

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/322158406>

Intelligent Twitter Spam Detection: A Hybrid Approach

Chapter · January 2018

DOI: 10.1007/978-981-10-6916-1_17

CITATIONS

9

READS

487

4 authors:



Varad Vishwarupe

Maharashtra Institute of Technology

14 PUBLICATIONS 19 CITATIONS

[SEE PROFILE](#)



Mangesh Bedekar

Dr. Vishwanath Karad MIT World Peace University

72 PUBLICATIONS 126 CITATIONS

[SEE PROFILE](#)



Milind S Pande

Dr. Vishwanath Karad MIT World Peace University

53 PUBLICATIONS 36 CITATIONS

[SEE PROFILE](#)



Anil S. Hiwale

Dr. Vishwanath Karad MIT World Peace University

29 PUBLICATIONS 114 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



Efficient Solar Charge Controller [View project](#)



Use of children in promotion and advertising [View project](#)

Intelligent Twitter Spam Detection: A Hybrid Approach

Varad Vishwarupe, Mangesh Bedekar, Milind Pande and Anil Hiwale

Abstract Over the years there has been a large upheaval in the social networking arena. Twitter being one of the most widely-used social networks in the world has always been a key target for intruders. Privacy concerns, stealing of important information and leakage of key credentials to spammers has been on the rise. In this paper, we have developed an Intelligent Twitter Spam Detection System which gives the precise details about spam profiles by identifying and detecting twitter spam. The system is a Hybrid approach as opposed to single-tier, single-classifier approaches which takes into account some unique feature sets before analyzing the tweets and also checks the links with Google Safe Browsing API for added security. This in turn leads to better tweet classification and improved as well as intelligent twitter spam detection.

Keywords Twitter • Spam • Machine learning • Google safe browsing
Hybrid classifiers

V. Vishwarupe (✉) · A. Hiwale
Department of Information Technology, MIT College of Engineering, Pune, India
e-mail: varad44@gmail.com

A. Hiwale
e-mail: anil.hiwale@mitcoe.edu.in

M. Bedekar
Department of Computer Engineering, MAEER's MIT, Pune, India
e-mail: mangesh.bedekar@mitpune.edu.in

M. Pande
MIT School of Telecom & Management Studies, Pune, India
e-mail: director@mitsot.com

1 Introduction and Related Work

Twitter is a free Micro-blogging service that allows users to post messages, called tweets, up to 140 characters in length. Its value is in its ease of sharing and accessing user-generated content, including opinions, news, and trending topics. Thus, Twitter provides an opportunity to generate large traffic and revenue, especially since it has hundreds of millions of users.

However, these opportunities make Twitter a prime target of spammers. It is easy for humans to distinguish spammers from actual users, but the existence of spammers wastes user time and attention, puts users at risk in accessing malicious and dangerous content, and devalues Twitter's services and the overall online social network. Some of the related work in the domain of spam detection is highlighted below:

The features extracted from each Twitter account for purpose of spam detection can be categorized into:

- (i) User-based features
- (ii) Content-based features.

User-based features are based on a user's relationships e.g. those whom a user follows (referred to as friends) [1–3], and those who follow a user (referred to as followers) or user behaviors e.g. the time periods and the frequencies when a user tweets.

For content-based features, we use some obvious features. The average length of a tweet, words, syntax etc. [4–8].

2 Problem Statement and Solution Approach

In Intelligent Twitter Spammer Detection we introduce features which exploit the behavioral-entropy, profile characteristics, spam analysis for spammer's detection in tweets. We take a supervised approach to the problem, but leverage existing hashtags in the Twitter data for building training data. Twitter is one such popular network where the short message communication (called tweets) has enticed a large number of users. Spammer tweets pose either as advertisements, scams and help perpetrate phishing attacks or the spread of malware through the embedded URLs [9]. In this system, we fetch twitters tweets for a particular hashtag. Each hashtag may have 1000s of comments and new comments are added every minute, in order to handle so many tweets we are using twitter4j API and perform preprocessing by removing quotes, hash symbols and spam analysis through URL, Number of Unique Mentions (NuMn), Unsolicited Mentions (UIMn), Duplicate Domain Names (DuDn) techniques and Google safe browsing API. The aforementioned features are described in [10] which are quite relevant to our proposed model for classification of tweets. Hence using these unique feature sets and additional

elements conceptualized by us such as AFINN dictionary mapping for some words, account type, etc. a better classifier is developed. This classifier hybrid uses Support Vector Machine (SVM) and Naïve Bayes (NB) for initial classification based on a 8-point feature set explained in 4.4. The malicious URLs if any are checked with Google Safe Browsing API. If a match is found, the tweet is more likely to be a SPAM tweet. A decision is made not only based on the URL found in the tweet but also takes into consideration other factors that might contribute to its classification.

