Volume 5, Issue 3, March 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Detection of Spam in Social Networks using Clustered k-Nearest Neighbour

Jyotika Verma, Dr. Sanjeev Dhawan
Department of Computer Science & Engineering,
UIET, Kurukshetra University, Kurukshetra, Haryana, India

Abstract— Social Networking sites are one of the services which are provided by the Internet.. Social Networking Sites (SNSs) example MySpace, Facebook, CyWorld, and Bebo are very much used. As it includes large number of users and it is a hub of information, this has become a potential channel for attackers to exploit or attack. Various techniques has been used for spam detection but it was experienced that there was a problem related to the accuracy and if one gives the accurate result then that technique was time consuming.. In this paper, we have came up with the idea of introducing a technique which involves the use of clustered KNN approach with the help of which the spam detection system will become time efficient and will give accurate result as well.

Keywords— Spam, Social Spam, k-Nearest Neighbour, Stop Word Removal, Term Frequency-Inverted Document Frequency.

I. INTRODUCTION

The Internet, its name defined it all there is no need to give the introduction of this term or this can be said that in today's time there is no need to elaborate about this term. Internet is defined as group of interconnected networks that uses the standard Internet Protocol Suite (TCP/IP) to link several billion devices worldwide [1].The most commonly used services of the Internet are Email and Social Networking Sites. Social Networking Sites (SNSs) example MySpace, Facebook, Twitter, CyWorld, and Bebo are very much used. These websites have attracted million of users, and most of them have taken these websites as the integral part of their daily life. Social Networking Sites example facebook, twitter etc have gained so much popularity as it becomes the daily routine of almost every person to check their profile at least once in a day. As it includes a large number of users and also consists of large amount of information, this has become very popular for attackers to exploit or attack. Everything in this world consists of some benefits and some drawbacks. Internet also consist some benefits as well as drawbacks. Single information can be easily broadcasted but sometimes situation arises when that information is not at all useful to one person, then it is said to be as "spam". There are many issues which are related to the Internet like Access, Children and the Internet, Human Rights, Standardization etc but Spam is one of the drawbacks of Internet which is faced by every person on the web [4].

A. Spam

According to the Oxford Dictionary the definition of the spam is given as "Irrelevant or unsolicited messages sent over the Internet, typically to large number of users, for the purpose of advertising, phishing, spreading malware, etc. [5]. Spamming is termed as the use of electronic messaging systems in order to send the unrequested messages (spam) especially advertising. And it is also termed as the process of sending the messages to the same site continuously. Here, the most importantly recognized spam is E-mail spam , well the spam also have been seen in many different areas like instant messaging spam[6], Usenet Newsgroup spam[7], Web Search Engine Spam[8] also known as the Spamdexing, and many more. A person who creates the electronic spam is referred to as spammer. Spam is the problem which is not only associated with the email but with the social networking sites too. Today the spammers are very difficult to identify and even more sophisticated. Facebook is used by everyone, sometimes the click on the "Like" button results very risky. Facebook and Twitter are two social networking websites with which everyone is familiar. This is the reason spammers mostly attack on these websites. Spammers send the false link messages to the user in the name of his/ her trusted contacts such as friends and family. "*As for Twitter, spammers usually gain credibility by following the verified accounts; when that account owner follows the spammer back, it legitimizes the spammer and allows him/her to proliferate"*[9]. Pinterest is the new site which is now very much used. And even this site is not free from the spam. Every site is facing the spamming problem in different manner. Spam is the problem which is faced by everyone who is using the web. Whether it is related to the email or with the social networking websites, spam can be easily found everywhere.

B. Email Spam

Email spam refers to the process of sending the unsolicited and unrequested email to a particular account. Sometimes the conditions occur when in the spam email links are provided, and if a click is performed on that link then it automatically get forwarded to some dirty websites or some phishing websites or the sites which are hosting the malware. Spam

includes very much of the malicious things for example it includes the phishing websites or it may include the executable file attachments. There is an existence of the opposite of the spam which is called as "ham" which everyone wants which means the legitimate mail in someone's account.

