IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 10, No. 1, March 2021, pp. 139~150
ISSN: 2252-8938, DOI: 10.11591/ijai.v10.i1.pp139-150

Analysis of spammers' behavior on a live streaming chat

Sawita Yousukkee, Nawaporn Wisitpongphan
Faculty of Information Technology and Digital Innovation, King Mongkut's University of Technology North Bangkok, Bangkok, Thailand

Article Info

Article history:

Received Jul 30, 2020
Revised Jan 20, 2021
Accepted Feb 10, 2021

Keywords:

Behavior analysis
Live chat behavior
Live chat message
Spammer's behavior
Viewer's behavior

ABSTRACT

Live streaming is becoming a popular channel for advertising and marketing. An advertising company can use this feature to broadcast and reach a large number of customers. YouTube is one of the streaming media with an extreme growth rate and a large number of viewers. Thus, it has become a primary target of spammers and attackers. Understanding the behavior of users on live chat may reduce the moderator's time in identifying and preventing spammers from disturbing other users. In this paper, we analyzed YouTube live streaming comments in order to understand spammers' behavior. Seven user's behavior features and message characteristic features were comprehensively analyzed. According to our findings, features that performed best in terms of run time and classification efficiency is the relevant score together with the time spent in live chat and the number of messages per user. The accuracy is as high as 66.22 percent. In addition, the most suitable technique for real-time classification is a decision tree.

Corresponding Author:

Sawita Yousukkee
Faculty of Information Technology and Digital Innovation
King Mongkut's University of Technology North Bangkok
Bangkok, Thailand
Email: s5807011966017@email.kmutnb.ac.th

1. INTRODUCTION

Spam is an unwanted message that consists of texts and links. Besides insulting posts, mass messaging, cruelty, humiliation, hate speech, malicious, fake hints [1], or fraudulent reviews, most spams messages are intended for advertising purposes: explicit advertising message or links to malicious websites in the form of long URLs or short URLs (Google URL Shortener, Bitly, TinyURL) redirecting to a product's website in order to increase the site's rating or promote products.

One good example of spam is clickjacking [2]: a malicious technique that tricks users into clicking on an encapsulated application disguised in the form of a picture, link, button, or something else that is intended for clicking. Users are usually unaware of any deception when they enter data into forms connected to clickjacking. Clicking on such a page, which seems innocuous, may result in the installation of a harmful program or disclosure of confidential information in a computer or URL redirection mechanisms that attacker can automatically redirect a visitor to a captured web site [3].

Video and comment spam is also another common unwanted content that may be generated and posted by an automated program, bot, or fake users [4]. Posted contents are usually unrelated to the topic of interest.

Recently, there has been dramatic growth in the amount of spam found in online social networks and live streaming media. On twitter network, spam or faux tweets [5-6] are in form of a hashtag that is not connected to topics or unrelated topics, hashtag may include offensive material such as ad hominem attacks or vulgar language. Also, the same hashtag may mean different things to different people and the motivations

behind spam may bring participants' attention to products or services. To promote the contents on social media, these online media allow users to input their opinions and express their views on the contents which gives rise to review spam or opinion spam. The opinion content may or may not contain valuable information in either positive or negative ways. If the content is related to consumer products, it may provide information that is valuable to the product owner and will assist the owner to investigate and identify product problems and improve customers' satisfaction. Review spam can be separated into 3 types [7]. The first is false opinions, where the review contains a false opinion to promote competitive products and encourage negative opinions to damage the product under review. Second is reviews on brand only, which comprise reviews which do not relate to the specific product being reviewed, but merely express an opinion on the brand. This type of review can often be highly biased. The third type is non-reviews, which are reviews that contain messages unrelated to the product being reviewed.

Live streaming is an easy way to broadcast videos in real-time, and it has become very popular because it provides a platform for interaction and content sharing to reach a large target audience. A public live streaming platform such as YouTube has a live chat feature that allows viewers to communicate in real-time by chatting with content owners and other viewers. The product owner can exploit the benefits of this medium and improve services to meet the needs of the customer in real-time. On the other hand, live chat has become a vulnerable access point in live streaming because it can be a channel where other marketers promote their websites or services. Spammers can use this service to spread unwanted content or spam content such as porn sites and malware.

There has been and continues to be, a great deal of research on the prevention of spam comments on social networks, especially YouTube. The work can be divided into 2 areas [8]: prevention and detection based. To prevent spam comments from appearing on the live chat, YouTube allows content owners to provide a list of blocked words which may indicate spam in addition to having a moderator to monitor live chats and manually remove unwanted conversations. If a message contains a blocked word, the system will delete the message or move it to a held for the review section. To detect comment spam, Benevenuto, *et al.* proposed a spam comment classifier [9] as a browser plugin by using a set of features such as stop word ratio, word duplication ratio, post-comment similarity, the number of words and the number of sentences in the comment. Their results suggested that random forest tree can accurately classify spam when used with all of the features or a subset of features. Similarly, tubespam purposed in [10] can automatically filter undesired comments posted on YouTube. In the experiment, the researchers collected text comments posted in five of the most viewed YouTube videos (Psy, KatyPerry, LMFAO, Eminem, and Shakira) and manually labeled each comment spam or ham (legitimate comment) using bag-of-words and frequency representation. The result of the experiment showed that most classification techniques such as decision tree, logistic regression, bernoulli naïve bayes, random forests, linear, and gaussian SVMs have spam detection and blocking accuracy rates higher than 90% and ham blocking rates of lower than 5%.

In this paper, we focused on the study of users' behavior on YouTube live streaming. In particular, we explored whether different potential markers in user comments, such as relevant score using bag-of-words, similarity between comments, polarity of the comments, number of chat messages, interarrival time of chat messages, and time duration that a user spent in a live chat can be used to effectively differentiate spammers from normal users.

