DISCERNING SPAM IN SOCIAL NETWORKING SITES

Sarita Yadav[1], Aakanksha Saini[1], Akanksha Dhamija[1] and Yoganta Narnauli[1]

[1]Department of Information Technology, GGSIPU, New Delhi, India

ABSTRACT

Social Networking Sites, in the present scenario, are an amalgam of knowledge and spam. As their popularity surges among the users day by day so does it among the spammers looking at easy targets for their campaigns. The threat due to spams causing atrocious harm to the bandwidth, overloading the servers, spreading malicious pages online et cetera has increased manifold making it necessary for researchers to foray into this field of spam detection and reduce their effect on the various social networking sites.

In this paper, we propose a framework for spam detection in the two largest social networking sites namely, Twitter and Facebook. We'll be utilizing the data publically available on these two giants of social networking era. Initially, we'll be citing the various approaches that have already been explored in this field. After that we'll briefly explain the two methods that we used to collect the datasets from these websites.

KEYWORDS

API's, Honeypots, Facebook, Social Networking websites, Spam, Twitter, SVM, Weka, Naïve Bayes Algorithm, Simple K means clustering.

1. INTRODUCTION

There are a large number of social networking sites booming on the internet these days. Some of these platforms are more popular than the other like Facebook and Twitter. This increase in social sites has made social media vulnerable to many kinds of online attacks. Among these, one of the leading problems engulfing the netizens is spam. Spam on online social sites or on any social media may include messages in bulk or repetition of messages, malicious links, and fake friend requests et cetera. This not only uses extra bandwidth but also as the volume of spam increases, the internet becomes more polluted and less useful. Earlier, spamming was carried out through emails, but now they have expanded their approach. Social websites which are used by users for communicating with one another is being targeted. Spam is increasing being used to distribute viruses, links to phishing websites et cetera. This has now become a security threat.

According to CNN, there are 83 million fake profiles on Facebook. This is the amount of spam covering one of the most popular social website. A study [1] shows that Facebook has 100 times more spam than any other social networks and 4 times more phishing attacks. According to Symantec International Security Threat Report 2014 [2], Adult Spam (70%) dominated in 2013. These are usually in the form of email inviting the user to connect to the scammer or a URL link. Such a scenario proves to be extremely harmful and dangerous for the young minds who wish to surf the internet dominantly for academic purposes.

DOI : 10.5121/avc.2016.3201

We are presenting this research paper with the aim to detect spam in the leading social networking sites Twitter and Facebook using unsupervised as well as supervised methods. Our objective of presenting the research paper titled "Discerning Spam in Social Networking Sites" is basically to diagnose the huge amount of spam available on the social websites which is predominantly used by each and every one of us so that this worrisome issue can be tackled.

2. BACKGROUND

2.1. TWITTER

Twitter is simple social website which gives access to its users for sending messages (called tweets) and to follow other users. It displays usernames on their profiles and their recent tweets. There were 307 million active twitter users in the first quarter of 2015 [3] and there are at least 2.5 million spam tweets every day. According to a report [4], almost 10% of twitter is spam! This indicates the level of unwanted and harmful material present on one of the trending social networking online websites which is used by millions of users worldwide. For this problem, twitter has taken a number of steps in the past. They've introduced "Reporting spam" option which could be used by users if they find any doubtful material on their site. Users can also flag any content which they find inappropriate. Some of these spams contain malicious links to dubious websites which tricks the users and their computers into thinking that it is all legit content. Unnecessary and unwanted tweets creates a lot of crowd and ultimately the user gets confused about what is real and what is not on the internet. These spammers take the advantage of these occasions and illegally and unknowingly collect all the data of the user on which they can put their hand upon. So, it is the need of the hour to detect the sources of these spams and take necessary measures so that a user can have a hassle free experience and their security remains the same.

2.2. FACEBOOK

Facebook is an online social networking website which people use to keep in touch with their friends, family, colleagues by posting statuses, uploading pictures, sharing links with one other, liking pages of their interest, joining public groups et cetera. with such an activity going on a large scale by 1.49 billion monthly active users [5], it is inevitable to safeguard each aspect of this networking site. Spam is the new harmful trend taking place on this platform. Facebook has a colossal value of around 170 million fake users [6]. Facebook took a number of security measures to combat these problems. They removed the "Likes" from the users which were inactive from a particular date. Such accounts were deleted by Facebook.

