# YouTube Spam Comment Detection Using Support Vector Machine and K–Nearest Neighbor

**Aqliima Aziz[1], Cik Feresa Mohd Foozy[2], Palaniappan Shamala[3], Zurinah Suradi[4]**
[2,3]Applied Computing Technology (ACT), Universiti Tun Hussein Onn Malaysia (UTHM),
Parit Raja, Batu Pahat, 86400 Johor, Malaysia.
[1,2,3]Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia (UTHM),
Parit Raja, Batu Pahat, 86400 Johor, Malaysia.
[4]Department of Management Information Systems, College of Commerce and Business Administrations,
Dhofar University, Salalah, Oman.

**Article Info**

***Article history:***

Received Apr 9, 2018
Revised May 20, 2018
Accepted Jul 11, 2018

***Keywords:***

Classification
Detection
Machine Learning
YouTube Spam

**ABSTRACT**

Social networking such as YouTube, Facebook and others are very popular nowadays. The best thing about YouTube is user can subscribe also giving opinion on the comment section. However, this attract the spammer by spamming the comments on that videos. Thus, this study develop a YouTube detection framework by using Support Vector Machine (SVM) and K-Nearest Neighbor (k-NN). There are five (5) phases involved in this research such as Data Collection, Pre-processing, Feature Selection, Classification and Detection. The experiments is done by using Weka and RapidMiner. The accuracy result of SVM and KNN by using both machine learning tools show good accuracy result. Others solution to avoid spam attack is trying not to click the link on comments to avoid any problems.

***Corresponding Author:***

Aqliima Aziz,
Faculty of Computer Science and Information,
Universiti Tun Hussein Onn Malaysia (UTHM),
Parit Raja, Batu Pahat, 86400 Johor, Malaysia.
Email: aqlimaaziz@gmail.com

## 1. INTRODUCTION

YouTube is one of the biggest site for user get information on the Internet [1]. Because of that, many spammers will trick the YouTube user by spamming the YouTube comments. According to Hamou [2], spam is now a trend attack and the YouTube defines spam as inappropriate comments, such as abuse or trolling and also people trying to sell things. Ham can be defined as “good comments” or YouTube free from spam comment.

Spam can be categorized as dangerous because spam has the potential of cyber security threat for end users. The spammer used this opportunity to spread malware through comment fields, which will exploit vulnerabilities in the user’s machines. Another intention includes seizing money transactions and hijacking credit card and banking information. Besides, spammer tends to ruin the content of web pages. This action will lead visitors to annoy overall of the posted content [3].

YouTube spam comments has potential to spread malware. The WannaCry issue is a representative example of malware used by spammer to exploit user’s vulnerabilities. Next, fileless malware attacks are being applied by attackers and cybercriminals. This attack might prevent detection and make difficult for forensic investigations. Usually, spammers making use of existing tools that already installed on users’ computers. For example, PowerShell, PSExec, WMI or running simple scripts and shellcode straight in memory. Fileless means creating a few files on hard disk, which less chance of being traced. Next, wipers are type malware that used by spammers for removing tracks after cyberespionage occurs [4]. Moreover, malware cause multiple breaches where leak millions of user records. An example of the details leaked such as usernames, email addresses and hashed passwords, probably use SHA-1 which is less secure. The main factor that leads to data breach is possibly weak password which can easily crack [5].

There are several studies to detect YouTube Spam such as [6]-[9] proposed to classify the YouTube comment as Spam and Ham by using Support Vector Machine (SVM). The significance of this research to develop a YouTube Spam detection framework and YouTube Spam features so the YouTube visitor able to identify the YouTube Spam characteristics. When the YouTube users able to identify the YouTube spam features, they will be more aware, and the malware spread can be reduced.

## 2. LITERATURE REVIEW

There are many types of spam, such as web spam, short message spam, email spam, social network spam and others. In this section, the YouTube spam detection studies will be focused.

### 2.1 Spam Detection Approach

YouTube is not excluded from malicious user who are often found to expose in spamming and promotional activities [10]. There are many approaches to detect Spam such as using Artificial Intelligent, Cryptography, Machine Learning and others. However, Manwar [7] said the machine learning also capable to detect YouTube spam.

The existing study in YouTube Spam Detection is Manwar [7] and Alberto [8] show that both of the authors used Support Vector Machine (SVM) as a classifier in classification phase. Manwar [7] stated that SVM classification is in binary-two class. Usually, class denoted by 0 and 1. However, the collection data have been classified into two classes. Hence, easy for pre-processing and feature selection to perform.

### 2.2 Framework detection

Basically, there are several phases in detection framework using machine learning techniques such as Data Collection, Feature Selection, Classification and Detection.