2. BACKGROUND OF THE STUDY

Spam, poses a huge challenge for researchers aiming to find ways to detect or exclude these unwanted messages from nearly every online media. Many studies focused on traditional methods, such as content-based analysis or extraction of features from the content or their information. Analysis of users' behavior is often, used to improve the accuracy and performance of spam detection.

In an experiment based on content analysis, Rathod and Pattewar [11] analyzed a body of the Gmail dataset to classify legitimate and spam email using the Bayesian classifier model. The proposed bayesian classifier was able to achieve as high as 96.46% accuracy. Moreover, similarity and relevance are important features used to distinguish spams from regular messages. Liu *et al.* [7] developed two algorithms to identify false reviews on Amazon.com based on the similarity of the reviews and how much the review content is related to the product and to describe some common behavior features of spammers in the spam review. According to the observation, it was found that if the similarity of the two reviews is greater than 70%, then the second review was identified as a copied review. The results showed that 54% of mobile phone reviews on Amazon.com are copied from existing reviews. Jindal and Liu [7] categorized customer reviews of Amazon.com into three types: false opinion, brand review, and non-reviews. They used a classification model to detect duplicate reviews and differentiate each review as spam or non-spam. In the analysis, they found a large number of duplicate reviews whose similarity scored were close to 90% and were classified as

spam. They, then performed manual label training for type 2 (brand review) and type 3 (non-reviews) and used classification models such as SVM, decision tree, naïve bayesian, and logistic regression to detect spam. Three types of features, characteristics of reviews, characteristics of reviewers, and characteristics of products, were used for learning. The result of the experiment showed that the logistic regression is a highly effective classification tool which resulted in the AUC value (area under ROC curve) of 98.7% for all spam types.

There are several studies that investigated the spammers' behavior on various social media platforms. Verma, *et. al.* [12] reported that spammers on Twitter tend to follow numerous users but have only a small number of followers. They often continue to create new accounts to avoid being detected. Other studies have found that typical spammers may post at a fixed schedule as is indicated by a large number of repetitive tweets in a short period of time or a more frequent posting during certain times [13]. The tweet content normally contains trendy keyword, latest articles from a website with unrelated messages. Similarity, spammers on YouTube repeatedly posted chat messages with emoji, links, promotion, or harmful content to direct uesrs to their website. Since YouTube can detect spam based on the text comment and users' behavior, some spammers try to avoid being detected by using several techniques: re-posting someone else's message and including advertising message in their usernames, using unusual Unicode characters that resemble English alphabets to compose certain spam words, and using multiple accounts.

This study, we analyze key features related to users' behavior characteristics (interarrival time or the time between two consecutive chat messages from the same user, polarity score, the duration of time a user spent on a live chat, and the number of chat message per user) and chat message characteristics (word count, relevant score, and similarity score) to understand the characteristics of users which can be used to identify the activity of bots and spammers on YouTube live chat.

3. METHODOLOGY

3.1. Data collection

In this study, we collected live chat messages on weather channel HD live stream which featured Hurricane IRMA from 8 September 2017 to 9 September 2017. A total of 60,115 chat messages were collected during a period of 17.30 hours. Our system tracked the number of viewers and the number of chat messages throughout the entire duration of the stream.

Figure 1 shows both the number of users watching the live streaming content and the number of chat messages during 00:00 to 17:30. As expected, as the number of live stream viewers increased, the number of chat messages posted on the live chat also increased. When the streaming ended at 17:30, the number of chat messages quickly dropped to zero because the live chat was disabled after live streaming ended.

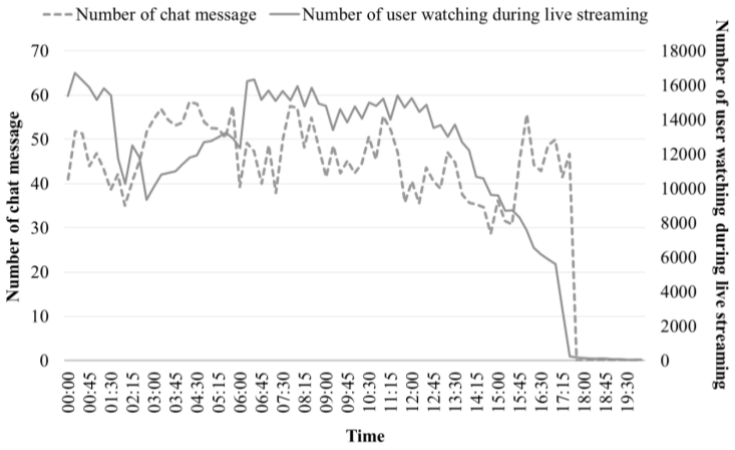

Figure 1. The number of chat messages and the number of users watching during live streaming.

3.2. Feature analysis

In this section, the differences between normal users and spammers were analyzed by considering multiple features to find the features that archive the best classification efficiency and low runtime overhead. Chat message characteristics are commonly used as features in many research studies: word count per chat message [14], relevance score using bag-of-words [15], polarity score, and similarity of chat message [16]. In

addition to the chat message characteristics, we proposed additional behavior-based features which include the duration that a user participates in live chat and the interarrival time or the time between two consecutive chat messages from the same user.

3.2.1. Number of chat messages per user

During the period of live streaming, there were a total of 60,115 chat messages from 11,554 users. According to our analysis, the majority of the users (46%) left the chat after leaving only one comment.

Figure 2 shows the number of chat messages of the top 20 users with the highest number of comments. By investigating their comments, we could separate these users into 3 types:

- Chat bot: Chatbot users who monitored the channel and answered common questions.
- Stop-by users: Users who left just one message. In this case, it was not possible to define the duration that a user participated in the live chat.
- Engaging users: Users who joined the live chat and engaged in interactive communication with others about the video topic or non-relevant topic. This type of user may leave spam or an advertising message.