3 System Modelling and Architecture

The above system architecture gives an overview of the Intelligent Twitter Spam Detection System which consists of unique feature vectors having features such as UIMn, NuMn, DuDn and comparison with Google Safe Browsing lists for detection and cross identification of URLs in Spam Tweets (Fig. 1).

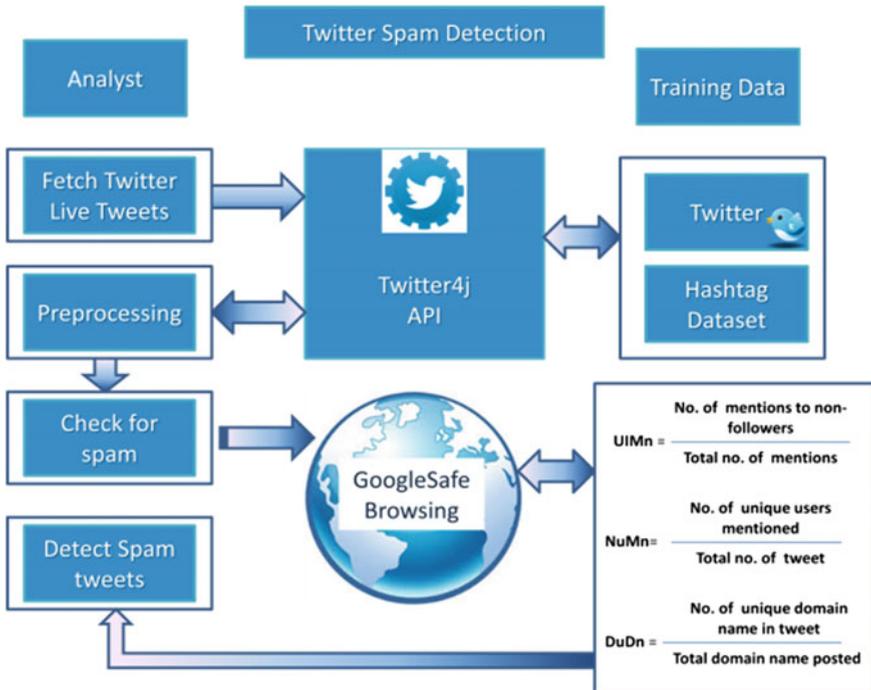


Fig. 1 System architecture

4 Algorithm and Tweet Analysis

4.1 Integrate the System with Twitter

The system will integrate with twitter and able to read the tweets for particular hash tags.

4.2 Data Set for Hashtags

Based on the trending topics on twitter (We used the trending topics in India during a period of 3 weeks in 2016).

4.3 Pre-processing

The first pre-processing technique is remove @ which means it scans the whole document of input dataset and after comparing it with @ it deletes @ from every available comment with @. The next step of pre-processing is remove URL where the whole input document gets scanned and compared with http:\\... and the comments having URL are deleted. Further we move on to stop word removal being the next step in data pre-processing. Stop word removal exactly means that from the whole statement after scanning it removes the words like and, is, the, etc. and only keeps noun and adjective. Tokenization and Normalization are carried out thereafter. Porter Stemmer Algorithm is used thereafter.

4.4 Analysis of Tweets

In this phase we are filtering the tweets on the basis of some criterion Checking that is tweets contains URL or not, if it contains then we are passing that URL to Google Safe Browsing API and getting the spam result from Google Safe Browsing API. Furthermore we are checking that is the tweet contains any spam keyword. We have a list spam keyword which are banned by worldwide social networking organization. The Basic criteria to declare a tweet is spam or not is if a tweet contains both a SPAM URL as well as SPAM WORDS. The steps in the analysis of tweets are as under:

- a. Calculating the duplicate tweet count.
- b. No of Unique Mentions.
- c. Duplicate domain names.
- d. Variance in tweet intervals.
- e. Unsolicited mentions.
- f. AFINN dictionary for finding word and their sense.

- g. Finding negative annotations in the sentence and reverse the weight.
- h. Detecting spam tweets.
- i. Detecting spam accounts.
- j. Blocking the spam accounts.
- k. Finally, positive, negative or neutral count for that particular #hashTag will be calculated.

