Intelligent Twitter Spam Detection: A Hybrid Approach

Conference Paper · February 2017

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.Vishawnath 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

Intelligent Twitter Spam Detection:
A Hybrid Approach

Varad Vishwarupe[1], Mangesh Bedekar[2], Milind Pande[3], Anil Hiwale[1],

[1] Department of Information Technology, MIT College of Engineering, Pune, India
[2] Department of Computer Engineering, MAEER's MIT, Pune, India
[3]MIT School of Telecom & Management Studies, Pune, India
{ varad44@gmail.com, mangesh.bedekar@mitpune.edu.in,
director@mitsot.com,anil.hiwale@mitcoe.edu.in }

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.

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.

Twitter users have different levels of awareness with respect to security threats hidden in social networking sites. For example, a previous study has showed that 45% of users on a social networking site readily click on links posted by any friend in their friend lists' accounts, even though they may not know that person in real life. Thus, spammers are attracted to use Twitter as a tool to send unsolicited messages to legitimate users, post malicious links, and hijack trending topics.

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 follow (referred to as friends), 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. Additional content-based features are described in subsequent subsections [4].

Based on the above identified features, we proceed to use traditional classifiers to help detect spammers. In this work, several classic classification algorithms such as Random Forest, Naïve Bayesian, Support Vector Machines, and K nearest neighbors are compared. The Random Forest classifier is known to be effective in giving estimates of what variables are important in the classification. This classifier also has methods for balancing error in class population unbalanced data sets. The naïve Bayesian classifier is based on the well- known Bayes theorem. The big assumption of the naïve Bayesian classifier is that the features are conditionally independent although research shows that it is surprisingly effective in practice without the unrealistic independence assumption[5-8].

2 Problem Statement and Solution Approach

In 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.[9-10] 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 twiter4j 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.

3 System Modelling and Architecture

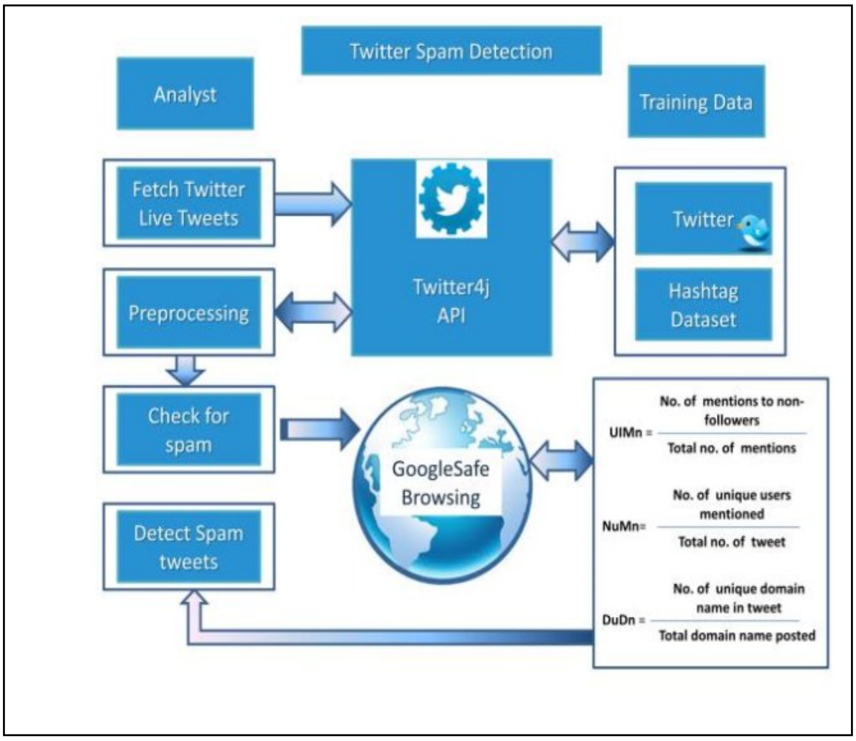

Twitter Spam Detection
Analyst
Training Data
Fetch Twitter Live Tweets
Twitter4j API
Twitter
Hashtag Dataset
Preprocessing
Check for spam
GoogleSafe Browsing
Detect Spam tweets
UIMn = No. of mentions to non-followers / Total no. of mentions
NuMn= No. of unique users mentioned / Total no. of tweet
DuDn = No. of unique domain name in tweet / Total domain name posted

Fig 1. System 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.

