iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 4 Issue: VII Month of publication: July 2016

DOI:

www.ijraset.com
IC Value: 13.98

Volume 4 Issue VII, July 2016
ISSN: 2321-9653

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Efficient Spam Detection on Social Network

Girisha Khurana[1], Mr Lalit kumar[2]

Department of computer science GNI Mullana, Kurushetra University

Abstract— With the growth of social networking sites for communicating, sharing, storing and managing significant information, it is attracting cybercriminals who misuse the Web to exploit vulnerabilities for their illicit benefits.spammers are the malicious users who contaminate the information presented by legitimate users and in turn pose a risk to the security and privacy of social networksTwitter is a social network designed as an information sharing service that allows users to exchange messages up to 140 characters. These messages are known as tweets. In our thesis we gather data from Twitter. This data will be used to analyze the features,test and train our data that will be used for supervised classification in order to detect real malicious profiles(spammers and non spammers).Based on dataset and feature collection, testing and training a supervised machine learning model is introduced for spammers identification. We use the KNN based classification(spammer detection) model and compared with other classification technique.we evaluate our data and optimize the system too.

Keywords— spammer,twitter,precision,recall,malicious,social network

I. INTRODUCTION

Social spam is unwanted spam content appearing on social networks and any website with user-generated content (comments, chat, etc.). It can be manifested in many ways, including bulk messages, profanity, insults, hate speech, malicious links, fraudulent reviews, fake friends, and personally identifiable information.Online Social Network (OSN) are websites where users can create profiles, es-tablish connection with other users and converse with them. There are hundreds of such OSN websites present today. Facebook, Twitter, etc. are the most popular ones boasting more than 500 million active users. Twitter, as an Online Social Network, is intended to help people converse using text-based posts called tweets. Popularity of Twitter and other OSNs has been rising in recent times having played crucial role in connecting people and providing a discussion forum on several occasions like protests in Syria.

A. *Types*

1) *Spam:* Commercial spam is a comment that has commercial content irrelevant to the discussion at hand. Many of the old email spam content resurfaced on social networks, from Viagra ads, to work-from-home scams, to counterfeit merchandise. Recent analysis showed social spammers content preferences changing slightly, with apparel and sports accounting for 36% of all posts. Others included: porn and pills (16%), SEO/web development (23%), and mortgage loans (12%).
2) *Social networking spam:* is spam directed specifically at users of internet social networking services such as Google+, Facebook, Pinterest, LinkedIn, or MySpace. Experts estimate that as many as 40% of social network accounts are used for spam. These spammers can utilize the social network's search tools to target certain demographic segments, or use common fan pages or groups to send notes from fraudulent accounts. Such notes may include embedded links to pornographic or other product sites designed to sell something. In response to this, many social networks have included a "report spam/abuse" button or address to contact. Spammers, however, frequently change their address from one throw-away account to another, and are thus hard to track.Facebook pages with pictures and text asking readers to e.g. "show your support" or "vote" are used to gather likes, comments and shares which improve the pages' ranking.The page is then slightly changed and sold for profit.
3) *Bulk:* submissions are a set of comments repeated multiple times with the same or very similar text. These messages, also called as spam-bombs can come in the form of one spammer sending out duplicate messages to a group of people in a short period of time, or many active spam accounts simultaneously posting duplicate messages. Bulk messages can cause certain topics or hashtags to trend highly. For example, in 2009, a large number of spam accounts began simultaneously posting links to a website, causing ‘ajobwithgoogle’ to trend.
4) *Profanity:* User-submitted comments that contain swear words or slurs are classified as profanity. Common techniques to circumvent censorship include “cloaking”, which works by using symbols and numbers in place of letters or inserting punctuation inside the word (for example, "w.o.r.d.s" instead of "words"). The words are still recognizable by the human eye, though are often missed by website monitors due to the misspelling.

5) *Threats:* User-submitted threats of violence are comments that contain mild or strong threats of physical violence against a person or group. In September 2012, Eric Yee was arrested for making threats in an ESPN comment section. He started out discussing the high price of LeBron James shoes, but quickly turned into a stream of racist and insulting comments, and threats against children.[17] This is a more serious example of social spam.

