Behavioural Based Online Comment Spammers in Social Media

G. Amudha , T. Jayasri , K. Saipriya , A. Shivani , C. H. Praneetha

Abstract: ***With the development of mobile Internet, it is changing the way we communicate with others. Internet media plays important application for information dissemination and user communication, including online news and social networks. However, the present advancement in technology provides opportunities to raise large number of spammers, who release false speech, advertisements and phishing websites on the media to gain commercial benefits, which seriously affects the experience of normal users. Therefore, in order to reduce the harm of false information, our paper focuses on the identification of spammers from normal users. However, the existing technologies of identifying spammers involve high data costs and poor effects, and most of them are concentrated in the field of social networks, while less research is carried out in the field of online news. In this paper, we propose an effective technology of identifying online news comment spammers based on the comment propagation algorithm (CPA), making full use of the user comment behaviours and contents. First of all, we extract few amount of information using scraping tool and label some users in the data as spammers or normal users manually to construct a labelled dataset. Next, we propose the identification technology based on the CPA. Finally, the set of values is input into the proposed technology in different combinations, and experiments and evaluations are carried out to determine possible spammers using behavioural features.***

I. INTRODUCTION

With the increase in the number of mobile Internet users and the high information of social life, there is a large amount of information at high value in mobile Internet, which is distributed in a variety of mobile crowd sourcing applications (5). At present, mobile crowd sourcing applications are generally composed of various Internet media, including online news, micro-blogs, blogs, and forums. The media are either social networks, or news media, or both , from which users can access a lot of information. But the potential business opportunities have made these Internet media flooded with a large amount of false information, such as false speech, numerous abusive advertisements. The false information has greatly reduced the user experience, affected the public opinions, reduced the quality of service, and impaired the economic interests, and thus it has become a serious problem of Internet media, which urgently needs to be regulated. And the false information derives from those people who commonly known as spammers. Spammer is a generic term of those malicious users who are driven by commercial interests to create and disseminate false information on Internet media by using special accounts registered on the platforms, in order to influence public opinions, disturb network order, and achieve other improper purposes . There have been spammers since the popularity of Internet media, and many technologies of spammer identification have been proposed in the industrial circles and academic circles. However, with the complexity of Internet environment, the behaviour of spammers is becoming more and more covert, the published false information is becoming more and more difficult to distinguish, and the data cost is becoming more and more high, which makes it more and more difficult to identify spammers . As a result, these technologies are either too limited or not suitable for the large-scale data sets nowadays, which need to be improved urgently. At present, most of technologies on the identification of spammers are focused on social networks, such as Twitter. There are few technologies on the identification of spammers on news media. Most of contents in social networks record users' private things, while the comments in news media are related to target news, so the standard for labelling spammers is different. Also, social networks contain social relationships such as followers, while news media does not have corresponding social relationships. Therefore, there are differences in the extraction of spammer features. Thus, it is necessary to study the identification technology of spammers on news media.

The development of mobile Internet has greatly expanded the proximity service, including public safety communications and commercial applications. The mobile crowd sourcing applications have also become an important research work of the proximity service but it is flooded with a large amount of false information.. This paper hands over all process of the identification technology to cloud computing.

II. DATASET COLLECTION AND PROCEESING

We extracted the data by using scrapping tool. For this purpose we use octoparse web scrapping tool. We followed the following steps:

1) Create a new task in advance mode.
2) Enter the URL and click the SAVE URL.
3) Build the workflow and extract the data.

Revised Manuscript Received on November 27, 2019.

