EFFECTIVE METHODS TO DETECT CYBERBULLYING AND INFLUENTIAL SPREADERS IN AN ONLINE SOCIAL NETWORK

MOHAMMED ALI DERHEM AL-GARADI

FACULTY OF COMPUTER SCIENCE AND
INFORMATION TECHNOLOG
UNIVERSITY OF MALAYA
KUALA LUMPUR

2017

EFFECTIVE METHODS TO DETECT CYBERBULLYING AND INFLUENTIAL SPREADERS IN AN ONLINE SOCIAL NETWORK

MOHAMMED ALI DERHEM AL-GARADI

THESIS SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

FACULTY OF COMPUTER SCIENCEAND INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2017

UNIVERSITY OF MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: Mohammed Ali Derhem Al-Garadi

Matric No: WHA130057

Name of Degree: Doctor of Philosophy

Title of Thesis : EFFECTIVE METHODS TO DETECT CYBERBULLYING AND INFLUENTIAL SPREADERS IN AN ONLINE SOCIAL NETWORK

Field of Study: Information Retrieval

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date:

Subscribed and solemnly declared before,

Witness's Signature Date:

Name:

Designation

ABSTRACT

Online social networks (OSNs) have become massively popular. The characteristics of OSNs communication is a revolutionary trend exploiting the expanded capabilities of Web 2.0, which provide users with the flexibility and freedom to post, write, and construct large social network relations. On one hand, OSNs provide users with novel and large-scale social interactions, which is a concept previously considered impossible in terms of scale and extent. On the other hand, OSNs can be used by criminals as a platform to commit cybercrimes without physically facing their victims. OSNs serve as a medium to commit cybercrimes as well as a delivery mechanism. To tackle these emerging problems, this work proposes effective methods to detect cyberbullying and identify influential spreaders in OSNs. First, an effective method to detect cyberbullying is proposed by offering a unique set of significant features, which show improvement in the performance of machine learning classifiers when compared to baseline features. Although any user in such massively connected networks can be vulnerable to online misbehavior, hence applying detection methods for every node (user) of a network is impractical. Therefore, an effective controlling method is required along with the detection method. The information spreading controlling method is achieved by proposing an effective method to identify influential spreaders in OSNs. Identifying these users is significant to either hinder the diffusion of unwanted information, such as rumor and cyberbullying, or accelerate spreading and distribution of precautionary messages as part of cyberbullying prevention strategies. Thus, interaction weighted k-core method (IWK_S) is developed to identify influential spreaders in OSNs. The degree centrality, PageRank, original K-core, and developed IWK_S are compared by calculating their respective imprecision functions, and recognition rate $r(f)$ real OSN networks to verify the performance of each algorithm in recognizing influential

spreaders. The developed IWK_S performs better than other methods in identifying the most influential spreaders and in quantifying the spreading effectiveness of nodes. The proposed methods can be used to minimize the role of OSNs in the rise of cybercrimes and analyze human behavior in cyberspace. The proposed methods can be implemented in a wide range of applications that can be used by parents, guardians, educational institutions, and organizations as well as non-government organizations, including crime prevention foundations, social chamber organizations, psychiatric associations, policy makers, and enforcement bodies.

ABSTRAK

Rangkaian-rangkaian sosial (OSNs) dalam talian telah menjadi sangat popular pada beberapa tahun lepas. Ciri-ciri komunikasi OSNs adalah satu aliran revolusi yang mengambil manfaat daripada Web 2.0 di dalam memperluaskan keupayaan, yang menyediakan fleksibiliti pengguna dan kebebasan untuk menghantar, menulis, dan membina hubungan rangkaian sosial yang besar. Secara khususnya dengan cara ini, OSNs boleh menyediakan novel pengguna dan interaksi sosial yang besar, yang sebelum ini dianggap mustahil dari segi skala dan tahap. Sebaliknya, OSN boleh digunakan oleh penjenayah sebagai landasan untuk melakukan jenayah siber dan merebak tanpa melibatkan fizikal mangsa. OSNs berfungsi sebagai medium untuk melakukan jenayah siber serta mekanisme penghantaran. Untuk menangani cabaran ini, kerja ini mencadangkan kaedah yang berkesan untuk mengesan pembuli siber dan pengaruh penyebar di dalam konteks OSNs. Pertama, kaedah pembelajaran mesin berkesan untuk mengesan pembuli siber dicadangkan. Matlamat ini dicapai dengan mencadangkan satu set ciri-ciri unik yang ketara, yang menunjukkan peningkatan dalam prestasi pengelasan pembelajaran mesin berbanding dengan ciri-ciri atas garis dasar. Walau bagaimanapun, membangunkan kaedah pengesanan yang berkesan sahaja di dalam rangkaian yang kompleks seperti (OSNs) yang terdiri daripada berbilion pengguna tidak mencukupi. Mana-mana pengguna dalam apa-apa rangkaian secara besar-besaran yang disambungkan boleh terdedah kepada kelakuan yang tidak senonoh di atas talian; Di samping itu, menggunakan kaedah pengesanan di setiap nod (pengguna) rangkaian adalah penyelesaian yang tidak praktikal. Oleh yang demikian, kaedah pengawalan yang berkesan bersama-sama dengan kaedah-kaedah pengesanan diperlukan. Kaedah kawalan dicapai dengan mencadangkan kaedah yang berkesan untuk mengenal pasti penyebar yang berpengaruh. Mengenal pasti pengguna-pengguna ini adalah penting untuk menghalang penyebaran maklumat yang tidak diingini, seperti,

desas-desus dan pembuli siber, sama ada atau untuk mempercepatkan penyebaran seperti menyebarkan mesej langkah berjaga-jaga untuk strategi pencegahan pembuli siber. Oleh itu, satu kaedah wajaran *K-core* (IWKs) dibangunkan untuk mengenal pasti penyebar yang berpengaruh di dalam OSN. Pengesanan penyebaran pautan dalam menyebarkan maklumat dinamik sebenar mengesahkan keberkesanan kaedah kami cadangkan untuk mengenal pasti penyebar yang berpengaruh di OSNs berbanding keutamaan pusat darjah, *PageRank*, dan *K-core*. Kaedah yang dicadangkan boleh dilaksanakan dalam pelbagai aplikasi yang boleh digunakan oleh ibu bapa, penjaga, institusi pendidikan, dan organisasi serta pertubuhan-pertubuhan bukan kerajaan, termasuk asas pencegahan jenayah, organisasi ruang sosial, persatuan psikiatri, pembuat dasar, dan badan-badan penguatkuasaan.

ACKNOWLEDGEMENTS

First, I would extend my thanks and immense gratitude to Allah for endowing me the strength, wisdom, and endless blessings to do my PhD. I would like also to sincerely express my deepest gratitude to my supervisors Dr. Kasturi Dewi Varathan and Dr. Sri Devi Ravana for their precious support, supervision, encouragement and inspirations to me during my PhD journey. Their continuous support and guidance helped me producing a valuable piece of research conveyed in this thesis.

As well, I would like to extend heartiest appreciation to my parents whom have sacrificed their time, efforts to train me to become a person of importance and value to the society. They are always offering me continuous helps and supports in every oppressive and repressive moment.

My gratitude to my darling wife Taghreed Abdallah Al-sharafi and my lovely daughter Nuha Mohammed for their patience throughout my studies. Thank you for your inspirations and motivations.

This thesis is dedicated to my father Ali Derhem , my mother Gareesa Abdallah for their endless support and motivation, and my beloved wife for her emotional support and unconditional love.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

OSNs	:	Online social networks
SVM	:	Support vector machine
LIBSVM	:	Library for Support Vector Machines
NB	:	Naive Bayes
RF	:	Random forest
KNN	:	K-nearest neighbors
SMOTE	:	Synthetic minority over-sampling technique
AUC	:	Area under curve
ROC	:	Receiver operating curve
TP	:	True positive
TN	:	True negative
FP	:	False positive
FN	:	False negative
IWK_S	:	Interaction weighted k-core method
$\epsilon(p)$	:	Imprecision functions
r(f)	:	Recognition rate

CHAPTER 1: INTRODUCTION

This chapter is organized into six sections. Section 1.1 discusses the introduction of OSNs and cybercrimes in OSNs. Section 1.2 provides the motivations for this research. Section 1.3 highlights the research gap and discusses the problem statement. Section 1.4 lists the objectives of the study. Section 1.5 presents the significance of this research. Section 1.6 states the organization of the thesis.

1.1 Introduction

Online social networks (OSNs) have become massively popular. Billions of users access these websites as innovative communication tools and real-time, dynamic data sources in which they can create profiles and communicate with other users regardless of geographical location and physical limitations. In this regard, these websites have become vital, ubiquitous communication platforms. Accordingly, human communications facilitated by OSNs can exceptionally address the temporal and spatial limitations of traditional communications. OSNs provide the researchers with novel insights into the construction of social networks and societies, which is previously considered impossible in terms of scale and extent. These digital tools can transcend the boundaries of the physical world that previously hindered or slowed down communication among people (H. W. Lauw, J. C. Shafer, R. Agrawal, & A. Ntoulas, 2010).

OSNs refer to a combination of three elements: content, user communities, and Web 2.0 technologies (Ahlqvist, Bäck, Halonen, & Heinonen, 2008). These networks provide the means of interaction among people in which they create, share, and exchange information in a virtual world. The popularity of OSNs is due to its simplicity; with basic internet skills, users can create and manage their social media account. The characteristics of OSNs transcend the boundaries of the physical world, allowing for an

in-depth observation of large-scale human relationships and behaviors (H. Lauw, J. C. Shafer, R. Agrawal, & A. Ntoulas, 2010).

OSNs have become dynamic social interaction platforms for billions of users worldwide. The information and ideas are disseminated among these users rapidly through online social interactions. The online interactions among OSN users generate a huge volume of data that provide opportunity to study human behavioral patterns (Ratkiewicz et al., 2011). In-depth investigation and understanding of OSNs are important to enhance knowledge on the relationships among people and help in answering several questions about society and sociality. This opportunity introduces OSNs as link in research between computer science and criminology, sociology, economy, and biological science, thus opening new and modern fields of research. OSN services reach billions of users and thus become fertile grounds for various research efforts (Ratkiewicz et al., 2011). OSNs offer a unique opportunity to study patterns of social interaction among populations far larger than those investigated before.

The anatomy of an OSN is a combination of two levels, that is, content and network (Centola, 2010; Kwak, Lee, Park, & Moon, 2010; Ngai, Moon, Lam, Chin, & Tao, 2015), as shown in Figure 1.1. The content level describes what users write, post, or express. The network level describes how users (nodes) are connected to and influence one another. Content analysis reveals what users post, whereas network analysis reveals how information spreads in a network. Each level may be analyzed separately with different techniques to understand or solve an issue related to OSNs. However, the association between analyses of these two levels can provide a full solution of a problem from several perspectives. Content and network analyses help to provide a cybercrime detection method and vital spreaders identification method, respectively.

These methods can be used to achieve effective solutions in large complex networks such as OSNs.

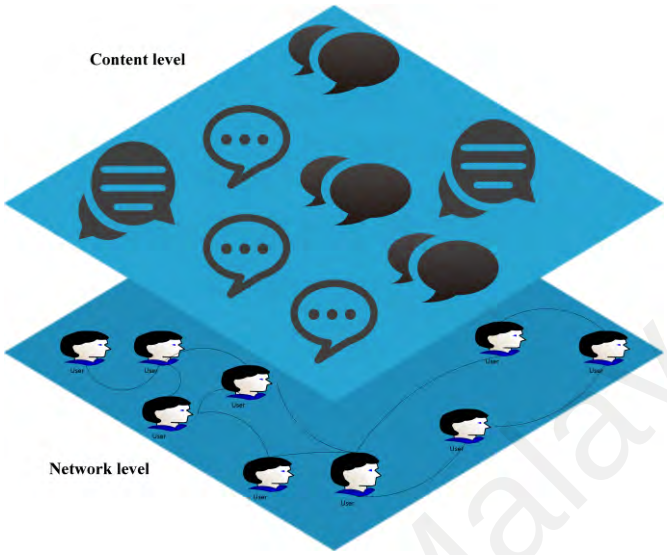

Figure 1.1: Anatomy of an OSN

1.1.1 Cybercrimes in OSNs

OSNs provide many benefits to users; individuals can effortlessly communicate and share experiences through these networks. Despite such advantages, OSNs act as a tool to commit various cybercrimes. OSN cybercrimes have increased recently as the number of OSN users grow. As OSNs become popular, they also become a platform for cybercrimes. The substantial intensification of cybercrimes and aggressive behavior in OSNs demonstrate a new challenge (BBC, 2012; Peterson & Densley, 2016). OSNs have contributed to cybercrime by becoming a platform for users to commit cybercrime as well as become a large spreading mechanism for cybercriminals (Weir, Toolan, & Smeed, 2011). Cybercriminals utilize OSNs as a new means for committing different types of cybercrimes. The main involvement of OSNs to cybercrimes can be concluded in two points (Fire, Goldschmidt, & Elovici, 2014a; Weir et al., 2011):

I. OSN communication is a revolutionary trend exploiting Web 2.0. Web 2.0 provides new features that allow users to create profiles and pages, which, in

turn, makes users active. Unlike Web 1.0, which limits the users to be only passive readers of content, Web 2.0 expands the capabilities, which allow users to be active as they post and write whatever comes in mind. OSNs exhibit four particular potential capacities, namely, collaboration, participation, empowerment, and time (Magro, 2012). These characteristics of OSNs enable criminals to use them as a platform to commit cybercrimes without confronting victims (Fire et al., 2014a; Weir et al., 2011). Examples of cybercrimes are committing cyberbullying (Chavan & Shylaja, 2015; Y. Chen, Zhou, Zhu, & Xu, 2012; Dadvar, Trieschnigg, Ordelman, & de Jong, 2013), financial fraud (Dong, Liao, Xu, & Feng, 2016), using malicious applications (Rahman, Huang, Madhyastha, & Faloutsos, 2012), and implementing social engineering and phishing (A. Aggarwal, Rajadesingan, & Kumaraguru, 2012).

II. OSNs can be described as a structure that enables the exchange and dissemination of information. OSNs are designed to allow a community of users to easily share information, such as messages, links, photos, and videos (Abu-Nimeh, Chen, & Alzubi, 2011). However, because OSNs connect billions of users, they, unfortunately, become delivery mechanisms for different cybercrimes at an extraordinary scale. OSNs help cybercriminals reach many users (Doerr, Fouz, & Friedrich, 2012).

OSNs can be used as a flexible means for criminals to commit cybercrimes. In addition, a highly connected network, such as OSNs, can be used to spread cybercrime to affect a large portion of users. Considering these two points as shown in Figure 1.2, the effective solutions for minimizing cybercrimes in OSNs must include the investigation and development of both detection and blocking methods.

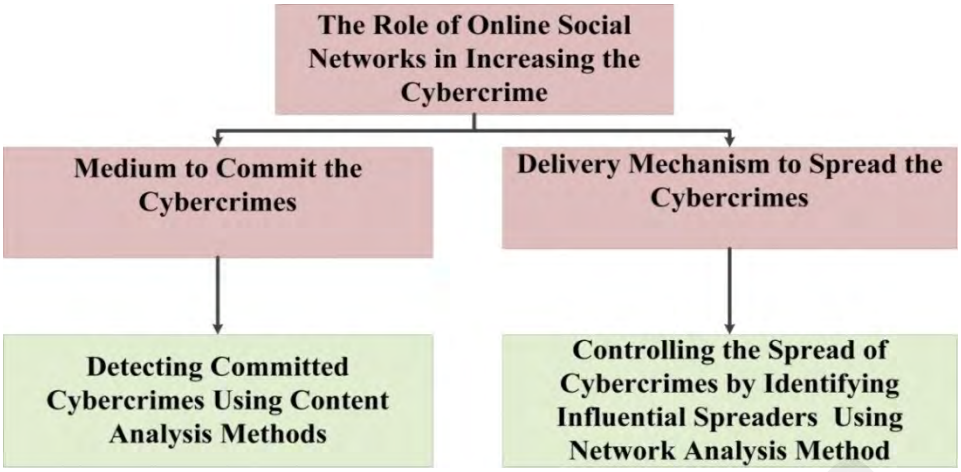

Figure 1.2 : Role of OSNs in increasing cybercrimes

Although researchers have developed effective methods to detect cybercrimes such as malware and fraud in OSNs, cybercrimes such as cyberbullying are arising, it has been represented as a rising as a serious ‘epidemic ‘in recent years (Wolke, Lee, & Guy, 2017). Detection of the cyberbullying is on rise because of the characteristics of OSNs that allow cyberbullies to commit their cybercrimes and spread them to a large scale of users with a high degree of simplicity and flexibility. However, the OSNs are complex networks containing billions of nodes (users); applying cybercrime detection methods to monitor all users in such network is practically impossible (Budak, Agrawal, & El Abbadi, 2011; Lü et al., 2016; Z.-K. Zhang et al., 2016). Such a complex network is required to identify the most imprtant nodes (users) which if they are monitored , the most of the networks are immunized (Basaras, Katsaros, & Tassiulas, 2013). Consequently, this thesis intends to propose effective methods for cyberbullying detection to detect and identifying influential spreaders identification in OSNs. The effectiveness of the method is defined as the improvement in the performance of a method compared to baseline methods. The performance of the method is calculated and compared using evaluation metrics.

The detailed motivation in proposing these methods are explained in the following section.

1.2 Motivation

Prior to the innovation of communication technology, social interaction tended to evolve within small cultural boundaries, such as locations and families (Quan, Wu, &

Shi, 2011). The recent development on communication technologies exceptionally transcends the temporal and spatial limitations of traditional communications. Over the last few years, online communication has shifted toward user-driven technologies, such as OSNs, blogs, online virtual communities, and online sharing platforms. New forms of aggression and violence occur exclusively online (Peterson & Densley, 2016). A huge rise in OSN cybercrimes, with high increments in cybercrimes and aggressive behavior, presents a new challenge (BBC, 2012; Peterson & Densley, 2016). The advent of Web 2.0 technologies, including OSNs, often accessed through mobile devices, has thoroughly transformed the functionality available to users (Watters & Phair, 2012). OSN characteristics, such as easy access, flexibility, freedom, and well-connected social networks, provide users with liberty and flexibility to post and write on their platform. Therefore, criminals can easily commit a cybercrime (Fire et al., 2014a; Shekokar & Kansara, 2016).

Moreover, the network structure of OSNs, which differs from that of traditional websites, provide a large-scale platform to disseminate misinformation (Chatfield, Reddick, & Brajawidagda, 2015; Z. He et al., 2016; Klausen, 2015; Tsugawa & Ohsaki, 2015; Wen et al., 2014a). OSNs assist in committing a cybercrime in two ways. First, these networks allow users to post an unwanted content (crime-related content), such as cyberbullying. Second, OSNs enhance the spread of such unwanted content within an enormous number of users with the help of their large-scale structural connections. Consequently, the motivation behind this thesis can be divided into two parts, that is, motivations to propose an effective method for detecting cyberbullying and effective method for identifying influential spreaders and are explained in the following subsections.

1.2.1 Motivations in Proposing Cyberbullying Detection Method

The motivations for carrying out research on detecting cyberbullying in OSNs are listed as follows:

***Motivation due to human security*:** Cyberbullying has been identified as an severe problem among youth in the last decade (Slonje, Smith, & Frisén, 2013). Cyberbullying has emerged as a major problem (O'Keeffe & Clarke-Pearson, 2011) and has been recognized as a serious national health problem (Xu, Jun, Zhu, & Bellmore, 2012) due to the recent growth of online communication and OSNs. Research shows that cyberbullying causes negative effects on psychological and physical health, and academic performance (Kowalski & Limber, 2013). Studies also show cyberbullying victims incur a significantly high risk of suicidal ideation (Hinduja & Patchin, 2010; Sampasa-Kanyinga, Roumeliotis, & Xu, 2014). These studies (Hinduja & Patchin, 2010; Sampasa-Kanyinga et al., 2014) reported an association between cyberbullying victimization and suicidal ideation risk. Consequently, developing a cyberbullying detection model that detects a cybercrime, which is related to the security of human beings is much important than developing a detection method for a cybercrime related to the security of machines.

Motivation due to cyberbullying nature: Cyberbullying can be committed anywhere and anytime. Escaping from cyberbullying is difficult; cyberbullying can reach victims anywhere and anytime. It can be committed using comments, post, status, and so on to a large potential audience, and the victims cannot stop the spread of such activities (Slonje et al., 2013).

Motivation due to specific OSN characteristics enhancing the severity of cyberbullying: Although OSNs have become integral parts of user lives, a study found that OSNs are the most common platforms for cyberbullying victimization

(Whittaker & Kowalski, 2015). A well-known characteristic of OSNs, such as Twitter, is that they allow the users to publicly express and spread their posts to a large audience, while remaining to be anonymous (Fire et al., 2014a). The effects of public cyberbullying are worse than those of private ones, and anonymous scenarios of cyberbullying are worse than non-anonymous cases (Sticca & Perren, 2013; Wen et al., 2014a). Consequently, the severity of cyberbullying increases in OSNs, which support both public and anonymous scenarios of cyberbullying. These characteristics make OSNs such as Twitter a dangerous platform for committing cyberbullying (Xu et al., 2012).

Motivation due to recommendation from experts and adolescents: Recent research concluded that most experts favored automatic monitoring of cyberbullying (Van Royen, Poels, Daelemans, & Vandebosch, 2015). A study that focused on 14 groups of adolescents confirmed the urgent need for automatic monitoring and detection method for cyberbullying (Van Royen, Poels, & Vandebosch, 2016) because the traditional strategies of coping with cyberbullying in the era of big data and networks do not work well. Also, analyzing huge complex data requires machine learning-based automatic monitoring.

1.2.2 Motivations in Proposing Influential Spreader Identification Method

The motivation for researching a method to identify influential spreaders in OSNs is detailed below.

OSNs promote a propagation platform for cybercrimes: From the spread of telegraphy messages to the extensive acceptance of OSNs, the development of digital communication tools has essentially changed the means of how information is created, shared, and consumed (S. Wu, Hofman, Mason, & Watts, 2011). The advancement on individual publishing technologies, such as the Web, blogs, OSNs, and photo sharing

sites, has made the process of spreading content easy and decentralized. Information diffusion refers to information dissemination among people in the society. Common examples of information diffusion include the spread of rumors, beliefs, and behaviors (L. Weng, Flammini, Vespignani, & Menczer, 2012). OSNs grow naturally; their network structure is not intended for any particular use but still permits the quick spread of their content (Doerr et al., 2012). Inconsolably, when it comes to dissemination, rumor (Oh, Kwon, & Rao, 2010), negative message (Tsugawa & Ohsaki, 2015), negative links (Leskovec, Huttenlocher, & Kleinberg, 2010), cyberbullying behavior (Peterson & Densley, 2016), and bad news (S. Wu, Tan, Kleinberg, & Macy, 2011), the mass–connected OSNs offer the potential of spreading with unbelievable speed (Z.-K. Zhang et al., 2016). Therefore, OSNs do not just provide users with the platform to commit a cybercrime; they also act as a platform to disseminate this cybercrime as well.

High diffusion of cybercrimes in OSNs magnifies the negative effect on human and society: The broad dissemination mechanism provided by OSNs has unfortunately amplified the effect of cybercrimes on society to involve a large number of infected victims. For example, as mentioned in the pervious section, cyberbullying has become a common phenomenon especially when the cyberbullies utilize the network structure to spread hurtful rumors about the victim and share embarrassing content with a large number of users (Fire et al., 2014a). The severity of cyberbullying has been amplified due to the huge propagation. In some cases, tragic consequences happen such as the story of Amanda Michelle Todd (Dean, OCTOBER 18, 2012) and Rebecca Ann Sedwick (Pearce, Sep. 2013) who committed suicide after being cyberbullied in an OSN. Similarly, a tweet on an explosion in White House led to a loss of billions in the US markets (Foster, April 2013). Complex network algorithms can be used to control the spread of unwanted content within large complex networks such as OSNs. Complex networks produce network immunization strategies. These techniques

are possible solutions to the control challenges. Currently, one of the most popular methods for network immunization is identifying influential spreaders in OSNs (Budak et al., 2011; Gao, Liu, & Zhong, 2011; Kitsak et al., 2010; Min, Liljeros, & Makse, 2015; Pei, Muchnik, Andrade Jr, Zheng, & Makse, 2014). These influential spreaders in a network are immunized (protected) so that they cannot be infected by the negative behavior, consequently reducing the spread of unwanted content such as cyberbullying, rumor, virus, and spams. Moreover, these influential spreaders can be targeted to spread the information on minimizing the cybercrime by spreading the awareness on cyberbullying effect, its prevention, and revealing the truth in the case of rumor (Lü et al., 2016; Wen et al., 2014a). Identifying the influential spreaders is significant in hindering the spread of unwanted content in a complex network such as OSNs (Lü et al., 2016).

1.3 Problem Statements

Generally, a profound comprehension of data on human behavior and interaction involves interdisciplinary angles and aspects, combining theorems and techniques from multidisciplinary and interdisciplinary fields. OSNs offer significant data on human behavior and interaction, which can be used by researchers to develop effective methods for detecting and blocking the cybercrime. Using traditional methods is challenged by scale and accuracy. These methods are commonly drawn from organized data on human behavior as well as from small-scale human networks (traditional social network). Consequently, applying these methods on large OSNs in both scale and extent has raised several problems.

This study addresses two main problems. The first problem is related to improving the cyberbullying detection performance in OSNs. The second problem lies in

developing the identification of influential spreader method in the OSNs context. Specifically, this research addresses the two following problems-

I. **Problem related to the cyberbullying detection method in OSNs:** Since cyberbullying is widely recognied as a serious national health problem (Xu et al., 2012), and the severity of cyberbullying behavior has been extensively increased with the introduction of OSNs such as Twitter (Kowalski, Limber, Limber, & Agatston, 2012). Existing studies (e.g., (Y. Chen et al., 2012; Dadvar & De Jong, 2012; Dadvar, Trieschnigg, Ordelman, et al., 2013; Dinakar, Jones, Havasi, Lieberman, & Picard, 2012; Galán-García, de la Puerta, Gómez, Santos, & Bringas, 2014; Hosseinmardi, Han, Lv, Mishra, & Ghasemianlangroodi, 2014; Huang, Singh, & Atrey, 2014; Kansara & Shekokar, 2015; Nalini & Sheela, 2015; Reynolds, Kontostathis, & Edwards, 2011; Sood, Antin, & Churchill, 2012)) focused on cyberbullying detection using limited features to construct cyberbullying detection methods. These features are considered inadequate and are not discriminative in detecting the of OSN data. For instance, survey studies observed that hostility significantly forecasts cyberbullying (Arıcak, 2009), and cyberbullying are strongly correlated to neuroticism (Connolly & O'Moore, 2003; Corcoran, Connolly, & O'Moore, 2012). Similarly a survey research observed a strong correlation between the cyberbullying behavior and sociability of users in online environments (Navarro & Jasinski, 2012). However, these observations are yet to be used as features to construct the cyberbullying detection method. Accordingly, these observations are useful in a practical form (features) that could be used to develop an effective detection method for cyberbullying in OSNs. Inadequacies in this area (features) must be addressed to construct an effective cyberbullying method. A set of features is required to fill this gap.Furthermore constructing detection methods depends on

many factors. The most important factor is the features used that can improve the discrimination power of the machine learning classifiers (Domingos, 2012). However the most of existing studies have used limited features to train the machine learning classifiers; consequently, their performance is yet to be developed.