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. #HashTagging data set

To create the hash tagged data set, we first filter out duplicate tweets, non-English tweets, and tweets that do not contain #hashTags. From the remaining set (about 4 million), we investigate the distribution of #hashTags and identify what we hope will be the sets of frequent #hashTags that are indicative of positive, negative and neutral messages. These #hashTags are used to select the tweets that will be used for development and training.

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. The Porter stemming algorithm (or 'Porter stemmer') is a process for removing the commoner morphological and inflexional endings from words in English. Its main use is as part of a term normalization process that is usually done when setting up Information Retrieval systems.

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 & 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.

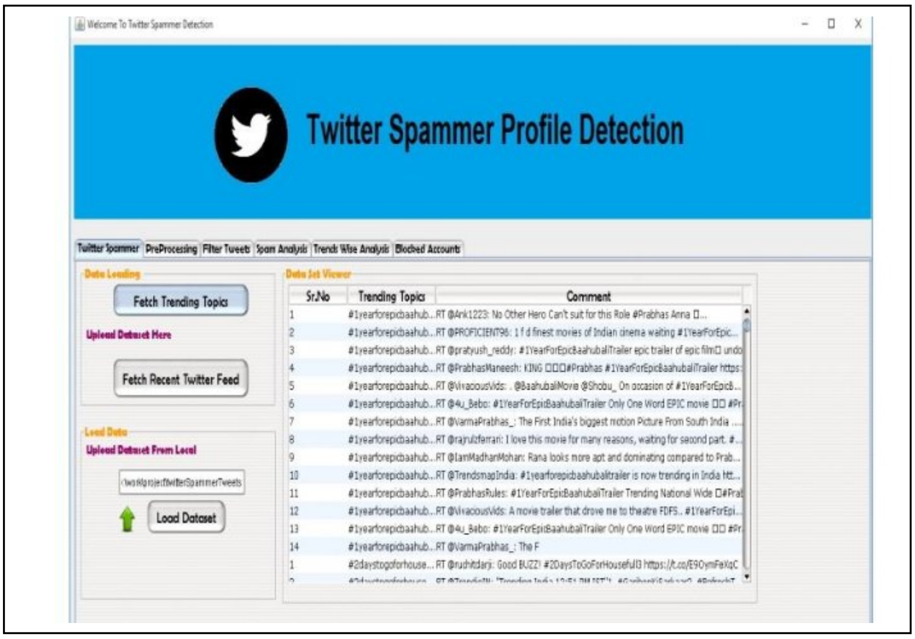
4.4 Feature Sets:

a. Calculating the duplicate tweet count
b. No of Unique Mentions
c. Duplicate domain names
d. Variance in tweet intervals
e. Unsolicited mentions