***** Correspondence Author

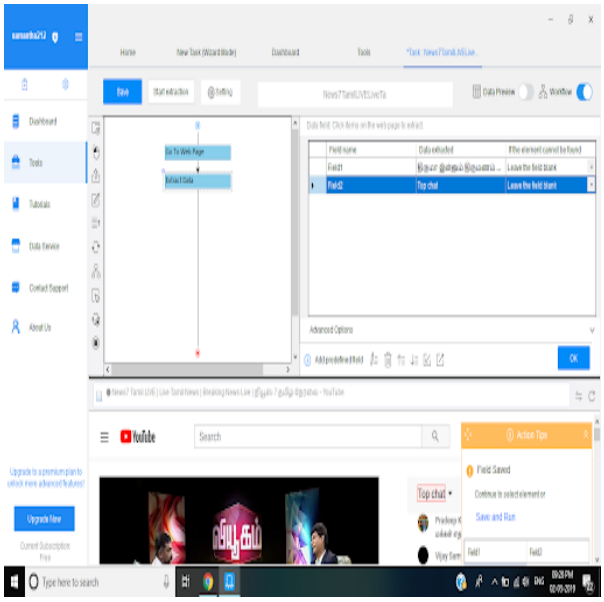
Dr. G. Amudha, Associate Professor, Department of Computer Science and Engineering, RMD Engineering College, Chennai. Email: gav.cse@rmd.ac.in

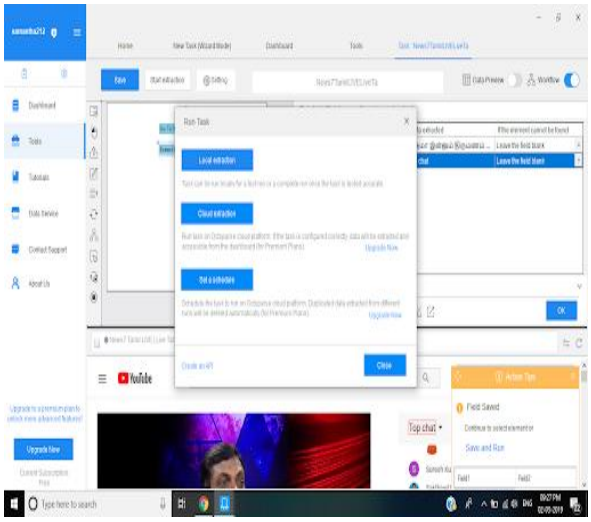
T. Jayasri, Department of Computer Science and Engineering, RMD Engineering College, Chennai. Email: ucs16132@rmd.ac.in

K. Saipriya, Department of Computer Science and Engineering, RMD Engineering College, Chennai. Email: ucs16201@rmd.ac.in

A. Shivani, Department of Computer Science and Engineering, RMD Engineering College, Chennai. Email: ucs17316@rmd.ac.in

C. H. Praneetha, Department of Computer Science and Engineering, RMD Engineering College, Chennai. Email: ucs17120@rmd.ac.in





III. DATA PREPROCESSING:

The data extracted will have some anonymous comments. These anonymous comments should be removed. So, it is also necessary to remove the users with fewer comments, because with manual checks, it is difficult to determine whether these users are normal or spammers. As a result our paper eliminates the users with fewer with 4 comments. The extraction technique is based on the semi supervised machine learning to label the users manually. All the users were sorted based on the number of comments from high to low. The current spammer identification technologies either use recognized datasets or use datasets collected by themselves. If they use the recognized datasets, there is no need to guarantee the authenticity of spammers' identity. Because its authenticity has been guaranteed by the data providing platform or domain experts. But most technologies use the datasets collected by themselves, they usually label which users are normal users and which users are spammers manually according to their own standards. It is impossible to ensure that the spammers labelled manually are real spammers, as unless the user acknowledges that he is a spammer or that the organization reveals him. That is, it often goes beyond the scope of technology to prove the authenticity of spammers that are manually labelled. This paper uses the dataset collected by us therefore; we call the manually labelled spammers as possible spammers to avoid discussing the authenticity.

