International Journal of Soft Computing 12 (1): 7-12, 2017
ISSN: 1816-9503

Detecting Malicious User in YouTube Using Edge Rank Based Feature Set

Omar Hadeb Sadoon and Yuhanis Yusof
University Utara Malaysia, Changlun, Malaysia

Abstract: Social media are websites that provide a network of people channels to make connections. An example of the media is YouTube that connects people through video sharing. Unfortunately, due to the explosive number of users and various content sharing there exist malicious users who aim to self-promote their videos or broadcast viruses and malwares. Even though the detection of malicious users is based on various features such as content details, social activity, social network analysing or a hybrid of features, the detection rate is still considered low (i.e., 46%). This study proposes a new set of features which is based on edge rank concept that focuses on affinity, weight and decay. The research is realized by analysing a set of YouTube users and their shared video prior to classifying the users using seven classifiers. Evaluation is performed by comparing the classification results of the proposed features against the existing feature set. Experiments showed that 86% of the classifiers obtained better results when using the proposed feature set as compared to using the existing features. The average classification accuracy is at 95.6%. Such a result indicates that the proposed work would benefit YouTube users in obtaining the required multimedia content and creating trust among users. In addition, system resources can be optimized as malicious accounts do not exist.

Key words: Social network, spam detection, malicious users, edge rank, feature construction

INTRODUCTION

The success of social media platforms such as Facebook, Tweeter and YouTube in the last few years encouraged more users to engage with these sites (Benevenuto *et al*., 2009; Zheng *et al*., 2015). YouTube, like the other social media platforms, depends on media contents that are created and shared by users. Such an approach allows malicious users (i.e., spammers) to exploit it (Zheng *et al*., 2015). According to a market survey on the impact of spammers over social media in 2008, 83% of social media users have received at least one message or friend request from unknown accounts (Kiran, 2015).

One of the main aims of malicious users is to spam videos over video-sharing platforms (Kiran, 2015). Video spammers are motivated to perform spamming in order to promote a specific content. A video spam occurs when a video which has a content that is completely unrelated to the video's title is posted as a response to an opening video (Benevenuto *et al*., 2008). Since, users cannot easily identify a video spam before watching at least a segment of it, they will waste their system resources, in particular, the bandwidth. Furthermore, it compromises user patience and satisfaction with the system. Thus, identifying video spam is a challenging problem in social video sharing systems (Kiran, 2015). Up to date, YouTube platform has not published any findings on handling malicious users. It only considers text comment as part of spam message (Chowdury *et al*., 2013). In addition, YouTube announced through its "Policy Centre" that in detecting spammers, it depends on user's engagement in reporting or flagging at a channel or comment. Such an approach may provide a reasonable result, especially when users respond and report on malicious content. Nevertheless, there are also users who abuse this approach. These users report any video they dislike as YouTube spam, hence resulting the topic to be closed immediately, even though their report is not valid. This problem needs to be solved as YouTube is becoming a prominent part of life's daily routine (Benevenuto *et al*., 2008).

Existing literature on online social media has proposed several malicious user detection approaches. However, the proposed methods focus on social content analysis that relies on keyword-based filtering and URL-based detection (Bhat *et al*., 2014; Burnap *et al*., 2015). The keyword-based filtering has a limitation, especially when malicious users use "cloaking". The generated cloaking terms are not filtered as spam. Furthermore, the existing detection methods only focus on the English language. Therefore, if a user provides comments in other languages, it will not be detected as spam. On the other hand, the URL-based detection may not function if the hyperlink destination is hidden or changed. Moreover, other methods have identified video spamming in Youtube using a hybrid approach. The hybrid approach integrates a set of features, namely video attributes, user attributes and social network metrics. However, this integration is only able to identify 44 and

Corresponding Author: Omar Hadeb Sadoon, University Utara Malaysia, Changlun, Malaysia

46% of the video spammers (Benevenuto *et al*., 2008; Kiran, 2015). This limitation is due to the usage of single video details with irrelevant features.