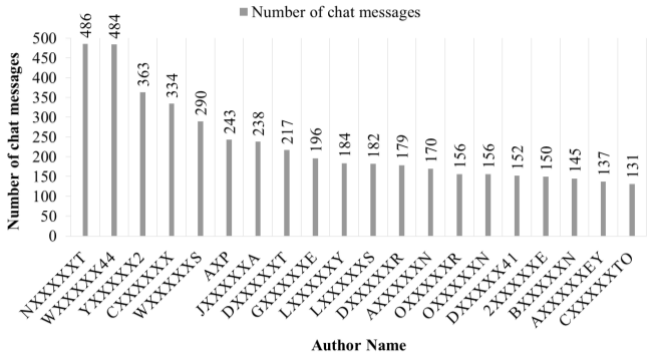

Figure 2. Top 20 users with the highest number of chat message

3.2.2. Duration of time that each user participates in the live chat

The time duration, as is displayed in Figure 3, is the amount of time a user participates in live chat. Let $T_{k,t}$ be the time that a user k leaves the first comment in the live chat and $T_{k,t+\Delta}$ be the time of user k's last comment, then time duration (TD_k) can be defined as

$$TD_k = T_{k,t+\Delta} - T_{k,t} \tag{1}$$

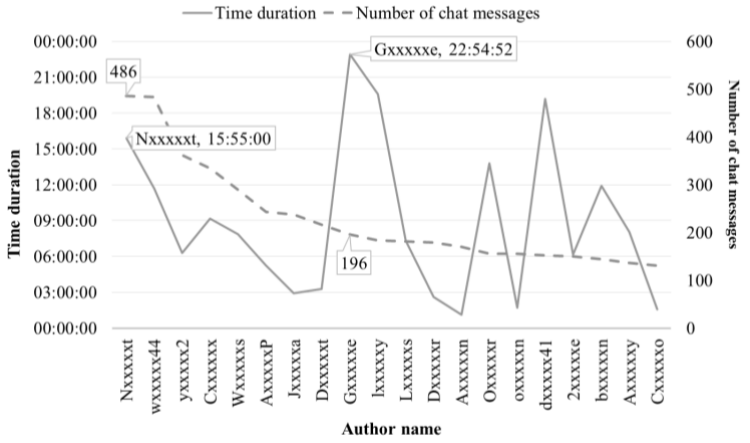

Figure 3. Time duration and number of chat messages of top-20 users during a live chat

Figure 3 depicts the duration of time the top 20 users who have the highest number of chat messages spent on the live chat. As can be observed, the amount of time users spent on the live chat does not correlate with the number of their comments. Hence, this feature, by itself, cannot be used to indicate the interest that a user has in the streaming content because it does not imply that a user remains in the live chat throughout the entire period: a user may have logged out and logged back in later to provide comments.

3.2.3. Word count per chat message

Word count is the number of words contained in each message. Figure 4 displays the number of chat messages divided by word count. The average word count per message from the collected data was 8 words with 7.10 standard deviation. Our result indicated that 63% of the messages contained fewer words than the average word count. Also, word count could not be used to identify relevancy between chat messages and streaming topics.

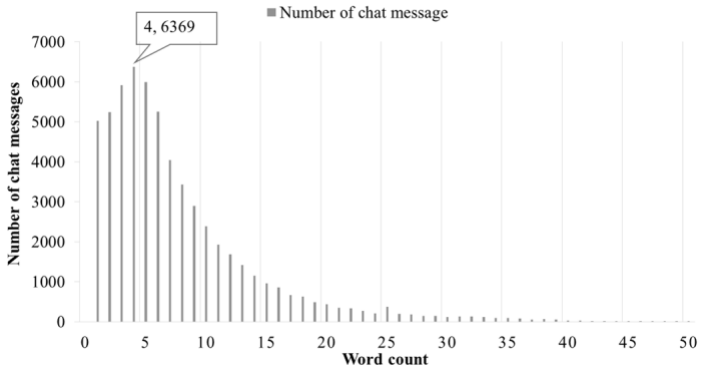

Figure 4. Number of chat messages divided by word count

3.2.4. Relevance score using bag-of-words

In this study, we assumed that relevant chat messages would likely contain some content that is related to the topic and the content of the live stream. Hence, related words, for example, hurricane, Irma, storm, surge, and Florida, were extracted from the live streaming content. The frequency of occurrence of such words found in each chat message was used as users' relevance score (*SR*).

Let *W* be a set of relevant word (bag-of-word), and *M* be a chat message which consists of a set of tokenized word $T = \{T_1, T_2, T_3, \ldots, T_N\}$, the relevant score of each chat message *j* can be defined as

$$SR_j = \sum_{i=1}^{N} T_i^I \tag{2}$$

Where

$$T_i^I = \begin{cases} 1 & if\ T_i \epsilon\ W \\ 0 & else \end{cases}$$

For this study, we could classify users into 3 types as shown in Table 1.

Table 1. Type of users categorized by relevance score

Type	Description	Number of user
0	Users whose chat messages were not related to the topic of the live stream.	6,551
1	Users who communicated with other users by specifically mentioning their user names with '@'.	203
2	Users who communicated with others and created chat messages related to the topic.	4,800

The results showed that a large number of users (Type 0) posted irrelevant chat messages. However, not all messages from Type 0 users could be considered as spam. This is due to the fact that a certain amount of messages were part of the on-going conversation which may or may not be related to the streaming

content. Messages from these users were similar to that of Type 1 users but without mentioning the recipient of the message. Therefore, the relevance score alone cannot be used for classifying spammers.