II. **Problem related the identification of influential spreader method in the OSNs:** The situation will be more critical in OSNs when the post that contains a negative behavior goes viral. In the well-connected OSN networks with billions of users, each user can be vulnerable to spread the unwanted information. However, controlling the entire complex network is impossible (Budak et al., 2011; Lü et al., 2016; Z.-K. Zhang et al., 2016). Such a complex network is required to identify if the most vital users (nodes) are protected given that most of the networks are immunized (Basaras et al., 2013). These nodes are known as influential spreaders, and must be identified to provide better immunization of the network (Basaras et al., 2013; Gao et al., 2011). Identifying influential spreaders in OSNs is significant in determining influential nodes in the network where the detection method can be applied to block the diffusion of annoying information (Kwon, Cha, Jung, Chen, & Wang, 2013; L. Zhao et al., 2011) such as spreading the cyberbullying viruses, rumors, and online negative behaviors or accelerating propagation such as clarifying a rumor or propagating the cyberbullying preventive message. Previous research tried to propose influential spreaders identification method using various complex network analysis algorithms (W. Chen, Cheng, He, & Jiang, 2012; Ding et al., 2013; Jabeur, Tamine, & Boughanem, 2012; Java, Kolari, Finin, & Oates, 2006; Nguyen & Szymanski, 2013; Pei et al., 2014; Räbiger & Spiliopoulou, 2015; Tunkelang, 2009; J. Weng, Lim, Jiang, & He, 2010; Yamaguchi, Takahashi, Amagasa, &

Kitagawa, 2010). However, comparative studies in (Kitsak et al., 2010; Lü et al., 2016; Pei et al., 2014), concluded that k-core is considered much suitable in identifying influential spreaders in complex networks. The drawback of the method (i.e. K-core) is that it deals with unweighted network (Ying Liu, Tang, Zhou, & Do, 2015; Pei et al., 2014). Nevertheless, most actual networks are weighted, and their weights describe the significant properties of underlying systems. Consequently, the major limitation of the current state of the art method in OSN context is that it considers all user links (connections that user has) in spreading the information equally regardless of the quantity of the interactions between the users. Interactions between the users are important factor to quantify the spread of information and calculating the influential spreaders in OSNs.

1.4 Objectives

This research is undertaken to propose an effective machine learning method for detecting cyberbullying in OSNs as well as to propose an effective method for identifying influential spreaders. The following are the objectives of this research.

1. To propose a set of features, which can provide discriminative power for detecting cyberbullying in OSNs.
2. To construct an effective method for detecting cyberbullying in OSNs based on the proposed features.
3. To develop an effective method for influential spreaders identification in OSNs.
4. To evaluate the effectiveness of the methods by comparing them with the baseline methods from the literature using the real data sets.

1.5 Significance of the Research

To minimize the role of OSNs in the rise of cybercrimes, this research considers both content and network analyses. These two methods can be used in a synergistic manner to accomplish the full functionality and provide significant solutions. In this research, an effective method is proposed to detect one of the most serious cybercrimes in OSNs (cyberbullying). An effective method to identify influential spreaders is also developed.

The effective method for cyberbullying can be used by the members of the organization, such as parents, guardians, educational institutions, and organizations (e.g., workplace) as well as non-government organizations, including crime-prevention foundations, social chamber organizations, psychiatric associations, policy makers, and enforcement bodies.

The effective method for identifying influential spreaders is significant in blocking the diffusion of annoying information (spreading of viruses, rumors, online negative behavior, and cyberbullying) in large networks, or in accelerating the dissemination of information that is useful for numerous applications, such as spreading the awareness and clarifying the truth. This method can be used with the cyberbullying detection method to handle large and complex networks such as OSNs to either spread the awareness to avoid involvement in cyberbullying activities or rumor activities and the spread of committed cyberbullying of important users.

These two methods are, in synergistic principle, effective detection methods that are important to be used at influential spreaders to deliver the best immunization technique in large complex networks and effectively identifying influential spreaders (important nodes) is important to identifiy the importnant nodes where effective detetion method can be applied.

The synergistic relationship between content analsysis and network analysis to detect cybercrimes. Content analysis mostly uses machine learning algorithms, which discover the pattern from OSN human–generated data to provide cybercrime detection methods, such as cyberbullying detection (Weir et al., 2011), phishing detection (A. Aggarwal et al., 2012), spam distribution (Yardi et al., 2009), and malware (C. Yang et al., 2012). Detecting cybercrimes by analyzing the content alone is not an effective solution in complex systems such as OSNs, where billons of nodes (users) can be vulnerable (detection method cannot be applied for every user in such large networks) to infection and involved in committing or spreading cybercrimes (Basaras et al., 2013; Kwon et al., 2013; Nahar, Li, & Pang, 2013; L. Zhao et al., 2011). Network analysis or graph theory use OSNs to construct and analyze complex human relationship (Lazer et al., 2009; L. Weng). Reseachers must understand how information is spread and find effective method to either minimize the spread of cybercrime by blocking the spread of rumor and cyberbullying (Kwon et al., 2013; Nahar et al., 2013; L. Zhao et al., 2011), or maximizing user awareness by spreading prevention strategies to a large number of users, such as cyberbullying prevention strategies by making kind words go viral (Ang, 2016; Patchin & Hinduja, 2013). Therefore this research addresses these two aspects.

1.6 Thesis Layout

The structure of this thesis is presented in Table 1.1, and the schematic of the thesis layout is presented in Figure 1.3.

Chapter 2: This chapter reviews the literature related to this research. This chapter focuses on the content and network methods in analyzing the OSNs. First, this chapter focused on content analysis to detect cyberbullying in OSNs. It comprehensively reviews the previous research regarding data collection, features used, machine learning algorithms commonly used in this field, and the evaluation of the machine learning

methods. The review of the issues and problems in previous research are identified in a separate section. Second, a review on network analysis of OSNs to identify the influential spreaders is preformed. The significance of identifying influential spreaders to restrain the misinformation in OSNs and applications to restrain the spread of negative information are discussed. Comprehensive review, comparison, and investigation between the current methods for identifying influential spreaders are performed. Finally, the issue and problems with the current methods in the OSN context are identified and presented in a separate section.

Chapter 3: This chapter presents the methodology used in this thesis to develop the proposed methods for detecting cyberbullying and identifying influential spreader in OSNs. First, it presents the methodology based on content analysis to detect cyberbullying in an OSN. Second, it presents a methodology-based network analysis for identifying influential spreaders.

Chapter 4: This chapter presents an effective method for cyberbullying detection in an OSN. A set of features derived from Twitter is proposed, and an effective method using these features is constructed. Using these features, a cyberbullying detection method is constructed. It comprehensively presents the different experiment set-ups and their results. The effectiveness of the proposed features in constructing the cyberbullying detection method is then evaluated.

Chapter 5: This chapter proposes and develops an effective method to identify influential spreaders in an OSN. The interaction weighted k-core decomposition method is developed and evaluated in this chapter. The developed interaction weighted k-core decomposition method to effectively identify influential spreaders is described in detail in this chapter. The evaluation models and evaluation metrics are described and the effectiveness of developed method is evaluated.

Chapter 6: This chapter concludes the thesis by reappraising the research objective. The main contributions are summarized. It discusses the limitations of the research and proposes future directions.

Table 1.1: Thesis layout

Chapter	Content	Reasons
Chapter 1 Introduction	❖ Background	➢ To define OSNs and describe how OSNs serve the field of cybercrimes
	❖ Motivation	➢ To signify the importance of the study
	❖ Statement of Problems	➢ To specify the problems to be addressed in this research
	❖ Objectives	➢ To state the aims and objectives of this research
	❖ Significance	➢ To highlight the significance of the work reported in this thesis
	❖ Thesis Layout	➢ To describe content structure presented in thesis
Chapter 2 Literature Review	❖ Introduction of OSNs	➢ To introduce OSNs characteristics and cybercrime in OSNs
	❖ Content-based analysis: Review of cyberbullying detection methods in OSNs	➢ To review current content methods for constructing cyberbullying detection in OSNs
	❖ Issues on current state-of-art cyberbullying detection methods in OSNs	➢ To identify the issues of current works in literature of constructing cyberbullying detection methods
	❖ Network-based analysis: Review of methods for identifying influential spreaders in OSNs	➢ To review current network based methods in identifying influential spreaders in OSNs
	❖ Issues on current state-of-art methods for identifying influential spreaders in OSNs	➢ To identify issues of current works in literature of influential spreaders in OSNs identification methods

Chapter 3 Methodology	❖ Introduction to methodology	➢ To schematically present the generalized flowchart of the methodology
	❖ Content-based methodology: Methodology for cyberbullying Detection in Twitter	➢ To describe the methodology in constructing the method for cyberbullying detection in OSNs, more specifically to describe the following stages: ▪ OSNs content data preparation ▪ Feature Extraction ▪ Construction of cyberbullying detection method ▪ Evaluation
	❖ Network-based methodology: Methodology for identifying influential spreaders in Twitter	➢ To describe the process in developing the method to identify influential spreaders in OSNs, more specifically to describe the following stages: ▪ OSNs network Data preparation ▪ Network construction ▪ Identification of influential spreaders ▪ Evaluation
Chapter 4 Cyberbullying Detection in an OSNs	❖ Proposed features	➢ To discuss the proposed features in detail
	❖ Experiment set-up	➢ To describe a set of experiments to construct effective cyberbullying detection method with the proposed features
	❖ Results	➢ To run the experiments and compare the performance of the machine classifiers based on proposed features, and select the best setting for the proposed features
	❖ Evaluation of the effectiveness detection method based on proposed features	➢ To compare the performance of the cyberbullying detection method based on proposed features under best setting with the cyberbullying detection methods based on baseline features under best setting using evaluation metrics
Chapter 5 Identification of Influential	❖ Proposed method	➢ To develop the proposed method, to present the pseudo code for developed method, and to provide an illustrative example of the proposed method
	❖ Results	➢ To run the proposed method and the baseline method on actual data sets from Twitter
	❖ Evaluation of the effectiveness of proposed method	➢ To compare and evaluate the effectiveness of the proposed method in identify influential spreaders compared to baselines methods using evaluation metrics

Chapter 6 Conclusions	❖ Reappraisal of the Research Objectives ❖ Contributions ❖ Limitations and future research directions	➢ To re-examine the research objectives ➢ To highlight the contribution of the research to the literature ➢ To discuss the limitations of the research and propose future research directions

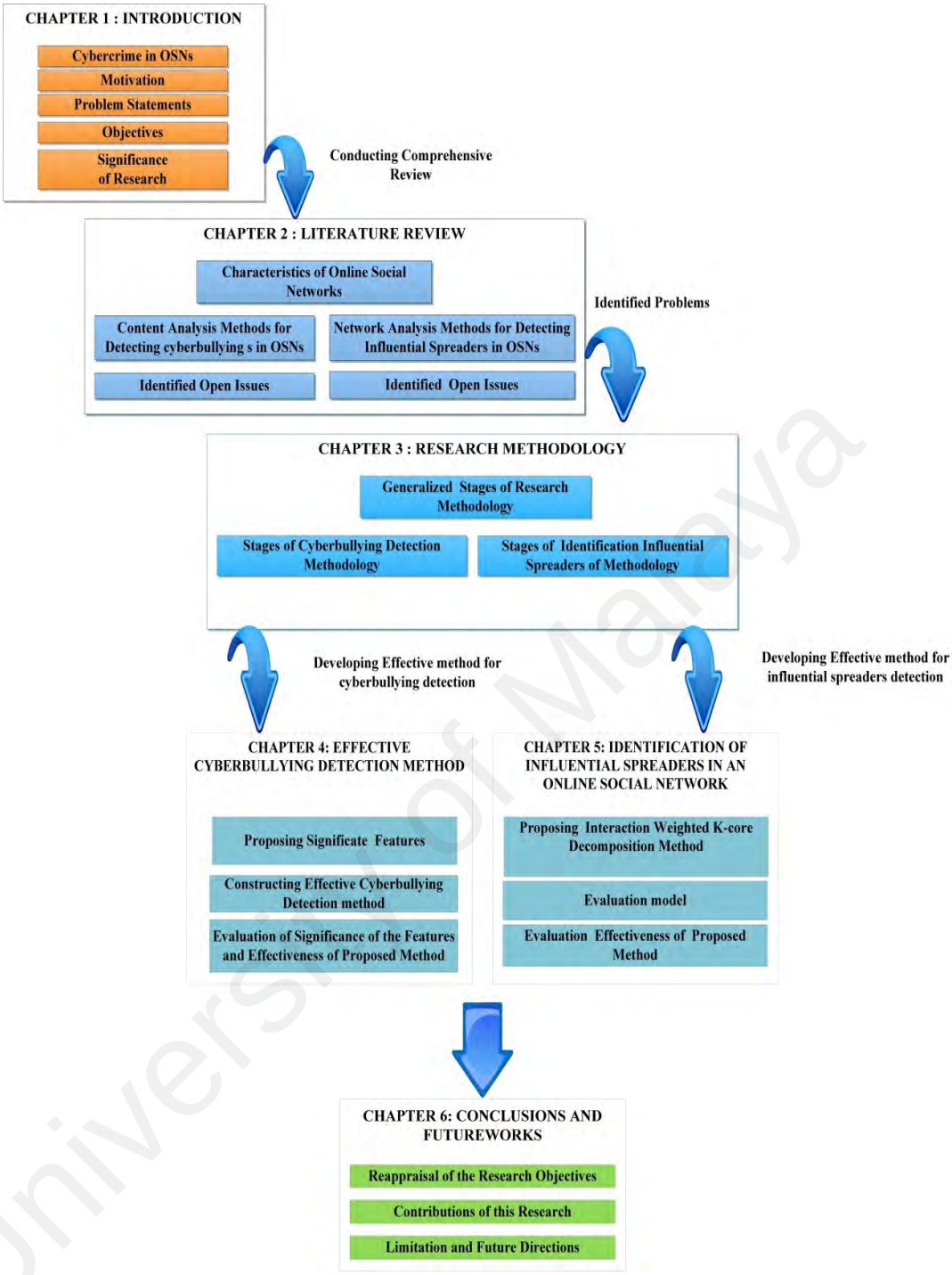

Figure 1.3: Schematic of the thesis layout

CHAPTER 2: LITERATURE REVIEW

In this chapter, background of OSN is presented in section 2.1, which includes cybercrime in OSNs and content and network analysis for cybercrime detection. Section, 2.2 comprehensively reviews of cyberbullying detection in OSNs. Section 2.3 reviews influential spreaders identification in OSNs.

2.1 Background

Online social network (OSN) services reach billions of users and thus become fertile ground for various research efforts (Ratkiewicz et al., 2011). Data extracted from OSNs can provide researchers huge and rich information about human networks and their societies that are not previously possible in both scale and extent (H. Lauw et al., 2010). OSN refers to any website that allows users to create their own profiles for making friends and communicating with other users regardless of geographic location (Ellison, 2007). Interaction through social networks (e.g., Facebook, Google+, LinkedIn, Twitter, etc.) generates huge and useful amount of data. Studying social phenomena with high precision is easy because of social network data available online. This opportunity creates new interest to introduce OSNs as link in research between computer science and criminology, sociology, economy, and biological science, thus opening a new and modern field of research.

As OSNs become essential in the lives of millions of people, these networks greatly influence people's education, jobs, day, and even relationships (Asur & Huberman, 2010). Although websites such as e-mail and chat groups provide people the means to communicate, users cannot visualize their social networks without perfect visualization environment. Currently, OSN sites create visible connections between friends and followers. These sites become a diary for many people, a place to post their daily activities, including what they do, where they go, and whom they meet, and even

updates on their personal relationship. OSNs introduce new environment in which people create their own ideas and knowledge and then post and share them with the network and communities of their online friends. Considering that people are exposed to various ideas, thoughts, cultures, and opinions, OSNs are regarded as an important factor in building society characteristic. OSNs are also utilized to forecast future outcomes (Asur & Huberman, 2010). Despite the belief that OSNs are a temporary fashion and will sooner or later be substituted by another Internet trend, current user statistics supports that OSNs are here to stay (Fire et al., 2014a). This evidence establishes the claim that OSNs are rooted in the daily lives of young children and teenagers, which can potentially result in abuse (Fire et al., 2014a). A European study claimed that approximately a quarter of the parents stated that they use monitoring tools (Livingstone, Haddon, Görzig, & Ólafsson, 2011).

Along with renovating the means through which people are influenced, OSNs serve as a place for severe form of misbehavior among users. Online complex networks, such as OSNs, have changed substantially over the last decade, and this change has been stimulated by the popularity of online communication through OSNs. Online communication is now also a tool of entertainment rather than only a means to communicate and interact with known and unknown users. Although OSNs deliver many benefits to users, unfortunately, cyber criminals have utilized OSNs as a new platform to commit different types of misbehaviors and/or cybercrimes. Common misbehaviors and/or cybercrimes on OSN sites include phishing (A. Aggarwal et al., 2012), spam distribution (Yardi, Romero, & Schoenebeck, 2009), malware spreading (C. Yang, Harkreader, Zhang, Shin, & Gu, 2012), and cyberbullying (Weir et al., 2011).

On one hand, the explosive growth of OSNs enhances how cybercrimes are committed and disseminated by providing platform to commit and networks to

propagate them. On the other hand, OSNs offer significant data for exploring human behavior and interaction in a large scale, which can be used by researchers to develop effective methods to detect and restrain misbehaviors and/or cybercrimes. As OSNs provide criminals with tools to perform their cybercrime and network to propagate the misconduct, methods for both angles (content and network) should be optimized to detect and restrain cybercrimes in such complex systems.

2.1.1 Cybercrime Detection in OSNs

To eliminate the role of OSNs in increasing cybercrimes, both content and network analyses are used. These two methods can be used synergistically (Zanin et al., 2016) to achieve full functionality and provide significant, well-understood, and complete solutions.

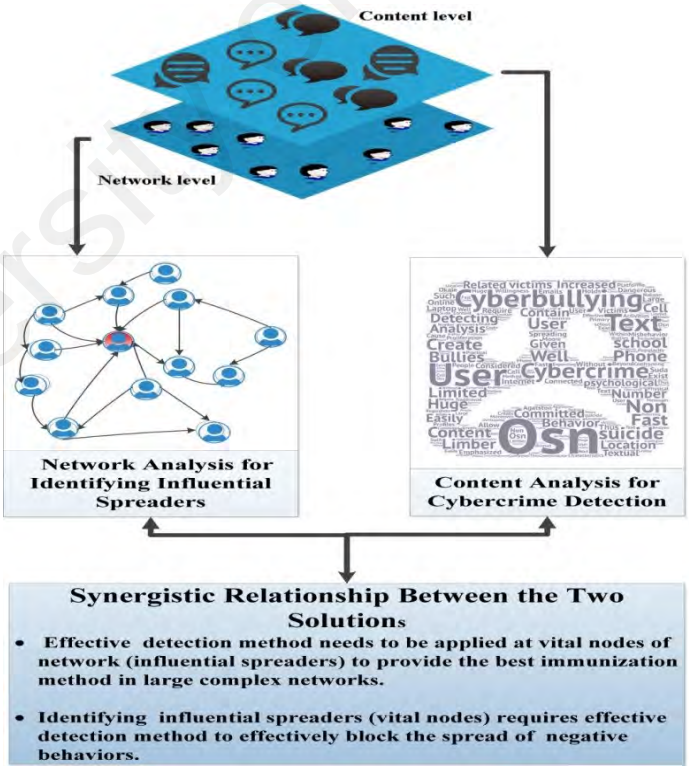

Figure 2.1: Schematic of the Synergistic Relation between Content and Network Analyses

Figure 2.1 shows the synergistic relationship between content analsysis and network analysis to detect cybercrimes. Content analysis mostly uses machine learning algorithms, which discover the pattern from OSN human–generated data to provide cybercrime detection methods, such as cyberbullying detection (Weir et al., 2011), phishing detection (A. Aggarwal et al., 2012), spam distribution (Yardi et al., 2009), and malware (C. Yang et al., 2012). Detecting cybercrimes by analyzing the content alone is not an effective solution in complex systems such as OSNs, where billons of nodes (users) can be vulnerable (detection method cannot be applied for every user in such large networks) to infection and involved in committing or spreading cybercrimes (Basaras et al., 2013; Kwon et al., 2013; Nahar, Li, & Pang, 2013; L. Zhao et al., 2011). Network analysis or graph theory use OSNs to construct and analyze complex human relationship (Lazer et al., 2009; L. Weng). Reseachers must understand how information is spread and find effective method to either minimize the spread of cybercrime by blocking the spread of rumor and cyberbullying (Kwon et al., 2013; Nahar et al., 2013; L. Zhao et al., 2011), or maximizing user awareness by spreading prevention strategies to a large number of users, such as cyberbullying prevention strategies by making kind words go viral (Ang, 2016; Patchin & Hinduja, 2013). The follwing subsections briefly describe the use of content analysis and network analysis for cybercrime detection.

2.1.1.1 Content Analysis of OSNs for Cybercrime Detection

The last few years witnessed a considerable amount of literature on the contribution of text classification methods to OSN content analysis. For instance machine learning research has become an important task in many application areas and successfully produced many methods, tools, and algorithms for handling large amounts of data to solve real-world problems (Patchin & Hinduja, 2013). Machine learning algorithms have been used extensively to analyze content OSNs for spam (S. Liu, Zhang, & Xiang, 2016; Miller, Dickinson, Deitrick, Hu, & Wang, 2014), phishing (Jeong, Koh, &

Dobbie, 2016), and cyberbullying detection (Chavan & Shylaja, 2015; Frommholz et al., 2016). The cybercrime includes phishing (A. Aggarwal et al., 2012), malware spread (C. Yang et al., 2012), and cyberbullying (Weir et al., 2011). In particular, textual cyberbullying has become a dominant cybercrime in OSNs due to the characteristics of OSNs, which allow users full freedom to post on their platform (Dadvar & De Jong, 2012; Dadvar, Trieschnigg, Ordelman, et al., 2013; Dinakar et al., 2012; Galán-García et al., 2014; Huang et al., 2014; Sood et al., 2012). Cyberbullying has emerged as a major problem along with the recent development of online communication and social media (O'Keeffe & Clarke-Pearson, 2011). Cyberbullying has also been extensively recognized as a one of serious health concerns in digital era (Xu et al., 2012) as victims show a significantly high risk of suicidal ideation (Sampasa-Kanyinga et al., 2014). With all these considerations, this thesis focuses on content-based analysis for detecting textual cyberbullying in OSNs.

2.1.1.2 Network Analysis of OSNs for Minimizing the Spread of Cybercrimes

In-depth analysis and understanding of OSNs are important to understand the relationship among people and help in answering many questions about society and sociality. Network analysis of OSNs is a key factor to understand information diffusion within a network. Previous research (Budak et al., 2011; Kitsak et al., 2010; Min et al., 2015; Pei et al., 2014) reported that identifying the most important nodes in networks hold significant applications. These influential nodes, if activated, can cause the spread of information to the entire network or, if immunized, can block the diffusion of large-scale information (Kovács & Barabási, 2015; Morone & Makse, 2015). These nodes in OSNs are called influential spreaders, and identifying these influential spreaders helps to hinder the diffusion of cybercrimes, such as viruses, online negative behaviors, cyberbullying, and rumors (Fire, Goldschmidt, & Elovici, 2014b; Kwon et al., 2013; Pei et al., 2014; L. Zhao et al., 2011). They can also enhance the spread of user awareness

on reducing the negative effects of cybercrimes, such as cyberbullying, by introducing prevention and intervention strategies (Ang, 2016; Patchin & Hinduja, 2013) to a large portion of users. Consequently, this thesis aims to propose an effective method for identifying influential spreaders in OSNs.

2.2 Cyberbullying in OSNs

Cyberbullying behaviors are defined as aggressive behaviors exhibited through electronic or digital media and intended to inflict harm or discomfort to a victim (Bauman, Toomey, & Walker, 2013; Kowalski, Giumetti, Schroeder, & Lattanner, 2014). According to a critical review and meta-analysis of cyberbullying (Kowalski et al., 2014), most researchers agree that cyberbullying involves the use of electronic communication technologies to bully others. Cyberbullying can take many forms, including posting hostile comments, frightening or harassing a victim, producing hateful or insulting posts, or abusing the victim (Qing Li, 2007; Tokunaga, 2010). Given that cyberbullying can be easily committed, it is considered a dangerous and fast-spreading cybercrime. Bullies only require a laptop or cell phone connected to the Internet and willingness to do the misbehavior without confronting the victims (Kowalski, Limber, et al., 2012). The popularity and proliferation of OSNs have increased online bullying activities. Cyberbullying in OSNs are committed to a large number of users due to the structural characteristics of OSNs(Whittaker & Kowalski, 2015). Cyberbullying in traditional platforms, such as emails or phone text messages, is committed to a limited number of people. OSNs allow users to create profiles to create friendships and communicate with other users regardless of geographic location, thus expanding cyberbullying beyond physical location. Moreover, anonymous users may exist within OSNs, and this is confirmed as a primary cause increasing aggressive user behavior (Nakano, Suda, Okaie, & Moore, 2016). The nature of OSNs allows cyberbullying to occur secretly, spread rapidly, and continue easily (Qing Li, 2007). Consequently

developing an effective detection method for detecting cyberbullying holds tremendous practical significance.

OSNs contain huge text or/and non-text contents as well as those related to cybercrimes. In this thesis, content analysis of OSNs for detecting cybercrimes is emphasized to the analysis of textual OSN content for detecting cyberbullying behavior.

- The first subsection comprehensively reviews the construction of cyberbullying detection methods in OSNs from data collection to evaluation metrics.
- The second subsection highlights the issues and problems in current cyberbullying detection methods.

2.2.1 Review on Construction of Cyberbullying Detection Methods

The most common method for constructing cyberbullying detection methods is the use of a text classification approach, which involves the construction of machine-learning classifiers from labeled instances of texts (Chavan & Shylaja, 2015; M. Dadvar, F. M. de Jong, R. Ordelman, & R. Trieschnigg, 2012b; Forman, 2003; Galán-García et al., 2014; Hosseinmardi et al., 2015). Another method is the use of a lexicon-based method, which involves computing orientation for a document from the semantic orientation of words or phrases in the document (Turney, 2002). Generally, in lexicon-based methods, the lexicon can be constructed manually similar to the approaches used in (R. M. Tong, 2001) or automatically using seed words to expand the list of words (Hatzivassiloglou & McKeown, 1997). However, cyberbullying detection with the use of the lexicon-based approach is rarely used in literature. The key reason is that the texts on OSNs are written in an unstructured way, which makes it difficult for the lexicon-based approach to detect cyberbullying based only on the lexicons (H. Chen, Mckeever, & Delany, 2017; Kontostathis, Reynolds, Garron, & Edwards, 2013; Nadali, Murad,

Sharef, Mustapha, & Shojaee, 2013). However, the lexicons are used to extract the features, which are usually utilized as inputs to machine-learning algorithms. For instance lexicon-based approaches, such as using a profane-based dictionary to detect the number of profane words in a post are used as profane features to machine-learning method (Reynolds et al., 2011). The main key to providing a cyberbullying detection performance is the set of features that are extracted and engineered (Nahar et al., 2013). Features and their combinations have an important role in the construction of an effective cyberbullying detection method (Nahar et al., 2013; Reynolds et al., 2011). Most studies in literature on cyberbullying detection (Chavan & Shylaja, 2015; Galán-García et al., 2014; Hosseinmardi et al., 2015; Mangaonkar, Hayrapetian, & Raje, 2015; Van Hee et al., 2015) have used machine-learning algorithms to construct a cyberbullying detection method. Machine-learning-based methods achieve a decent performance on cyberbullying detection (Sanchez & Kumar, 2011). Consequently, the present paper focuses on reviewing the construction of cyberbullying detection methods based on machine learning.

A machine-learning field focuses on the development and application of computer algorithms that improve with experience (Andrieu, De Freitas, Doucet, & Jordan, 2003; Libbrecht & Noble, 2015). The objective of machine learning is to identify and define the patterns and correlations between data. The power of content analysis of big data has been shown in finding hidden information through thorough learning and mining of raw data (Ni, Tan, & Xiao, 2016). The concept of machine learning can be described as the adoption of computational methods to improve machine performance by detecting and describing meaningful patterns in training data and the acquisition of knowledge from experience (Langley & Simon, 1995). With the application of this concept for OSN content, the potential of machine learning lies in exploiting historical data to detect, predict, and understand large OSN data. For example, in supervised machine

learning for classification application, classification is learned using appropriate exemplars from a training data set. In the testing stage, new data are fed into the model, and instances are classified to the specified class learned during the training stage. Subsequently, classification performance is evaluated.