This study employs the concept of Edge Rank Checker (ERC) used on Facebook that provides custom-tailored recommendations in order to learn about user's interests (Zheng *et al*., 2015). In detail, this study adopts ERC concept to construct new features of malicious users in Youtube network.

Literature review: People's lives have been changed and become intertwined over time through online communication. When the Internet appeared and online activity started, many options were presented to create and maintain online relationships. This can be seen in online social media sites. Unfortunately, these entertainments create an opportunity for cyberattack and online threats due to opening windows. Online social media sites offer details of the users, hence users become a target for spammers, scams and other various attacks. Users' information are extracted when they update their status, post short texts, links, images, videos and send messages. Furthermore, these attacks primarily flourish on social media networks such as Youtube. Besides, the popularity of these sites makes them perfect spots for performing cybercriminal activities (Alberto *et al*., 2015).

Current Online Social Media (OSM) or Online Social Network (OSN) offers two major characteristics. The first one is content sharing whereby contents that are created and shared by any user are available to other users for viewing, providing opinions, rating and bookmark. For example, in Youtube media, an uploaded video can be given a rating (Like, Dislike) and comments from registered users. Second, the OSN also offers levels of relationships between users. These are typically framed as follows, friendships, subscriptions where it specifies the interest of a user towards another user. For example, in order to keep updated with the latest activity for a specific channel on Youtube, a user can subscribe other user's channels. A relationship between any two users of such systems is typically asymmetric, i.e., a friendship link from user A to user B means that the former is interested in the latter's activity but not necessarily vice-versa (Chiluka *et al*., 2011).

Many studies on detection of malicious users over online social media have been conducted by mining social media contents and analysing them (i.e., content-based approach) (Alberto *et al*., 2015). For instance, mining comments activity of users and then use a supervised learning method to extract patterns to detect malicious contents. However, there is also a user-based approach that focuses on a number of friends, followers and the number of likes. This is also known as profile-based approach (Chowdury *et al*., 2013). Lee and Kim (2012) conducted a study that mines URL and produce URL redirecting patterns in order to detect malicious users.

On the another hand, there is also work that mines the social activity of users either based on posting behaviours or user behaviours (Benevenuto *et al*., 2008). Some other detection methods are based on learning classification models from social network analysis or network-based topological features of the interacting users/nodes over online social networks (Bhat and Abulaish, 2013). Each approach has many limitations to be considered for it to be an idealistic approach to detection. For instance, it is not easy to apply content-based features (Razmara *et al*., 2012) in malicious detection of video objects as textual features are also required. This also applies to the other approaches such as detecting spammers based on analysing user activities or social networks because user privacy could be affected and the behaviours of malicious users can quickly change (Zhu *et al*., 2012). Besides, malicious users are often seen to mimic legitimate user patterns of interaction behaviour, making it difficult to characterize them (Bhat and Abulaish, 2013).

The hybrid analysis is another approach of detecting malicious users where it uses a group of different features or an ensemble of classifiers. Besides, there is also work that integrates both the features and classifiers to enhance the classification results (Bhat *et al*., 2014).

Edge rank algorithm: Edge Rank Checker (ERC) is an algorithm used by Facebook to decide which posts/stories should appear in each user's newsfeed. The main function of this algorithm is to evaluate each post and try to understand the actual content of the post through its score. It is learned that the higher the ERC score, the less possibility it is to be a spammer (Zheng *et al*., 2015).