6) *Hate speech:* User-submitted hate speech is a comment that contains strongly offensive content directed against people of a specific race, gender, sexual orientation, etc. According to a Council of Europe survey, across the European Union, 78% of respondents had encountered hate speech online; 40% felt personally attacked or threatened; and 1 in 20 have posted hate speech themselves.

7) *Malicious links:* User-submitted comments can include malicious links that will inappropriately harm, mislead, or otherwise damage a user or computer. These links are most commonly found on video entertainment sites, such as YouTube.What happens when you click on malicious links can range from downloading malware to your device, to directing you to sites designed to steal your personal information, to drawing unaware users into participating in concealed advertising campaigns.Malware can be very dangerous to the user, and can manifest in several forms: virus, worm, spyware, Trojan horse, or adware.

II. WHAT IS TWITTER?

Twitter is a social network designed as an information sharing service that allows users to exchange messages up to 140
characters. These messages are known as tweets.

Twitter allows unidirectional relationships among users. This results in users following (known as followers) other users (known as friends, this is, the users a given user follows).If user A □ B, we define the relationship as A follows B, A will become a follower of B and B will be in the list of friends for A. Followers receive any tweet (message up to 140 characters) sent from their friends.Twitter is designed as a microblogging service showing all this data openly. Anyone can browse any user´s profile and check her tweets and relationships unless the user requested this data to be private. Direct messages are not public. The percentage of private accounts is not disclosed byTwitter, however they are a clear minority.

Following is the standard terminology used in Twitter and relevant to our work:

Tweets : A message on Twitter containing maximum length of 140 characters.

Followers & Followings : Followers are the users who are following a particular user and followings are users whom user follows.

Retweet: A tweet that has been reshared with all followers of a user.

Hashtag : The # symbol is used to tag keywords or topics in a tweet to make it easily identifiable for search purposes

Mention : Tweets can include replies and mentions of other users by preceding their usernames with @ sign.

Lists : Twitter provides a mechanism to list users you follow into groups

Direct Message [17]: Also called a DM, this represents Twitter's direct messaging system for private Communication amongst user.

III. EXISTING RESEARCH

In the past ten years, email spam detection and filtering mechan-isms have been widely implemented. The main work could be summarized into two categories: the content-based model and the identity-based model. In the first model, a series of machine learning approaches[1,2] are implemented for content parsing according to the keywords and patterns that are spam potential. In the identity-based model, the most commonly used approach is that each user maintains a whitelist and a blacklist of email addresses that should and should not be blocked by anti-spam mechanism[3,4]. More recent work is to leverage social network into email spam identifica-tion according to the Bayesian probability[5].

With the rapid development of social networks, social spam has attracted a lot of attention from both industry and academia. In industry, Facebook proposes an EdgeRank algorithm[6] that assigns each post with a score generated from a few feature (e.g., number of likes, number of comments, number of reposts, etc.). Therefore, the higher EdgeRank score, the less possibility to be a spammer. The disadvantage of this approach is that spammers could join their networks and continuously like and comment each other in order to achieve a high EdgeRank score.

In academia, Yardi et al[7]. studies the behavior of a small part of spammers in Twitter, and find that the behavior of spammers is different from legitimate users in the field of posting tweets, foll-owers, following friends and so on. Stringhini et al[8]. further inv-estigates spammer feature via creating a number of honey-profiles in three large social network sites (Facebook, Twitter and Myspace) and identifies five common features (followee-to-follower, URL ratio, message similarity, message sent, friend number, etc.) potential for spammer detection. However, although both of two approaches introduce convincible framework for spammer detection, they lack of detailed approaches specification and prototype evaluation.Wang[9] proposes a naïve Bayesian based spammer classifica-tion algorithm to distinguish suspicious behavior from normal ones in Twitter, with the precision result (F-measure value) of 89%.Gao et al[10]. adopts a set of novel feature for effectively reconstructing spam messages into campaigns rather than examining them indivi-dually (with precision value over 80%). The disadvantage of these two approaches is that they are not precise enough. Benevenuto et al[11] collects a large dataset from Twitter and identify 62 feature related to tweet content and user social beh-avior. These characteristics are regarded as attributes in a machine learning process for classifying users as either spammers or non-spammers.