C. Social Spam

This is the problem which is directed specifically to the users of the social networking site such as Google+, Facebook, Pinterest, LinkedIn, or MySpace. This has been said that out of all the accounts around 40% of the accounts are used for the spam. The spammers mostly use the common fan pages to send notes from the fraudulent accounts. And these notes may contain the very disturbing content like the pornographic videos or any other sites which wants to sell their products. Many social networking sites have added the "report abuse" button in response to this. But spammers are very intelligent, they frequently change their address and due to this it becomes difficult to track them. There are many techniques which are used for the spam filtering but broadly it can be classified into two categories i.e.

(a) Content Based Filtering – is the method which includes The Header, Subject, Body. The header part of any email includes the following subsection:-

- The Sender email id or address.
- The Receiver email id or address.
- The Subject.

The Body is that which includes the complete content which a sender wants to convey to the receiver including the complete text, the images etc. In this Technique one thing is detected whether the received mail is spam or ham depending on the body part of the mail, that what type of content it contains. The header section is completely ignored in this technique. For performing this various filtering techniques are used like Fisher- Robinson, Inverse chi square function technique, KNN classifier, AdaBoost Classifier etc.

(b) Metadata based – it includes various list on basis of which the text is categorized as spam or ham. The various list which comes under this category are black list, white list, gray list etc. This categorization is shown in Table I [12].

TABLE: I TECHNIQUES USED FOR THE SPAM FILTERING [12]

Content Based	Metadata Based
SVM	Black List
KNN	White List
AIS	Forging Based
GA	Gray List
NAÏVE BAYES	

So, it can be said that spam disturbs and interrupts people in doing their work. Sometimes it adds irrelevant things to process while one try to get the actual information. It usually piles up the work and thus sometimes reduces the efficiency. So, to overcome these issues, many technologies and methodologies have been introduced to reduce or remove spam. Various techniques have been used for the spam filtering; one of them is KNN, which is widely used as it provides most accurate results.

II. RELATED WORK

There is availability of plenty of research in literature related to the spam in social networking sites. One can easily find out the various techniques for the detection of the spam in the social networking sites. Here we are going to review some of the studies or researches which have been done in the field of detection of the spam. Classifiers like NB, AdaBoost, and KNN are used for the detection of the spam. Out of them NB is the simple and easy to use but if we see the KNN then we will find out that KNN gives us the result with the high accuracy but it takes a lot of time to process the text. So, we can say that KNN is very time consuming process. Firstly, the spam detection process is itself very time consuming and on this if one uses the KNN which is also very time consuming then the work takes a lot longer to complete. So, to overcome this issue KNN was later on used with resampling and it was done by Loredana *et al.* [10]. There are some incremental and non-incremental strategies available for the detection of the spam. Before, doing any work on the incremental strategy, the batch learning was used for the detection of the spam. But later on the incremental algorithm for the detection of spam was introduced [11]. And this method performed better than any of the other incremental or non-incremental algorithms. KNN algorithm is very much useful in detection of the spam text but it takes a lot of time in processing this is the only disadvantage with this algorithm otherwise it always provide the high accuracy in results. There is also one technique which can be used for the detection of the spam is the Genetic Algorithm. There are various benefits of the Genetic Algorithm it doesn't require any of the well known rules for performing its work but infact it does all of its work on it internal rules. But in this aspect the efficiency of the GA was not that good. So, to improve its efficiency a new model was proposed and that named was Enhanced Genetic Algorithm (EGA) and this was proposed by Saber Salehi *et al.* [12]. Another concept of ANN (Artificial Neural Network) comes in light and this time the study was performed using another new algorithm which was named as Mimetic Algorithm [13]. They used the UCI Spambase dataset and the concept of simulated annealing was also used and all these things have been used so that the parameters of the ANN gets optimized and to improve the spam detection the hybrid approach algorithm was used. Naïve Bayes is