The Data Collection are collected from social media. For example, Facebook, Twitter, Sina Weibo (Instagram), YouTube and Email. Thus, UCI will collect those comments and form a dataset according to social media categories such as YouTube, Facebook and others. Figure below shows raw data collected from UCI machine learning repository. Next, identify whether the comment is spam or ham. Based on Table 1, raw data already classified in spam and ham.

Table 1. Datasets for YouTube spam comment [8]

| Datasets | YouTube ID | Spam | Ham | Total |
|---|---|---|---|---|
| Psy | 9bZkp7q19f0 | 175 | 175 | 350 |
| KatyPerry | CevxZvSJLk8 | 175 | 175 | 350 |
| LMFAO | KQ6zr6kCPj8 | 236 | 202 | 438 |
| Eminem | uelHwf8o7_U | 245 | 203 | 448 |
| Shakira | pRpeEdMmmQ0 | 174 | 196 | 370 |

In machine learning, feature selection is used to classify the class. Several studies by Afzal [11], used the URL as features. Wu, F., & Huang, Y. [12] used content- based features to detect the spam comments in the user's message. The features are URLs, keywords, hashtags and bad comments. Meanwhile other studies applied other types of features.

Classification will be used to classify the dataset into several classes based on the suitable features. According to [13], SVM is one of the techniques that can classify the problems [13]. Meanwhile, K-NN is a simple yet efficient classification algorithms for data mining [14].

Lastly, the results obtain. The purpose of this research is to compare which techniques provide better accuracy result in detecting the YouTube spam comment.

## 3. METHODOLOGY

There are (5) steps in this detection framework such as Data Collection, Data Pre-processing, Feature Extraction, Classification and Comparison of Results, as shown in Figure 1. This framework is chosen from [6] because it can provide the result with good accuracy. This framework also provides the phase to compare the results of SVM technique and k-NN technique.

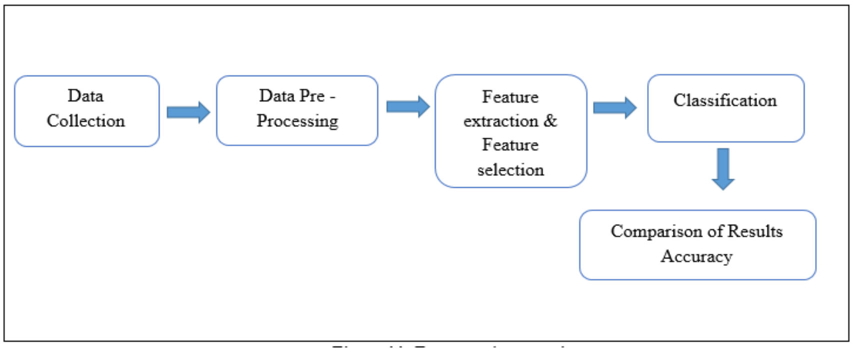
Data Collection
Data Pre - Processing
Feature extraction & Feature selection
Classification
Comparison of Results Accuracy

Figure 1. YouTube Spam Detection Framework [6]

The description for every phase in YouTube Spam detection framework:

a) Data Collection

In this phase, the dataset for experiments is downloaded from UCI machine learning repository. The dataset contained of five (5) selected videos and were downloaded from YouTube through API [8]. The comments are from PSY, KatyPerry, LMFAO, Eminem and Shakira. The total for spam and ham in Psy video is 350, followed by Katy Perry is 350, LMFAO is 438, Eminem is 448 and Shakira is 370.

b) Pre – processing

For Pre-processing phase, the raw dataset will be executed the data cleaning such as tokenization, stopwords removal and stemming are performed. The clean dataset will be used for next process of feature selection and extraction.

c) Feature Selection and Extraction

Feature selection is a process before classification class. The suitable features will be identified based on the dataset.

d) Classification

There is training and testing process in this phase. 60% will be used for training and 40% for testing. After completing the step iii, supposed to be there is features that is considered as spam. Thus, the dataset needs to train based on machine learning techniques.

SVM is successfully suitable in differentiating positive and negative problem such as spam. SVM is a supervised learning model that analyzes data used for classification and regression. SVM mostly used in classification problems. SVM is used for binary classification problem and used kernel functions.

K-NN is a supervised learning method. Data is appearing in a vector space in the K-NN algorithm. K–NN emphasize *k* most similar training data points to a testing data point. After determining the K-Nearest Neighbors, the algorithm will combines the neighbors' to decide the label of testing data point. For implementation, labels are combined as the labels used simple majority vote.

e) Comparison of results

The result performance will be used Accuracy, Precision, Recall and F-measure.