Zhu et al[12] proposes a matrix factorization based spam classification model to collaboratively induce a succinct set of latent feature (over 1000 items) learned through social relation-ship for each user in RenRen site (www.renren.com). However, these two approaches are based on a large amount of selected feature that might consume heavy computing capability and spend much time in model training. Zachary Miller et al[13] proposes due to the increasing popularity and heavy use of social networks like Twitter, the number of spammers is rapidly growing. This has resulted in the development of several spam detection techniques[14-18] . This study has made three new contributions to the field of spam detection on Twitter. First we view spam identification as an anomaly detection problem. Secondly, we introduce 95 one-gram features from tweet text to the task of spam detection on Twitter. Finally, we use the stream of real-time tweets as well as user profile information with two stream-based clustering algorithms, DenStream and StreamKM++. When tested, these two approaches achieved 97.1% accuracy and 84.2% F-Measure and 94.0% accuracy and 74.8% F-Measure respectively. Our findings suggest the addition of one-gram features enhances spam detection. Although these algorithms independently demonstrated good detection, the combination of the two further improved all our metrics particularly recall and false positive rate to 100% and 2.2%, showing the value of the multi-layer approach to spam detection.

A. Technical development: data gathering

The first step for our analysis is to gather data from Twitter. This data will be used to analyze the features that will be used for supervised classification in order to detect real malicious profiles. We should find the way to get both clean and malicious data, and a method to tag it properly in our system. we get the twitter data through R package

B. Preprocessing and feature extraction

This section describes all the work related to the analysis of the data gathered from the previous section, including the analysis of features

We collect the online data of twitter through R package,twitter api.The data we get is not processed.firstly we have to proceesed the data and analyze its features

1) *Preprocessing:* In the preprocessing ,we processed 150 records and we manually labeled them spammer or non spammer data.We analyze 53 out of 150 are spam.

During preprocessing we remove lowercasetext,stopwords,special symbol and perform stamming using porter algorithm

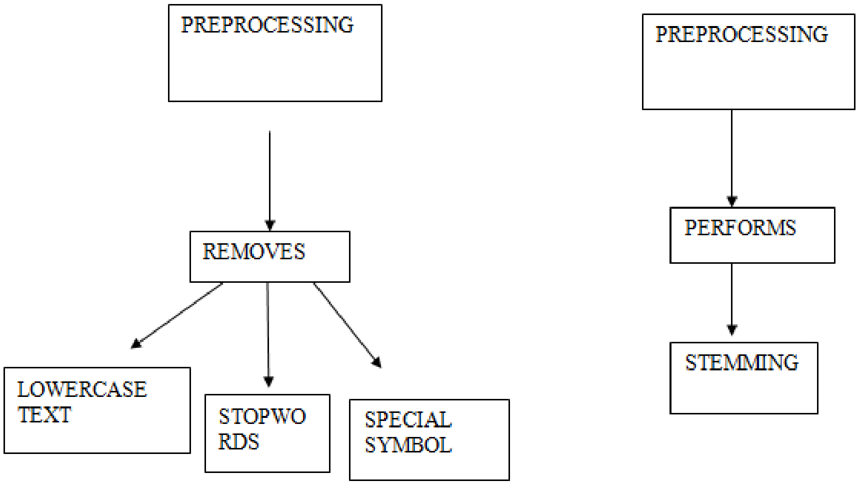


Fig. 1 Example of Pre-processing

C. Feature Extraction

In our thesis we basically consider content based features to find spammer and non spammer.we use 6 content based features like numSpamWords, total words, num url per words, num urls, Retweet count, num hash tagper words.