IV. BEHAVIOURAL FEATURE ANALYSIS

After labelling possible spammers and normal users manually, this paper made a statistical analysis of user behavioural features (4), in order to find out the objective features which are helpful to identify possible spammers. In this paper, mainly the following four behavioural features are analyzed, namely, the number of new comments by users, the number of news commented, the number of comments during working hours and the frequency of comments. For the convenience of discussing the features of this paper, *Qn* and *Qs* are used to represent the number of normal users and that of possible spammers, and *Pn* and *Ps* are used to represent the user proportion of normal users and that of possible spammers. We adopted manual graph based analysis to feature our values between 0 and 1.

1) NUMBER OF NEW COMMENTS BY USERS

In daily life, people tend to get inspired when reading a piece of news, and then make some comments. For normal users, they will make new comments about news, but they are more about discussing with others, because one of the major functions provided by online news is the interaction between users. For possible spammers, they tend to just post new comments, instead of replying to others' comments, in order to increase the number of comments to finish their tasks as soon as possible. Therefore, this paper argues that the feature of the number of new comments by users can be used to distinguish between normal users and possible spammers, and the number of new comments by normal users is lower than that of possible spammers.

We calculated the ratio of number of new comments per user to their total comments number of comments as follows:

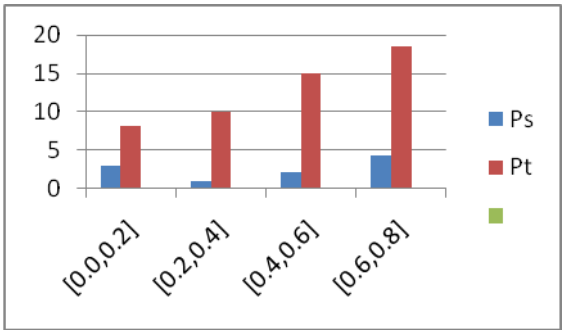
1) Count the number of new comments per user and define as Cn.

2) Count the total number of comments and define as Ct.

3) Calculate the ratio of number of new comments to the total comments and define it as A.

A=Cn/Ct

RANGE	Qs	Ps	Qn	Pn
[0.0,0.2]	11	2.84%	211	8.08%
[0.2,0.4]	3	0.78%	261	9.99%
[0.4,0,6]	8	2.07%	391	14.96%
[0.6,0.8]	16	4.13%	482	18.45%



2) NUMBER OF NEWS COMMENTED

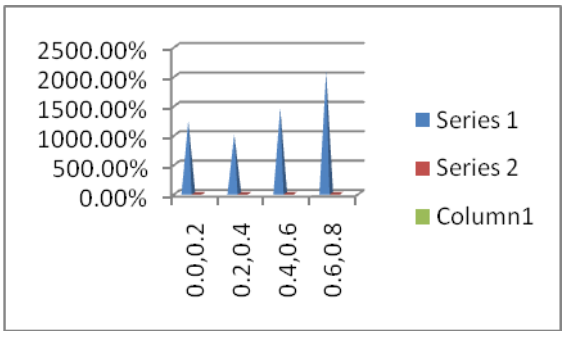
People have a variety of interests and are curious about different things, and the same is true when they read news. For normal users, they tend to read more news and comment on it instead of just making comments on several pieces of news. For possible spammers, they tend to comment on certain news for some purpose, such as raising the popularity of some news. Therefore, this paper argues that the feature of the number of news commented can be used to distinguish normal users from possible spammers, and the number of news commented by normal users is higher than that by possible spammers. We calculated the ratio of the number of news commented per user to their number of comments. The steps are as follows:

1. Count the number of news commented per user, and define it as *Nn*.
2. Count the total number of comments per user, and define it as *Na*.