ERC is like a credit rating, although it is invisible, it is very important to each user. In the Facebook developer's conference they exposed three elements of the algorithm as shown in Eq. 1. This study will adopt this concept and implement it to understand the actual content of each post (i.e., video shared in Youtube) by constructing hybrid features based on the integration of content analysis and user behaviour approaches:

$$\text{EdgeRank} = \sum U_e \times W_e \times D_e \quad (1)$$

Where:

U_e = Affinity (the score between the viewing user and edge creator)

W_e = Weight (the weight for this edge type such as comments, likes and shares)

D_e = Decay (the decay factor based on how long the age of the edge has been created)

MATERIALS AND METHODS

Our goal is to construct a feature set to be used in classifying users of social video-sharing systems into either legitimate or malicious. In achieving this goal, this study crawled through Youtube to create a dataset on users. From the obtained data, the study then proposes hybrid features to be used in classifying types of users (i.e., malicious or legitimate). Details on the steps undertaken are presented in the following subsections.

Data collection: In order to collect YouTube data, this study uses the Web Scraper (WSC) that extracts data from web pages. The crawling strategy inspects users with an account on YouTube and the crawling duration employed is a 4 months period as implemented by Tan *et al.* (2013) and Alberto *et al.* (2015). The data collection process is divided into three phases; the first phase involves the process of randomly crawling YouTube main pages and picking up a list of channels addresses. To do so, searching technique based on keywords such "music", "movie", "game" and "cartoon" is used to enforce the web scraper to crawl through different categories. In total, there are 500 channels that were selected for the study. The second phase focuses on crawling through the contents of the identified channel addresses and stored data on the user profile. Last but not least, the third phase includes the process of scraping information on the videos and creating a file for each channel (Fig. 1).

Data pre-processing: Once all data is collected, the pre-processing phase is performed. The obtained data is cleaned by checking for missing values. This is done on the two produced files of data collection phase; channel and video. Then, all channel information is stored by calculating the average value of all videos shared by each user.

Feature construction: As identified from the literature (Benevenuto *et al*., 2008; Sureka, 2011; Chowdury *et al*., 2013; Kiran, 2015) various features could be extracted from Youtube for classification purposes. This study includes reported features of user profile and user behaviour. Table 1 demonstrates features that have been used for malicious detection (as in the literature). In addition, this study proposes new features that are extracted from Youtube and these are presented in Table 2.

The key aspect of feature construction is to explore the benefits of a new feature in improving classification accuracy. Hence, this study adopts edge rank (Zheng *et al*., 2015) in creating new features for malicious user classification. Table 3 demonstrates the proposed features based on edge rank concept of affinity, weight and decay.

Table 1: Traditional features extracted from Youtube

User-based features (dataset-UB)	User-behaviour features (dataset-UA)
Channel name (Sureka, 2011)	Total videos views (Benevenuto *et al*., 2008; Chowdury *et al*., 2013)
Channel ID (Sureka, 2011)	Total videos likes (Benevenuto *et al*., 2008; Chowdury *et al*., 2013)
Total channel subscribed (Benevenuto *et al*., 2008;	Total videos dislikes (Benevenuto *et al*., 2008; Chowdury *et al*., 2013)
Total channel videos number (Benevenuto *et al*., 2008)	-

Table 2: New features extracted from YouTube

User-based features (dataset-UB)	User-behaviour features (dataset-UA)
Joined date	Total videos shared
Discussions	
Playlist	
Description	
Country	
Links	
Profile picture	
Background picture	

Based on the findings, the required features can be constructed as shown in Table 4. Consequently, this study will use an equation to derive a new value from the extracted data for each new feature. Originally, the Affinity concept represents the trust level between users and channel owners. Similarly, the weight concept illustrates the value of each event based on users' engagements. Likewise, the decay concept shows the value of each event based on the event's age. Table 4 illustrates all 16 of the proposed features and their equations.

Classification: For classification purposes, a total of 7 algorithms implemented in WEKA were employed as classifiers in detecting Youtube malicious users. The classifiers include the KStar random committee (Selvakuberan *et al*., 2008) bayes net, naive bayesian (Muralidharan and Sugumaran, 2012) J48 (Mohamed *et al*., 2012) multilayer perceptron, LibSVM (Kumar and Kumar, 2011). All of the mentioned classifiers were fed by the same dataset in classifying malicious users in YouTube. This includes dataset-UB and UA that represent user based approach and user-behaviour approach, respectively. In addition, the experiment includes dataset-ER which contains the 16 constructed features based on edge rank concept.