3. LITERATURE REVIEW

A.H. Wang [7] uses Twitter to build their own three graph-based and content based features from 20 most recent tweets for spam bots detection. He observed that if an account posts duplicate messages on one account, it could be termed as a spam account. For this, he used a classifier called Bayesian classifier, since it is noise robust and has a better performance based on user's specific pattern. He used Twitter's API methods and developed his own web crawler for the collection of data sets for his experiments. The result of A.H. Wang research paper showed that there is approximately 1% spam account in the datasets collected by him and approximately 3% spam on Twitter.

Maarten Bosma et al. [8] basically used HITS link analysis algorithm for their research. They also used spam reports for the purpose of spam detection. They used three unsupervised models,

namely, reporter-model, author-reporter model and similarity-author-reporter model. In the reporter model, Maarten Bosma et al. created a bipartite graph where links could be seen between two nodes, namely, reporter and messages. They evaluated spam score and hub score. In the second model, i.e., author-reporter model, they further extended the previous model and introduced a new node called author. They calculated a new score called author score which implied that an author i.e., the author of spam messages could post more number of messages which could be put into the category of spam in comparison to a user who does not posts any spam messages. In the last model which they studied for their paper, they introduced links in between messages so that similar contents could be detected because authors of these spam messages tend to post many duplicate messages online. In the end of their research paper, Maarten Bosma et al. compared these models with one another. Their conclusion was that similarity-author-reporter model was the best performing model because it could detect spam in almost all the scenarios and took over the drawbacks from their previous models.

Ritesh Kumar et al. [9] took the approach of machine learning for the detection of spam from the metadata and logs of social sites. Further, Ritesh Kumar et al. applied the concept of data mining. They observed any new activity of posting on these websites. For every message that is posted, their system makes predictions based on the algorithm. If this does not prove to be successful, they've used manual detection process. Their model can be applied to Twitter, Facebook and YouTube so that they could reach at a correct estimation of spam over these social sites.

Gianluca Stringhini [10] analyzed the method that the spammers use to target the social sites. For this, they created approximately 900 honey-profiles and witnessed a huge deal of spam. They even used various different techniques by which spam profiles can be recognized without creating honey-profiles.

Enhua Tan et al [11] has used a defence based protection scheme SD2 (Sybil Defense based Spam Detection) and has designed an unsupervised method called UNKI Unsupervised social network spam Detection) for the same. These two are the new approaches proposed by Enhua Tan et al in their paper. The UNKI approach was implemented when SD2 was not able to handle more number of attacks. This new approach used social and user-link graphs. They concluded that their UNIK approach could detect spammers with a false positive rate of 0.6% and a false negative rate of 3.7%.

4. SPAM AND ITS CATEGORIES

Spam is the use of electronic medium to send unrequested bulk and duplicate messages, malicious links to fake phishing websites in an attempt for the spammers to earn money or to access any personal user data. This is the real curse of the internet today. Spam can be broadly divided into the following categories:

4.1. SOCIAL NETWORK SPAM

With the increase in social networking sites, more people are using them as a means to communicate and connect with others. With such a hype in the number of users and hence in the amount of data shared, it is almost inevitable to stop spammers to come luring on these social websites. This huge increase in the growth of spammers has forced these large scale networking websites to deploy some spam filtering techniques. Bayesian classification algorithm is also applied to differentiate between the suspicious behaviours from normal users. It has proved to be one of the best algorithms in finding spam.

4.2. EMAIL SPAM

Email spam is the oldest, or you can say is the "classical" technique of spam available on the web. The proportion of spam in email in the first quarter of 2015 was 56.17% [12]. This means that this amount of spam consumes resources on a large level, time spent reading unwanted messages increases, bandwidth and disk storage et cetera are also being wasted. Every mail sent on the internet passes through SMTP (Simple Mail Transfer Protocol), but this is not a secure route. Any email can be made to change its way. This deficiency of SMTP is exploited by spammers for doing fraud.

4.3. IMAGE SPAM

Spammers are now using a different technique in which they use images containing human readable material. These images could be inserted in the mails or posted anywhere on social media. The detection of image spam is a difficult task

4.4. CLICK SPAM

Click spam is the most commonly used spam technique where multiple number of fraud clicks are generated with the intention of directing the users to different locations. Click spams are generally used for promotion of their websites or products and this has proved to be a successful method by spammers in directing users to their websites.