3. Calculate the ratio of the number of news commented to the total number of comments, and define it as *B*, and its quantification equation is as follows:

B=Nn/Na

RANGE	Qs	Ps		
[0.0,0.2]	48	12.40%	10	0.38%
[0.2,0.4]	39	10.08%	20	0.77%
[0.4,0.6]	56	14.47%	57	2.18%
[0.6,0.8]	80	20.67%	231	8.84%



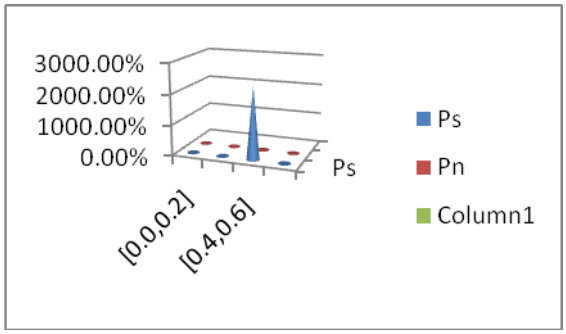
3) NUMBER OF COMMENTS DURING WORKING

In daily life, people will engage in a variety of jobs, and stick to their own posts. Reading news and making comments are entertainment activities, and such behaviour r tends to take place less frequently during working hours. For normal users, they tend to be busy with their work and spend less time reading news and making comments during their working hours. For possible spammers, the time they spend commenting on news often varies irregularly, as they may make comments during working hours or during non-working hours, which depends on whether possible spammers receive certain tasks. Therefore, this paper argues that the feature of the number of comments during working hours can be used to distinguish normal users from possible spammers, and the number of comments posted by normal users during working hours is lower than that of possible spammers.

We calculated the ratio of the number of comments posted during the working hours per user to their total number of comments. The steps are as follows:

1. Count the number of comments per user that are posted between the required time limits and define it as *Nw*.
2. Count the total number of comments per user, and define it as *Na*.
3. Calculate the ratio of the number of comments during working hours to the total number of comments, and define it as *Rw*, and its quantification equation is as follows:

RANGE	Qs	Ps	Qn	Pn
[0.0,0.2]	70	18.09%	515	19.71%
[0.2,0.4]	96	24.81%	935	35.78%
[0.4,0.6]	92	23.77%	722	27.63%
[0.6,0.8]	49	12.66%	308	11.79%



4) FREQUENCY OF COMMENTS:

As mentioned above, users comment on news because of their interest in it. If a user reads an interesting piece of news one day, he will have a heated discussion with other users, and thus this user will make a lot of comments on that day. Correspondingly, if the user does not read any news that he is interested in, there will be very few comments on the day. Such case is very common for normal users. After all, it is impossible to have news that users are interested in every day, and thus the time intervals between comments posted by normal users tend to be longer. For possible spammers, they tend to comment on news for some purpose, and they are likely to receive the task of commenting every few days or even every day, so the time intervals between comments posted by possible spammers are often shorter. Therefore, this paper argues that the

feature of the frequency of comments can be used to distinguish normal users from possible spammers, and the frequency of comments made by normal users is lower than that by possible spammers.

We calculated the ratio of the number of days when each user posted comments to the total number of days. The steps are as follows:

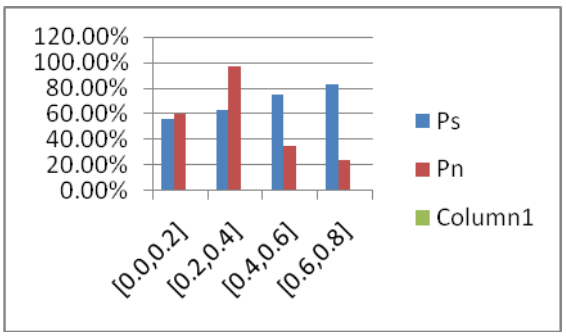
1. Sort the comments of each user according to the posting time, and extract the time of the earliest comment and the latest comment. Then subtract the two time values to get a time period, that is, the user has only posted comments within that time period. Define the time period as *Tn*

2. Count the number of days in which each user actually posts comments in his respective time period. For example, the time interval between the earliest comment and the latest comment posted by a user is 30 days, that is, *Tn*is 30. But this user has actually posted comments only in five days of *Tn*, so the actual number of days he has posted comments is 5, which is defined as Tf.