3.2.5. Interarrival times

Interarrival time is the time between arrivals of chat messages from the same user on live chat. Let $t_{k,n}$ be an arrival time of chat message n from user k at time t and $t_{k,n-1}$ be an arrival time of previous chat message from user k, then the interarrival time of message n from user k ($IT_{k,n}$) can be defined as

$$IT_{k,n} = t_{k,n} - t_{k,n-1} \tag{3}$$

Figure 5 shows the probability mass function (PMF) of interarrival time between consecutive chat messages for each user type excluding stop-by users or ones that leave only one message on the live chat. The number of messages from these stop-by users amounted to 45% of the total number of chat messages. For this study, the focus was on Type 2 users who actively engaged in the conversation. According to the result, the message interarrival time of type 2 users tended to be longer than that of the other types, i.e., the interarrival time of 55% of the messages was on the order of minutes to several minutes. Furthermore, interarrival time of 19% of the messages is longer than one hour, which indicated that users might have left the live stream and returned later to rejoin the chat.

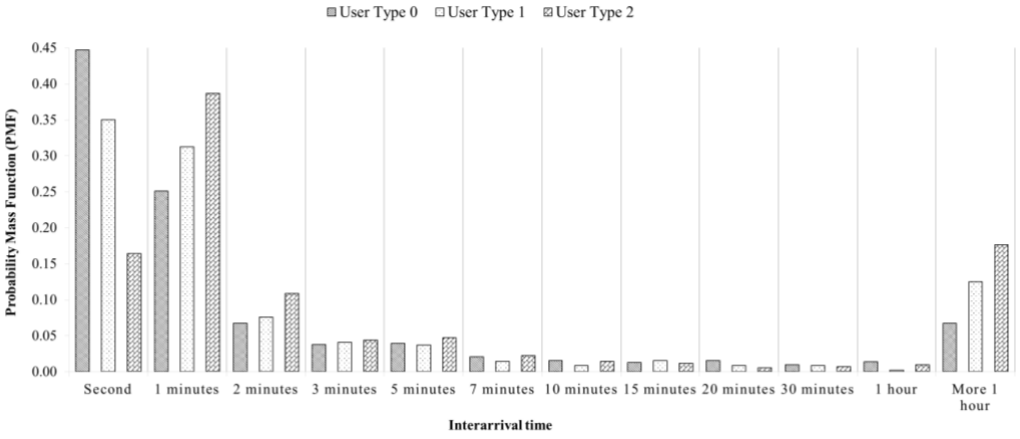

Figure 5. Probability mass function (PMF) of the message interarrival time

3.2.6. The similarity of chat messages

The similarity of chat message (*Sim*) is an estimate of the degree of similarity between two chat messages from the same user. In this study, we used the result of the calculation of cosine similarity [17] to represent the degree of message similarity. Cosine similarity [18-19] is the traditional method used to measure the degree of similarity between two vectors, obtained from the cosine angle multiplication. The Cosine similarity [8] can be calculated using term frequency and inverse document frequency (TF-IDF) formulas. A result is a number that ranges from 0 to 1. The two vectors have no similarities when the value is 0, while the value of 1 indicates that they are identical. Let $M_{k,t}$ be a message at time t from user k and $M_{k,t+\Delta}$ be the next message from user k at time $t+\Delta$. Let the features of the two messages be denoted by the vectors $x(M_{k,t})$ and $y(M_{k,t+\Delta})$, where $x(M_{k,t}) = \{x_1, x_2, \ldots, x_n\}$ and $y(M_{k,t+\Delta}) = \{y_1, y_2, \ldots, y_n\}$. The cosine similarity between the two vectors can be defined as

$$Sim(x(M_{k,t}), y(M_{k,t+\Delta})) = \frac{\sum_{i=1}^{n} x_i^2 * y_i^2}{\sqrt{\sum_{i=1}^{n} x_i^2 * \sum_{i=1}^{n} y_i^2}} \tag{4}$$

Be assuming that a spam user creates multiple chat messages with the same or similar content, we can use this method to classify spams by marking the messages with a cosine similarity of greater than 0.8 as spam.

Table 2 shows the result of the classification using the similarity of chat messages. Only 1,803 messages or 3% out of 60,115 chat messages were identified as spam using this method and 19% of the messages cannot be classified as either spam or ham (normal). This is because the similarity score is based on similarity of messages from the same user. Therefore, only spammers who posted multiple similar messages will be detected.

Table 2. Classification result using the similarity of chat messages

Type of chat message	Number of chat message
Spam chat messages	1,803 (3%)
Normal chat message	46,890 (78%)
Chat messages that cannot be classified (because users only left one message during the live streaming)	11,422 (19%)

According to our observation, this is not an efficient way to classify spam messages because many spammers usually alter messages slightly or completely when posting to prevent them from being detected. Besides, spammers may also disguised themselves by using multiple usernames. Therefore, this method, by itself, can be quite ineffective in such cases.

3.2.7. Polarity score

SentiStrength [20], SentiWordNet and AFINN are sentiment lexicon classification tools that are often used to estimate the strength of positive and negative sentiment of words. These methods can be applied to identify the polarity of the comments as being positive, negative, or neutral [17, 21].

Figure 6 displays the polarity of the messages as being positive, negative, or neutral. Because the streaming content in this study was related to hurricane Irma, many people participated in the live chat to express their feeling or concern. While one might anticipate that the streaming about the devastating hurricane to contain primarily negative comments, there were quite many encouraging messages from the people who did not face the situation directly. These people sent out encouraging messages to the victims. Hence, the topic of the streaming content in this case strongly affected the polarity of the message.