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

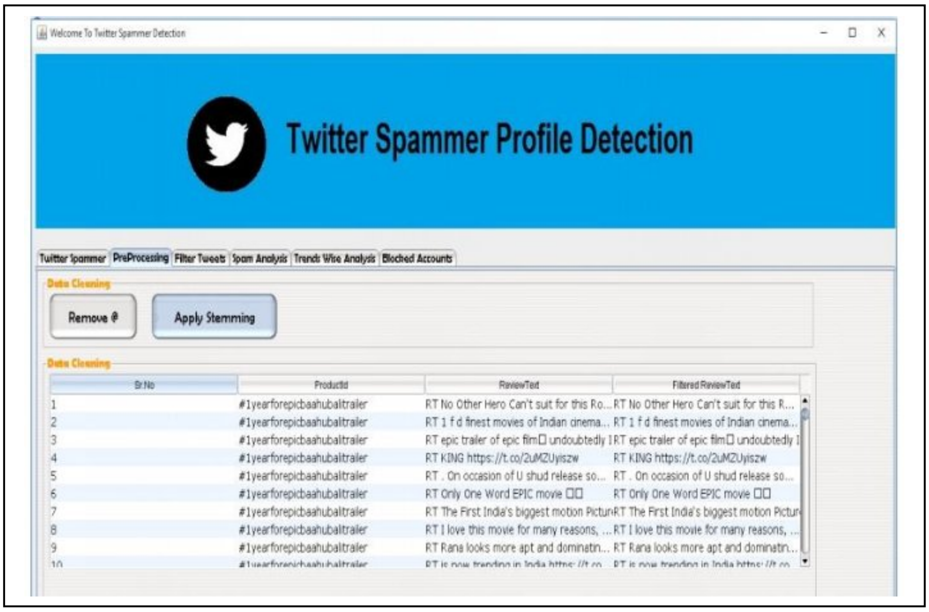
5 Implementation and Results

In the implementation part of Intelligent Twitter Spam Detection using a Hybrid Approach, we used Twitter 4J 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:



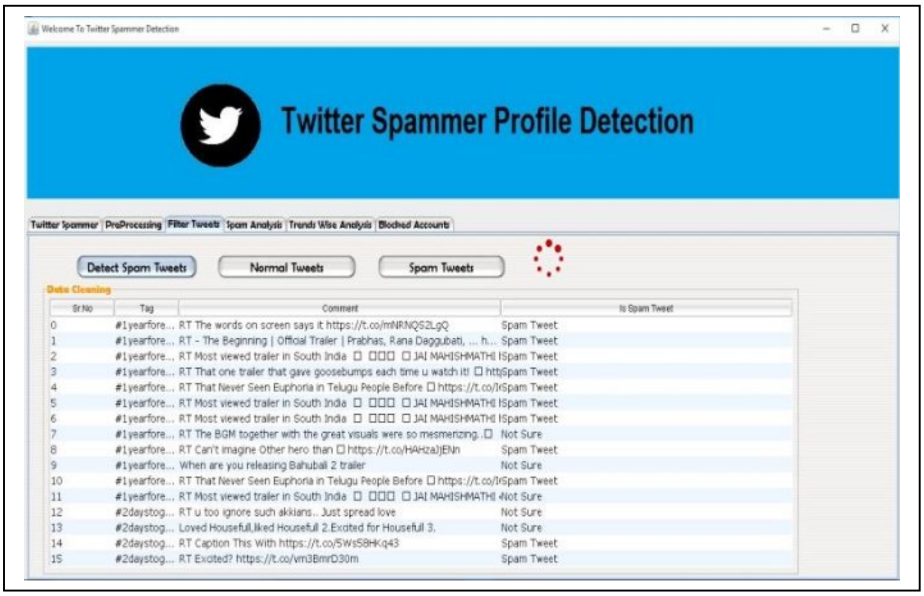
Sr.No	Trending Topics	Comment
1	#1yearforepicbaahub...	RT @Ank1223: No Other Hero Can't suit for this Role #Prabhas Anna □...
2	#1yearforepicbaahub...	RT @PROFICIENT96: 1 f d finest movies of Indian cinema waiting #1YearForEpic...
3	#1yearforepicbaahub...	RT @pratyush_reddy: #1YearForEpicBaahubaliTrailer epic trailer of epic film□ undo
4	#1yearforepicbaahub...	RT @PrabhasManeesh: KING □□□#Prabhas #1YearForEpicBaahubaliTrailer https:
5	#1yearforepicbaahub...	RT @VivaciousVids: . @BaahubaliMovie @Shobu_ On occasion of #1YearForEpicB...
6	#1yearforepicbaahub...	RT @4u_Bebo: #1YearForEpicBaahubaliTrailer Only One Word EPIC movie □□ #Pr
7	#1yearforepicbaahub...	RT @VarmaPrabhas_: The First India's biggest motion Picture From South India ...
8	#1yearforepicbaahub...	RT @rajrulzferrari: I love this movie for many reasons, waiting for second part. #...
9	#1yearforepicbaahub...	RT @IamMadhanMohan: Rana looks more apt and dominating compared to Prab...
10	#1yearforepicbaahub...	RT @TrendsmapIndia: #1yearforepicbaahubalitrailer is now trending in India htt...
11	#1yearforepicbaahub...	RT @PrabhasRules: #1YearForEpicBaahubaliTrailer Trending National Wide □#Pra
12	#1yearforepicbaahub...	RT @VivaciousVids: A movie trailer that drove me to theatre FDFS.. #1YearForEpi...
13	#1yearforepicbaahub...	RT @4u_Bebo: #1YearForEpicBaahubaliTrailer Only One Word EPIC movie □□ #Pr
14	#1yearforepicbaahub...	RT @VarmaPrabhas_: The F
1	#2daystogoforhouse...	RT @rudhitdarji: Good BUZZ! #2DaysToGoForHousefull3 https://t.co/E9OymFeXqC