3. Calculate the ratio of the number of days when each user posted comments to the total number of days, and define it as *C*, and its quantification equation is as follows:

C=Tf/Tn

RANGE	*Qs*	*Ps*	*Qn*	*Pn*
[0.0.0.2]	***130***	***33.59%***	***1549***	***59.28%***
[0.2,0.4]	***127***	***32.82%***	***719***	***27.52%***
[0.4,0.6]	***38***	***9.82%***	***202***	***7.73%***
[0.6,0.8]	***25***	***6.46%***	***81***	***3.31%***



V. COMMENT PROPAGATION ALGORITHM:

Comment propagation is a graph based manual algorithm to distinguish the normal users with the possible spammers.

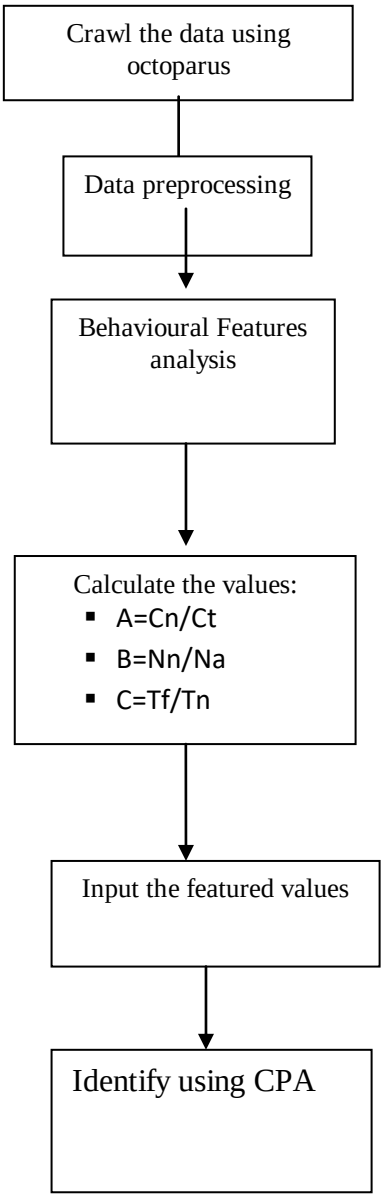
The data are collected using scraping tool and labelled as required fields. The comments are defined variously for better understanding of the algorithm. Our evaluation metrics uses the following entities such as,

- Direct Positive-number of possible spammers that are correctly classified.
- Direct Negative-number of normal users that are correctly classified.
- Indirect Positive-number of normal users that are wrongly classified as possible spammers.
- Indirect Negative-number of possible spammers that are wrongly classified as normal users.

From the above Behavioural analysis, the following data are calculated manually,

- A=Cn/Ct
- B=Nn/Na
- C=Tf/Tn

VI. ALGORITHM SEQUENCE FLOW:



VII. CONCLUSION:

The network needs security against attackers and hackers. Network Security includes two basic securities. The first is the security of data information i.e. to protect the information from unauthorized access and loss. And the second is computer security i.e. to protect data and to thwart hackers. This paper presents a technology of spammer identification based on CPA, which is suitable for the field of news media. By using the data crawled , a set of user behavioural features are extracted and input into CPA. Based on the experiments and evaluations, the results show that the identification technology in this paper is reliable and effective, and has higher accuracy, lower data cost. On the other hand, the features extracted in this paper are based on manual selection, which is likely to overlook some useful features that have not been discovered. Therefore, we can consider how to apply the method of deep learning into the automatic feature extraction in the future. Also, we will improve the spammer identification technology to apply to more industrial applications in the future.