1) *No. of hashtages(#) per word*- spammers tweet multiple unrelated updates to the most mentioned topics on Twitter using # to lure legitimate users to read their tweets.
2) *No. of URLs per word*- spammer's tweets consist of large number of URLs per word of malicious sites.
3) *Retweets count*- Retweets are the replies to any tweet using @RT symbol and spammers use maximum @RT in their tweets.
4) *Spam Words*- Spammer's tweets mainly consist of spam words.
5) *No of URL*-- spammer's tweets consist of large number of URLs of malicious sites.
6) *Total words*-It will consider all the words in the message

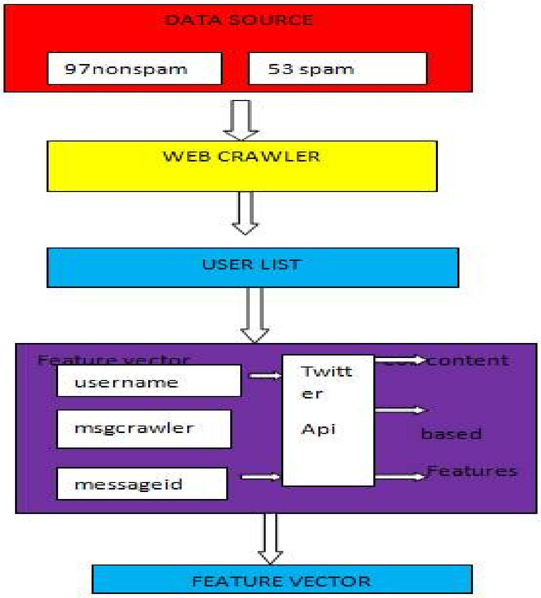


Fig 2.(Dataset and feature collection procedure)

D. Testing and Training of data

After data gathering ,preprocessing and feature selection we will perform testing and training of data .Test set and validation set are used for evaluating whether the discovered relationships hold.

E. Spammer detection

Based on dataset and feature collection ,testing and training described in the previous section, a supervised machine learning model is introduced for spammers identification. Supervised learning is the machine learning task of inferring a function from labeled training data that consists of a set of training examples. Inside supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called supervisory signal). Through analysis of the training data, supervised learning solution produces a classification model for predicting new examples.

F. KNN Based Spammer Detection Model

Instance-based classifiers such as the kNN classifier operate on the premises that classification of unknown instances can be done by relating the unknown to the known according to some distance/similarity function. The intuition is that two instances far apart in the instance space defined by the appropriate distance function are less likely than two closely situated instances to belong to the same class

In this spammer detection model we have consider two matrix Feature matrix and class matrix.In the class matrix we consider
1 for non spam
2 for spam

Then we train the data in feature matrix and class matrix

We input test, train data into Knn classifier then we get predicted class.Then we evaluate the predicted class and test class.

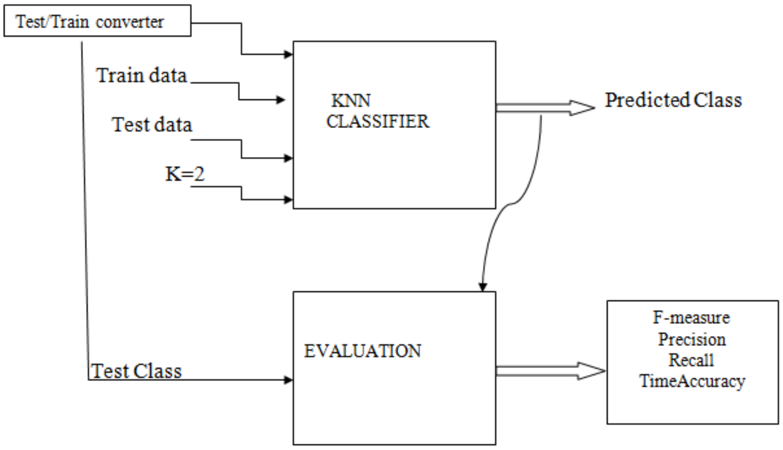


Fig 3(Thesis model)

G. Evaluation metrics

In our thesis we evaluate our classifier in terms of F-measure,Recall,Precision,Accuracy and then we compared with existing approach and calculate their percentage. We get the following graphs:

Algo	Accuracy%	Precision%	Recall%	FMeasure%
Existing	96	98.7	92.9	94.7
Proposed	98	99	97	98

Fig 4 (comparison of existing and proposed approach)

F-measure is the harmonic mean between precision and recall, and is defined as F¼2PR/ (PþR).

