Detecting Spammers on Twitter by Identifying User Behavior and Tweet-Based Features

Dewi W. Wardani and Yulia Wardhani
Department of Informatics,
Universitas Sebelas Maret (UNS), Surakarta, Indonesia
dww_ok@uns.ac.id

***Abstract*—Spam is a problem in the delivery of news and communication networks. It has various forms and definitions depend on the type of the network. With millions of users across worldwide, Twitter provides a variety of news and events. However, with the ease of dissemination of news, and allowing users to discuss the stories in their status, these services also open opportunities for another kind of spam. In this study, the proposed spammer detection classifies accounts into a spammer or non-spammer by studying/identifying user behavior and tweet-based features (number of followers, following, mentions and hashtag). The results showed that our proposed approach returns better scores comparing to the result of C5.0 algorithm.**

***Index Terms*—C5.0; Spammer; Detection; Tweet-Based Features; Twitter.**

I. INTRODUCTION

Recently, a new form of spam appearing on social networking sites due to their wide popularity and tight integration in the daily life [1]. It happens as well on Twitter [2]. A study explained that more than 3% of tweets are spam [3]. Spammer attacks caused temporary negate Twitter trending topic. Unfortunately, deleting those tweets are not polite [4]. Twitter has set the Twitter Rules, which explains the definition of spam. Some of the factors that are considered as a spam by Twitter Rules are: if posting in large numbers using hashtag, mentions, URL, and if it has a number of followers that is less than the following.

Spammers usually disseminate information by posting a URL with the intention that users of Twitter will click the URL [5]. Figure 1 is an example of how spammers attack a verified account. Spammers also indicated to make a lot of mentions into non-follower account. The results of the study [6] also mentioned that Twitter spams more successful forcing the user to click on a URL to email spam with a 0.13% click through.

One solution for these problems is by applying data mining techniques for identifying the characteristics of the data. In this study, we propose a new approach which follows rules of Twitter to identify spammer. Our approach studies the behavior of the user and the tweet-based some features (followers, following, URLs, mentions and hashtag). The application is built to evaluate the performance comparing C5.0 algorithm.

This paper is organized as follows, after the introduction, in Section 2 presents a few closest related works then in Section 3 presents the proposed approach. Section 4 explains the experiment and finally we summarize our work and proposes future work in Section 5.

Figure 1: Example of Twitter Verified Account Attack

II. RELATED WORK

A work studies to detect spammers who post at least a tweet which contains unrelated URL with the real content of the tweet. For instance, a URL contains an advertisement which has different contents with the hashtag of the posted tweet. The other case is changing the real URL with illegitimate one by shortening the illegitimate URL [7]. This work utilizes Support Vector Machine to detect a spammer. Some features are used to be analyzed. They are as follows: the number of words which are listed as the word of spam, the number of URL, the number of hashtag, the number of words, the number of numeric character, the number of number character, the number of URL, the number of hashtag, the number of mention, the number of tweets which are a mention or retweet and the number of reply. The other work utilizes Logic Regression, Naïve Bayes and RBF Network to categorize the spammer account [8] [9]. A few features of tweet have been used as well. Eventually, Naïve Bayes returns a better performance.

One work proposed a novel approach to detect spambot in Twitter [10]. The tweets usually contain malicious link. It proposed graph-based feature and content-based feature. These three graph-based features are the number of followers, the number of following and the ratio of follower. A graph approach is used by Wang [11] within several algorithms. Some comparisons have been performed as well, and Naïve Bayes returns much better performances comparing the other classification algorithms.

III. Proposed Approach

A. Spammer Detection Algorithm

We studied the explanation of a spammer by Twitter. The result of the study drives this study to focus on using user and tweet-based features. They are as follows; number of following; number of followers; number of URLs on the 20 most recent tweets; number of mentions at 20 most recent tweets and number of hashtag at 20 most recent tweets.

