# HYBRID FEATURES FOR DETECTION OF MALICIOUS USER IN YOUTUBE

**OMAR HADEB SADOON**

**MASTER OF SCIENCES (INFORMATION TECHNOLOGY)**
**UNIVERSITI UTARA MALAYSIA**
**2017**

## Permission to Use

In presenting this dissertation in fulfilment of the requirements for a postgraduate degree from University Utara Malaysia, I agree that the University Library may make it freely available for inspection. I further agree that permission for the copying of this dissertation in any manner, in whole or in part, for scholarly purpose may be granted by my supervisor(s) or, in their absence, by the Dean of Awang Had Salleh Graduate School of Arts and Sciences. It is understood that any copying or publication or use of this dissertation or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to University Utara Malaysia for any scholarly use which may be made of any material from my dissertation.

Requests for permission to copy or to make other use of materials in this dissertation, in whole or in part, should be addressed to:

Dean of Awang Had Salleh Graduate School of Arts and Sciences
UUM College of Arts and Sciences
University Utara Malaysia
06010 UUM Sintok

# Abstrak

Media sosial adalah laman web yang menyediakan tempat untuk manusia berhubung. Salah satu contoh media sosial adalah *YouTube*, iaitu satu medium yang menghubungkan manusia melalui perkongsian video. Malangnya, akibat daripada bilangan pengguna komputer yang tinggi dan perkongsian video yang pelbagai, wujud segolongan pengguna hasad yang mempromosikan video sendiri atau menyebarkan virus dan perisian berbahaya. Walaupun pengesanan pengguna hasad telah dilakukan berdasarkan pelbagai ciri seperti maklumat kandungan, aktiviti sosial pengguna, analisa rangkaian sosial atau ciri hibrid, kadar pengesanan masih dianggap rendah (iaitu 46%). Kajian ini mencadangkan satu set ciri yang terdiri daripada ciri pengguna, kelakuan pengguna dan ciri yang direka berdasarkan konsep *Edge Rank*. Kajian ini direalisasikan dengan menganalisis satu set pengguna *YouTube* dan kandungan video yang dikongsi. Ini diikuti dengan proses mengklasifikasikan pengguna dengan menggunakan 22 pengklasifikasi berdasarkan set ciri yang telah dicadangkan. Penilaian dibuat dengan membandingkan keputusan klasifikasi bagi ciri hibrid yang dicadangkan dengan ciri yang bukan-hibrid. Eksperimen yang telah dijalankan menunjukkan bahawa kebanyakan pengklasifikasi memperoleh keputusan yang lebih baik apabila menggunakan ciri hibrid berbanding dengan ciri bukan-hibrid. Purata kejituan klasifikasi adalah 95.6% bagi set ciri hibrid. Keputusan ini menunjukkan bahawa kajian yang dicadangkan akan memanfaatkan pengguna *YouTube* kerana pengguna hasad yang berkongsi kandungan yang tidak relevan boleh dikenal pasti. Ini akan membawa kepada pengoptimuman sumber sistem dan mewujudkan kepercayaan antara pengguna.

**Kata kunci:** Pengguna hasad, Pengesanan spam, *Edge Rank*, Pembinaan ciri

# 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 viruses and malware. Even though detection of malicious users have been done using various features such as the content, user social activity, social network analyses, or hybrid features, the detection rate is still considered low (i.e., 46%). This study proposes a new set of features that includes features of the user, user behaviour and also features created based on Edge Rank concept. The work was realized by analysing a set of YouTube users and their shared video. It was followed by the process of classifying users using 22 classifiers based on the proposed feature set. An evaluation was performed by comparing the classification results of the proposed hybrid features against the non-hybrid ones. The undertaken experiments showed that most of the classifiers obtained better result when using the hybrid features as compared to using the non-hybrid set. The average classification accuracy is at 95.6% for the hybrid feature set. The result indicates that the proposed work would benefit YouTube users as malicious users who are sharing non-relevant content can be detected. The results also lead to the optimization of system resources and the creation of trust among users.



**Keywords**: Malicious users, Spam detection, Edge Rank, Features construction

## Acknowledgement

First and foremost, all praise is to Allah, who by his grace and blessings the author have completed his dissertation.

The author would like to express his deepest gratitude to the advisor and supervisor Assoc. Prof. Dr. Yuhanis Yusof for the useful comments, remarks, and engagement through the learning process of this master study. It has been an honour to be one of her master students.

The author would like to take this opportunity to thanks, Ministry of Higher Education and Scientific Research and the University of Technology - Iraq for the generosity in funding scholarship. He is very honoured to be the recipient of this award. Receiving this scholarship motivates the author to maintain his GPA and complete his dissertation. He looks forward to being able to give back to the community once he begins his career teaching. All the thanks for confidence and willingness to help him achieve his goals.

The author would like to thanks, all UUM staff especially at School of Computing, College of Arts and Science, University Utara Malaysia and those that contributed indirectly towards the success of his studies.

The author would like to thank his family, who have supported him throughout the entire process, both by keeping him harmonious and helping him putting pieces together. I will be grateful forever for them love.

To his beloved wife, Noor and his princesses; Ban and Vian; the author appreciates their understanding and for being there with him while he is sailing through the arduous journey.

The author would like to express special thanks to his colleagues Dr. Athraa and Dr. Khalil for their support and encouragement.

To his friends; the author values their words of encouragement and he would like to thank them for walking with him when he needed support.

# Table of Contents

# List of Tables

# List of Figures

# List of Abbreviations

| | |
|---|---|
| OSM | Online Social Media |
| OSN | Online Social Network |
| SNA | Social Network Analysis |
| ERC | Edge Rank Checker |
| URL | Uniform Resource Locator |
| AVC | Advanced Video Coding |
| RT | Retweet |
| API | Application Programming Interface |
| HTTP | Hypertext Transfer Protocol |
| WSC | Web Scraper |
| Weka | Waikato Environment For Knowledge Analysis |
| MPEG | Moving Picture Experts Group |
| AVC | Advanced Video Coding |
| IT | Information Technology |
| RT | Retweeting |
| FeatureSet-UB | User Based Feature Set |
| FeatureSet-UBA | User Behaviour Feature Set |
| FeatureSet-ER | Edge Rank Feature Set |
| FeatureSet-H | Hybrid Feature Set |
| FeatureSet-HF | Hybrid Feature Set After Feature Selection |
| ML | Machine Learning |
| K-NN | K-Nearest Neighbours Algorithm |

| | |
|---|---|
| MLP | Multilayer Perceptron |
| ANN | Artificial Neural Network |
| K* | K Star Algorithm |
| NNge | Nearest Neighbour Like Algorithm |
| SMO | Sequential Minimal Optimization |
| FT | Functional Trees |
| SVM | Support Vector Machines |
| LibSVM | Library For Support Vector Machines |
| LibLINEAR | Library For Large Linear Classification |
| CFS | Correlation Feature Selection |
| CV | Cross Validation |

# CHAPTER ONE
# INTRODUCTION

## 1.1 Background

One decade ago, various online social media platforms (OSM) appeared and this includes the Hi5, LinkedIn, Facebook, YouTube, etc. Up to date, significant developments of this online interaction can be seen due to the explosive services that the network offers. Nevertheless, the popularity of these platforms led to malicious users target (Wuest, 2010; Zheng, Zeng, Chen, Yu, & Rong, 2014).

During the first age of social media, email filters were employed to detect malicious messages, where it caught over 95% of these messages. This is supported by another study that shows that email spam has dropped by half during 2010 according to Tan et al. (2013). Hence, malicious users try to find a new target for their activity (Tynan, 2012). Based on the easiest way to create a fake account and connect with people was social media, so they moved to social media networks. In the same time, malicious users are now able to reach more of personal information through using social media. In addition, they are now able to publish comments, links, fake detail, videos or follow people, like the post, add friends. These were new features used by malicious users to gain what they want from other legitimate users.

The threat of social malicious users is on the rise during the first half of 2013. According to Nexgate study in 2013, there is a 355% growth of social spam on a typical social media account. Even though most platforms have their own spammer detection technique, spammers always change their strategies and invent new scam techniques to circumvent most of the spam detection algorithms. However, various spam detection studies focus on

Facebook, Twitter, Renren, LinkedIn, and YouTube where these studies were conducted based on exposing different features of each site.

YouTube is one of the famous sites in social media network. It has over a billion users and almost a third of them are on the Internet every day (YouTube, 2015). YouTube has become the major channel for sharing video and delivering multimedia contents, where around 4 billion hours of video content are watched by more than 1 billion unique users visiting YouTube every month. It offers and supports a new feature of interaction among users, including video chats, political debates, video emails and video blogs (Chowdury, Adnan, Mahmud, & Rahman, 2013).

YouTube users watch hundreds of millions of hours of videos and generate a billion of views. The number of people watching YouTube each day has increased by 40% since March 2014. Furthermore, in March 2015, creators filming in YouTube Spaces have produced over 10,000 videos which have generated over 1 billion views and 70+ million hours of watch time (YouTube, 2015). According to the twice-yearly Global Internet Phenomena Report (Sandvine, 2015), during peak-period, real-time entertainment traffic is by far the most dominant traffic category, accounting for 40% of the downstream bytes on the network. Where in the same study, YouTube accounted for 17.7% of peak downstream traffic and year later that figure saw a significant increase to 21.2%.

Since malicious users like to scams other users, they will keep updating their channel by fake videos or with most popular movies to increase their score and make other users wanted to subscribe their channels. There are cases where most of the videos updated by the malicious user do not contain the media that is supposed to contain (Chowdury et al.,

2013). This means the threat of those users not only have an effect on legitimate users but also led to network bandwidth consumption (Benevenuto et al., 2008). Hence, there is a need to automatically detect spammers among YouTube channels.

The malicious user usually publishes a spam in order to promote a specific content, advertisements to generate sales or to increase view count for some websites to make them more credible (Tan et al., 2012). There are a number of spam detection techniques that exploit characteristics present in the contents for instance email body and comments in social sites (Hu, Tang, Gao, & Liu, 2015).

On the other hand, malicious users also publish video spamming, and this is commonly found in social video sharing systems Such as YouTube, In this case, it can be much more challenging to detect (Chowdury et al., 2013). However, it is not easy to apply content based features (Razmara et al., 2012; McCord & Chuah, 2011) in malicious detection over video objects while the content based required textual features in order to be analysed. There are some studies to detect malicious users based on social media features of engagements (Zhu et al., 2012). These features include social activity where it focuses on capturing the interaction or behaviour between users (Yardi, Romero, Schoenebeck, & Boyd, 2010). In addition, there are also work on underneath communication (i.e. relationship between users, network usage) in order to distinguish spammers (O'Callaghan, Harrigan, & Carthy, 2012).

Nevertheless, the aforementioned studies become less effective as a result of the rapid evolution of the techniques used by malicious users (Hu, Tang, & Liu, 2014). Hence, recent work for malicious user detection is based on the combination of features. For example,

Zheng et al. (2015) had been conducting a study on detecting spammers in social networks using 18 features. Similar work can also be seen in Zhu et al. (2012) conducted a study that proposes the combination of users' social action and social relations features. On the other hand, Benevenuto et al. (2008) and Kiran (2015) conducted studies using three subsets of features: user details, social network, and video attributes. These features are of user-based and user-behaviour. However, these studies could not obtain accuracy higher than 46%. This may be due to the employment of irrelevant features.

### 1.2 Problem Statement

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 (YouTube, 2015). YouTube like the other social media platform depends on media contents that are created and shared by users. Such an approach allows malicious users (i.e. spammers) to exploit it (Wuest, 2010; Zheng, Zeng, Chen, Yu, & Rong, 2014). According to a market survey on the impact of spammer over social media in 2008, 83% of social media users have received at least one message or friend request from unknown accounts in (UK, 2008; Kiran, 2015).

One of the main aims of 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. 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 (Kiran, 2015). 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).

Existing literature on online social media has proposed several malicious user detection approaches. However, the proposed methods focused on social content analysis that relies on keyword-based filtering and URL-based detection (Bhat, Abulaish, & Mirza, 2014; Burnap, Javed, Rana, & Awan, 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 Markus and Ratkiewicz; Alex and Jakobsson diaries (as cited in Soman & Murugappan, 2014). Moreover, other methods have identified spam video in YouTube using hybrid analysis. The hybrid analysis integrates a set of features, namely video attributes, user attributes, and social network metrics. However, this integration only able to identify 44% and 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.

In order to address the aforementioned limitations, this dissertation employs the concept of Edge Rank Checker (ERC) (Socialbakers, 2015; Zheng et al., 2015) used on Facebook to construct a feature set for malicious user detection.

### 1.3 Research Questions

The research questions of this dissertation are as follows:

a. What are the features to be included in YouTube repository?
b. How to construct new features based on the content of the created YouTube repository?
c. How to classify whether a YouTube user is a malicious user or legitimate?

### 1.4 Research Objectives

The aim of this dissertation is to propose a set of features that can be used to detect malicious users in YouTube. This can be achieved by the followings:

a. To create a repository of YouTube channels and its contents.
b. To construct hybrid features based on the integration of user-based and user-behaviour features.
c. To classify YouTube users as either malicious or legitimate using the proposed hybrid features.

## 1.5 Scope of the Study

This dissertation is focused on detecting malicious users among YouTube channel owners who publish spam videos on their account. The proposed features have been used to classify the user's account into either legitimate user or malicious user. In particular, this dissertation adopts the concept of Edge Rank Checker (ERC) (Socialbakers, 2015) that has been used in Facebook to assign a score for each post. In Facebook, the higher the ERC score, the less possibility it is to be a spammer (Zheng et al., 2015).

Furthermore, this dissertation employed features crawled from YouTube platform for a period of four months. These features includes the ones based on the integration of content (user-based) and social activities (user-behaviour).

## 1.6 Significance of the Study

YouTube is overstuffed with various unwanted videos and comments that infiltrate existing method that YouTube depends in detecting malicious contents (Benevenuto et al., 2008; Kiran, 2015). While, malicious user tries to publish video or comments to increase the popularity of his channel, YouTube employed limited tools for video and comment analysis. Hence, malicious users volume increases and lead owners of famous channels to disable the comments section in their videos as a sign of protest (Alberto, Lochter, & Almeida, 2015).

This dissertation introduces new feature set that represents information on the user, user-behaviour and the shared content. The combination of these information is useful in differentiating between malicious and legitimate users. This will later benefit YouTube community in obtaining the required multimedia content and create trust among users and

the channel owners. Furthermore, system resources can be optimized as irrelevant content will not be retrieved.

### 1.7 Organization of Dissertation

The rest of this dissertation is organized as follows. In the next chapter, this dissertation reviews exiting work in malicious user detection. Chapter 3 contains the employed methodology while the proposed features are described in Chapter 4. Chapter 5 presents the results of the experiment along with its discussion. The final chapter demonstrates the contribution, future work, and limitation of the work.

# CHAPTER TWO
# LITERATURE REVIEW

This chapter presents the review of related literature. Section 2.1 gives an introduction which is the general information on social media platform. Online social media and their features have been discussed in Section 2.2. While Section 2.3 includes the threats of malicious users on social media community along with existing approaches of malicious users' detection. Section 2.4 illustrate common techniques of classification that was used by other studies. Then, Section 2.5 describes the fundamentals and requirements of create a features set. Moreover, research gap and summary of the chapter is presented in Section 2.6 and 2.7 respectively.

## 2.1 Introduction

Through online communication, people lives have been changed and become intertwined over time. While the Internet appeared and online activity started, many options 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 (Browsing). Online social media sites offer details about users, hence that the users become target for spammers, scams and other different attacks. However, the media have several options to create and share content, where users can update status, post short text, links, images, videos and send messages. Furthermore, these attacks primarily grew on social media networks such as YouTube (Benevenuto et al., 2008; Chowdury et al., 2013). The popularity of these sites makes them perfect spots for performing cybercriminal activities (Alberto et al., 2015).

## 2.2 Online Social Media

Current online social media (OSM) or online social network (OSN) offers two major characteristics. The first one is the content sharing; contents that are created and shared by any user are available to other users for viewing, add opinions, rating, and bookmark. In YouTube, 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 interest of a user towards another user's activity. For example, in order to keep updated with latest activity for the 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 a user A to user B means that the former is interested in the latter's activity, but not necessarily vice-versa (Chiluka, Andrade, & Pouwelse, 2011).

A study shows that Web 2.0 platforms and social media websites such as YouTube, discussion forums, blogs, and video sharing platforms are an easy target for malicious users based on the anonymity and no barrier policy to post content (Heymann, Koutrika, & Garcia-Molina, 2007). Table 2.1 shows the most popular social media sites, sorted by date of launched.

Table 2.1

*Most Popular Online Social Media*

| **OSN** | **Date launched** | **Focus** | **Registered users** |
|---|---|---|---|
| **LinkedIn** | May 2003 | Business and professional networking | + 400 M (LinkedIn, 2016) |

| OSN | Date launched | Focus | Registered users |
|---|---|---|---|
| **hi5** | June 2003 | General. | 100 M (Hi5, 2016) |
| **Myspace** | Aug 2003 | General | 50.6 M (Myspace, 2015) |
| **Facebook** | Feb 2004 | General: photos, videos, blogs, apps. | 1.59 B (Statista, 2016) |
| **YouTube** | Feb 2005 | Video sharing website | + 1 B (YouTube, 2015) |
| **Twitter** | July 2006 | General, Micro-blogging | 1 B (Twitter, 2015) |
| **Academia.edu** | Sep 2008 | Social networking site for academics and researchers | + 32.6 M (Academia, 2016) |
| **Google+** | Dec 2011 | General | 1.6 B (Digitalinsights, 2014) |

### 2.2.1 YouTube

Nowadays, the most popular website on the Internet for video sharing system and social network features is the YouTube. Figure 2.1 shows video contextual features of YouTube, where it uses WebM, H.264/MPEG-4 AVC and AdobeFlash Video technology to play a wide range of videos generated by users and corporate media. Through the YouTube platform, users are able to view, upload videos, share videos also subscribe a channel, like or dislike any published video, and post a short textual comment on a published video. According to YouTube Statistics in 2015, around 4 billion hours of video content are watched by more than 1 billion unique users every month. The same statistics show that over 100 million users participate in actively of either liking, disliking a video or by replying a comment, where every minute around 72 hours of new videos are uploaded on YouTube.

The popularity of such video sharing provides a platform for malicious users like spammers and promoters to post unrelated, low quality and irrelevant content, either as video a

response or as a related video to the most popular videos either to gain popularity or to promote their sites or products (Benevenuto, Rodrigues, Almeida, Gonçalves, & Almeida, 2009). Malicious users become a serious problem as there is an enormous amount of data that streams on YouTube platform every minute. The presence of spam in such case could lead to losing bandwidth, time waste, and degraded user experience which is undesirable.

Figure 2.1 illustrates the channel various details on YouTube. A video has a contextual feature includes a title, a brief description of the video, textual comments, category of the video (entertainment, music, people). Whereas the total number of channel subscribers illustrate the success of the particular channel. Furthermore, the total number of view integrated with rating details refer to the quality of the content and state how many other YouTube users satisfied with such video.


*Figure 2.1.* Channel and Video Detail on YouTube

### 2.2.2 Facebook

One of the largest Social media networks in the world is Facebook (Smith, 2015). Facebook allows users to set up their profiles that contain some personal information such as name, marital status, birthday, and other personal interests or hobbies, also found bidirectional ("friending") or unidirectional ("following") social links with other users.

Each user in Facebook has an asynchronous messaging mechanism between friends illustrate as message board called "wall". Usually, friends are able to contact over Facebook by posting messages on their walls. The posts are the main way for users to share information on Facebook. The content of posts can be text, URL or a photo shared by a user. Like is a widget that gives to an object in Facebook, such as a post, a page, or an application. When users click the Like button, the corresponding object will appear in their friends' newsfeed and thus allows information about the object to spread across Facebook. Facebook has a newsfeed page, which it responsible for showing a summary of friends' social activities. Furthermore, Facebook allows third-party developers to create and develop their own applications to serve users (Facebook, 2016).

### 2.2.3 Twitter

Twitter is another well-known social media network that focuses on information sharing. It allows users to share short messages, which are called tweets and should be not more than of 140 characters. Over Twitter, the relationship between users is called "following", where this relationship has not required a reciprocation in the process of accept following and being followed. Any user on Twitter can be a follower or a followed, and a user being followed need not follow back. In the case of a follower, he/she will able to receive all tweets sent by him/ her followers. When a followed publish or shares a tweet, this event

will be distributed to all others followers. Retweet, it's the process of re-sends someone's tweets (RT) for other users, that mean the followers can receive this event as well. Furthermore, a user is able to send a tweet to specific users by mentioning them in the tweet using ("@") before the identifier name of twitter receivers. This well-defined markup vocabulary combined with a strict limit of 140 characters per tweet conveniences users with brevity in expression.

The number of malicious users on the two popular social media networks namely Facebook and Twitter, have been compared with YouTube based on Digital Media Ramblings statistics as shown in Table 2.2 (Smith, 2015).

Table 2.2

*Fake Users, View Statistics of YouTube vs Facebook vs Twitter*

| **Statistics** | **Facebook** | **YouTube** | **Twitter** |
|---|---|---|---|
| **Total number of registered users** | 1.59 billion (Statista, 2016) | Over one billion (YouTube, 2015) | One billion (Twitter, 2015) |
| **Fake users, view** | 140 Million Fake users (Smith, 2015) | Over one billion Fake user/view (Gayle, 2012) | 20 Million Fake users (Smith, 2015) |

### 2.3 Malicious Users

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. 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

information. They do not need to be computer savvy to compromise sensitive information. Table 2.3 illustrates the different types of items created by malicious users.

Table 2.3

*Types of Malicious Items over OSN*

| Type | Details |
|---|---|
| **Spam** | The comment that has commercial content unrelated to the discussion at hand or it involves contacting users with unwanted content (Profanity, Insults, Bulk, Hate speech, Threats) or requests (Facebook Help Center, 2016) |
| **Video spam** | A video spam is illustrated when a video posted as a response to an opening video, but their content is completely unrelated to their title or videos without content (Benevenuto et al., 2008) |
| **Malicious links** | User submitted a comment with malicious links that could mislead, inappropriately harm, or otherwise, damage a user account or computer (Burnap et al., 2015) |
| **Fraudulent reviews** | Reviews of a service or product from a fake user where they never used it, and thus misleading or insincere (Hsu, 2012) |
| **Fake friends, Subscribe** | It happens when several fake accounts become "friends/Subscribe". Usually, these users are spambots try to gain credibility by following/subscribe certified accounts/channel, such as those of popular celebrities and public figures (Fernandes, Patel, & Marwala, 2015) |

Many studies on detection of malicious users over online social media have been conducted by mining social media content and analysing it (i.e. Content-based) (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 like. This

approach is also known as profile-based (Chowdury et al., 2013). Lee and Kim (2012) conducted a study by mining URL and came out with URL redirecting patterns and detect malicious users.

On the another hand, there is also work through mining social activity 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 & Abulaish, 2013).

The hybrid analysis is another approach, where it uses a group of different features or ensemble classifiers group. It could be considered a hybrid analysis also by integrated both of them for enhancing the classification results of a proposed system. It is founded on the assumption that grouping of multiple classifiers based on different features, may be able to produce an overall classifier which is more accurate and stable than of individual one (Bhat, Abulaish, & Mirza, 2014). Figure 2.2 shows the common feature analysis approach of malicious users based on the examined literature, which is four approaches namely content analysis, social activity analysis, social network analysis, and hybrid analysis.

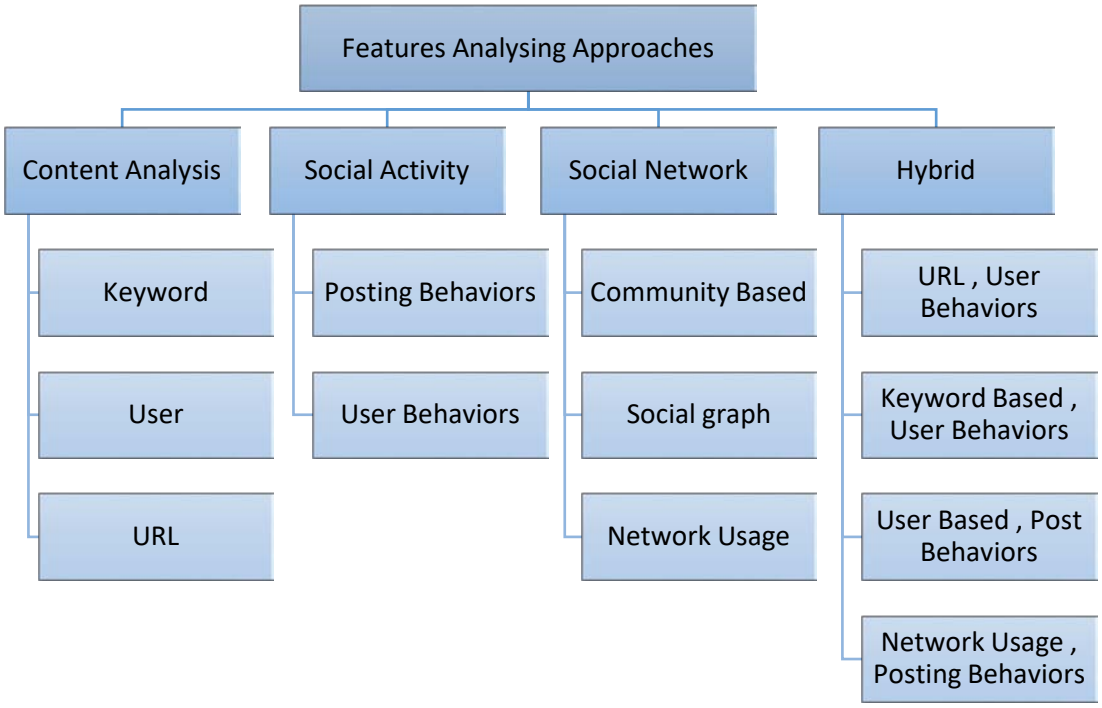


*Figure 2.2.* Features Analysis Approaches

### 2.3.1 Content Analysis Approach

Many malicious users' detection techniques have been proposed based on content features, whereas the most common technique is keyword-based filtering and users' interaction data. However, many counters filtering techniques based on the frequent use of non-dictionary words and images in malicious objects (Bhat et al., 2014). Table 2.4 illustrates the most common features used in content analysis approach.

Table 2.4

*Common Features Used in Content Analysis Approach*

| Analysis Features | Method | Limitation |
|---|---|---|
| **Keyword-Based** | Mining keywords from given text and either compare it with a specific list (Unsupervised method), or extract a frequent pattern of some keywords (supervised method) to detect malicious users (Razmara, Asadi, Narouei, & Ahmadi, 2012; Hu et al., 2014) | Requires large computational cost. Moreover, the issue of user content privacy (private messages, posts, profile details) is usually held against it (Bhat & Abulaish, 2013) |
| **URL-Based** | URL technique, works by mining the URL and also it works using an unsupervised method by comparing URL with black list, or use a supervised method by extracting either URL redirecting patterns or posting behaviours (Cao & Caverlee, 2015) | Most of the URL is shortened using the link shortening service making the detection very difficult. Also, sometimes the problem is not with the first link; it could be led to another malicious link. (Rodrigues, Benevenuto, Cha, Gummadi, & Almeida, 2011) |
| **User-Based (Profile-Based)** | User profile-based methods build a classifier using some features extracted from account profiles, such as about me, number of friends, and number of views. (Singh, Bansal, & Sofat, 2014) | Using user-based alone will not generate a good classification cause most malicious users mimic legitimate user profiles (Kiran, 2015) |

### 2.3.2 Social Activity Analysis Approach

This approach depends on social activity between users inside social media. Since users in social media are allowed for interaction with other users, many studies are conducted to distinguish the user behaviours. The behaviour of a small part of malicious users had been examined over Twitter (Yardi et al., 2010), while the researchers found that the behaviour of malicious users is different from the legitimate users in terms of posting tweets,

following friends, followers and so on. Table 2.5 demonstrates the most common features used in the social activity approach.