4.5 Feature Sets

- a. Duplicate tweet count.
- b. No. of Unique Mentions.
- c. Duplicate domain names.
- d. Variance in tweet intervals.
- e. Unsolicited mentions.
- f. Time interval of user being online.
- g. Annotations.
- h. Links to malicious URLs.
- i. Account type: Verified/Unverified.

On the basis of the above constraints we will block/suspend the account.

5 Implementation and Results

In the implementation part of Intelligent Twitter Spam Detection using a Hybrid Approach, we used Twitter4J API, Google Safe Browsing Toolkit, A combination of classifiers including NB and SVM and unique feature sets that provide an intelligent spam detection solution. The system consists of 6 tabs which were designed in NetBeans using JavaSwing for the front end part. The Twitter Spammer tab consists of buttons that Fetch Live Tweets and Recent Twitter Feed related to the trending hashtags. The tweets are then stored on a local dataset which are then loaded for further analysis. Thereafter pre-processing takes place and the tweets are sent to the next stage for implementation of classification algorithms using the Hybrid Approach. A decision is made based on the code that runs on the filtered tweets using the aforementioned feature vectors and the tweets are classified as SPAM. Furthermore there is a provision for showing Trends-Wise Analysis for a particular hashtag in the next stage after which the final stage shows the suspended twitter accounts which were labelled as SPAM by the system. The analysis for this research was conducted by mining 10,782 tweets comprising of 72 hashtags that were trending on twitter. The system was able to classify 2,466 tweets as SPAM while the others were found legitimate. After cross-checking with Google Safe Browsing it was found that 2,153 tweets did contain a malicious URL that re-directed the user to suspicious websites. This affirmed that the accuracy of the said system stands at **87.30%** with this multi-tier approach which is greater than

systems which use a single-classifier and non-hybrid approaches. The snapshots of the system at various stages are as follows (Figs. 2, 3 and 4):