referred to as one of the popular techniques which has been used for the detection of the spam but the time which is taken by this method is lot longer. A technique was then introduced to improve its time taken for the work of processing. And that was applied on the Paul Graham's Bayesian spam detection method [14]. His main aim was to reduce the time which was taken for performing the computations and it was done using the radix encoded fragmented database approach. Datasets which have been used were LingSpam and SpamAssasin. The computation time was very much improved after using this technique. For the classification of the spam another technique was also proposed and it was done by Nadir Omer Fadl Elssied *et al.* [15]. They have proposed a concept which makes the use of the K-means clustering and the Support Vector Machines for the spam classification and have used them in the hybrid way. The model which they have been proposed includes the improved detection rate and the time cost have also been reduced and the false positive also. They have provided the measurements also. Comparisons have been performed between the proposed system and the spam detection system which is based on the Support Vector Machine. Later on, Kunal Jain *et al.* [16] have then proposed the technique for the classification of the spam which was based on the combination of the local concentration based feature extraction method with k-means clustering. They have performed the experiments on the PU Corpora which was in a controlled environment and have concluded that there proposed system was better than the existing one. They have also provided the future scope that the performed work can be made better by applying the concept of the multiclass classification.

III. PROBLEM FORMULATED

All the research work which has been studied on the basis of that most important concept which is used for the spam detection is the classification. There are various classification techniques are available which can be applied on the incoming text so to properly separate the spam and ham. But when the different classification techniques are applied their respective advantages and the disadvantages came into light. After performing the study related to the spam detection, a conclusion is made that even after using any of the technique for the detection of the spam the result was not that of the considerable accuracy and if it was achieved then the processes which have been used were time consuming. So, there was no such technique which has given the great accuracy and as well as was time efficient. The problems which can be formulated from the study can be named as Accuracy and Complexity. For performing the spam detection work there always exists a requirement for the dataset on which the work is to be performed or you can take the live data also. Various datasets have been used like LingSpam, SpamAssasin, and PU Corpora etc. There was a study performed for the spam detection which was based on the dataset PU Corpora. They have used the Tokenization and Dictionary Generation for the preprocessing of the Dataset. The work of the tokenization includes the processes like encryption and decryption which is itself a very time consuming process. Feature Selection strategies are used in designing the spam filters. Under this it is said that every email or text is represented by a Vector Space Model (VSM) means in this every email or any text is considered as vector of word terms. There are various Feature Selection Strategies available which we can be used to perform the spam detection. After that the K-means Clustering technique is applied for the classification of the text whether it is a spam or ham. But again there is also an advantage and one disadvantage associated with the K-means Clustering Technique and that is, although it gives the accurate result but it is very time consuming.

IV. PROPOSED METHODOLOGY

This has been observed that the technique which has been used very much for the filtration process of the spam is Content Based Filtering Technique. It works with the header and the body part of the email. The model which has been proposed for the detection of the spam by using the K-means clustering give the result which is accurate but the method is very time consuming. To overcome this problem some steps can be performed which can be divided in the four strategies and all the strategies are independently working. The process should be performed step by step such that the output of one step is fed as the input to another step. For performing any work related to the spam detection, a dataset or any real time data is required. Datasets like LingSpam, SpamAssasin, PU Corpora etc can be used. So, the first step is there should be availability of the dataset or one can take the real time data. Then the preprocessing of the data is performed, we will perform this work with the help of the Stop Word Removal. Before commencing the classification of the text, the preprocessing should be performed. There are two ways to perform it. If we are working in the real time environment then the incoming mails or the text are to be processed but ever if we are working for the experiment purpose then we have to make use of the datasets. Then those datasets are to be processed. We will use the Stop Word Removal in this phase for performing this preprocessing. Stop Word Removal is the process of removing the common words so to decrease the complexity of the process with the help of this the system will become time efficient. After performing this, the data will be transmitted to the next stage for further refinement. After that, the term selection is to be performed and this is done by identifying the "number of bits" gained by knowing the term is present or absent. Its output is then fed as the input to the next stage and on the next stage the feature extraction work is to be performed. Feature Selection strategies are used in designing the spam filters. Under this it is said that text is represented by a Vector Space Model (VSM) means in the text is considered as vector of word terms. There are various Feature Selection Strategies available which we can be used to perform the detection of the spam. There are various feature selection algorithms, TF-IDF (Term Frequency- Inverted Document Frequency) is one from them, this technique should be used to perform the feature extraction, this will help to make the system time efficient. After that the classification is to be performed. It is referred to as the fourth stage of the model and it is also the most important stage as it will give us the result that whether the incoming email or the text is spam or ham (Legitimate email). The technique which is to be used for performing the classification is the Clustered KNN. If we consider the whole process in the name of steps, these can be written as:

Step 1: The incoming email/text or if we have used the dataset should be taken as the input.
Step 2: Preprocessing of the Dataset should be performed with Stop Word Removal.
Step 3: Term should get selected with the Information Gain.
Step 4: Then, the Feature Selection should be performed using the TF-IDF (Term Frequency-Inverted Document Frequency)
Step 5: Clustered KNN (K- Nearest Neighbor) classification technique should be used.
Step 6: Result whether the email/text is spam or ham.