Spammer detection algorithm is constructed based on the important features that are analyzed in an earlier study which has been explained before. It seems that our approach is much simple than the previous approaches, but later we will show its effectiveness. The pseudo-code of the algorithm is explained as below:

Input: U (tweet), A (twitter account)
Output: SA (spammer account)

(1) Preprocessing of U
(2) if isfriend = true then SA = spammer;
(3) elseif isfriend = false AND isurl = true then A = SA;
(4) elseif isfriend = false AND isurl = false AND ismention = true then A = SA;
(5) elseif isfriend = false AND isurl = false AND ismention = false AND ishashtag = true then A = SA;
(6) elseif isfriend = false AND isURL = false AND ismention = false AND ishashtag = false then A = SA.

B. Features Importance Analysis for C5.0

Features importance analysis is an analysis of the features to get the features which have the most important role in the process of identifying spammers. This research used information gain to determine the ranking of important features. We obtain the rank of features from the highest to the lowest as follow: ishashtag, ismention, isurl and isfriend. Table 1 explains the chosen features to be analyzed by considering the 20 most recent tweets.

Table 1
Description of attributes

Attribute	Value	Note
isfriend	TRUE	follower < following
	FALSE	follower >= following
isurl	TRUE	URL >= 20
	FALSE	URL < 20
ismention	TRUE	Mention >= 20
	FALSE	Mention < 20
ishashtag	TRUE	Hashtag >= 20

IV. Experiment

A. Obtaining Dataset

The dataset is tweets which are related to verified accounts. In this case, we use tweets which are related to Justin Bieber. We use tweets which are addressed to @spam and data which are related to verified account as the training data. We use tweets which related to verified account as the testing data. We obtained three datasets form three verified accounts: @justinbieber, @BarackObama and @ladygaga.

Labeling data is done manually (by some volunteers) by checking on the 20 most recent tweets that have a number of followers, following, number of URLs, the number of mentions, and the number of hashtags. To simplify the process of labeling data, a web-based application has been built to ease the labeling process. Volunteers are asked to handle the process. Each account is classified based on majority voting. Figure 2 until Figure 6 are the example descriptions of datasets (Dataset 1).