Precision = True Positive/(False Positive + True Positive) (3.1)

Recall = True Positive/ (False Positive + True Positive) (3.2)

F-measure = 2*Recall*Precision / Recall + Precision (3.3)

Accuracy = (True Negative + True Positive) / (False Positive + True Positive + False Negative + True Negative) (3.4)

## 4. RESULTS AND ANALYSIS

There are several results that discussed in this section such as Data Collection, Pre-Processing, Feature Selection, Classification and Detection Result.

### 4.1 Results in Data Collection

In order to collect raw data, this research uses UCI machine learning repository. Those data already classes in attribute such as users with an account on YouTube when importing into Excel before going through pre-processing. This data collection contains 1005 spam and 935 ham (legitimate) comments [15]. After that, the data need to change file type first which is .txt before testing on RapidMiner and Weka.

### 4.2 Results Pre- Processing

Once all the data is collected, the pre-processing step is performed. Once data.txt is inserted into the tool, the tool will be performed tokenization, stopwords removal and stemming phase. Tokenization separates the string block by block. Thus, the tokenization makes the process of stopwords removal become easy. Stopwords eliminate the commonly used word such as “a”, “an”, “the” and numbers in the sentences. The purpose of stopwords removal is to shorten the pre-processing time and avoid those words taking space in the database. Next, stemming purpose is to get the root word used in query, avoid from having equal meaning and become incomplete sentences. For example, the words subscribe and please respectively become “subscrib” and “pleas”. Hence, the obtained data after pre-processing is cleaned. Next, the process data need to extract in excel to facilitate next step which features extraction ad feature selection [16].

### 4.3 Results in Feature Extraction and Feature Selection

As features identified from the literature review, various features may be extracted from YouTube classification purposes. Besides, the data already consists of two (2) classes where the classes are “spam” and “ham” [15]. Thus, easy to choose features that certainly label as spam. YouTube comments may contain hyperlinks, text, uppercase and lowercase characters. However, those uppercase characters do not exist after pre-processing phase. After pre-processing, this study decides to use keywords as feature selection. The aim for feature extraction is to explore the advantages of new features in order to gain high accuracy.

### 4.4 Classification Result

In classification, a total of seven (7) algorithms implemented in RapidMiner and Weka were set as classifiers in detecting YouTube spam comments. The purpose of implemented the 7 of algorithms are to compare the accuracy. The classifiers were fed and tested by the same datasets in classifying YouTube spam but 6 different algorithms [16].

The classification of accuracy across seven (7) different classification algorithms such as Naïve Bayes, Decision Tree, Support Vector Machine, Random Tree, Random Forest, k-Nearest Neighbor and Logistic using data proportion of 70:30. 70:30 means 70% for training and 30% for testing.

Meanwhile, in Weka, Naïve Bayes, Decision Tree, Support Vector Machine, Random Tree, Random Forest and Logistic using data proportion of 60:40. 60:40 means 60% for training and 40% for testing.

### 4.5 Results Detection

The table 2 and 3 below show the experiment with a data proportion of the percentage split training and testing. The result shows that Naïve Bayes classifier gives the highest accuracy when testing in RapidMiner among other classifiers. In general, Naïve Bayes ranks the first, followed by Decision Tree and Logistic. Meanwhile, in Weka, the result shows the accuracy 90% and above. Thus, the results of accuracy as:

Table 2. Classification Accuracy (%) in Weka

| Classifier | Precision | Recall | F-measure | Accuracy |
|---|---|---|---|---|
| Naïve Bayes | 0.928% | 0.928% | 0.928% | 92.78% |
| Decision Tree | 0.922% | 0.920% | 0.920% | 92.01% |
| Logistic | 0.928% | 0.928% | 0.928% | 92.78% |
| Support Vector Machine | 0.918% | 0.915% | 0.915% | 91.49% |
| Random Forest | 0.907% | 0.906% | 0.906% | 90.59% |
| Random Tree | 0.904% | 0.902% | 0.902% | 90.20% |
| k-NN | 0.909% | 0.906% | 0.906% | 90.59% |

Table 3. Classification accuracy (%) in RapidMiner

| Classifier | Accuracy |
|---|---|
| Naïve Bayes | 92.78% |
| Decision Tree | 90.38% |
| Logistic | 88.32% |
| Support Vector Machine | 74.40% |
| Random Forest | 73.54% |
| Random Tree | 52.92% |
| k-Nearest Neighbor | 56.70% |

The goal of this research is to find which algorithms provide high and best in accuracy, precision and recall to help in detecting unwanted comments on YouTube. Besides, the result for this project may as a baseline for people who interested in the YouTube spam comment and improve the results for future comparisons [8].