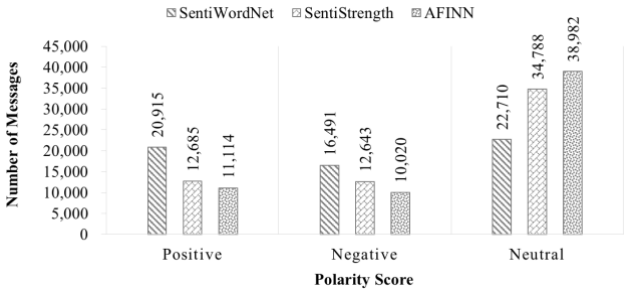

Figure 6. The polarity of messages grouped by detection methods

4. CLASSIFICATION METHOD AND RESULT

Based on the live chat dataset and feature selection described in the previous section, we created a training dataset that consisted of 7 features: interarrival time (C1), polarity score (C2), duration of times (C3), number of chat message per user (C4), word count (M1), relevant score (M2), and similarity score (M3). Each message was manually labeled as being *relevant* or *irrelevant*. In the experiment, evaluation of the classification of user behavior was analyzed with 10-fold cross-validation using 6 different types of classifiers [1] that is naïve bayes [15, 22-24], decision trees [15, 24], random forest [15, 25], k-nearest neighbors [26], support vector machine (SVM) [27-28], and artificial neural network (ANN) [29]. The result of this classification was analyzed to find suitable features for identifying spam during live streaming.

Correlation is a popular technique that is often used for finding the association between the two features. Table 3 shows the correlation between each feature and the actual result (spam or ham). According to the result, none of the features was highly correlated with the actual result. The relevant score (M2) exhibited the highest correlation value of 0.25 and could correctly classify 60.03% of the messages on average (averaging over the 6 classification techniques). The overall accuracy of a single feature classification approach was only 54.93% averaging over different features and classification techniques.

These results indicated that none of the features could effectively classify spam messages. Therefore, in order to increase the efficiency of classification we had to find a set of features that results in the highest accuracy.

Table 3. Single feature analysis

No	Feature	Correlation	Average Classification Accuracy
1	Word count (M1)	0.124992	56.28%
2	Relevant score (M2)	0.256252	60.03%
3	Similarity score (M3)	-0.009899	52.11%
4	Interarrival time (C1)	0.100143	54.97%
5	Polarity score (C2)	-0.014852	53.18%
6	Time duration (C3)	0.131724	55.07%
7	No of message (C4)	0.140127	53.37%

Figure 7 depicts the top 5 of the classification accuracy of the 2-feature approach. The result showed that the naïve bay yielded the highest accuracy (66.74%) when using the relevant score (M2) and time duration (C3) as classification features. Interestingly, we observed that when M2 is combined with other features, the classification accuracy seemed to be higher when compared with the other set of features without M2. In addition, the 2-feature approach increased the accuracy by approximately 9% from that of the single-feature approach.

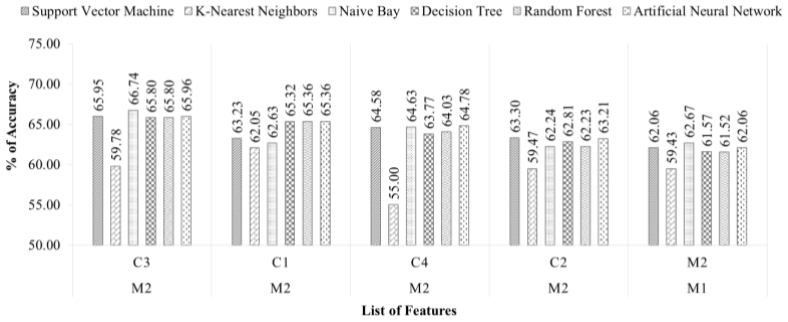

Figure 7. Top 5 classification results using 2 features

As we increased the number of features from 3-7, we observed similar results, i.e., dominant features tend to improve the overall performance of the classifier. As shown in Figures 8-12, M2 and C3 appeared in the set that yielded the highest accuracy. When using 3 or 4 features, ANN performed slightly better than the other classification techniques, followed by SVM and decision tree. However, it can be seen from Figures 10-12 that adding more features did not seem to significantly improve the classification accuracy. In fact, techniques other than SVM seemed to perform worse when using more than 3 features.