Recall (R) is the ratio of the number of instances correctly classified to the total number of predicted instances and is expressed with formula R¼a/(aþb).

Precision (also called positive predictive value) is the fraction of retrieved instances that are relevant.

Accuracy is the proportion of true results (both true positives and true negatives) among the total number of cases examined.

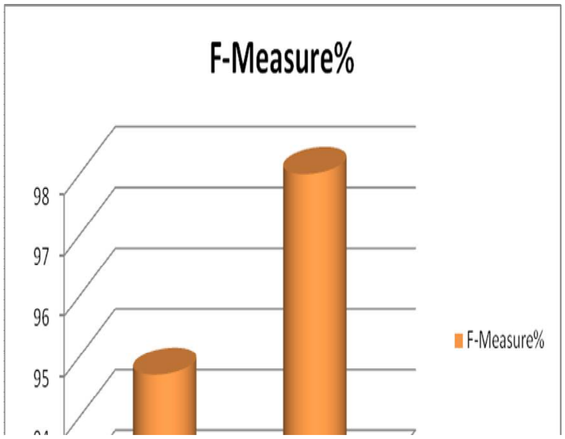


Fig 4(F-measure%)

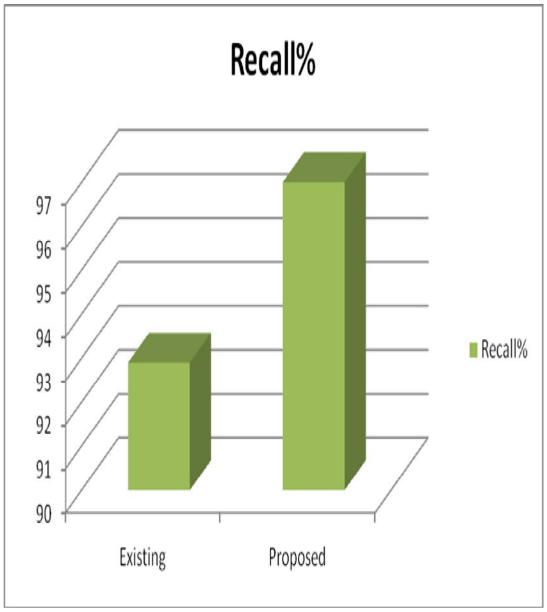


Fig 5(Recall%)

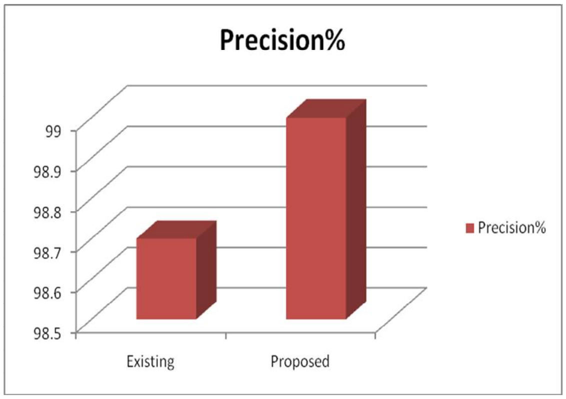


Fig 6(Precision%)

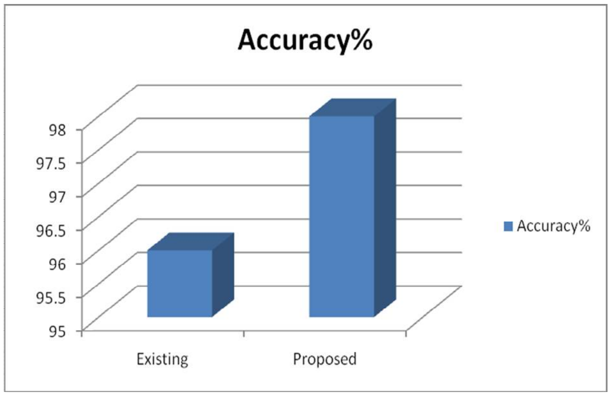


Fig 7(accuracy%)