Table 3: YouTube features based on edge rank concept (dataset-ER)

Affinity	Weight	Decay
Channel average upload	Like rate based on total views	Channel age
Subscribe rate based on total views	Dislike rate based on total views	Subscriber rate based on channel age
View rate based on total videos number	Like rate based on total videos number	View rate based on channel age
Subscribe rate based on total videos number	Dislike rate based on total videos number	Share rate based on channel age
Share rate based on total views		Like rate based on channel age
Share rate based on total videos number		Dislike rate based on channel age

Table 4: Youtube features construction equations based on edge rank aspects

Data driven name	Equation	
Channel age	$x = \text{Joined date-scraped date}$	(2)
Channel average upload	$x = \frac{\sum \text{Channel videos}}{\text{Joined date-scraped date}}$	(3)
Subscriber rate based on total videos number	$x = \frac{\text{Channel subscriber}}{\sum \text{Channel videos}}$	(4)
Subscriber rate based on channel age	$x = \frac{\text{Channel subscriber}}{\text{Joined date-scraped date}}$	(5)
Subscriber rate based on total views	$x = \frac{\text{Channel subscriber}}{\text{Channel views}}$	(6)
View rate based on channel age	$x = \frac{\text{Channel views}}{\text{Joined date-scraped date}}$	(7)
View rate based on total videos number	$x = \frac{\sum \text{Videos share}}{\sum \text{Channel videos}}$	(8)
Share rate based on total views	$x = \frac{\sum \text{Videos share}}{\text{Channel views}}$	(9)
Share rate based on channel age	$x = \frac{\sum \text{Videos share}}{\text{Joined date-scraped date}}$	(10)
Share rate based on total videos number	$x = \frac{\sum \text{Videos share}}{\sum \text{Channel videos}}$	(11)
Like rate based on total views	$x = \frac{\sum \text{Channel Likes}}{\text{Channel Views}}$	(12)
Like rate based on channel age	$x = \frac{\sum \text{Channel likes}}{\text{Joined date-scraped date}}$	(13)
Like rate based on total videos number	$x = \frac{\sum \text{Channel likes}}{\sum \text{Channel videos}}$	(14)
Dislike rate based on total views	$x = \frac{\sum \text{Channel dislikes}}{\text{Channel views}}$	(15)
Dislike rate based on channel age	$x = \frac{\sum \text{Channel dislikes}}{\text{Joined date-scraped date}}$	(16)
Dislike rate based on total videos number	$x = \frac{\sum \text{Channel dislikes}}{\sum \text{Channel videos}}$	(17)

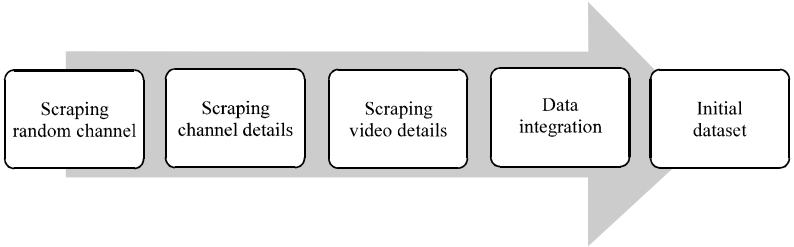

Fig. 1: Data collection process

Evaluation: To evaluate the proposed features, this study compares the effectiveness of the proposed feature set against the user-based and user-behaviour features. This is achieved by submitting the different feature sets (i.e., dataset-UB, UA, ER) to the 7 classifiers. Comparison of the classification accuracy (Singh *et al*., 2014) is made between these classifiers.