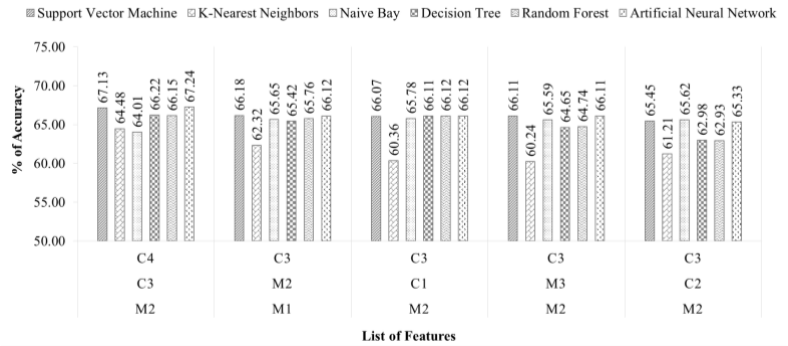

Figure 8. Top 5 classification results using 3 features

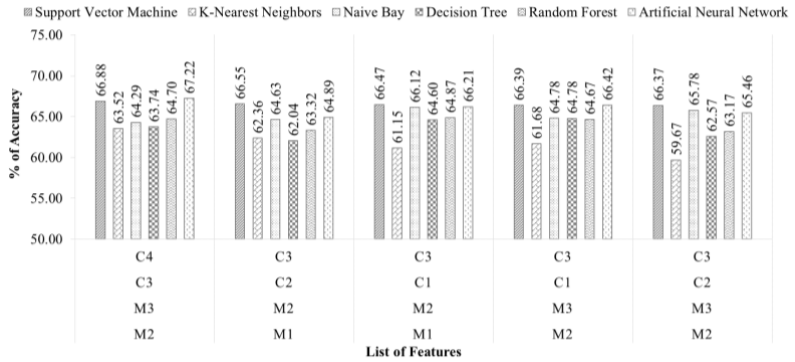

Figure 9. Top 5 classification results using 4 features

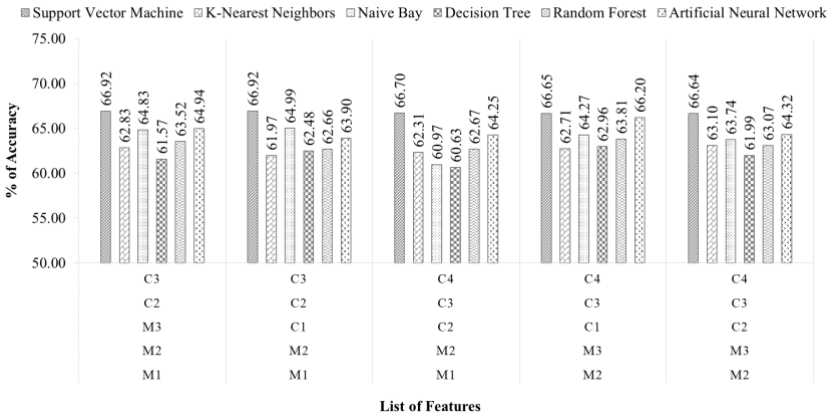

Figure 10. Top 5 classification results using 5 features

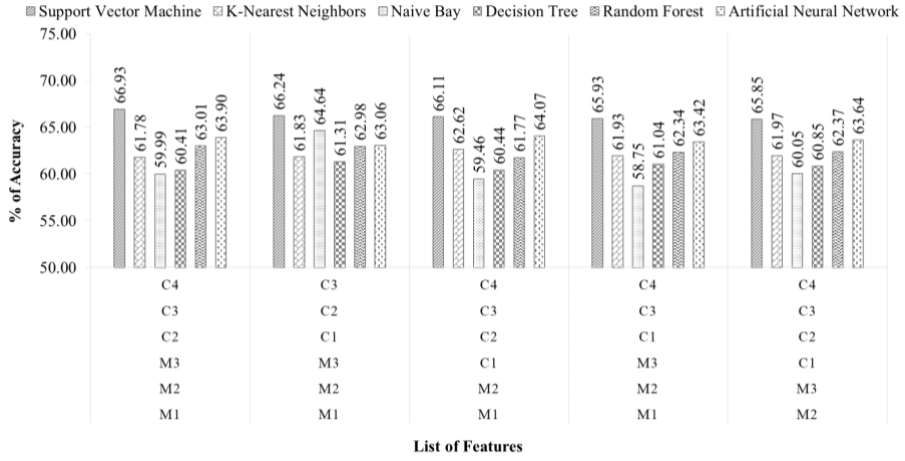

Figure 11. Top 5 classification results using 6 features

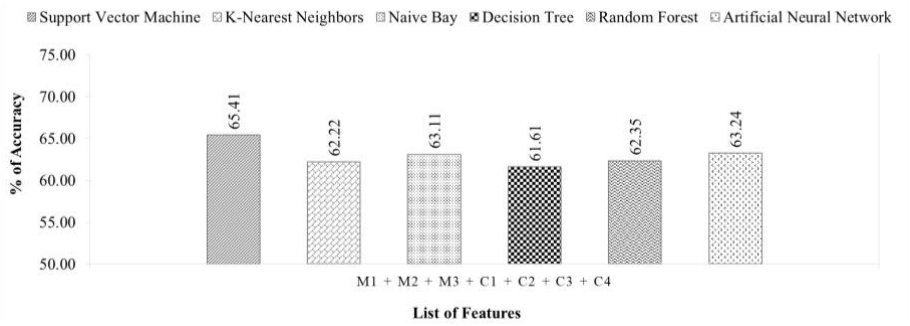

Figure 12. Classification result of combination 7 features

According to Figure 13, it can be observed that the ANN and SVM had similar performance. More specifically, the accuracy of these two techniques dropped significantly when using 5-7 features. ANN performed slightly worse than SVM when the number of features is more than 5. All techniques performed equally well when using M2, C3, and C4 as features. Therefore, these three features seem to be an optimal choice for classifying spam messages.

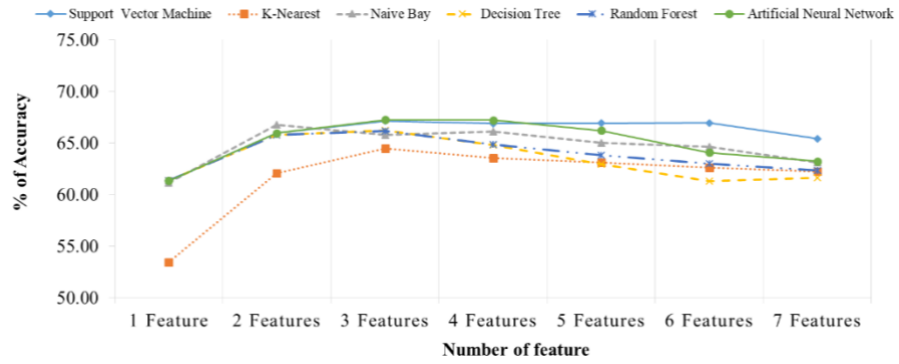

Figure 13. Highest classification result of each approach.