Common graph in terms of accuracy,precision,recall,F-measure

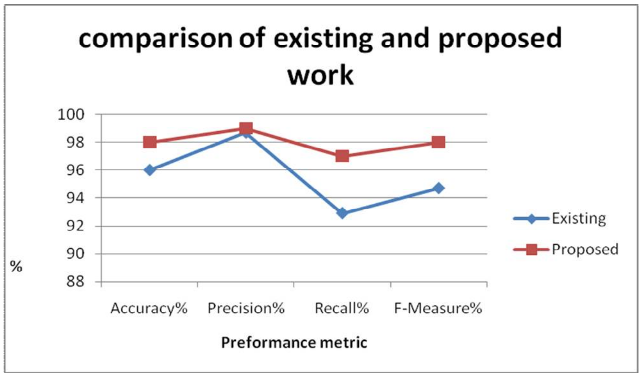


Fig 8(comparison of existing and proposed work)

In this we evaluate the classifier according to the training percentage

Training%	accuracy%	recall%	precision%	f_measure%	Time(sec)
20	86.6667	94.8454	85.9813	90.1961	0.019
40	89.3333	88.6598	94.5055	91.4894	0.019
60	98.6667	97.9381	100	98.9583	0.02
80	98.6667	100	97.9798	98.9796	0.02
100	100	100	100	100	0.022

Fig 9(Comparision of evaluation features acc to training%)

When we train the data 20%,40%,60%,80%,100% then we evaluate accuracy,Precision,f-measure,time(sec) the graphs we get according to training percentage

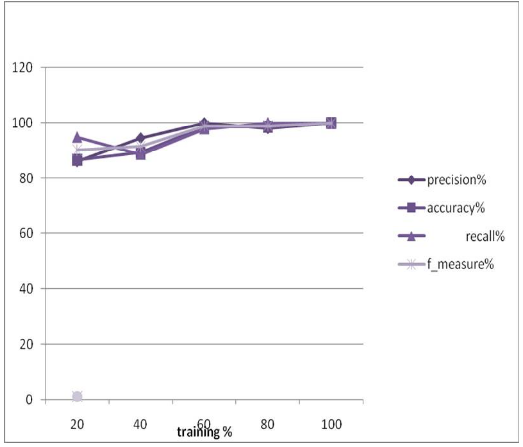


Fig 10(common graph acc to training %)

We can say our timing complexity is still less according to training percentage.

IV. CONCLUSIONS

In thispaper, we have introduced a machine learning based spa-mmer detection model for social network(Twitter). The solution considers the user's content and behavior feature, and apply them into KNN based algorithm for spammer classification. We work on live data.This paper has less time complexity and accuracy is higher than our existing approach.

In future we can improve the classifier by joining some optimization technique and can add some other relevant features.It would be interesting adding some degree of semantic analysis to create new features. Also this would only be possible using a central system so semantic information could be crosschecked with information of malicious ongoing campaigns.

V. ACKNOWLEDGMENT

I, wish to express my sincere thanks and deep sense of gratitude to Mr Harish Saini , Associate professor, Department of Computer Science and Engineering; my esteemed teacher and guide Mr Lalit kumar, assistant Professor, Department of Computer Science and Engineering, GNIT, Mullana (Ambala) for his constant inspiration, scholarly guidance and helpful suggestions throughout the course of this pre-thesis work. I am indebted to his for allowing me to draw upon his precious and 'valuable time. I will always remember his quick analysis, thoughtful solutions and critical reviews he has given throughout the period.

I, also thank other faculty members of Department of Computer Science and 'Engineering, GNIT, Mullana (Ambala) for their intellectual support throughout the course of stay at GNIT, Mullana (Ambala).

Finally, I am thankful to all whosoever has contributed in this Pre-thesis work directly or indirectly in completing my present study.