V. CONCLUSION

Based on the research done in context of spam detection we can conclude that most important concept which is used for the spam detection is the classification .When the different classification techniques are applied there respective advantages and the disadvantages came into light. It has been observed that the techniques which have been used for the detection of the spam gives the accurate results but the processes which have been used for the spam detection were time consuming. So, it can be concluded that the techniques which have been used for the spam detection are time consuming. The objective of this paper is to come up with a system which is time efficient as well as gives the accurate result.

ACKNOWLEDGMENT

With the overwhelming sense of respect, I would like to convey my heart-felt gratitude to Dr. Sanjeev Dhawan (Asst. Prof.), for continuous encouragement and support they provided me during the tenure of my work at UIET, and giving me the opportunity to carry out my project/ report training in this organization.

REFERENCES

[1] Wikipedia, "Internet",2014, http://en.wikipedia.org/wiki/Internet, Accessed on 29-January-2015.
[2] The Free Dictionary, "email definition", 2014, www.thefreedictionary.com/e-mail, Accessed on 29-January-2015.
[4] Internet Issues – Ethical & Technology Issues, 2015 http://www.internetsociety.org/what-we-do/internet-issues Accessed on 2-February-2015.
[5] Oxford Dictionary, "Definition of spam", http://www.oxforddictionaries.com/definition/english/spam Accessed on 29-January-2015
[6] Wikipedia, "Messaging spam" http://en.wikipedia.org/wiki/Messaging_spam. Accessed on 29-January-2015
[7] Wikipedia,"Newsgroup_spam"2014http://en.wikipedia.org /wiki/ Newsgroup_spam. Accessed on 29-January-2015.
[8] Wikipedia,"Spamdexing", 2014, http://en.wikipedia.org/wiki/Spamdexing. Accessed on 29-January-2015.
[9] Wikipedia, "Spamming," 2015, http://en.wikipedia.org/wiki/Spamming Accessed on 3-February-2015.
[10] Loredana Firte, Camelia Lemnaru, Rodica Potolea "Spam Detection Filter using KNN Algorithm and Resampling", IEEE International Conference on Intelligent Computer Communication and Processing, Cluj-Napoca, 26-28 August 2010, pp. 27-33.
[11] Elham Ghanbari, Hamid Beigy "An Incremental Spam Detection Algorithm", IEEE International Symposium on Artificial Intelligence and Signal Processing, Tehran, 15-16 June 2011, pp. 31-36.
[12] Saber Salehi, Ali Selamat, Mohammad Bostanian " Enhanced Genetic Algorithm for Spam Detection in Email", IEEE 2nd International Conference on Software Engineering and Service Science, Beijing, 15-17 July 2011, pp. 594-597.
[13] Shaveen Singh, Anish Chand, Sunil Pranit Lal, "Improving Spam Detection Using Neural Networks Trained by Memetic Algorithm", IEEE 5th International Conference on Computational Intelligence, Modelling and Simulation, Seoul, 24-25 September 2013, pp. 55-60.
[28] Nishtha Jatana, Kapil Sharma, "Bayesian Spam Classification: Time Efficient Radix Encoded Fragmented Database Approach", IEEE International Conference on Computing for Sustainable Global Development, New Delhi, 5-7 March 2014, pp. 939-942.
[15] Nadir Omer Fadl Elsseid, Othman Ibrahim, Waheeb Abu-Ulbeh "An Improved of Spam E-mail Classification Mechanism using K-means Clustering", Journal of Theoretical and Applied Information Technology, Vol. 60 No.3, February 2014.
[16] Kunal Jain, Sanjay Aggarwal "A Hybrid Approach for Spam Filtering using Local concentration based K-means Clustering", IEEE 5th International Conference on Confluence The Next Generation Information Technology Summit(Confluence), Noida, 25-26 September 2014, pp. 194-199.

