How to cite this paper:

Yuhanis Yusof & Omar Hadeb Sadoon. (2017). Detecting video spammers in youtube social media in Zulikha, J. & N. H. Zakaria (Eds.), Proceedings of the 6th International Conference of Computing & Informatics (pp 228-234). Sintok: School of Computing.

DETECTING VIDEO SPAMMERS IN YOUTUBE SOCIAL MEDIA

Yuhanis Yusof[1], and Omar Hadeb Sadoon[2]

[1]*Universiti Utara Malaysia, Malaysia*, yuhanis@uum.edu.my, ir7aqi@yahoo.com

ABSTRACT. Social media is any site that provides a network of people with a place 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 unrelated content. Even though the detection of malicious users is based on various features such as content details, social activity, social network analysing, or hybrid, the detection rate is still considered low (i.e. 46%). This study proposes a new set of features by constructing features based on the EdgeRank algorithm. Experiments were performed using nine classifiers of different learning; decision tree, function-based and bayesian. The results showed that the proposed video spammers detection feature set is beneficial as the highest accuracy (i.e average) is as high as 98% and the lowest was 74%. The proposed work would benefit YouTube users as malicious users who are sharing non-relevant content can be automatically detected. This is because system resources can be optimized as YouTube users are presented with the required content only.

Keywords: spammers detection, EdgeRank, malicious users, cybersecurity.

INTRODUCTION

Malicious users try to compromise computers and sensitive information from the inside as authorized and "trusted" users. Malicious users go for systems they believe they can compromise for illegal gains or revenge (Zheng, Zeng, Chen, Yu, & Rong, 2015). Malicious attackers are, generally speaking, both hackers and malicious users. Malicious users are often the worst enemies of IT and information security professionals because they know exactly where to go for getting the goods. They do not need to be computer savvy to compromise sensitive information. There are various malicious items over the social network and this depicted in Table 1.

Table 1. Types of Malicious Items over OSN.

Type	Details
Spam	Text comments that have commercial content unrelated to the discussion at hand or it involve contacting users with unwanted content (Profanity, Insults, Bulk, Hate speech, Threats) or requests (Facebook Help Center, 2016)
Video spam	A video with unrelated content compared to its title or a video without content (Benevenuto et al., 2008)
Malicious links	A link that misleads users, or a link with inappropriate harm, or damage a user account or computer (Burnap, Javed, Rana, & Awan, 2015)

Type	Details
Fraudulent reviews	Fake reviews on a service or product from a user who has never used it, hence giving misleading information (Hsu, 2012)
Fake friends, Subscribers	Fake accounts that are owned by users who try to gain credibility by following/subscribing to certified accounts/channel, such as those of popular celebrities and public figures (Fernandes, Patel, & Marwala, 2015)

One of the main aims of the malicious user is video spamming over video-sharing platforms (Hu, Tang, & Liu, 2014; Kiran, 2015). Video spammers are motivated to perform spamming in order to promote specific content (Alberto, Lochter, & Almeida, 2015; Chowdury, Adnan, Mahmud, & Rahman, 2013). A video spam occurs when a video posted as a response to an opening video. Whereas, the content is completely unrelated to the video's title (Benevenuto et al., 2008). Since users cannot easily identify a video spam before watching at least a segment of it, users 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 (Benevenuto et al., 2008; Chowdury et al., 2013; Kiran, 2015) such as in the YouTube. 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 Center" (Google, 2016), to detect 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 it. These users report any dislike video 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 daily life routine (Benevenuto et al., 2008; Sandvine, 2015).

This study proposes a new feature to be used in detecting video spammers. The proposed work is based on the employment of EdgeRank that was introduced on Facebook. Edge Rank checker (ERC) is an algorithm used by Facebook to decide which post/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 can be seen that the higher ERC score, the less possibility to be a spammer (Zheng et al., 2015). ERC is like a credit rating, although it's invisible, but it's very important to each user (Jeff, 2015). In the Facebook developer conference (Facebook, 2010), they exposed three elements of the algorithm; affinity, weight, and decay. This study will adopt this concept and implement it to understand the actual content of each post (i.e video) over YouTube by constructing features based on the integration of content analysis and user behaviour approaches.

RELATED WORK

Many studies on the detection of malicious items on the social network have been conducted and this includes mining the media content and analysing it (i.e. Content-based) (Alberto et al., 2015). For instance, mining the comments provided by users and learn the pattern to detect malicious contents. However, mining the keywords requires large computational cost and it is limited to the listed words only. As comments could be written using formal and informal language, relying on keywords would not be efficient.