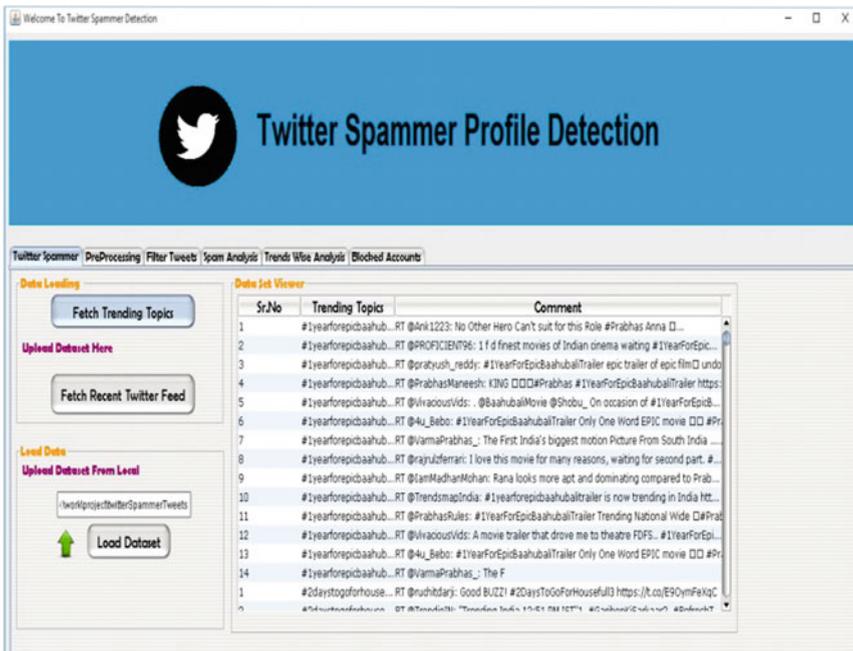


Fig. 2 Phase 1: fetching of tweets and trends

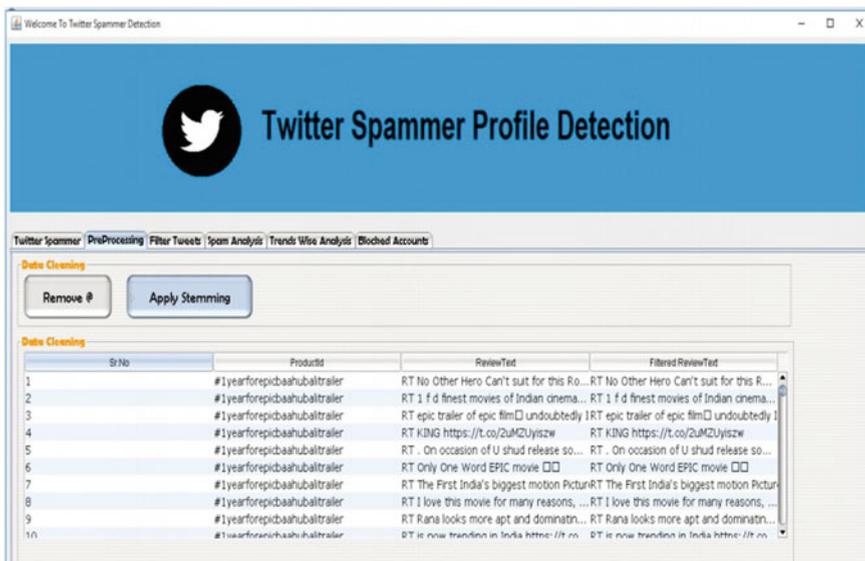


Fig. 3 Phase 2: preprocessing

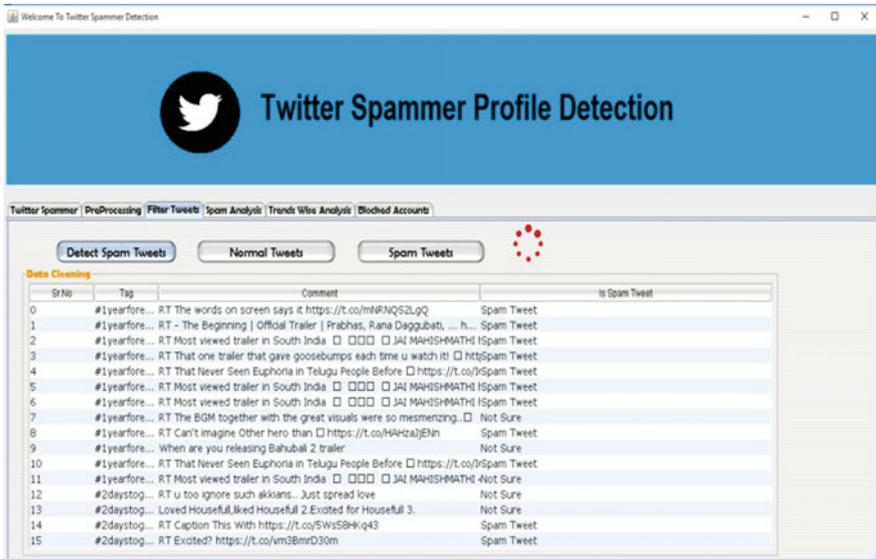


Fig. 4 Phase 3: analysis and detection of SPAM tweets

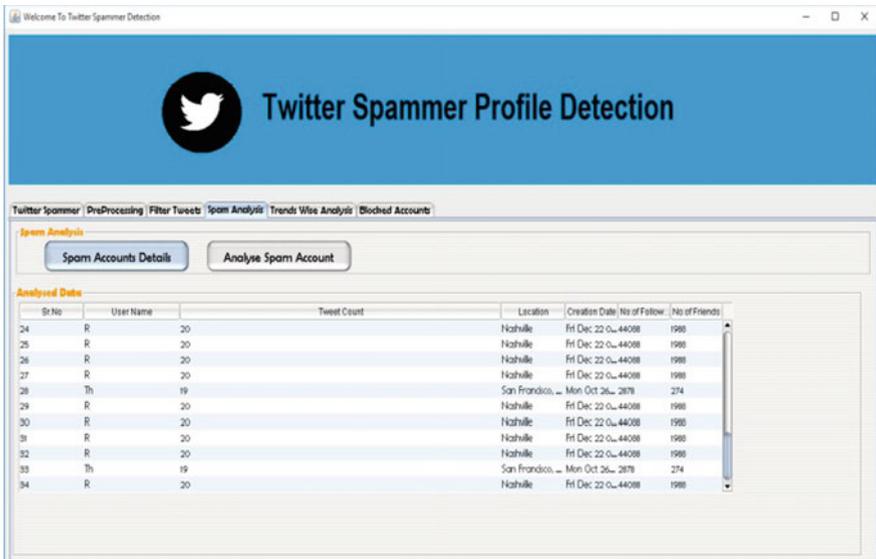


Fig. 5 Phase 4: SPAM account details

After this, the details of SPAM account with the Tweets and Location are displayed (Figs. 5 and 6).