4.5. LINK SPAM

Comment spam or blog spam can be known as link spam. This targets social networking websites, blogs, discussion forums. In this, the spammer uses URLs on the comment section of any trending discussion with a hope of luring as many users as possible. Such instances could be seen on pages of popular groups and discussion groups on social networking websites where any trending discussion topic is taking place.

5. DATASET COLLECTION

Data collection is very important in such an experiment. For performing the task of finding the spammers on the social websites, a large amount of data is required so that correct analysis and inference could be reached upon. The information required for performing our analysis is publically available on both of the social networking sites that we analysed i.e. Twitter and Facebook.

From Twitter, the data was accessed using its API's. These API's allow the users to access mostly all of the data which the user asks for. For this purpose, Twitter provides Consumer Key (API Key), Consumer Secret (API Secret), Access Token and Access Token Secret. These could be used for data accessing.

From Facebook again the data was collected using its Graph API (version 2.0). The approach for Facebook for data collection in our work was highly indirect due to the fact that Facebook deprecated its Graph API in an attempt to strengthen the security, integrity and privacy of its users.

6. DATA FROM TWITTER

6.1. DATA USING API AND IMPLEMENTATION OF ALGORITHMS

Vast amount of data in the form of profiles, tweets, photos, comments et cetera is available on Twitter. There are 307 million active users on twitter as of in the third quarter of 2015 [13]. Some Twitter data is available for the public as well. Anyone with a good intention can have a look at this data for academic purposes. Twitter provides API's (Application Program Interface) to assist developers in data collection. The information required can be directly saved in a database using these API's. The data is stored in csv file format. It is much easier than on Facebook.

We have used "Tweepy" , which is Twitter in Python for the layman, to make calls to the API. Python version 2.7.10 was installed on our machine to facilitate its efficient working.

For accessing Tweepy, the user has to first get his hands on Consumer Key (API Key), Consumer Secret (API Secret), Access Token and Access Token Secret. These keys are then used by in the python code which we've written for accessing Twitter data. In the code we can search for any particular term for example any trending word which can be thought of as being a spam terminology. For example, we can use the term "bit.ly" which is the shortening of URL being used these days. This term is being used by us because bit.ly is generally being seen in URL links which direct the user to some other websites which he/she is not interested in. One other term which is also frequently seen in duplicate messages on posts, comments et cetera is ".com". We've also used this in our code for identifying any possible spammer activity. This is an example of how spam is being used on social websites. We observed during our study of the social networking website that spammers on this site have particular characteristics based on which we picked random accounts containing these specific characteristics. These accounts were randomly picked from the Twitter. Twitter limits the API calls and once a particular limit is reached, Twitter automatically disables the user's connection.

In the implementation part there is a requirement of a data mining and machine learning tool such as Weka to analyze the data. The streamed data is collected and stored as a csv file and then converted into arff format. This data would be fed into Weka for further mining and classification process. The data is changed to arff format just to ensure that the input to the software is in numeric. Now with the data loaded into Weka, Naïve Bayes Classifier is being implemented to distinguish between a spam and a non-spam by evaluating the number of occurrences of each particular probable spam word in the tweets of different accounts. After Naïve Bayes Classifier, Simple K Means clustering algorithm is applied on the test cases which shows the percentage of spam and non-spam in our dataset.

For instance – let say an anonymous account whose tweets have mostly contained a word 'share' would be our probable spam. The main task is to count the number of times it has occurred in the tweets which gives the percentage count of the attributes.

Examining the classifier model in Weka requires the concept of having two classes – Class 0 to indicate a non-spam and Class1 to indicate a spam. Compute the percentage once the percentage count is done. To do this, count the instances that are produced within the Classifier output screen under the Classify tab. The general format of the Weka count output is in Table 1.
The Table 1 means that 4 instances (i.e. tweets of an account) contain that particular attribute value (e.g. the word "share") in Class 1 (Spam). 2 instances didn't have that attribute in Class 1. 3 instances of Class 0 contained that word, while1 instance of Class 0 (e.g. Not Spam) didn't

contain that attribute value. The totals will depict the number of instances that belong to both classes e.g. the number of accounts that are Spam and Not Spam.

Table 1. Classes Assignment

	Class	
	0	***1***
0	1	2
1	3	4
total	4	6

6.2. SVM CLASSIFIER

Supervised Vector Machines are learning models with a learning algorithm that analyses our data which is being used for classification. We will have our training attributes, each marked for belonging to one of two categories. The SVM training algorithm builds a model that assigns new attributes into one category or the other. In our case, it will be “spam” or “non-spam”. An SVM model will map the attributes in separate categories which will be divided by a clear wide gap. New attributes are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

In Weka, SVM will be implemented with the help of SMO (Sequential minimal optimization), under classifiers-functions option. We are using SVM after executing with Naïve Bayes as SVM has proved to be more efficient.

