Information Sciences 260 (2014) 64–73

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Twitter spammer detection using data stream clustering

Zachary Miller [a], Brian Dickinson [a], William Deitrick [a], Wei Hu [a,*], Alex Hai Wang [b]

[a] Department of Computer Science, Houghton College, Houghton, NY, United States
[b] College of Information Sciences and Technology, The Pennsylvania State University, Dunmore, PA, United States

ARTICLE INFO

Article history:
Received 8 August 2012
Received in revised form 9 October 2013
Accepted 1 November 2013
Available online 17 November 2013

Keywords:
Twitter
Spam detection
Clustering
Data stream

ABSTRACT

The rapid growth of Twitter has triggered a dramatic increase in spam volume and sophistication. The abuse of certain Twitter components such as “hashtags”, “mentions”, and shortened URLs enables spammers to operate efficiently. These same features, however, may be a key factor in identifying new spam accounts as shown in previous studies. Our study provides three novel contributions. Firstly, previous studies have approached spam detection as a classification problem, whereas we view it as an anomaly detection problem. Secondly, 95 one-gram features from tweet text were introduced alongside the user information analyzed in previous studies. Finally, to effectively handle the streaming nature of tweets, two stream clustering algorithms, StreamKM++ and DenStream, were modified to facilitate spam identification. Both algorithms clustered normal Twitter users, treating outliers as spammers. Each of these algorithms performed well individually, with StreamKM++ achieving 99% recall and a 6.4% false positive rate; and DenStream producing 99% recall and a 2.8% false positive rate. When used in conjunction, these algorithms reached 100% recall and a 2.2% false positive rate, meaning that our system was able to identify 100% of the spammers in our test while incorrectly detecting only 2.2% of normal users as spammers.

1. Introduction

Social networking sites have exhibited exponential growth over the last few years. Among the most popular sites are MySpace, Facebook, and most recently Twitter. While the popularity of MySpace is declining and Facebook’s membership has plateaued at around 130 million, Twitter is still growing with around 25 million active users. Facebook’s slowing expansion due to market saturation stands in contrast to Twitter’s growth rate of about 30% annually [9].

Unfortunately, the proliferation of social networking has contributed to an increase in spam activity [15]. Spammers send unsolicited messages to users with varying purposes, which include, but are not limited to advertising, propagating pornography, phishing, spreading viruses, and degrading Twitter’s reputation [4]. “Hashtags”, “mentions”, and shortened URLs are frequently abused by spammers, making them indicators that a tweet may be from a spam user. Mentions are normally used to reply to another user or to send a message, which will appear on his profile. A spammer may exploit this by mentioning other users in order to spread spam beyond his limited following. Hashtags are used to denote trending topics making them more searchable for users and are abused by spammers, who attach the most popular hashtags to messages with links to unrelated topics. URL attacks are aided by Twitter’s 140 character limit on tweets as many legitimate users need to use link-shortening services to reduce the length of their URLs. This allows spammers to deceive users as the shortened links do not appear to be malicious, which is particularly important because links are a necessary part of almost any spamming attack. On Twitter, links are especially dangerous because it is very difficult to determine their destination without following

[*] Corresponding author. Tel.: +1 5855679612.
E-mail address: wei.hu@houghton.edu (W. Hu).

0020-0255/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2013.11.016

them to potentially harmful sites [14]. The ability to disguise URL destinations has made Twitter a particularly attractive target for spammers, which has motivated the development of several spam detection techniques.

In the first study focusing on Twitter spam detection, a data set of approximately 25,000 Twitter accounts was collected over several weeks with a web-crawler using Twitter's API [15]. A variety of predictors were then used for classification using 10-fold cross validation with a set of user and content based features. The users classified as spammers by the algorithms were then manually checked to determine the false positive rate. The Naïve Bayes Algorithm provided the best results with 91.7% precision, recall, and F-measure.

Researchers in [4] collected an enormous data set comprised of every known active public account on Twitter as of August 2009. This data set included millions active and inactive users and 1.7 billion tweets. Tweets containing one of the three most popular hashtags (#musicmonday, #MichaelJackson, and #SusanBoyle) were then identified and removed for manual labeling. A Support Vector Machine running 5-fold cross validation was used on this subset to evaluate the effectiveness of a similar feature set to [15]. They were able to identify 70% of spammers while maintaining a relatively low false positive rate of about 4%.

One of the most recent studies [13] used a collection of learning algorithms including Random Forest, Support Vector Machine, K-Nearest Neighbor, and Naïve Bayes. Each was used to evaluate a labeled data set of approximately 1000 Twitter accounts utilizing a similar feature set to the other previously mentioned studies [4,15]. The Random Forest algorithm produced the top results with over 95% precision, recall and F-measure. SVM performed nearly as well, achieving around 93% in each category.

In each of these previous studies, classification was utilized for spam detection on Twitter. The aim of this study is to develop an anomaly detection system for identifying spammers on Twitter using account information and streaming tweets. Typically binary classification applies to a data set that has balanced class labels, while anomaly detection is effective when the majority of a data set is one class; data outside that class are outliers. Due to the speed and volume of Twitter traffic, stream mining is the most natural technique for spam detection. Data streams are characterized by large volumes of continuous data evolving over time. Because of this, stream mining must make use of only one pass over the data. Therefore it is preferable for use in the Twitter environment where millions of tweets are posted every day.

2. Materials and methods

2.1. Data and their representation

The data set for this study included 3239 user accounts with a sample tweet from each account. Among the 3239 users, 208 were spam accounts from [15]. Because an estimated 6% of all Twitter accounts are spammers, our 208 spam users were combined with 3031 randomly selected verified normal users to form our complete data set. The 3031 normal users were randomly selected from a set of 37,000 accounts that tweeted during a one-hour period on March 3rd 2012 using the Twitter Streaming API. Each instance, consisting of the collected user account information and a sample tweet, was manually labeled as spam or non-spam in order to enable precise evaluation of our spam detection algorithms' performance. This process takes a great deal of time due to hourly limitations on requests made to Twitter. While studies such as [15,4] have opted for a larger data set with incomplete results, studies like [13] have chosen smaller data sets to facilitate more thorough analysis. Because we wanted more conclusive results, we elected to use a smaller data set that could be manually labeled.

After being labeled, the instances were divided into two separate sets: training and testing. Each set contained approximately 1500 normal and 100 spam users, which reflects the ratio of spam to non-spam users on Twitter (Table 1). The training set was used to tune the parameters of the algorithms that would be used in the final tests. Each instance was then parsed and represented by a vector of 107 numerical feature values.

The features used to represent each instance fall into three principal categories: content, user information, and tweet text (Table 2). Content-based features include link, mention, and hashtag counts denoted by "http", "@", and "#" respectively. Links are a necessary tool for distributing spam content and as such are included in nearly every spam tweet. The presence of hashtags and mentions in a large percentage of tweets is indicative of a spammer [15] because these tools increase audience size. User-based features include the number of followers, the number of accounts followed, and the follower ratio.

Table 1
Training and testing set data.

	Label	Number of users
Training	Spam	104
	Non-spam	1483
	Total	1587
Testing	Spam	104
	Non-spam	1548
	Total	1652