First and foremost, a dataset of five (5) YouTube comments were collected using public and non-encoded data [8]. This data will be going to test with data mining tool for comparison of result's accuracy, by using different types of algorithms. Indirectly, will prove which algorithms provide the best result and more accurate. Based on observation, spam comments found more than legitimate comments.

For future work, since not all of the algorithms use as classifiers give best accuracy for every single dataset in RapidMiner, this proves that those top three (3) algorithms of Naïve Bayes, Decision Tree and Logistic are more accurate. These top three (3) also gives an accuracy above 80%. However, in Weka those seven (7) algorithms indicate high accuracy. Hence, Weka is recommended to detect YouTube spam comment. Weka provides more accuracy. In addition, hybrid technique for these three (3) algorithms may improve performance in getting high accuracy. Furthermore, the more features used, the higher the percentage of accuracy. Besides, to enhance performance, create new tools, especially for YouTube spam can be made for future research such as TubeSpam. TubeSpam is an example of new tool in detecting spam [8].

## ACKNOWLEDGEMENT

We would like to say thank you to Universiti Tun Hussein Onn Malaysia (UTHM) for kindly proving us with the internal funding (Vot B29000).

## REFERENCES

[1] Scheltus, P., Dorner, V., & Lehner, F. (2013). Leave a Comment! An In-Depth Analysis of User Comments on YouTube. *Wirtschaftsinformatik*, *42*.

[2] Hamou, R. M., Amine, A., & Tahar, M. (2017). The Impact of the Mode of Data Representation for the Result Quality of the Detection and Filtering of Spam. In *Ontologies and Big Data Considerations for Effective Intelligence*(pp. 150-168). IGI Global.

[3] Alsaleh, M., Alarifi, A., Al-Quayed, F., & Al-Salman, A. (2016). Combating comment spam with machine learning approaches. Proceedings - 2015 IEEE 14th International Conference on Machine Learning and Applications, ICMLA 2015, 295–300. https://doi.org/10.1109/ICMLA.2015.192

[4] European Union Agency for Network and Information Security. (2017). ENISA threat landscape report 2017 - EU Law and Publications. https://doi.org/10.2824/967192

[5] Data, G., & Regulation, P. (2018). Fraud & security, (April).

[6] Sah, U. K., & Parmar, N. (2017). An approach for Malicious Spam Detection in Email with comparison of different classifiers.

[7] Manwar, S. R., Lambhate, P., & Patil, J. (2017). Classification Methods for Spam Detection In Online Social Network.

[8] Alberto, T. C., Lochter, J. V., & Almeida, T. A. (2015, December). Tubespam: Comment spam filtering on youtube. In *Machine Learning and Applications (ICMLA), 2015 IEEE 14th International Conference on* (pp. 138-143). IEEE.

[9] Hayati, P., & Potdar, V. (2009, June). Toward spam 2.0: an evaluation of web 2.0 anti-spam methods. In Industrial Informatics, 2009. INDIN 2009. 7th IEEE International Conference on (pp. 875-880). IEEE.

[10] Kaushal, R., Saha, S., Bajaj, P., & Kumaraguru, P. (2016, December). KidsTube: Detection, characterization and analysis of child unsafe content & promoters on YouTube. In *Privacy, Security and Trust (PST), 2016 14th Annual Conference on* (pp. 157-164). IEEE.

[11] Afzal, H., & Mehmood, K. (2016, January). Spam filtering of bi-lingual tweets using machine learning. In *Advanced Communication Technology (ICACT), 2016 18th International Conference on* (pp. 710-714). IEEE.

[12] Wu, F., & Huang, Y. (2017). Social Spammer and Spam Message Detection in an Online Social Network: A Codetection Approach. *Social Network Analysis: Interdisciplinary Approaches and Case Studies*, 225.

[13] Phan, A. V., Le Nguyen, M., & Bui, L. T. (2017). Feature weighting and SVM parameters optimization based on genetic algorithms for classification problems. *Applied Intelligence*, *46*(2), 455-469.

[14] Gursoy, M. E., Inan, A., Nergiz, M. E., & Saygin, Y. (2017). Differentially private nearest neighbor classification. *Data Mining and Knowledge Discovery*, *31*(5), 1544-1575.

[15] Uysal, A. K., Gunal, S., Ergin, S., & Gunal, E. S. (2013). The impact of feature extraction and selection on SMS spam filtering. Elektronika Ir Elektrotechnika, 19(5), 67–72. https://doi.org/10.5755/j01.eee.19.5.1829

[16] Sadoon, O. H., & Yusof, Y. (2017). Detecting Malicious User in YouTube Using Edge Rank Based Feature Set. *International Journal of Soft Computing*, *12*(1), 7-12.