Fig 2. Phase 1: Fetching of Tweets and Trends



Sr.No	ProductId	ReviewText	Filtered ReviewText
1	#1yearforepicbaahubalitrailer	RT No Other Hero Can't suit for this Ro...	RT No Other Hero Can't suit for this R...
2	#1yearforepicbaahubalitrailer	RT 1 f d finest movies of Indian cinema...	RT 1 f d finest movies of Indian cinema...
3	#1yearforepicbaahubalitrailer	RT epic trailer of epic film□ undoubtedly I	RT epic trailer of epic film□ undoubtedly I
4	#1yearforepicbaahubalitrailer	RT KING https://t.co/2uMZUyiszw	RT KING https://t.co/2uMZUyiszw
5	#1yearforepicbaahubalitrailer	RT . On occasion of U shud release so...	RT . On occasion of U shud release so...
6	#1yearforepicbaahubalitrailer	RT Only One Word EPIC movie □□	RT Only One Word EPIC movie □□
7	#1yearforepicbaahubalitrailer	RT The First India's biggest motion Pictur	RT The First India's biggest motion Pictur
8	#1yearforepicbaahubalitrailer	RT I love this movie for many reasons, ...	RT I love this movie for many reasons, ...
9	#1yearforepicbaahubalitrailer	RT Rana looks more apt and dominatin...	RT Rana looks more apt and dominatin...
10	#1yearforepicbaahubalitrailer	RT is now trending in India https://t.co...	RT is now trending in India https://t.co...