Table 5: Classification accuracy (%) for data proportion of 70:30

Classifier	Dataset-UB	Dataset-UA	Dataset-ER
KStar	92.66	98.00	98.00
Random committee	78.00	75.33	96.00
Bayes net	92.66	25.33	95.33
Naïve bayesian	92.66	30.66	95.33
J48	94.00	75.33	95.33
Multilayer perceptron	89.33	94.00	94.66
LibSVM	75.33	75.33	94.66

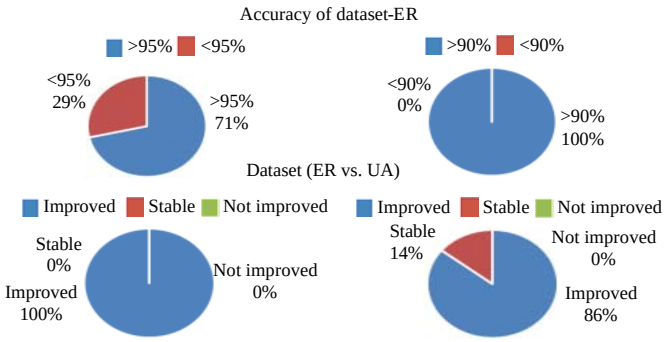

Fig. 2: The accuracy result out of total classifiers for data spilt into 70:30

RESULTS AND DISCUSSION

This section presents the experiment results of 7 different algorithms. The results are presented based on the different data proportion that was used in the experiments. Figure 2 visualized the statistics laid out by Table 5.

Table 5 shows the classification accuracy across 7 different classification algorithms based on 3 datasets using data proportion of 70:30 (i.e., 70% for training and 30% for testing). It is noted that the proposed feature set (i.e., dataset-ER) leads all classifiers to produce higher accuracy. Results showed that around 86% out of the total classifiers was improved by comparing the accuracy rate between hybrid features against non-hybrid ones. Besides, the average accuracy result of classification achieved a significant result at 95.6%. Moreover, 5 out of 7 algorithms were able to achieve classification accuracy over 95% based on EdgeRank. In constraining, only one classifier retained the same accuracy as using single features approach.

CONCLUSION

This study studies the relevant features to be used in classifying Youtube users. A new set of features is proposed and it consists of 16 features that were constructed based on Edgerank concept as applied in Facebook. The 7 classifiers were engaged to verify that the proposed features were able to recognize malicious and legitimate users. The new set of features is able to identify a significant fraction of video spam with an average accuracy of (95.6%) in our test collection. Such a result indicates promising utilization in detecting spammers in online social media particularly the one that involves video sharing.

REFERENCES

Alberto, T.C., J.V. Lochter and T.A. Almeida, 2015. Tubespam: Comment spam filtering on Youtube. Proceedings of the 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), December 9-11, 2015, IEEE, Sorocaba, Brazil, ISBN: 978-1-5090-0287-0, pp: 138-143.

Benevenuto, F., T. Rodrigues, V. Almeida, J. Almeida and C. Zhang *et al*., 2008. Identifying video spammers in online social networks. Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web, April 22, 2008, ACM, New York, USA., ISBN: 978-1-60558-159-0, pp: 45-52.

Benevenuto, F., T. Rodrigues, V. Almeida, J. Almeida and M. Goncalves, 2009. Detecting spammers and content promoters in online video social networks. Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, July 19-23, 2009, ACM, Boston, Massachusetts, ISBN: 978-1-60558-483-6, pp: 620-627.

Bhat, S.Y. and M. Abulaish, 2013. Community-based features for identifying spammers in online social networks. Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), August 25-28, 2013, IEEE, New Delhi, India, ISBN: 978-1-4503-2240-9, pp: 100-107.

Bhat, S.Y., M. Abulaish and A.A. Mirza, 2014. Spammer classification using ensemble methods over structural social network features. Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), Vol. 2, August 11-14, 2014, IEEE, New Delhi, India, ISBN: 978-1-4799-4143-8, pp: 454-458.