REFERENCES

[1] M. Uemura, T. Tabata, Design and evaluation of a Bayesian-filter-based image spam filtering method, in: Proceedings of the International Conference on Information Security and Assurance (ISA), IEEE, 2008, pp. 46–51.

[2] B. Zhou, Y. Yao, J. Luo, Cost-sensitive three-way email spam filtering, J. Intell. Inf. Syst. 42 (1) (2013) 19–45.

[3] J. Jung, E. Sit, An empirical study of spam traffic and the use of DNS black Lists, in: Proceedings of the 4th ACM SIGCOMM Conference on Internet Measure-ment, ACM, 2004, pp. 370–375.

[4] M. Antonakakis, R. Perdisci, D. Dagon, W. Lee, N. Feamster, Building a dynamic reputation system for DNS, in: Proceedings of the Third USENIX Workshop on Large-scale Exploits and Emergent Threats (LEET), 2010.

[5] Trust evaluation based content filtering in social interactive data, in: Proceed-ings of the 2013 International Conference on Cloud Computing and Big Data (CloudCom-Asia), IEEE, 2013, pp. 538–542

[6] J. Kincaird, Edgerank: the secret sauce that makes Facebook's news feed tick, TechCrunch, 2010, ⟨http://techcrunch.com/2010/04/22/facebook-edgeran⟩.

[7] S. Yardi, D. Romero, G. Schoenebeck, Detecting spam in a Twitter network, First Monday 15 (1) (2009).

[8] G. Stringhini, C. Kruegel, G. Vigna, Detecting spammers on social networks, in: Proceedings of the 26th Annual Computer Security Applications Conference, ACM, 2010, pp. 1–9.

[9] A.H. Wang, Don't follow me: spam detection in Twitter, Security and Cryptography (SECRYPT), in: Proceedings of the 2010 International Conference on. IEEE, 2010, pp. 1–10.

[10] H. Gao, Y. Chen, K. Lee, D. Palsetia, A. Choudhary, Towards online spam filtering in social networks, in: Proceedings of the Symposium on Network and Distributed System Security (NDSS), 2012.

[11] F. Benevenuto, G. Magno, T. Rodrigues, V. Almeida, Detecting spammers on Twitter, in: Proceedings of the Seventh Annual Collaboration, Electronic

messaging, Anti-abuse and Spam

a. Conference (CEAS), 2010.

[12] Y. Zhu, X. Wang, E. Zhong, N.N. Liu, H. Li, Q. Yang, Discovering spammers in social networks, in: Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI), 2012.

[13] Zachary Miller , Brian Dickinson , William Deitrick , Wei Hu ,Alex Hai Wang ,Twitter spammer detection using data stream clustering,Pennsylvania State University, Dunmore, PA, United States,2013,elsevier.

[14] Chao Yang, Robert Harkreader, Jialong Zhang, Seungwon Shin, Guofei Gu, Analyzing spammers' social networks for fun and profit, in: Proceedings of the 21st International Conference on, World Wide Web, 2012, pp. 71–80.

[15] Zi Chu, Indra Widjaja, Haining Wang, Detecting social spam campaigns on twitter, in: Applied Cryptography and Network Security, Springer, Berlin Heidelberg, 2012, pp. 455–472.