How the SVM Algorithm fares in comparison to Naïve Bayes is illustrated in Figures 1 and 2.

True Positive (TP): It is defined as the number of instances predicted positive that are actually positive.
False Positive (FP): It is defined as the number of instances predicted positive that are actually negative.
True Negative (TN): It is defined as the number of instances predicted negative that are actually negative
False Negative (FN): It is defined as the number of instances predicted negative that are actually negative.
The confusion matrix of both the classifiers was

Confusion Matrix of Naïve Bayes	**Confusion Matrix of SMO**
a b <-- classified as 2 7 | a = 0 8 7 | b = 1	a b <-- classified as 2 7 | a = 0 6 9 | b = 1

Figure 1. Confusion Matrices (Twitter Dataset)

After taking the values of TP, FP, TN, FN from the above matrix, Accuracy, Recall and Precision can be calculated as

Accuracy is calculated by $\frac{TP+TN}{TP+TN+FP+FN}$

Recall is calculated by $\frac{TP}{TP+FN}$

Precision is calculated by $\frac{TP}{TP+FP}$

From the above calculated data one can see that SMO is more efficient as compared to Naïve Bayes in these parameters.

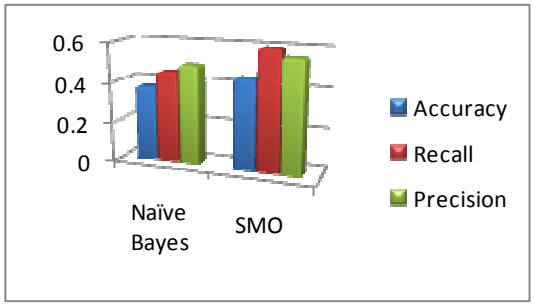

Figure 2. Comparison between the Values for both the classifiers (Twitter Dataset)

7. DATA FROM FACEBOOK

7.1. LIMITATIONS OF FACEBOOK DATA COLLECTION

Facebook is the other social giant which we've used for our project. For extracting data from Facebook, we were approaching for Facebook's API called as Graph API. It is considered as one of the most efficient tools for extracting data from Facebook. This is so because everything on Facebook is represented as a social graph. The nodes of the graph being the users, the things they upload which include pictures, comments, shares et cetera. The links between these nodes being the "friends" relationship between users, likes, reactions et cetera. Facebook Graph API is a structured API for fetching data and the connections between users from Facebook's social graph. But on the contrary, Automated Data Collection from Facebook without the company's express written permission is illegal. Crawling the website is also a criminal offence. The other API's, except the Graph API, can't be accessed without written permissions which are mostly granted only to business, marketing and analytic firms. Also, Graph API can be used to gather only the

publicly available data. No personal information of the users can be gathered using the particular API. In the versions earlier than the current version which is version 2.5 of the API, Facebook allowed for publicly available user data to be accessed using the particular user name or the unique user id that Facebook assigns to each node in its social graph. This capability has been deprecated in the current version of the API due to security reasons. The API does not allow for gathering personally identifiable information either.

Rate limiting for the calls made to the API is another factor that restricts the amount of data that can be collected from the social giant. Unlike in Twitter, rate limiting on Facebook isn't just done on a per user basis. It is calculated by taking the number of users our app had previous day and adding today's new logins which gives the base number of users for our app. Each app is allocated 200 API calls per user in any 60 minute window. For instance, our app had 10 users yesterday and 5 new logins today, that would give us a base of 15 users. This means that our app can make ((10+5)*200) = 300 API calls in any 60 minute window.

7.2. DATA USING APIS AND IMPLEMENTATION OF ALGORITHMS

Facebook doesn't gives any third party the admin authorization of a user's account. Moreover, with the deprecation brought in the data that can be fetched using APIs, getting a user's data through his or her "user id" or "username" isn't possible. Although, the public data of the pages can be collected by making calls to the API. Hence, instead of directly trying to access the unauthorized data, we collected the comments and posts of users from public pages. These pages we selected by randomly searching spam words on Facebook using its "Search" tab. For instance, "rewards" is a spam word. All the pages that resulted by searching were scanned by us. Using the calls to the API, we collected the comments made from user accounts on these pages. All the comments were further cleaned by us to zero down on only those which had the highest content of spam words. After identifying these users, we went to each user's profile on Facebook. All those users whose profiles were public, we found the pages they had liked and further collected their comments from these pages. When sufficient amounts of each account's comments were collected, we found the probability of spam words in those comments and made our dataset. Naïve Bayes and SVM algorithms were then applied on this dataset to detect the amount of spam in our data set.