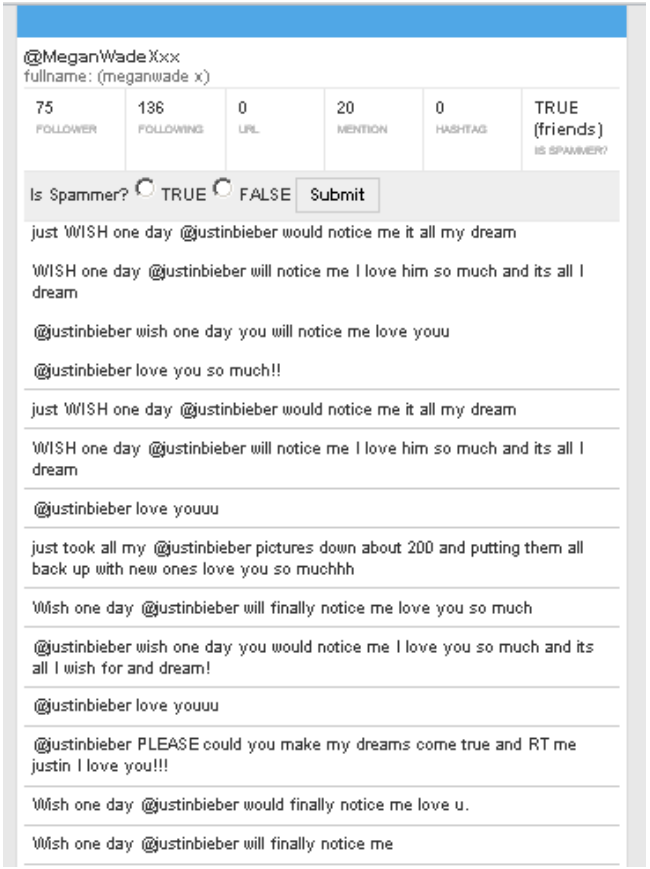

Figure 2: Spammers Detection Application

B. Experimental Result

Figure 3 shows that spammers do not attack hashtag too frequent. The majority of related tweets are tweets from Justin Bieber's fans which include the hashtag in their tweets. Figure 4 indicates that many spammers attack by mention to other users. They include verified account in their 20 most recent tweets to increase their follower. Figure 5 describes that the spammers do not include more than 20 URLs in 20 most recent tweets they post. This result is slightly different from previous studies with the result that many spammers posting URL. Figure 6 shows that spammers have a lot of following than their followers. Spammers follow multiple accounts to get a lot of followers and promote their spammer accounts.

We conducted three experiments for three datasets. Dataset 1 (100 records) is tweets which are related to verified account Justin Bieber (@justinbieber), dataset 2 (150 records) which are related to verified account Barack Obama (@BarackObama) and dataset 3 (300 records) which are related to account Lady Gaga (@ladygaga).

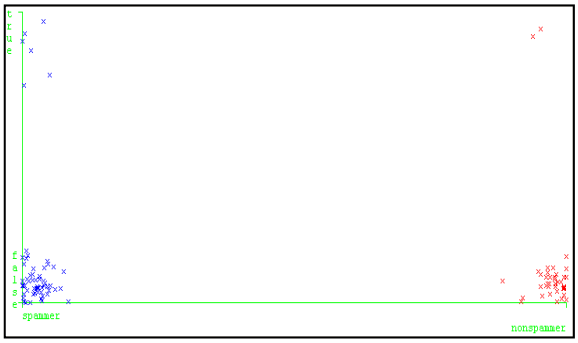

Figure 3: The distribution of ishashtag

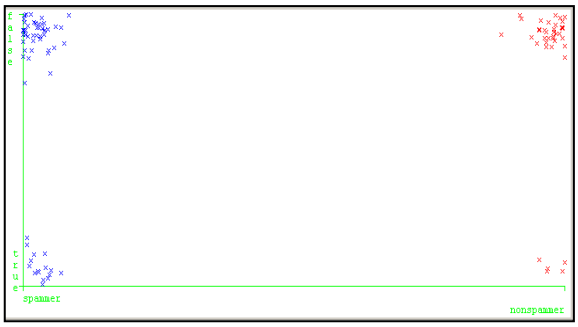

Figure 4: The distribution of ismention

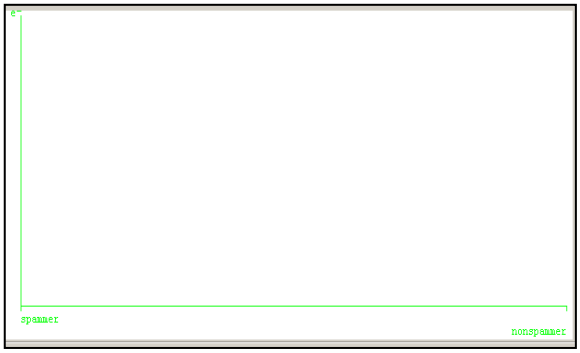

Figure 5: The distribution of URL

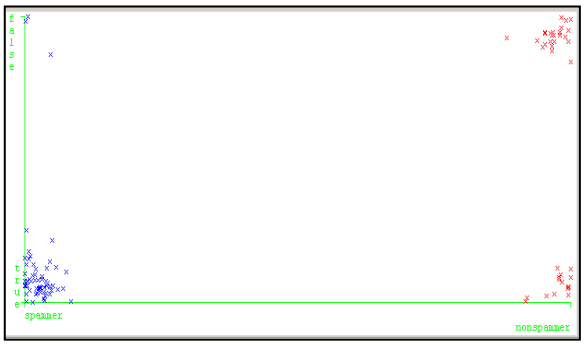

Figure 6: The distribution of isfriend

Table 3 is the detection result of the proposed approach and table 4 is the detection result of C5.0. In summary, the result of testing is explained in Table 5. The experimental results show that the proposed spammer detection returns more stable results comparing to C5.0. Overall, the scores are surpassing the scores of C5.0 except for the recall. Although the recall score of the proposed approach does not surpass all recall scores of C5.0, it returns stable high scores. These results show that the proposed approach, although it seems simpler than the other work, it shows promising approach.

