International Journal of Latest Engineering and Management Research (IJLEMR)
ISSN: 2455-4847
www.ijlemr.com || Volume 04 - Issue 05 || May 2019 || PP. 122-126

Spam – Ham Classification along with Bad URL Identification in Social Networking Sites

Sreelakshmi K.U[1], Abey Abraham[2]
[1,2]*Department of Information Technology,*
Rajagiri School of Engineering and Technology,
Ernakulam, Kerala

Abstract: Social Media or Social Networking sites is one which offers a great platform to its users. People can share their views, opinions etc. through these SNS without any barrier or restriction. Because of its popularity fraudulent activities are also increasing which focuses the users who are not aware of social media attacks. Spam content and false URL is one among this.in this paper we are proposing a best method which can detect Spam data as well as Bad URL. Twitter is taken as the example for this research. Twitter data is collected and the proposed algorithm shows better results with great accuracy and precision.

Index Terms: Social Networking Sites, Social Networking Communities, Spam- Ham Classification, Bad URL

I. INTRODUCTION

The wide popularity and technological benefits of social media make more and more people attracted towards it. In this world it is easy to connect people digitally rather than geographically. This is why there is a considerable growth in the use of social media usage. Nowadays nearly everyone have account in any of the social media like Facebook, Twitter, and LinkedIn etc. Even though these sites use great technologies and security policies the attackers will find more vulnerability and crash the system.

The users of online social networks can create account and shares data like photos, videos, text messages etc. The users can be either legitimate users or spammers. Legitimate users are the victim of spammers. Spammers find multiple ways to attract the trusted or normal users. The revolution in smart devices is also a reason behind the cyber-attacks. Among these smart device users majority are not aware of the spamming activities inside social media.

Spam is electronic junk mail or junk newsgroup postings. Some people define spameven more generally as any unsolicited email. However, if a long-lost brother finds your email address and sends you a message, this could hardly be called spam, even though it is unsolicited [1]. Spam content is difficult to detect because it will initially looks like normal data but some of its properties make it spam. It will appear either as bulk messages or it may contain malicious URLs.

Twitter is one of the social networking sites where users communicate through short messages called tweets. Only registered users can tweet but others can read the tweet. One of important thing to be noticed is that the only 140 characters are permitted for one tweet. This is why people insert URLs in tweets. This is to get more attention from others. Attackers will send spam tweets along with malicious URL. Whenever a person clicks on this link it may proceeded to a phishing site or any other malicious site.

URL is an address to a content or resource in World Wide Web. In the case of malicious URL the link will leads to potential damages. Irrelevant or uninvited messages sent over the Internet which aims to reach typically a large number of users. The high click rate and the effective message propagation make social media as an attractive platform for spammers. Increase in spamming activities affects the people who are using social media adversely.

Spam – ham classification separates tweets into either spam or ham. Ham tweets are nothing but the tweets which doesn't contain any spam words. The proposed method classifies the tweets into wither spam or ham and the URL contained in it is good or bad.

II. RELATED SEARCH

In the paper "Detecting spam and promoting campaigns in twitter ", [2] spam and promoting campaigns are detected using URL based methods. The framework consists of three steps: the first step links accountswho post URLs for similar purposes; the second step extracts candidate campaigns that may be for spam or promotion purposes; and the third step classifies the candidate campaigns into normal, spam, and promotion groups. The key point of the framework is how to measure the similarity between accounts' purposes of posting URLs. Here present two measure methods based on Shannon information theory: the first one uses the URLs posted by the users, and the second one considers both URLs and timestamps.

In the paper " Towards detecting malicious links in online social networks through user behavior "[3] , The suggested model is driven by a user's interest. This approach is inspired by online personalization, where recommender systems focus mainly on the user's interest in order to recommend relevant topics or products to them. In OSNs, involving in any activities is completely driven by user interest and social habit. Data collection, feature extraction and classification are the main components in this framework.URL Classification Classifiers will be constructed based on features extracted from user profiles and messages together with features of URLs and OSNs with the hope of identifying malicious URLs with a lower rate of false positives.

Email Spam Classification using Neighbor Probability based Naïve Bayes Algorithm[6] by P.V Anitha and C.V Guru Rao they classifies into ham and spam emails with an optimized and well efficient classification technique. Ham holds emails that are legitimate or legally valid message can get accepted by users. Spam emails are unwanted emails that a user doesn't want and to get rid of it. This study emphasizes on the improvement in classifying all mails into these two groups with minimal requirement of training and with an accuracy of hundred percent. Here in this study, Modified Naïve Bayes (MNB) classifier ensured the requirements with very low percentage of training and produces accurate results than existing Naïve Bayes (NB) or Supporting Vector Machine (SVM) classifier

Table 1 shows the list of general spam words. The tweets or twitter direct messages which contain these type of words or if the probability of these words is high in the given text then that will be a spam content. During tweet extraction tweets which contain the spam words are focused.