In addition, there is also a user-based approach that examines the relation that a user has with other people, such as a number of friends, followers, and the number of like that a user obtained. Such an approach is known as profile-based (Chowdury et al., 2013). On the another hand, there is also work that mines users social activity either based on posting behaviours

or user behaviours (Benevenuto et al., 2008). The behaviour of malicious users in Twitter has been examined and it learned that behaviour of malicious users is different from the legitimate users in terms of posting tweets, following friends, followers and so on (Yardi, Romero, Schoenebeck, & Boyd, 2010). On the other hand, there is also work that uses the features extracted from social network graph (Bhat & Abulaish, 2013). An example of a study conducted using this approach was to detect campaigns of malicious users and it relies on network usage features (O'Callaghan, Harrigan, & Carthy, 2012).

METHOD

This study undergoes five main phases; data collection, data preprocessing, feature construction, spam detection, and evaluation. Details of the activities involved in the phases are illustrated in Figure 1.

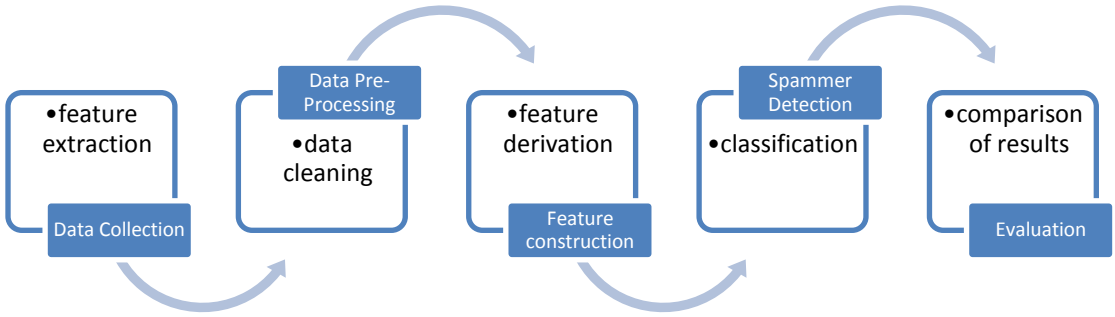

Figure 1. Research Activities.

In the first phase, the study employed Web Scraper (WSC) (Web Scraper, 2016) that extracts data from web pages. The crawling inspects users with an account on YouTube where the crawling duration is of the period of four months as implemented in (O'Callaghan et al., 2012; Tan, Guo, Chen, Zhang, & Zhao, 2013; Alberto et al., 2015). Once the YouTube users have been identified, features of the channel owner and the shared content are extracted. In this study, a total of 30,621 videos that belongs to 500 YouTube users were analyzed. Once the features are available, there is a need to check for any missing values, and this process is known as data cleaning. This study has replaced the missing values using the mean based imputation.

The contribution of this study is depicted in the 3rd phase that includes feature construction. The process was based on Edge Rank algorithm employed by Facebook in determining a recommendation for users. Using the extracted features, this study proposes new features to be used in detecting video spammers. Details of the features are presented in Table 2.

Table 2. Proposed Features for Detection of Video Spammers.

Data Driven Name	Equation
Channel age	$x = \text{Joined Date} - \text{Scraped Date}$
Channel average upload	$x = \frac{\sum \text{Channel Videos}}{\text{Joined Date} - \text{Scraped Date}}$