5. DISCUSSION

In the real-time streaming scenario where spams need to be filtered out quickly, the processing time is an important factor. Processing time, in this case, is composed of the time used for extracting the required features from the incoming chat message and the time that the classification engine needs to process and output the classification results based on the given set of features. The feature extraction time depends on the complexity of the algorithms used for deriving the feature. For example, it takes at most *O(n)* to find relevant keywords in chat messages and compute the relevant score (M2). To find the time duration (C3) a user spent in a live chat and the number of chat messages (C4), it takes only *O(d)*. Note that in practice, C3 is the time elapsed between the first chat message and the current chat message of a user and C4 is an accumulated number of chat messages up to the current one. Each classification technique considered in this study also has a different level of complexity. Table 4 shows that ANN is the most time-consuming technique and Decision Tree consumes the least amount of time when compared to the other methods. Hence, Decision Tree seems to be the best choice for implementation as it has comparable performance to ANN and SVM in terms of accuracy using 3 features and it runs much faster than any other classification techniques.

Table 4. Processing time (ms) of each classification technique

Feature	Support Vector Machine	K-Nearest	Naive Bay	Decision Tree	Random Forest	Artificial Neural Network
1 Feature	12263	1591	85	12	588	6995
2 Features	12755	992	99	10	785	10258
3 Features	13461	897	106	18	1058	34362
4 Features	13767	908	106	29	1456	119366
5 Features	14464	990	109	40	1664	137092
6 Features	15161	1138	111	54	1806	137211
7 Features	15960	1287	103	43	1637	138227

6. CONCLUSION

The objective of this study was to investigate and analyze YouTube viewers' behavior on live chat. The experiment considered users' behavior and chat message characteristics for classifying spammer. The features considered in this study include users' behavior characteristics (interarrival time or the time between two consecutive chat messages from the same user, polarity score, the duration of time a user spent on a live chat, and the number of chat message per user) and chat message characteristics (word count, relevant score, and similarity score). Our study shows that each feature cannot efficiently be used to classify spams. The best combination of features that are suitable for the real-time spam filtering purpose is the relevant score, the time the user spent in the live chat, and the number of the user's chat messages. Given these three features, the decision tree seems to be the best choice for spam filtering because it runs much faster than the other techniques and has quite a very high accuracy. Note that the results presented in this study are subjective to the topic of the live streaming which is about Hurricane Irma. In the future, we plan to further generalize our approach to cover multiple types of streaming content.

REFERENCES

[1] N. Alias, C. Feresa, M. Foozy, and S. N. Ramli, "Video spam comment features selection using machine learning techniques," *International Journal of Electrical and Computer Science*, vol. 15, no. 2, pp. 1046-1053, 2019. doi: 10.11591/ijeecs.v15.i2.pp1046-1053.

[2] L. Wu, B. Brandt, X. Du and Bo Ji, "Analysis of clickjacking attacks and an effective defense scheme for Android devices," *2016 IEEE Conference on Communications and Network Security (CNS)*, Philadelphia, PA, 2016, pp. 55-63. doi: 10.1109/CNS.2016.7860470.

[3] O. Çıtlak, M. Dörterler, and İ. A. Doğru, "A survey on detecting spam accounts on Twitter network," *Soc. Netw. Anal. Min.*, vol. 9, 2019. doi: 10.1007/s13278-019-0582-x.

[4] C. Rădulescu, M. Dinsoreanu and R. Potolea, "Identification of spam comments using natural language processing techniques," *2014 IEEE 10th International Conference on Intelligent Computer Communication and Processing (ICCP)*, Cluj Napoca, 2014, pp. 29-35. doi: 10.1109/ICCP.2014.6936976.

[5] J. P. Carpenter, K. B. Staudt Willet, M. J. Koehler, and S. P. Greenhalgh, "Spam and Educators' Twitter Use: Methodological Challenges and Considerations," *TechTrends*, vol. 64, no. 3, pp. 460-469, 2020. doi: 10.1007/s11528-019-00466-3

[6] J. P. Carpenter, M. J. Koehler, K. B. S. Willet, and S. P. Greenhalgh, "Spam, Spam, Spam, Spam: Methodological Considerations and Challenges for Studying Educators ' Twitter Use," *Proceedings of Society for Information Technology & Teacher Education International Conference*, pp. 2702-2711, 2019.

[7] N. Jindal and B. Liu, "Analyzing and Detecting Review Spam," *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, Omaha, NE, 2007, pp. 547-552. doi: 10.1109/ICDM.2007.68.

[8] M. Alsaleh, A. Alarifi, F. Al-Quayed and A. Al-Salman, "Combating Comment Spam with Machine Learning Approaches," *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, Miami, FL, 2015, pp. 295-300. doi: 10.1109/ICMLA.2015.192.

[9] F. Benevenuto, T. Rodrigues, A. Veloso, J. Almeida, M. Goncalves and V. Almeida, "Practical Detection of Spammers and Content Promoters in Online Video Sharing Systems," in *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 3, pp. 688-701, June 2012. doi: 10.1109/TSMCB.2011.2173799.

[10] T. C. Alberto, J. V. Lochter and T. A. Almeida, "TubeSpam: Comment Spam Filtering on YouTube," *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, Miami, FL, 2015, pp. 138-143. doi: 10.1109/ICMLA.2015.37.

[11] S. B. Rathod and T. M. Pattewar, "Content based spam detection in email using Bayesian classifier," *2015 International Conference on Communications and Signal Processing (ICCSP)*, Melmaruvathur, 2015, pp. 1257-1261. doi: 10.1109/ICCSP.2015.7322709.

[12] M. Verma, D. Divya, and S. Sofat, "Techniques to Detect Spammers in Twitter- A Survey," *Int. J. Comput. Appl.*, vol. 85, no. 10, pp. 27-32, 2014. doi: 10.5120/14877-3279.