This section reviews the most common processes in the construction of cyberbullying detection methods in OSNs based on machine learning, starting from data collection, feature engineering, feature selection, and machine-learning algorithms.

2.2.1.1 Data collection

Data are the important ingredient of all machine learning based detection methods. However, data, even "Big Data," are useless on their own until one extracts knowledge or implications from them. Data extracted from OSNs are used to select training and testing data sets. Supervised detection methods aims to provide techniques for computers to increase their prediction performance at defined tasks based on observed instances (labeled data) (Ghahramani, 2015). Developing machine learning methods for a certain task primarily aims to generalize; the success of method should not only be limited to examples in a training data set (Domingos, 2012) but also include unlabeled real data. Data quantity is inconsequential; whether the extracted data well represent activities in OSNs is significant (Cheng & Wicks, 2014; González-Bailón, Wang, Rivero, Borge-Holthoefer, & Moreno, 2014; Yabing Liu, Kliman-Silver, & Mislove, 2014). The main data collection strategies in previous cyberbullying detection studies on OSNs can be categorized into data extracted from OSNs either using keywords, that is, words, phrases, or hashtags (e.g., (Bellmore, Calvin, Xu, & Zhu, 2015; Chavan & Shylaja, 2015; Margono, Yi, & Raikundalia, 2014; Xu et al., 2012; R. Zhao, Zhou, & Mao, 2016))or using user profile (e.g., (Galán-García et al., 2014; García-Recuero, 2016; Hosseinmardi et al., 2015; Reynolds et al., 2011)). The issues of such data

collection strategies and their effect on the performance of machine learning algorithms are highlighted in section 2.2.2.2 Data Collection related issues.

2.2.1.2 Feature engineering

Feature is a measurable property of a task being observed (Anzai, 2012). The main purpose of engineering feature vectors are to provide machine learning algorithms with a set of learning vectors through which these algorithms learn how to discriminate between different types of classes (Libbrecht & Noble, 2015). Feature engineering is a key factor behind the success and failure of most machine learning methods (Domingos, 2012). The success and failure of detection may be based on various factors. The most important factor is the features used and whether the method display many independent features that correlate well with the class (Ghahramani, 2015). Most of the effort in constructing cyberbullying detection method using a supervised machine learning algorithms is devoted to this task (Dadvar et al., 2012b; Hosseinmardi et al., 2015; Van Hee et al., 2015) . In this context, the design of the input space (that is, the features and their combination that are provided as input to the classifier) is vital. Proposing a set of discriminative features, which will be inputs to the machine learning classifier, is the main step toward constructing effective classifier in many applications (Libbrecht & Noble, 2015). Feature sets can be proposed based on human-engineered observations, which rely on how these feature correlate with the occurrences of classes (Libbrecht & Noble, 2015). For instance, recent survey cyberbullying studies (Arıcak, 2009; Calvete, Orue, Estévez, Villardón, & Padilla, 2010; Connolly & O'Moore, 2003; Corcoran et al., 2012; Slonje & Smith, 2008; Vandebosch & Van Cleemput, 2009; Williams & Guerra, 2007) identify the correlation between different variables, such as age, gender, and user personality, and of cyberbullying occurrences. These observations can be engineered into practical form (feature) to learn the classifier to discriminate between cyberbullying and non-cyberbullying and thus can be used to develop effective cyberbullying

detection methods. Proposing features is an important step toward improving the discrimination power of detection methods (Domingos, 2012; Libbrecht & Noble, 2015). Similarly, proposing a set of significant features of cyberbullying engagement in OSNs is important to develop effective detection methods based on machine learning algorithms (Dinakar, Reichart, & Lieberman, 2011; Kontostathis et al., 2013).

State-of-art research has developed features to improve the performance of cyberbullying detection. In dealing with detecting offensive language, a lexical syntactic feature was proposed, which performed better than traditional learning-based approach in terms of precision (Y. Chen et al., 2012). Dadvar et al (2012). focused on gender information from the profile information to develop a gender-based approach for cyberbullying detection using data sets from Myspace as basis; the gender feature was selected to improve the discrimination capacity of a classifier. In other studies (M. Dadvar, F. de Jong, R. Ordelman, & R. Trieschnigg, 2012a; Dadvar, Trieschnigg, Ordelman, et al., 2013), Age and gender were included as features, but these features were limited to information mentioned by users in their online profile. Several works focused on cyberbullying detection based on profane words as a feature (Dinakar et al., 2012; Dinakar et al., 2011; Kontostathis et al., 2013; Reynolds et al., 2011; D. Yin et al., 2009) ; similarly, a constructed lexicon of profane words indicates bullying, and these words are used as features input for machine learning algorithms (Ptaszynski et al., 2010; Raisi & Huang, 2016). Using profane-related features demonstrate significant improvement in model performance. For instance, both the number of "bad" words and the density of "bad" words were proposed as features for input to machine learning (Reynolds et al., 2011). This study concluded that the percentage of "bad" words in a post is indicative of cyberbullying. A research (R. Zhao et al., 2016) intended to expand a list of pre-defined profane words and allocate different weights to obtain bullying

features; these features were concatenated with bag-of-Words and latent semantic features to represent feature input for machine learning algorithm.

Study (Chavan & Shylaja, 2015) proposed features such as pronouns and skip-grams as additional features to traditional methods, such as bag-of-words (n-gram n=1); they claimed that adding these features improves overall classification accuracy. Another study (Hosseinmardi et al., 2015) analyzed textual cyberbullying associated to comments on images in Instagram and developed a set of features from text comprising traditional bag of words features, comment counts for an image, and post counts within less than one hour of posting the image. Features mined from user and media information, including the number of followers and likes, as well as shared media and features from image content, such as image types, were added (Hosseinmardi et al., 2015). The combination of all features improves overall classification performance (Hosseinmardi et al., 2015). Context-based approach is better than list-based approach to developing feature vector (Sood et al., 2012). However, the diversity and complexity of cyberbullying does not always support this conclusion. Studies (Kontostathis et al., 2013; Squicciarini, Rajtmajer, Liu, & Griffin, 2015; Van Hee et al., 2015; D. Yin et al., 2009) discussed how sentiment analysis can improve the discrimination power of the classifier to distinguish between cyberbullying and normal posts. These studies assumed the sentiment features are good signal for cyberbullying occurrences. In other study which aimed to find ways for reducing cyberbullying activities by detecting troll profiles, researchers proposed a method to identify and associate troll profiles in Twitter; they assumed that detecting troll profiles is an important step toward detecting and stopping cyberbullying occurrences in OSNs (Galán-García et al., 2014). This study proposed features based on tweeted text as well as posting time, language, and location to improve the identification of authorship of posts and determine whether a profile is troll. (Squicciarini et al., 2015) merged features from the structure of OSNs, (e.g.,

degree, closeness, betweenness, and eigenvector centralities as well as clustering coefficient) with features from users (e.g., age and gender) and content (e.g., length and sentiment of post). Combining these features improves the final machine learning accuracy (Squicciarini et al., 2015). The following Table 2.1 compares the different features used in cyberbullying detection literature.

Table 2.1: Summary of Feature Types Used in Cyberbullying Detection literature

Studies	Content based features						Profile based Features			
	Bag of words	**Skip-grams**	**Profanity Features**	**General cyberbullying-related features**	**Sentiment features**	**Pronouns**	**Age or gender features from profile data**	**Number of friends or followers**	**Timestamp of posts**	**Location of posts**
(Chavan & Shylaja, 2015)	✓	✓	✗	✗	✗	✓	✗	✗	✗	✗
(Y. Chen et al., 2012)	✓	✗	✓	✓	✗	✓	✗	✗	✗	✗
(Dadvar et al., 2012b)	✗	✗	✓	✗	✗	✓	✓	✗	✗	✗
(Dinakar et al., 2011)	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗
(Van Hee et al., 2015)	✓	✗	✗	✗	✓	✗	✗	✗	✗	✗
(Hosseinmardi et al., 2015)	✓	✗	✓	✗	✗	✗	✗	✓	✗	✗

(Kontostathis et al., 2013)	✓	✗	✓	✗	✓	✗	✗	✗	✗	✗
(Sanchez & Kumar, 2011)	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
(R. Zhao et al., 2016)	✓	✗	✓	✓	✗	✗	✗	✗	✗	✗
(Squicciarini et al., 2015)	✓	✗	✓	✗	✓	✓	✓	✓	✗	✗
(Reynolds et al., 2011)	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗
(D. Yin et al., 2009)	✓	✗	✓	✗	✓	✗	✗	✗	✗	✗
(Xu et al., 2012)	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
(Galán-García et al., 2014)	✓	✗	✗	✗	✗	✗	✗	✗	✓	✓
(Nahar et al., 2013)	✓	✗	✓	✓	✗	✗	✗	✗	✗	✗

This review of the cyberbullying detection methods presented in previous studies emphasizes the importance of proposing significant features to training machine learning algorithms. Most of the studies adopted similar machine learning algorithms; the difference lies in the features used, which affect detection performance. If the constructed features contain a large set of features that individually associate well with class, then the learning process is effective. This condition explains why most of the discussed studies aim to produce many features. The input features should reflect the behavior related to the occurrence of textual cyberbullying. However, the set of features should be analyzed using feature selection algorithms. Feature selection algorithms are used to determine which features are most likely to be relevant or irrelevant to classes. The issues and limitations of the features currently used to train machine learning algorithms are investigated in subsection 2.2.2.1 fFeature Engineering

2.2.1.3 Feature selection

In any machine learning application, the selection of features as input to a machine learning algorithm is an important part of learning. Selecting the most significant set of features from all proposed set of features (the process of selecting the significant features from the extracted features using feature selection algorithms) occurs either before applying a machine learning method or during training (Libbrecht & Noble, 2015). Whether all proposed features or subsets of proposed features are the most discriminative is decided by the improvement in classifier performance (Prieto, Matos, Alvarez, Cacheda, & Oliveira, 2014). Feature selection algorithms eliminate redundant features. Redundant features are those features that are irrelevant and complicate the identification of meaningful patterns (Prieto et al., 2014). However, selecting the most significant features does not always deliver the optimal choice. In some circumstances, the best performance of classifier is realized when all proposed features are used.

Features that appear irrelevant when used separately may be relevant when used together (Domingos, 2012).

Feature selection algorithms are rarely adopted by state-of- -art research to develop cyberbullying detection in OSNs based on machine learning (all extracted features are used to train the classifiers). Most of the discussed studies (e.g., (Y. Chen et al., 2012; Dadvar et al., 2012b; Dinakar et al., 2011; Kontostathis et al., 2013; Nahar et al., 2013; Reynolds et al., 2011; Squicciarini et al., 2015; Van Hee et al., 2015; D. Yin et al., 2009; R. Zhao et al., 2016)) did not use feature selection to determine the most significant features. Studies (Chavan & Shylaja, 2015) (Hosseinmardi et al., 2015) used chi-square and PCA respectively for selecting the significant feature form the extracted features.

Apart from testing the significance of features in constructing a cyberbullying detection method, selecting the best set of features ensures that only significant features are used for learning the classifier. As discussed above although many studies on textual cyberbullying did not use feature selection algorithms, this step is a good practice to ensure that the features used in constructing a detection method are significant and remove redundant features (Prieto et al., 2014). Comprehensive comparative studies (C. C. Aggarwal & Zhai, 2012; Lee, Mahmud, Chen, Zhou, & Nichols, 2015; Prieto et al., 2014; Y. Yang & Pedersen, 1997) on different feature selection approaches for textual classification have been conducted to provide performance comparison of feature selection algorithms in text classification applications. Prominent feature selection algorithms, which are used, for text classification are information gain pearson correlation and chi-square test. These algorithms are briefly discussed in following subsections.

(a) ***Information gain***

Information gain is the estimated decrease in entropy produced by separating the examples according to given features. Entropy is a well-known concept in information theory, and it describes the (im)purity of an arbitrary collection of examples (Gray, 1990).

Information gain is used to calculate the strength or significance of features in a classification method according to class attribute. Information gain (Qabajeh & Thabtah, 2014) evaluates how well a specified feature divides training data sets with respect to class labels as explained by the following equations. Given a training data set (Tr).

$$Entropy\ for\ all\ training\ data\ (Tr) = I\ (Tr) = -\sum P_n\ log_2\ P_n \quad (2.1)$$

where P_n is the probability that Tr belongs to class n.

For attribute Att data sets, the expected entropy is calculated as

$$Expected\ entropy\ for\ Att\ = I(Att) = \sum \left(\frac{Tr_{att}}{Tr}\right) * I(Tr_{att}). \quad (2.2)$$

The information gain of attribute Att data sets is

$$information\ gain(Att) = I(Tr) - I(Att). \quad (2.3)$$

(b) ***Pearson correlation***

Correlation-based feature selection is a commonly used method for reducing feature dimensionality and evaluating the discrimination power of a feature in classification methods. It is also a straightforward method for choosing significant features. Pearson

correlation measures the relevance of a feature by computing the Pearson correlation between it and a class. Pearson correlation coefficient measures the linear correlation between two attributes (Benesty, Chen, Huang, & Cohen, 2009). The subsequent value lies between -1 and +1, with -1 implying absolute negative correlation (as one attribute increases, the other decreases), +1 denoting absolute positive correlation (as one attribute increases, the other also increases), and 0 meaning no linear correlation between the two attributes. For two attributes or features X and Y, Pearson correlation coefficient measures the correlation (M. A. Hall, 1999) as

$$r_{xy} = \frac{\sum(x_i - \bar{x})\,(y_i - \bar{y})}{(n-1)S_x S_y}, \tag{2.4}$$

where $\bar{x}$ and $\bar{y}$ are the sample means for X and Y, respectively, S_x and S_y are the sample standard deviations for X and Y, respectively, and n is the size of the sample used to compute the correlation coefficient (M. A. Hall, 1999).

(c) ***Chi-square test***

Another common feature selection method is chi–square test. This test is used in statistics, among other variables, to test the independence of two occurrences. More precisely in feature selection, chi-square is used to test whether the occurrences of a particular feature and class are independent. Thus, the following quantity is assumed for each feature, and they are ranked by their score.

$$N = \frac{N[P(f,c_i)P(\bar{f},\bar{c_i}) - P(f,\bar{c_i})P(\bar{f},c_i)]^2}{P(f)P(\bar{f})P(c_i)P(\bar{c_i})} \tag{2.5}$$

Chi-square test (Zheng, Wu, & Srihari, 2004) assesses the independence between feature f and class c_i in which N is the total number of documents.

2.2.1.4 Machine learning algorithms

Many types of machine learning algorithms exist, but nearly all studies on cyberbullying detection in OSNs used the most established and widely used type, that is, supervised machine learning algorithms (Nadali et al., 2013; Squicciarini et al., 2015). The accomplishment of a machine learning algorithms is determined by the degree to which the method accurately converts various types of prior observation or knowledge about the task. Certainly, much of the practical application of machine learning considers the details of a particular problem and then selects an algorithmic method that allows accurate encoding of those facts. However, no optimal machine learning algorithm works best for all problems (Buczak & Guven, 2015; Mangaonkar et al., 2015; Wolpert & Macready, 1997). Therefore, most researchers select and compare between many supervised classifiers to find the best ones for their problem. The selection of classifiers is commonly based on the most commonly used classifiers in the field as well as data features available for the experiment. However, researchers can only decide which algorithms to adopt for constructing a cyberbullying detection method by using comprehensive practical experiment as basis. Table 2.2 summarizes commonly used machine learning algorithms for constructing cyberbullying detection method.

Table 2.2 : Summary of Machine Learning Algorithms Tested in the Cyberbullying Literature

Study	SVM classifier family	Naïve Bayes	Random forest	Decision tree classifier family	K-nearest neighbor (KNN)	Logistic regression	Association rules	Rule-based algorithm (JRip)
(Chavan & Shylaja, 2015)	✓	✗	✗	✗	✗	✓	✗	✗
(Y. Chen et al., 2012)	✓	✓	✗	✗	✗	✗	✗	✗
(Dadvar et al., 2012b)	✓	✗	✗	✗	✗	✗	✗	✗
(Dinakar et al., 2011)	✓	✓	✗	✓	✗	✗	✗	✓
(Galán-García et al., 2014)	✓	✓	✓	✓	✓	✗	✗	✗
(García-Recuero, 2016)	✗	✗	✓	✗	✗	✗	✗	✗
(Van Hee et al., 2015)	✓	✗	✗	✗	✗	✗	✗	✗
(Hosseinmardi et al., 2015)	✓	✗	✗	✗	✗	✗	✗	✗
(Sanchez & Kumar, 2011)	✗	✓	✗	✗	✗	✗	✗	✗
(Mangaonkar et al., 2015)	✓	✓	✗	✗	✗	✓	✗	✗
(Margono et al., 2014)	✗	✗	✗	✗	✗	✗	✓	✗

(Nahar et al., 2013)	✓	✗	✗	✗	✗	✗	✗	✗
(R. Zhao et al., 2016)	✓	✗	✗	✗	✗	✗	✗	✗
(Squicciarini et al., 2015)	✓	✓	✗	✗	✗	✓	✗	✗
(Reynolds et al., 2011)	✓	✗	✗	✓	✗	✗	✗	✓
(D. Yin et al., 2009)	✓	✗	✗	✗	✗	✗	✗	✗

The following sections describe the machine learning algorithms, which are commonly used for constructing cyberbullying detection method as shown in Table 2.2.

(a) ***Support vector machine***

SVM was used in construction of cyberbullying detection methods in (Chavan & Shylaja, 2015; Y. Chen et al., 2012; Dadvar et al., 2012b; Dinakar et al., 2011; Mangaonkar et al., 2015; Van Hee et al., 2015). SVM is a supervised machine learning classifier and commonly used in text classification (Joachims, 1998). SVM is constructed by finding a separating hyperplane in the feature attributes between two classes in which the distance between the hyperplane and the nearest data point of each class is maximized (Hsu, Chang, & Lin, 2003). Theoretically, SVM has been developed from statistical learning theory (S. Tong & Koller, 2001). In SVM algorithm, the optimal separation hyperplane pertains to the separating hyperplane that minimizes misclassifications that is achieved in the training step. The approach is based on a minimized classification risk (Buczak & Guven, 2015; Vapnik, 2013).

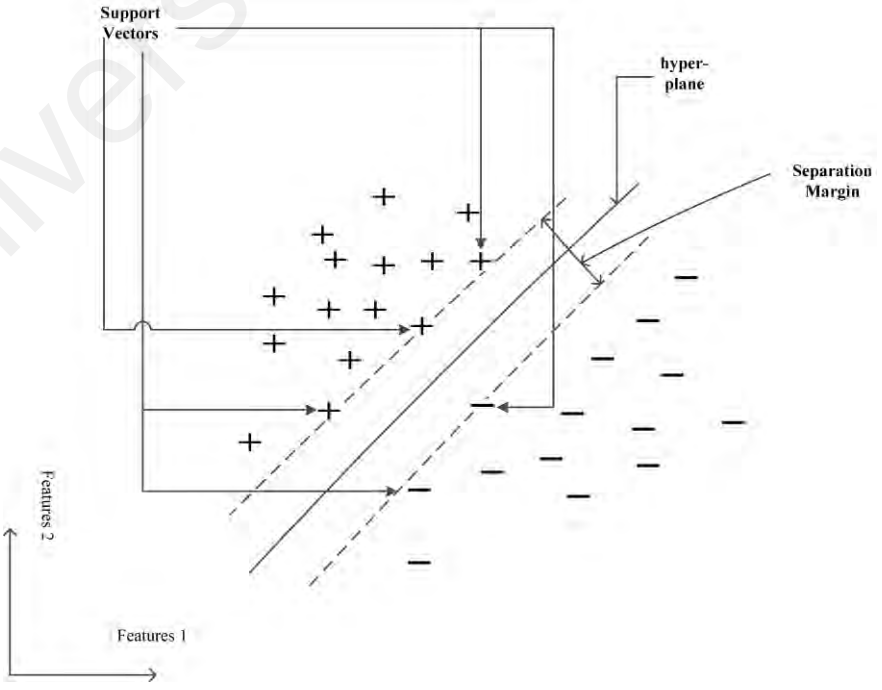

Figure 2.2: SVM linear separation in feature space

SVM is initially established to classify linearly separable classes of objects, as shown in Figure 2.2. A two-dimensional plane comprises linearly separable objects from different classes (e.g., positive or negative). SVM primarily aims to separate the two classes effectively. SVM identifies the exceptional hyperplane, which provides maximum margin, by maximizing the distance among the hyperplane and the nearest data point of each class.

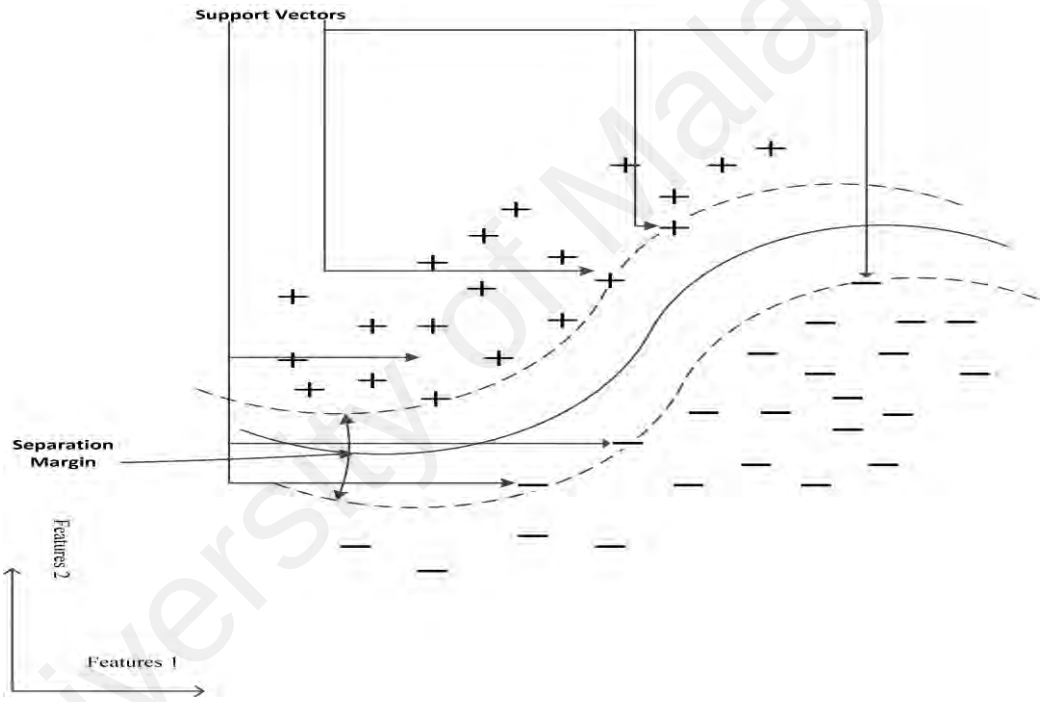

Figure 2.3: SVM non-linear separation in feature space

In real-time application, precisely determining the separating hyperplane becomes difficult and impossible in some cases. Consequently, SVM has been developed to adapt to these cases and can now be used as a classifier for non-separable classes. SVM has become a capable classification algorithm because of its significant characteristics. That is, SVM can powerfully separate non-linearly separable patterns by transforming them to a high-dimensional space using the kernel model (Boser, Guyon, & Vapnik, 1992) as shown in Figure 2.3.

The advantage of SVM is high speed and scalability as well as the capability of detecting intrusions in real time and updating the training patterns dynamically.

(b) ***Naïve Bayes algorithm***

NB was used in construction of cyberbullying detection method in (Y. Chen et al., 2012; Dinakar et al., 2011; Galán-García et al., 2014; Mangaonkar et al., 2015; Sanchez & Kumar, 2011). By applying Bayes' theorem between features, NB classifiers are constructed. Bayesian learning is commonly used for text classification. This method assumes that the text is generated by a parametric model and utilizes training data to compute Bayes-optimal estimates of the model parameters. With these approximations, it categorizes generated test data (McCallum & Nigam, 1998).

NB classifiers can deal with arbitrary number of continuous or categorical independent features (Buczak & Guven, 2015). Using the assumption that the features are independent, a high-dimensional density estimation task is reduced to one-dimensional kernel density estimation (Buczak & Guven, 2015).

NB algorithm is a classification algorithm that is based on the application of Bayes theorem with strong (naive) independence assumptions. Given a class variable $\boldsymbol{y}$ and a dependent feature vector $\boldsymbol{x_1}$ through $\boldsymbol{x_n}$, Bayes' theorem states the following relationship:

$$p(y|\, x_1, x_2, \cdots, x_n) = \frac{p(y)p(x_1, x_2, \cdots, x_n \mid y)}{p(x_1, x_2, \cdots, x_n)}. \tag{2.6}$$

Using the naive independence assumption, the following equation is obtained:

$$p(x_i|\, y, x_1, \cdots x_{i-1}, x_i + 1, \ldots., x_n) = p(x_i|\, y). \tag{2.7}$$

for all i, this relationship is simplified as follows:

$$p(y|\ (x_1, x_2, \cdots, x_n) = \frac{p(y)\prod_{i=1}^{n} p(x_i|\ y)}{p(x_1, x_2, \cdots, x_n)} \quad (2.8)$$

Given that $p(y|\ (x_1, x_2, \cdots, x_n)$ is a constant because of the input, the following classification rule is used:

$$p(y|\ (x_1, x_2, \cdots, x_n) \propto p(y)\prod_{i=1}^{n} p(x_i|\ y). \quad (2.9)$$

This technique is discussed in further detail by Zhang (H. Zhang, 2004b).

NB algorithm is one of the most effective and inductive machine learning algorithms (H. Zhang, 2004a), and it has been used as a classifier in several studies on social media (Bora, Zaytsev, Chang, & Maheswaran, 2013; D. M. Freeman, 2013; A. H. Wang, 2010).

(c) ***Random forest***

Random forest (RF) was used in the construction of cyberbullying detection methods in (García-Recuero, 2016; Van Hee et al., 2015). RF is a machine-learning method that combines decision trees and ensemble learning (Breiman, 2001). This method fits several classification trees to a data set and then combines the predictions from all the trees (Cutler et al., 2007). Therefore, RF consists of many trees that are used randomly to select feature variables for the classifier input. The construction of RF is achieved by the following simplified steps:

I. Consider the number of examples (cases) in training data to be *N* and the number of attributes in the classifier to be *M*.

II. Create a number of random decision tress by selecting attributes randomly. Select a training set for each tree by choosing *n* times from all

N available instances. The rest of the instances in the training set are used to approximate the error of the tree by forecasting their classes.

III. For each tree's nodes, select random *m* variables on which to base the decision at that node. Compute the finest split using these *m* attributes in the training set. Each tree is completely built and is not pruned, as can be done in building a normal tree classifier.

IV. Subsequently, an enormous number of trees are created. These decision trees vote for the most popular class. These processes are called RFs (Breiman, 2001).

RF can be defined as a classifier comprising a group of tree-structured classifiers $(h(\boldsymbol{x}, Q_K), k = 1$, where Q_K refers to independent identically distributed random vectors, and each tree votes for the most popular class at input x (Breiman, 2001).

(d) ***Decision tree***

Decision tree classifiers were used in construction of cyberbullying detection method in (Dinakar et al., 2011; Galán-García et al., 2014). A decision tree is a graphic method in which each branch node represents a choice between alternatives. Graphic approach is used in decision trees to compare competing alternatives (Safavian & Landgrebe, 1991). Decision trees are samples to understand and easy to interpret; hence, the decision tree algorithm can be used to analyze the data and build a graphic model for classification. The most commonly improved version of decision tree algorithms used for cyberbullying detection is C.45 (Dinakar et al., 2011; Galán-García et al., 2014; Reynolds et al., 2011). C4.5 can be explained as follows: Given that N number of examples, C4.5 first produces an initial tree through the divide-and-conquer algorithm as follows (X. Wu et al., 2008):

If all the examples in N belong to the same class or N is small, the tree is a leaf labeled with the most frequent class in N. Otherwise, a test is selected based on, for example, the mostly used information gain test on a single attribute with two or more outputs. As the test is the root of the tree creation partition of N into subsets N_1, N_2, N_3 ..... regrading to the outputs for each examples, the same procedure is applied recursively to each subset (X. Wu et al., 2008).