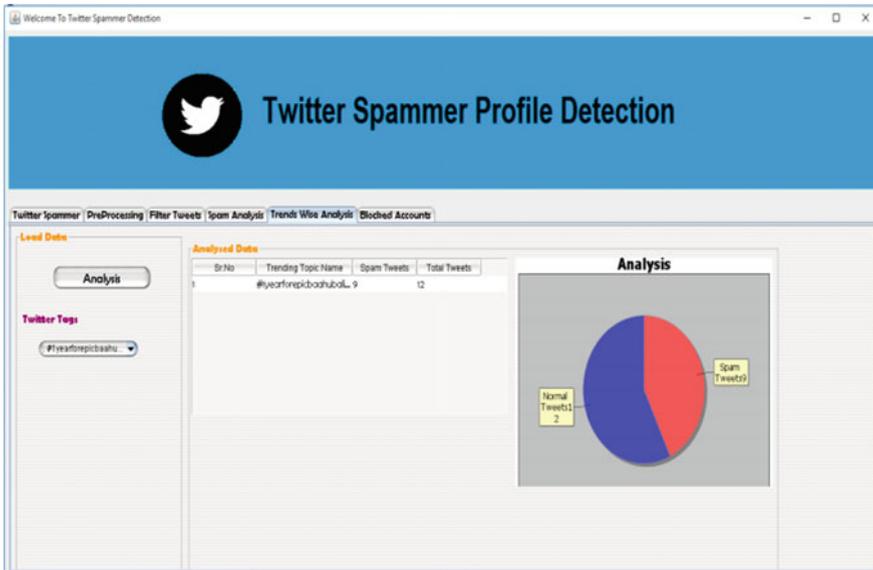


Fig. 6 Phase 5: Trend-wise analysis for a particular hashtag

6 Conclusion and Future Work

Twitter spammer profile detection makes the efficient use of classifiers so as to differentiate between legitimate accounts and spam profiles. Our approach is novel in its own way due to the inclusion of enhanced feature vectors for building the classification which shall be a proponent as compared to the other existing models. Thus it will enhance the user experience on twitter the way for a nuisance free social networking.

Future work in this regard can comprise of mining Tweets to ascertain even more unique feature sets and discover new techniques of classification that can make twitter spam detection even more accurate in the long run.

References

1. Stringhini, G., Kruegel, C., Vigna, G.: Detecting spammers on social networks. In: Proceedings of the 26th Annual Computer Security Applications Conference, pp. 1–9. ACM (2010)
2. Sala, A., Cao, L., Wilson, C., Zablit, R., Zheng, H., Zhao, B.: Measurement-calibrated graph models for social network experiments. In: Proceedings of the 19th International Conference on World Wide Web (WWW'10), pp. 861–870. ACM (2010)

3. Ahmed, F., Muhammad, A.: An MCL-based approach for spam profile detection in online social networks. In: 2012 IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications, pp. 1–7
4. Halim, Z., Gul, M., ul Hassan, N., Baig, R., Rehman, S., Naz, F.: Malicious users' circle detection in social network based on spatio-temporal co-occurrence. In: 2011 International Conference on Computer Networks and Information Technology (ICCNIT), pp. 35–39, July 2011
5. Perez, C., Lemercier, M., Birregah, B., Corpel, A.: SPOT 1.0: scoring suspicious profiles on Twitter. In: 2011 International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 377–381. IEEE (2011)
6. Tian, X., Guang, Y., Peng-yu, L.: Spammer detection on Sina micro-blog. In: 2014 (21st) International Conference on Management Science and Engineering, pp. 1–6, 17–19 August 2014
7. Qunyan, Z., et al.: Duplicate detection for identifying social spam in microblogs. In: 2013 IEEE International Congress on Big Data (BigData Congress) (2013)
8. Amleshwaram, A.A., et al.: CATS: characterizing automation of Twitter spammers. In: 2013 Fifth International Conference on Communication Systems and Networks, New York. IEEE (2013)
9. Chen, C., et al.: Battling the internet water army: detection of hidden paid posters. In: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. ACM (2013)
10. Joshi, S., et al.: Detection of spam tweets by using machine learning. *Int. J. Adv. Res. Comput. Sci. Softw. Eng.* 7(4). ISSN: 2277 128X