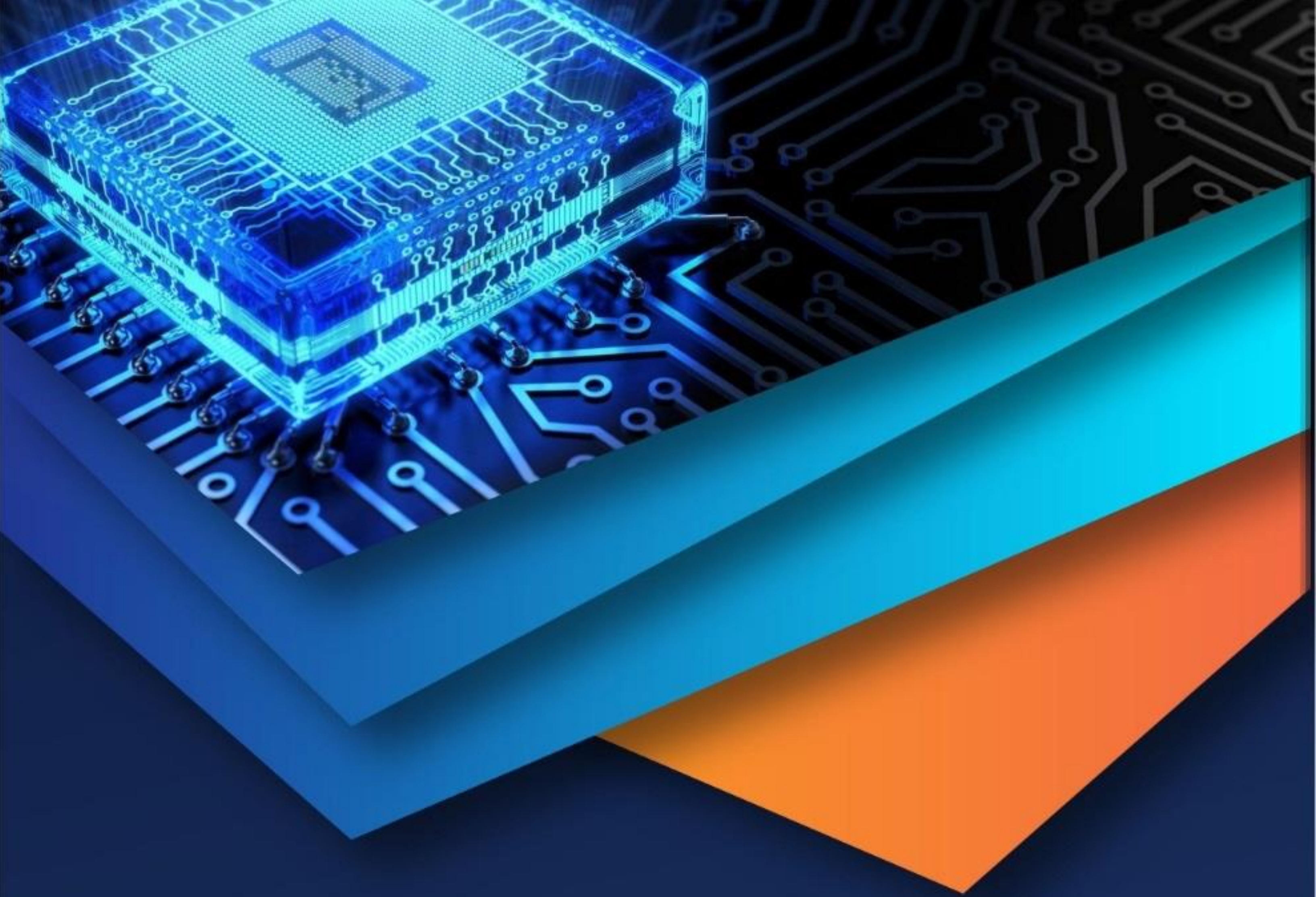
[16] Chris Grier, Kurt Thomas, Vern Paxson, Michael Zhang, @ Spam: the underground on 140 characters or less, in: Proceedings of the 17th ACM Conference on Computer and Communications Security, ACM, 2010, pp. 27–37.

[17] Kurt Thomas, Chris Grier, Justin Ma, Vern Paxson, Dawn Song, Design and evaluation of a real-time url spam filtering service, in: 2011 IEEE Symposium on Security and Privacy (SP), IEEE, 2011, pp. 447–462.

[18] Fabrıcio Benevenuto, Gabriel Magno, Tiago Rodrigues, Virgılio Almeida, Detecting spammers on twitter, in: Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS), 2010.

[19] Facebook, ⟨http://www.facebook.com/⟩.

[20] Welcome to Twitter, ⟨http://twitter.com/⟩.



doi crossref

10.22214/IJRASET

INDEX COPERNICUS

45.98

ISRA JIF

IMPACT FACTOR:
7.129

TOGETHER WE REACH THE GOAL

IMPACT FACTOR:
7.429

ISI

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY