". In International Journal of Network Security & Its Applications (IJNSA), Volume 2, Number 2, April 2010.

[4] Alex Hai Wang. "Don't Follow Me: Spam Detection In Twitter". In Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference.

[5] Xin Jin, Cindy Xide Lin, Jiebo Luo and Jiawei Han. "A Data Mining-based Spam Detection System for Social Media Networks". Published in 2011.

[6] M. McCord and M. Chuah. "Spam Detection on Twitter Using Traditional Classifiers". In ATC'11 Proceedings of the 8th international conference on Autonomic and trusted computing.

[7] Kurt Thomas, Chris Grier, Vern Paxson and Dawn Song. "Suspended Accounts in Retrospect: An Analysis of Twitter Spam". In IMC'11 proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference.

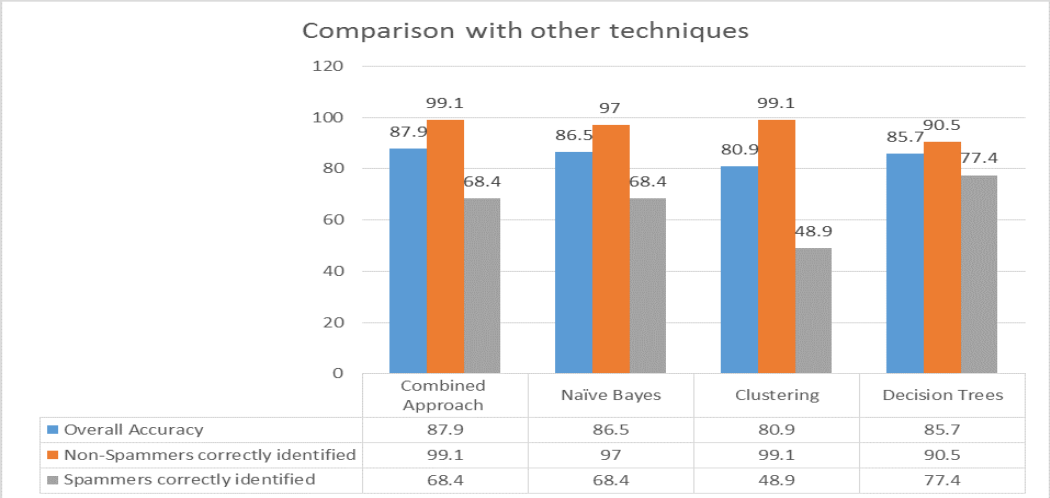

Figure 1: Comparison of Improvement in Spam Detection using four learning approaches