(e) ***K-nearest neighbors (KNNs)***

KNN is a nonparametric technique that attempts to decide the KNNs of x_0 and uses a majority vote to calculate the class label of x_0. The KNN classifier usually uses Euclidean distances as the distance metric (Soucy & Mineau, 2001). To demonstrate a KNN classification, classifying a new input posts (from a testing set) is considered using a number of known manually labeled posts. This example is shown in Figure 2.4, which shows positive and negative examples and unknown examples to be classified as either positive or negative. The main task of KNN is to classify the unknown example grounded on a nominated number of its nearest neighbors, that is, to finalize the class of unknown examples as either a positive class or a negative class. KNN classifies the class of unknown examples using majority votes for nearest neighbors of the unknown classes. For example, in Figure 2.4, if KNN is one nearest neighbor [estimating the class of an unknown example using the one nearest neighbor vote (k = 1)], then KNN will classify the class of the unknown example as positive (because the closest point is positive). For two nearest neighbors (estimating the class of an unknown example using the two nearest neighbor vote), KNN is unable to classify the class of the unknown example because the second closest point is negative (positive and negative votes are equal). For four nearest neighbors (estimating the class of an unknown example using the four nearest neighbor vote), KNN will classify the class of the unknown example as positive (because the three closest points are positive and only one vote is negative).

The KNN algorithm is one of the simplest classification algorithms, but despite its simplicity, it can provide competitive results (Deng, Zhu, Cheng, Zong, & Zhang, 2016). KNN was used in the construction of cyberbullying detection methods in (Galán-García et al., 2014).

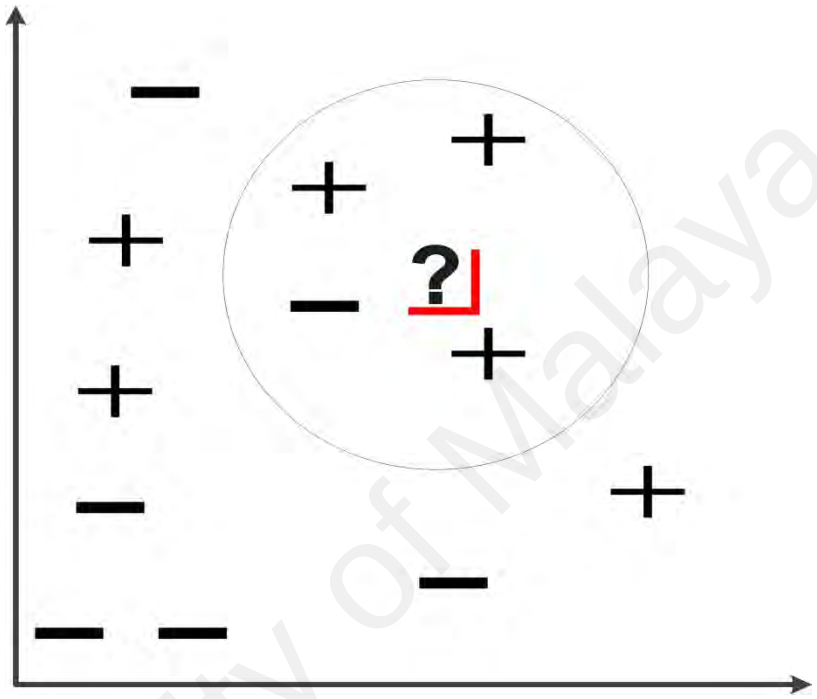

Figure 2.4: KNN algorithm

(f) ***Logistic regression classification***

Logistic regression is one of common techniques imported by machine learning from the statistics field. Logistic regression is an algorithm that builds a separating hyperplane between two data sets by means of the logistic function (Dreiseitl et al., 2001). The logistic regression algorthim takes inputs (features) and generates a forecast according to the probability of the input being appropriate for A class. For example, if the probability is >0.5, the classification of the instance will be a positive class; otherwise, the prediction is for the other class (negative class) (Hosmer Jr, Lemeshow, & Sturdivant, 2013). Logistic regression was used in the construction of cyberbullying detection methods in (Chavan & Shylaja, 2015; Mangaonkar et al., 2015).

2.2.1.5 Evaluation

The essential objective of constructing detection method based on machine learning is to generalize more than the training data set (Domingos, 2012). When a machine learning model is applied to real example, it can perform well. Accordingly, the data is divided into two parts. The first part is the training data that are used to train the machine learning algorithms; the second part is the testing data that are used to test the machine learning algorithms. However, Separately dividing the data into training and testing data is not widely employed (Domingos, 2012), especially in applications in which deriving training and testing data are difficult. For example, in cyberbullying detection, most state-of-art studies manually labeled data; hence, creating labeled data is expensive. These issues can be reduced by cross-validation, that is, randomly dividing the training data into, for example, ten subsets, and this process is called 10-fold cross-validation. Cross-validation involves the following steps: keep a fold separate (the model does not see it) and train data on the model using the remaining folds; then test each learned classifier on the fold which it did not see; and average the results to see how well the particular parameter setting performs (Domingos, 2012; Kohavi, 1995).

(a) ***Evaluation metric***

Researchers measure the effectiveness of a proposed method to determine how successfully the method can distinguish cyberbullying from non-cyberbullying by using various evaluation measures. Reviewing common evaluation metrics in the research community is important to understand the performance of conflicting models. The most commonly used metric for evaluating cyberbullying classifier in OSNs are as follows:

Accuracy: it has been used to evaluate the cyberbullying detection methods in (Dinakar et al., 2011; Hosseinmardi et al., 2015; Mangaonkar et al., 2015; Reynolds et al., 2011), and it is calculated as follows:

$$accuracy = \frac{(tp+tn)}{(tp+fn+fp+tn)} \quad (2.10)$$

Precision, recall, and F-measure: they have been used to evaluate the cyberbullying detection methods in (Y. Chen et al., 2012) (Dadvar et al., 2012b) (Van Hee et al., 2015) (Mangaonkar et al., 2015). They are calculated as follows:

$$\text{Precision} = \frac{\text{tp}}{(\text{tp}+\text{fp})}, \quad (2.11)$$

$$\text{Recall} = \frac{\text{tp}}{(\text{tp}+\text{fn})}, \quad (2.12)$$

$$\text{F-measure} = \frac{2*\text{precision}*\text{recall}}{\text{precision}+\text{recall}}. \quad (2.13)$$

- True positive (TP) is the instances belonging to class X and correctly classified as X.
- False negative (FN) is the instances that belong to class X and incorrectly classified as Y.
- True negative (TN) is the instances that belong to class Y and correctly classified as to Y.
- False positive (FP) is the instances that belong to class Y and incorrectly classified as X.

AUC: The area under the curve provides a signal of the discriminatory rate of the classifier at various operating points (Chavan & Shylaja, 2015; Galán-García et al., 2014). The principal advantage of AUC is that it is more robust than accuracy in class imbalance situations

2.2.2 Issues Related Current Cyberbullying Detection Methods in OSNs

In this section, the issues extracted from the reviewed studies are identified. The main issues are related to feature engineering. Other issues are found in data collection, and evaluation metric selection are identified and discussed in following subsections.

2.2.2.1 Feature Engineering

Features are vital components in improving the effectiveness of machine learning detection methods (Domingos, 2012). Most of the discussed studies attempted to provide effective machine learning solution to cyberbullying in OSNs by providing significant features (see Table 2.1). However, these studies overlooked other important features. Online cyberbullies change the way they use words and acronyms to engage in cyberbullying. OSNs help create cyberbullying acronyms that have never been found in traditional bullying or are beyond social media norms (Dailymail, 2014). Recent survey response studies (questionnaire-based studies) have reported positive correlations between different variables, such as personality (Connolly & O'Moore, 2003; Corcoran et al., 2012) and sociability of a user in an online environment (Navarro & Jasinski, 2012), and cyberbullying occurrences. Observations by these studies are important to understand such behavior in online environment. However, these observations are yet to be used with machine learning algorithm to provide significant methods. These observations can be useful when transformed to practical form (features) that can be employed to develop effective machine learning detection methods for cyberbullying in OSNs. Rich information provided by OSNs about users' information, textual information by users and network information of users should be utilized to convert observations into a set of features. For example, two studies (Dadvar et al., 2012a; Dadvar, Trieschnigg, Ordelman, et al., 2013) attempted to improve the machine learning classifier performance by including features such as age and gender, which show improvement in classifier performance, but these features were limited to direct user

information mentioned on the online profiles of users. By contrast, most studies found that only a few users provide complete details in their online profiles (D. Nguyen, R. Gravel, D. Trieschnigg, & T. Meder, 2013; Peersman, Daelemans, & Van Vaerenbergh, 2011). These studies have suggested the useful practice of utilizing words expressed in the content (tweet) to identify user age and gender (D. Nguyen et al., 2013; Peersman et al., 2011). Cyberbullying is associated with the aggressive behavior of a user. A survey study demonstrated that hostility significantly predicts cyberbullying (Arıcak, 2009). Among the five personality traits, bullying and cyberbullying are strongly related to neuroticism (Connolly & O'Moore, 2003; Corcoran et al., 2012). Therefore, predicting if a user has used words related to neuroticism is significant to detect cyberbullying. Predicting neurotic personality using neurotic-related words as features can provide useful discriminative features to construct machine-learning classifier.

Similarly, a strong correlation is found between cyberbullying behavior and sociability of a user in an online environment (Navarro & Jasinski, 2012). Users who are more active in an online environment are more likely to engage in cyberbullying (Balakrishnan, 2015). On the basis of these observations, OSNs own features that can be used as signals to measure the sociability of a user, such as number of friends, number of posts, URLs in posts, hashtags in posts, and number of users engaged in conversation (mentioned). The combination of these features with traditionally used ones, such as profanity features, can provide comprehensive discriminative features. The reviewed studies (see Table 2.1) focused on using either traditional feature method (e.g., bag-of-words) or information such as age or gender limited to user profile information (information written by users in their profile). As such information is limited, comprehensive features should be proposed to improve classifier performance.

Moreover, maintaining a precise and accurate process in constructing machine learning method from start (data collection) to end (evaluation metric selection) is important to ensure that the proposed features hold significance in improving classifier performance. The following subsection analyses the other issues related to constructing effective machine learning method for cyberbullying detection in OSNs.

2.2.2.2 Data Collection

Cyberbullying detection studies collected their data sets using specific keywords or profile IDs. However, by merely tracking tweets that contained specific keywords, these studies introduced a potential sampling bias (Cheng & Wicks, 2014; Morstatter, Pfeffer, Liu, & Carley, 2013), limited their detection coverage to such tweets, and disregarded many other tweets relevant to cyberbullying. These data collection approaches narrow the detection range of cyberbullying. The selection of keywords for tracking tweets is also subject to the author's perception on cyberbullying. Classifiers must be extended from "enriched data" to the complete range of tweets (Xu et al., 2012). The important objective of machine learning is to generalize and not to limit the examples in a training data set (Domingos, 2012). Researchers should investigate whether the sampled data is extracted from data that well represents all activities in OSNs (Morstatter et al., 2013). Extracting well-representative data from an OSN is the first step toward building effective machine learning detection method. The OSNs' public application program interface (API) only provides access to a small sample of all relevant data in some instances, and thus poses a potential for sampling bias (Cheng & Wicks, 2014; González-Bailón et al., 2014; Y. Liu et al., 2014). For instance, a previous research (Morstatter et al., 2013) discussed whether data extracted from Twitter's streaming API is a sufficient representation of the activities in the Twitter network as a whole; he compared keyword (words, phrases, or hashtags), user ID, and geo-coded sampling. Twitter's streaming API returns data set with some bias when keyword or user ID

sampling is used. By contrast, using geo-tagged filtering provides good data representation (Morstatter et al., 2013). Bearing these points in mind, researchers should ensure minimum bias as possible when they extract data to guarantee that the examples selected to be represented in training data are generalized and provide effective model when applied to testing data. Bias in data collection can impose bias in selected training data set based on specific keywords or users, and such bias consequently introduces overfitting issues that affect the ability of a machine learning method to make reliable predictions on general untrained data.

2.2.2.3 Evaluation Metric Selection

Accuracy, precision, and recall are commonly used as evaluation metrics, along with AUC (Chavan & Shylaja, 2015; Galán-García et al., 2014). Evaluation metric selection is an important task. The selection is on the basis of understanding the nature of manually labeled data. Selecting inappropriate evaluation metric may result in claiming a better performance according to the selected evaluation metric. Then, the researcher may find the results to be significantly improved although an investigation on how the machine learning model is evaluated may find contradicting results and may not truly reflect the improvement of performance. For instance, cyberbullying posts are commonly considered as abnormal cases, whereas non-cyberbullying posts are considered the normal cases. The ratio between cyberbullying and non-cyberbullying is normally large. Generally, the non-cyberbullying posts comprise the large portion. For example, if 1000 posts were manually labeled as cyberbullying and non-cyberbullying. Assuming non-cyberbullying posts are 900, and the remaining 100 posts are cyberbullying. If a machine learning classifier classifies all 1000 posts as non-cyberbullying and unable to classify any posts (0) as cyberbullying, then this classifier is considered impractical. By contrast, if researchers use accuracy metric as the main

evaluation metric, the accuracy of this classifier calculated as mentioned in equation 2.10 is 90%.

In the example, the classifier failed to classify any cyberbullying posts but obtained an accuracy of 90%. Knowing the nature of manually labeled data is important to selecting evaluation metric. In case the data are imbalanced, then researchers may need to select AUC as the main evaluation metric. The key advantage of AUC is that it is more robust than accuracy, precision, recall, and f-measure in class imbalance situations (Japkowicz & Shah, 2011). Cyberbullying and non-cyberbullying data are commonly imbalanced data sets (non-cyberbullying posts are more than cyberbullying ones) that closely represent the real life data that machine learning algorithms need to train on. Accordingly, the learning performance of these algorithms is independent of data skewness (Mangaonkar et al., 2015). Special care should be taken in selecting the main evaluation metric to avoid uncertain results and appropriately evaluate the performance of machine learning algorithms.

2.3 Influential Spreaders in OSNs

Understanding how the certain rumors or negative behaviors spread in the OSNs is beneficial for developing effective methods for controlling such adverse practices. Therefore, the influential spreaders identification holds the remarkable practical importance, and has recently attracted researcher's attention (Gao et al., 2011; Wen et al., 2014b; Z.-K. Zhang et al., 2016)

Section 2.3.1 describes the significance of identification influential spreader for minimizing the cybercrimes in OSN. Section 2.3.2 comprehensively reviews different methods that are widely used to identify influential spreaders in OSNs. The issues of

each method are also outlined in section 2.3.3. Section 2.3.4 compares different influential spreaders methods. Section 2.3.5 discusses the performance evaluations of the identification of influential spreaders. Finally, section 2.3.6 provides a summary and taxonomy of the identification of influential spreaders researches in the OSN Context.

2.3.1 Significance of Influential Spreaders Identification for Minimizing the Cybercrimes in OSNs

The significance of identifying influential spreader for minimizing the cybercrimes in OSN can be categorized into disinformation restraint and information dissemination applications as following:

(a) ***Disinformation Restraint***

The OSN platforms also allow cyberbullying, spam, viruses, negative behaviors, rumors, gossips, and other forms of disinformation to spread to the users (Wen et al., 2014b). The challenging task is to control the propagation of the unwanted contents is such large networks. Influential spreaders are identified and the detection methods are applied at them to restrain the unwanted contents in the OSNs (Pastor-Satorras & Vespignani, 2002; Yan, Chen, Eidenbenz, & Li, 2011; Zou, Towsley, & Gong, 2007) (Comin & da Fontoura Costa, 2011; Y.-Y. Liu, Slotine, & Barabási, 2011; Nepusz & Vicsek, 2012). The identification of the influential spreaders is useful for proposing the immunization strategies for large networks. The strategy is to identify the influential spreaders in the entire network. The immunization (detection methods) is subsequently applied on the selected users to achieve the maximal effect of the cybercrime containment at the minimal cost (W. Yang, Wang, & Yao, 2015). Consequently influential spreaders identification methods are important to restrain the spread of cybercrime in large social networks

(b) ***Information dissemination***

OSNs have become a popular means of sharing and disseminating information. The massive popularity of OSNs has introduced many applications that aim to spread information. These applications maximize the spread of products (Weinberg, 2009) and news (Ho & Dempsey, 2010; Zhu, 2013). More importantly, these influential spreaders can be targeted for enhancing information spread to prevent cybercrimes. These influential spreaders can develop user consciousness by spreading prevention strategies to a large number of users. These strategies include creating an application for spreading cyberbullying prevention by making kind words go viral (Patchin & Hinduja, 2013; Z.-K. Zhang et al., 2016). or clarifying the truth in case of rumors (Budak et al., 2011). Identifying influential users in information dissemination is significant for designing the strategies to accelerate information spread.

2.3.2 Influential Spreader Identification Methods for OSNs

Identifying influential spreaders in a social network community is essentially related to user influence measurement; the users with higher influence are the more influential spreaders (Zhu, 2013). User influence is measured based on various factors with many techniques. Identifying the most important factors and using the most appropriate techniques are challenging tasks, particularly in OSNs, because of the diverse characteristics of such networks. Many previous studies have used different user influence measurement approaches for identifying the influential spreaders in complex networks. However, these measurements may not be directly applied to OSNs because of their diverse characteristics. Therefore, this section investigates the methods for identifying the influential spreaders in the OSN context.

2.3.2.1 Degree centrality

Degree centrality is a straightforward and widely used topological measure of user influence. Generally in a network, a high-degree node is assumed to be in authority for the largest spread processes (Albert, Jeong, & Barabási, 2000; Pastor-Satorras & Vespignani, 2001). Users with high connectedness have the opportunity to influence the behavior of others (Albert & Barabási, 2002). A network can be either directed or undirected. Directed networks can be described by in-degree and out-degree centrality. In-degree centrality pertains to the number of links that link to the node from other nodes, whereas out-degree centrality denotes the number of links that link from the node to other nodes. In directed networks, in-degree centrality usually refers to the popularity of the user, whereas out-degree centrality typically connotes the sociality of the user (Jiang et al., 2013; Mislove, Marcon, Gummadi, Druschel, & Bhattacharjee, 2007).

In the OSN context, degree counts refer to the size of the audience for the users, number of social relationships, or the amount of interaction. Several studies have used the degree measure to identify the most influential users in OSNs (Bakshy, Hofman, Mason, & Watts, 2011; Cha, Haddadi, Benevenuto, & Gummadi, 2010; Kim & Han, 2009; Romero, Galuba, Asur, & Huberman, 2011). However, the degree measurement alone cannot accurately reveal the influence of the users, and having high degree is not necessary to be considered an influential spreader (Cha et al., 2010). Consequently, high connectedness is also related to other social factors rather than influence.

2.3.2.2 Closeness centrality

Closeness centrality calculates how close a node is to all of the other nodes in the network. It is based on the length of the average shortest path between a user and all of the users in the network. At a node, i is the inverse of the average distance to all other

nodes (Bavelas, 1950; L. C. Freeman, 1979). Users are categorized according to their closeness centrality from the highest to the lowest ranking.

2.3.2.3 Betweenness centrality

Betweenness centrality is defined as the shortest path number from all of the nodes to all of the other nodes through a network. For example, study (Catanese, De Meo, Ferrara, Fiumara, & Provetti, 2012) have applied betweenness centrality to data graphs from Facebook to identify the central nodes of the network.

2.3.2.4 Eigenvector centrality

Eigenvector centrality is used in measuring the importance of a node in a network. It is adopted to identify the most influential node(s) in the graph (Borgatti & Everett, 2006; Duda, Hart, & Stork, 2012; H. He, 2007).

The previously reviewed classical centrality measures (i.e., degree, closeness centrality, betweenness centrality, and eigenvector centrality) are direct methods for identifying influential spreaders. However, given their limitations, these measures are not commonly used in the OSN context compared with other method , such as PageRank, k-core, and learning methods.

2.3.2.5 PageRank-like methods

PageRank is a network-based diffusion algorithm. It is the famous Google algorithm for ranking websites that was initially proposed by Brin *et al*. (Brin & Page, 2012). PageRank is a global ranking of all web pages, regardless of their contents, based solely on their connected links and locations on the web graph. PageRank scores recursively and two key metrics are considered, namely, incoming links counts and the PageRank value of all incoming links. PageRank was initially used in ranking the pages on the World Wide Web. It has created a revolution in the web search field, contributing to the

search engine Google. PageRank is implemented in many applications to rank a wide-ranging array of data. It is characterized by simple assumptions, direct implementation, and comparatively low computational complexity; thus, scholars are motivated to use PageRank to recognize the critical influential spreaders in OSNs in numerous practical situations. In the simplified algorithm, PageRank is expressed as follows.

$$P\,R(u) = (1 - d)/N + d \sum_{v \in M(u)} PR(v)/L(v), \qquad (2.14)$$

where N is the total number of web pages in the network; $L\ (v)$ is the number of outgoing links from page v; $M\ (u)$ refers to the set of web pages pointing to web page u; and d (with $0 \leq d \leq 1$) is a damping factor that is usually set to 0.85 (Brin & Page, 2012).

PageRank has been used to identify the influential spreaders in OSN websites (Java et al., 2006; Nguyen & Szymanski, 2013). Furthermore, a study (Z. Yin & Zhang, 2012) in the Sina Microblog proposed a model that takes into account the personal characteristics of users, such as their level of activity and willingness to retweet, to calculate the influence between a pair of users (influence score). User influence is subsequently measured using a weighted PageRank: a user-to-user network is generated and influence score as edge weights is used. In contrast to other studies, this research aimed to measure users between two users (pairwise influence) rather than to measure global influence (the user influence in the entire network).

The characteristic of OSNs differ from those of traditional web pages; thus, the PageRank algorithm has been extended. Many extensions are made to PageRank to improve the identification of influential spreaders in OSNs. The different PageRank extensions are discussed below.

TunkRank (Tunkelang, 2009): The idea behind this algorithm can be described as follows. First, the amount of influence of a user in Twitter is evenly distributed among his/her followers; that is, if (i)a user in a Twitter network and if (i) is a member of followers(j), then there is a $1/\|\text{Following}\ (i)\|$ probability that (i) will read a tweet posted by (j), where $\text{Following}\ (i)$ is the set of people that (i)follows. Second, if (i) reads $(j)'s$ tweets, the constant probability (p) exists that (i) is going to retweet the tweets. The preceding concept is mathematically represented as

$$Influence(i) = \sum_{j\ \text{Follower of}\ i} \frac{1+p*Influence(j)}{\|following\ of(j)\|} \tag{2.15}$$

Where p is the probability that (i) is going to retweet (j)'s tweets.

TURank (Yamaguchi et al., 2010): This algorithm studies both the relationship between the users and their posts and that between users and their friends' networks. TURank network is constructed using two different nodes. The first type of nodes is represented by the users, whereas the second type is represented by the tweets. The edges are constructed such that post-relation links between the users and their tweets, following–relations links between the users and retweets links between the tweets. A network called user–tweet network is used to model the information flow in Twitter and calculate the users' ranking scores. This ranking algorithm is constructed based on several observations; for instance, a user who is followed by many influential spreader is likely to be an influential spreader, a tweet retweeted by many influential spreaders is likely to be a valuable tweet, and a user who posts many valuable tweets is likely to be an influential user.

TwitterRank (J. Weng et al., 2010): This algorithm uses both network structure and topics to rank the influence of the users in Twitter. TwitterRank first computes each user's topic distribution based on tweet content using latent dirichlet allocation (Blei,

Ng, & Jordan, 2003) (Griffiths & Steyvers, 2004). It computes topic similarity scores between each pair of users, user i and user$j$, and each topic *t*, denoted by($sim_t\ (i,j)$). TwitterRank subsequently measures the user influence in topic *t*, denoted as $\overrightarrow{TR_t}$, and it is calculated iteratively as follows.

$$\overrightarrow{TR_t} = \gamma p_t \times \overrightarrow{TR_t} + (1-\gamma)Et \tag{2.16}$$

Where Et is the teleportation vector, γ is the value between 0 and 1 to control the probability of teleportation, and p_t is the transition probability defined as

$$p_t(i,j) = \frac{T.tweets_j}{\sum_{a:\ i\ \ follows\ \ a} T.tweets_a} * sim_t(i,j) \tag{2.17}$$

Where $T.tweets_j$ denotes the total number of tweets by user$j$, $\sum_{a:\ i\ \ follows\ \ a} T.tweets_a$ is the total number of tweets posted by all of the other user i friends, and $sim(i,j)$ is topic similarity between user i and user j.

LeaderRank (Lü, Zhang, Yeung, & Zhou, 2011): This algorithm has introduced a ground node. The ground node is connected to each node in the network using bidirectional links. The ranking process starts with assigning one unit to each node in the network except for the ground node. Through the direct link, the unit is evenly distributed to the node neighbors. The process will continue similar to random walk for a directed network until the steady state is reached. The authors claimed that the proposed LeaderRank has an advantage over PageRank in the following aspects: given that the network is strongly connected, it converges more rapidly; it is more tolerant of noisy data, such as spurious and missing links; it is applicable to any type of network; and it is robust against spammers.

LeaderRank has been further improved by Li *et al*. in weighted LeaderRank (Qian Li, Zhou, Lü, & Chen, 2014). The weighted LeaderRank has outperformed the original LeaderRank. The improvement has been achieved by making the ground node more biased toward the nodes with more fans using a biased random walk. Weighted LeaderRank is capable of identifying the more influential spreaders. It is more tolerant of noisy data and more robust against intentional attacks compared with the original LeaderRank.

InfluenceRank (W. Chen et al., 2012): This algorithm is calculated using two models; the first model measures users' relative influence, whereas the second model measures user network global influence. The users' relative influence model is calculated based on three factors, namely, quality of the tweet, ratio of retweets, and topic similarity between users. This model can be mathematically presented as follows.

$$Relative\ Influence\ RI\ (v_i, v_j) = Q_{vi} + R\ (v_i, v_j) + sim(v_i, v_j) \tag{2.18}$$

User $Relative\ Influence\ RI\ (v_i, v_j)$ measures the influence of user v_i on user v_j

Where Q_{vi} refers to the quality of tweet, and is calculated by

$$Q_{vi} = \frac{number\ of\ retweets\ (vi) + number\ of\ comments(vi)}{Total\ number\ of\ tweets\ (vi)} \tag{2.19}$$

$R\ (v_i, v_j)$ refers to the retweet ratio of user v_j to v_i, and is calculated by

$$R\ (v_i, v_j) = \frac{Retweets\ (v_i, v_j)}{Retweets\ (v_j)}. \tag{2.20}$$

$sim(v_i, v_j)$ refers to topic similarity between users v_i and v_j.

The user network global influence is PageRank-like algorithm; it is calculated by replacing $RI\left(v_i, v_j\right)$ value into the PageRank algorithm. It is mathematically calculated as follows.

$$Influence(v) = (1-\lambda)\frac{|Followers\,(v)|}{N} + \lambda \times \sum_{v_j \in Followers\,(v_i)} \frac{RI\,(v_i,v_j) \times Influence(v_j)}{Followees(v_j)}, \quad (2.21)$$

Where λ is damping factor.

The only difference between the above-represented method and the original PageRank algorithm is that scoring value is unequally distributed to all followers; it uses a biased random walk, and the scoring value is determined by$Relative\ Influence\ \left(RI\left(v_i, v_j\right)\right)$.

InfRank (Jabeur et al., 2012): InfRank is a PageRank-like algorithm that models the online social network micro-blogger to identify influencers, leaders, and discussers. This study measures users' influence by initially measuring their ability to spread the information in the network (i.e., by having a high number of retweets) and subsequently determining the good influential users in their retweet list (i.e., their tweets are retweeted by influential users). First, InfRank constructs a graph network using users as nodes, and an edge exists between v_i and v_j if at least one tweet of v_j is retweeted by v_i. This retweet has advantages and disadvantages compared with the follower graph. The retweet graph represents a strong social connection because one user can follow another without ever retweeting them and one user can retweet another without following them. However, this graph is relatively sparser than the follower graph. Second, InfRank distributes the ranking score retweet edges in terms of weights. The weighted edge is calculated as follows.

$$w\left(v_i, v_j\right) = \frac{\left|total\ number\ of\ tweets\ by\ v_j \cap total\ number\ of\ retweets\ by\ v_i\right|}{\left|total\ number\ of\ tweets\ by\ v_i\right|}. \quad (2.22)$$

SpreadRank (Ding et al., 2013): SpreadRank has been introduced to identify the influential spreaders in microblogs. This method is a variant of the PageRank method. SpreadRank constructs the network using users as nodes, and an edge exists between $v_i\ and\ v_j$ if at least one tweet of v_j retweeted is by v_i. The network whose edges are weighted by unique weight equal to $r/logT$, where (r) is the total number of repost and (T) is the total number of tweets. The time interval of repost is also considered (the author hypothesized that the faster the tweets are reposted, the higher is the diffused rate). Hence, the influence transition from $v_i\ to\ v_j$ is calculated as follows.

i. $$P\left(v_j, v_i\right) = \frac{\sum_{rij} f(t_{ij})}{logT_j} \quad (2.23)$$

where $\left(t_{ij}\right)$ is the time interval of its retweets, and T_j refers to the number of user v_j tweets.

The study used the location of users in information cascades to measure the teleport vector, which indicates that a location closer to the root (main tweets) will obtain higher scores.

ProfileRank (Silva, Guimarães, Meira Jr, & Zaki, 2013): ProfileRank is model inspired by PageRank. ProfileRank has introduced an integrated view of user influence and content relevance in information diffusion. The working principle of ProfileRank is that the influential spreaders can be identified by measuring their ability to create and propagate relevant content to a significant portion of the community. This algorithm is computed by random walks on a user-content bipartite network.

The preceding techniques are summarized in Table 2.3, and their methodologies, objectives, input parameters, network types, and weights are discussed.

Table 2.3: Comparison of PageRank-like algorithms

0	Methodology	Objective	Input Parameters	Network Type
TunkRank (Tunkelang, 2009)	This algorithm has introduced *p* constant probability that users retweet a tweet. TunkRank measures user's influence as the expected number of users who will read a tweet that they publish.	To recursively measure user influence by considering both the attention distribution and the retweeting probability	Number of followers and probability that users retweet a tweet	Follower network
TURank (Yamaguchi et al., 2010)	TURank constructs a user–tweet network in which users, their tweets, and followers are linked with their corresponding edges. User–tweet graph studies the information spread, and then performs a structural analysis to calculate a user's influence.	To recursively measure user influence by considering that a user followed by many influential users is likely to be an influential user, a tweet retweeted by many influential users is likely to be a valuable tweet, and a user who posts many valuable tweets is likely to be an influential user	Number of followers, number of tweets, and number of reposts	User–tweet network
TwitterRank (J. Weng et al., 2010)	TwitterRank uses both network structure and topic similarity to measure the influence of users in Twitter.	To recursively measure user influence by considering the topic similarity between users and the network structure	Number of followers and topic similarity	Follower network
LeaderRank (Lü et al., 2011)	LeaderRank constructs a user-to-user network using a directed network, and it has introduced a ground node. The ground node is connected to each node in the network using bidirectional links.	To propose algorithms that can effectively quantify user influence, and should be more tolerant of noisy data and robust against spammers	Number fans (directed link from fans to their leaders) and ground node (bidirectional links to every node in the network)	Fan–leader–ground node network
Weighted LeaderRank (Q. Li et al., 2014)	This algorithm is an extended version of LeaderRank. It uses biased random walk instead of that used in the original LeaderRank to make the ground node more biased toward nodes with more fans.	To improve the original LeaderRank by biasing the ground node toward the nodes that are more popular (with more fans)	Number of fans (directed link from fans to their leaders) and ground node (bidirectional links to every node in the network)	Fan–leader–ground node network
InfluenceRank(W. Chen et al., 2012)	This algorithm works in two steps; the first step involves measuring a user's relative influence, whereas the second step involves measuring user network global influence. It uses a biased random walk, and the scoring value is unequally distributed to all followers; it is determined by a user's relative influence.	To measure users' influences recursively by considering user's relative influence.	Number of followers, quality of followers, quality of tweet, retweet ratio, and topic similarity	Follower network

InfRank (Jabeur et al., 2012)	This study measures the users' influence by initially measuring their ability to spread the information in the network (i.e., by having a high number of retweets) and second by subsequently determining the good influential users in their retweet list (i.e., their tweets are retweeted by influential users).	To identify influencers, leaders, and discussers in online social network microblogs	Number of retweets and number of influential users in their retweet list	Retweet network
SpreadRank (Ding et al., 2013)	This method is a variant of the PageRank method. SpreadRank constructs the network using users as a node, and an edge exists between v_i *and* v_j, if at least one tweet of v_j is retweeted by v_i.	To recursively measure user influence by considering both the weights and time interval of the retweets	Number of reposts and time interval of retweets	Retweet network
ProfileRank (Silva et al., 2013)	The working principle of this algorithm is that users' influence can be calculated by measuring their ability to create and propagate a relevant content to a significant portion of the community.	To recursively measure user influence by considering user's influence and content relevance	Users and content	User–content bipartite directed network

2.3.2.6 K-core (k-shell) method

K-core ranking is based on the k-shell decomposition of the network. Each node is assigned the k-shell number, k_s, that is, the order of the shell to which it belongs. In k-shell decomposition, all of the nodes with degree $k = 1$ are initially removed, and pruning processes will continue until no node with $k = 1$ exists. Similarly, the pruning processes will be applied to the next k-shells. This process will continue until the k-core of the network is found (Batagelj & Zaversnik, 2003).

K-shell decomposition methods have been proved effective techniques for identifying the influential spreader in complex networks. The most influential nodes are those that are located within the core of the network, and they can be successfully identified by the k-shell decomposition method (Kitsak et al., 2010). This aspect confirms that the influential spreaders and highly connected users correspond to each other. To overcome the limitations related to the original k-shell decomposition such as considering only the residual degree (i.e., the links between the remaining nodes) and entirely ignoring the exhausted degree (i.e., the links connected to the removed nodes), mixed degree decomposition (MDD) has been introduced (Zeng & Zhang, 2013).

For Weighted network studies (Garas, Schweitzer, & Havlin, 2012) and (Wei, Liu, Wei, Gao, & Deng, 2015) have extended the k-shell decomposition to weighted complex networks; these methods consider both the weight and the degree of a network. However, these methods only consider the weight to be the degree of connected nodes; that is, if the nodes with high degree are connected, then the weight of the edge between these two nodes will be the total degree of these two nodes.

OSNs differ from many complex systems because of the unavailability of complete network data. This unavailability prevents the direct evaluation of the efficiency of user's influence measurements and their comparison with other approaches. Pei *et al*. applied different user influence measurements to identify the influential spreaders in OSNs, and noted that k-core outperformed other approaches, such as PageRank and degree (Pei et al., 2014). K-shell method was modified to measure the user influence in Twitter (Feng, 2011). Logarithmic mapping was applied, in which each k-shell level represented roughly the log value of the analyzed connection count. The difference between the original k-shell and this modified method is that the former decomposes the network in such a way that nodes with a degree equal to or less than k are placed in the k-level. While this modified method decomposes the network in such a way that nodes with a degree $2^k - 1$ or less are placed in the same k-level. Although the author claimed that this modified k-shell method effectively identified a small group of users and was faster than the original method, the result was evaluated only against Twitter usage data (i.e., average tweets/retweets from Twitter usage data). Several studies have shown the plausible circumstances in which influential spreaders do not correspond to Twitter usage data (Cha et al., 2010). This measure may be susceptible to some forms of self-promotion (Feng, 2011).

Recent research (Pei et al., 2014) conducted with large datasets from OSNs has reported that the most influential spreaders are located in the k-core. The k-core method not only calculates the influence of users more effectively than other approaches but also distinguishes most influential spreaders more accurately. studies (Kitsak et al., 2010; Morone & Makse, 2015; Pei et al., 2014) conducted comprehensive comparison between complex networks methods in complex networks such as OSNs. These studies in general, and study reported in (Pei et al., 2014) in specific, concluded. K-core is outperformed other methods such PageRank and degree and found to be more suitable for identifying influential spreaders in OSNs.

2.3.2.7 Machine learning methods

While ranking methods are currently built on known metrics to identify the influential spreaders, the learning approach is based on machine learning algorithms, which uses metrics as features to predict influential spreaders. The most common form of machine learning is supervised learning (Hinton, Osindero, & Teh, 2006) and the most commonly used supervised algorithm are Naïve Bayes, support vector machine, and decision tree. The effective learning approach requires a robust set of features that can provide significant discriminative power for better prediction results. Learning approaches require sufficient training and testing datasets to train and test the machine learning method.

The majority of current studies on predication of influential spreaders (Bigonha, Cardoso, Moro, Gonçalves, & Almeida, 2012; Chai, Xu, Zuo, & Wen, 2013; Cossu, Dugué, & Labatut, 2015; N. Liu, Li, Xu, & Yang, 2014; Mei, Zhong, & Yang, 2015; Xiao, Zhang, Zeng, & Wu, 2013) focus on proposing significant features that can improve the overall predication model rather than proposing any specific algorithms. Predicting the influential spreaders in OSNs has generated controversial discussions on

the specific features that should be selected to effectively predict user influence. Features such as number of followers, number of retweets, number of mentions are basic, and direct features are used in predicting the influential spreaders. However, a study (Mei et al., 2015) shows that aside from these direct features, other effective features can predict the influential spreaders, such as number of public lists, new tweets, and the ratio of followers to friends.

Several features (post-feature based approach) were extracted to train a (SVM). These features are used to calculate user influence through three different means of aggregation, namely, score-based aggregation, list-based aggregation, and SVM-based aggregation (N. Liu et al., 2014). Another approach combines user location in a network with the user's opinion polarity and tweet quality to obtain an aggregated influence score (Bigonha et al., 2012). Moreover, logistic regression analysis was applied to identify the significant features for predicting user influence (Xiao et al., 2013); these features are used for training four machine learning algorithms, and then selecting the most suitable algorithm. The ACQR framework was proposed in (Chai et al., 2013); this framework extracts sets of features that are considered discriminatory attributes to identify efficient spreaders in OSNs. These features are extracted from four aspects, namely, activeness, centrality, quality of post, and reputation. These features are subsequently used to train SVM.

However, a study (Cossu et al., 2015) investigated a large selection of traditional features extracted from OSNs, such as features based on user activity, local topology, stylistic aspects, tweet characteristics, and occurrence-based term weighting. The study concluded that these traditional features did not provide significant results. It proposed a set of new features that demonstrated better performance. However, this study could not be generalized because it did not use comprehensive traditional features from the

previous literature. Furthermore, this study used a specific dataset, and the results were only valid for the considered dataset. The following Table 2.4 compares different features used in training the learning method to identify the influential spreaders in OSNs.

Table 2.4: Comparison of Different Features Used in Training the Learning Model to Identify the Influential Spreaders in OSNs

Studies	Propagation Features				Network Features				User Information Features				Quality Features			Topic Features				Activity Features			
	F 1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F 14	F15	F 16	F 17	F18	F19	F20	F21	F22	F23
(Mei et al., 2015)	✓	✓	×	×	×	×	×	×	✓	✓	✓	×	✓	✓	×	×	×	×	×	✓	✓	×	×
(N. Liu et al., 2014)	✓	×	×	×	×	×	×	×	✓	✓	×	×	×	✓	×	×	×	×	×	✓	✓	×	✓
(Bigonha et al., 2012)	✓	✓	✓	×	✓	✓	✓	✓	✓	✓	×	×	×	✓	✓	×	✓	×	×	✓	×	×	✓

(Xiao et al., 2013)	✓	✓	✓	✓	×	×	×	×	✓	✓	×	×	×	×	×	×	×	✓	✓	✓	×	×	×
(Chai et al., 2013)	✓	✓	×	×	✓	×	×	×	✓	✓	×	×	✓	✓	×	×	×	×	×	✓	✓	✓	✓
(Cossu et al., 2015)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	×	✓	×	✓	×	✓	✓

F1	**Number of reposts (e.g., sharing, retweets) by others**	**F 13**	**Number of (likes or favorites)**
F2	**Number of tags (e.g., tagging or mentioning others)**	**F 14**	**Number of (comments or replies)**
F3	**Hashtag (#)**	**F 15**	**Content quality**

F4	**Shared URL links**	**F16**	**Text feature (TF×IDF or bag or words)**
F5	**In-degree**	**F17**	**Polarity features (positive, negative, or neutral)**
F6	**Betweenness centrality**	**F18**	**Users' topic similarity features**
F7	**Closeness centrality**	**F19**	**Topic distribution**
F8	**Eigenvector centrality**	**F20**	**Number of posts (e.g., status and tweets)**
F9	**Number of friends (or followers) in the list**	**F21**	**Number of reposting others' posts**
F10	**Number of followers**	**F22**	**Number of acknowledging others' posts (e.g., like or favorite other posts)**
F11	**Account information (e.g., official, verified, age of the account)**	**F23**	**Number of interactions with others (e.g., number of comments and replies to other posts)**

2.3.3 Issues of Current Influential Spreader Identification Methods

This section discusses the drawbacks of influential spreader methods. Generally, the major drawback of heuristic methods, such as degree, eigenvector centrality, closeness centrality, PageRank, and k-core, is that they do not enhance the global function of influence. Hence, an assurance of their results is lacking (Morone & Makse, 2015). A recent study (Morone & Makse, 2015) showed that the set of vital influential spreaders is considerably smaller than that detected by heuristic methods. Remarkably, numerous previously ignored weakly connected nodes appear among vital influential users. These nodes are identified as low-degree nodes based on the topological analysis of the network, surrounded by hierarchical coronas of hubs, and are exposed only through the optimal collective exchange of all of the influential users in the network (Morone & Makse, 2015). However, the supervised learning approaches of influential user identification suffer from their dependency on the training data. Obtaining training data on "which users are influential" is difficult, expensive, and time consuming.

Moreover, betweenness centrality, k-core, and PageRank are global measures because their measurement can be implemented with the entire network as opposed to the local degree. Extracting the complete network of most OSNs is difficult because of ethical and technical reasons, which causes difficulty in claiming the most influential spreaders based on those measurements. The subsequent section discusses the drawbacks and open issues of applying these measurements to OSNs to identify the influential spreaders or users.

2.3.3.1 Degree centrality

Degree centrality is commonly in complex networks, the most connected nodes are generally considered to be in authority for the largest information dissemination and are viewed as the most influential spreaders or nodes (Albert et al., 2000). A limitation of

this method is that hubs may form tightly-knit groups called "rich-clubs" (Colizza, Flammini, Serrano, & Vespignani, 2006). Strategies based on degree measures will highly rank these rich-club hubs (Morone & Makse, 2015). Contrary to usual assumptions, reasonable circumstances exist in which the influential spreaders do not correspond to the most highly connected users (Kitsak et al., 2010). Furthermore, research has reported the invalidity of degree measures to identify influential spreaders (Pei et al., 2014) because the degree measure only reflects the quantity of the adjacent neighbors of a user. A hub that is located on the edge of a network may have a trivial effect on the influential spreading process because its neighbors are restricted in spreading capability. A weakly connected user who is strategically located in the core of the network will have an important effect that induces diffusion through a huge portion of the users.

2.3.3.2 Closeness centrality

Closeness centrality is likely to give a high ranking to individuals who are close the center of local clusters; therefore, it over-allocates spreaders next to each other (Morone & Makse, 2015). More importantly, closeness centrality has high computational complexity; hence, it is unsuitable to be applied into significantly large-scale OSNs (D. Chen, Lü, Shang, Zhang, & Zhou, 2012).

2.3.3.3 Betweenness centrality

Betweenness centrality is a popular technique in complex network analysis, particularly in community detection. However, this technique suffers from a high computational time that stops the analysis in/of large-scale networks. The best algorithm for betweenness centrality requires computational time equal to $O(NM)$ for unweighted networks with N nodes and M edges (Morone & Makse, 2015). This limitation has rendered betweenness centrality impractical for large OSNs.

2.3.3.4 Eigenvector centrality

This method is inefficient, particularly in scale-free networks, because the weight is assigned to a few nodes (hub), whereas the remaining majority have considerably small weights; therefore, they will not be ranked accurately (Morone & Makse, 2015). However, the degree distribution for networks such as the Internet (Barabási & Albert, 1999), e-mail (Ebel, Mielsch, & Bornholdt, 2002), and Facebook (Catanese et al., 2012) are proved to be scale-free networks. Consequently, eigenvector centrality may result in improper ranking if applied to such networks.

2.3.3.5 PageRank-like methods

PageRank and PageRank-like measurement are network-based diffusion algorithms and are computed by random walks on the network graph. The algorithm's desirability is attributed to its known efficiency to rank web pages; and it is easy to comprehend its concept (Ghoshal & Barabási, 2011).

Thus, the shortcoming of PageRank is attributed to its consideration for the node's score when computing others' scores. In other words, a user with a high PageRank may confer a significantly higher score to otherwise weakly influential spreaders to whom he/she links.

The complete OSN structure is unavailable due to the inherent limitations of OSNs caused by API restrictions and user privacy. Thus, the PageRank algorithm is an unreliable measurement for OSNs. The study (Ghoshal & Barabási, 2011) reported that for random networks, the measurements given by PageRank are responsive to perturbations in network topology, rendering it unreliable for incomplete or noisy networks.

PageRank is commonly used in ranking web pages, and many circumstances exist in which it is unsuccessful in detecting influential spreaders in real-world social networks. A study (Pei et al., 2014) showed that PageRank is unreliable in identifying influential spreaders in OSNs. PageRank is frequently used to identify influential spreaders based on the assumption of the random diffusion of information in the network. However, in reality, the diffusion of information processes is not totally based on random walks (Goel, Watts, & Goldstein, 2012). This aspect could induce a substantial divergence between the PageRank outcomes and the actual outcomes.

2.3.3.6 K-core (k-shell) method

Even though, studies in (Kitsak et al., 2010) (Pei et al., 2014) (Morone & Makse, 2015) reported that K-core method is effective for identifying influential spreaders in OSNs as compared with degree centrality, PageRank. The major limitation of the k-core method is that it is proposed to deal with unweighted networks. Nevertheless, most real networks are weighted, and their weights describe significant properties of the underlying systems. Attempts to eliminate this limitation have been studied in (Garas et al., 2012; Wei et al., 2015) , these studies proposed weighted k-core , the proposed edge weights are based on the degree of nodes only, if a node is connected to high degree nodes then this node will receive high weight. Conversely, the degree of the nodes as weight in OSN context does not always provide accurate influence of the users (Cha et al., 2010). In other hand, the observation that interaction among users is a substantial factors in quantifying the diffusion ability of a user in OSNs (L. Weng et al., 2012). Interaction among users need to be integrated with K-core to combine both the capability of K-core with most important features of OSNs for quantifying the diffusion ability in order to propose an effective method for influential spreaders identifications.

Furthermore, various numerical simulations have been implemented (Ying Liu et al., 2015) to understand the relationship between the influential nodes identified by the k-core method and their influence in real networks. These implementations have spurred the realization that not all nodes with high shells are highly influential. Two sets of core nodes exist, namely, the true influential nodes whose shell level correctly reflects their influence in real networks, and nodes with high shell levels but are not influential spreaders (i.e., core-like group) (Ying Liu et al., 2015). Understanding these observations will enhance the understanding of the real network structure and influential nodes, as well as improve the k-core method by removing redundant connections (weak links with weak interaction strength) that cause core-like group issues.

2.3.3.7 learning methods

Supervised learning approaches for identifying the influential spreaders have the drawback of their dependency on the training data. Creating training data (i.e., labeled samples) is more difficult, expensive, and time consuming. A robust learning approach for detecting influential user requires large amounts of labeled training data to effectively learn the different classes of models. To alleviate this drawback, semi-supervised approaches might be used with only a small amount of labeled data. However, in some applications, labeled samples are very limited, and obtaining effective knowledge from such small samples is inaccurate in most cases (Bouguessa, 2011). The scarcity of trained data is one of most challenging tasks that face current supervised machine learning approaches. A drawback of most machine learning-based studies to predict user influence is the evaluation of the proposed machine learning method. This situation is attributed to the absence of a ground truth on which supervised learning can be performed (Räbiger & Spiliopoulou, 2015). This fact has spurred a deep argument in the previous research, with a number of studies claiming different results on the best algorithms and sets of features that provide the best prediction of

users' influence according to their particular dataset and features used to train their model.

2.3.4 Comparison between Influential Spreaders methods

According to above reviewed influential spreaders identification methods, in this section the abovementioned methods are compared based on their advantages, disadvantages.

Table 2.5: Comparison between Influential Spreaders Identification Methods

methods	Advantages	Disadvantages
Degree centrality	• Simple assumptions	• It only measures the local features of users (number of direct friends)
Closeness centrality	• Perform on part of the network nodes and result in a global impact	• It suffers from a high computational time. Therefore, it is not applicable for most of large OSNs.
Betweenness centrality	• Reveal shortest paths between all pairs of nodes. • Global measure	• It suffers from a high computational time. Therefore, it is not applicable for most of large OSNs.
Eigenvector centrality	• It is simple and effective for a network where degree is biased in such way the *node is important if it is linked to other important nodes.*	• Eigenvector centrality may result in improper ranking if applied to OSNs as the degree distribution for most OSNs such as Facebook are scale-free networks (Catanese et al., 2012).
PR methods	• Simple assumptions • Direct implementation • Comparatively low computational complexity • Global measure	• For random networks, the measurements given by PageRank are responsive to perturbations in network topology, rendering it imprecise for incomplete or noisy networks (Ghoshal & Barabási, 2011) • Unreliable in identifying influential users in OSNs(Pei et al., 2014) • It is based on the assumption of the random diffusion of information in the network. However, in reality, the diffusion of information processes is not totally based on random walks (Goel et al., 2012)

K-core method	• Simple assumptions • Global measure • In OSNs with incomplete data, the k-core algorithm calculates the influence of users more effective than other approaches (Pei et al., 2014).	• Designed for unweighted network • The output of k-core has two sets of core nodes, namely, the true influential nodes whose shell level correctly reflects their influence in real networks, and nodes with high shell levels but are not influential spreaders (i.e., core-like group) (Ying Liu et al., 2015)
Machine learning methods	• It can predicate the influential users based in users characteristics.	• It requires sufficient training data • It is difficult to extract global features to train the learning model, therefore in most of the studies the learning model is trained based on local features of the users

2.3.5 Performance Evaluations of the Identification of Influential Spreaders

Influence diffusion can be modeled in probabilistic frameworks (Cosley, Huttenlocher, Kleinberg, Lan, & Suri, 2010). However, OSNs have inherent properties that differentiate them from traditional social networks. These properties are challenging tasks for developing a model that can efficiently illustrate the influence spread in the OSN context. Different influence models have been discussed in (AlFalahi, Atif, & Abraham, 2014; Sun & Tang, 2011). Most studies that analyzed the spread of information in the structure of social networks reported that the spread of information has full equivalence with the spread of infectious diseases (Leskovec, Adamic, & Huberman, 2007; Watts, Peretti, & Frumin, 2007). This has idea resulted in the intensive implementation of classical disease models, such as the susceptible-infectious-recovered (SIR) and susceptible-infectious-susceptible (SIS) models in information diffusion studies (Pei, Muchnik, Tang, Zheng, & Makse, 2015a). These models, inspired by the spread of contagious diseases (Hethcote, 2000), are proposed based on the basic belief of human behavior, which might not be representative and illustrative of the real dynamics of information diffusion (Pei et al., 2014; Pei et al., 2015a). Therefore, these models may provide a deceptive output in terms of which measure is better. For example, (Pei et al., 2014; Pei et al., 2015a) demonstrate that in the implementation of the SIR and SIS models in real-world networks, k-core shows

better results than other measures, such as degree and betweenness centrality in (Kitsak et al., 2010). By contrast, (Borge-Holthoefer & Moreno, 2012) used the rumor dynamics model and suggested that k-core was invalid because of the absence of influential spreaders. Studies have also reported that the measurement that is based on models is unsuitable in practice (Goldenberg, Libai, & Muller, 2001; Singh, Sreenivasan, Szymanski, & Korniss, 2013). Moreover, the spread of diseases and spread of information are different in practice (Centola & Macy, 2007; Singh et al., 2013). Therefore, to eliminate the dependence of the identification of influential spreaders based on the particular model used to simulate the dynamics, A study used the real dynamics of information diffusion in real-world social networks (Pei et al., 2014). Another study introduced coverage to verify the effectiveness of the proposed method; coverage considered the link structure and the time interval to identify the spreadability of a node within a specified period. Kendall's Tau algorithm or Spearman's rank are commonly used to quantify the correlation between ranking lists obtained by artificial stochastic models.

Another evaluation approach is to compare the results from various ranking lists obtained by different identification algorithms with the ranking lists obtained from manual annotation ranking lists. Analysts are asked to categorize the users into influential and non-influential spreaders via manual processes or real recommendation systems. Accuracy, F-measure, precision, and area under curve (AUC) metrics are subsequently used to compare performances of these algorithms with manually labeled data (Chai et al., 2013; Cossu et al., 2015; Xiao et al., 2013). The drawback of this evaluation approach is that it requires human intelligence to label the user into influential spreaders or non-influential spreaders, which is time consuming and expensive. Furthermore, humans can only judge the influence of a user based on static information on user features, and they cannot judge based on a deep analysis of the user

position in the entire network and how far the user posts spread in the entire network. Consequently, this evaluation is based on the local rather than the global features of user influence.

Some studies evaluated their proposed algorithm by comparing user rankings obtained by different identification algorithms with their OSN characteristics, such as in-degree number, volume of propagated content, number of replies that their posts have, and volume of content that the users post. The algorithm is subsequently claimed to be the best if the ranking list that it generated is highly correlated with user characteristics (Feng, 2011). This evaluation approach is simple and straightforward; however, the majority of studies have shown that the influential spreaders do not always highly correlate with their OSN characteristics (Cha et al., 2010; Morone & Makse, 2015; Räbiger & Spiliopoulou, 2015; Xiao et al., 2013). Hence, this evaluation is not always applicable. Furthermore, this evaluation can be biased to the authors' preferences in selecting user characteristics to evaluate their proposed algorithm.

2.3.6 Summary and Taxonomy of the Identification of Influential Spreaders Researches in the OSN Context

Figure 2.5 shows the thematic taxonomy of the identification of the influential users in the OSNs. The studies on the identification of influential users in the OSNs are categorized based on five characteristics, namely, objectives, identification algorithms, metrics used, type of networks, and evaluation models. These parameters are used to compare different studies as shown in Table 2.6.

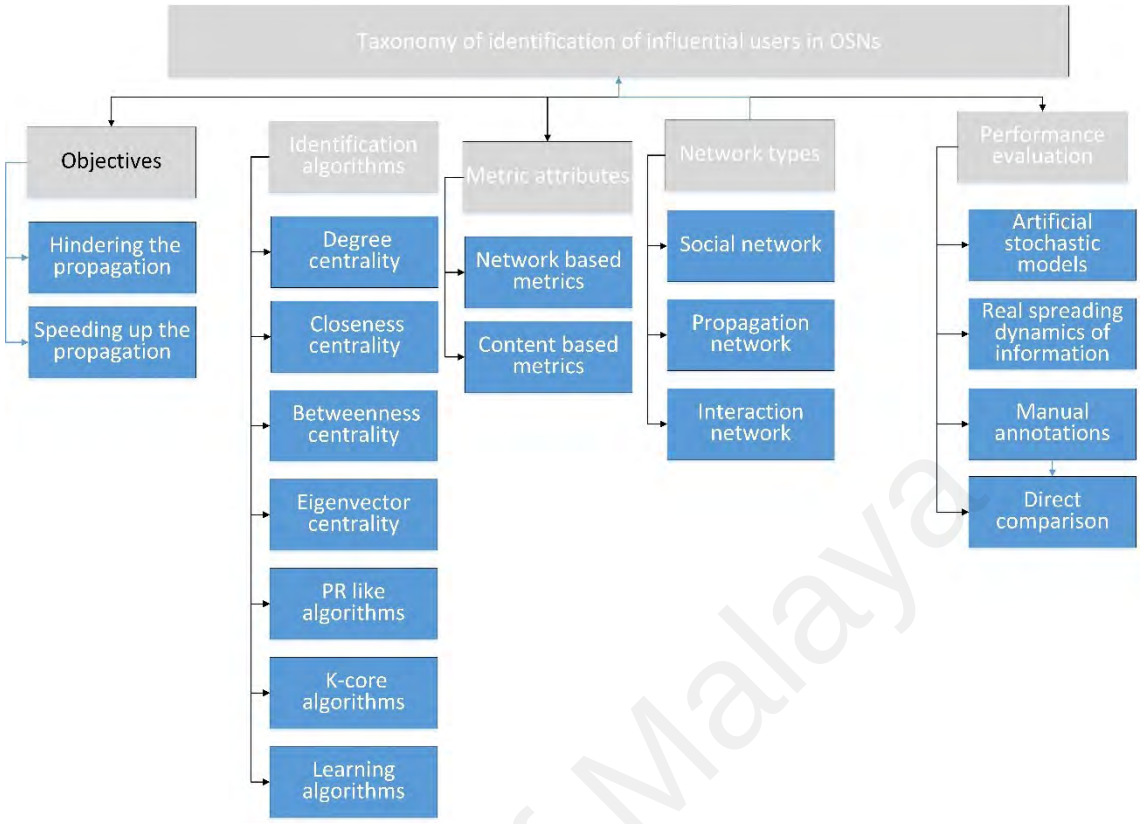

Figure 2.5: Taxonomy of the Identification of Influential Spreaders studies in OSNs.

A. Objectives

The OSN can be used as platform for spreading positive or negative social activities. The objective of identifying the influential users can be classified into accelerating the propagation of information and hindering the propagation of information.

(1) Hindering the spread of information in the OSNs: Blocking the spread of the unwanted contents, such as cyberbullying, rumors, viruses, and spam, to the influential users is one of the strategies for restraining the spread of an unwanted content (Gao et al., 2011; Wen et al., 2014b; Z.-K. Zhang et al., 2016) .

(2) Accelerating the spread of information in the OSNs: Identifying and targeting the influential users is significant to enhance the spread of specific information within the OSNs. The objective has several applications, for instance prevention strategies such as

cyberbullying prevention by spreading kind words and awareness (Patchin & Hinduja, 2013; Z.-K. Zhang et al., 2016) .

B. Identification methods

The OSNs have created a massive communication and social interaction among the users. In the recent years, the OSNs have attracted millions of the users. Consequently, the OSNs have become the large networks that contain millions of the nodes and links. Some algorithms such as the greedy algorithms are found to be accurate for identifying the influential users in the small networks. However, the greedy algorithms are unsuitable to be applied in the networks with the large numbers of nodes and links, such as OSNs, due to the inherent limitations of the greedy algorithms. The limitations include the inefficiency due to a high computational time (H. Li, Bhowmick, Sun, & Cui, 2015). The state-of-the-art method s for identifying the influential spreaders in the OSN context include degree, closeness centrality, betweenness centrality, eigenvector centrality, PageRank-like methos, k-core method, and learning methods as discussed in above sections.

C. Metric attributes

The metrics extracted for the OSNs are important identification parameters for the influential user techniques. The different metrics used in the methods generates varied ranking results (Cha et al., 2010). Therefore, understanding the correlation of the metrics with the influential users is important. classified metrics extracted from the OSNs are classified into network-based and post-based metrics.

The *Network based metrics* deal with the question of how users are connected with one another and describe the interaction network between the users in the network. The metrics are related to the structure of the network. For example, the in-degree metric

directly indicates the size of the audience for a user. The propagation metric describes that how the information propagates through the network in the OSNs and indicates the ability of a user in an OSN to generate and propagate the content throughout the network. In addition, the engagement metric indicates the ability of a user to involve others in a conversation within the network. For example, the in-degree metrics in the Facebook and Twitter are the number of friends and the number of followers, respectively. The propagation metrics in the Facebook and Twitter are measured through the sharing process and the retweeting process, respectively. The engagement metrics are expressed in the Facebook and Twitter by tagging and by mentioning other users, respectively. Furthermore, the output of applying the algorithm on the OSNs, such as degree centrality, betweenness centrality, closeness centrality, eigenvector centrality, and PageRank, can be used as a network metric input to another identification algorithm such as the machine-learning algorithm (Bigonha et al., 2012; Chai, Xu, Zuo, & Wen; Cossu et al., 2015).

The *Post-based metrics* deal with the quality of the user's post. The post-based metrics focus on the content features that make a user's posts viral. Considering the content similarity between the users may improve the ranking of an influential user. The content metrics, such as content similarity between the users and content quality of a post, are combined with the network-based metrics to improve the identification of the influential users (W. Chen et al., 2012) however this textual-based metrics are not always a good option since the diversity of text related to spreaders in OSNs are varied.

D. Network types

The explicit and implicit OSN connections induce connection diversity. The different network connections are created within the OSNs that describe various relational connections between the users. Examples are social networks that describe the social relationship between the users, such as following the relationships in the Twitter or friendships in the Facebook. Moreover, the propagation networks that describe the diffusion networks between the users, for example, retweet networks in the Twitter or shared networks in the Facebook. Therefore, applying an identification algorithm on the different OSNs provides different rankings of the users according to the type of a network constructed. The most common network types with state-of-the-art features can be classified as the social networks, propagation networks, and engagement networks,

The *Social networks* in the OSNs describe the social connections between the users within the OSNs. For example, in Facebook, if a user A is a friend of another user B, then a social connection exists between them. Similarly, in Twitter, if the user A follows the user B, then a direct social connection exists between them.

The *Propagation networks* describe how a piece of information propagates from one user to another. For example, if a user A shares or retweets a user B's post, then the post is propagated from the user B to the user A. and a propagating or diffusing connection is created.

The *Engagement networks* describe the ability of a user to involve the others in a conversation. An engagement network is constructed if a user A tags or mention a user B, then an engaging connection is created from the user A to the user B. An interaction network can also be constructed if the user A replies to the post of the user B.

E. Performance evaluation

The straightforward evaluation of the effectiveness of the influential user identification algorithms is not possible due to the unavailability of the full diffusion information in the OSNs. The unavailability of the data is due to the technical and privacy issues set by the users. The evaluation approaches adopted in the literature are summarized into the following four categories:

1) Evaluation through artificial stochastic models: The artificial stochastic models include the susceptible-infectious-recovered (SIR) model and susceptible-infectious-susceptible (SIS) model (Pei et al., 2015a), rumor dynamics model (Borge-Holthoefer & Moreno, 2012), linear threshold model, and independent cascade model (AlFalahi et al., 2014).

2) Evaluation through real spreading dynamics: In this evaluation approach, the real diffusion of information is tracked to obtain the ranking list based on the real spread dynamic of the information. The ranking list obtained by the different identification algorithms is correlated with the list on the real spread dynamic of the information (Ding et al., 2013; Pei et al., 2014). A higher correlation corresponds to an effective algorithm.

3) Evaluation by manual annotations: The results from the different ranking lists obtained by the various identification algorithms are compared with the manual annotation ranking lists. The measures, such as Accuracy, F-measure, precision, and AUC metrics are used to compare the performances of the algorithms with manually labeled data (Chai et al., 2013; Cossu et al., 2015; Xiao et al., 2013).

4) Evaluation by direct comparison: The user rankings obtained by the different identification algorithms are compared with the OSN characteristics, such as in-degree

number, volume of propagated content, number of replies that their posts have, and volume of content posted by users. The algorithm is claimed to be the best if the algorithm's generated ranking list is highly correlated with the user characteristics (Feng, 2011).

Table 2.6: Comparison Summary of Influential Spreaders Identification Researches In the OSNs context.

Comparison parameters	Identification method							Metrics Attributes		Network Type			Performance Evaluation			
Reference	Degree centrality	Closeness centrality	Betweenness	Eigenvector	PR-like methods	K-core method	Learning methods	Network	Post	Social network	Propagation	Engagement	Artificial	Real spreading dynamics	Manual annotations	Direct comparison
(Cha et al., 2010)	✓	×	×	×	×	×	×	✓	✓	✓	✓	✓	×	×	×	✓
(Kim & Han, 2009)	✓	×	×	×	×	×	×	✓	✓	✓	×	×	✓	×	×	×
(Catanese et al., 2012)	✓	×	✓	×	✓	×	×	✓	×	✓	×	×	×	×	×	×
(Tunkelang, 2009)	×	×	×	×	✓	×	×	✓	×	✓	×	×	×	×	×	×
(Yamaguchi et al., 2010)	×	×	×	×	✓	×	×	✓	×	✓	✓	✓	×	×	✓	×
(J. Weng et al., 2010)	×	×	×	×	✓	×	×	✓	✓	✓	×	×	×	×	✓	×
(Lü et al., 2011)	×	×	×	×	✓	×	×	✓	×	✓	×	×	✓	×	×	×
(Q. Li et al., 2014)	×	×	×	×	✓	×	×	✓	×	✓	×	×	✓	×	×	×
(W. Chen et al., 2012)	×	×	×	×	✓	×	×	✓	✓	✓	×	×	×	×	×	✓
(Jabeur et al., 2012)	×	×	×	×	✓	×	×	✓	×	×	✓	×	×	×	✓	×
(Ding et al., 2013)	×	×	×	×	✓	×	×	✓	×	×	✓	×	×	✓	×	×
(Silva et al., 2013)	×	×	×	×	✓	×	✓	✓	✓	×	×	×	×	×	✓	×
(Pei et al., 2014)	✓	×	×	×	✓	✓	×	✓	×	×	×	✓	×	✓	×	×
(Feng, 2011)	×	×	×	×	×	✓	×	✓	×	✓	×	×	×	×	×	✓

(Mei et al., 2015)	×	×	×	×	×	×	✓	✓	✓	×	×	×	×	×	×	✓
(N. Liu et al., 2014)	×	×	×	×	×	×	✓	✓	✓	×	×	×	×	×	✓	×
(Bigonha et al., 2012)	✓	✓	✓	✓	×	×	✓	✓	✓	✓	✓	✓	×	×	✓	×
(Xiao et al., 2013)	×	×	×	×	×	×	✓	✓	✓	×	×	×	×	×	✓	×
(Chai et al., 2013)	✓	×	×	×	✓	×	✓	✓	✓	✓	×	×	×	×	✓	×
(Cossu et al., 2015)	✓	✓	✓	✓	×	×	✓	✓	✓	✓	×	×	×	×	✓	×

2.4 Conclusion

A cyberbully can harass his/her victims before an entire online community. Online social media, such as social networking sites (e.g., Facebook and Twitter) have become integral components of a user's life. Increase in cyberbullying occurrences is commonly attributed to the fact that traditional bullying is more difficult to practice than cyberbullying, in which perpetrators bully their victims without direct confrontation by using a laptop or a cellphone connected to the Internet. The characteristics of OSNs have also expanded the reach of cyberbullies to previously unreachable locations and countries. Most of the reviewed studies applied machine learning algorithms to construct cyberbullying detection methods. However, applying machine learning may prove successful or unsuccessful in predicting cyberbullying because building a successful machine learning model depends on many factors. The most important of these factors are the features used and the presence of independent features in the model that correlate well with the class. Selecting the best features with high discriminative power between cyberbullying and non-cyberbullying tweets is a complex task that requires considerable effort. To construct a cyberbullying detection method, discriminative features that can be used in machine learning schemes should be identified to distinguish cyberbullying tweets from non-cyberbullying ones. A set of comprehensive features should be proposed to enhance the discriminative power of classifiers. This set will be a key novel contribution to the literature. By identifying the most significant features and using them as inputs to different machine learning classification algorithms, cyberbullying can be detected with high accuracy. A cyberbullying detection method with extensive detection coverage and a substantially accurate and effective cyberbullying detection method must be proposed to avoid inconvenience from normal posts. For example, raw extracted tweets contain substantial information that needs to be utilized to achieve considerably accurate and effective

cyberbullying detection. Simultaneously, rigorous processes, from extracting representative data to evaluating the model, should be maintained to ensure the effectiveness of any proposed method.

However proposing a cybercrime detection method is an ineffective solution alone. An effective method for identifying influential spreaders in networks to be immunized is needed. The reviewed papers in literature concluded that k-core method is considered more suitable for identifying influential spreaders in OSNs than other algorithms in the current literature review. Generally, the major limitation of k-core method is that it deals with unweighted graphs. Nevertheless, most real networks are weighted, and their weights describe significant properties of underlying systems. To overcome the original k-core method issues related to treating all links equally, the number of connected links to the nodes should be considered rather than the quality of links among the nodes. The effectiveness of k-core method should be investigated after considering the interaction between users while identifying the influential spreaders.

CHAPTER 3: RESEARCH METHODOLOGY

3.1 Introduction

This chapter discusses the general methodology used for developing proposed methods for detecting cyberbullying and identifying influential spreaders in OSNs.

Understanding complex systems such as OSNs involves interdisciplinary aspects (content and network) that work together to achieve full functionality. The explosive evolution of OSNs has enhanced the ways cybercrimes are committed and spread; OSNs provide tools to commit cybercrimes and largely connected networks of users to propagate them. On the upside, OSNs offer noteworthy data on human behavior and interaction that can be analyzed by researchers to develop methods for effectively detecting and tracking cybercrimes. This chapter introduces the general methodology used for developing the proposed methods. Content level deals with information that can be directly extracted locally from OSNs post and information. Network level deals with network representation by extracting the relationship between users to analyze the information spreading in the network globally. Figure 3.1 explains the general methodology. In the succeeding sections, the content analysis methodology stages for cyberbullying detection and network analysis methodology stages for the identification of influential spreaders are presented. Specific details for each method and their contribution are comprehensively explained in Chapters 4 and 5, respectively.

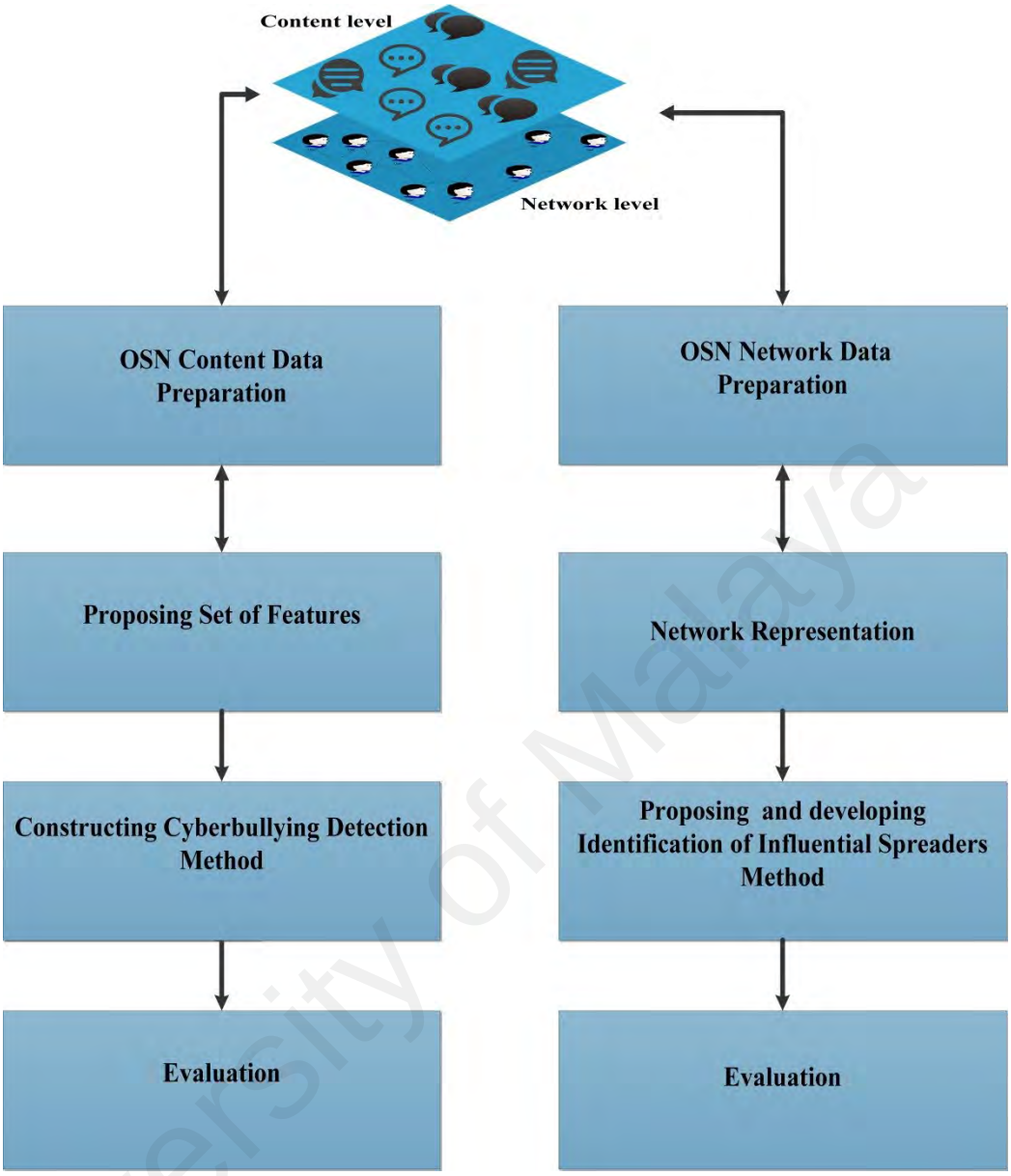

Figure 3.1: Stages of research methodology

Figure 3.1 elaborates the methodology stages for cyberbullying detection and identification of influential spreaders. These stages are explained in the succeeding sections.

3.2 Cyberbullying Detection in OSNs

The characteristics of OSNs also expand the range of cyberbullies to previously unreachable locations and countries. This study aims to improve cyberbullying

detection performance to realize an effective cyberbullying detection method. The methodology used for cyberbullying detection is divided into the following stages.

3.2.1 OSNs Content Data Preparation

3.2.1.1 Data Collection

In this research, the data were collected from Twitter from January 2015 to February 2015. The collected data set contains 2.5 million tweets. a study showed that Twitter is turning into a "cyberbullying playground" (Xu et al., 2012). Consequently, this research collected data from Twitter. This research adopts geo-code filtering to extract the data from twitter. The reason of selection geo-code filtering can be explained as follows: the public streaming API of Twitter only provides access to a small sample of relevant data in a few instances, thereby introducing a potential sampling bias (Cheng & Wicks, 2014; González-Bailón et al., 2014; Y. Liu et al., 2014). Morstatter et al.(2013) analyzed whether the data extracted from this API could sufficiently represent the activities in the Twitter in general. They determined that when geo-code filtering was used, API would return a nearly complete set of geo-tagged tweets despite the geo-coded sampling (Morstatter et al., 2013). By contrast, the API would return a data set with certain bias if keyword (i.e., words, phrases, or hashtags) or user ID sampling was adopted. Alternatively, geo-tagged filtering can be applied for collecting data to achieve a favorable representation of the activities in Twitter. Researchers who adopt geo-tagged filtering are confident that they are working with an almost complete sample of Twitter data (Morstatter et al., 2013). Therefore, the present research adopts geo-code filtering to minimize bias in data collection.

Thereafter, the data were preprocessed. The tweets were converted to lowercase, www.* orhttp://* was converted to URL, @username was converted to AT_USER, additional white spaces were removed, two or more repetitions of characters were

replaced by the same character, and misspelled words were corrected using a spelling corrector.

3.2.1.2 Manual Data Set Annotation

In the present study, the tweets are labeled with the assistance of three experts using the definition of cyberbullying mentioned in literature chapter as basis. The tweets are classified as follows:

- Cyberbullying: the tweet content indicates the presence of cyberbullying behavior.
- Non-cyberbullying: the tweet content does not indicate the presence of cyberbullying behavior.

The tweets are considered cyberbullying if at least two of the assigned experts regard them as such. If these experts do not agree on the classification of a tweet, the tweet will be deleted from the data set.

Experts are used to code the tweets rather than coding through Mechanical Turk website (MTurk) to improve the quality of labeling. Using experts also avoids online spam tracker, which simply classifies the tweets without actually reading them (Ipeirotis, 2010). The drawback is that using experts is more time consuming than using MTurk.

The tweets in the present study were labeled with the assistance of three experts who used the above-mentioned definition as basis for the process. These experts were oriented about the abbreviations, slang words, and acronyms commonly used in social networks and online communications to assist them in further understanding the tweet contents. 10606 tweets are manually classified. These tweets are classified into 10,007 non-cyberbullying and 599 cyberbullying ones.

3.2.2 Proposing Set of Features

In the present research, comprehensive features related to cyberbullying behavior is proposed based on network information, activity information, user information, and tweet content{Lee, 2015 #2412}. feature engineering is an important factor to provide effective detection methods (Domingos, 2012). Proposing a set of significant features is the main step toward constructing effective classifier in many applications (Libbrecht & Noble, 2015). In this study as a key novel contribution to the literature, is set of features related to cyberbullying, which are extracted and used to construct cyberbullying detection method with high performance accuracy. These features, which are comprehensively explored in Chapter 4 (Section 4.2), are used in conjunction with supervised machine learning algorithms to create a cyberbullying detection method.

3.2.3 Construction of Cyberbullying Detection Method

In this stage, the different algorithms, which are utilized to construct cyberbullying detection method using the proposed features, are selected.

3.2.3.1 Machine Learning Algorithms

In this stage, a machine learning algorithm is selected to be trained on the proposed features. However, deciding which classifier performs best for a specific data set is difficult. In the present research, more than one machine learning classifiers are used. Three points are used as guide to narrow the selection of machine learning algorithm to be used. First, a specific literature on machine learning for cyberbullying detection is important to select specified classifier. The preeminence of a classifier may be circumscribed to a given domain (Macià, Bernadó-Mansilla, Orriols-Puig, & Ho, 2013) . Therefore, the literature review in Chapter 2 is used as a guide to select the machine learning algorithm. Second, literature review in text mining (Korde & Mahender, 2012; Sebastiani, 2002) is also employed as a guide. The performance comparison on

comprehensive data set (Fernández-Delgado, Cernadas, Barro, & Amorim, 2014) is also used as basis to select machine learning algorithms. The machine learning algorithms selected are NB, SVM (LIBSVM), RF, and KNN. NB, SVM (LIBSVM), RF, and KNN are tested using WEKA (M. Hall et al., 2009). Detailed descriptions of these methods are presented in Chapter 2.

3.2.3.2 Feature Selection Algorithms

In this research three feature selection algorithms are to selected, namely, chi-square test, information gain, and Pearson correlation, to determine the discriminative power of each feature (Y. Yang & Pedersen). As discussed in chapter 2, these three features selection algorithms are the prominent feature selection algorithms and mostly widely used for text classification. Detailed descriptions of these algorithms are presented in Chapter 2. Feature selection is performed to select the set of features from all proposed features to be used as input to the classifiers.

3.2.3.3 Handling of Imbalanced Class Distribution (SMOTE and Cost sensitive techniques)

In real-world applications, data sets often contain imbalanced data in which the normal class forms the majority and the abnormal class forms the minority. Examples of imbalanced data are fraud detection, instruction detection, and medical diagnosis. The number of cyberbullying tweets is expected to be much less than non-cyberbullying tweets, and this assumption will generate imbalanced class distribution, in which the data set of non-cyberbullying contains much more tweets than that of cyberbullying. Such imbalanced class distribution can prevent the model from accurately classifying the instances. Machine learning algorithms with imbalanced class distribution tend to be overwhelmed by the major class and ignore the minor one. Several approaches for overcoming this issue have been proposed, such as a combination of oversampling the

minority (abnormal) class and undersampling the majority (normal) class (Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, & W Philip Kegelmeyer, 2002) as well as weight adjusting (X.-Y. Liu & Zhou, 2006). Both approaches are employed in the present study.

SMOTE technique (Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, & W Philip Kegelmeyer, 2002) is applied to avoid overfitting which happens when particular replicas of minority classes are added to the main dataset. A subdivision of data is reserved from the minority class as an example and then new synthetic similar classes are generated. These synthetic classes are then added to the original dataset. The created dataset is used to train the machine learning methods.

Cost sensitive technique {Chawla, 2009 #2411} for controlling the imbalance class. Cost sensitive is based on creating is cost-matrix, which defines the costs experienced in false positives and false negatives. Over-sampling of the minority (abnormal) class and under-sampling of the majority (normal) class (SMOTE) are used along with weights adjusting approaches (cost-sensitive) to handle the imbalanced class distribution in our manually labeled data set. The outputs of these techniques are fed to the machine learning algorithms.

3.2.4 Evaluation

This research uses AUC, an area under the receiver operating characteristic (ROC) curve as the main performance measure. Other performance measures, such as precision, recall, and F-measure, are used as additional reference measures. AUC possesses a significant statistical property. The AUC of a classifier is also equal to the

probability that the classifier will rank a randomly selected positive instance higher than a randomly selected negative instance (Fawcett, 2006). AUC is commonly used in medical decision making. In recent years, AUC has been used intensively in machine learning and data-mining studies, where imbalanced class distribution exists (Fawcett, 2006).Considering that cyberbullying tweets is much less than non-cyberbullying tweets, manually labeled data usually contains an imbalanced class distribution; therefore, the selection of an evaluation metric is important. As explained in Chapter 2 (2.2.2.3; Issues Related to Evaluation Metric Selection), evaluation metric should be carefully selected; otherwise, this may lead to misleading evaluation metric. This issue is well known in machine learning with imbalanced class distribution. The key advantage of AUC is that it is more robust than accuracy, precision, recall, and f-measure in class imbalance situations. Given a 95% imbalance (e.g., in favor of the positive class), the accuracy of the default classifier that consistently issues "positive" will be 95%, whereas a considerably interesting classifier that actually deals with the issue is likely to obtain a worse score. The ROC curve denotes the rate of TP versus FP at different threshold settings. The area under the curve provides a signal of the discriminatory rate of the classifier at various operating points (Fawcett, 2004, 2006; Prieto et al., 2014; Provost & Fawcett, 1997). Therefore, in this research AUC is selected as the main performance measure because of its high robustness for evaluating machine-learning classifiers.

This research also uses 10-fold cross-validation. In applications where obtaining training and testing data are difficult, such as cyberbullying detection, most state-of-art studies used single manually labeled data set because creating labeled data is expensive and sharing the data set containing user contents is not allowed by OSN providers. These issues can be resolved with cross-validation, that is, randomly dividing the training data into, for example, ten subsets, and this process is called 10-fold cross-

validation. Cross-validation involves the following steps: keep a fold separate (the model does not see it), and train data on the model using the remaining folds; then test each learned classifier on the fold which it did not see; and average the results to see how well the particular parameter setting performs (Domingos, 2012; Kohavi, 1995). All experiments in this research will be based on 10-fold cross-validation.

3.3 Influential Spreaders Identification in OSNs

Identifying influential spreaders holds practical significance, and it has attracted much attention (Kitsak et al., 2010; Pei et al., 2014). Targeting these influential spreaders is significant for either speeding up the propagation of useful information or hindering the diffusion of unwanted content, such as preventing the spread cyberbullying , virus, online negative behavior, and rumors (Kwon et al., 2013; L. Zhao et al., 2011).

The methodology that is used in identifying influential spreaders is divided into the following stages.

3.3.1 OSNs Network Data Preparation

In this research, Twitter networks are used to verify the effectiveness of the proposed influential spreader identification methods compared to existing methods. Twitter permits millions of users to broadcast short messages through social connections (L. Weng, 2014). This OSN is a highly popular service, which provides a natural situation for studying diffusion processes. Different from other OSNs, Twitter is particularly dedicated to spreading information in that users follow the information broadcasted by other users (S. Wu, 2013); thus, the network of information spreading can be reconstructed by crawling the corresponding followers' network. In contrast to the collection of tweet content, Twitter data policy allows researchers to share anonymized network data set, which contains user relationships. Consequently publicly available

networks are used to verify the effectiveness of the proposed method compared with existing methods.

To check the performance of the improved method, two large real online social networks from Twitter are used. These networks are (1) a directed twitter network used in [44] and (2) a reciprocal following relationship network from Twitter used in [45].

Network 1 in [44] is the Higgs dataset that was created on July 4, 2012. This dataset contains data extracted from Twitter between July 1 and 7, 2012. Specifically, these dates are before, during, and after the announcement of the discovery of a new particle with the features of the elusive Higgs boson on. This dataset contains three data of the same IDs. These data are the social network of 456626 nodes and 14855842 edges, the retweet network of 256491 nodes and 328132 edges, and the mention network of 116408 nodes and 150818 edges. This dataset has been anonymized such that the same user IDs are used for all networks (social, retweet, and mention). This arrangement allows the use of this dataset in studies on large-scale interdependent/interconnected networks. In these networks, social network accounts for the single social structure and the retweet and mention networks are used to weigh the social network.

Network 2 in [45] includes a dataset containing 121,807,378 tweets posted by 14,599,240 unique users. The authors in [45] created an undirected and unweighted social network based on the reciprocal relationships among 595,460 randomly selected users. Two other types of networks are constructed based on retweets and mentions. This dataset has been anonymized such that the same user IDs are used for all networks (social, retweet, and mention).

On the basis of these data sets, social network nodes are used to construct the network and we extract the number of retweets and mentions corresponding to each user from the retweet and mention networks. Then, the data is used to construct weights for this social network. For example, for a directed network in network 1, if user 1 follows user 2, user 1 retweets user 2 (2 times) and user 1 mentions user 2 (1 time). The network is constructed in such way a directed link is created from user 1 and user 2 with weight equal to 3 (total number of interactions). for an undirected network in network 2, if user 1 and user 2 follow each other, user 1 and user 2 retweet each other (2 times), and user 1 and user 2 mention each other (1 time). The network is constructed in such way an undirected link is created between user 1 and user 2 with weight equal to 3 (total number of interactions).

These two large data sets are used in this study. These data sets comprehensively represent all social, retweet, and mention networks among the same users. These data sets have been anonymized such that the same user ID is used for all networks (social, retweet, and mention). Using these actual two large-scale networks, containing actual large-scale interdependent/interconnected networks (one network represents the social structure and two networks translate different types of user interaction dynamics) are sufficient to test the effectiveness of the proposed algorithm. Previous studies (Lü et al., 2011; L. Weng et al., 2012) used only a single network from Twitter to draw their conclusions.

3.3.2 Network Representation

In this research, OSNs network is exemplified as networks in which nodes are connected by links. In this study, nodes represent the users, and links represent the relationship of users throughout the networks. Assuming that a network can be viewed as network G = *(V, E)*, where *V* = nodes (users) and *E* = links (relationship), the users

are nodes and the single network (such as following connections) between the users represent the link.

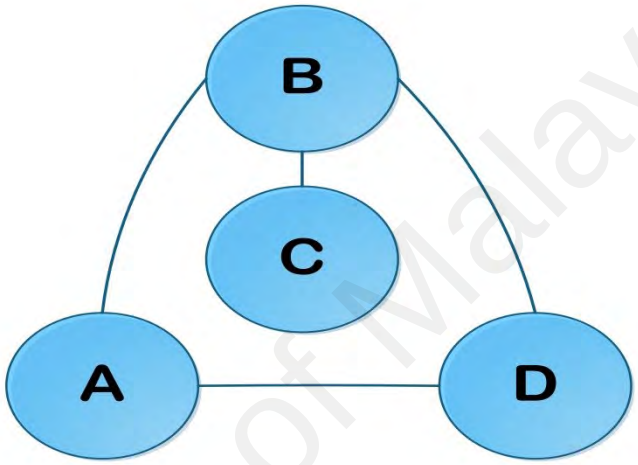

Figure 3.2 : Network Representation

Consider that A, B, C, and D, are users of OSNs; social links (e.g., the following connections) among (A, B), (A, D), (B, C), and (B, D) exist. The network of these users can be represented as shown in Figure 3.2.

3.3.3 Proposing Identification of Influential Spreaders method

This research intends to develop an effective method for influential spreaders identification , consequently Interaction weighted k-core method(IWK_S) is proposed by presenting a novel link-weighting method based on the interaction among users. Social network connection among users is constructed, and the link among users is weighted

using the number of interaction among users extracted from interaction activities (the retweet and mention information among users). The users are then connected using their social network relationship, and these links are weighted using the number of interaction among them. The proposed weighting method is based on the observation that interaction among users is a significant factor in quantifying the spreading ability of a user in OSNs (L. Weng et al., 2012).

As reviewed in chapter 2 , a number of different complex network method have been applied to the constructed OSNs' networks to identify influential spreaders (Pei & Makse, 2013). The most prominent ones include classical centrality measures in complex networks such as degree centrality (Barabási & Albert, 1999; Jiang et al., 2013; Mislove et al., 2007), betweenness centrality (L. C. Freeman, 1977), closeness centrality (Faust, 1997), and eigenvector centrality (Borgatti & Everett, 2006; Duda et al., 2012; H. He, 2007) , PageRank (Brin & Page, 2012; Q. Li et al., 2014; Lü et al., 2011; J. Weng et al., 2010) and k-core algorithm (Batagelj & Zaversnik, 2003; Dorogovtsev, Goltsev, & Mendes, 2006; Kitsak et al., 2010; Pei et al., 2014). However closeness and betweenness centrality has high computational complexity; hence, it is unsuitable to be applied into significantly large-scale OSNs. Also eigenvector centrality is inefficient, particularly in scale-free networks, (Catanese et al., 2012; Morone & Makse, 2015). Consequently, because betweenness centrality, closeness centrality, and eigenvector centrality are infeasible to be applied to large-scale social networks, in this research degree, PageRank, and k-core are used as baselines to be compared with the proposed method(IWK_S). Detailed discussion on these baselines method and their shortcomings are presented in chapter2 (literature review chapter) and detailed discussion on developed method is presented in chapter 5.

3.3.4 Evaluation

In the present study, the proposed influential spreaders identification method is evaluated using real spreading dynamics of information evaluation model as proposed in (Pei et al., 2014) .

Many evaluation models are established in literature, selecting evaluation model for identifying influential spreaders is an important step to ensure the effectiveness of the proposed method. Evaluation model are evaluation through artificial models, such as susceptible-infectious-recovered (SIR) (Kitsak et al., 2010; Pei & Makse, 2013), susceptible-infectious-susceptible (SIS) (Hethcote, 2000), rumor-spreading models (Borge-Holthoefer & Moreno, 2012), and Evaluation through real spreading dynamics of information (Pei et al., 2014). SIR and SIS as well as rumor spreading models are used to verify the effectiveness of different influential users' measurements in complex networks (Kitsak et al., 2010; J.-G. Liu, Ren, & Guo, 2013; Wei et al., 2015). These models have been active in simulating information spread (Kim & Han, 2009; Q. Li et al., 2014; Lü et al., 2011).

However, OSNs have inherent properties that differentiate them from traditional social networks. These properties are challenging tasks for developing a model that can efficiently model the influence spread in the OSN context. Studies have concluded that such artificial models fail to generate accurate diffusion patterns (Pei et al., 2014; Pei, Muchnik, Tang, Zheng, & Makse, 2015b). Consequently, this conclusion explains the intensive argument in previous researches that claims the best approaches in identifying influential spreaders. Previous studies claim inconsistent outcomes according to specific models used in information diffusion. The models are inspired by the spread of a contagious disease (Hethcote, 2000) and are proposed on the basis of a basic hypothesis of human behavior that cannot be illustrative and representative of real dynamic

information diffusion in OSNs(Pei et al., 2014; Pei et al., 2015b). Moreover tracking real diffusion processes shows that the spread of diseases and the spread of information are different (Centola & Macy, 2007; Singh et al., 2013). Therefore, in the present study real spreading dynamics of information (Pei et al., 2014) is used to evaluate influential spreaders identification methods. Detailed description on the evaluation model is explained in chapter 5.

3.4 Conclusion

This chapter presents the general methodology used in the design and implementation of the proposed methods for cyberbullying detection and identification of influential spreaders in OSNs. Specific details for each method and their contribution are provided in Chapters 4 and 5, respectively.

CHAPTER 4: EFFECTIVE CYBERBULLYING DETECTION METHOD

4.1 Introduction

This chapter presents an effective method for detecting cyberbullying in an OSN. The significant contribution is the proposed comprehensive features derived from Twitter, including network, activity, user, and tweet content. On the basis of these features, a supervised machine learning method is constructed for detecting cyberbullying in Twitter.

For developing the effective cyberbullying detection method, this research utilizes useful information in tweets to extract comprehensive features. In particular, useful features in Twitter, such as those related to network, activity, user, and tweet content, are proposed and used to construct cyberbullying detection method. Figure 4.1 shows the experimental processes in constructing the cyberbullying detection method based on the proposed features.

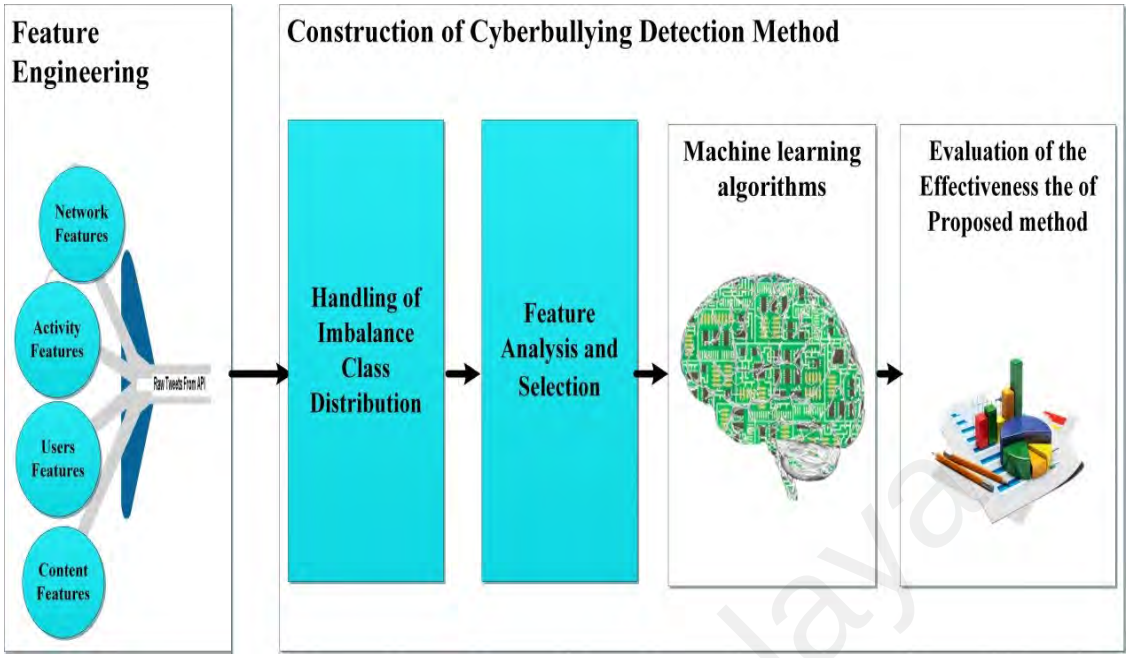

Figure 4.1: Experimental construction of cyberbullying detection method using proposed features

Data from Twitter are extracted and preprocessed as discussed in chapter 3. As shown in Figure 4.1, a set of features is then proposed from the extracted data that includes network information, activity information, user information, and tweet content. The synthetic minority over-sampling technique (SMOTE) and the weight adjusting approach (cost-sensitive) are used to balance the classes in the data set. Three feature selection algorithms, namely, chi-square test, information gain, and Pearson correlation are used, to select the significant features as input to the classifiers. Thereafter, the performance of four classifiers is compared under four different settings to select the best setting for the proposed features. These four classifiers are NB, SVM, RF, and KNN. AUC is used as the main evaluation metric. The following sections explain each block in Figure 4.1 in detail.

4.2 Feature Engineering

Previous studies (see Table 2.1) focused on constructing cyberbullying detection using based limited features such as content based features (i.e. bag of word, skip-

grams, profanity related word general cyberbullying-related words, sentiment features and pronouns features) and profile based features (i.e. as age or gender features (limited to the information in the profile), number of friends, timestamp and location of posts). However, these features are not comprehensive to construct effective cyberbullying detection method (refer to chapter 2 section feature related issue). There are features are yet to be used to construct more effective detection methods

In this section, comprehensive features are presented for constructing cyberbullying detection method based on four categories, namely, network, activity, user, and content. These features are mainly derived from tweet content (tweet text) and information, such as network and activity. These features, as explored in the succeeding subsections, are used in conjunction with a supervised machine learning algorithms to construct a cyberbullying detection method.

4.2.1 Network Features

Set of network related features are extracted which includes the number of friends following a user (followers), number of users being followed by a user (following), following–followers ratio, and account verification status. A survey research observed a strong correlation between the cyberbullying behavior and sociability of users in online environments (Navarro & Jasinski, 2012). These observations from survey study prompted the inclusion of network-related features in construction the cyberbullying detection method in the present study. These proposed features measure the sociability of Twitter users (Lee, Mahmud, Chen, Zhou, & Nichols, 2014). As shown in Table 2.1 the previous studies mostly use only number of followers as feature to construct cyberbullying detection method. In this study, the set of network related features mentioned above are used.

4.2.2 Activity Features

Activity features measure the online communication activity of a user (Pennacchiotti & Popescu, 2011). The number of posted tweets, favorited tweets, and URLs as well as hashtags and mentioned users (e.g., AT_USERname) in a tweet was extracted to measure the activeness of Twitter users. A survey research determined that users who are considerably active in online environments tend to engage in cyberbullying (Balakrishnan, 2015) therefore these observations encouraged the inclusion of activity-related features in construction the cyberbullying detection method in the present study.

4.2.3 User Features

This subsection explores the features related to users.

4.2.3.1 Personality Features

OSNs have become a place where users can present and introduce themselves to the virtual world. Many researchers have utilized OSN data to predict the personality of users within such networks (Adali & Golbeck, 2012; Golbeck, Robles, & Turner, 2011; Quercia, Kosinski, Stillwell, & Crowcroft, 2011). In particular, these researchers predict personality by analyzing online communication data from social media, and personality prediction in social media can facilitate the understanding of human behaviors (Mahmud, Zhou, Megiddo, Nichols, & Drews, 2013).

Survey studies show that hostility significantly predicts cyberbullying (Arıcak, 2009), and both bullying and cyberbullying are strongly related to neuroticism (Connolly & O'Moore, 2003; Corcoran et al., 2012). Neuroticism is characterized as anxiety, anger, and moodiness. Neurotic people are angrier, moodier, and more tense than normal, thereby indicating that a neurotic user is likely to engage in cyberbullying. Predicting neurotic personality and neurotic-related text can provide useful discriminative information. These observations from survey study prompted the

inclusion of neurotic -related features (words) in construction the cyberbullying detection method in the present study.

The words in writings, such as blogs and essays, are also related to user personality (Fast & Funder, 2008; Gill, Nowson, & Oberlander, 2009; Mairesse & Walker; Tausczik & Pennebaker, 2010). Previous studies (Golbeck, Robles, Edmondson, & Turner, 2011; Golbeck, Robles, & Turner, 2011; Mahmud et al., 2013) encouraged the use of social media text for personality prediction even if the text contains only a few words. Previous studies on personality prediction in social media revealed a positive correlation between neuroticism and usage of anxiety- and anger-related words (Adali & Golbeck, 2012; Golbeck, Robles, & Turner, 2011; Quercia et al., 2011; Schwartz et al., 2013). Therefore, predicting neuroticism can provide a beneficial discriminative feature in detecting cyberbullying. The previous study (Schwartz et al., 2013) developed the word related to neuroticism. In this research to predict neuroticism, the one hundred most common words used in social media positively correlated with neuroticism and the one hundred most common words used in social media negatively correlated with neuroticism are used. Using neuroticism related words as features to detect cyberbullying occurrence was overlooked. According to the best author's knowledge, this study is first study to introduce these features to detect cyberbullying.

4.2.3.2 Gender

Many survey researchers have investigated the relationship between gender and engagement in cyberbullying. A few studies (Calvete et al., 2010; Vandebosch & Van Cleemput, 2009) show that males are more likely to engage in cyberbullying than females, but other studies indicate the opposite (Dilmac, 2009; Sourander et al., 2010). Another study determines no significant difference between males and females in terms of tendency to engage in cyberbullying (Kowalski, Giumetti, Schroeder, & Reese,

2012). Although survey studies do not clearly confirm the relationship between gender and cyberbullying behavior, another study (Van Royen et al., 2015) that conducted a survey involving numerous experts in the field of cyberbullying suggested including gender to build effective detection methods (Van Royen et al., 2015). Another similar study showed that males engage in cyberbullying to a considerable extent than females do (Calvete et al., 2010). Accurate cyberbullying detection in online social networks is improved by using gender-related information (information from profiles) as a feature (Dadvar et al., 2012a). These observations from survey studies prompted the inclusion of gender-related features in building the machine learning method in the present study.

Unfortunately, most users do not mention their gender in their profiles, and not all gender information provided in user profiles are correct (Peersman et al., 2011). Recent studies have applied natural language processing to predict the gender of users based on their writing styles. Previous studies showed that males and females use specific words to distinguish themselves from the opposite gender. For example, females use the word "shopping" more frequently than males (Schwartz et al., 2013). Accordingly, in this research features related to the gender of users were extracted and used to detect cyberbullying. First to predict gender based on tweet text, using 100 most common words that are used/not used by males/females as proposed in (Schwartz et al., 2013),. Second to predict gender based on the first name of the user was also adopted using a large gender-labeled data set proposed in (W. Liu & Ruths, 2013) which include the large number of male and female number , whether a user was male or female was predicted based on the first name reported in his/her tweet.

4.2.3.3 Age

Survey studies discussed the effects of age on cyberbullying engagement. These studies contended that cyberbullying decreases as age increases age, and the highest rate

of cyberbullying is among teenage users (Slonje & Smith, 2008; Williams & Guerra, 2007). Nevertheless, the older age group must be considered. Most OSN users do not provide information on their age or date of birth, and not all users provide accurate age information (Peersman et al., 2011). Similar to gender, the lack of information on age imposes a challenge.

However, user age can be predicted by analyzing the language used by users from different age levels. In a most OSN users do not clearly state their age and gender in their online profiles (Peersman et al., 2011); thus, many studies (Hosseini & Tammimy, 2016; Peersman et al., 2011; Rangel & Rosso, 2013; Santosh, Bansal, Shekhar, & Varma, 2013; Talebi & Kose, 2013) have focused on predicting age and gender in OSNs by using text analysis and natural language processing on user posts. The aforementioned study (Schwartz et al., 2013) has developed an open vocabulary by analyzing 700 million words, phrases, and topic instances collected from social media and determined the words related to age, gender, and personality. In this study (Schwartz et al., 2013) , age levels are classified into age level 1: 13 to 18 years; age level 2: 19 to 22 years; age level3 : 23 to 29 years; and age level 4: 30 years and above and every set of words related to age level were developed. The study (Schwartz et al., 2013) provided comprehensive exploration of language that distinguishes people, thereby determining connections that are not obtained with traditional closed vocabulary word category analyses, such as Linguistic Inquiry and Word Count. The proposed open vocabularies related to gender and age have been used in various studies. For example, a study used these vocabularies to predict county-level heart disease mortality in Twitter (Eichstaedt et al., 2015). Another study used these vocabularies to propose a feature related to age and gender to detect mental illnesses, such as depression and post-traumatic stress disorder, in Twitter (Preotiuc-Pietro et al., 2015). These vocabularies were also used for predicating age and gender features from user posts on social media

websites to build a machine learning classifier for different applications (Burger, Henderson, Kim, & Zarrella, 2011; L. Li, Sun, & Liu, 2014; Miller, Dickinson, & Hu, 2012; D.-P. Nguyen, R. Gravel, R. Trieschnigg, & T. Meder, 2013; Rangel & Rosso, 2013; Rao, Yarowsky, Shreevats, & Gupta, 2010). The present study uses the open vocabularies related to age and gender proposed in the aforementioned study (Schwartz et al., 2013) to construct cyberbullying detection method. According to the best author's knowledge, this study is the first study to use these open vocabularies to predict the gender and age to be used as features for cyberbullying detection. The same age levels were used in the present study as age-related features. Nearly 800 age-related words in social media are used to distinguish users from different age levels. For example, the word "school" is significantly related to age level 1, whereas "job" is substantially related to age level 3.

4.2.4 Content Features

4.2.4.1 Vulgarity Features

Using vulgar words in online communication can be used to detect cyberbullying because they may signal hostility and offensive behaviors. Similarly, tweets containing a vulgar or profane word may be considered a cyberbullying tweet (Xiang, Fan, Wang, Hong, & Rose, 2012).

Vulgarity is a useful discriminative feature for detecting offensive and cursing behaviors in Twitter (W. Wang, Chen, Thirunarayan, & Sheth, 2014; Xiang et al., 2012) and cyberbullying in YouTube (Dadvar, Trieschnigg, Ordelman, et al., 2013). These features were extracted from the contents of a user post. The number of profane words was measured in a post using a dictionary of profanity. The words in this dictionary were compiled from sources cited in previous studies (Reynolds et al., 2011; W. Wang

et al., 2014). This feature is used in this study to detect the number of profane words in the posts in order to detect cyberbullying.

4.2.4.2 Special OSNs Acronym and Abbreviation Features

Technology-mediated communication has immensely contributed to the increasing number of novel acronyms and abbreviations. By introducing new terms, such as "unfriend" and "selfie," social media evidently affect language. Acronyms and abbreviations generated from online social media communication assist users in communicating easily and rapidly with one another. These terms can also be easily entered in mobile phones with tiny keypads. In Twitter, acronyms assist users to make the most out of 140 characters. Acronyms and abbreviations have also become popular among mobile phone users who use these terms to reduce their effort and time for typing. Social media have made their presence by introducing new words, adding new meanings to old words and changing the manner users communicate.

Similarly, cyberbullies change the method they use words and acronyms to engage in cyberbullying. OSNs facilitate in creating cyberbullying-related acronyms that have never been used in traditional bullying or beyond social media. The words, acronyms, and abbreviations commonly used in cyberbullying were collected from a previous study (Dailymail, 2014). In this study these words, acronyms, and abbreviations are used as features in order to train the machine learning algorithms.

4.2.4.3 First and Second Person Pronouns

First and second person pronouns in tweets are also used as features that provide useful information about to whom the text is directed. A text containing cyberbullying-related features and a second person pronoun are most likely meant for harassing others (Dadvar, Trieschnigg, & Jong, 2013).

This above process for creating feature vectors from above-mentioned features can be explained in following pseudo code (Figure 4.2): -

Pseudo code for creating features vectors

❖ ***Variables definition***

$r.d$	:	*Raw data*
c	:	Number of class to be classified
ld	:	Number of labeled data
$d.f$	:	*Extraction of Direct features*
$d.b.f$	:	*Dictionaries based features*
f	:	*Frequency count of each feature*
Fn_s	:	*Spelling-checker function*
Fn_{lc}	:	*Lower-case conversion function*
Fn_t	:	*Tokenization*
Fn_{sw}	:	*Stop words removal function*
Fn_r	:	*Repetition of character removal*
Fn_w	:	*white space removal*
$f.arff$	:	*Final feature vectors arff file*

❖ **Algorithm:-**

- **Input: Raw data from Twitter API.**
- **Output: Set of features as input to machine learning classifiers.**

1. For (i): *1 to ld*
2. For (j): *1 to c*
3. LOAD $T \leftarrow r.d(i,j)$
4. $T_{d.f} \leftarrow d.f$
5. End for
6. End for
7. For (i)*: 1 to ld*
8. For (j): *1 to c*
9. LOAD $T \leftarrow r.d(i,j)$
10. $T_s \leftarrow Fn_s$
11. $T_{lc} \leftarrow Fn_l$
12. $T_t \leftarrow Fn_t$
13. $T_{sw} \leftarrow Fn_{sw}$
14. $T_r \leftarrow Fn_r$
15. $T_w \leftarrow Fn_w$
16. $P\ (i,j) = T_s \cup T_{lc} \cup T_t \cup T_{sw} \cup T_r \cup T_w$

```
 17.      End for
18. End for
19. For (i) to ld
20.     For (j) to c
21.       For k = 1 to ld
22.        LOAD D ← d.b.f
23.              f(k) = ∑P(i,j) = D
24.        End for
25.       T_d.bf = f(k)
26.      End for
27. End for
28.  f.arff ← T_d.f + T_d.bf
```

Figure 4.2 : Pseudo Code for Creating Features Vectors

The process to create the feature vectors for the proposed features can be described as follows .First, for every tweet in manually labeled data; a set of feature vectors is extracted based on feature engineering. For every tweet in manually labeled data, direct feature from raw tweets are extracted. These features include number of followers, number of followed accounts, and account verification status. Second, dictionary-based features are created for features, such as age, gender, personality, pronoun, vulgarity and cyberbullying-specific features. The required raw data corresponding to the labeled examples are pre-processed in functions, and then the token in each tweet is converted to feature vector space using term frequency matching from corresponding dictionary of each feature; the count is considered as weight of the feature. Finally, both direct- and dictionary-based features are combined in ARFF file to be provided to the machine learning algorithms in the succeeding section.

4.3 Experimental Construction of Cyberbullying Detection Method Using Proposed Features

An extensive set of experiments were run to measure the performance of the four classifiers (i.e., NB, LIBSVM, RF, and KNN). All experiments were performed using WEKA.

4.3.1 Experiment Settings

All four classifiers were tested in four different settings, namely, basic classifiers, classifiers with feature selection techniques, classifiers with SMOTE alone and with feature selection techniques, and classifiers with cost-sensitive alone and with feature selection techniques. All experiments were based on a 10-fold cross-validation (Kohavi, 1995; Refaeilzadeh, Tang, & Liu, 2009).

All four classifiers were tested in four different settings as following

I. Basic classifiers (extracted features are input directly into the classifiers); see Figure 4.3.

II. Classifiers with feature selection techniques (feature algorithms are applied to the extracted features and then input into to the classifiers); see Figure 4.4.

III. Classifiers with SMOTE alone and with feature selection techniques. First, SMOTE algorithms are applied to the extracted feature. The output of applying SMOTE is provided directly as input into the classifiers without applying feature selection algorithms. Second, SMOTE algorithms are applied to the extracted feature; the output is provided as input into the classifiers after applying feature selection algorithms; see Figure 4.5.

IV. Classifiers with cost-sensitive algorithms alone and with feature selection techniques. First, cost-sensitive algorithms are applied to the extracted feature. The output of applying the cost-sensitive algorithm is provided directly as input into the classifiers without applying feature selection algorithms. Second, cost-sensitive algorithms are applied to the extracted features, and the output is provided as input into the classifiers after applying feature selection algorithms; see Figure 4.6.