F r e e	Fast cash	For just $XXX
Hidden assets	hidden charges	Income
Incredible deal	Insurance	Investment
Acceptance	Accordingly	Avoid
Chance	Dormant	Freedom
Here	Hidden	Home
Leave	Lifetime	Lose
Maintained	Medium	Miracle
Never	Passwords	Problem
Remove	Reverses	Sample
Satisfaction	Solution	Stop
Loans	Lowest price	Million dollars
Success	Teen	Wife

Table 1. List of general spam words

II. PROPOSED METHOD

Proposed system mainly contain two major module one is spam-ham classification module and URL classification module

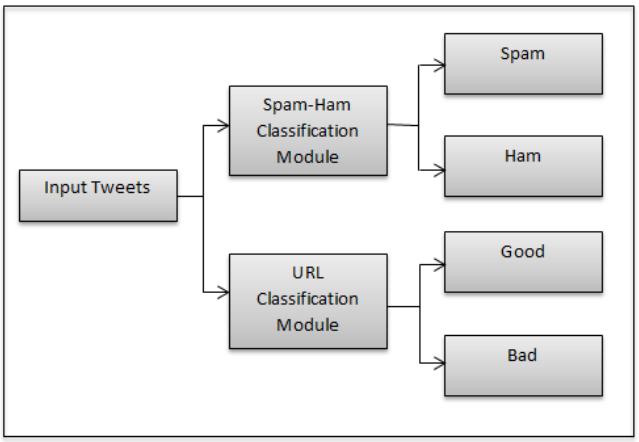

Figure 1 Module classification

The above figure shows the major modules of the proposed scenario. First module spam-ham classification module as the name implies it classifies the tweets into either spam or ham using TF-IDF probability. Content based spam detection is used here. In URL classification module the URL contained in the tweet is good or bad. Bad URL means it may lead to phishing site or dead end.

A. Spam – Ham Classification Module

The system begins with spam- ham classification module. After successful classification URL classification module is is triggered.

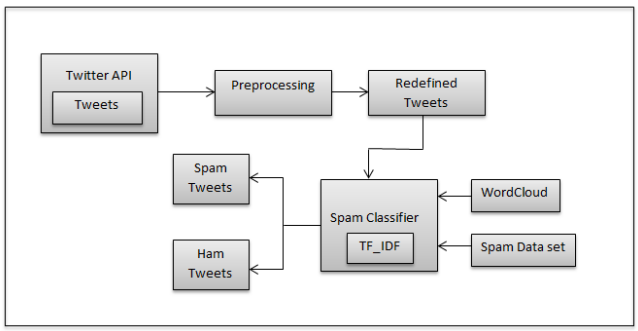

Figure 2. Spam-Ham Classification Module.

Figure 2 shows the details of the spam-ham classification module. The first component is the Twitter API. Input to the system is twitter data. Tweets are needs to be collected. For this an account is created in twitter application and after that they will provide credentials for collecting twitter data Tweets are collected using random hashtags and stored in CSV file. This file is the testing input that is after spam- ham classification another output CSV file is generated and it contains another additional column which labels the tweets into either spam or ham.

Preprocessing means separation of text and URL using regular expression. After this process redefined tweets that are the text part is considered first for next stage.Word Cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance. Significant textual data points can be highlighted using a word cloud. Word clouds are widely used for analyzing data from social network websites. The word cloud is generated from the spam data set. Spam data set contains spam message which is already categorized into either spam or ham. Messages which have label 1 is considered as spam and those have label 0 I considered as ham. Ham is nothing but data which doesn't show any spam behavior is called ham. Word cloud visualizes spam words as well as ham words using matplotlib.

Words from the data set which are labeled as spam in the data set are assumed to be spam words. By considering the frequency of those words word cloud of spam words is generated and same is done for ham words. The words which appear in bigger font denote its frequency. That means if the test data contain these bigger font words then that data is assumed to be more spam.

Figure 3 Example for spam word cloud

Spam classifier is used for the classification of spam and ham. It uses TF-IDF for classification. TF-IDF stands for term frequency-inverse document frequency, and the TF-IDF weight is a weight often used in

information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears TF-IDF weighting scheme are often used by search engines as a central tool in scoring andranking a document's relevance given a user query. TF-IDF can be successfully used for stop-words filtering in various subject fields including text summarization and classification.

A spam data set which contains spam and ham data is split into test and train data randomly. Using this word cloud is generated. So eventually the word cloud contains spam and ham words. Then a new input file is considered for the classification. The input file is csv file. Each row is processed and TF-IDF is calculated. After TF-IDF calculation probability of each spam and ham data is calculated. If the probability of spam is more than probability of ham then that sentence (here it is tweet) is classified in to spam otherwise ham.