REFERENCES

1. *Zhu, Xiaojin. "Learning From Labeled and Unlabeled Data With Label Propagation".*
2. Jump up to:[a] [b] [c] [d] U.N.Raghavan – R. Albert – S. Kumara Near linear time algorithm to detect community structures in large-scale networks 2007
3. M. E. J. Newman, "Detecting community structure in networks", 2004
4. Amudha, G. & Narayanasamy, P. Wireless Pers Commun (2018) 102:3303. https://doi.org/10.1007/s11277-018-5369-2
5. Lavanya Ramprasad , G Amudha International Conference on Information Communication and Embedded Systems (ICICES2014)10.1109/ICICES.2014.7033826.

AUTHORS PROFILE

Dr. G. AMUDHA, B.E, M.E, Ph.D., pursued her Bachelors of Engineering (CSE) in the year 2002 from Periyar University and Master of Engineering in Computer Science and Engineering in the year 2007 from Anna University, Chennai. She bagged Ninth University Rank in M.E(CSE).Ph.D., in the area of Wireless Sensor Networks under the Faculty of Information Science and Technology, CEG Campus, Anna University, Chennai. She has 17 years of working experience in the teaching profession. She is coordinating Information Security Centre of Excellence activities. She obtained IBM - DB2, Tivoli, and RAD value added certifications. She bagged NPTEL Elite certificate in Introduction to Internet of Things and Cryptology. Her areas of interest are Cryptography and Network Security, Compiler Design, and Sensor Networks. She has guided eight Master of Engineering projects. She had coordinated the. She was associated as Co-coordinator with AICTE Sponsored Faculty Development Programme on “Provision of Urban Amenities in Rural Areas” and National Level Conference RING 2015. She has published nine research papers in journals and conferences. She was invited as a Guest Speaker in Anna University Sponsored Faculty Development Training Programme on Compiler Design in the topic of Code Optimization. She coordinates Cyber Security Centre of Excellence and conducts various training in the domain of vulnerability assessment. Completed various security based NPTEL faculty development programmes and completed online courses like Internet of Things, Cryptography and Network Security, Internetwork Security, Industrial Internet of things, Cryptology etc., She bagged the topper award in IoT NPTEL online course and CNS. In all courses she nagged Elite certificate. She also bagged CEH certification.

T.JAYASRI, student of R.M.D Engineering College pursuing her final year in the department of Computer Science and Engineering. She has presented paper on security issues in the prestigious institution of MIT campus. She has a student membership in the cyber security centre of excellence sponsored by Tata Consultancy Services. She has actively participated in the prominent “Capture The Flag” contest. She is working on the mini project “MENACING VIGILANT” to detect the sudden threats. She has competed in the Smart India Hackathon event. She attended the internship on “ETHICAL HACKING” in the RedBack IT solutions.

Kondula Saipriya, the student of RMD Engineering College, from the department of computer science and engineering. She has given talks for her paper presentation at a prestigious institution named MIT. Apart from the works she has done in the institution, she has participated in implantation training in Bharat Sanchar Nigam Limited .Also, she has successfully attended multiple workshops that was conducted at IIT-M Chennai and Global Techno solutions . She has also been a part at the prominent contest named “Capture the flag” conducted by the TCS .Also, she has been a part in the cyber security COE sponsored by TCS.

SHIVANI A is a Student of RMD Engineering college pursuing third year Bachelor of Engineering in Computer science. She has a Membership in Cyber Security Centre of Excellence Sponsored by Tata Consultancy Services Limited. She has done internship in Redback IT solutions Limited on PHP and Sql tools. She has developed demo web pages for companies. She has done mini project in Employment management systems using PHP and HTML.She is currently developing web page for online learning courses. She has done paper presentation in MIT Chennai on ”Recent security issues”.

Praneetha Ch, a student of RMD engineering college studying third year in the department computer science and engineering. She has participated in Capture the flag contest conducted by Tata Consultancy Service. She attended a Smart India Hackathon contest on the concept of farm dairy. She had certified with a Database Management System completion course. She had done her workshop on Programming in Java in IITM. She is working on a mini project of birthday wishes name print. She was offered with a BEC certification. She has a student membership in Cyber Security centre of excellence sponsored by Tata Consultancy Services.