Table 3
The result of the Proposed Spammer Detection

Testing	Actual	Prediction Spammer	Prediction NonSpammer
Testing1	Spammer	58	8
	NonSpammer	3	31
Testing2	Spammer	84	3
	NonSpammer	18	45
Testing3	Spammer	175	4
	NonSpammer	15	106

Table 4
The result of C5.0

Testing	Actual	Prediction Spammer	Prediction NonSpammer
Testing1	Spammer	61	39
	NonSpammer	0	0
Testing2	Spammer	86	1
	NonSpammer	63	0
Testing3	Spammer	0	179
	NonSpammer	0	121

Table 5
The Comparison of the result between the Proposed Spammer Detection and C5.0

Method	Testing (T)	Accuracy	Recall	Precision	Error
The proposed Spammer Detection	T1	0.89	0.95	0.878	0.11
	T2	0.86	0.9655	0.8235	0.14
	T3	0.9367	0.9776	0.9210	0.63
C5.0	T1	0.61	1	0.61	0.39
	T2	0.573	0.988	0.577	0.4267
	T3	0.4033	0	0	0.5967

V. Conclusion And Future Work

This study concludes that some features (number of followers, number of following, number of URL, number of mention, number of hashtags) can be used to determine the classification of spammer or non-spammer account. We proposed the new approach and overall it shows quite good scores compared to the scores of C5.0. The near future work is the study to use more diverse features. They are needed for investigating the other kinds of attacks on Twitter which are related to verified accounts.

Acknowledgment

Thanks to Fabricio Benevenuto for the advice and support.

References

[1] N. Spirin, “Mutually Reinforcing Spam Detection on Twitter and Web,” in 2010 VIII All-Russian scientific conference, pp. 1–7.
[2] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, “Detecting spammers on Twitter,” in 2010 Collaboration, electronic messaging, anti-abuse and spam conference (CEAS), pp. 12–22.
[3] P. Analytics, “Twitter study,” HYPERLINK, vol.1, no.1, pp. 1–13, 2009.
[4] N. Pemmaraju, R. A. Mesa, N. S. Majhail, and M. A. Thompson, “The use and impact of Twitter at medical conferences: best practices and Twitter etiquette,” in 2017 Seminars in hematology, pp. 184–188.
[5] J. Song, S. Lee, and J. Kim, “Spam filtering in twitter using sender-receiver relationship,” in 2011 Recent advances in intrusion detection, pp. 301–317.
[6] C. Grier, K. Thomas, V. Paxson, and M. Zhang, “@spam: the underground on 140 characters or less,” in 2010 Proceedings of the 17th ACM conference on Computer and communications security, pp. 27–37.

[7] I. Ernawati, "Prediksi Status Keaktifan Studi Mahasiswa dengan Algoritme C5. 0 dan K-Nearest Neighbor," unpublished.
[8] E. Thomas, "Data mining: Definitions and decision tree examples," in 2004 the Association for Institutional Research and Planning Officers (AIRPO), pp. 1-13.
[9] R. Kohavi and J. R. Quinlan, "Data mining tasks and methods: Classification: decision-tree discovery," in 2002 Handbook of data mining and knowledge discovery, pp. 267–276.
[10] R. Kohavi and F. Provost, "Confusion matrix," Machine learning, vol. 30, no. 2–3, pp. 271–274, 1998.
[11] A. H. Wang, "Detecting Spam Bots in Online Social Networking Sites: A Machine Learning Approach.," DBSec, vol. 10, pp. 335–342, 2010.