B. URL Classification Module

In content based spam classification when a tweet contain more spam words then that will be classified as spam. But in some situations this may not be a correct classification. For example "Congratulations you are rewarded 10000000 click here to grab your prize money" is most probably a spam tweet. The words like congratulations, rewarded are mostly appearing in spam messages or emails. So according to this algorithm these words are assumed to be spam words. The problem arises when a genuine content contains these words. Because of the presence of those words this tweets will be also categorized as spam tweet. To avoid this disadvantage URL Classification module is introduced.

Tweets have one limitation. That is it should not be more than 140 characters. So spammers will tries to insert URL to get more attention. One way of spreading spam tweets is they will send bulk tweets within short time interval. They will do this wither manually or with the help of machines. When a user click on the URL which is provided with spam tweet it will leads to serious issue. Sometimes it can be phishing sites, malicious or blacklisted sites, and dead sites. So if a tweet is classified as spam and the URL contained in it is a bad URL then we can confirm that it is spam tweet. Thus accuracy can be increased.

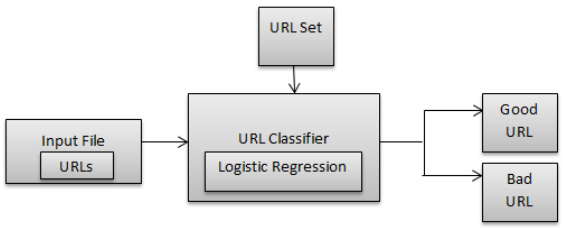

Figure 4 URL Classification Module

Tokenization is the first step in this classification. Slash, dot, com are removed from the URLs and make it into tokens. There is a URL set which is already labeled as good or bad.

By considering this URL set and logistic regression score the URL is classified into either good or bad. URLs which end up in a good result are considered as good URLs. These URLs can be trusted and doesn't create any issue. Bad URL mostly goes to dead end. It will show message like "Sorry this page cannot be displayed". Even though the displayed message is like this there can be chance for background activities

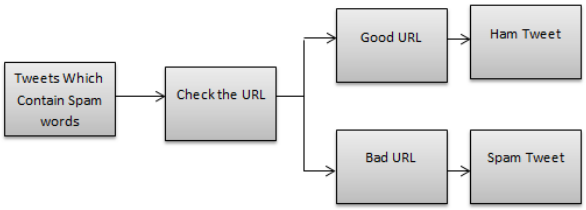

Figure 5 Final decision

After successful classification of tweets the one which contains spam words are further considered for the URL Classification. URL from spam tweet is fetched and goes proceed to the second verification that is URL module. If a tweet contain spam words and bad URL then that tweet is assumed to a spam tweet otherwise it is classified in to a ham tweet. Figure 5 explains this scenario.

III. CONCLUSION

Social Networking Sites or simply social media is one of the popular things which provides best platform for its users to share their views, ideas, opinion etc. Because of its popularity and ease of use more and more people are attracted towards the social media like twitter, Facebook, LinkedIn. The rapid increase in usage of smart devices is one of the reasons behind this. Among 100 people who are using smart phones or smart devices at least 80 have account in any of these sites. Out of this some have good technical knowledge but others don't have.

Even though social media developers uses better technologies and security mechanism attackers also find vulnerability to crash the system. Sometimes it can be through a spam message, fake or malicious URL or through games. Whenever a person got a message or post which offers gifts or money he or she may believe that it is correct and shares or follows the instructions in it.

In the proposed algorithm twitter is taken as an example. Tweets are collected and processed in order to identify the spam tweets. To attain efficiency and for better accuracy URL contained in the tweet is also evaluated and the final result is determined by the output received from the two modules in the algorithm that is spam-ham classification module and URL classification module.

REFERENCES

[1] https://www.webopedia.com/TERM/S/spam.htmlaccessed on 20/05/2019
[2] Luis Alberto B. Pacheco, Jo˜ao J. C. Gondim, Priscila A. Solis Barreto and Eduardo Alchieri , "Detection of spam promoting campaigns " , IEEE 15th International Symposium on Network Computing and Applications , 2016
[3] Bandar Alghamdi, Jason Watson, YueXu, "Toward Detecting Malicious Links in Online Social Networks through User Behavior " International Journal of Advanced Research in Computer Science and Software Engineering , 2016.
[4] Nupur S. Gawale, Nitin N. Patil, " Implementation of A System To Detect Malicious URLs for Twitter Users" , IEEE 14th Intl Conf on Dependable, Autonomic and Secure Computing, 14th Intl Conf on Pervasive Intelligence and Computing
[5] Prabhakaran Kasinathan, Claudio Pastrone, Maurizio A. Spirito , Mark Vinkovits ," Spam detection using deep learning" , IEEE 9th International Conference on Wireless and Mobile Computing, Networking and Communications 2013.
[6] P. U. Anitha, C. V. Guru Rao "Email Spam Classification using Neighbor Probability based Naïve Bayes Algorithm attack " , 39th annual IEEE conference on Local Computer Networks, 2014.