Table 2.5

*Common Features Used in Social Activity Approach*

| Analysis Features | Method | Limitation |
|---|---|---|
| **User-Behaviours** | Builds a classifier using features extracted from the user's activity on social media that describe user's relationship (Benevenuto et al., 2008), such as number of comments, number of tags, number of like/dislike, number of friends request. | Using thees features alone will not able to detect malicious users (Kiran, 2015). Thus, because of user privacy issues |
| **Posting-Behaviours** | Focuses on user's posting behaviours. For instance, the behaviour of post large number of unrelated/repeated comments (Zhu et al., 2012) | User privacy could be effected on detection, while the behaviours of malicious users can quickly change make them hard to detect(Zhu et al., 2012) |

### 2.3.3 Social Network Analysis Approach

The process of examining social structures over the use of the social graph and network structure or usage is defined as social network analysis (SNA). As an alternative to traditional approaches, this approach depends on structures of the network in terms of people, nodes, or other things within a network combine to the ties or edges of relationships or interactions that connect them (Ulrike, 2001). Basically, this features used in detection through arranges nodes, people, or other things inside the network in clusters, based on node interaction it will determine the malicious nodes. Furthermore, some study conducted using this approach to detect campaigns of malicious users depend on network usage

features (O'Callaghan et al., 2012). Table 2.6 shows the most common features used in the social network approach.

Table 2.6

*Common Features Used in Social Network Approach*

| Analysis Features | Method | Limitation |
|---|---|---|
| **Community Based** | Depends on network-based topological features of the interacting users/nodes over online social networks (Bhat & Abulaish, 2013) | Malicious users are often seen to mimic legitimate user patterns of interaction behaviour making it difficult to characterize them (Bhat & Abulaish, 2013) |
| **Graph-Based** | Employs social network relationship into consideration (Tan et al., 2013) | Using this approach alone may be worthless unless joining with other features this is mean this approach does not have powerful features (Bhat et al., 2014; Kiran, 2015) |
| **Network Usage Based** | Focus on network usage. For instance, recurring campaigns derived from a user's comment activity posted by users (O'Callaghan et al., 2012) | |

### 2.3.4 Hybrid Analysis Approach

Traditional methods of malicious user detection become less effective as a result of the rapid evolution of the techniques used by malicious users. First, malicious users over social media show dynamic patterns due to post content and posting behaviours inside social media. Usually, malicious users' content information and behaviour change too fast to be detected by a static system based on off-line modelling (Hu et al., 2014).

The existing analysing systems as shown in Table 2.7 rely on building a new detecting model to capture content analysis approach combined with another approach to extract a pattern of malicious users. Given the rapidly evolving nature, it is essential to have an

efficient framework that reflects the effect of newly emerging data. There are many studies in the literature to find efficient ways to handle malicious users' activity through classification methods using hybrid feature extraction from online social media (Chowdury et al., 2013; Bhat et al., 2014; Zheng et al., 2015). In brief, these studies usually used different approaches to creating a learning classification machine that is able to distinguish users based on hybrid features.

Table 2.7

*Summary of the Work on Hybrid Approach*

| No. | Author/Year | Features Used | Methodology | Remarks |
|---|---|---|---|---|
| 1 | Benevenuto et al. (2008) | User based, Social activity, Social network | Analysis set of features using SVM | 46% of spammers was detected |
| 2 | Stringhini et al. (2010) | User based, URL based, Keyword based | A new metric that allows predicting the success of a campaign using honeypot. Analysis the integrated features using Random Forest algorithm | Identify single spam bots, as well as large-scale campaigns |
| 3 | Benevenuto et al. (2010) | User based, URL based, Keyword based | Analysis set of features using SVM | 70% of spammers and 96% of non-spammers were correctly classified |
| 4 | McCord and Chuah (2011) | User based , Keyword based | Compared Random Forest, SVM, Naive Bayesian, K-NN | RFC giving highest accuracy at 96% |
| 5 | Zhu et al. (2012) | Social Activity, Social network | Analysis set of features using SVM | 95% of spammers was detected |

| No. | Author/Year | Features Used | Methodology | Remarks |
|---|---|---|---|---|
| 6 | Fernandes et al. (2015) | Content analysis, Social network | compares SVM classification and a number of clustering approaches to separate human from not human users in Twitter | around 90% F1 accuracy was the scores for both of the classification and clustering approaches |
| 7 | Kiran (2015) | User based, Social activity, Social network | Analysis set of features using SVM | 47% of spammers was detected |
| 8 | Zheng et al. (2015) | Content analysis, Social activity | Analysis set of features using SVM, Decision Tree, Naïve Bayes and Bayes Network | 99% of spammers was detected |

Based on the examined literature, it is learned that the hybrid analysis approach becomes the accepted approach in classifying malicious users. Since, most of recent work employed features based on the integration of two or more approaches, this dissertation follows the same. In addition, the proposed feature set includes features created based on the Edge Rank concept.

Stringhini, Kruegel, and Vigna (2010) presented a hybrid mining approach for detecting malicious profiles over Facebook, MySpace, and Twitter. They used content analysis approach based on user messages, URL ratio and keyword based for choosing effective features. Other researchers achieve good performance result by building SVM classification model focused on 18 feature items using content analysis and social analysis over Weibo site (Zheng et al., 2015).

Benevenuto et al. (2010) achieved approximately 70% spammers correctly classified. They have conducted the study over Twitter using hybrid features, this is illustrated integration a set of features based on user based, URL-based, keyword-based. Furthermore, another

study conducted based on an effective way for detection malicious users by building a classifier based on content analysis approach and social network approach (Hu et al., 2014).

McCord and Chuah conduct a study in 2011 to detect spammer over Twitter. They extract hybrid features from the social content approach. Then they analyse these features to distinguish between malicious users and legitimate ones. Such integration is able to achieve the very good result through analysis it using a set of the existing algorithm (McCord & Chuah, 2011).

Using 18 features from the integration of content-based and social activity approach allow Zheng et al. (2015) to achieve a very good result. The best performance result was around 99.1% accuracy for spammer detecting using SVM classification based on proposed integration. Moreover, some other studies achieve a good result by focusing on users' social actions and social relations (Zhu et al., 2012).

Fernandes et al. (2015) proposed a slightly different method by classification of users over two stages. The first stage tries to classified users into two class either human or not human users over Twitter to identify normal human activity. The second stage is to classify not human into brands, celebrities, and promoters. Also, they compared classification and clustering approaches to separate human from not human users in Twitter. However, around 90% F1 accuracy was the scores for both of the classification and clustering approaches (Fernandes et al., 2015).

Other researchers conducted studies to capture video spam in YouTube. For instance, Benevenuto et al. (2008) and Kiran (2015) used a set of features, including video attributes, user attributes, and social network metrics. Unfortunately, they failed to achieve a good

result even they used hybrid analysis approach. However, such integration was able to identify only 44% and 46% of the video spammers, this because using a single video analysis and irrelevant features (Benevenuto et al., 2008; Kiran, 2015).

### 2.4 Classification Methods

The aim of classification is to categorize data to certain group. The classifier able to categorize each instance based on features values into one of a set of possible classes. There are many popular classification algorithms have been used for detection such as K*, Random Committee, Random Forest, Bayes Net, JRip, PART, LibLINEAR, Multilayer Perceptron, NNge, SMO, Hoeffding Tree, FT, J48graft, Naïve Bayesian, Logistic, Multi Class Classifier, Attribute Selected Classifier, AdaBoostM1, J48, Decision Tree, J48 Consolidated, LibSVM (Korb & Nicholson, 2011; Tretyakov, 2004; Chowdury et al., 2013; Fan, Chang, Hsieh, Wang, & Lin, 2008; Chang & Lin, 2011; Strano & Colosimo, 2006; Dreyfus, 2005; Platt, 1998; John G. Cleary, 1995; Freund & Schapire, 1996; George-Nektarios, 2013; Salih & Abraham, 2014; Witten & Frank, 2005; Cohen, 1995; Martin, 1995; Frank & Witten, 1998; Pfahringer, Holmes, & Kirkby, 2007; Quinlan, 2014; Ibarguren, Pérez, Muguerza, Gurrutxaga, & Arbelaitz, 2014; Webb, 1999; Breiman, 2001).

Chowdury et al. (2013) conducted a study to detect spammer over YouTube. Chowdury study's employed Decision Tree plus two other methods to detect spammer users. The study was achieved a good result through using Decision Tree method. In the end, the study concluded that Decision Tree is more accurate for predicting spammers especially for the higher number of test cases (Chowdury et al., 2013).

According to Tan et al. (2013) that classified labelled feature set using different classifiers (Naive Bayes, Logistic Regression, and Decision Tree) based on all features that they have, Decision Tree has 98.6% true positive rate with the lowest false positive rate at 1.6% (Tan et al., 2013). Later, Singh et al. (2014) conducted a study to detect malicious users. They were able to achieve a good result using features extracted from the user-based approach. They used some common method for classification such as BayesNet, Naïve Bayes, SMO, J48, and Random Forest. Whereby using Naïve Bayes in classification did not achieve the best result through experiment, but it is still able to get high accuracy at 95.8% (Singh et al., 2014).

Saad et al. (2014) conducted a comparative study of classification algorithms of e-mail filtering, where they used ANN in the comparative study. Through experiment, they found the effectiveness of ANN was slightly increased by using more features (Saab, Mitri, & Awad, 2014).

According to Zheng et al. (2015) study on spammers, they compare the effectiveness of proposed features across SVM with other existing algorithms. They obvious found that SVM classifier is capable of achieving better result compare with other classifiers.

Tretyakov study gives an overview of some of the common machine learning methods (e.g. k-NN, ANNs, Bayesian classification, SVMs) and of their applicability to the problem of spam filtering (Tretyakov, 2004). Based on experiment both of ANN and SVM was able to achieve a good result at 98%. However, the researcher surprised with ANN performance where it gets a better result.

The following subsection includes a brief description of the types of classifiers used in this study; Bayes network, Function-based, Lazy, Metaheuristics, Rule-based and Trees. The 22 classifiers consider the best known algorithms based on examine literatures, while they have been used in classification different tasks include of spammers (Korb & Nicholson, 2011; Tretyakov, 2004; Chowdury et al., 2013; Fan, Chang, Hsieh, Wang, & Lin, 2008; Chang & Lin, 2011; Strano & Colosimo, 2006; Dreyfus, 2005; Platt, 1998; John G. Cleary, 1995; Freund & Schapire, 1996; George-Nektarios, 2013; Salih & Abraham, 2014; Witten & Frank, 2005; Cohen, 1995; Martin, 1995; Frank & Witten, 1998; Pfahringer, Holmes, & Kirkby, 2007; Quinlan, 2014; Ibarguren, Pérez, Muguerza, Gurrutxaga, & Arbelaitz, 2014; Webb, 1999; Breiman, 2001).

### 2.4.1 Bayes

- **BayesNet**

  A Bayes net is a model. It reflects the status of some parts of an object or domain that is being modelled and it describes how those states are related by probabilities. The objector domain may possibly be a house, or a car, a human body, a community, an ecosystem, a stock-market, etc. Obviously, anything can be modelled by a Bayes net. All the possible states of the model represent all the possible domains that can exist, that is, all the possible ways that the parts or states can be configured. Bayes nets may be used in any walk of life where modelling an uncertain reality is involved (and hence probabilities are present), and, in the case of decision nets, wherever it is helpful to make intelligent, justifiable, quantifiable decisions that will maximize the chances of a desirable outcome. In short, Bayes nets are useful universally (Korb & Nicholson, 2011).

- **Naïve Bayes**

  Naive Bayes classifiers are popular statistical methods of spam filtering. They typically use a bag of word features to identify spam e-mail, an approach commonly used in text classification. Bayesian algorithm does not have to set rules in advance and does not need to analyse the content of the e-mail. Through the analysis of characteristic words and the category of text, it gets the statistical models (Guo et al., 2014). Many studies were conducted using this method and produced good results in classification. (Tretyakov, 2004; Chowdury et al., 2013).

### 2.4.2 Functions

- **LibLINEAR**

  The LibLinear classifier represents an open source library that is developed for large-scale linear classification. The logistic regression and linear support vector machine are also supported. It is considered as a very efficient classifier on large sparse of feature set (Fan et al., 2008). Yuan, Ho, and Lin (2012) investigated the efficiency of large linear classification. They discovered that linear classifiers may give equal results to nonlinear classifiers, but yet with less computational time.

- **LibSVM**

  LibSVM which stands for Library for Support Vector Machines illustrates an integrated software for support vector classification, (C-SVC, nu-SVC), regression (epsilon-SVR, nu-SVR) and distribution estimation (one-class SVM). It supports multi-class classification (Chang & Lin, 2011).

Zheng et al. (2015) who conducted a research on spammers, made a comparison between the effectiveness of the proposed features across LibSVM and other existing algorithms. They evidently found that LibSVM classifier was capable of achieving a better result as compared to other classifiers (Zheng et al., 2015).

- **Logistic**

  This is a class for building and using a multinomial logistic regression model with a ridge estimator. The binary logistic model is used to estimate the probability of a binary response based on one or more predictor (or independent) variables. The binary logistic model is used to estimate the probability of a binary response based on one or more predictor (or independent) variables (features). Logistic regression can be binomial, ordinal or multinomial. Binomial or binary logistic regression deals with situations in which the observed outcome for a dependent variable can have only two possible types (Strano & Colosimo, 2006).

  Kumar et al. (2015) conducted a study in order to measure the intent of ads clicking for online users. They had selected Logistic Regression classifier to estimate that. The predicted in Logistic Regression will be always estimated as probabilities, lying between 0 and 1. Using this prescribed they were able to achieve around 90% accuracy rate.

- **Multilayer Perceptron**

  A multilayer perceptron (MLP) is a feedforward artificial neural network (ANN) model that maps sets of input data onto a set of appropriate outputs. Whereas an

MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one.

Artificial neural network (ANN) is a popular learning algorithm and it was inspired by biological neural networks. ANN is used to estimate functions that can depend on a large number of inputs and is generally unknown. Nowadays, many variants of ANN exist with improved performance.

The enormous power of artificial neural networks (ANNs) as pattern classifiers or feature selectors has been used in many fields, for instance, image compression, character recognition, market prediction and loan applications (Dreyfus, 2005). Many of these applications utilise huge neural networks with thousands of neurons.

- **SMO**

  Sequential Minimal Optimization (SMO) is a simple algorithm that considers the solution for SVM-QP problem. It can quickly solve a problem without any extra matrix storage and without using numerical QP optimization steps at all. SMO decomposes the overall QP problem into QP sub-problems, using Osuna's theorem to ensure convergence (Platt, 1998). Singh et al. (2014) conducted a study in order to detect malicious users that harm genuine ones. Singh study is employee several classifiers include that SMO to classified 7434 users. While using SMO was able to achieve 81.6% accuracy rate.

### 2.4.3 Lazy

- **K Star**

  K* represents an instance-based classifier, the class of a test instance that is based upon the class of those training instances similar to it, as determined by a certain similarity function. The K* differs from other instance-based learners because it uses the entropy-based distance function (John G. Cleary, 1995).

  Nisa and Ahsan (2015) conducted a study to predict fault for software using soft computing techniques. They were applying various machine learning (ML) classifiers, K* on of these algorithms. The prediction models were improved by handling ML such as feature reduction and class imbalance and K* performed very well.

### 2.4.4 Meta

- **AdaBoost M1**

  "Boosting" represents a general method that is able to improve the performance of any learning algorithm. In theory, the error of any "weak" learning algorithm could be reduced by boosting. Boosting, which is slightly better than random guessing, consistently generates classifiers (Freund & Schapire, 1996). Shams and Mercer (2013) conducted a study to classify email spam based on content-language and readability combined with content-based task features. They used five well-known algorithms to detect the spammers that include AdaBoostM1 classifier. The extensive experiments imply that the examined classifiers generated using meta-

learning classifier perform better than other classifiers such as trees, functions, and probabilistic methods.

- **Attribute Selected Classifier**

  Attribute Selected Classifier is the dimensionality of training and test data that is reduced by attribute selection before being transmitted to a classifier (George-Nektarios, 2013). Villuendas, Yanez, and Rey (2015) applied this concept in selecting relevant objects and features before implementing classification. This concept has been found to be more efficient in the correct discrimination of objects. The pre-processing contributes to increasing the desired efficiency and robustness of the classifier.

- **Multiclass Classifier**

  A metaclassifier is to manage multi-class datasets with 2-class classifiers. This classifier is also capable of applying error correcting output codes for increased accuracy (George-Nektarios, 2013). Babu and Pradeepa (2013) conducted a comparative study for underwater target classification based on multiclass classifiers. They address that by comparing different techniques of multiclass classification. That was determined by using a particular feature derived from the real datasets. The performance of multiclass classifiers is analysed in details, with different methods.

- **Random Committee**

  Random Committee classifier is used to build an ensemble of random base classifiers. Each base classifier is built with a different random number seed. The

final prediction is a straight average of the predictions generated by the individual base classifiers (Salih & Abraham, 2014). A comparison study has been conducted by Amasyali and Ersoy (2011) to compare of single and ensemble classifiers, and it is learned that Random Committee produces good results in terms accuracy and execution time.

### 2.4.5 Rule

- **Decision Table**

  Decision Tree is one of the learning methods that is commonly used in the field of data mining, with the aim of creating an effective model that enables the prediction of the value of a target variable based on several input variables. Each node agrees on one of the input variables; there are arc/edges for each possible value of that input variable. Each leaf indicates a value of the target variable given the values of the input variables represented by the path from the root to the leaf (Witten & Frank, 2005). A data mining methods applied by Chowdury et al. (2013) to classified YouTube users into malicious or legitimate. They applied these methods to understand and predict the behaviour of a YouTube video. That includes DT, where it was able to obtained accuracy rate up to 98.66%.

- **JRip**

  The JRip is a classifier that implements a propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER), which was proposed by William W. Cohen as an optimized version of IREP (Cohen, 1995). Hassan and El Fattah Hegazy (2015) worked on a big dataset which encompasses two data

sources. The first is Learning Management System (LMS) while the second is social media related to that LMS. In order to do classification as well as a prediction for the learner's learning style LS, they employed WEKA as well as they did analysis through different machine learning classifiers includes JRip. While this is done in order to know which one will fit for the used dataset. The results showed that JRip able to achieve a good rate at 90%.

- **NNge**

  The nearest-neighbour-like algorithm uses non-nested generalized exemplars (which are hyper-rectangles that can be viewed as if-then rules) (Martin, 1995). A study has been conducted by Weber and Mateas (2009) to demonstrate a data mining approach for strategy prediction in real-time strategy games. They applied data mining over a large number of game traces, then they developed an opponent model that is not limited to a single opponent, set of maps or style of gameplay.

- **Part**

  Part is a classifier for generating a PART decision list. It applies the concept of separate-and-conquer, whereby it builds a partial C4.5 decision tree in each iteration and makes the "best" leaf into a rule (Frank & Witten, 1998). Diplaris, Tsoumakas, and Mitkas (2005) published a study of classifying new proteins to structural families by classification models. They studies data concerning patterns of proteins with known structure. Several approaches had been applied, that combine multiple learning algorithms to increase the accuracy of predictions that

includes Part classifier. The result showed that by using Part classifier, a low error rate was obtained, that is 0.026.

### 2.4.6 Trees

- **Functional Trees**

  FT is a classifier for building 'Functional trees', which are classification trees that can have logistic regression functions at the inner nodes and/or leaves. The algorithm can deal with binary and multi-class target variables, numeric and nominal attributes and missing values (Frank & Witten, 1998). A novel intelligent application was introduced by Khakham, Chumuang, and Ketcham (2015) to classify the traditional handwritten datasets using Functional Trees. They have developed a new approach for classifying images, using several of feature extraction methods for recognition system. The dataset is separated into two groups for training and testing, respectively. They used nine features in this process of classification with average accurate at 82.33%.

- **Hoeffding Tree (VFDT)**

  A Hoeffding tree (VFDT) is an incremental, anytime decision tree induction algorithm that is capable of learning from massive data streams, assuming that the distribution generating examples does not change over time. Hoeffding trees exploit the fact that a small sample can often be sufficient in order to choose an optimal splitting attribute. This idea is supported mathematically by the Hoeffding bound, which quantifies the number of observations (in our case, examples) needed

to estimate some statistics within a prescribed precision (in our case, the goodness of an attribute) (Pfahringer et al., 2007).

- **J48**

  The J48 is a class for generating a pruned or unpruned C4.5 decision tree (Quinlan, 1993). A community-based feature has been identified by Bhat and Abulaish, in 2013. The main aim of the study has proposed a framework to detect spammers from online social networks. While the main element of their study is learning a classifier (i.e J48) from the features of a community-based node of online social networks. In this regard, they present the performance of some classification models learned using the proposed features include that J48. The results showed that J48 was one of the top classifiers in detecting spammers.

- **J48 Consolidated**

  J48 Consolidated is a class for generating a pruned or unpruned C45 consolidated tree. It uses the Consolidated Tree Construction (CTC) algorithm: a single tree is built based on a set of subsamples. New options are added to the J48 class to set the Resampling Method (RM) for the generation of samples to be used in the consolidation process.

  Recently, a new method has been added to determine the number of samples to be used in the consolidation process which guarantees the minimum percentage, the coverage value of the examples of the original sample to be contained by the set of built subsamples (Ibarguren et al., 2014).

- **J48 Graft**

  The J48 graft is a class for generating a grafted (pruned or unpruned) C4.5 decision tree. Decision tree grafting adds nodes to an existing decision tree with the objective of reducing prediction error (Webb, 1999). Hassan and El Fattah Hegazy (2015) worked on a big dataset which encompasses two data sources. The first is Learning Management System (LMS) while the second is social media related to that LMS. In order to do classification as well as a prediction for the learner's learning style LS, they employed WEKA to utilize J48Graft. The results showed that J48Graft was able to achieve a good rate which is at 89%.

- **Random Forest**

  Random Forest is a class for constructing a forest of random trees. It considers a combination of tree predictors whereby each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest (Breiman, 2001). In the work done by Singh et al., (2014), Random Forest produces the highest accuracy rate which is at 99.8% and this was employed on an imbalanced dataset to detect malicious users.

### 2.5 Creating Feature Set for YouTube

Feature extraction, feature construction, and feature selection are the major techniques used to prepare data before any knowledge extraction process can be performed. They are usually used to transform the initial representation of the data into a better representation that can be processed by existing data mining algorithms (Sia, Alfred, Yu, & Fun, 2012).

### 2.5.1 Repository of YouTube Channels

Each social media platform has unique characteristics that make it different from other platforms. The conducting studies in this field focus on analysis special platform and try to understand their content and how users engage with it. In order to do that, they need to extract specific data from a social media site related to their work. Nowadays, many ways to collect social media content was proposed, while the common one is using API provided by each social media provider. Unfortunately, the data that's available on a specific social media website are not available through the API (Abdesslem, Parris, & Henderson, 2012). There is another alternative method such as crawl the social media website with an automated tool that explores the website and collects data using HTTP requests (Zheng et al., 2015). This dissertation will be used Web extracting method in order to build a YouTube feature set.

### 2.5.2 Feature Construction: Edge Rank Algorithm

Feature construction process is very important when working with a real world feature set especially if it does not contain enough meaningful features for beneficial analysis (Freitas, 2001). The main goal of feature construction is to get new features which are able to improve classification task (Bermejo, Joho, Jose, & Villa, 2009). There are some feature construction methods that have been proposed in literature such as Decision Tree Related (Pagallo, 1989), Genetic Programming Related (Vafaie & De Jong, 1995), Inductive Logic Programming (Lavrač, Džeroski, & Grobelnik, 1991), and Annotation Based (Roth & Small, 2009). However, these methods have limitations such as overfitting, difficulty in comparison, and incorporating domain knowledge. Hence choosing the right method for feature construction becomes a problem (Sondhi, 2010). This dissertation employed the

three aspects of Edge Rank concept for feature construction leveraging from both user and user-behaviour features.

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). Therefore, this dissertation adopt this concept and implement it to understand the actual content of each post (video) over YouTube by constructing hybrid features, this is employed based on the three aspects Affinity, Weight, and Decay.

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 as shown in Equation 2.1 (Jeff, 2015).

$$EdgeRank = \sum U_e \times W_e \times D_e \quad \text{(Equation 2.1)}$$

Where, $U_e$ is illustrated the Affinity, the score between the viewing user and edge creator. While $W_e$ represent the Weight, the weight for this edge type such as comments, like, and shares. $D_e$ is shown the Decay, the decay factor based on how long age the edge was created.

### 2.5.3 Feature Selection

Feature Selection is a process of identifying subset from the input features that are relevant to a particular learning (or data mining) (Guyon & Elisseeff, 2003). Feature selection carries out tasks of removing the most irrelevant and redundant features from the feature set according to the class without incurring much loss of information (Bermejo et al., 2009).

According to Sondhi (2010), there are three methods for feature selection: filters, wrappers, and embedded method. In addition, there is also Correlation-based Feature Selection (CFS) (Hall, 1999).

#### 2.5.3.1 Filters

Filters are the process of selecting the feature subsets independent of the predictor. They essentially work as a data pre-processing step before a predictor is trained. Features ranking approaches, which is ranking individual feature using information theoretic or correlation criteria, then constructing a subset of high-scoring features fit in each category. The filters have an advantage in term of speed, they are faster than wrappers. However, a disadvantage is that the chosen subset may not be the best suited for the predictor to be used in the next step (Sondhi, 2010).

#### 2.5.3.1 Correlation-Based Feature Selection

CFS is the process of select a subset of features based on features worth evaluate. Whereas evaluates the worth of a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between them. The subsets of features that are highly correlated with the class while having low intercorrelation are preferred (Hall, 1999). According to Hall results, CFS can drastically reduce the dimensionality of data sets while maintain or improving the performance of learning compared with others methods, besides it is faster (Hall, 2000).

#### 2.5.3.2 Wrappers

Wrappers are another method of feature selection that use the learning method for prediction as a black box to select subsets feature. These methods divide the training set

into two sets: a train and validation set. For any given feature subset, the predictor is trained on the train set and tested on the validation set, where the accuracy of prediction based on the validation set is considered as the score of each feature subset. Thus, we would ultimately want to choose the highest scoring feature subset. Due to repeated train and test cycles for every feature subset, wrappers tend to be much more computationally intensive compared to filters. The goal usually is to traverse the feature space such that the number of subsets to be tested is minimized. An obvious advantage is that the chosen subset is tuned to the predictor (Sondhi, 2010).

#### 2.5.3.3 Embedded Methods

Embedded methods combine the process of feature selection and model learning. These methods are highly specific to the learning machine. Such methods are often fast and lead to accurate predictors. They are however not directly generalizable to any predictor (Sondhi, 2010).

## 2.6 Research Gap

Existing malicious user detection methods usually rely on some features extracted from the content-based approach. However, techniques in content classification are hard to be applied on video objects (Benevenuto et al., 2008). On the other hand, there is also a method based on features of the social network. This method is reported to be useful for malicious campaign detection (O'Callaghan et al., 2012). However, a study that integrates content-based features and social-network features is only able to detect 46% of spammers (Kiran, 2015). Moreover, studies in malicious users' detection, especially for video spam detection, detect malicious users based on the analysis of a single video that the user shared. This means if they classify a video as a spam video, then they will assume that the user is

malicious (Benevenuto et al., 2008). This may not be fair to the user. Hence, there is a need for a method that analyses overall channel details based on user and his behaviour.

### 2.7 Summary

This chapter shows a wide view of existing works and their limitations. Based on the examined literature, it is learned that the hybrid analysis becomes the standard approach in classifying malicious users. The low classification accuracy obtained using the approach has lead this study to investigate for new feature set to be used in malicious user detection. Hence, this dissertation focused on understanding the actual content of each post, leveraging from content features and social activity features. Furthermore, to achieve better results, this dissertation adopted the concept of Edge Rank to construct new features.

# CHAPTER THREE
# RESEARCH METHODOLOGY

The aim of this chapter is to present the procedures to classify YouTube channel owners into legitimate users or malicious users. This has been done by using features extracted and derived from channel content and users' engagement. In detail, Section 3.1 describes an overview of the proposed methodology. Then, Section 3.2 illustrates crawling strategies that were used to extract information from YouTube website. Data pre-processing method and feature construction process are discussed in Section 3.3 and 3.4 respectively. Classification methods that were used in classification phase are included in Section 3.5 while the evaluation phase is presented in Section 3.6. Finally, the chapter ends with a summary.

## 3.1 Introduction

YouTube offers video contents created and shared by users, and these users may be of legitimate or malicious (Chowdury et al., 2013). So, in order to distinguish them, this dissertation proposed a set of hybrid features. The new features have been used for understanding the actual content of specific YouTube channel. While the proposed features have been examined among a set of classifiers. Basically, the new features illustrate the integration of user-based, user-Behaviour and Edge Rank concept. While the main used of ERC is to construct a new set of hybrid features. Figure 3.1 illustrates the general steps of the undertaken methodology.

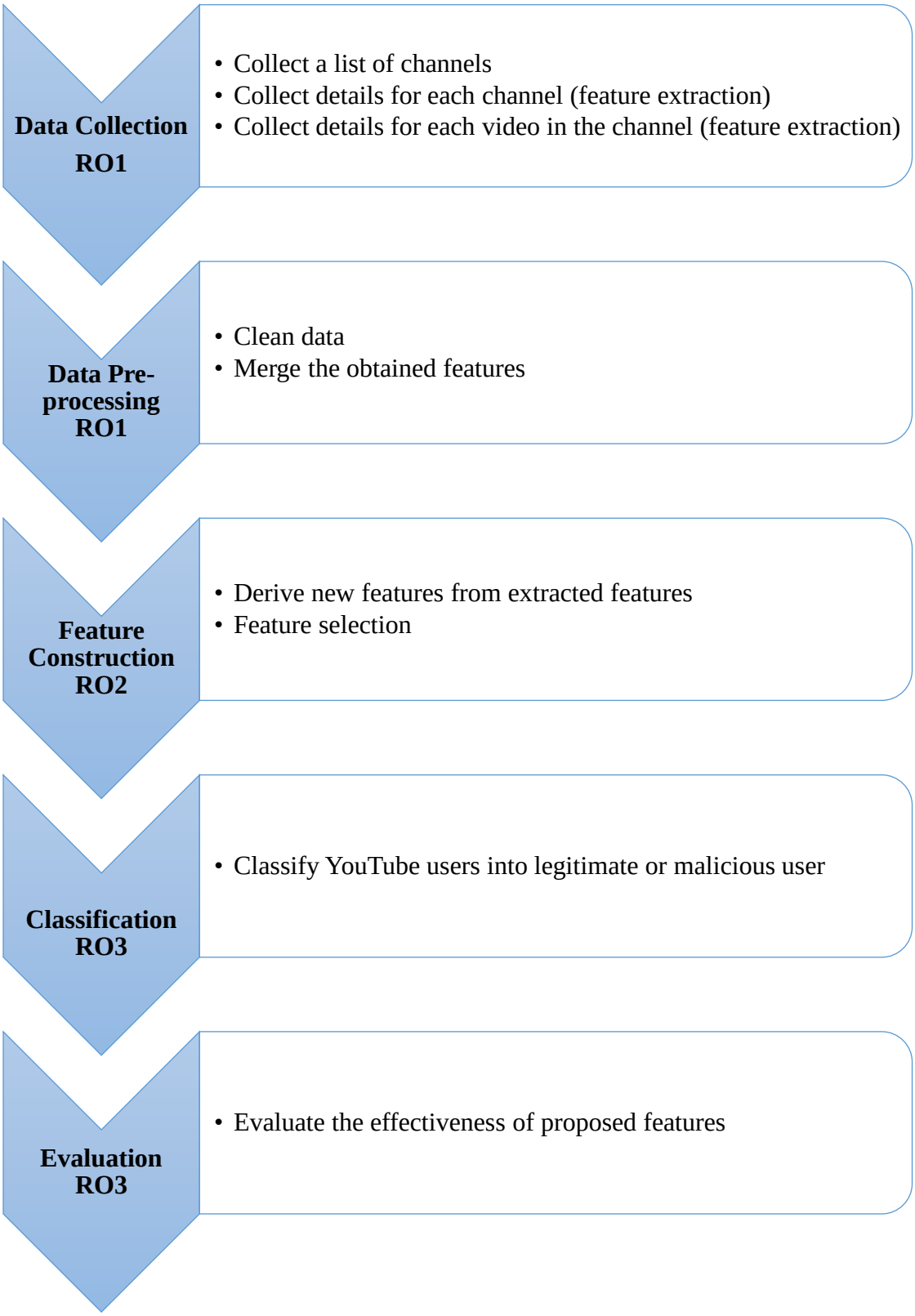


*Figure 3.1.* General Steps of Methodology

### 3.2 Data Collection

In order to collect YouTube data, this dissertation uses the Web Scraper (WSC) to extract data from web pages (Web Scraper, 2016). The crawling strategy inspects users with an account on YouTube and the employed crawling duration is of the period of four months as implemented by O'Callaghan et al. (2012), Tan, Guo, Chen, Zhang, and Zhao. (2013), and Alberto et al. (2015). The data collection process is divided into three phases as illustrated in Figure 3.2. The first phase involves the process of randomly crawling YouTube main page and picking up a list of channels addresses as shown in Figure 3.3. 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 this dissertation. The second phase focuses on crawling through the contents of the identified channel addresses and extract data on user profile as demonstrated in Figure 3.4. Last but not least, the third phase includes the process of scraping information on the video and creating a file for each channel. Figure 3.5 illustrates an example of crawling graph design for scraping videos details.

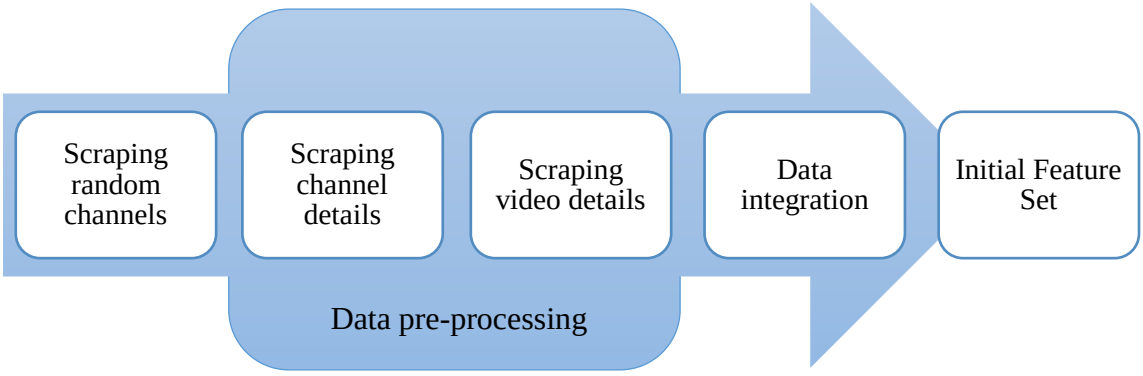


*Figure 3.2.* Process of Creating Initial Feature Set

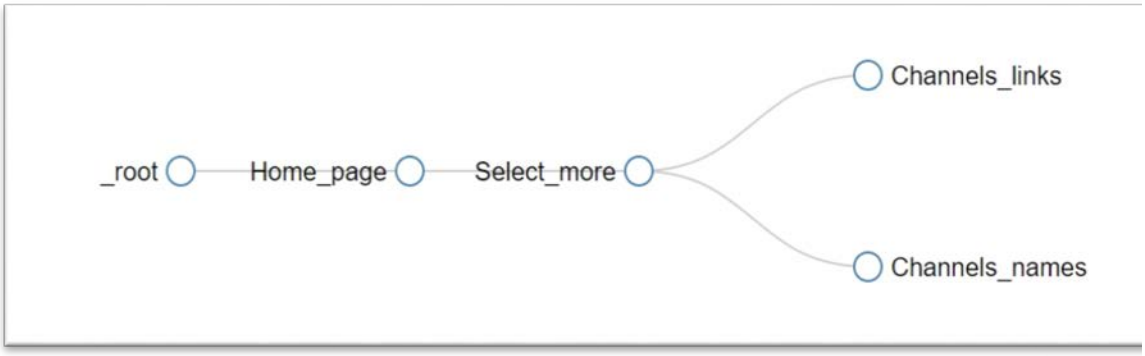


*Figure 3.3*. Crawling Graph for Scraping Channels List

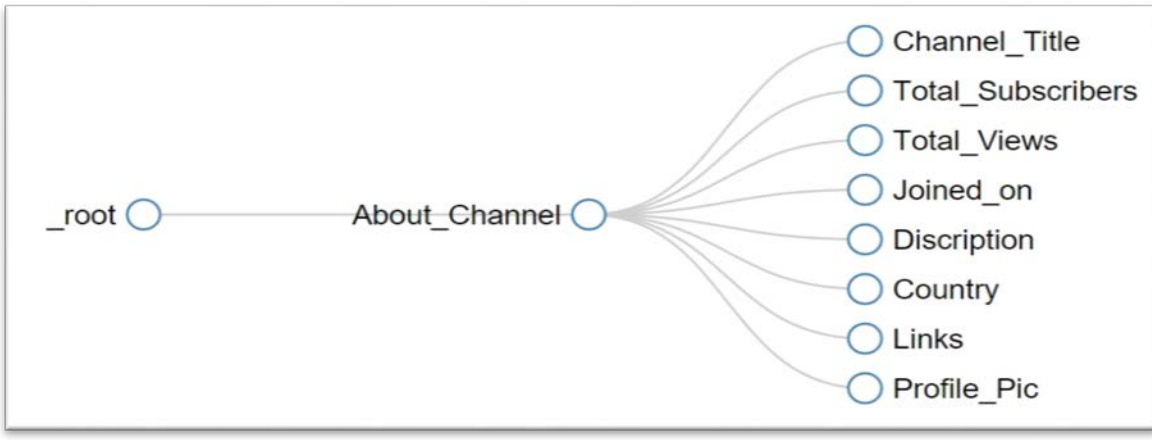


*Figure 3.4.* Crawling Graph for Scraping Channels Details

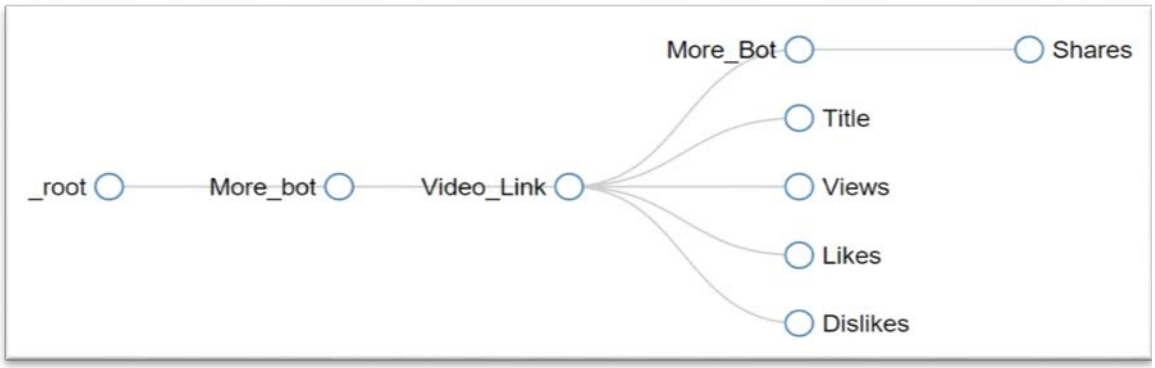


*Figure 3.5.* Crawling Graph for Scraping Videos Details

The obtained crawling data were arranged into three files, the first file stores a list of selected channels to be crawled with their address (URL) and date of scraping as shown in Figure 3.6. The second file store each channel details as demonstrated in Figure 3.7. The third one store each video details such as title, total video views, the number of likes/dislikes, and shared number as shown in Figure 3.8. As a result of the three files, the final file contains around sixteen different attributes that illustrate channel features.

| | A | B | C |
|---|---|---|---|
| 1 | Channel Name | Channel Link | Date of Collecting |
| 2 | Mohamed AlSalim \| | https://www.youtube.com/channel/UCd3PaZEcxNru29bA4--WL_A | 27/2/2016 |
| 3 | albasheer show | https://www.youtube.com/channel/UCjxrFnMg_scE7fkw_Ip0_yA | 29/2/2016 |
| 4 | Kimberly C. Sinclair | https://www.youtube.com/channel/UCbakp3km-hUazI0wBDZEiZQ/videos?sh | 29/2/2016 |
| 5 | AdeleVEVO | https://www.youtube.com/user/AdeleVEVO | 29/2/2016 |
| 6 | Discovery Documentary HD | https://www.youtube.com/channel/UCCwBMVm11X_XKCtYYoOR50w | 29/2/2016 |
| 7 | Amal Ramahy | https://www.youtube.com/user/AmalRamahiChef | 29/2/2016 |
| 8 | Mateusz M | https://www.youtube.com/channel/UCnJ-KJLPlRw90rGs_6XfmmQ | 29/2/2016 |

*Figure 3.6.* Sample of Channels Lists

| 1 | Channel_Name | Total_Subscribers | Totao_Views | Total_Ch_Videos | Ch_Joined_on |
|---|---|---|---|---|---|
| 2 | Mohamed AlSalim \| | 190350 | 58535481 | 62 | 14-Sep-14 |
| 3 | albasheer show | 290,167 | 35,532,789 | 100 | 27-Jan-14 |
| 4 | Kimberly C. Sinclair | 53 | 13,903 | 19 | 24-Jan-16 |
| 5 | AdeleVEVO | 10,840,601 | 3725951730 | 24 | 15-Oct-09 |
| 6 | Discovery Documentary HD | 2,330 | 736375 | 23 | 3-Jun-15 |
| 7 | Amal Ramahy | 21314 | 4,226,461 | 19 | 1-Jul-12 |
| 8 | Mateusz M | 708,499 | 134,244,707 | 21 | 11-Nov-11 |
| 9 | G8-ALLVIDEOS | 548 | 876,776 | 5 | 26-Jul-14 |

*Figure 3.7.* Sample of Channels Details

| B | C | D | E | F | G |
|---|---|---|---|---|---|
| Video_list-href | Video_Title | Viedo_views | Video_Likes | Video_Disli | Shares_no |
| https://www.youtube.com/watch?v=I | Pacific Ocean Paradise - Nature Docu | 86 | 0 | 0 | 0 |
| https://www.youtube.com/watch?v=\ | What If The Earth Suddenly STOPPED | 6,465 | 16 | 4 | 11 |
| https://www.youtube.com/watch?v=\ | Extreme Rare Lightning: Sprites (Edge | 1,262 | 5 | 2 | 1 |
| https://www.youtube.com/watch?v=4 | Decoding the Universe: The Great Ma | 1,122 | 9 | 1 | 4 |
| https://www.youtube.com/watch?v=( | The End of the Universe: Big Crunch, I | 88 | 0 | 0 | 0 |
| https://www.youtube.com/watch?v=( | Glacier National Park - Nature Docum | 58 | 1 | 0 | 0 |
| | | | 1254 | 540 | 1508 |

*Figure 3.8.* Sample of Videos Details

### 3.3 Data Pre-processing

After data being collected using web scraper tools as discussed in the previous section, it's now ready for the pre-processing stage. During this stage, the collected data are pre-processed by cleaning and preparing the data to be stored in one main file. Furthermore, the data have been examined to detect any missing data or mistakes during the capturing phase for each file separately. This process requires high effort and consume time as the data are stored in various files and thousands of records were crawled. To achieve this goal, the data pre-processing stage is arranged into three sub-processes. First, examine channels' files, followed by checking each videos in the channel. Then, in the third sub-process the main file that stores all channel information is created.

#### 3.3.1 Pre-processing of Channels File

Once channel's data being collected, it needs to be examined manually in order to detect any possible mistakes. During this stage, the researcher examined each channel online to retrieve missing data and add some features that can't be captured by the scraper. For instance, remove the "Views" term from channel view's column. Also, new features such as the existence of profile picture, background picture, discussion, playlist are included.

These new features cannot be captured during the first stage of collecting data. Figure 3.9 shows a sample of data in the pre-processing of channels file.

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| | Channel_ID | Channel_Name | tal_Subscrib | Totao_Views | Total_Ch_Videos | Ch_Joined_on | Ch_Discussion | Ch_Playlist |
| | UCd3PaZEcxNru29bA4--WL_A | Mohamed AlSalim \| | 190,350 | 58,535,481 | 62 | 14-Sep-14 | Yes | Yes |
| | UCjxrFnMg_scE7fkw_Ip0_yA | albasheer show | 290,167 | 35,532,789 | 100 | 27-Jan-14 | Yes | Yes |
| | UCbakp3km-hUazI0wBDZEiZQ | Kimberly C. Sinclair | 53 | Vie s 13903 | 19 | 24-Jan-16 | No | No |
| | UComP_epzeKzvBX156r6pm1Q | AdeleVEVO | 10,840,601 | 3,725,951,730 | 24 | 15-Oct-09 | Yes | Yes |
| | UCCwBMVm11X_XKCtYYoOR50w | Discovery Document | 2,33 | 736,375 | 23 | 3-Jun-15 | Yes | No |
| | UCRHd9bQqKDxJVmqOAdUq-CQ | Amal Ramahy | | 4,226,461 | 19 | 1-Jul-12 | Yes | Yes |
| | UCnJ-KJLPlRw90rGs_6XfmmQ | Mateusz M | | 34,244,707 | 21 | 11-Nov-11 | Yes | Yes |
| | UCVztq6eujLZjywe26qh8uqQ | G8-ALLVIDEOS | | 876,776 | 5 | 26-Jul-14 | No | No |
| | UCORp0Fxa5lHFuZfLHxfeLuQ | Basquiat Picasso | | 21,850,631 | 83 | 2-Sep-12 | Yes | Yes |
| | UCdbd4O5KnIHRWUTtpoF8slA | Stun_gravy | | 26,643,107 | 96 | 10-Nov-07 | Yes | Yes |
| | UC8cXcdMGLC8lO1E_l6qWcag | HaiDeR MaDLoL | 5,705 | 927,238 | 42 | 7-Apr-15 | Yes | No |
| | UCw3EildEnRRQTMBeFeUdJOQ | Tube Star Network | 9,794 | 162,082 | 35 | 24-Feb-13 | Yes | Yes |

Remove "View"

*Figure 3.9.* Sample of Data in Pre-processing of Channels File

### 3.3.2 Pre-processing of Videos Files

Each channel has many videos that are published by the owner and these videos are investigated using a web scraper as illustrated in Figure 3.5. The crawling graph shows the data collected from YouTube videos in each channel. During the scraping process, the crawler might miss capturing some features. This could appear when one page failed to be loaded or one feature or more was not accessible due to "Privacy Issue". For instance, the total number of shared for a specific video is available in the "Video statistic". If the scraper returns "Null value" as shown in Figure 3.10, this dissertation uses two ways to replace the "Null" value; mean based imputation and auto calculation using tool.

| | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | videos_linl | videos_linl | Video_Titl | Viedo_viev | Video_Like | Video_Disl | Shares_no | | | | |
| 2 | البشير شو الـ | https://wv | البشير شو الـ | 1,363,625 | 17,105 | 1,571 | 2,294 | | | | |
| 3 | Albasheer | https://wv | Albasheer | 612,793 | 8,058 | 391 | 881 | | | | |
| 4 | Albasheer | https://wv | Albasheer | 170,556 | 2,453 | 48 | 919 | | | | |
| 5 | Albasheer | https://wv | Albasheer | 865,695 | 13,473 | 733 | 1,192 | | | | |
| 6 | Albasheer | https://wv | Albasheer | 43,486 | 669 | 17 | null | | | | |
| 7 | albasher sl | https://wv | albasher sl | 40,042 | 652 | 14 | 35 | | | | |
| 8 | Albasheer | https://wv | Albasheer | 191,321 | 2,258 | 100 | 498 | | | | |
| 9 | Albasheer | https://wv | Albasheer | 1,017,064 | 13,626 | 504 | 1,716 | | | | |
| 10 | Albasheer | https://wv | Albasheer | 143,270 | 2,180 | 51 | null | | | | |

*Figure 3.10.* Missing Value in Videos File

#### 3.3.2.1 Mean Based Imputation

Imputation is a refilling method to represent the missing values with estimated ones. The process is to calculate the median for all non-missing values of that variable, then replace missing values with an average value (Acuña & Rodriguez, 2004). In this case, the new value estimates the average value of a total number of shares as shown in Equation 3.1. This method was used to estimate missing value if it was below than 20%. If the missing data is more than 20%, tools will be used to represent the values.

$$X = \frac{\sum_{i=0}^{n} Xi}{n} \qquad \text{(Equation 3.1)}$$

Where, X represent the missing value which it consider an average value and n illustrate instances number.

#### 3.3.2.2 Social Analytics Online Tool

Social analytic is an online tool used to estimate different activity over social media sites. It is available online as an extension can be installed in the browser. In our case, if the missing data was more than 20%, then tool were used to estimate the value. However, this tool requires time as it needs to check each of the collected videos. Figure 3.11 demonstrates the snapshot of using the social analytic tool.

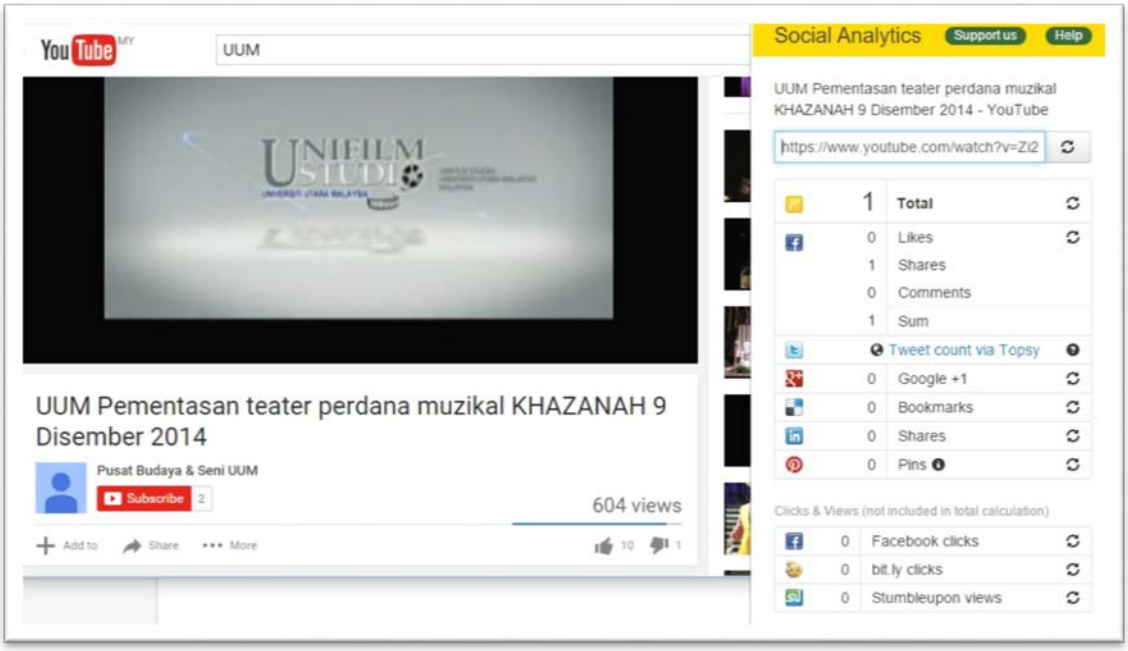


*Figure 3.11.* Social Analytics Tool

### 3.3.3 Data Integration

Once the data have been pre-proceesed, the obtained files need to be integrated. The process is performed by creating one file that combines all of the features contained in the previous files. Figure 3.12 illustrates sample of the YouTube repository based on the integration of User-based and User-behaviour features. In this integration file, another attribute is included which is the ‘Class’. The value of “Class” is either “False” if it is legitimate or “True” if it is malicious. A user is classified as either a legitimate or malicious based on human judgment (Zheng et al., 2015). A video is considered as a spam video if its content is unrelated to their title or the video does not have any content (Benevenuto et al., 2008; Google, 2016). The channel owner is considered as a malicious user if he publishes the aforementioned video more than once.

| A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ID | Total_Subscribers | Total_Views | Total_Ch_ | Ch_Joined_c | Scraped_Dat | Total_Videos_ | Total_Videos_L | Total_Videc | Ch_Dis | Ch_Pla | Ch_Desc | Ch_Cou | Ch_Links | Ch_Profile | Ch_Backg | Class |
| UCd3PaZEcv | 190,350 | 58,535,481 | 62 | 14-Sep-14 | 27-Feb-16 | 842,646 | 364,943 | 29,283 | Yes | Yes | Yes | Yes | Yes | Yes | Yes | FALSE |
| UCjxrFnMg_ | 290,167 | 35,532,789 | 100 | 27-Jan-14 | 29-Feb-16 | 97,461 | 586,158 | 26,337 | Yes | Yes | Yes | Yes | Yes | Yes | Yes | FALSE |
| UCbakp3km | 53 | 13,903 | 19 | 24-Jan-16 | 29-Feb-16 | 2 | 55 | 126 | No | No | No | No | No | No | No | TRUE |
| UComP_epz | 10,840,601 | ######### | 24 | 15-Oct-09 | 29-Feb-16 | 15,742,117 | 19,580,695 | 702,417 | Yes | Yes | Yes | Yes | Yes | Yes | Yes | FALSE |
| UCCwBMVn | 2,330 | 736,375 | 23 | 3-Jun-15 | 29-Feb-16 | 1,505 | 1,254 | 540 | Yes | No | Yes | No | No | No | No | FALSE |
| UCRHd9bQc | 21,314 | 4,226,461 | 19 | 1-Jul-12 | 29-Feb-16 | 20,196 | 18,462 | 1,794 | Yes | Yes | Yes | No | Yes | Yes | Yes | FALSE |
| UCnJ-KJLPlR | 708,499 | 134,244,707 | 21 | 11-Nov-11 | 29-Feb-16 | 148,363 | 1,130,754 | 20,207 | Yes | Yes | Yes | Yes | Yes | Yes | Yes | FALSE |
| UCVztq6eujl | 548 | 876,776 | 5 | 26-Jul-14 | 29-Feb-16 | 89 | 1,674 | 149 | No | No | Yes | No | Yes | Yes | Yes | FALSE |

*Figure 3.12.* Sample of Repository of YouTube Features

### 3.4 Features Construction

Once, the repository of YouTube features is available, new features can be constructed. In this dissertation, new features are derived based on the concept of affinity, weight, and decay as discussed in Section 2.5.2. This includes the construction of "age of channel", "view rate" and "average upload" features. The final file contains more than 30 different features that are used later in the classification process. Figure 3.13 shows part of the initial feature set while the detail implementation of the feature construction is presented in Chapter 4.

| A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Channel_II | Total_Subs | Total_View | Total_Ch_ | Ch_Age | Avrage_Up | Subscribe_ | Subscribe_ | Subscribe_ | View_Rate | Total_Vide | Total_Vide | Total_Vide | Share_Rat | Share_Rat | Share_Rat | Like_Rate_ |
| UCd3PaZE | 190,350 | ######### | 62 | 531 | 0.116761 | 3070.161 | 358.4746 | 0.003252 | 110236.3 | 842,646 | 364,943 | 29,283 | 0.014395 | 1586.904 | 13591.06 | 0.006235 |
| UCjxrFnMg | 290,167 | ######### | 100 | 763 | 0.131062 | 2901.67 | 380.2975 | 0.008166 | 46569.84 | 97,461 | 586,158 | 26,337 | 0.002743 | 127.7339 | 974.61 | 0.016196 |
| UCbakp3k | 53 | 13,903 | 19 | 36 | 0.527778 | 2.789474 | 1.472222 | 0.003812 | 386.1944 | 2 | 55 | 126 | 0.000144 | 0.055556 | 0.105263 | 0.003956 |
| UComP_e | ######## | ######## | 24 | 2328 | 0.010309 | 451691.7 | 4656.616 | 0.002909 | 1600495 | ######## | ######## | 702,417 | 0.004225 | 6762.078 | 655921.5 | 0.005255 |
| UCCwBMV | 2,330 | 736,375 | 23 | 271 | 0.084871 | 101.3043 | 8.597786 | 0.003164 | 2717.251 | 1,505 | 1,254 | 540 | 0.002044 | 5.553506 | 65.43478 | 0.001703 |
| UCRHd9bC | 21,314 | 4,226,461 | 19 | 1338 | 0.0142 | 1121.789 | 15.92975 | 0.005043 | 3158.79 | 20,196 | 18,462 | 1,794 | 0.004778 | 15.09417 | 1062.947 | 0.004368 |
| UCnJ-KJLP | 708,499 | ######### | 21 | 1571 | 0.013367 | 33738.05 | 450.986 | 0.005278 | 85451.75 | 148,363 | 1,130,754 | 20,207 | 0.001105 | 94.43857 | 7064.905 | 0.008123 |
| UCVztq6e | 548 | 876,776 | 5 | 583 | 0.008576 | 109.6 | 0.939966 | 0.000625 | 1503.904 | 89 | 1,674 | 149 | 0.000102 | 0.152659 | 17.8 | 0.001909 |
| UCORp0Fx | 86,040 | ######### | 83 | 1276 | 0.065047 | 1036.627 | 67.42947 | 0.003938 | 17124.32 | 60,174 | 181,828 | 3,778 | 0.002754 | 47.15831 | 724.988 | 0.008321 |
| UCdbd4O5 | 65,507 | ######## | 96 | 3034 | 0.031641 | 682.3646 | 21.59097 | 0.002459 | 8781.512 | 18,449 | 396,486 | 4,611 | 0.000692 | 6.080751 | 192.1771 | 0.014881 |
| UC8cXcdM | 5,705 | 927,238 | 42 | 330 | 0.127273 | 135.8333 | 17.28788 | 0.006153 | 2809.812 | 138 | 4,042 | 154 | 0.000149 | 0.418182 | 3.285714 | 0.004359 |

*Figure 3.13.* Sample of Initial Feature set

### 3.4.1 Feature Selection

Feature selection is the process of choosing a subset of relevant features, then using only that subset for classification task (Guyon & Elisseeff, 2003). While the feature set contains many features/attributes that are either redundant or irrelevant, using feature selection method remove them without incurring much loss of information. For this purpose, this dissertation employed the Waikato Environment for Knowledge Analysis (Weka) to evaluate the attributes using CFS method to score feature subsets (Witten et al., 1999) while using "GreedyStepwise" as a searching method. Besides than utilizing minimal computational time, the CFS also reduces the dimensionality of data sets while maintaining or improving the performance of learning compared with other methods (Hall, 2000).

## 3.5 Classification

The aim of this dissertation is to classify whether a YouTube user is malicious or legitimate. The classification process is performed using classification algorithms included in Weka. In particularly, this dissertation used 22 classifiers in order to perform the process of classification such as K*, Random Committee, Random Forest, Bayes Net, JRip, PART, LibLINEAR, Multilayer Perceptron, NNge, SMO, Hoeffding Tree, FT, J48graft, Naïve Bayesian, Logistic, Multi Class Classifier, Attribute Selected Classifier, AdaBoostM1, J48, Decision Tree, J48 Consolidated, LibSVM. All of these classifiers were fed with the same feature set obtained in the aforementioned stages.

## 3.6 Evaluation

To evaluate the proposed features, this dissertation compares the effectiveness of the hybrid features versus User-based features and User-behaviour features (Cao & Caverlee, 2015). Additionally, this dissertation determine the effectiveness across a set of classification

algorithms in Weka such as K*, Random Committee, Random Forest, Bayes Net, JRip, PART, LibLINEAR, Multilayer Perceptron, NNge, SMO, Hoeffding Tree, FT, J48graft, Naïve Bayesian, Logistic, Multi Class Classifier, Attribute Selected Classifier, AdaBoostM1, J48, Decision Tree, J48 Consolidated, LibSVM. In detail, the comparison was done based on classification accuracy as an evaluation metric (Singh et al., 2014).The accuracy calculation is as shown in Equation 3.2 (Chawla, 2005).

$$accuracy = \frac{a + d}{a + b + c + d} \quad \text{(Equation 3.2)}$$

Where, “a” is illustrate the total number of true positive predictions, “d” is illustrated the total number of true negative predictions. Furthermore, “c, b” are illustrating the total number of false positive and false negative prediction respectively. Therefore, the total number of correct predictions is (a + d) while the total number of incorrect predictions is (c + b).

In addition, to ensure the results were not obtained by chance, this dissertation performed statistical analysis (i.e T-Test) on the result obtained using different feature sets (Dong & Wang, 2009).

### 3.7 Summary

The process of creating a new feature set that is able to be used in classifying YouTube channel owners has been presented. In general, this chapter includes on how to collect, pre-processed and construct features based on user profile, behaviour and also the Edge Rank concept. This is followed by the phases of classification and evaluation.

# CHAPTER FOUR
# HYBRID FEATURES

This chapter details the procedures of constructing new features based on data extracted from YouTube. The new features represents a combination of channel content and users' engagement details. In particular, these features are of the integration of User-based and User-behaviour. Section 4.1 includes a brief information on procedures of feature set construction. Then, YouTube features are explained in Section 4.2. Section 4.3 illustrates YouTube features construction based on Edge Rank concept while Section 4.4 and 4.5 details the utilized feature sets. Finally, the chapter ends with a summary.

## 4.1 Introduction

The process of creates a new feature set requires some steps in order to achieve the objectives of this dissertation. Figure 4.1 illustrates the general steps of creating a new YouTube feature set.

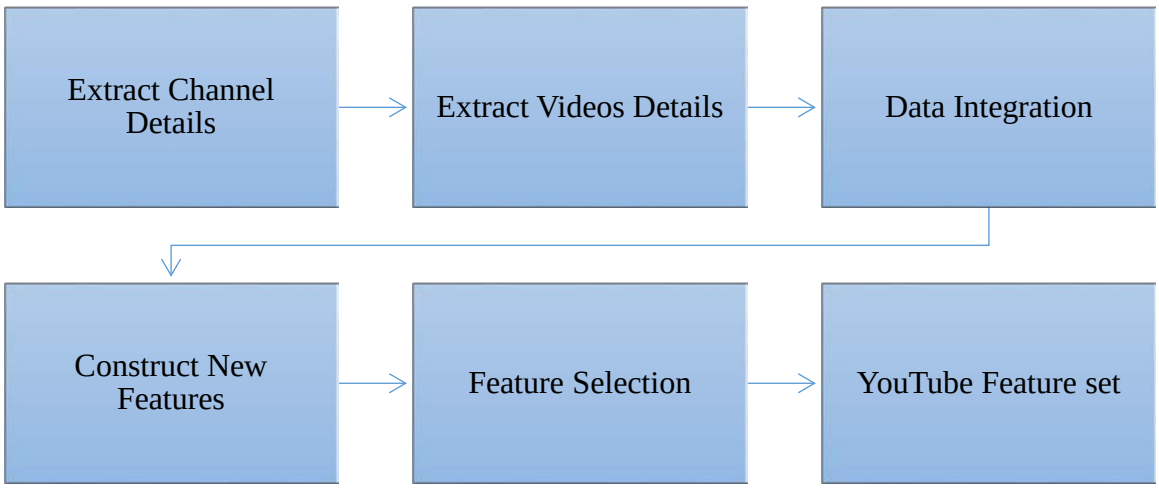


*Figure 4.1.* General Steps of Creating YouTube Channels' Feature set

## 4.2 YouTube Repository

This dissertation creates the initial repository of YouTube features by employing features that were reported in the literature; user-based and user behaviour. The included features are shown in Table 4.1. The FeatureSet-UB consist of four features while FeatureSet-UBA includes three features.

Table 4.1

*Traditional Features Extracted from YouTube*

| User-based Features (FeatureSet-UB) | User-Behaviour Features (FeatureSet-UBA) |
|---|---|
| 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 4.2

*New Features Extracted from YouTube*

| User-based Features (FeatureSet-UB) | User-Behaviour Features (FeatureSet-UBA) |
|---|---|
| Joined Date | Total Videos Shared |
| Discussions | |
| Playlist | |
| Description | |
| Country | |
| Links | |
| Profile Picture | |
| Background Picture | |

As this dissertation uses the scraper tool to extract information from YouTube, there are additional features that are found to be useful in representing user profile and user behaviour. Data in Table 4.2 illustrates the new features of FeatureSet-UB and FeatureSet-UBA respectively. When data in Table 4.1 and Table 4.2 are combined, there are 12

features in FeatureSet-UB and 4 features in FeatureSet-UBA. On the other hand, data in Table 4.3 and 4.4 show the characteristics of both feature sets.

Table 4.3

*Summary of YouTube User-Based FeatureSet (FeatureSet-UB)*

| Characteristics | YouTube User-Based FeatureSet-UB |
|---|---|
| **Sample Period** | 20-04-2016/20-08-2016 |
| **# of videos** | 30,621 |
| **# of distinct users** | 500 |
| **# of features** | 12 |

Table 4.4

*Summary of YouTube User-Activity FeatureSet (FeatureSet-UBA)*

| Characteristics | YouTube User-Based FeatureSet-UB |
|---|---|
| **Sample Period** | 20-04-2016/20-08-2016 |
| **# of videos** | 30,621 |
| **# of distinct users** | 500 |
| **# of features** | 4 |

### 4.3 YouTube Features Construction

As discussed before in Section 2.5.2, Edge Rank algorithm gives a score based on Affinity, Weight, and Decay. Table 4.5 demonstrates the proposed features based on Edge Rank concept.

Table 4.5

*YouTube Constructed Features based on Edge Rank Concept*

| 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 |

| Affinity | Weight | Decay |
|---|---|---|
| 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 |

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 user's engagements. Likewise, the Decay concept shows the value of each event based on event's age. Table 4.6 illustrates the equations for the proposed features and these are derived using data in Table 4.1 and 4.2. In total, there are 16 features included as FeatureSet-ER

Table 4.6

*Equations for YouTube based on Edge Rank Aspects*

| | Data Driven Name | Equation | |
|---|---|---|---|
| **Affinity** | Channel average upload | $x = \frac{\sum \text{Channel Videos}}{\text{Joined Date} - \text{Scraped Date}}$ | (Equation 4.1) |
| | Subscriber rate based on total views | $x = \frac{\text{Channel Subscriber}}{\text{Channel Views}}$ | (Equation 4.2) |
| | View rate based on total videos number | $x = \frac{\sum \text{Videos Share}}{\sum \text{Channel Videos}}$ | (Equation 4.3) |
| | Subscriber rate based on total videos number | $x = \frac{\text{Channel Subscriber}}{\sum \text{Channel Videos}}$ | (Equation 4.4) |
| | Share rate based on total views | $x = \frac{\sum \text{Videos Share}}{\text{Channel Views}}$ | (Equation 4.5) |
| | Share rate based on total videos number | $x = \frac{\sum \text{Videos Share}}{\sum \text{Channel Videos}}$ | (Equation 4.6) |

| | Data Driven Name | Equation | |
|---|---|---|---|
| Weight | Like rate based on total views | $x = \frac{\sum \text{Channel Likes}}{\text{Channel Views}}$ | (Equation 4.7) |
| | Dislike rate based on total views | $x = \frac{\sum \text{Channel Dislikes}}{\text{Channel Views}}$ | (Equation 4.8) |
| | Like rate based on total videos number | $x = \frac{\sum \text{Channel Likes}}{\sum \text{Channel Videos}}$ | (Equation 4.9) |
| | Dislike rate based on total videos number | $x = \frac{\sum \text{Channel Dislikes}}{\sum \text{Channel Videos}}$ | (Equation 4.10) |
| Decay | Channel age | $x = \text{Joined Date} - \text{Scraped Date}$ | (Equation 4.11) |
| | Subscriber rate based on channel age | $x = \frac{\text{Channel Subscriber}}{\text{Joined Date} - \text{Scraped Date}}$ | (Equation 4.12) |
| | View rate based on channel age | $x = \frac{\text{Channel Views}}{\text{Joined Date} - \text{Scraped Date}}$ | (Equation 4.13) |
| | Share rate based on channel age | $x = \frac{\sum \text{Videos Share}}{\text{Joined Date} - \text{Scraped Date}}$ | (Equation 4.14) |
| | Like rate based on channel age | $x = \frac{\sum \text{Channel Likes}}{\text{Joined Date} - \text{Scraped Date}}$ | (Equation 4.15) |
| | Dislike rate based on channel age | $x = \frac{\sum \text{Channel Dislikes}}{\text{Joined Date} - \text{Scraped Date}}$ | (Equation 4.16) |

| | Channel_ID | Channel_Name | Total_Subscribers | Total_Views | Total_Ch_Videos | Ch_Joined_on | Scraped_Date | Ch_Age | Avrage_Upload |
|---|---|---|---|---|---|---|---|---|---|
| 0 | UC9tpEi6q77wuQ_6ImEYQqyg | freedoor8 | 2,217 | 2,029,492 | 23 | 8-Jan-11 | 13-Mar-16 | 1891 | 0.012162877 |
| 1 | UCAUtMll4-Bi2sWnbqtMVrTg | scaffoal | 19,334 | 3,338,696 | 42 | 22-Aug-07 | 14-Mar-16 | 3127 | 0.013431404 |
| 2 | UCB7BryuXaMe1pUMznYAq4Jg | MotivationGrid | 329,021 | 47,448,636 | 27 | 24-Oct-13 | 14-Mar-16 | 872 | 0.030963303 |
| 3 | UCQu40D09NBpnvnqPjTW1hAw | Box Office | 171 | 157,346 | 33 | 29-Jun-15 | 15-Mar-16 | 260 | 0.126923077 |
| 4 | UCHINpD9tTZc6fq3wVFcU3ig | YunaVEVO | 74,739 | 9,284,846 | 9 | 30-Jun-11 | 15-Mar-16 | 1720 | 0.005232558 |
| 5 | UCznHh6gk3ewFhaTZtdQFWWw | AsapTHOUGHT | 645,885 | 25,801,022 | 93 | 4-Jan-14 | 15-Mar-16 | 801 | 0.116104869 |
| 6 | UCsYS0EXyoQEnKQkr1TPJHew | Joanne cKee | 31 | 8,018 | 48 | 26-Feb-16 | 15-Mar-16 | 18 | 2.666666667 |
| 7 | UC9Ya7yz4uaEGLFQEUNqsC3Q | Deadpool Full Movie | 398 | 328 | 5 | 15-Dec-15 | 17-Mar-16 | 93 | 0.053763441 |
| 8 | UCHkj014U2CQ2Nv0UZeYpE_A | JustinBieberVEVO | 20,000,542 | 9,894,135,926 | 119 | 25-Sep-09 | 18-Mar-16 | 2366 | 0.050295858 |
| 9 | UC0VOyT2OCBKdQhF3BAbZ-1g | ArianaGrandeVevo | 10,266,487 | 3,433,117,544 | 46 | 21-Oct-10 | 18-Mar-16 | 1975 | 0.023291139 |
| 0 | UCi-SIlZFFRKex7vpOVb1UOw | **WLive16** | 14,470 | 968,779 | 12 | 31-May-15 | 18-Mar-16 | 292 | 0.04109589 |
| 1 | UCYGneKXGKoU5JakC-Rsbv1w | EnormousVIDS | 41,785 | 9,222,857 | 5 | 19-May-15 | 18-Mar-16 | 304 | 0.016447368 |
| 2 | UCRKwxkAaQWPhbLuE1KDbyeg | HoobastankVEVO | 320,873 | 259,667,557 | 22 | 11-May-09 | 18-Mar-16 | 2503 | 0.008789453 |
| 3 | UCM9UCLBfAssegYZXRvHXCEg | Ultimate Videos | 26,597 | 25,261,422 | 11 | 25-Aug-08 | 19-Mar-16 | 2763 | 0.00398118 |
| 4 | UC31TgQ6y2qdgkyaBL77m3zA | Carl Toon | 12,580 | 5,228,498 | 2 | 4-Nov-07 | 19-Mar-16 | 3058 | 0.000654022 |
| 5 | UC4v2tQ8GqP0RbmAzhp4IFkQ | April Wilkerson | 208,615 | 15,532,482 | 109 | 23-Sep-12 | 21-Mar-16 | 1275 | 0.085490196 |
| 6 | UCRix1GJvSBNDpEFY561eSzw | Laura Kampf | 14,865 | 348,996 | 26 | 15-Oct-15 | 21-Mar-16 | 158 | 0.164556962 |
| 7 | UCtaykeSsGhtn2o2BsPm-rsw | Make Something | 136,779 | 7,510,723 | 89 | 26-Jan-13 | 21-Mar-16 | 1150 | 0.077391304 |
| 8 | UC_MFUMilp949SWNM-9Q2opg | The Fox | 43,394 | 37,709,602 | 37 | 25-Sep-15 | 21-Mar-16 | 178 | 0.207865169 |
| 9 | UCpPZggubTs5NvcMCHfRCVKw | Echosmith | 567,956 | 131,534,756 | 98 | 17-Aug-10 | 21-Mar-16 | 2043 | 0.047968674 |

*Figure 4.2.* Part of Initial YouTube Feature Set

### 4.4 YouTube Features Selection

When all the feature sets are combined (i.e FeatureSet-UB, FeatureSet-UBA, and FeatureSet-ER), there exist around of 30 features. Hence, this dissertation investigates if the size of feature set could be reduced. This is realized using the feature selection offered in Weka The Correlation Feature Selection (CFS) measure evaluates subsets of features and in detail, the 'CfsSubsetEval' attribute evaluator have been used along with 'GreedyStepwise' as a searching method.

Before starting the evaluation process, some features need to be removed from the initial feature set such as (Channel Name, Channel Join Date, Channel Scraping Date) because it does not have any effect as it exists. First, the initial feature set that includes the data for features from FeatureSet-H needs to be formatted in an acceptable format. Then, once the file is loaded, it's necessary to configure Weka based on attribute selection. Next, select the required evaluator and searching method as stated before in Section 3.4.1. The feature selection results showed that only thirteen features were evaluated as best subset features (known as FeatureSet-HF). Table 4.7 illustrates the outcome of the feature selection process.

Table 4.7

*List of Features in FeatureSet-H and FeatureSet-HF*

| No. | Initial FeatureSet | Data Source | Final FeatureSet |
|---|---|---|---|
| 1 | Channel ID | User Based | Channel ID (address) |
| 2 | Total Subscribers | | |
| 3 | Total Channel Videos | | |
| 4 | Channel Discussion | | *Channel Discussion* |
| 5 | Channel Playlist | | *Channel Playlist* |

| No. | Initial FeatureSet | Data Source | Final FeatureSet |
|---|---|---|---|
| 6 | Channel Description | | |
| 7 | Channel Country | | |
| 8 | Channel Links | | *Channel Links* |
| 9 | Channel Profile Picture | | |
| 10 | Channel Background Picture | | |
| 11 | Total Views | User Behaviour | |
| 12 | Total Videos Shared | | |
| 13 | Total Videos Likes | | |
| 14 | Total Videos Dislikes | | |
| 15 | Channel Age | Edge Rank Concept | *Channel Age* |
| 16 | Channel Average Upload | | *Channel Average Upload* |
| 17 | Subscriber Rate Based on Total Videos Number | | |
| 18 | Subscriber Rate Based on Channel Age | | |
| 19 | Subscriber Rate Based on Total Views | | *Subscribe Rate Based on Total Views* |
| 20 | View Rate Based on Channel Age | | *View Rate Based on Channel Age* |
| 21 | View Rate Based on Total Videos Number | | |
| 22 | Share Rate Based on Total Views | | *Share Rate based on Total Views* |
| 23 | Share Rate Based on Channel Age | | |
| 24 | Share Rate Based on Total Videos Number | | *Share Rate based on Total Videos Number* |
| 25 | Like Rate Based on Total Views | | *Like Rate based on Total Views* |
| 26 | Like Rate Based on Channel Age | | *Like Rate Based on Channel Age* |
| 27 | Like Rate Based on Total Videos Number | | *Like Rate based on Videos Number* |
| 28 | Dislike Rate Based on Total Views | | *Dislike rate based on total views* |
| 29 | Dislike Rate Based on Channel Age | | |

| No. | Initial FeatureSet | Data Source | Final FeatureSet |
|---|---|---|---|
| 30 | Dislike Rate Based on Total Videos Number | | |
| 31 | Class | Manual Judgement | Class |

In brief, the final feature set has only thirteen features excluding channel ID and the Class. Data in Table 4.7 showed that the most significant features are the ones constructed from Edge Rank concept. Table 4.8 illustrates a summary of YouTube feature set. Furthermore, the final feature set represents FeatureSet-HF which is the hybrid feature set after feature selection process.

Table 4.8

*Summary of YouTube Channels' FeatureSet*

| Characteristics | FeatureSet-UB | FeatureSet-UBA | FeatureSet-ER | FeatureSet-H | FeatureSet-HF |
|---|---|---|---|---|---|
| **Sample Period** | 20-04-2016/20-08-2016 | | | | |
| **# of videos** | 30,621 | | | | |
| **# of distinct users** | 500 | | | | |
| **# of legitimate users** | 366 | | | | |
| **# of malicious users** | 134 | | | | |
| **# of features** | 12 | 4 | 16 | 30 | 13 |

## 4.5 Summary

This chapter has presented the process in constructing the feature set. The outcome of the process are of feature sets which includes FeatureSet-UB that represents features of a user, FeatureSet-UBA that includes information on user behaviour, FeatureSet-ER that includes hybrid features derived based on Edge Rank, FeatureSet-H that combines the earlier three sets and FeatureSet-HF that contains the relevant features. Later in Chapter 5, results of classification by various classifiers are presented.

# CHAPTER FIVE
# RESULT

This chapter presents the evaluation results of the proposed feature set. In the undertaken experiments, the evaluation was performed between single approach feature set and hybrid approach features. Moreover, the effectiveness features derived from Edge Rank concept was also included. Then, Section 5.3 illustrates the results discussion while the summary is presented at the end of the chapter.

## 5.1 Experiment and Results

This section examined the proposed features by employing 22 different classifiers namely K*, Random Committee, Random Forest, Bayes Net, JRip, PART, LibLINEAR, Multilayer Perceptron, NNge, SMO, Hoeffding Tree, FT, J48graft, Naïve Bayesian, Logistic, Multi Class Classifier, Attribute Selected Classifier, AdaBoostM1, J48, Decision Tree, J48 Consolidated and LibSVM. Two testing methods were used; percentage split and cross validation (Hu et al., 2014). The split percentage method divides data into two parts with a specific portion, normally the largest portion goes for training and the balance is assumed to unseen data. On the other hand, the cross-validation method used the whole data for training while its data split into different fold. By using different testing methods and different data portion, the results are verified to get a sense of how likely it is effective. Furthermore, by using different number of training samples (i.e different data proportion), it helps to avoid bias brought by the sizes of the training data. (Hu et al., 2014; Hu et al., 2015; Burnap et al., 2015). Experiments in this dissertation utilize 5 feature sets that were created based on single approach or hybrid approach. The experiment starts by analysing the performance of single approach through FeatureSet-UB and FeatureSet-UBA that

represent user-based approach and user-behaviour approach respectively. Then, the analysis was done on the performance of hybrid approach using FeatureSet-H, as well as FeatureSet-HF that represents the hybrid features after the process of feature selection. Similarly, the experiment also includes FeatureSet-ER that contains the 16 features constructed based on Edge Rank concept. In the statistical test, this dissertation considers 95% as the threshold of confidence level (Alberto et al., 2015).

### 5.1.1 Experiments using Percentage Split

The testing experiment for percentage split considers various data proportion; 70%-30%, 80%-20%, and 90%-10% (Hu et al., 2014). The bigger portion is for training while the smaller one includes the testing data. Table 5.1 - 5.3 summarize the performance of the proposed features across 22 classification algorithms for the 70-30% data proportion. Figure 5.1 – 5.3 visualize the performance statistics laid out by Table 5.1 – 5.3, respectively.

Table 5.1

*Classification Accuracy (%) for Experiment of Data Split into 70:30*

| Classifier | FeatureSet -UB | FeatureSet -UBA | FeatureSet -ER | FeatureSet -H | FeatureSet -HF |
|---|---|---|---|---|---|
| Bayes Net | 92.66 | 25.33 | 95.33 | 96.66 | 98 |
| Naïve Bayesian | 92.66 | 30.66 | 95.33 | 96 | 98 |
| LibLINEAR | 92 | 95.33 | 94.66 | 96.66 | 94.66 |
| LibSVM | 75.33 | 75.33 | 94.66 | 94 | 94.66 |
| Logistic | 64.66 | 93.33 | 94.66 | 95.33 | 96.66 |
| Multilayer Perceptron | 89.33 | 94 | 94.66 | 96.66 | 94.66 |
| SMO | 95.33 | 94.66 | 94.66 | 96.66 | 94.66 |

| Classifier | FeatureSet -UB | FeatureSet -UBA | FeatureSet -ER | FeatureSet -H | FeatureSet -HF |
|---|---|---|---|---|---|
| K* | 92.66 | 98 | 98 | 98.66 | 95.33 |
| AdaBoost M1 | 93.33 | 95.33 | 95.33 | 95.33 | 96.66 |
| Attribute Selected Classifier | 94 | 75.33 | 96 | 95.33 | 96 |
| Multi Class Classifier | 64.66 | 93.33 | 94.66 | 95.33 | 96.66 |
| Random Committee | 78 | 75.33 | 96 | 98 | 96 |
| Decision Tree | 92.66 | 98 | 95.33 | 95.33 | 95.33 |
| JRip | 93.33 | 96 | 95.33 | 96.66 | 94 |
| NNge | 87.33 | 96.66 | 95.33 | 96.66 | 95.33 |
| PART | 94 | 75.33 | 95.33 | 96.66 | 96 |
| FT | 94 | 95.33 | 95.33 | 96 | 95.33 |
| Hoeffding Tree | 92 | 75.33 | 90.66 | 96 | 96.66 |
| J48 | 94 | 75.33 | 95.33 | 95.33 | 95.33 |
| J48graft | 94 | 75.33 | 95.33 | 96 | 96 |
| J48 Consolidated | 93.33 | 75.33 | 94.66 | 94.66 | 95.33 |
| Random Forest | 78 | 98.66 | 96.66 | 97.33 | 96.66 |
| **Average** | **88.05** | **82.14** | **95.14** | **96.14** | **95.81** |

The analysis begins by examining the performance of FeatureSet-ER against FeatureSet-UB and FeatureSet-UBA. Based on the results, it is learned that the 16 features included in FeatureSet-ER give better insight on classifying malicious users as half of the classifiers' produce higher accuracy with average rate at 95%. For example, for the Bayes Net classifier, using FeatureSet-UB, the obtained accuracy was only 92.66 while FeatureSet-UBA performed worst by only achieving 25.33%. But, when the classifier was fed with FeatureSet-ER, the accuracy has increased to 95.33%. On the other hand, only three of the classifiers (K*, AdaBoost M1, and FT) have produced the same accuracy while using

FeatureSet-UB, UBA, and ER. The ER features were of useful as 68% of the classifiers obtained accuracy that is higher than 95%. However, there are seven classifiers that obtained lower accuracy rate when fed with FeatureSet-ER namely LibLINEAR, SMO, Decision Tree, JRip, Hoeffding Tree, and Random Forest.

As FeatureSet-ER produces better accuracy, its performance is compared against FeatureSet-H that combines all features from FeatureSet-UB, FeatureSet-UBA, and FeatureSet-ER. Experiments showed that the proposed FeatureSet-H is equal or better than using FeatureSet-ER with average rate at 96.14%. Classification results stated that around 73% of the classifiers produce higher accuracy. Moreover, around 91% of the classifiers achieved classification accuracy over 95% as shown in Figure 5.1. Only 2 out of the 22 classifiers had a drop in their accuracy.

As the hybrid feature set contains a large number of features (i.e 30), this dissertation performs feature selection to reduce the size of FeatureSet-H. This feature set (i.e FeatureSet-HF) is then fed to all the 22 classifiers and the results are in the most right column in Table 5.1. Comparison of accuracy when using FeatureSet-H and FeatureSet-HF shows that approximately 55% of the classifiers was able to either maintain the accuracy or improve it. On the other hand, 11 of the classifiers are better off by using FeatureSet-H as compared to FeatureSet-HF.

Additionally, all hybrid feature sets (ER, H, and HF) achieved accuracy higher than 90%. In particular, almost all of the hybrid feature sets produces relatively stable results between 90.66 % - 98.66 % as shown in Figure 5.1.

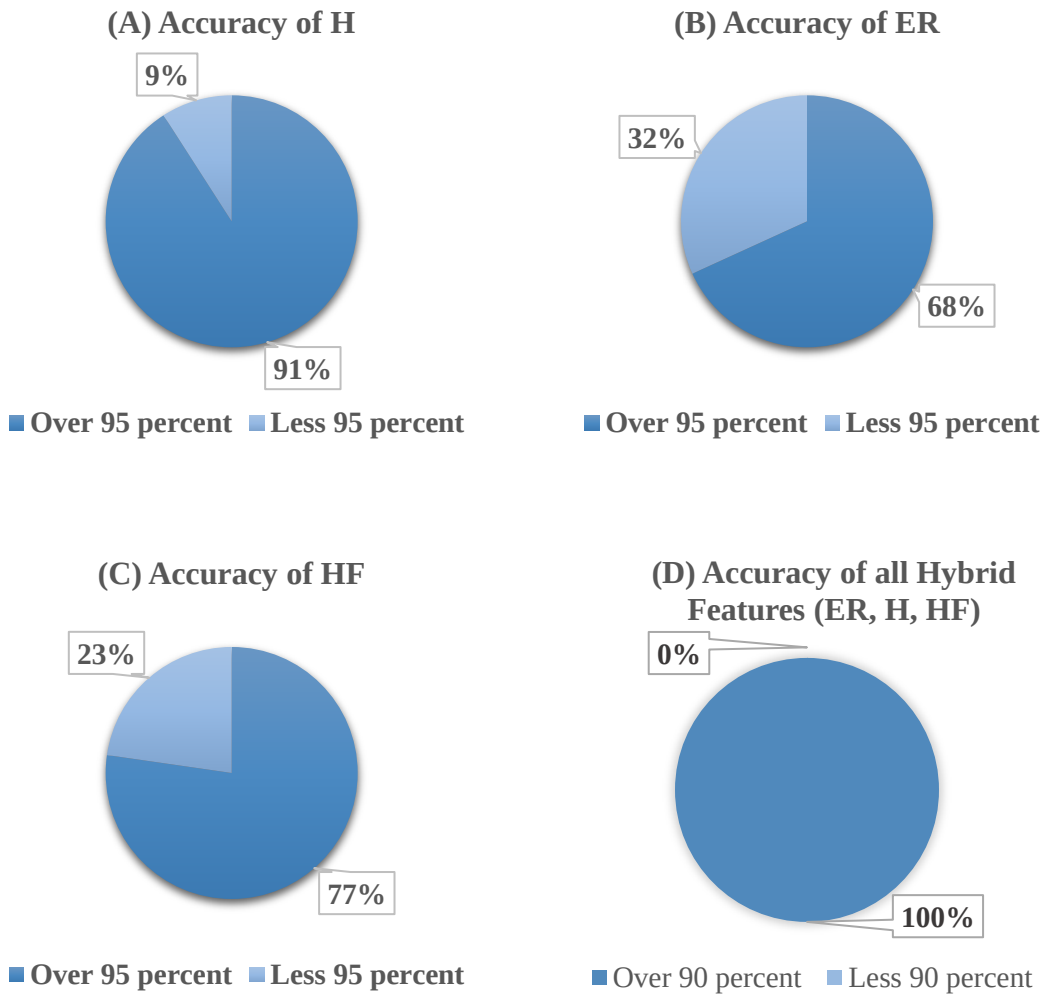


*Figure 5.1.* Classification Accuracy: Data Proportion of 70:30.

Table 5.2

*Classification Accuracy (%) for Experiment of Data Split into 80:20*

| Classifier | FeatureSet -UB | FeatureSet -UBA | FeatureSet -ER | FeatureSet -H | FeatureSet -HF |
|---|---|---|---|---|---|
| Bayes Net | 94 | 25 | 99 | 99 | 99 |
| Naïve Bayesian | 94 | 31 | 99 | 99 | 99 |
| LibLINEAR | 94 | 97 | 95 | 98 | 95 |
| LibSVM | 75 | 75 | 96 | 95 | 96 |
| Logistic | 96 | 95 | 94 | 94 | 96 |

| Classifier | FeatureSet -UB | FeatureSet -UBA | FeatureSet -ER | FeatureSet -H | FeatureSet -HF |
|---|---|---|---|---|---|
| Multilayer Perceptron | 91 | 95 | 89 | 95 | 94 |
| SMO | 96 | 94 | 94 | 97 | 95 |
| K* | 95 | 98 | 99 | 100 | 96 |
| AdaBoost M1 | 94 | 95 | 99 | 97 | 98 |
| Attribute Selected Classifier | 95 | 75 | 97 | 97 | 97 |
| Multi Class Classifier | 96 | 95 | 94 | 94 | 96 |
| Random Committee | 96 | 98 | 97 | 98 | 96 |
| Decision Tree | 95 | 98 | 98 | 97 | 98 |
| JRip | 94 | 97 | 98 | 99 | 98 |
| NNge | 85 | 98 | 82 | 98 | 82 |
| PART | 90 | 75 | 97 | 96 | 97 |
| FT | 95 | 96 | 96 | 96 | 96 |
| Hoeffding Tree | 93 | 75 | 91 | 98 | 75 |
| J48 | 94 | 75 | 97 | 97 | 97 |
| J48graft | 94 | 75 | 98 | 98 | 98 |
| J48 Consolidated | 92 | 75 | 96 | 98 | 96 |
| Random Forest | 77 | 98 | 94 | 97 | 97 |
| **Average** | **92.04** | **83.40** | **95.40** | **97.13** | **95.04** |

Table 5.2 presents the results for data proportion of 80:20. First, the results were on the performance of FeatureSet-ER against UB and UBA. Based on that, the constructed features based on ER was able to obtain around 60% out of total classifiers, whose was able to either maintain the accuracy or improve it. In contrast, 40% of the classifiers got lower accuracy rate. However, the ER features can still be considered relevant as 68% of the total classifiers obtained accuracy rate over 95% with average rate at 95.40%.

This experiment also reveals that FeatureSet-H is of equal or better than using FeatureSet-UB, UBA, and ER. Classification results stated that around 68% of the classifiers produce either equal or higher accuracy when FeatureSet-H was employed. Moreover, around 91% of the classifiers achieved classification accuracy over 95% with average rate at 97.13% as shown in Figure 5.2. Only 18% of the classifiers had a drop in their accuracy comparing with the single approach.

Again, this dissertation performs the feature selection method to reduce the size of FeatureSet-H. This Feature set (i.e FeatureSet-HF) is then fed to all the 22 classifiers and the results as in Table 5.2. Comparison of accuracy when using FeatureSet-H and FeatureSet-HF shows that approximately 59% of the classifiers was able to either maintain the accuracy or improve it with average rate at 95%. On the other hand, 41% of the classifiers are better off by using FeatureSet-H as compared to FeatureSet-HF.

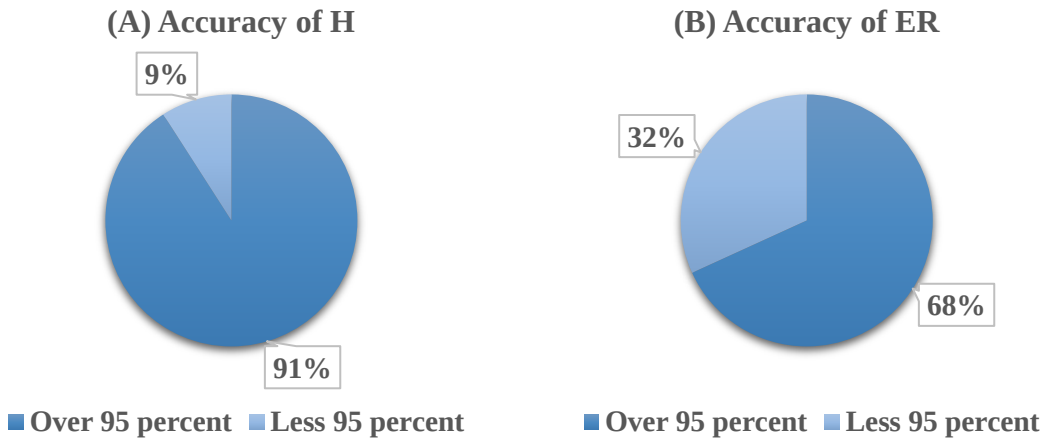


*Figure 5.2.* Classification Accuracy: Data Proportion of 80:20

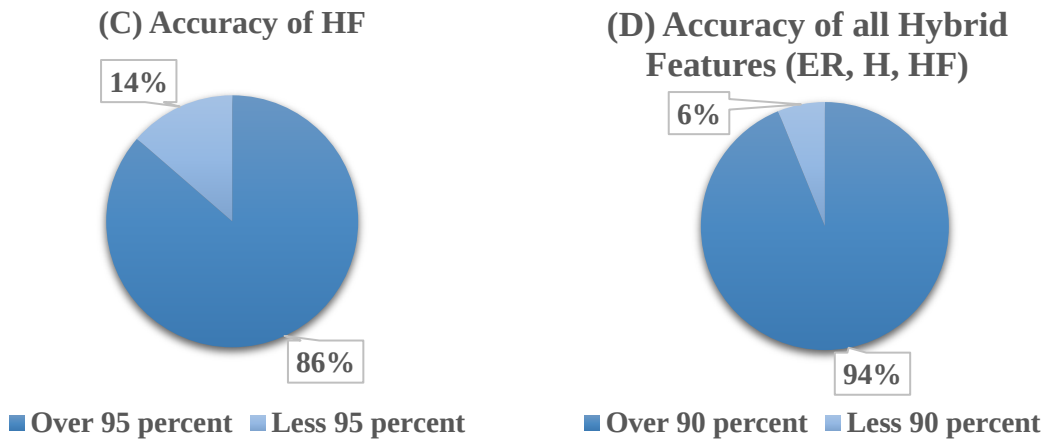


*Figure 5.2.* Classification Accuracy: Data Proportion of 80:20 (Continuous)

Table 5.3

*Classification Accuracy (%) for Experiment of Data Split into 90:10*

| Classifier | FeatureSet -UB | FeatureSet -UBA | FeatureSet -ER | FeatureSet -H | FeatureSet -HF |
|---|---|---|---|---|---|
| Bayes Net | 96 | 26 | 100 | 100 | 100 |
| Naïve Bayesian | 96 | 26 | 100 | 99 | 100 |
| LibLINEAR | 92 | 96 | 92 | 96 | 94 |
| LibSVM | 72 | 72 | 92 | 92 | 92 |
| Logistic | 96 | 96 | 98 | 94 | 98 |
| Multilayer Perceptron | 92 | 96 | 88 | 96 | 90 |
| SMO | 96 | 92 | 92 | 96 | 92 |
| K* | 94 | 96 | 98 | 100 | 98 |
| AdaBoostM1 | 92 | 92 | 100 | 98 | 100 |
| Attribute Selected Classifier | 94 | 72 | 98 | 98 | 98 |
| Multi Class Classifier | 96 | 96 | 98 | 98 | 98 |
| Random Committee | 76 | 96 | 96 | 94 | 98 |

| Classifier | FeatureSet -UB | FeatureSet -UBA | FeatureSet -ER | FeatureSet -H | FeatureSet -HF |
|---|---|---|---|---|---|
| Decision Tree | 92 | 96 | 96 | 94 | 100 |
| JRip | 92 | 96 | 98 | 98 | 98 |
| NNge | 90 | 32 | 92 | 98 | 100 |
| PART | 92 | 72 | 96 | 98 | 98 |
| FT | 94 | 98 | 96 | 98 | 94 |
| Hoeffding Tree | 94 | 72 | 92 | 100 | 92 |
| J48 | 92 | 72 | 98 | 98 | 98 |
| J48graft | 92 | 72 | 98 | 98 | 98 |
| J48Consolidated | 90 | 28 | 98 | 98 | 98 |
| Random Forest | 88 | 96 | 100 | 94 | 96 |
| **Average** | **91.27** | **76.81** | **96.18** | **97.04** | **96.81** |

In the same way, the performance of various classifiers using data proportion of 90:10 is depicted in Table 5.3. A closer look indicates that performance of hybrid features (FeatureSet-ER, FeatureSet-H, and FeatureSet-HF) are good. For instance, the Bayes Net classifier, using FeatureSet-UB, the accuracy is 96 while FeatureSet-UBA only achieve 26%. But, when the classifier was fed with FeatureSet-ER, H, and HF the accuracy has increased to 100% for all three feature sets. Again the results analysis begins by examining the performance of FeatureSet-ER against FeatureSet-UB and FeatureSet-UBA. Based on the results, around 68% out of classifiers give better results in classifying malicious users with accuracy rate as high as 100% with average rate at 96.18%. On the other hand, only two of the classifiers have produced the same accuracy while using FeatureSet-UB, UBA, and ER. Despite this, there are only five classifiers that have a decrease in their accuracy. Additionally, the ER features are relevant as 77% of the classifiers obtained accuracy rate over 95% as shown in Figure 5.3.

Then again, the performance of FeatureSet-H is compared against all features of FeatureSet-UB, UBA, and ER. Experiments showed that the proposed FeatureSet-H is equal or better than using FeatureSet-ER. Classification results stated that around 86% of the classifiers produce higher or equal accuracy compared to using FeatureSet-ER. Likewise, around 73% of the classifiers produce equal or higher accuracy compared with UB and UBA with average rate at 97%. Moreover, around 82% out of the classifiers achieved classification accuracy over 95% as shown in Figure 5.3.

The comparison of accuracy when using FeatureSet-H and FeatureSet-HF shows that approximately 68% out of the classifiers was able to either maintain the accuracy or improve it with average rate at 96.81%. Additionally, 98% out of available feature sets based on hybrid features were obtained accuracy rate over 90%. While the performance of hybrid features was relatively stable almost over 95 % as shown in Figure 5.3.

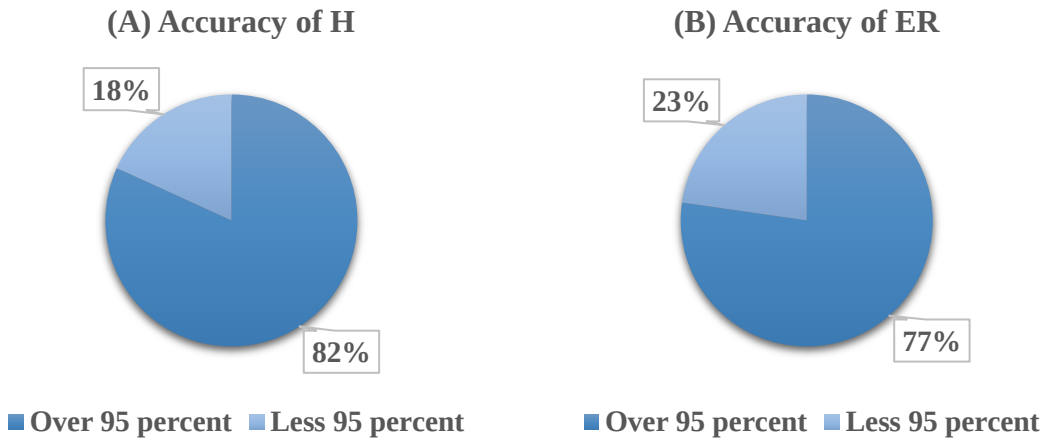


*Figure 5.3* Classification Accuracy: Data Proportion of 90:10

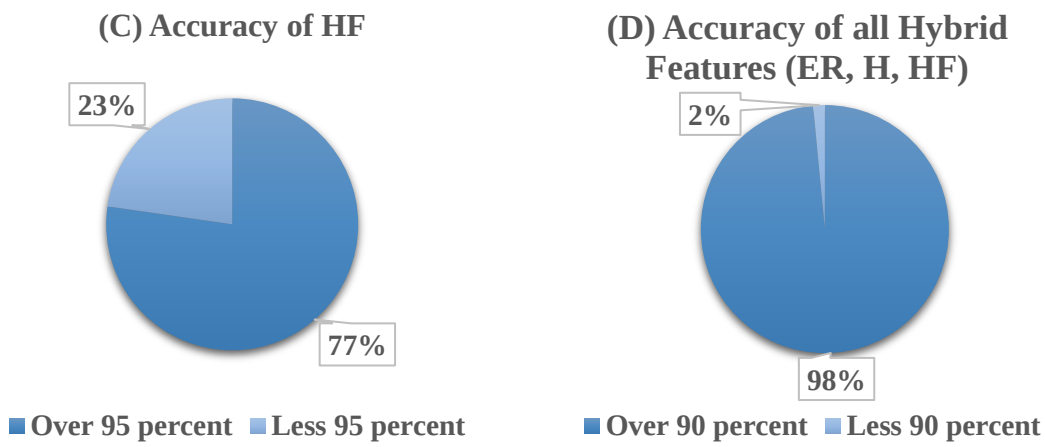


*Figure 5.3.* Classification Accuracy: Data Proportion of 90:10 (Continuous)

This dissertation came out with different feature sets, while the main one is hybrid features (FeatureSet-H), which is consider integration of UB, UBA, and ER. So, in order to see the effect of hybrid features based on different data proportion, this dissertation made a comparison between the obtained results (i.e. FeatureSet-H). Table 5.4 and Figure 5.4 illustrate the comparison of classifiers accuracy across FeatureSet-H.

Table 5.4

*Comparison of Classifiers Accuracy (%) based on Hybrid Features*

| **Classifier** | **FeatureSet-H (A) (70:30)** | **FeatureSet-H (B) (80:20)** | **FeatureSet-H (C) (90:10)** |
|---|---|---|---|
| Bayes Net | 96.66 | 99 | 100 |
| Naïve Bayesian | 96 | 99 | 99 |
| LibLINEAR | 96.66 | 98 | 96 |
| LibSVM | 94 | 95 | 92 |
| Logistic | 95.33 | 94 | 94 |
| Multilayer Perceptron | 96.66 | 95 | 96 |
| SMO | 96.66 | 97 | 96 |

| Classifier | FeatureSet-H (A) (70:30) | FeatureSet-H (B) (80:20) | FeatureSet-H (C) (90:10) |
|---|---|---|---|
| K* | 98.66 | 100 | 100 |
| AdaBoostM1 | 95.33 | 97 | 98 |
| Attribute Selected Classifier | 95.33 | 97 | 98 |
| Multi Class Classifier | 95.33 | 94 | 98 |
| Random Committee | 98 | 98 | 94 |
| Decision Tree | 95.33 | 97 | 94 |
| JRip | 96.66 | 99 | 98 |
| NNge | 96.66 | 98 | 98 |
| PART | 96.66 | 96 | 98 |
| FT | 96 | 96 | 98 |
| Hoeffding Tree | 96 | 98 | 100 |
| J48 | 95.33 | 97 | 98 |
| J48graft | 96 | 98 | 98 |
| J48Consolidated | 94.66 | 98 | 98 |
| Random Forest | 97.33 | 97 | 94 |

Table 5.4 and Figure 5.4 compare the classification accuracy for FeatureSet-H based on data proportion 70:30 (A), 80:20 (B), and 90:10 (C) respectively. After examining the results, it can be seen that the hybrid features were able to achieve significant accuracy result among a set of classifiers. According to the accuracy results, FeatureSet-H (B) improved by 68% out of total classifiers comparing with FeatureSet-H (A). While 23% of classifiers slightly decreased and only 9% of classifiers got same results over different data proportion. Likewise, FeatureSet-H (C) also improved by 64% and 41% comparing with FeatureSet-H (A) and FeatureSet-H (B) respectively. However. The decreased in the results was at 36% and 32% comparing with FeatureSet-H (A) and FeatureSet-H (B) respectively.

Additionally, the outcome shows that more data for training will steadily increase the result accuracy for most of the classifiers.

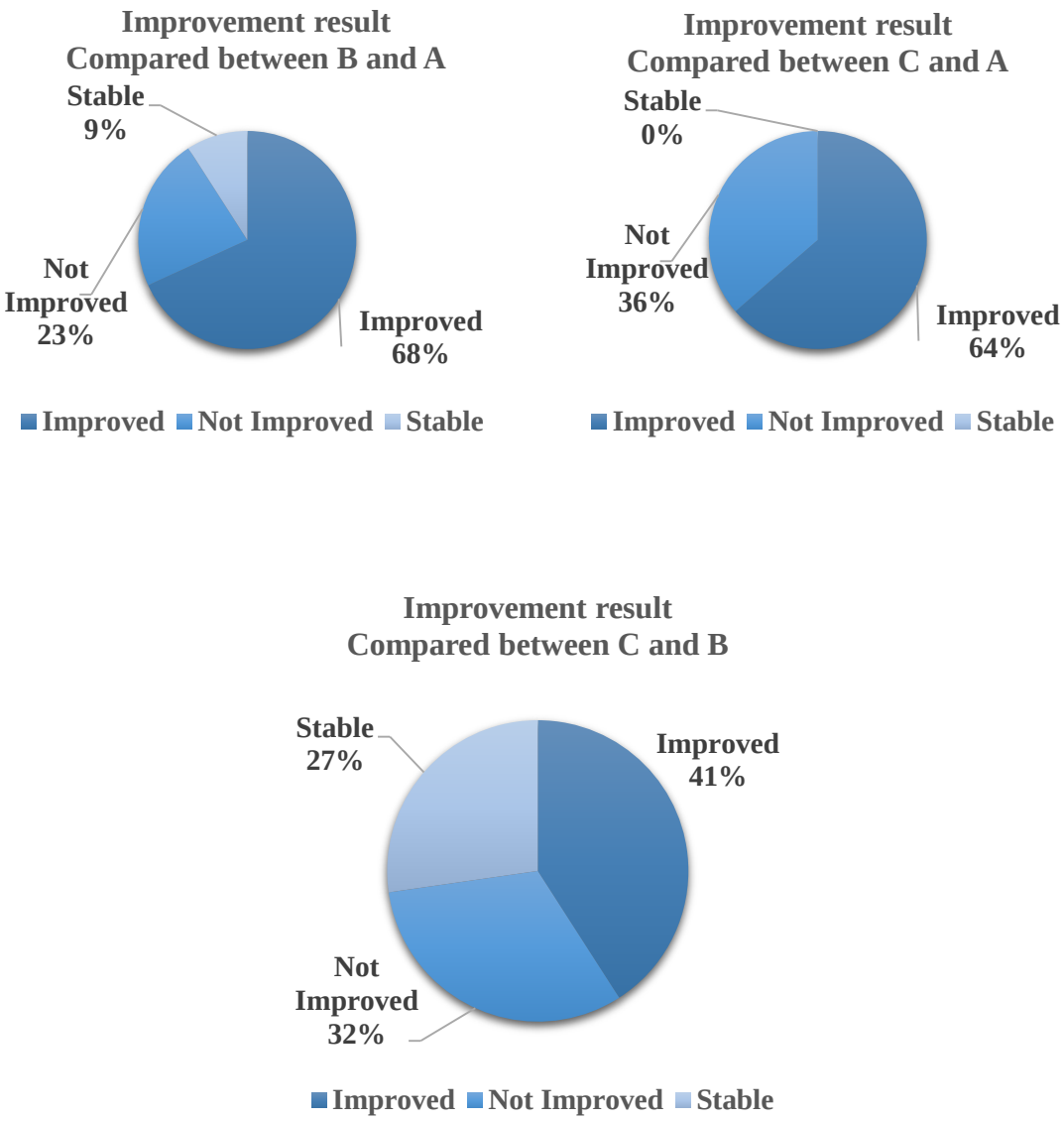


*Figure 5.4.* Classification Accuracy of FeatureSet-H using Different Data Proportion

In order to get a better understanding of the results and determine which feature set is the best, this dissertation performed a T-Test experiment to return the probability associated with a sample. For each paired of hybrid feature set (HA) and single feature set (SA), this

study performed T-Test based on different data portion groups 70%-30%, 80%-20%, and 90%-10%, while using 0.05 as P-value. For instance, this dissertation starts by examining all feature sets with data portion 70%-30%, where the analysis is between FeatureSet-ER vs FeatureSet-UB, then FeatureSet-H vs FeatureSet-UB, followed by FeatureSet-HF vs FeatureSet-UB. Similar analysis is performed on other data portion, for example, the probability under data portion 90%-10% is examined between FeatureSet-ER vs FeatureSet-UBA, and then FeatureSet-H vs FeatureSet-UBA, follow by FeatureSet-HF vs FeatureSet-UBA. Table 5.5 illustrates the T-Test results, where each result shows the probability associated with a feature set's paired T-Test.

Table 5.5
*T-Test Results for Each Feature Sets Paired Based on Split Percentage*

| SA / HA | FeatureSet-UB | | | FeatureSet-UBA | | |
|---|---|---|---|---|---|---|
| | **70:30** | **80:20** | **90:10** | **70:30** | **80:20** | **90:10** |
| FeatureSet-ER | 0.00218 | 0.03012 | 0.00214 | 0.00640 | 0.01334 | 0.00206 |
| FeatureSet-H | 0.00066 | 0.00053 | 0.00026 | 0.00373 | 0.00514 | 0.00139 |
| FeatureSet-HF | 0.00098 | 0.08745 | 0.00056 | 0.00446 | 0.01721 | 0.00156 |

Based on T-Test results on all experiments of various data proportion; 70%-30%, 80%-20%, and 90%-10%, it is learned that the hybrid feature sets FeatureSet-ER, FeatureSet-H, and FeatureSet-HF gives better insight on classifying malicious users as most of feature sets produce better performance compared with FeatureSet-UB and FeatureSet-UBA. Moreover, a closer look at the T-Test results indicated that FeatureSet-H produces highly significant performance compared with other feature sets. Furthermore, this dissertation conduct another T-Test between FeatureSet-H and FeatureSet-HF based on various data proportion 70%-30%, 80%-20%, and 90%-10% as shown in Table 5.6.

Table 5.6
*T-Test Results for FeatureSet-H vs FeatureSet-HF Based on Split Percentage*

| Feature Set | FeatureSet-H | | |
|---|---|---|---|
| | **70:30** | **80:20** | **90:10** |
| **FeatureSet-HF** | 0.30319 | 0.10567 | 0.78022 |

Data in Table 5.6 shows that accuracy difference between FeatureSet-H and FeatureSet-HF is not significant. In all percentage split, the values are greater than 0.005 (i.e 0.30319, 0.10567 and 0.78022). This shows that there isn't enough evidence to claim that using FeatureSet-HF is better than using FeatureSet-H.

### 5.1.2 Experiments using Cross Validation

The second phase of experiments employs the cross-validation (CV) method in different fold such as 10, 15, and 20 (Hu et al., 2014). For each fold, this method is randomly assigned data points to two sets d0 and d1, so that both sets are equal size. Then train on d0 and test on d1, followed by training on d1 and testing on d0. This has the advantage that the training and test sets are both large, and each data point is used for both training and validation on each fold. Table 5.7 - 5.9 summarized the performance of the proposed features across 22 classification algorithms. Figure 5.5 – 5.7 visualized the performance statistics laid out by Table 5.7 – 5.9, respectively.

Table 5.7

*Classification Accuracy (%) for Experiment of Data CV 10 Fold*

| Classifier | FeatureSet -UB | FeatureSet -UBA | FeatureSet -ER | FeatureSet -H | FeatureSet -HF |
|---|---|---|---|---|---|
| Bayes Net | 90 | 28.4 | 95.2 | 95.6 | 97.2 |

| Classifier | FeatureSet -UB | FeatureSet -UBA | FeatureSet -ER | FeatureSet -H | FeatureSet -HF |
|---|---|---|---|---|---|
| Naïve Bayesian | 91 | 28.8 | 74.2 | 78.6 | 88.4 |
| LibLINEAR | 93.2 | 95.8 | 93.8 | 96 | 94.8 |
| LibSVM | 73.2 | 73.2 | 91.4 | 91.6 | 92.6 |
| Logistic | 92 | 91.8 | 96.6 | 93.8 | 98 |
| Multilayer Perceptron | 87.8 | 88.4 | 85.6 | 91 | 89 |
| SMO | 93.2 | 96 | 94.2 | 96 | 94.6 |
| K* | 92.8 | 95.8 | 97.6 | 98.4 | 97.6 |
| AdaBoostM1 | 93.2 | 93.8 | 96.8 | 97.8 | 97.4 |
| Attribute Selected Classifier | 93.8 | 73.2 | 96.4 | 95.8 | 96.4 |
| Multi Class Classifier | 92 | 91.8 | 96.6 | 93.8 | 98 |
| Random Committee | 88.4 | 85 | 95.2 | 95.6 | 96.2 |
| Decision Tree | 91.4 | 95.4 | 96 | 95.6 | 95.4 |
| JRip | 92.2 | 95.2 | 96.8 | 96.6 | 97 |
| NNge | 84.4 | 83.4 | 85.2 | 95.2 | 86.4 |
| PART | 92.6 | 73.2 | 96.6 | 95.4 | 97 |
| FT | 93.8 | 95.8 | 97.8 | 94.6 | 96.6 |
| Hoeffding Tree | 90.2 | 73.2 | 88.8 | 94.4 | 95.4 |
| J48 | 93.2 | 73.2 | 95 | 95.6 | 95.8 |
| J48graft | 93 | 73.2 | 95.2 | 96.4 | 96.4 |
| J48Consolidated | 90 | 31.6 | 95.4 | 96.8 | 96.6 |
| Random Forest | 77.4 | 90.8 | 96 | 94.8 | 96 |
| **Average** | **89.94** | **78.5** | **93.47** | **94.51** | **95.12** |

Table 5.7 shows the performance of the proposed features across 22 classifiers using cross-validation 10 fold. As shown in Figure 5.5 the performance of the hybrid features across most classifiers was magnificent compare with UB and UBA. For instance, the Bayes Net

classifier, using FeatureSet-UB, the accuracy was only 90% while FeatureSet-UBA performed worst by only achieving 28.4. But, when the classifier was fed with FeatureSet-ER, H, and HF the accuracy has increased to 95.2%, 95.6%, and 97.2% respectively. First, the results analysis begins by examining the performance of FeatureSet-ER against FeatureSet-UB and FeatureSet-UBA. Based on the results, ER gives better insight on the classifying malicious users as 77% out of the classifiers' accuracy has increased. While only 22% out of these classifiers slightly decreased their results. In addition, the ER features have achieved good performance results as 73% out of the total classifiers obtained accuracy rate over 95%.

It is clear that FeatureSet-ER produces better accuracy, its performance is compared against FeatureSet-H that includes all features of FeatureSet-UB, UBA, and ER. Experiments showed that the proposed FeatureSet-H is equal or better than using FeatureSet-ER. Classification results stated that around 60% and 90% of the classifiers produce higher or equal accuracy compares with ER, UB, and UBA from one side, UB and UBA from another sides. Moreover, around 68% of the classifiers achieved classification accuracy over 95% with average rate at 94.51% as shown in Figure 5.5.

Similarly, this dissertation performs the feature selection method to reduce the size of FeatureSet-H. This feature set (i.e FeatureSet HF) is then fed to all the 22 classifiers and the results as in Table 5.7. Comparison of accuracy when using FeatureSet-H and FeatureSet-HF shows that approximately 64% of the classifiers was able to either maintain the accuracy or improve it with average rate at 95.12%. On the other hand, only 8 of the classifiers are better off by using FeatureSet-H as compared to FeatureSet-HF.

Additionally, 91% of available feature sets based on hybrid features were achieved accuracy result more than 90% as shown in Figure 5.5.

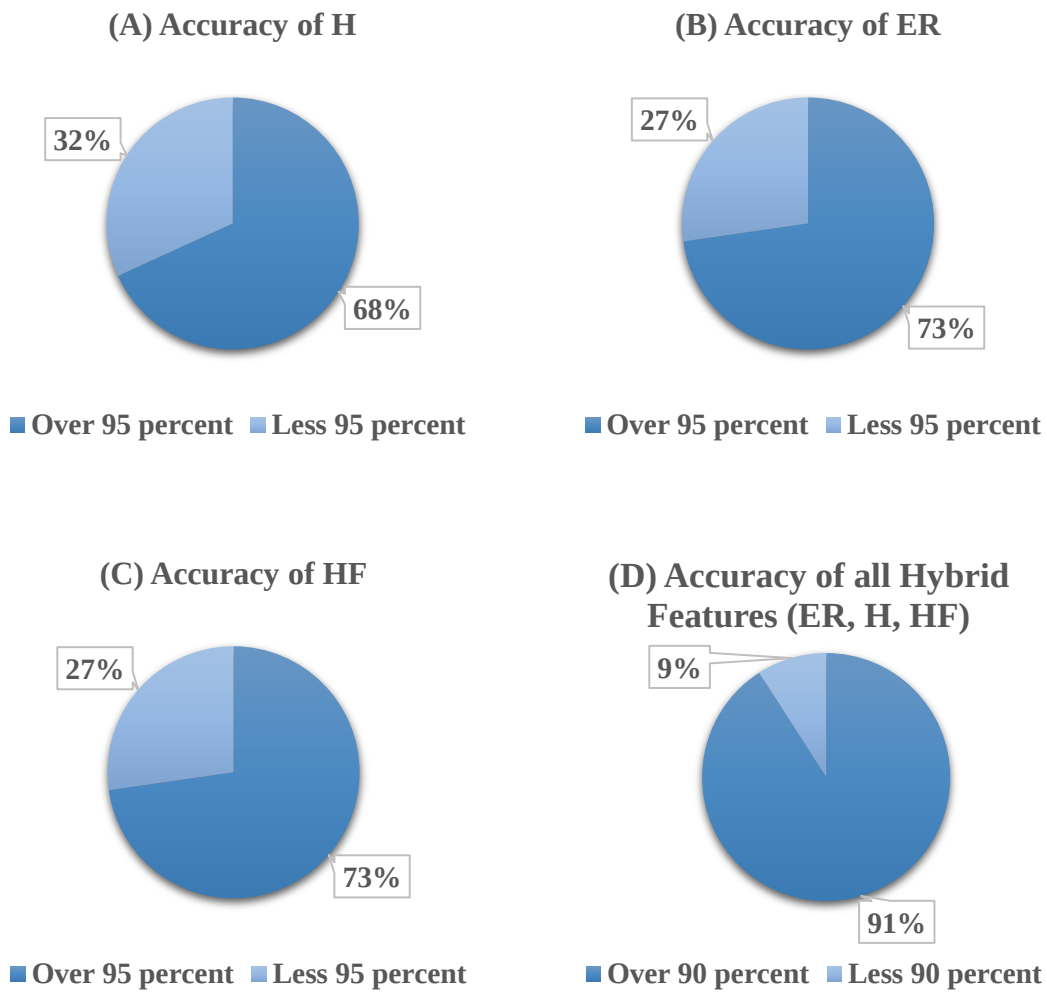


*Figure 5.5.* Classification Accuracy: CV 10 Fold

Table 5.8

*Classification Accuracy (%) for Experiment of Data CV 15 Fold*

| Classifier | FeatureSet -UB | FeatureSet -UBA | FeatureSet -ER | FeatureSet -H | FeatureSet -HF |
|---|---|---|---|---|---|
| Bayes Net | 90.4 | 28.8 | 95 | 96 | 97.6 |
| Naïve Bayesian | 91.4 | 29 | 74.4 | 78.6 | 88.6 |

| Classifier | FeatureSet -UB | FeatureSet -UBA | FeatureSet -ER | FeatureSet -H | FeatureSet -HF |
|---|---|---|---|---|---|
| LibLINEAR | 93.2 | 96.2 | 93.4 | 96.6 | 94.4 |
| LibSVM | 73.2 | 73.2 | 91.4 | 91.6 | 92.6 |
| Logistic | 91.6 | 91.8 | 95.4 | 93.8 | 97 |
| Multilayer Perceptron | 87 | 91.6 | 90.2 | 85.8 | 80.2 |
| SMO | 93.2 | 96.4 | 94 | 96.6 | 94.6 |
| K* | 92.8 | 95.8 | 97.6 | 98.2 | 97.6 |
| AdaBoostM1 | 93.4 | 93.8 | 96.8 | 97 | 97 |
| Attribute Selected Classifier | 94.2 | 73.2 | 96.2 | 96 | 97.2 |
| Multi Class Classifier | 91.6 | 91.8 | 95.4 | 93.8 | 97 |
| Random Committee | 81.8 | 86 | 92.8 | 95 | 96.2 |
| Decision Tree | 91.8 | 95 | 96 | 94.8 | 95.6 |
| JRip | 92.8 | 96.2 | 96 | 97.6 | 96.8 |
| NNge | 84.2 | 83 | 82.6 | 95.2 | 83.8 |
| PART | 93.4 | 73.2 | 96 | 96.6 | 96.4 |
| FT | 93.8 | 95.6 | 97 | 95.4 | 96.6 |
| Hoeffding Tree | 90.6 | 73.2 | 88.6 | 94.8 | 94.8 |
| J48 | 93.2 | 73.2 | 95.8 | 95.6 | 96.4 |
| J48graft | 93 | 73.2 | 95.4 | 96.2 | 96.4 |
| J48Consolidated | 89.8 | 29.8 | 95.2 | 94.2 | 94.6 |
| Random Forest | 79.2 | 92.4 | 96 | 95.2 | 96.4 |
| **Average** | **89.8** | **78.74** | **93.23** | **94.3** | **94.44** |

Table 5.8 and Figure 5.6 compare the performance for the proposed features across 22 different classifiers using cross-validation 15 fold. Approximately 88% of all classifiers based on hybrid features has been achieved accuracy rate over 90% as shown in Figure 5.6. The performance of the hybrid features across most classifiers still outstanding compared

with UB and UBA. As can be seen in Bayes Net classifier, using FeatureSet-UB, the accuracy was only 90.4% while FeatureSet-UBA performed worst by only achieving 28.8%. But, when the classifier was fed with FeatureSet-ER, H, and HF the accuracy has increased to 95%, 96%, and 97.6% respectively.

Initially, the results analysis begins by examining the performance of FeatureSet-ER against FeatureSet-UB and FeatureSet-UBA. Based on the results, the ER give better results as 73% out of the classifiers' accuracy has increased with average rate at 93.23%. While only 27% out of these classifiers slightly decreased their results. In addition, the ER features have achieved good performance as 68% out of the total classifiers obtained accuracy rate over 95%.

Subsequently, the FeatureSet-ER results produced good accuracy rate. Hence, its performance is compared against FeatureSet-H that includes all features of FeatureSet-UB, UBA, and ER. Experiments showed that the proposed FeatureSet-H is equal or better than using FeatureSet-ER. Classification results stated that around 55% and 82% of the classifiers produce higher or equal accuracy compares with (ER, UB, and UBA), (UB and UBA) respectively. Moreover, around 68% of the classifiers achieved classification accuracy over 95% with average rate at 94.3% as shown in Figure 5.6.

Then again, this dissertation performs the feature selection method to reduce the size of FeatureSet-H. FeatureSet-HF is then fed to all the 22 classifiers and the results as in Table 5.6. Comparison of accuracy when using FeatureSet-H and FeatureSet-HF shows that approximately 68% of the classifiers was able to either maintain the accuracy or improve it. On the other hand, only 7 of the classifiers are better off by using FeatureSet-H as

compared to FeatureSet-HF. Additionally, 88% of available feature sets based on hybrid features were achieved accuracy result more than 90% as shown in Figure 5.6.

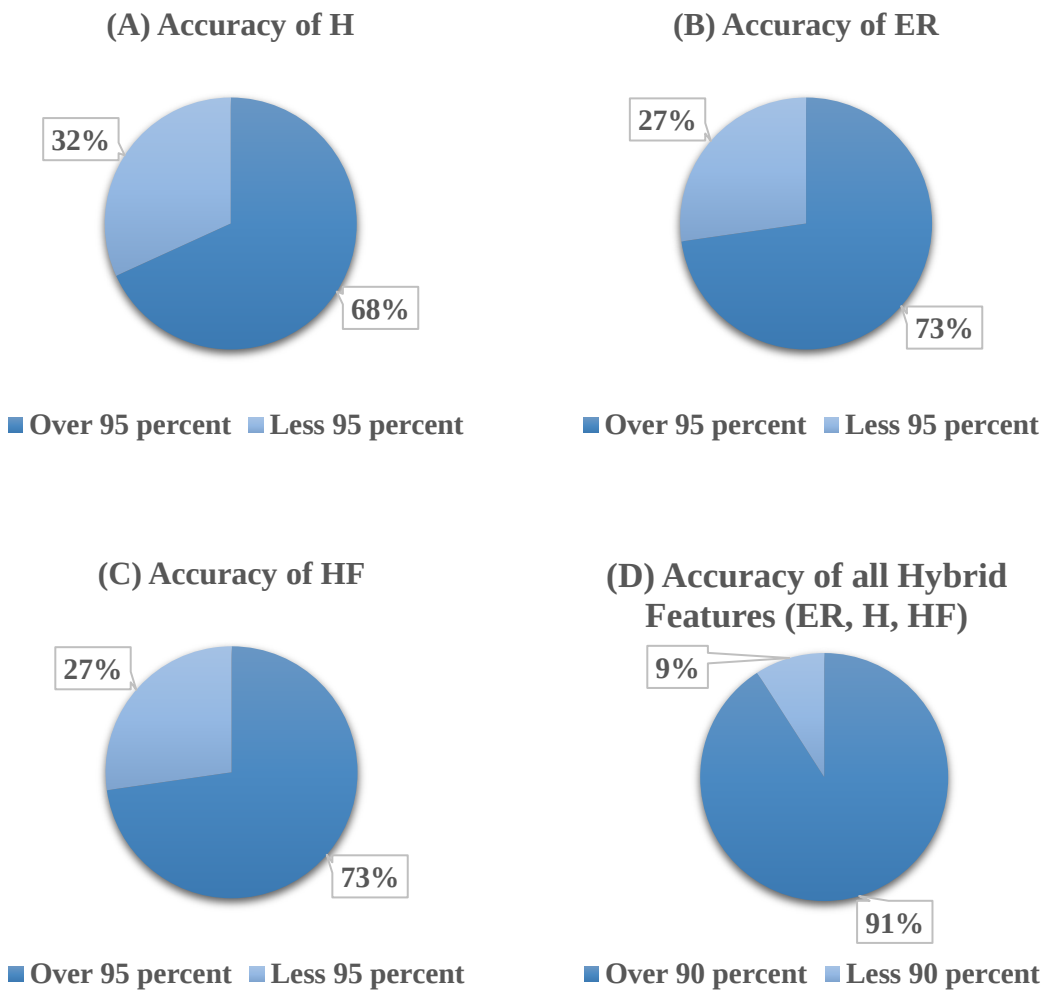


*Figure 5.6.* Classification Accuracy: CV 15 Fold

Table 5.9

*Classification Accuracy (%) for Experiment of Data CV 20 Fold*

| Classifier | FeatureSet -UB | FeatureSet -UBA | FeatureSet -ER | FeatureSet -H | FeatureSet -HF |
|---|---|---|---|---|---|
| Bayes Net | 90 | 28.6 | 95.2 | 96 | 97.2 |
| Naïve Bayesian | 90 | 28.8 | 74.6 | 78.6 | 88.6 |

| Classifier | FeatureSet -UB | FeatureSet -UBA | FeatureSet -ER | FeatureSet -H | FeatureSet -HF |
|---|---|---|---|---|---|
| LibLINEAR | 93.8 | 96.2 | 93.8 | 96.4 | 95 |
| LibSVM | 73.2 | 73.2 | 91.8 | 92.2 | 93.2 |
| Logistic | 92 | 92.4 | 97.2 | 94 | 97 |
| Multilayer Perceptron | 88.4 | 92.2 | 89.6 | 91.8 | 91.2 |
| SMO | 93.2 | 96.4 | 93.8 | 96.4 | 94.8 |
| K* | 93.2 | 96 | 97.8 | 98.4 | 97.6 |
| AdaBoost M1 | 93.4 | 93.8 | 96.6 | 97.8 | 97.6 |
| Attribute Selected Classifier | 94.2 | 73.2 | 97 | 96 | 97.2 |
| Multi Class Classifier | 92 | 92.4 | 97.2 | 94 | 97 |
| Random Committee | 80.4 | 83 | 95.8 | 93.6 | 96.2 |
| Decision Tree | 91.6 | 95.6 | 95.6 | 95.6 | 95.2 |
| JRip | 93 | 95.6 | 97 | 97.4 | 97.2 |
| NNge | 84.2 | 73 | 85 | 95.2 | 86.4 |
| PART | 92.8 | 73.2 | 96.4 | 95.6 | 96.8 |
| FT | 93.4 | 96 | 97.6 | 96.2 | 96 |
| Hoeffding Tree | 90.8 | 73.2 | 89.2 | 94.8 | 94.8 |
| J48 | 93 | 73.2 | 96 | 96.4 | 97.4 |
| J48graft | 92.8 | 73.2 | 96.6 | 97.4 | 98 |
| J48Consolidated | 91 | 29 | 96.4 | 95.4 | 96.4 |
| Random Forest | 78.2 | 90.8 | 96.2 | 95 | 96.4 |
| **Average** | **89.75** | **78.13** | **93.92** | **94.73** | **95.32** |

Table 5.9 show the performance of the proposed features across 22 classifiers using cross-validation 20 fold. As shown in Figure 5.7 the performance of the hybrid features across most classifiers considered excellent compared with UB and UBA. For instance, the J48 classifier, using FeatureSet-UB, the accuracy was only 93% while FeatureSet-UBA

performed poorest by only achieving 73.2%. But, when the classifier was fed with FeatureSet-ER, H, and HF the accuracy has increased to 96%, 96.4%, and 97.4% respectively.

At first, the results analysis begins by examining the performance of FeatureSet-ER against FeatureSet-UB and FeatureSet-UBA. Based on the results, ER gives good rate at 77% out of the classifiers' accuracy has increased with average rate at 93.92%. While only 18% out of these classifiers slightly decreased their results. In addition, the ER features have been achieved good performance results as 68% out of the total classifiers obtained accuracy rate over 95%.

It is clear that FeatureSet-ER produces good accuracy rate, its performance is compared against FeatureSet-H that includes all features of FeatureSet-UB, UBA, and ER. Experiments showed that the proposed FeatureSet-H is equal or better than using FeatureSet-ER. Classification results stated that around 60% and 91% of the classifiers produce higher or equal accuracy compares with (ER, UB, and UBA), (UB and UBA) respectively. Moreover, FeatureSet-H obtained rate at 68% out of the classifiers, which achieved classification accuracy over 95% with average rate at 94.73% as shown in Figure 5.7.

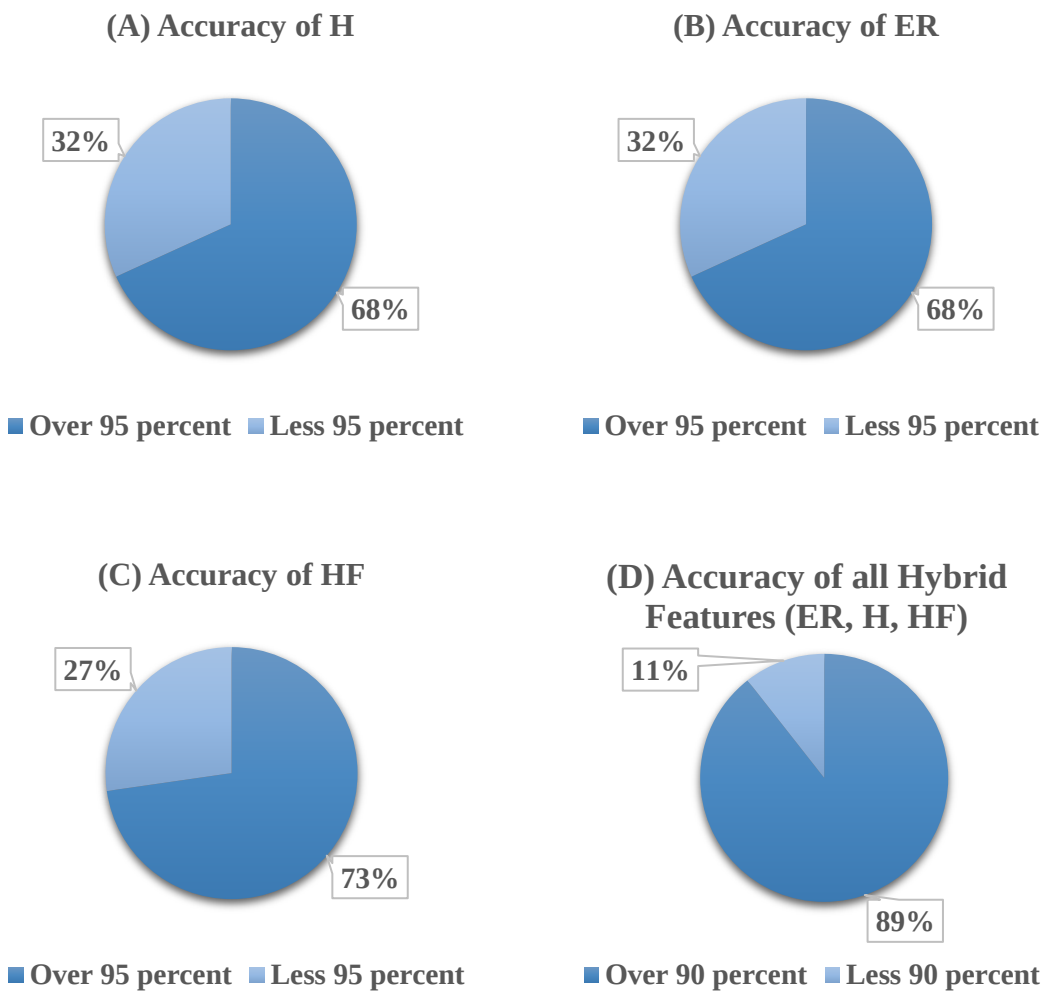


*Figure 5.7.* Classification Accuracy: CV 20 Fold

In order to see the effect of hybrid features based on different cross-validation proportion, this dissertation made a comparison between the obtained results (i.e. FeatureSet-H). Table 5.8 and Figure 5.8 illustrate the comparison of classifiers accuracy across FeatureSet-H.

Table 5.10

*Comparison of Classifiers Accuracy (%) based on Hybrid Features*

| Classifier | FeatureSet-H (A) (CV-10) | FeatureSet-H (B) (CV-15) | FeatureSet-H (C) (CV-20) |
|---|---|---|---|
| Bayes Net | 95.6 | 96 | 96 |
| Naïve Bayesian | 78.6 | 78.6 | 78.6 |
| LibLINEAR | 96 | 96.6 | 96.4 |
| LibSVM | 91.6 | 91.6 | 92.2 |
| Logistic | 93.8 | 93.8 | 94 |
| Multilayer Perceptron | 91 | 85.8 | 91.8 |
| SMO | 96 | 96.6 | 96.4 |
| K* | 98.4 | 98.2 | 98.4 |
| Decision Tree | 95.6 | 94.8 | 95.6 |
| JRip | 96.6 | 97.6 | 97.4 |
| NNge | 95.2 | 95.2 | 95.2 |
| PART | 95.4 | 96.6 | 95.6 |
| AdaBoostM1 | 97.8 | 97 | 97.8 |
| Attribute Selected Classifier | 95.8 | 96 | 96 |
| Multi Class Classifier | 93.8 | 93.8 | 94 |
| Random Committee | 95.6 | 95 | 93.6 |
| FT | 94.6 | 95.4 | 96.2 |
| Hoeffding Tree | 94.4 | 94.8 | 94.8 |
| J48 | 95.6 | 95.6 | 96.4 |
| J48graft | 96.4 | 96.2 | 97.4 |
| J48Consolidated | 96.8 | 94.2 | 95.4 |
| Random Forest | 94.8 | 95.2 | 95 |

Table 5.10 and Figure 5.8 compare the classification accuracy for FeatureSet-H based on cross-validation 10 (A), 15 (B), and 20 (C) respectively. After examining the results, it can

be seen that the hybrid features were able to achieve significant accuracy result among a set of classifiers with the average result at 94.5%. According to the accuracy results, FeatureSet (B) improved by 23% out of total classifiers comparing with FeatureSet (A). In contrast, FeatureSet (C) was significantly improved by 68% out of total classifiers comparing with FeatureSet (A). Additionally, this result drops down at 50% out of total classifiers comparing with FeatureSet (B). This is mean the hybrid features (FeatureSet-H) have a steady improvement at 20%. Likewise, the outcome showed that more data for training will steadily increase the result accuracy for most of the classifiers.

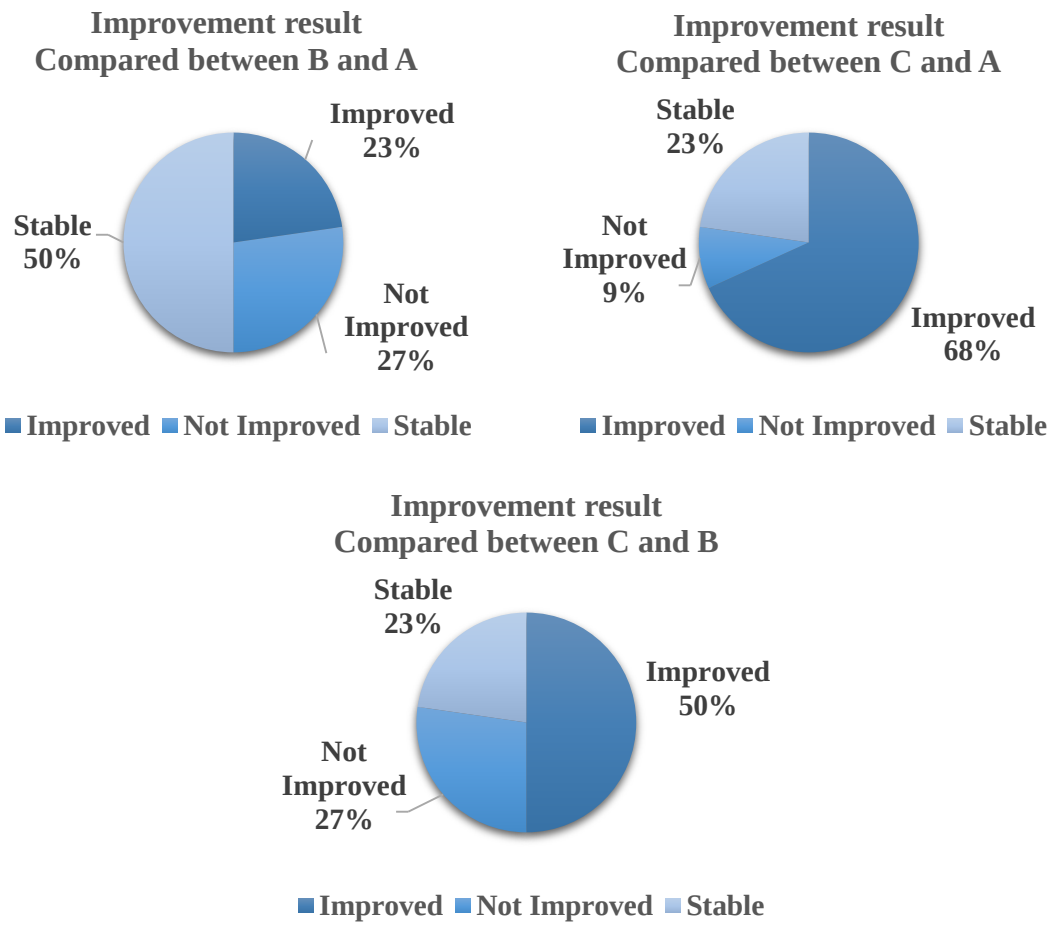


*Figure 5.8.* Classification Accuracy of FeatureSet-H using Different Fold Proportion

In the same way that abled in different data portion, this dissertation performed a T-Test experiment to return the probability associated with a sample. For each paired of hybrid feature set (HA) and single feature set (SA), this study abled the T-Test based on different fold values 10, 15, and 20, while P-value is 0.05. For instance, this dissertation start by examining all feature sets with fold value 10, where the probability compared between FeatureSet-ER vs FeatureSet-UB, then FeatureSet-H vs FeatureSet-UB, followed by FeatureSet-HF vs FeatureSet-UB. In the same way for other data portion, for example, the probability under data portion 15 is compared between FeatureSet-ER vs FeatureSet-UBA, then FeatureSet-H vs FeatureSet-UBA, follow by FeatureSet-HF vs FeatureSet-UBA. Table 5.11 illustrates the T-Test results, where each result shows the probability associated with a feature set's paired T-Test.

Table 5.11
*T-Test Results for Each Feature Sets Paired Based on Cross-Validation*

| SA / HA | FeatureSet-UB | | | FeatureSet-UBA | | |
|---|---|---|---|---|---|---|
| | **10** | **15** | **20** | **10** | **15** | **20** |
| **FeatureSet-ER** | 0.03721 | 0.04240 | 0.01621 | 0.00475 | 0.00658 | 0.00362 |
| **FeatureSet-H** | 0.00240 | 0.00430 | 0.00168 | 0.00262 | 0.00373 | 0.00229 |
| **FeatureSet-HF** | 0.00039 | 0.00385 | 0.00028 | 0.00188 | 0.00350 | 0.00166 |

Based on T-Test results across all experiments of various fold values; 10, 15, and 20, it is learned that the hybrid feature sets FeatureSet-ER, FeatureSet-H, and FeatureSet-HF gives better insight on classifying malicious users as all groups of the classifiers produce better performance compared with FeatureSet-UB and FeatureSet-UBA. Moreover, a closer look at the T-Test results indicated that FeatureSet-HF produces highly significant performance

compared with FeatureSet-UB (i.e 0.0039, 0.00385, 0.00028), FeatureSet-UBA, and FeatureSet-ER, while its results slightly improved compared with FeatureSet-H. However, the proposed FeatureSet-H is still producing highly significant performance compared with FeatureSet-UB, FeatureSet-UBA, and FeatureSet-ER.

Furthermore, this dissertation conducts another T-Test between FeatureSet-H and FeatureSet-HF based on various fold values 10, 15, and 20 as shown in Table 5.12.

Table 5.12
*T-Test Results for FeatureSet-H vs FeatureSet-HF Based on Cross-Validation*

| **FeatureSet** | **FeatureSet-H** | | |
|---|---|---|---|
| | **10** | **15** | **20** |
| FeatureSet-HF | 0.57672 | 0.91367 | 0.58123 |

Even though by using FeatureSet-HF, most classifiers obtained a higher classification accuracy (refer to Table 5.7 – 5.9), the difference is not significant at P-value 0.005. Data in Table 5.12 shows that the values are greater than the utilized P-value.

## 5.2 Discussion

In this dissertation, the discussion centres on how the hybrid features is able to improve the classification results. It is clear that hybrid features were able to achieve good results as the obtained accuracy is as high as 100%. Furthermore, most of the data proportion experiments (i.e 85%) produces better accuracy when using hybrid features as to the non-hybrid features. In the same way, around 89% cross-validation experiments (i.e 10, 15, and 20) that uses hybrid features achieve better classification. Moreover, the experiment shows that the classifier accuracy rate increases in parallel to the amount of data used for training.

However, there are a few cases where the hybrid features fail to produce higher than non-hybrid. Nevertheless, these results can still be considered good as most of them are over 90% accuracy. This may be explained by the nature of the utilized classifiers such as Decision Tree.

In addition, this dissertation examined the features constructed based on Edge Rank concept where it is of 16 features (i.e. FeatureSet-ER). By using ER features only, the classification accuracy of the employed classifiers can be improved. In particular, around 68% of the experiments in various data proportion (70:30, 80:20, and 90:10) obtained better result as compared to using non-hybrid features. In addition, results on 77% of cross validation experiments were also improved. However, the success rate of FeatureSet-H is still considered better compared to FeatureSet-ER, this is because FeatureSet-ER only relies on 16 features while FeatureSet-H employs 30 features.

Additionally, as the hybrid feature set contains a large number of features (30), this dissertation performs the feature selection method to reduce the size of FeatureSet-H. The results of hybrid features after process of feature selection (FeatureSet-HF) shows a slight improvement comparing with H. The experiments based on the data proportion 70:30, 80:20, and 90:10 present that around 64% of the experiments was either improved or have similar results. Moreover, only 2 experiments did not produce higher than 90% accuracy. Additionally, testing evaluation using cross-validation (10, 15, and 20) also shows that the reduced feature set is better. Only 5 experiments show a little decrease in accuracy. This dissertation examined the features of HF in order to simplify the process of training and testing. Through the experiments, this dissertation observed that the computational time

was reduced for most of the classifiers. However, this dissertation did not consider time as an evaluation metric as the main evaluation is made based on classification accuracy.

In order to support the findings, this dissertation performed T-Test analysis to determine whether there is a statistically significant difference on the classification accuracy between the feature sets. Based on T-Test results, it is learned that the hybrid feature sets FeatureSet-ER, FeatureSet-H, and FeatureSet-HF gives better insight on classifying malicious users as all classifiers produce better performance compared with FeatureSet-UB and FeatureSet-UBA. Moreover, a closer look at the T-Test results indicated that FeatureSet-H and FeatureSet-HF produce highly significant performance compared with FeatureSet-UB, FeatureSet-UBA, and FeatureSet-ER. In addition, the results also indicate that the accuracy difference between FeatureSet-H and FeatureSet-HF is not significant. Since this dissertation does not consider computational effort in evaluating the proposed features, it recommends the utilization of FeatureSet-H in detecting malicious users.

### 5.3 Summary

In this chapter, the classification experimental results and discussion have been presented. It started with the findings of split percentage methods for each feature set and followed by the results of cross-validation. In the next chapter, the conclusion, contribution, and future recommendation of this dissertation are presented.

# CHAPTER SIX
# CONCLUSION

This chapter presents the conclusion of the dissertation besides limitation of the study and future work that could be implemented. Section 6.1 includes the dissertation contribution whereas Section 6.2 illustrates the limitation. In the end, Section 6.3 describes recommendations for future work.

## 6.1 Contribution

Social media networks have become extremely popular and this creates the opportunity for the malicious user to publish unwanted content such as spam video. This dissertation presents two contributions that includes; 1) repository of YouTube channels and 2) a set of features to be used in detecting malicious users. The first contribution is as presented in Figure 3.12 where it contains data on users, how the users interact with other users and also information on the content that they shared on YouTube. On the other hand, the proposed feature set which comprises of 30 features (refer to Table 4.7) is based on user based, user behaviour, and Edge Rank concept which was employed in Facebook.

This dissertation examines the effectiveness of Edge Rank concept to construct new features for spam detection. The process of construction is performed by using three aspects of Edge Rank which is Affinity, Weight, and Decay (refer to Table 4.5). Total of 16 features have been constructed, where it was evaluated by CFS method and it seems that has the most interesting features. Hence, 10 out of 13 features of FeatureSet-HF is based on Edge Rank concept. This highlights the new features that never exist on YouTube before.

The empirical results revealed that the classification accuracy is more than 95% for most of the employed classifiers (i.e 22 classifiers). This indicates that the proposed features can be used to detect video spam in YouTube. Hence, creating the possibility for the features to be used in an anti-spam video system.

### 6.2 Limitations of Study

Although this dissertation has achieved its objectives, there is still some limitations. At the first, due to scraper tool and time limit, this study has examined only spam video threat. Second, the feature extraction technique that is employed in this dissertation could not automatically extract all features, as some features were collected manually. Besides that, there are also missing values in the identified features; data cannot be obtained due to poor connection or privacy issues. Third, this dissertation relies on human judgment to assign class value, hence, increasing computational effort and not sufficiently reliable.

### 6.3 Future Work

Even though this dissertation has presented a feature set to identify spam video, there are other challenges that still requires attention. This includes features on video replication and low-quality videos. What are the features that can be used to represent these types of videos? Furthermore, this dissertation focused on the video component only, however, there are other interesting sections such as the comments section in the YouTube. The information included in this section need to be examined in order to build a better recognition system.

## REFERENCES


Abdesslem, F. Ben, Parris, I., & Henderson, T. (2012). Reliable online social network data collection. In *Computational Social Networks: Mining and Visualization* (pp. 183–210). http://doi.org/10.1007/978-1-4471-4054-2_8

Academia. (2016). About academia.edu. Retrieved January 1, 2016, from http://www.academia.edu/about

Acuña, E., & Rodriguez, C. (2004). The Treatment of Missing Values and its Effect on Classifier Accuracy. In *Classification, Clustering, and Data Mining Applications* (pp. 639–647). incollection, Springer. http://doi.org/10.1007/978-3-642-17103-1_60

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.

Amasyali, M. F., & Ersoy, O. K. (2011). Comparison of single and ensemble classifiers in terms of accuracy and execution time. In *INISTA 2011 - 2011 International Symposium on INnovations in Intelligent SysTems and Applications* (pp. 470–474). IEEE. http://doi.org/10.1109/INISTA.2011.5946119

Babu, T. A. F., & Pradeepa, R. (2013). Comparative study of multiclass classifiers for underwater target classification. In *2013 Third International Conference on Advances in Computing and Communications* (pp. 400–403). IEEE. http://doi.org/10.1109/ICACC.2013.85

Benevenuto, F., Magno, G., Rodrigues, T., & Almeida, V. (2010). Detecting spammers on twitter. In *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)* (Vol. 6, p. 12). http://doi.org/10.1.1.297.5340

Benevenuto, F., Rodrigues, T., Almeida, J., Gonçalves, M., & Almeida, V. (2009). Detecting spammers and content promoters in online video social networks. In *Proceedings - IEEE INFOCOM*. http://doi.org/10.1109/INFCOMW.2009.5072127

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.

Bermejo, P., Joho, H., Jose, J. M., & Villa, R. (2009). Comparison of feature construction methods for video relevance prediction. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 5371 LNCS*, 185–196.

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

Bhat, S. Y., Abulaish, M., & Mirza, A. a. (2014). Spammer classification using ensemble methods over structural social network features. In *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)* (pp. 454–458). http://doi.org/10.1109/WI-IAT.2014.133

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. http://doi.org/10.1023/A:1010933404324

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

Cao, C., & Caverlee, J. (2015). Detecting spam URLs in social media via behavioral analysis. *Advances in Information Retrieval*, *9022*, 703–714. http://doi.org/10.1007/978-3-319-16354-3_77

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM. *ACM Transactions on Intelligent Systems and Technology*, *2*(3), 1–27. http://doi.org/10.1145/1961189.1961199

Chawla, N. V. (2005). Data Mining for Imbalanced Datasets: An Overview. In *Data Mining and Knowledge Discovery Handbook* (pp. 853–867). http://doi.org/10.1007/0-387-25465-X_40

Chiluka, N., Andrade, N., & Pouwelse, J. (2011). A link prediction approach to

recommendations in large-scale user-generated content systems. In *Advances in Information Retrieval* (pp. 189–200).

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).

Cohen, W. W. (1995). Fast effective rule induction. In *Twelfth International Conference on Machine Learning* (pp. 115–123). http://doi.org/10.1.1.50.8204

Digitalinsights. (2014). Social Media 2014 Statistics. Retrieved January 1, 2016, from http://blog.digitalinsights.in/social-media-users-2014-stats-numbers/05205287.html

Diplaris, S., Tsoumakas, G., Mitkas, P. A., & Vlahavas, I. (2005). Protein classification with multiple algorithms. In *Panhellenic Conference on Informatics* (pp. 448–456). http://doi.org/10.1007/11573036_42

Dong, A., & Wang, B. (2009). Feature selection and analysis on mammogram classification. *IEEE Pacific RIM Conference on Communications, Computers, and Signal Processing - Proceedings*, 731–735.

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

Facebook, I. (2016). Help Center. Retrieved February 18, 2016, from https://www.facebook.com/help/

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

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., & Lin, C.-J. (2008). LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, *9*, 1871–1874. http://doi.org/10.1038/oby.2011.351

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

Frank, E., & Witten, I. (1998). Generating accurate rule sets without global optimization. Retrieved from http://researchcommons.waikato.ac.nz/handle/10289/1047

Freitas, A. A. (2001). Understanding the crucial role of attributeInteraction in data mining. *Artificial Intelligence Review*, *16*(3), 177–199. Retrieved from http://portal.acm.org/citation.cfm?id=508382.508383

Freund, Y., & Schapire, R. (1996). Experiments with a new boosting algorithm. *International Conference on Machine Learning*, *96*, 148–156.

Gayle, D. (2012). YouTube cancels billions of music industry video views after finding they were fake or "dead." Retrieved January 1, 2016, from http://www.dailymail.co.uk/sciencetech/article-2254181/YouTube-wipes-billions-video-views-finding-faked-music-industry.html

George-Nektarios, T. (2013). *Weka Classifiers Summary*. *Athens University of Economics and Bussiness Intracom-Telecom*. Athens.

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

Guo, Y., Zhou, L., He, K., Gu, Y., & Sun, Y. (2014). Bayesian spam filtering mechanism based on decision tree of attribute set dependence in the mapreduce framework. *Open Cybernetics & Systemics Journal*, *8*, 435–441.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*, 1157–1182. http://doi.org/10.1162/153244303322753616

Hall, M. (1999). Correlation-based Feature Selection for Machine Learning. *Methodology*, *21i195-i20*(April), 1–5. http://doi.org/10.1.1.37.4643

Hall, M. (2000). Correlation-based feature selection of discrete and numeric class machine learning. article.

Hassan, S., & El Fattah Hegazy, A. (2015). A model recommends best machine learning algorithm to classify learners based on their interactivity with moodle. In *2015 Second International Conference on Computing Technology and Information Management (ICCTIM)* (pp. 49–54). IEEE. http://doi.org/10.1109/ICCTIM.2015.7224592

Heymann, P., Koutrika, G., & Garcia-Molina, H. (2007). Fighting spam on social websites:

A survey of approaches and future challenges. *IEEE Internet Computing*, *11*(6), 36–45.

Hi5. (2016). About us. Retrieved January 1, 2016, from http://www.hi5.com/

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., Gao, H., & Liu, H. (2015). Social spammer detection with sentiment information. In *Proceedings - IEEE International Conference on Data Mining, ICDM* (Vol. 2015–Janua, pp. 180–189). http://doi.org/10.1109/ICDM.2014.141

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

Ibarguren, I., Pérez, J. M., Muguerza, J., Gurrutxaga, I., & Arbelaitz, O. (2014). Coverage-based resampling: Building robust consolidated decision trees. *Knowledge-Based Systems*, *79*, 51–67. http://doi.org/10.1016/j.knosys.2014.12.023

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

John G. Cleary, L. E. T. (1995). K*: An Instance-based Learner Using an Entropic Distance Measure. In *Proceedings of The 12th International Conference on Machine Learning* (pp. 108--114).

Khakham, P., Chumuang, N., & Ketcham, M. (2015). Isan Dhamma Handwritten Characters Recognition System by Using Functional Trees Classifier. In *2015 11th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)* (pp. 606–612). IEEE. http://doi.org/10.1109/SITIS.2015.68

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.

Korb, K. B., & Nicholson, A. E. (2011). *Bayesian artificial intelligence* (Second Edi). CRC Press.

Kumar, R., Naik, S. M., Naik, V. D., Shiralli, S., Sunil V.G, & Husain, M. (2015). Predicting clicks: CTR estimation of advertisements using Logistic Regression

classifier. In *2015 IEEE International Advance Computing Conference (IACC)* (pp. 1134–1138). IEEE. http://doi.org/10.1109/IADCC.2015.7154880

Lavrač, N., Džeroski, S., & Grobelnik, M. (1991). Learning nonrecursive definitions of relations with LINUS. In *European Working Session on Learning* (pp. 265–281). inproceedings.

Lee, S., & Kim, J. (2012). W ARNING B IRD : Detecting suspicious URLs in Twitter stream. In *NDSS Symposium 2012* (pp. 1–13).

LinkedIn. (2016). About Us. Retrieved January 1, 2016, from https://www.linkedin.com/about-us?trk=uno-reg-guest-home-about

Martin, B. (1995). *Instance-based learning: nearest neighbour with generalisation* (No. 95/18).

McCord, M., & Chuah, M. (2011). Spam detection on twitter using traditional classifiers. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *6906 LNCS*, 175–186. Journal Article.

Myspace. (2015). Press Releases. Retrieved January 1, 2016, from https://myspace.com/pressroom/pressreleases

Nexgate. (2013). *State of social media spam*.

Nisa, I. U., & Ahsan, S. N. (2015). Fault prediction model for software using soft computing techniques. In *2015 International Conference on Open Source Systems & Technologies (ICOSST)* (pp. 78–83). IEEE. http://doi.org/10.1109/ICOSST.2015.7396406

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

Pagallo, G. (1989). Learning DNF by Decision Trees. In *IJCAI* (Vol. 89, pp. 639–644). inproceedings.

Pfahringer, B., Holmes, G., & Kirkby, R. (2007). New options for Hoeffding Trees. In *AI*

*2007: Advances in Artificial Intelligence* (pp. 90–99). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-540-76928-6_11

Platt, J. (1998). Sequential Minimal Optimization: A fast algorithm for training Support Vector Machines. In *Advances in kernel methods* (pp. 185–208). techreport. Retrieved from https://www.microsoft.com/en-us/research/publication/sequential-minimal-optimization-a-fast-algorithm-for-training-support-vector-machines/

Quinlan, J. (1993). C4. 5: programs for machine learning. In *Machine Learning* (p. 302). Elsevier.

Razmara, M., Asadi, B., Narouei, M., & Ahmadi, M. (2012). A novel approach toward spam detection based on iterative patterns per text. In *International eConference on Computer and Knowledge Engineering (ICCKE)* (Vol. 3, pp. 3–8). IEEE.

Rodrigues, T., Benevenuto, F., Cha, M., Gummadi, K., & Almeida, V. (2011). On word-of-mouth based discovery of the web. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference - IMC '11* (p. 381).

Roth, D., & Small, K. (2009). Interactive feature space construction using semantic information. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning* (pp. 66–74). inproceedings.

Saab, S. A., Mitri, N., & Awad, M. (2014). Ham or spam? A comparative study for some content-based classification algorithms for email filtering. In *Proceedings of the Mediterranean Electrotechnical Conference - MELECON* (pp. 439–443). http://doi.org/10.1109/MELCON.2014.6820574

Salih, A., & Abraham, A. (2014). Novel ensemble decision support and health care monitoring system. *Journal of Network and Innovative*, *2(2014)*, 041–051.

Sandvine. (2015). *Global internet phenomena report: North America and latin America*.

Shams, R., & Mercer, R. E. (2013). Classifying spam emails using text And readability features. In *2013 IEEE 13th International Conference on Data Mining* (pp. 657–666). IEEE. http://doi.org/10.1109/ICDM.2013.131

Sia, F., Alfred, R., Yu, L., & Fun, T. S. (2012). A variable length feature construction

method for data summarization using DARA. In *Computing and Convergence Technology (ICCCT), 2012 7th International Conference on* (pp. 881–887). Seoul: IEEE.

Singh, M., Bansal, D., & Sofat, S. (2014). Detecting malicious users in Twitter using classifiers. In *7th International Conference on Security of Information and Networks* (p. 247).

Smith, C. (2015). Statistics of Social Networking Sites. *DMR*. Retrieved from http://expandedramblings.com/index.php/march-2013-by-the-numbers-a-few-amazing-twitter-stats/3/#.U3xVv9KSyuE

Socialbakers. (2015). EdgeRank checker. Retrieved from https://www.socialbakers.com/edgerankchecker/edgerank/learn

Soman, S. J., & Murugappan, S. (2014). A study of spam detection algorithm on social media networks. *Journal of Computer Science*, *10*(10), 2135–2140.

Sondhi, P. (2010). Feature construction methods: a survey. *Sifaka. Cs. Uiuc. Edu*, *69*, 70–71.

Statista. (2016). Number of monthly active Facebook users worldwide as of 4th quarter 2015 (in millions). Retrieved January 1, 2016, from http://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/

Strano, M., & Colosimo, B. M. (2006). Logistic regression analysis for experimental determination of forming limit diagrams. *International Journal of Machine Tools and Manufacture*, *46*(6), 673–682. http://doi.org/10.1016/j.ijmachtools.2005.07.005

Stringhini, G., Kruegel, C., & Vigna, G. (2010). Detecting spammers on social networks. In *Proceedings of the 26th Annual Computer Security Applications Conference* (pp. 1–9). inproceedings.

Sureka, A. (2011). Mining user comment activity for detecting forum spammers in youtube. *Arxiv - Computers & Society*, 0–3. Retrieved from http://arxiv.org/abs/1103.5044

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).

Tan, E., Guo, L., Chen, S., Zhang, X., & Zhao, Y. E. (2012). Spammer behavior analysis and detection in user generated content on social networks. In *Distributed Computing Systems (ICDCS), 2012 IEEE 32nd International Conference on* (pp. 305–314). IEEE.

Tretyakov, K. (2004). Machine learning techniques in spam filtering. *Data Mining Problem-Oriented Seminar, MTAT*, (May), 60–79.

Twitter. (2015). COMPANY FACTS. Retrieved January 1, 2016, from https://about.twitter.com/company

Tynan, D. (2012). Social spam is taking over the Internet. Retrieved February 1, 2016, from http://www.itworld.com/article/2832566/it-management/social-spam-is-taking-over-the-internet.html

UK, H. I. (2008). *A study of social networks scams*.

Ulrike, G. (2001). *Social network analysis: Introduction and resources*. *Analysis*. Retrieved from http://lrs.ed.uiuc.edu/tse-portal/analysis/social-network-analysis/

Vafaie, H., & De Jong, K. (1995). Genetic algorithms as a tool for restructuring feature space representations. In *Tools with Artificial Intelligence, 1995. Proceedings., Seventh International Conference on* (pp. 8–11). inproceedings.

Villuendas, Y., Yanez, C., & Rey, C. (2015). Attributes and cases selection for social data classification. *IEEE Latin America Transactions*, *13*(10), 3370–3381. http://doi.org/10.1109/TLA.2015.7387244

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

Webb, G. (1999). Decision tree grafting from the all-tests-but-one partition. In *IJCAI* (Vol. 2, pp. 702–707).

Weber, B. G., & Mateas, M. (2009). A data mining approach to strategy prediction. In *2009 IEEE Symposium on Computational Intelligence and Games* (pp. 140–147). IEEE. http://doi.org/10.1109/CIG.2009.5286483

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. (Elsevier, Ed.) (2 edition). book, San Francisco: Elsevier.

Witten, I. H., Frank, E., Trigg, L., Hall, M., Holmes, G., & Cunningham, S. J. (1999). Weka : Practical machine learning tools and techniques with Java implementations. *Seminar*, *99*, 192–196.

Wuest, C. (2010). *The Risks of Social Networking*. *Symantec Security Response*.

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

YouTube. (2015). Statistics. Retrieved February 1, 2016, from https://www.youtube.com/yt/press/en-GB/statistics.html

Yuan, G.-X., Ho, C.-H., & Lin, C.-J. (2012). Recent advances of large-scale linear classification. In *Proceedings of the IEEE* (Vol. 100, pp. 2584–2603). http://doi.org/10.1109/JPROC.2012.2188013

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

Zhu, Y., Wang, X., Zhong, E., Liu, N., Li, H., & Yang, Q. (2012). Discovering spammers in social networks. In *Association for the Advancement of Artificial Intelligence* (pp. 171–177).

# Appendix A

## Sample FeatureSet-UB

| | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Channel_ID | Total_Subscribers | Total_Views | Ch_Discussion | Ch_Playlist | Ch_Description | Ch_Country | Ch_Links | Ch_Profile_Pic | Ch_Background_Pic | Class |
| 2 | UCd3PaZEcxNru29bA4--WL_A | 190,350 | 58,535,481 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | FALSE |
| 3 | UCjxrFnMg_scE7fkw_Ip0_yA | 290,167 | 35,532,789 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | FALSE |
| 4 | UCbakp3km-hUazI0wBDZEiZQ | 53 | 13,903 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | TRUE |
| 5 | UComP_epzeKzvBX156r6pm1Q | 10,840,601 | 3,725,951,730 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | FALSE |
| 6 | UCCwBMVm11X_XKCtYYoOR50w | 2,330 | 736,375 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | FALSE |
| 7 | UCRHd9bQqKDxJVmqOAdUq-CQ | 21,314 | 4,226,461 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | FALSE |
| 8 | UCnJ-KJLPlRw90rGs_6XfmmQ | 708,499 | 134,244,707 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | FALSE |
| 9 | UCVztq6euJLZjywe26qh8uqQ | 548 | 876,776 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | FALSE |
| 10 | UCORp0Fxa5lHFuZfLHxfeLuQ | 86,040 | 21,850,631 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | FALSE |
| 11 | UCdbd4O5KnIHRWUTtpoF8sJA | 65,507 | 26,643,107 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | FALSE |
| 12 | UC8cXcdMGLC8lO1E_l6qWcag | 5,705 | 927,238 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | FALSE |
| 13 | UCw3EildEnRRQTMBeFeUdJOQ | 9,794 | 162,082 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | FALSE |
| 14 | UC9Yno23nxsL3KN2d40DRDtw | 3,214 | 593,711 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | FALSE |
| 15 | UCPeLZYqHrQdV4xdZ2Lj3Reg | 98,526 | 75,247,201 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | FALSE |
| 16 | UCa8nqCmiWvaA8rnrRCySQsw | 4,198 | 787,438 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | FALSE |
| 17 | UCnW8ZhxajcBZfGBiurL1mQQ | 321 | 94,944 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | FALSE |
| 18 | UCEB2wiZCXUa8-hD-jsKBk6g | 101 | 14,802 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | FALSE |
| 19 | UCe1wNQ3iVR-LjT-nGelDRuQ | 20 | 395 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | TRUE |
| 20 | UCbgIijbm5VgB2dy6142HEbg | 3,778 | 995,339 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | FALSE |
| 21 | UCfh05rH15bSAlUjE0xt08HQ | 50 | 8,466 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | TRUE |
| 22 | UC0CKsihx4jpE6k6XpZyVbGQ | 5 | 9,744 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | TRUE |
| 23 | UCjYuzWEBTpM4llxH0mLNLGw | 108,913 | 783,796 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | FALSE |
| 24 | UCpmZQGTZXn9xd4nN59pblWQ | 256,325 | 23,740,090 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | FALSE |
| 25 | UCYtiPkKhgkFzMa0q3Gdn5sQ | 0 | 3,204 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | TRUE |
| 26 | UCurN9ADMkzOTkQ6MyB5iRcA | 1,293 | 923 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | FALSE |
| 27 | UCP6YrHRe_gjDuBayaP6dNzQ | 3,532 | 616,857 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | FALSE |
| 28 | UCqQHMj1WtxOoJ6N7NRFKu-A | 171,968 | 174,440,213 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | FALSE |
| 29 | UCa8nqCmiWvaA8rnrRCySQsw | 4,210 | 789,888 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | FALSE |
| 30 | UCxueNeCdh5_4-JuCrvWsjrA | 21,230 | 2,302,742 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | FALSE |

# Appendix B

## Sample FeatureSet-UBA

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | Channel_ID | Total_Subscribers | Total_Videos_Shared | Total_Videos_Likes | Total_Videos_Dislikes | Class |
| 2 | UCd3PaZEcxNru29bA4--WL_A | 190,350 | 842,646 | 364,943 | 29,283 | FALSE |
| 3 | UCjxrFnMg_scE7fkw_Ip0_yA | 290,167 | 97,461 | 586,158 | 26,337 | FALSE |
| 4 | UCbakp3km-hUazI0wBDZEiZQ | 53 | 2 | 55 | 126 | TRUE |
| 5 | UComP_epzeKzvBX156r6pm1Q | 10,840,601 | 15,742,117 | 19,580,695 | 702,417 | FALSE |
| 6 | UCCwBMVm11X_XKCtYYoOR50w | 2,330 | 1,505 | 1,254 | 540 | FALSE |
| 7 | UCRHd9bQqKDxJVmqOAdUq-CQ | 21,314 | 20,196 | 18,462 | 1,794 | FALSE |
| 8 | UCnJ-KJLPlRw90rGs_6XfmmQ | 708,499 | 148,363 | 1,130,754 | 20,207 | FALSE |
| 9 | UCVztq6eujLZjywe26qh8uqQ | 548 | 89 | 1,674 | 149 | FALSE |
| 10 | UCORp0Fxa5lHFuZfLHxfeLuQ | 86,040 | 60,174 | 181,828 | 3,778 | FALSE |
| 11 | UCdbd4O5KnlHRWUTtpoF8slA | 65,507 | 18,449 | 396,486 | 4,611 | FALSE |
| 12 | UC8cXcdMGLC8lO1E_l6qWcag | 5,705 | 138 | 4,042 | 154 | FALSE |
| 13 | UCw3EildEnRRQTMBeFeUdJOQ | 9,794 | 393 | 1,566 | 94 | FALSE |
| 14 | UC9Yno23nxsL3KN2d40DRDtw | 3,214 | 2,535 | 3,170 | 457 | FALSE |
| 15 | UCPeLZYqHrQdV4xdZ2Lj3Reg | 98,526 | 67,959 | 77,901 | 27,753 | FALSE |
| 16 | UCa8nqCmiWvaA8rnrRCySQsw | 4,198 | 416 | 3,325 | 100 | FALSE |
| 17 | UCnW8ZhxajcBZfGBiurL1mQQ | 321 | 23 | 232 | 52 | FALSE |
| 18 | UCEB2wiZCXUa8-hD-jsKBk6g | 101 | 100 | 153 | 6 | FALSE |
| 19 | UCe1wNQ3iVR-LjT-nGeIDRuQ | 20 | 3 | 10 | 21 | TRUE |
| 20 | UCbglijbm5VgB2dy614zHEbg | 3,778 | 2,871 | 7,248 | 2,471 | FALSE |
| 21 | UCfh05rH15bSAlUjE0xt08HQ | 50 | 3 | 5 | 9 | TRUE |
| 22 | UC0CKsihx4jpE6k6XpZyVbGQ | 5 | 4 | 2 | 7 | TRUE |
| 23 | UCjYuzWEBTpM4llxH0mLNLGw | 108,913 | 3,734 | 74,050 | 6,365 | FALSE |
| 24 | UCpmZQGTZXn9xd4nN59pblWQ | 256,325 | 67,951 | 212,234 | 3,685 | FALSE |
| 25 | UCYtiPkKhgkFzMa0q3Gdn5sQ | 0 | 1 | 2 | 9 | TRUE |

# Appendix C

## Sample FeatureSet-ER

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Channel_ID | Ch_Age | Avrage_Upload | Subscribe_Rate | Subscribe_ | Subscribe_ | View_Rate | Share_Rat | Share_Rat | Share_Rat | Like_Rate | Like_Rate | Like_Rate | Dislike_Ra | Dislike_Ra | Dislike_Ra | Class |
| 2 | UCd3PaZEcxNnu29bA4--WL_ | 531 | 0.116760829 | 3070.16129 | 358.4746 | 0.003252 | 110236.3 | 0.014395 | 1586.904 | 13591.06 | 0.006235 | 687.275 | 5886.177 | 0.0005 | 55.14689 | 472.3065 | FALSE |
| 3 | UCjxrFnMg_scE7fkw_Ip0_yA | 763 | 0.131061599 | 2901.67 | 380.2975 | 0.008166 | 46569.84 | 0.002743 | 127.7339 | 974.61 | 0.016496 | 768.228 | 5861.58 | 0.000741 | 34.51769 | 263.37 | FALSE |
| 4 | UCbakp3km-hUazIOwBDZEiz | 36 | 0.527777778 | 2.789473684 | 1.472222 | 0.003812 | 386.1944 | 0.000144 | 0.055556 | 0.105263 | 0.003956 | 1.527778 | 2.894737 | 0.009063 | 3.5 | 6.631579 | TRUE |
| 5 | UComP_epzeKzvBX156r6pm | 2328 | 0.010309278 | 451691.7083 | 4656.616 | 0.002909 | 1600495 | 0.004225 | 6762.078 | 655921.5 | 0.005255 | 8410.951 | 815862.3 | 0.000189 | 301.7255 | 29267.38 | FALSE |
| 6 | UCCwBMVm11X_XKCtYYoO | 271 | 0.084870849 | 101.3043478 | 8.597786 | 0.003164 | 2717.251 | 0.002044 | 5.553506 | 65.43478 | 0.001703 | 4.627306 | 54.52174 | 0.000733 | 1.99262 | 23.47826 | FALSE |
| 7 | UCRHd9bQqKDxlVmqOAdUc | 1338 | 0.014200299 | 1121.789474 | 15.92975 | 0.005043 | 3158.79 | 0.004778 | 15.09417 | 1062.947 | 0.004368 | 13.79821 | 971.6842 | 0.000424 | 1.340807 | 94.42105 | FALSE |
| 8 | UCnJ-KULPlRw90rGs_6Xfmm | 1571 | 0.013367282 | 33738.04762 | 450.986 | 0.005278 | 85451.75 | 0.001105 | 94.43857 | 7064.905 | 0.008423 | 719.767 | 53845.43 | 0.000151 | 12.86251 | 962.2381 | FALSE |
| 9 | UCVztq6eujLZjywe26qh8uqC | 583 | 0.008576329 | 109.6 | 0.939966 | 0.000625 | 1503.904 | 0.000102 | 0.152659 | 17.8 | 0.001909 | 2.871355 | 334.8 | 0.00017 | 0.255575 | 29.8 | FALSE |
| 10 | UCORp0Fxa5lHFuZfLHxfeLuQ | 1276 | 0.065047022 | 1036.626506 | 67.42947 | 0.003938 | 17124.32 | 0.002754 | 47.15831 | 724.988 | 0.008321 | 142.4984 | 2190.699 | 0.000173 | 2.960815 | 45.51807 | FALSE |
| 11 | UCdbd4O5KnIHRWUTtpoF8t | 3034 | 0.031641397 | 682.3645833 | 21.59097 | 0.002459 | 8781.512 | 0.000692 | 6.080751 | 192.1771 | 0.014881 | 130.6809 | 4130.063 | 0.000173 | 1.519776 | 48.03125 | FALSE |
| 12 | UC8eXcdMGLC8lO1E_l6qWc | 330 | 0.127272727 | 135.8333333 | 17.28788 | 0.006153 | 2809.812 | 0.000149 | 0.418182 | 3.285714 | 0.004359 | 12.24848 | 96.2381 | 0.000166 | 0.466667 | 3.666667 | FALSE |
| 13 | UCw3EildEnRRQTMBeFeUdJ | 1102 | 0.031760436 | 279.8285714 | 8.887477 | 0.060426 | 147.0799 | 0.002425 | 0.356624 | 11.22857 | 0.009662 | 1.421053 | 44.74286 | 0.00058 | 0.085299 | 2.685714 | FALSE |
| 14 | UC9Yno23nxsL3KN2d4ODRD | 346 | 0.095375723 | 97.39393939 | 9.289017 | 0.005413 | 1715.928 | 0.00427 | 7.32659 | 76.81818 | 0.005339 | 9.16185 | 96.06061 | 0.00077 | 1.320809 | 13.84848 | FALSE |
| 15 | UCPeLZYqHrQaV4xdZ2Lj3Re | 404 | 0.212871287 | 1145.651163 | 243.8762 | 0.001309 | 186255.4 | 0.000903 | 168.2153 | 790.2209 | 0.001035 | 192.8243 | 905.8256 | 0.000369 | 68.69554 | 322.7093 | FALSE |
| 16 | UCa8nqCmiWvaA8rnrRCySQ | 404 | 0.170792079 | 60.84057971 | 10.39109 | 0.005331 | 1949.104 | 0.000528 | 1.029703 | 6.028986 | 0.004223 | 8.230198 | 48.18841 | 0.000127 | 0.247525 | 1.449275 | FALSE |
| 17 | UCnW8ZhxajcBZfGBiurL1mC | 255 | 0.007843137 | 160.5 | 1.258824 | 0.003381 | 372.3294 | 0.000242 | 0.090196 | 11.5 | 0.002444 | 0.909804 | 116 | 0.000548 | 0.203922 | 26 | FALSE |
| 18 | UCEB2wiZCXUa8-hD-jsKBk6j | 43 | 0.023255814 | 101 | 2.348837 | 0.006823 | 344.2326 | 0.006756 | 2.325581 | 100 | 0.010336 | 3.55814 | 153 | 0.000405 | 0.139535 | 6 | FALSE |
| 19 | UCe1wNQ3iVR-LjT-nGelDRu | 1636 | 0.000611247 | 20 | 0.012225 | 0.050633 | 0.241443 | 0.007595 | 0.001834 | 3 | 0.025316 | 0.006112 | 10 | 0.053165 | 0.012836 | 21 | TRUE |
| 20 | UCbgIjbm5VgB2dy614zHEbg | 973 | 0.076053443 | 51.05405405 | 3.882837 | 0.003796 | 1022.959 | 0.002884 | 2.950668 | 38.7973 | 0.007282 | 7.449126 | 97.94595 | 0.002483 | 2.539568 | 33.39189 | FALSE |
| 21 | UCfh05rH15bSAlUjEDxt08HC | 250 | 0.012 | 16.66666667 | 0.2 | 0.005906 | 33.864 | 0.000354 | 0.012 | 1 | 0.000591 | 0.02 | 1.666667 | 0.001063 | 0.036 | 3 | TRUE |
| 22 | UCOCKsihx4jpE6k6XpZyVbGC | 264 | 0.003787879 | 5 | 0.018939 | 0.000513 | 36.90909 | 0.000411 | 0.015152 | 4 | 0.000205 | 0.007576 | 2 | 0.000718 | 0.026515 | 7 | TRUE |
| 23 | UCjYuzWEBTpM4lIxH0mLNL | 22 | 0.454545455 | 10891.3 | 4950.591 | 0.138956 | 35627.09 | 0.004764 | 169.7273 | 373.4 | 0.094476 | 3365.909 | 7405 | 0.008121 | 289.3182 | 636.5 | FALSE |
| 24 | UCpmZQGTZXn9xd4nN59pbl | 756 | 0.055555556 | 6102.97619 | 339.0542 | 0.010797 | 31402.24 | 0.002862 | 89.88228 | 1617.881 | 0.00894 | 280.7328 | 5053.19 | 0.000155 | 4.874339 | 87.7381 | FALSE |
| 25 | UCYtiPkKhgkFzMaOq3Gdn5st | 22 | 0.090909091 | 0 | 0 | 0 | 145.6364 | 0.000312 | 0.045455 | 0.5 | 0.000624 | 0.090909 | 1 | 0.002809 | 0.409091 | 4.5 | TRUE |

# Appendix D

## Sample FeatureSet-H

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Channel_ID | Ch_Age | Avrage_Upload | Subscribe_Rate | Subscribe_ | Subscribe_ | View_Rate | Share_Rat | Share_Rat | Share_Rat | Like_Rate | Like_Rate_ | Like_Rate_ | Dislike_Ra | Dislike_Ra | Dislike_Ra | Class |
| 2 | UCd3PaZEcxNru29bA4--WL_ | 531 | 0.116760829 | 3070.16129 | 358.4746 | 0.003252 | 110236.3 | 0.014395 | 1586.904 | 13591.06 | 0.006235 | 687.275 | 5886.177 | 0.0005 | 55.14689 | 472.3065 | FALSE |
| 3 | UCjxrFnMg_scE7fkw_IpO_yA | 763 | 0.131061599 | 2901.67 | 380.2975 | 0.008166 | 46569.84 | 0.002743 | 127.7339 | 974.61 | 0.016496 | 768.228 | 5861.58 | 0.000741 | 34.51769 | 263.37 | FALSE |
| 4 | UCbakp3km-hUazI0wBDZEiZ | 36 | 0.527777778 | 2.789473684 | 1.472222 | 0.003812 | 386.1944 | 0.000144 | 0.055556 | 0.105263 | 0.003956 | 1.527778 | 2.894737 | 0.009063 | 3.5 | 6.631579 | TRUE |
| 5 | UComP_epzeKzvBX156r6gom | 2328 | 0.010309278 | 451691.7083 | 4656.616 | 0.002909 | 1600495 | 0.004225 | 6762.078 | 655921.5 | 0.005255 | 8410.951 | 815862.3 | 0.000189 | 301.7255 | 29267.38 | FALSE |
| 6 | UCCwBMVm11X_XKCtYYoO | 271 | 0.084870849 | 101.3043478 | 8.597786 | 0.003164 | 2717.251 | 0.002044 | 5.553506 | 65.43478 | 0.001703 | 4.627306 | 54.52174 | 0.000733 | 1.99262 | 23.47826 | FALSE |
| 7 | UCRHd9bQqKDxJVmqOAdUc | 1338 | 0.014200299 | 1121.789474 | 15.92975 | 0.005043 | 3158.79 | 0.004778 | 15.09417 | 1062.947 | 0.004368 | 13.79821 | 971.6842 | 0.000424 | 1.340807 | 94.42105 | FALSE |
| 8 | UCnJ-KILPlRw9OrGs_6Xfmm | 1571 | 0.013367282 | 33738.04762 | 450.986 | 0.005278 | 85451.75 | 0.001105 | 94.43857 | 7064.905 | 0.008423 | 719.767 | 53845.43 | 0.000151 | 12.86251 | 962.2381 | FALSE |
| 9 | UCVztq6eujLZjywe26qh8uqC | 583 | 0.008576329 | 109.6 | 0.939966 | 0.000625 | 1503.904 | 0.000102 | 0.152659 | 17.8 | 0.001909 | 2.871355 | 334.8 | 0.00017 | 0.255575 | 29.8 | FALSE |
| 10 | UCORp0Fxa5lHFuZfLHxfeLuC | 1276 | 0.065047022 | 1036.626506 | 67.42947 | 0.003938 | 17124.32 | 0.002754 | 47.15831 | 724.988 | 0.008321 | 142.4984 | 2190.699 | 0.000173 | 2.960815 | 45.51807 | FALSE |
| 11 | UCdbd4O5KnlHRWUTtpoF8: | 3034 | 0.031641397 | 682.3645833 | 21.59097 | 0.002459 | 8781.512 | 0.000692 | 6.080751 | 192.1771 | 0.014881 | 130.6809 | 4130.063 | 0.000173 | 1.519776 | 48.03125 | FALSE |
| 12 | UC8cXcdMGLC8lO1E_l6qWc | 330 | 0.127272727 | 135.8333333 | 17.28788 | 0.006153 | 2809.812 | 0.000149 | 0.418182 | 3.285714 | 0.004359 | 12.24848 | 96.2381 | 0.000166 | 0.466667 | 3.666667 | FALSE |
| 13 | UCw3EildEnRRQTMBeFeUdJ | 1102 | 0.031760436 | 279.8285714 | 8.887477 | 0.060426 | 147.0799 | 0.002425 | 0.356624 | 11.22857 | 0.009662 | 1.421053 | 44.74286 | 0.00058 | 0.085299 | 2.685714 | FALSE |
| 14 | UC9Yno23nxsL3KN2d4ODRD | 346 | 0.095375723 | 97.39393939 | 9.289017 | 0.005413 | 1715.928 | 0.00427 | 7.32659 | 76.81818 | 0.005339 | 9.16185 | 96.06061 | 0.00077 | 1.320809 | 13.84848 | FALSE |
| 15 | UCPeLZYqHrQdV4xdZ2Lj3Re | 404 | 0.212871287 | 1145.651163 | 243.8762 | 0.001309 | 186255.4 | 0.000903 | 168.2153 | 790.2209 | 0.001035 | 192.8243 | 905.8256 | 0.000369 | 68.69554 | 322.7093 | FALSE |
| 16 | UCa8nqCmiWvaA8rnrRCySQ | 404 | 0.170792079 | 60.84057971 | 10.39109 | 0.005331 | 1949.104 | 0.000528 | 1.029703 | 6.028986 | 0.004223 | 8.230198 | 48.18841 | 0.000127 | 0.247525 | 1.449275 | FALSE |
| 17 | UCnW8ZhxajcBZfGBiurL1mC | 255 | 0.007843137 | 160.5 | 1.258824 | 0.003381 | 372.3294 | 0.000242 | 0.090196 | 11.5 | 0.002444 | 0.909804 | 116 | 0.000548 | 0.203922 | 26 | FALSE |
| 18 | UCEB2wiZCXUa8-hD-jsKBK6j | 43 | 0.023255814 | 101 | 2.348837 | 0.006823 | 344.2326 | 0.006756 | 2.325581 | 100 | 0.010336 | 3.55814 | 153 | 0.000405 | 0.139535 | 6 | FALSE |
| 19 | UCe1wNQ3iVR-LjT-nGelDRu | 1636 | 0.000611247 | 20 | 0.012225 | 0.050633 | 0.241443 | 0.007595 | 0.001834 | 3 | 0.025316 | 0.006112 | 10 | 0.053165 | 0.012836 | 21 | TRUE |
| 20 | UCbgIijbm5VgB2dy614zHEbg | 973 | 0.076053443 | 51.05405405 | 3.882837 | 0.003796 | 1022.959 | 0.002884 | 2.950668 | 38.7973 | 0.007282 | 7.449126 | 97.94595 | 0.002483 | 2.539568 | 33.39189 | FALSE |
| 21 | UCfh05rH15bSAlUjE0xt08HC | 250 | 0.012 | 16.66666667 | 0.2 | 0.005906 | 33.864 | 0.000354 | 0.012 | 1 | 0.000591 | 0.02 | 1.666667 | 0.001063 | 0.036 | 3 | TRUE |
| 22 | UCOCKsihx4ipE6k6XpZyVbGC | 264 | 0.003787879 | 5 | 0.018939 | 0.000513 | 36.90909 | 0.000411 | 0.015152 | 4 | 0.000205 | 0.007576 | 2 | 0.000718 | 0.026515 | 7 | TRUE |
| 23 | UCjYuzWEBTpM4IlxH0mLNL | 22 | 0.454545455 | 10891.3 | 4950.591 | 0.138956 | 35627.09 | 0.004764 | 169.7273 | 373.4 | 0.094476 | 3365.909 | 7405 | 0.008121 | 289.3182 | 636.5 | FALSE |
| 24 | UCpmZQGTZXn9xd4nN59pbl | 756 | 0.055555556 | 6102.97619 | 339.0542 | 0.010797 | 31402.24 | 0.002862 | 89.88228 | 1617.881 | 0.00894 | 280.7328 | 5053.19 | 0.000155 | 4.874339 | 87.7381 | FALSE |
| 25 | UCYtiPkKhgkFzMa0q3Gdn5s | 22 | 0.090909091 | 0 | 0 | 0 | 145.6364 | 0.000312 | 0.045455 | 0.5 | 0.000624 | 0.090909 | 1 | 0.002809 | 0.409091 | 4.5 | TRUE |

# Appendix E

## Sample FeatureSet-HF

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Channel_II | Ch_Age | Avrage_Uf | Subscribe_ | View_Rate | Share_Rat | Share_Rat | Like_Rate_ | Like_Rate_ | Like_Rate_ | Dislike_Ra | Ch_Discus | Ch_Playlis | Ch_Links | Class |
| 2 | UCd3PaZE | 531 | 0.11676 | 0.00325 | 110236 | 0.0144 | 13591.1 | 0.00623 | 687.275 | 5886.18 | 0.0005 | 1 | 1 | 1 | FALSE |
| 3 | UCjxrFnMf | 763 | 0.13106 | 0.00817 | 46569.8 | 0.00274 | 974.61 | 0.0165 | 768.228 | 5861.58 | 0.00074 | 1 | 1 | 1 | FALSE |
| 4 | UCbakp3k | 36 | 0.52778 | 0.00381 | 386.194 | 0.00014 | 0.10526 | 0.00396 | 1.52778 | 2.89474 | 0.00906 | 0 | 0 | 0 | TRUE |
| 5 | UComP_ef | 2328 | 0.01031 | 0.00291 | 1600495 | 0.00422 | 655922 | 0.00526 | 8410.95 | 815862 | 0.00019 | 1 | 1 | 1 | FALSE |
| 6 | UCCwBMV | 271 | 0.08487 | 0.00316 | 2717.25 | 0.00204 | 65.4348 | 0.0017 | 4.62731 | 54.5217 | 0.00073 | 1 | 0 | 0 | FALSE |
| 7 | UCRHd9bC | 1338 | 0.0142 | 0.00504 | 3158.79 | 0.00478 | 1062.95 | 0.00437 | 13.7982 | 971.684 | 0.00042 | 1 | 1 | 1 | FALSE |
| 8 | UCnJ-KJLP | 1571 | 0.01337 | 0.00528 | 85451.8 | 0.00111 | 7064.9 | 0.00842 | 719.767 | 53845.4 | 0.00015 | 1 | 1 | 1 | FALSE |
| 9 | UCVztq6eL | 583 | 0.00858 | 0.00063 | 1503.9 | 0.0001 | 17.8 | 0.00191 | 2.87136 | 334.8 | 0.00017 | 0 | 0 | 1 | FALSE |
| 10 | UCORp0Fx | 1276 | 0.06505 | 0.00394 | 17124.3 | 0.00275 | 724.988 | 0.00832 | 142.498 | 2190.7 | 0.00017 | 1 | 1 | 1 | FALSE |
| 11 | UCdbd4O5 | 3034 | 0.03164 | 0.00246 | 8781.51 | 0.00069 | 192.177 | 0.01488 | 130.681 | 4130.06 | 0.00017 | 1 | 1 | 1 | FALSE |
| 12 | UC8cXcdM | 330 | 0.12727 | 0.00615 | 2809.81 | 0.00015 | 3.28571 | 0.00436 | 12.2485 | 96.2381 | 0.00017 | 1 | 0 | 1 | FALSE |
| 13 | UCw3EildE | 1102 | 0.03176 | 0.06043 | 147.08 | 0.00242 | 11.2286 | 0.00966 | 1.42105 | 44.7429 | 0.00058 | 1 | 1 | 1 | FALSE |
| 14 | UC9Yno23 | 346 | 0.09538 | 0.00541 | 1715.93 | 0.00427 | 76.8182 | 0.00534 | 9.16185 | 96.0606 | 0.00077 | 0 | 0 | 1 | FALSE |
| 15 | UCPeLZYql | 404 | 0.21287 | 0.00131 | 186255 | 0.0009 | 790.221 | 0.00104 | 192.824 | 905.826 | 0.00037 | 1 | 1 | 1 | FALSE |
| 16 | UCa8nqCn | 404 | 0.17079 | 0.00533 | 1949.1 | 0.00053 | 6.02899 | 0.00422 | 8.2302 | 48.1884 | 0.00013 | 1 | 1 | 1 | FALSE |
| 17 | UCnW8Zh | 255 | 0.00784 | 0.00338 | 372.329 | 0.00024 | 11.5 | 0.00244 | 0.9098 | 116 | 0.00055 | 0 | 0 | 0 | FALSE |
| 18 | UCEB2wiZ | 43 | 0.02326 | 0.00682 | 344.233 | 0.00676 | 100 | 0.01034 | 3.55814 | 153 | 0.00041 | 0 | 0 | 0 | FALSE |
| 19 | UCe1wNQ | 1636 | 0.00061 | 0.05063 | 0.24144 | 0.00759 | 3 | 0.02532 | 0.00611 | 10 | 0.05316 | 0 | 0 | 1 | TRUE |
| 20 | UCbgIijbm | 973 | 0.07605 | 0.0038 | 1022.96 | 0.00288 | 38.7973 | 0.00728 | 7.44913 | 97.9459 | 0.00248 | 1 | 0 | 1 | FALSE |
| 21 | UCfh05rH | 250 | 0.012 | 0.00591 | 33.864 | 0.00035 | 1 | 0.00059 | 0.02 | 1.66667 | 0.00106 | 0 | 0 | 0 | TRUE |
| 22 | UCOCKsihx | 264 | 0.00379 | 0.00051 | 36.9091 | 0.00041 | 4 | 0.00021 | 0.00758 | 2 | 0.00072 | 0 | 0 | 0 | TRUE |
| 23 | UCjYuzWE | 22 | 0.45455 | 0.13896 | 35627.1 | 0.00476 | 373.4 | 0.09448 | 3365.91 | 7405 | 0.00812 | 0 | 1 | 1 | FALSE |
| 24 | UCpmZQG | 756 | 0.05556 | 0.0108 | 31402.2 | 0.00286 | 1617.88 | 0.00894 | 280.733 | 5053.19 | 0.00016 | 1 | 1 | 1 | FALSE |
| 25 | UCYtiPkKh | 22 | 0.09091 | 0 | 145.636 | 0.00031 | 0.5 | 0.00062 | 0.09091 | 1 | 0.00281 | 0 | 0 | 0 | TRUE |