Subscriber rate based on total number of videos	$x = \frac{\text{Channel Subscriber}}{\sum \text{Channel Videos}}$
Subscriber rate based on channel age	$x = \frac{\text{Channel Subscriber}}{\text{Joined Date} - \text{Scraped Date}}$
Subscriber rate based on total views	$x = \frac{\text{Channel Subscriber}}{\text{Channel Views}}$
View rate based on channel age	$x = \frac{\text{Channel Views}}{\text{Joined Date} - \text{Scraped Date}}$
View rate based on total number of videos	$x = \frac{\sum \text{Videos Share}}{\sum \text{Channel Videos}}$
Share rate based on total views	$x = \frac{\sum \text{Videos Share}}{\text{Channel Views}}$
Share rate based on channel age	$x = \frac{\sum \text{Videos Share}}{\text{Joined Date} - \text{Scraped Date}}$
Share rate based on total number of videos	$x = \frac{\sum \text{Videos Share}}{\sum \text{Channel Videos}}$
Like rate based on total views	$x = \frac{\sum \text{Channel Likes}}{\text{Channel Views}}$
Like rate based on channel age	$x = \frac{\sum \text{Channel Likes}}{\text{Joined Date} - \text{Scraped Date}}$
Like rate based on total number of videos	$x = \frac{\sum \text{Channel Likes}}{\sum \text{Channel Videos}}$
Dislike rate based on total views	$x = \frac{\sum \text{Channel Dislikes}}{\text{Channel Views}}$
Dislike rate based on channel age	$x = \frac{\sum \text{Channel Dislikes}}{\text{Joined Date} - \text{Scraped Date}}$
Dislike rate based on total number of videos	$x = \frac{\sum \text{Channel Dislikes}}{\sum \text{Channel Videos}}$

In total, there are 16 features derived based on the three aspects of Edge Rank; six features illustrate Affinity, four features represent Weight and another six features show Decay. Originally, the Affinity refer to the trust level between users; the Weight score the value of each event based on user's engagements while Decay shows the value of each event based on event's age. These features are then fed to different types of classifiers; decision trees, function-based and bayes. These three categories of learners are represented by 2 bayes, 5 function-based and 6 decision trees algorithms. Information on the employed classifiers is as in Figure 2.

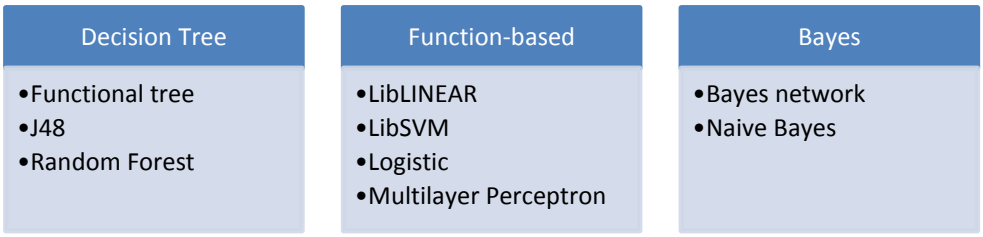

Figure 2. Classifiers for Detecting Video Spammers.

In the evaluation phase, detection accuracy obtained by all of the employed classifiers (i.e 9 algorithms) are compared. In addition, the comparison is also based on the testing strategy utilized during the experiments; percentage split and cross-validation. The aim was to investigate which classifier produces the best result while using the same feature set. This will then show the effectiveness of the proposed detection features.

RESULTS

Results for the undertaken experiments are depicted in Table 3 and Table 4. The experiment that includes three data proportion for the percentage split testing strategy reveals that the Bayes classifiers are best to utilize the proposed features. Data in Table 3 shows that the average detection accuracy for both the Bayes Network and Naïve Bayesian is as high as 98%. While using the 90:10 data proportion, both of the classifiers produced 100% accuracy.

On the other hand, the Multilayer Perceptron produces the lowest detection result, which is 88%. The highest accuracy was 95% and this is obtained while using the 70:30 data proportion. In general, it can note that bayes ranks the first, followed by decision tree and function-based classifiers.

Table 3. Detection Accuracy: Percentage Split.

Classifier	70:30	80:20	90:10	average
Functional Tree	95	96	96	95.67
J48	95	97	98	95.67
Random Forest	97	94	100	97
LibLINEAR	95	95	92	94
LibSVM	95	96	92	94.33
Logistic	95	94	98	95.67
Multilayer Perceptron	95	89	88	90.67
Bayes Network	95	99	100	98
Naïve Bayesian	95	99	100	98

The other performed experiment was on using a various number of folds for cross-validation testing. Details of the results are shown in Table 4. In this experiment, a different result is obtained. Data in the table showed that the highest accuracy was obtained while using the Functional tree classifier (average 97.67%). The classifier shows a consistent perfor-

mance as it produced similar accuracy while using a different number of folds for the cross-validation strategy. On the other hand, the performance of the bayes classifiers are different when cross validation is employed. Both the bayes classifiers could not perform as good as in the percentage split testing. In fact, the Naïve Bayesian produced the lowest accuracy which is 74%.

Table 4. Detection Accuracy: Cross Validation.

Classifier	k=10	k=15	k=20	average
Functional Tree	98	97	98	97.67
J48	95	96	96	95.67
Random Forest	96	96	96	96
LibLINEAR	94	93	94	93.67
LibSVM	91	91	92	91.33
Logistic	97	95	97	96.33
Multilayer Perceptron	86	90	90	88.67
Bayes Network	95	95	95	95
Naïve Bayesian	74	74	74	74

CONCLUSION

Social media networks have become extremely popular and this creates the opportunity for the malicious user to publish unwanted content such as video spam. This study has introduced the feature set to be used in detecting video spammers that exist in the YouTube media. The features were constructed based on the features obtained from the user profile and the content that they shared. Based on the undertaken experiments, it is learned that existing classifiers that were widely used in the data mining community could utilize the features in detecting video spammers. The average detection accuracy was as high as 98% while the lowest value was only 74%. Such a result provides insight on the usefulness of the proposed video spammer feature set. In order to investigate more, additional experiments need to be performed, comparing the results against existing feature set for spammer detection.

REFERENCES

Alberto, T. C., Lochter, J. V, & Almeida, T. A. (2015). TubeSpam: Comment Spam Filtering on YouTube. *Proceedings of the 14th IEEE International Conference on Machine Learning and Applications (ICMLA'15)*, 138–143. Journal Article.

Benevenuto, F., Rodrigues, T., Almeida, V., Almeida, J., Zhang, C., & Ross, K. (2008). Identifying video spammers in online social networks. *Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web AIRWeb 08*, 45.

Bhat, S. Y., & Abulaish, M. (2013). Community-based features for identifying spammers in online social networks (12). In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining - ASONAM '13* (pp. 100–107). http://doi.org/10.1145/2492517.2492567

Burnap, P., Javed, A., Rana, O. F., & Awan, M. S. (2015). Real-time classification of malicious URLs on Twitter using machine activity data. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2015* (pp. 970–977). http://doi.org/10.1145/2808797.2809281

Chowdury, R., Adnan, M. N., Mahmud, G. A. N., & Rahman, R. M. (2013). A data mining based spam detection system for YouTube. In *Proceedings of the 8th International Conference on Digital Information Management (ICDIM'13)* (pp. 373–378).

Facebook, I. (2010). *Facebook Developer Conference.* Retrieved January 1, 2016, from https://www.fbf8.com/

Facebook Help Center. (2016). *What is spam?* Retrieved January 1, 2016, from https://www.facebook.com/help/1461986764019121

Fernandes, M. A., Patel, P., & Marwala, T. (2015). Automated Detection of Human Users in Twitter. *Procedia Computer Science*, *53*, 224–231.

Google. (2016). *Policy Center*. Retrieved January 10, 2016, from https://support.google.com/youtube/topic/2803176?hl=en&ref_topic=2676378

Hsu, T. (2012). Yelp's new weapon against fake reviews: *User alerts.* Retrieved from http://www.latimes.com/business/la-fi-mo-yelp-fake-review-alert-20121018-story.html

Hu, X., Tang, J., & Liu, H. (2014). Online social spammer detection. In *Twenty-Eighth AAAI Conference on Artificial* (pp. 59–65). http://doi.org/10.1109/ICDM.2014.141

Jeff. (2015). *EdgeRank.* Retrieved January 1, 2016, from http://edgerank.net/

Kiran, P. S. (2015). Detecting spammers in YouTube : A study to find spam content in a video platform. *IOSR Journal of Engineering (IOSRJEN)*, *5*(7), 26–30.

O'Callaghan, D., Harrigan, M., & Carthy, J. (2012). Network analysis of recurring youtube spam campaigns. *arXiv Preprint arXiv:* Retrieved from http://arxiv.org/abs/1201.3783

Tan, E., Guo, L., Chen, S., Zhang, X., & Zhao, Y. (2013). UNIK: unsupervised social network spam detection. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management - CIKM '13* (pp. 479–488).

Web Scraper. (2016). *Web Scraper*. Retrieved January 1, 2016, from http://webscraper.io/

Yardi, S., Romero, D., Schoenebeck, G., & Boyd, D. (2010). Detecting spam in a Twitter network. *First Monday*, *15*(1). Retrieved from http://firstmonday.org/ojs/index.php/fm/article/viewArticle/2793

Zheng, X., Zeng, Z., Chen, Z., Yu, Y., & Rong, C. (2015). Detecting spammers on social networks. *Neurocomputing*, *159*, 27–34.