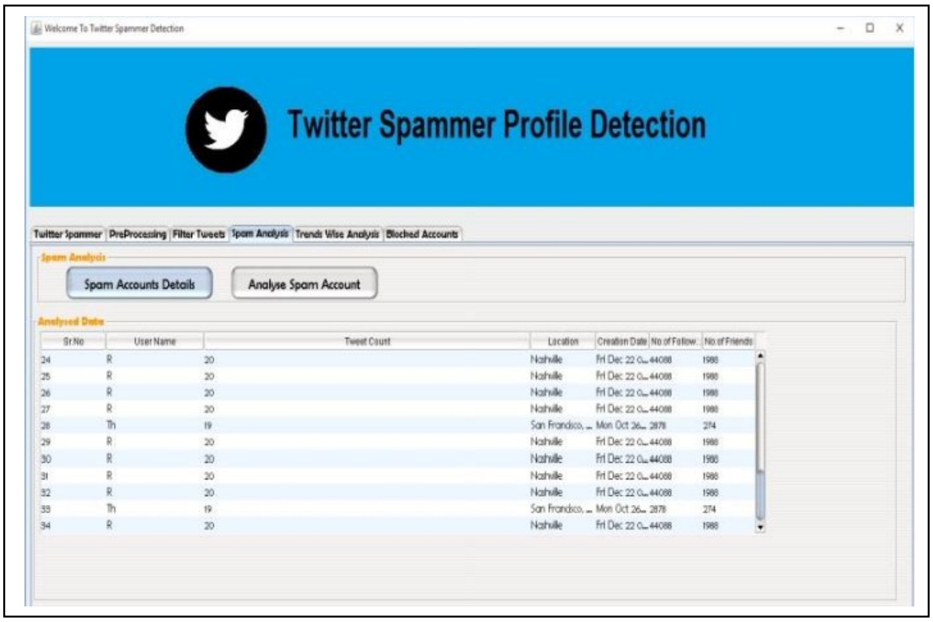
Fig 3. Phase 2: Preprocessing



Sr.No	Tag	Comment	Is Spam Tweet
0	#1yearfore...	RT The words on screen says it https://t.co/mNRNQS2LgQ	Spam Tweet
1	#1yearfore...	RT - The Beginning | Official Trailer | Prabhas, Rana Daggubati, ... h...	Spam Tweet
2	#1yearfore...	RT Most viewed trailer in South India □ □□□ □JAI MAHISHMATHI I	Spam Tweet
3	#1yearfore...	RT That one trailer that gave goosebumps each time u watch it! □ htt	Spam Tweet
4	#1yearfore...	RT That Never Seen Euphoria in Telugu People Before □ https://t.co/I	Spam Tweet
5	#1yearfore...	RT Most viewed trailer in South India □ □□□ □JAI MAHISHMATHI I	Spam Tweet
6	#1yearfore...	RT Most viewed trailer in South India □ □□□ □JAI MAHISHMATHI I	Spam Tweet
7	#1yearfore...	RT The BGM together with the great visuals were so mesmerizing..□	Not Sure
8	#1yearfore...	RT Can't imagine Other hero than □ https://t.co/HAHzaJjENn	Spam Tweet
9	#1yearfore...	When are you releasing Bahubali 2 trailer	Not Sure
10	#1yearfore...	RT That Never Seen Euphoria in Telugu People Before □ https://t.co/I	Spam Tweet
11	#1yearfore...	RT Most viewed trailer in South India □ □□□ □JAI MAHISHMATHI	Not Sure
12	#2daystog...	RT u too ignore such akkians.. Just spread love	Not Sure
13	#2daystog...	Loved Housefull,liked Housefull 2.Excited for Housefull 3.	Not Sure
14	#2daystog...	RT Caption This With https://t.co/5WsS8HKq43	Spam Tweet
15	#2daystog...	RT Excited? https://t.co/vm3BmrD30m	Spam Tweet

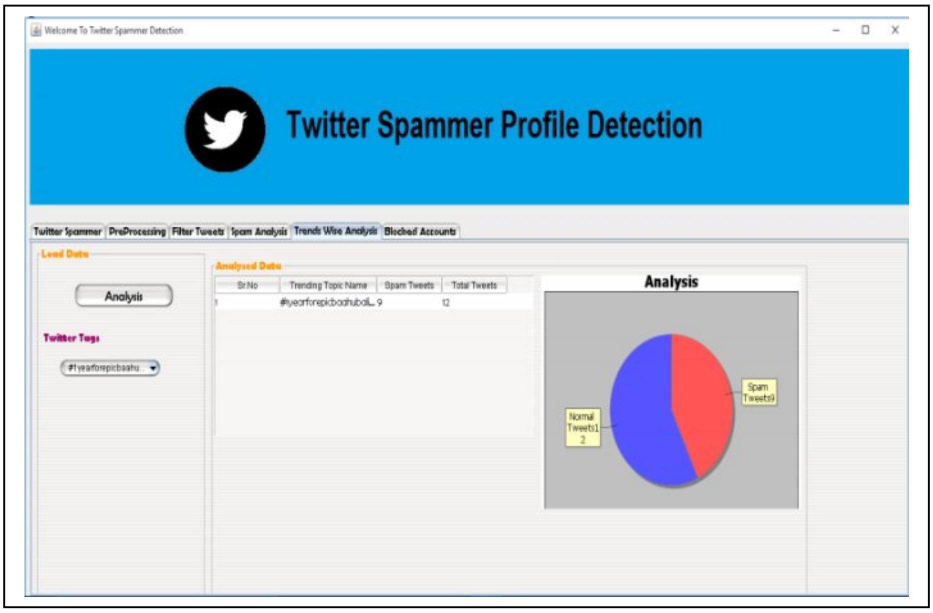
Fig 4. Phase 3: Analysis and Detection of SPAM tweets