[13] N. Eshraqi, M. Jalali and M. H. Moattar, "Spam detection in social networks: A review," *2015 International Congress on Technology, Communication and Knowledge (ICTCK)*, Mashhad, 2015, pp. 148-152. doi: 10.1109/ICTCK.2015.7582661.

[14] A. Serbanoiu and T. Rebedea, "Relevance-Based Ranking of Video Comments on YouTube," *2013 19th International Conference on Control Systems and Computer Science*, Bucharest, 2013, pp. 225-231. doi:

10.1109/CSCS.2013.87.
[15] W. Etaiwi and A. Awajan, "The Effects of Features Selection Methods on Spam Review Detection Performance," *2017 International Conference on New Trends in Computing Sciences (ICTCS)*, Amman, 2017, pp. 116-120. doi: 10.1109/ICTCS.2017.50.
[16] C. Kale, D. Jadhav, and T. Pawar, "Spam Review Detection Using Natural Language Processing Techniques," vol. 3, no. 1, pp. 1-6, 2016.
[17] H. A. Santoso, E. H. Rachmawanto, A. Nugraha, A. A. Nugroho, D. R. I. M. Setiadi, and R. S. Basuki, "Hoax classification and sentiment analysis of Indonesian news using Naive Bayes optimization," *Telkomnika (Telecommunication Comput. Electron. Control)*, vol. 18, no. 2, pp. 799-806, 2020. DOI: 10.12928/TELKOMNIKA.v18i2.14744.
[18] W. H.Gomaa and A. A. Fahmy, "A Survey of Text Similarity Approaches," *Int. J. Comput. Appl.*, vol. 68, no. 13, pp. 13-18, 2013. doi: 10.5120/11638-7118.
[19] A. K. Nikhath and K. Subrahmanyam, "Feature selection, optimization and clustering strategies of text documents," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 2, pp. 1313-1320, 2019. DOI: 10.11591/ijece.v9i2.pp1313-1320.
[20] H. Bhuiyan, J. Ara, R. Bardhan and M. R. Islam, "Retrieving YouTube video by sentiment analysis on user comment," *2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, Kuching, 2017, pp. 474-478. doi: 10.1109/ICSIPA.2017.8120658.
[21] G. Suciu, A. Pasat, T. Uşurelu, and E.-C. Popovici, "Social Media Cloud Contact Center Using Chatbots," International Conference on Future Access Enablers of Ubiquitous and Intelligent Infrastructures, FABULOUS 2019: Future Access Enablers for Ubiquitous and Intelligent Infrastructures, pp. 437-442, 2019. doi: 10.1007/978-3-030-23976-3_39
[22] Y. Vernanda, M. B. Kristanda, and S. Hansun, "Indonesian language email spam detection using n-gram and naïve bayes algorithm," *Bull. Electr. Eng. Informatics*, vol. 9, no. 5, pp. 2012-2019, 2020. doi: 10.11591/eei.v9i5.2444.
[23] N. Ain, M. Samsudin, C. Feresa, N. Alias, and P. Shamala, "Youtube spam detection framework using naïve bayes and logistic regression," vol. 14, no. 3, pp. 1508-1517, 2019. DOI: 10.11591/ijeecs.v14.i3.pp1508-1517.
[24] S. Shrivastava and R. Anju, "Spam mail detection through data mining techniques," *ICCT 2017 - Int. Conf. Intell. Commun. Comput. Tech.*, vol. 2018-January, pp. 61-64, 2018. DOI: 10.1109/INTELCCT.2017.8324021.
[25] S. Aiyar and N. P. Shetty, "N-Gram Assisted Youtube Spam Comment Detection," *Procedia Comput. Sci.*, vol. 132, pp. 174-182, 2018. doi: 10.1016/j.procs.2018.05.181.
[26] A. M. Al-Zoubi, J. Alqatawna, and H. Faris, "Spam profile detection in social networks based on public features," in *2017 8th International Conference on Information and Communication Systems (ICICS)*, pp. 130-135), 2017. doi: 10.1109/IACS.2017.7921959.
[27] K. Zainal and Z. Jali, "An Analysis of Various Algorithms For Text Spam Classification and Clustering Using RapidMiner and Weka," *International Journal of Computer Science and Information Security*, vol. 13, no. 3, 2015.
[28] S. C. Hui, Y. He and Haichao Dong, "Text mining for chat message analysis," *2008 IEEE Conference on Cybernetics and Intelligent Systems*, Chengdu, 2008, pp. 411-416. doi: 10.1109/ICCIS.2008.4670827.
[29] D. Puniškis, R. Laurutis, and R. Dirmeikis, "An artificial neural nets for spam e-mail recognition," *Electron. Electr. Eng.*, vol. 5, no. 5, pp. 1215-1392, 2006.

BIOGRAPHIES OF AUTHORS

Sawita Yousukkee received B.S. degree in Computer Science from Burapha University in 2002 and M.S. degree in Information Technology from King Mongkut's Institute of Technology Ladkrabang in 2005, respectively. Since March 2015, she is with the Faculty of Information Technology from King Mongkut's University of Technology North Bangkok as a Ph.D. candidate. She is currently a lecturer in Phranakhon Si Ayutthaya Rajabhat University, Thailand. Her research interests include data mining, machine learning, and deep learning.

Nawaporn Wisitpongpun received her B.S., M.S., and Ph.D.degrees in electrical and computer engineering from Carnegie Mellon University in 2000, 2002, and 2008, respectively. From 2003 to 2009 she was also a research associate in the Electrical and Control Integration Laboratory, General Motors Corporation. Presently, She is an assistant to the president for research and information technology and a lecturer in the Faculty of Information Technology at King Mongkut's University of Technology North Bangkok, Thailand. Her research interests include traffic modeling, chaos in the Internet, and cross-layer network protocol design for wireless networks, social network analysis, and digital government.