Burnap, P., A. Javed, O.F. Rana and M.S. Awan, 2015. Real-time classification of malicious URLs on Twitter using machine activity data. Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), August 25-28, 2015, IEEE, Cardiff, UK., ISBN: 978-1-4503-3854-7, pp: 970-977.

Chiluka, N., N. Andrade and J. Pouwelse, 2011. A link prediction approach to recommendations in large-scale user-generated content systems. Proceedings of the European International Confederence on Information Retrieval, April 18-21, 2011, Springer, Dublin, Ireland, pp: 189-200.

Chowdury, R., M.N.M. Adnan, G.A.N. Mahmud and R.M. Rahman, 2013. A data mining based spam detection system for youtube. Proceedings of the 2013 8th International Conference on Digital Information Management (ICDIM), September 10-12, 2013, IEEE, Dhaka, Bangladesh, ISBN: 978-1-4799-0615-4, pp: 373-378.

Kiran, P.S., 2015. Detecting spammers in YouTube: A study to find spam content in a video platform. IOSR. J. Eng., 5: 26-30.

Kumar, G. and K. Kumar, 2011. AI based supervised classifiers: An analysis for intrusion detection. Proceedings of the International Conference on Advances in Computing and Artificial Intelligence, July 21-22, 2011, ACM, NEW York, USA., ISBN: 978-1-4503-0635-5, pp: 170-174.

Lee, S. and J. Kim, 2012. WarningBird: Detecting suspicious URLs in Twitter Stream. Pohang University of Science and Technology, Pohang, South Korea. https://pdfs.semanticscholar. org/9353/31d53a4f2f19b3e4e862a183eca6dc61203d.pdf.

Mohamed, W.N.H.W., M.N.M. Salleh and A.H. Omar, 2012. A comparative study of reduced error pruning method in decision tree algorithms. Proceedings of the 2012 IEEE International Conference on Control System, Computing and Engineering (ICCSCE), November 23-25, 2012, IEEE, Batu Pahat, Malaysia, ISBN: 978-1-4673-3143-2, pp: 392-397.

Muralidharan, V. and V. Sugumaran, 2012. A comparative study of Naive Bayes classifier and Bayes net classifier for fault diagnosis of monoblock centrifugal pump using wavelet analysis. Appl. Soft Comput., 12: 2023-2029.

Razmara, M., B. Asadi, M. Narouei and M. Ahmadi, 2012. A novel approach toward spam detection based on iterative pat-terns per text. Msc Thesis, Islamic Azad University, Arak.

Selvakuberan, K., M. Indradevi and R. Rajaram, 2008. Combined feature selection and classification a novel approach for the categorization of web pages. J. Inf. Comput. Sci., 3: 083-089.

Singh, M., D. Bansal and S. Sofat, 2014. Detecting malicious users in twitter using classifiers. Proceedings of the 7th International Conference on Security of Information and Networks, September 09-11, 2014, ACM, New York, USA., ISBN: 978-1-4503-3033-6, pp: 247-247.

Sureka, A., 2011. Mining user comment activity for detecting forum spammers in Youtube. Indraprastha Institute of Information Technology Delhi, New Delhi, India. https://arxiv.org/pdf/1103.5044.pdf.

Tan, E., L. Guo, S. Chen, X. Zhang and Y. Zhao, 2013. Unik: Unsupervised social network spam detection. Proceedings of the 22nd ACM International Conference on Information and Knowledge Management, October 27-November 01, 2013, ACM, New York, USA., ISBN: 978-1-4503-2263-8, pp: 479-488.

Zheng, X., Z. Zeng, Z. Chen, Y. Yu and C. Rong, 2015. Detecting spammers on social networks. Neurocomputing, 159: 27-34.

Zhu, Y., X. Wang, E. Zhong, N.N. Liu and H. Li *et al*., 2012. Discovering spammers in social networks. Hong Kong University of Science and Technology, Hong Kong. http://home.cse.ust.hk/~qyang/Docs/2012/AAAI_2012_Spammers.pdf.