After this, the details of SPAM account with the Tweets & Location are displayed.



Sr.No	User Name	Tweet Count	Location	Creation Date	No of Follow..	No of Friends
24	R	20	Nashville	Fri Dec 22 0...	44088	1988
25	R	20	Nashville	Fri Dec 22 0...	44088	1988
26	R	20	Nashville	Fri Dec 22 0...	44088	1988
27	R	20	Nashville	Fri Dec 22 0...	44088	1988
28	Th	19	San Francisco, ...	Mon Oct 26...	2878	274
29	R	20	Nashville	Fri Dec 22 0...	44088	1988
30	R	20	Nashville	Fri Dec 22 0...	44088	1988
31	R	20	Nashville	Fri Dec 22 0...	44088	1988
32	R	20	Nashville	Fri Dec 22 0...	44088	1988
33	Th	19	San Francisco, ...	Mon Oct 26...	2878	274
34	R	20	Nashville	Fri Dec 22 0...	44088	1988

Fig. 5 Phase 4: SPAM Account Details



Sr.No	Trending Topic Name	Spam Tweets	Total Tweets
1	#1yearforepicbaahubali...	9	12

Fig 6. Phase 5: Trends 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. G. Stringhini, C. Kruegel, and G. Vigna, “Detecting spammers on social networks,” in Proceedings of the 26th Annual Computer SecurityApplications Conference. ACM, 2010, pp. 1–9.
2. A. Sala, L. Cao, C. Wilson, R. Zablit, H. Zheng, and B. Zhao, “Measurement-calibrated graph models for social network experiments,” in Proceedings of the 19th International Conference on World Wide Web (WWW’10). ACM, 2010, pp. 861–870.
3. F. Ahmed, A. Muhammad, An MCL-Based Approach for Spam Profile Detection in Online Social Networks 2012 IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications, pp 1-7
4. Z. Halim, M. Gul, N. ul Hassan, R. Baig, S. Rehman, and F. Naz, “Malicious users’ circle detection in social network based on spatio-temporal co-occurrence,” in Computer Networks and Information Technology (ICCNIT), 2011 International Conference on, July, pp. 35–39.
5. C. Perez, M. Lemercier, B. Birregah, and A. Corpel, “SPOT 1.0: Scoring Suspicious Profiles on Twitter,” in Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on. IEEE, 2011, pp. 377–381.
6. X. Tian ,Y. Guang, L.Peng-yu, “Spammer Detection on Sina Micro-Blog”, 2014 International Conference on Management Science & Engineering (21st), August 17-19, 2014,pp 1-6
7. Qunyan, Z., et al. Duplicate Detection for Identifying Social Spam in Microblogs. 2013 IEEE International Congress on Big Data (BigData Congress), 2013.
8. Amleshwaram, A. A., et al. CATS: Characterizing Automation of Twitter Spammers. 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. Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ACM, 2013.
10. Gao, H., et al. Detecting and characterizing social spam campaigns. Proceedings of the 10th ACM SIGCOMM conference on Internet measurement, ACM, 2010.