8. RESULTS

Naïve Bayes and Support Vector Machines were used to train the classifier. Simple k means segregated the data set into spam and non-spam categories. Naïve Bayes and Support Vector Machines were compared on the basis of precision, recall and accuracy. Support Vector Machines showed better results with respect to all the above three parameters, both qualitatively and quantitatively.

8.1. TWITTER RESULTS

Table 2. Naïve Bayes

Parameters	Values
Accuracy	0.375
Precision	0.5
Recall	0.46

Table 3. SMO

Parameters	Values
Accuracy	0.45
Precision	0.5625
Recall	0.6

8.2. FACEBOOK RESULTS

Table 4. Naïve Bayes

Parameters	Values
Accuracy	0.3
Precision	0.6
Recall	0.2

Table 5. SMO

Parameters	Values
Accuracy	0.8
Precision	0.6
Recall	1

9. CONCLUSION AND FUTURE SCOPE

In this paper we presented different ways for detecting spam in social networking websites. Our first approach was Tweepy for extracting data from Twitter and second was Graph API for Facebook. Naïve Bayes mining algorithm was applied to both the datasets for segregation of spam from non-spam. SMO was also applied to present its performance better than Naïve Bayes.

This approach gave us the final result as the percentage of spam and non-spam in our data. For segregating particular accounts and labelling them as spam and non-spam, admin authorisation from the websites will be required. On having access to admin roles from written express permission of each website, our approach can be used to develop a tool which can carve out the spam and non-spam accounts from a particular dataset.

This tool can be deployed as an inherent part of the website or as another admin tool. It will be of great advantage to both the websites as spam is a menace that needs to be tackled with utmost efficiency.

ACKNOWLEDGEMENTS

We are indebted to Prof. (Dr.) Vanita Jain, Head of Department, Bharati Vidyapeeth College of Engineering and our mentor, Mrs Sarita Yadav for the for helpful guidance. We gratefully acknowledge our friends for continuous support and discussions.

REFERENCES

[1] http://www.adweek.com/socialtimes/nexgate-spam-study/428835
[2] http://www.symantec.com/content/en/us/enterprise/other_resources/b-istr_main_report_v19_21291018.en-us.pdf
[3] http://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

[4] http://www.fastcompany.com/3044485/almost-10-of-twitter-is-spam
[5] http://www.statista.com/statistics/264810/number-of-monthly-active-Facebook-users-worldwide/
[6] http://www.huffingtonpost.com/james-parsons/Facebooks-war-continues-against-fake-profiles-and-bots_b_6914282.html?ir=India&adsSiteOverride=in
[7] Alex Hai Wang, "Detecting Spam Bots in Online Social Networking Sites: A Machine Learning Approach", Proceedings of the 24th annual IFIP WG 11.3, Berlin, Germany 2010, pp. 335-342
[8] Maarten Bosma, Edgar Meij and Wouter Weerkamp, "A Framework for Unsupervised Spam Detection in Social Networking Sites", Proceedings of the 34th European Conference on Information Retrieval, Berlin, Germany, 2012, pp. 602-608
[9] Ritesh Kumar, Shital Ghadge, G.S. Navale, "Spam Detection using Approach of Data Mining for Social Networking Sites", International Journal Of Computer Applications, 2014.
[10] Gianluca Stringhini, Christopher Kruegel, Giovanni Vigna, "Detecting Spammers on Social Networks", Proceedings of the 26th Annual Computer Security Applications Conference, New York, USA, 2010, pp. 1-9
[11] Enhua Tan, Lei Guo, Songqing Chen, Xiaodong Zhang and Yihong(Eric) Zhao, "UNIK: Unsupervised Social Network Spam Detection", Proceedings of The 22nd ACM International Conference On Information and Knowledge Management (CIKM 2013), San Francisco, CA, USA, October 27-November 1, 2013
[12] https://securelist.com/analysis/quarterly-spam-reports/69932/spam-and-phishing-in-the-first-quarter-of-2015/
[13] http://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/