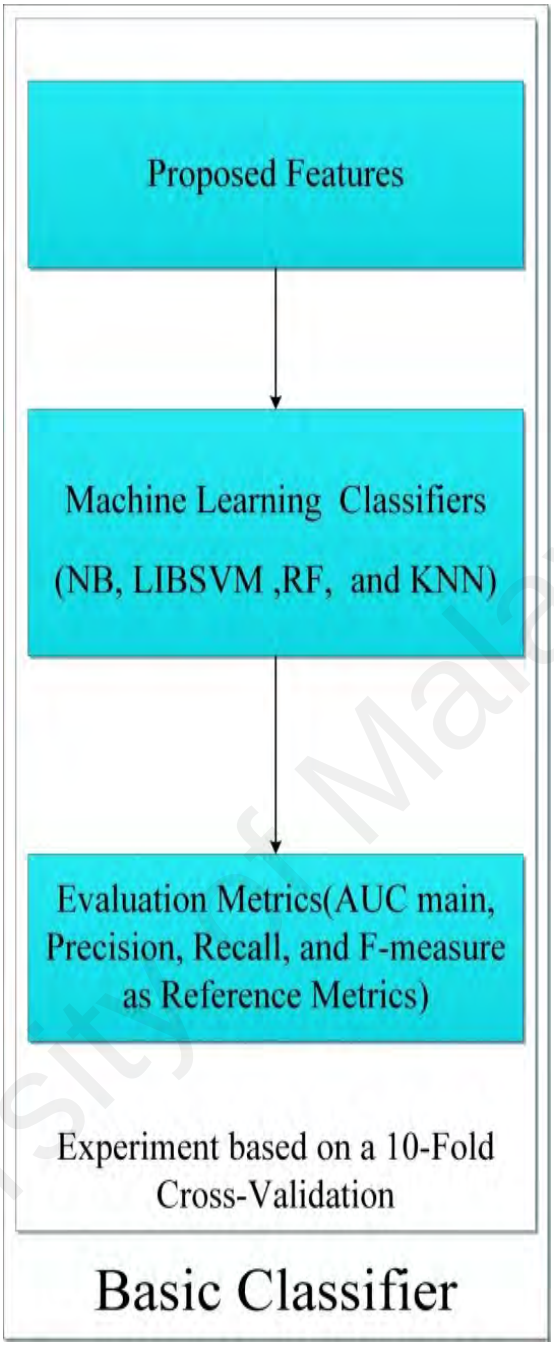

Figure 4.3: Experiment Setting 1 (basic classifiers)

Figure 4.3 shows that the proposed features are input into the (i.e., NB, LIBSVM, RF, and KNN). AUC, Precision, Recall, and F-measure of all these experiments are represented in Table 4.1.

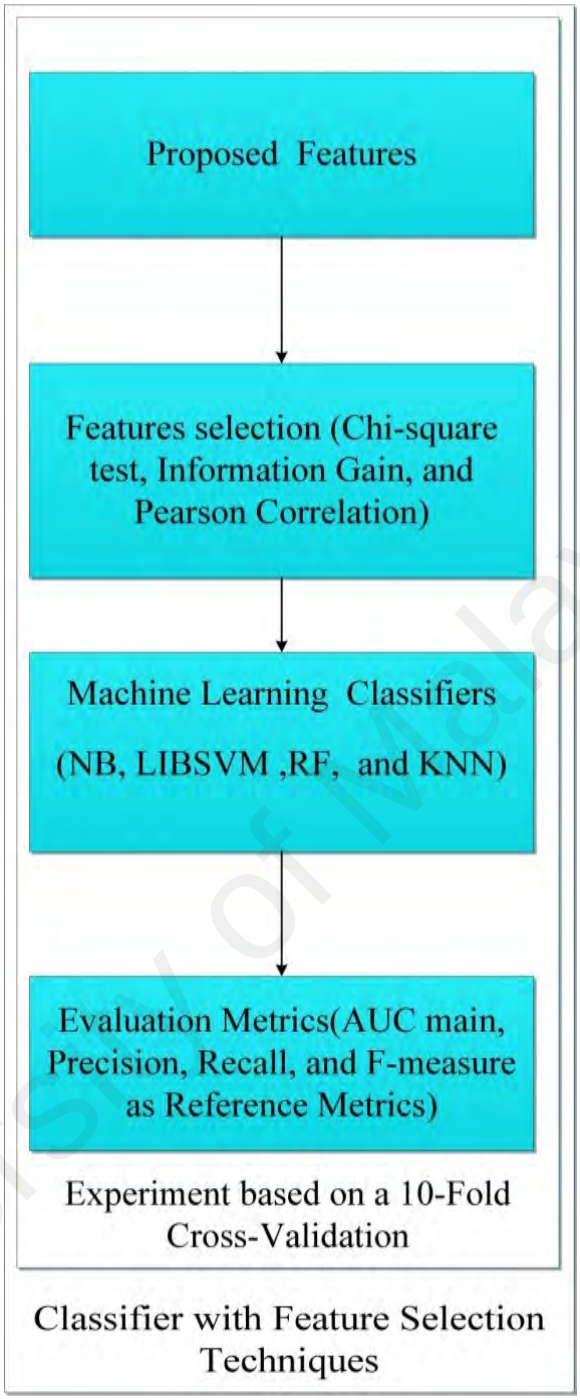

Figure 4.4: Experiment Setting 2 (Classifiers with Feature Selection Techniques)

Figure 4.4 shows firstly feature selections algorithm (chi-square test, information gain, and Pearson correlation) are applied to the proposed features then the outputs of each feature selection algorithm is fed into the classifiers (i.e., NB, LIBSVM, RF, and KNN). AUC, Precision, Recall, and F-measure of all these experiments are reported in in Table 4.2, Table 4.3. and Table 4.4.

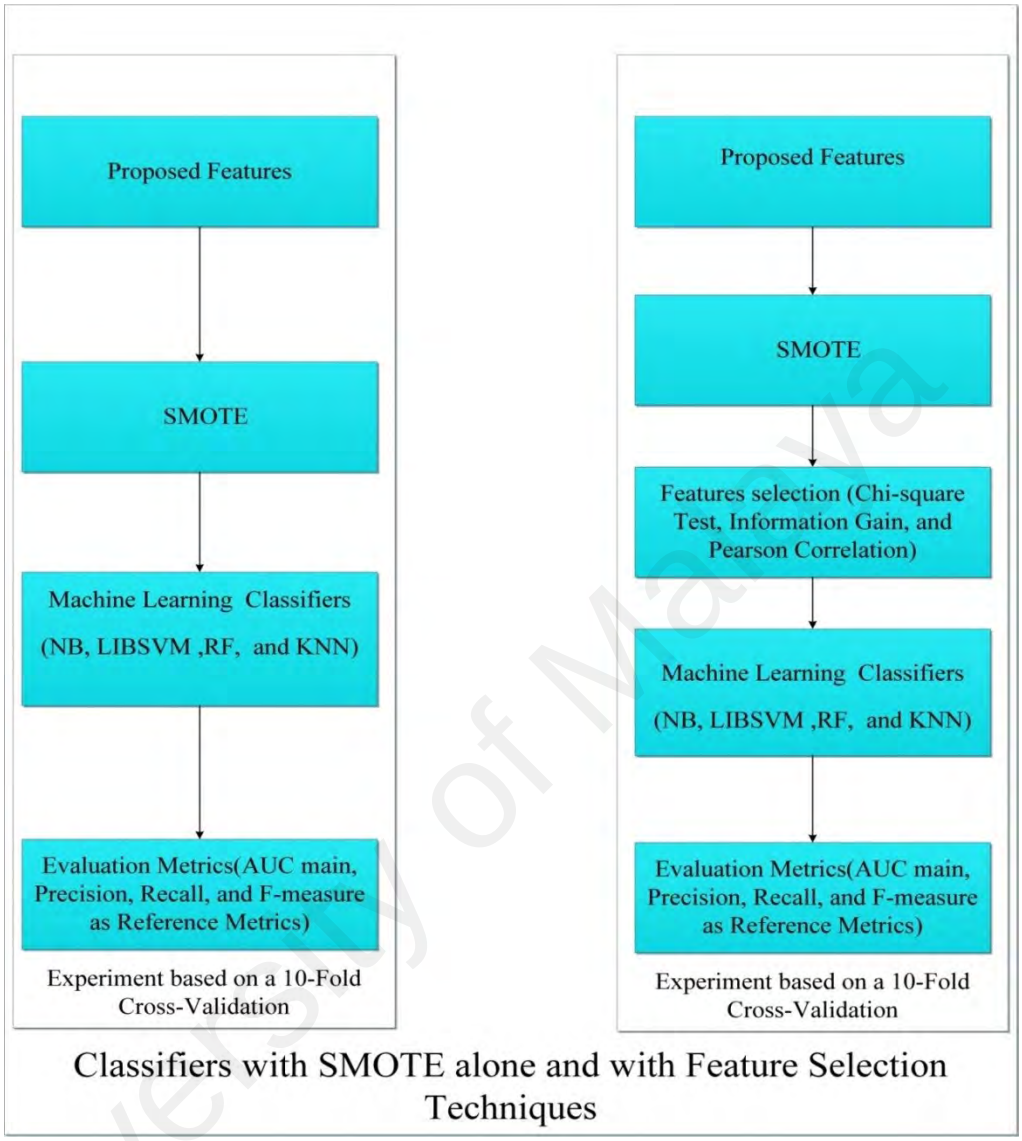

Figure 4.5: Experiment Setting 3 (Classifiers with SMOTE alone and with Feature Selection Techniques)

In Figure 4.5 First SMOTE is applied to the proposed features and the outputs are fed into the classifiers (i.e., NB, LIBSVM, RF, and KNN). Secondly SMOTE algorithm is applied on the proposed feature then feature selection algorithms (square test, information gain, and Pearson correlation) are applied . The outputs are then fed into classifiers (i.e., NB, LIBSVM, RF, and KNN). AUC, Precision, Recall, and F-measure of all of these experiments are reported in result section Table 4.5.

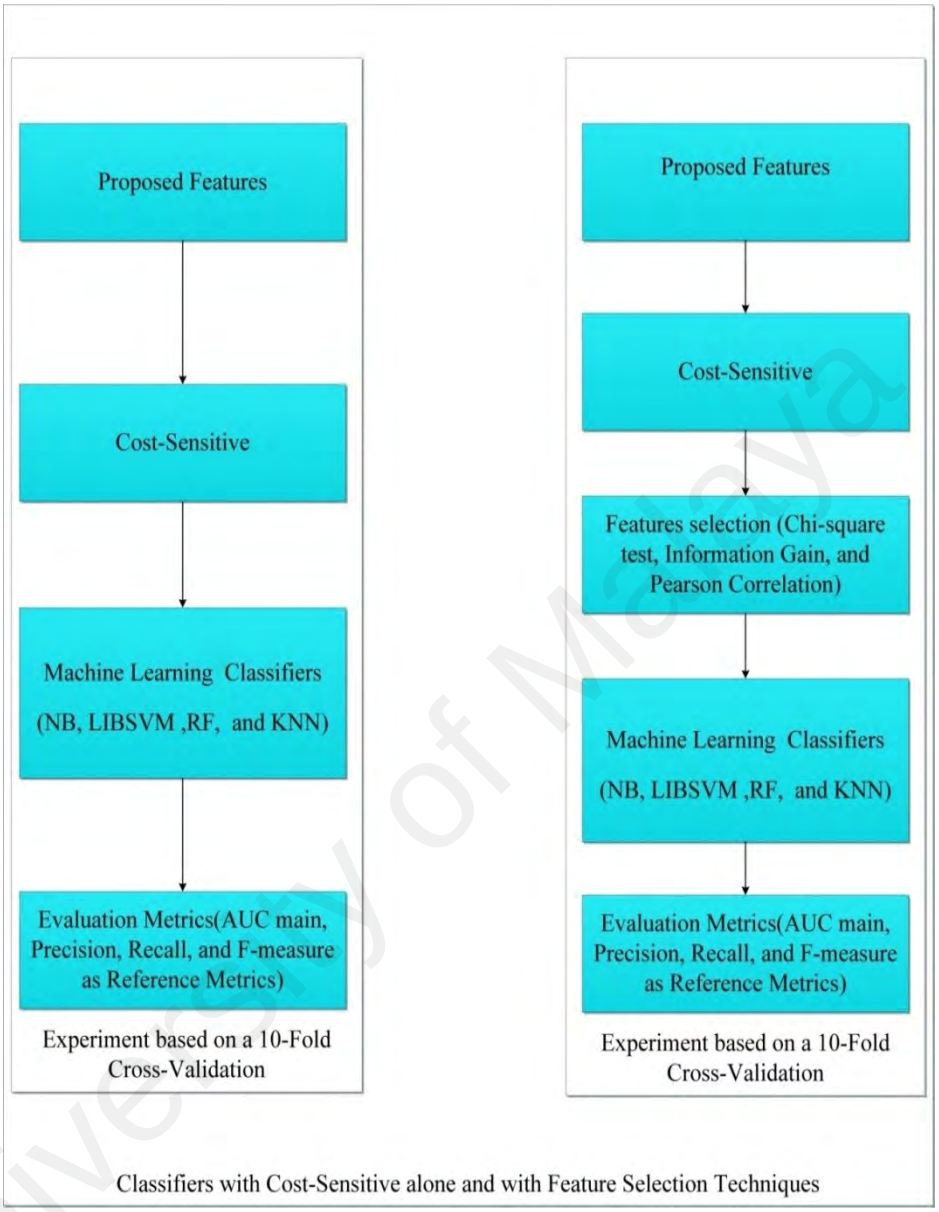

Figure 4.6: Experiment Setting 4 (Classifiers with Cost-Sensitive alone and with Feature Selection Techniques)

In Figure 4.6 First cost-sensitive was applied to the proposed features and the outputs are fed into the classifiers (i.e., NB, LIBSVM, RF, and KNN). Secondly cost-sensitive algorithm was applied on the proposed feature then feature selection algorithms (square test, information gain, and Pearson correlation) are applied. The outputs then fed into classifiers (i.e., NB, LIBSVM, RF, and KNN). AUC, Precision ,Recall ,and F-measure of all of these experiments were reported in result section Table 4.6.

4.4 Result and Discussion

First, the results of classifiers using basic setting were obtained. Second, the results of these classifiers with feature selections (i.e., chi-square test, information gain, and Pearson correlation) were obtained. Third, the results of these classifiers with SMOTE alone and with feature selection techniques were obtained. Fourth, the results of these classifiers with cost-sensitive alone and with feature selection techniques were obtained.

Results Obtained Using Basic Classifiers: Table 4.1 reports the results of the features are directly as input into the classifiers (i.e., NB, LIBSVM, RF, and KNN).

Table 4.1: Results Obtained Using Basic Classifiers

Classifier	Precision	Recall	F-measure	AUC
NB	0.909	0.897	0.903	0.690
LIBSVM	0.890	0.944	0.916	0.500
RF	0.908	0.942	0.917	0.626
KNN	0.910	0.937	0.920	0.588

Table 4.1 presents the results of all four classifiers were run using the proposed features based on a 10-fold cross-validation. Table 4.1 provides the results for each classifier. The AUC results vary between 0.5 and 0.69. NB showed the best overall performance under basic setting with an f-measure varying between 0.903 and 0.920.

Results Obtained by Classifiers with Feature Selection: Table 4.2, 4.3, 4.4 report the results of applying the feature selections algorithm (square test, information gain, and Pearson correlation) respectively into the proposed features then it is fed into the (i.e., NB, LIBSVM, RF, and KNN). All four classifiers with feature selection were run to determine the most significant feature that might improve the performance of the

classifier. Three feature selection algorithms, namely, chi-square test, information gain, and Pearson correlation, were tested in the experiment. Different feature combinations were tested, and different numbers of features were iteratively selected to determine a combination with a significant discriminative power that can provide an improved outcome. Tables 4.2 to 4.4 compare the four classifiers with each feature selection method.

Table 4.2: Results Obtained Using Chi-square Test

Classifier	Precision	Recall	F-measure	AUC
NB	0.909	0.901	0.905	0.704
LIBSVM	0.890	0.943	0.916	0.500
RF	0.903	0.940	0.917	0.629
KNN	0.907	0.935	0.918	0.568

Table 4.2 presents the results obtained using chi-square test to select significant features. Compared with the results in Table 4.1, AUC for NB (0.704) and RF (0.629) slightly improve, but AUC slightly decreases for KNN (0.568). AUC for SVM (0.500) holds.

Table 4.3: Results Obtained Using Information Gain

Classifier	Precision	Recall	F-measure	AUC
NB	0.909	0.901	0.904	0.705
LIBSVM	0.944	0.890	0.943	0.500
RF	0.904	0.940	0.917	0.637
KNN	0.904	0.933	0.916	0.570

Table 4.3 shows the results obtained using information gain to select the significant features. Similar to the results in the Table 4.2, AUC for NB (0.705) and RF (0.637)

slightly improve from the results in Table 4.1. AUC for KNN (0.570) slightly decreases, whereas that for SVM (0.500) holds.

Table 4.4: Results Obtained Using Pearson Correlation

Classifier	Precision	Recall	F-measure	AUC
NB	0.909	0.898	0.904	0.701
LIBSVM	0.890	0.944	0.916	0.500
RF	0.901	0.941	0.916	0.646
KNN	0.910	0.937	0.920	0.588

Table 4.4 shows the results obtained using Pearson correlation to select the significant features. Compared with the results in Table 4.1, AUC for NB (0.701) and RF (0.646) slightly improve, whereas those for SVM (0.500) and KNN (0.588) hold.

In summary, using the three feature selection techniques only slightly improves the AUC results compared with the results using basic setting (see Table 4.1).

Results Obtained by Classifiers with Imbalanced Data Distribution: Oversampling the minority (abnormal) class and undersampling the majority (normal) class (SMOTE) were applied, along with weight adjusting approaches (cost-sensitive), to handle imbalanced data distribution. The four classifiers with these two approaches were tested with and without feature selection to obtain the best result.

Table 4.5: Results Obtained Using SMOTE

Cases	Classifier	Precision	Recall	F-measure	AUC
SMOTE only	NB	0.763	0.774	0.768	0.692
	LIBSVM	0.820	0.831	0.786	0.583
	RF	**0.941**	**0.939**	**0.936**	**0.943**
	KNN	0.870	0.873	0.871	0.866
SMOTE and chi-square test	NB	0.760	0.769	0.764	0.703
	LIBSVM	0.820	0.833	0.792	0.593
	RF	0.934	0.934	0.930	0.924
	KNN	0.863	0.871	0.865	0.846
SMOTE and information gain	NB	0.764	0.774	0.769	0.717
	LIBSVM	0.823	0.835	0.796	0.599
	RF	0.897	0.897	0.897	0.929
	KNN	0.861	0.869	0.863	0.843
SMOTE and\Pearson correlation	NB	0.762	0.781	0.770	0.708
	LIBSVM	0.829	0.836	0.796	0.598
	RF	0.939	0.938	0.934	0.932
	KNN	0.864	0.871	0.866	0.850

Table 4.5 shows the results obtained using the four classifiers with SMOTE. In summary, using SMOTE significantly improves AUC for all classifiers, except for NB, which only shows a small improvement.

Table 4.6: Results Obtained using Cost-Sensitive

Cases	Classifier	Precision	Recall	F-measure	AUC
Cost-sensitive only	**NB**	0.909	0.897	0.903	0.690
	LIBSVM	0.919	0.944	0.916	0.502
	RF	0.908	0.942	0.917	0.626
	KNN	0.910	0.937	0.920	0.588
Cost-sensitive and chi-square test	**NB**	0.909	0.901	0.905	0.704
	LIBSVM	0.919	0.944	0.916	0.502
	RF	0.903	0.940	0.917	0.629
	KNN	0.907	0.935	0.918	0.568
Cost-sensitive and information gain	**NB**	0.909	0.901	0.904	0.705
	LIBSVM	0.913	0.943	0.916	0.502
	RF	0.904	0.940	0.917	0.637
	KNN	0.904	0.933	0.916	0.570
Cost-sensitive and Pearson correlation	**NB**	0.909	0.898	0.904	0.701
	LIBSVM	0.947	0.944	0.917	0.502
	RF	0.930	0.901	0.941	0.646
	KNN	0.912	0.938	0.921	0.573

Table 4.6 shows the results obtained using the four classifiers with the weight adjusting approach (cost-sensitive). In summary, using classifiers with cost-sensitive does not significantly improve the AUC of the classifiers.

Summary of Results:

Table 4.5 shows that the best overall classifier performance is achieved using SMOTE. In particular, RF using SMOTE alone demonstrates the best AUC (0.943) and f-measure (0.936).

The following Figures 4.7 and 4.8 compare the ROC results of all classifiers under the basic and best performance settings (i.e., classifiers using SMOTE).

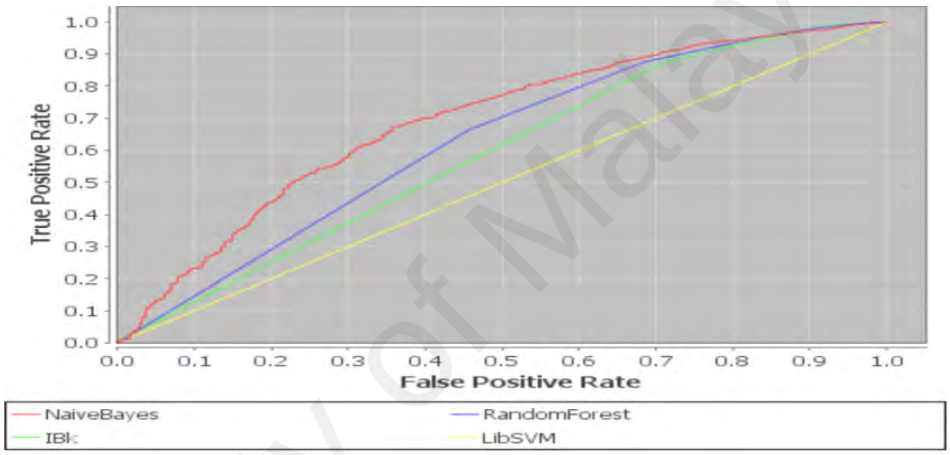

Figure 4.7: ROC results for the four classifiers under the Basic setting

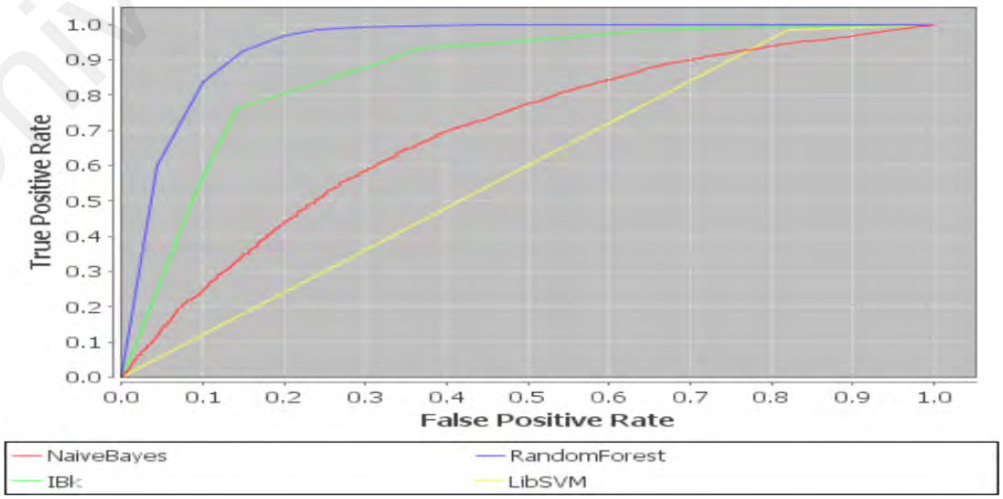

Figure 4.8: ROC results for the four classifiers using SMOTE alone

The confusion table for the best performance classifier (i.e., RF using SMOTE) is presented in Table 4.7.

Confusion table		Classified	
		Non-Cyberbullying	**Cyberbullying**
Actual	Non-Cyberbullying	99.4 % (TP)	0.6 % (FN)
	Cyberbullying	28.6 % (FP)	71.4 % (TN)

Table 4.7: Confusion Table

- TP is the percentage of instances that are non-cyberbullying and correctly classified as non-cyberbullying.
- FN is the percentage of instances that are non-cyberbullying and incorrectly classified as cyberbullying.
- TN is the percentage of instances that are cyberbullying and correctly classified as cyberbullying.
- FP is the percentage of instances that are cyberbullying and incorrectly classified as non-cyberbullying.

Overall, machine learning working in an online communication environment should be balanced between providing effective methods for detecting cyberbullying content or any negative behavior that does not ethically harm other innocent contents or users. The Table 4.7 shows that the proposed cyberbullying detection method classified 99.4% of non-cyberbullying as non-cyberbullying; therefore, this result meets the ethical challenges (Vayena, Salathé, Madoff, Brownstein, & Bourne, 2015) because the content falsely detected as cyberbullying but is actually not is at a low rate of 0.6 % (percentage of non-cyberbullying tweets classified as cyberbullying.. The performance results (AUC = 0.943) and confusion table of RF using SMOTE emphasize that the constructed

method based on the proposed features provides a feasible solution to detecting cyberbullying in online communication environments.

4.4.1 Discussion

As described in Chapters 2 and 3, deciding which and why a machine learning algorithm performs better on a given data set is a complex task that depends both in the fundamental theory of the algorithm and the manner in which it matches better with the characteristic of the data; no optimal classifier for all data sets exists (Wolpert & Macready, 1997). Machine learning algorithms are composite, and they often consist of many components (Vanschoren, Blockeel, Pfahringer, & Holmes, 2012). The superiority of a classifier may be limited to a given domain that may display similar data characteristics where a specified classifier may perform well compared with other classifiers (Macià et al., 2013). Therefore, the literature in cyberbullying detection reviewed to select the machine-learning algorithm to be used from is narrow. As described in Chapter 3, the three points used to narrow down the selection of machine learning algorithm are merely helping steps to finalize the selection of classifier. With these helping points as basis, the above machine learning algorithms are used with the proposed features. The results after applying the selected machine learning algorithms yield that RF using SMOTE provides preeminent performance.

RF (Breiman, 2001) is machine learning method that combines decision trees and ensemble learning. RF is one of the most powerful supervised classifiers available (Fernández-Delgado et al., 2014). It runs effectively on many data sets, including huge data sets, and can handle numerous input features without parameter removal. RF approximates what features are significant in the processes of classification, and it produces an internal unbiased approximation of the generalization error as the forest construction evolves. RF is an effective technique for approximating missing data

because it preserves the accuracy when a large amount of the data is missing and balances errors in data sets with imbalanced class distribution. SMOTE improves performance because it facilitates the classifier to construct large decision regions with many training instances to learn from, thus enhancing the performance of RF (Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, & W. Philip Kegelmeyer, 2002).

NB classifiers are constructed by applying Bayes' theorem between features. Bayesian learning is commonly used for text classification. This method assumes that the text is generated by a parametric model and utilizes training data to compute the Bayes-optimal estimates of the model parameters. With these approximations, NB categorizes incoming test data (McCallum & Nigam, 1998). The advantage of NB is easy implementation that requires a small amount of training data for estimating parameters. However, the NB assumption is class-conditional independent, thus reducing accuracy in practice; dependencies exist among variables. For example, the existence of insulting words and words related to children's age (e.g., "stupid" and "school") in a sentence increases the probability of being a cyberbullying case. Therefore, B is assumed to work better if the features are with fewer dependencies. Such condition can be met if the features are extracted by traditional method such as bag-of-words in which every word in the sentence forms separate variables, reducing the dependencies between them.

As explained in the literature, SVM is constructed by finding a separating hyperplane in the feature attributes between two classes in which the distance between the hyperplane and the nearest data points of each class is maximized (Hsu et al., 2003). SVM is one of the commonly used classifiers in literature. In the experiments of this research, imbalanced class distribution plays an important role to test the ability of the SVM in finding the separating hyperplane. The experimental results show that the SVM

is not effective compared to other classifiers, thus failing to find the separating hyperplane that maximizes the distance between the instances of two classes for a given data set. Therefore, SVM achieves worse than the other machine learning algorithm techniques. This observation is consistent with the experimental results of a previous study when SVM was applied to data with class imbalance (Mangaonkar et al., 2015).

KNN is regarded as a lazy machine learning technique that classifies data sets based on their similarity with neighbors. The experimental results indicate that KNN does not work well except when used with SMOTE. KNN with SMOTE produces a competitive result, and this may be due to the fact that KNN produces accurate results with a large number of examples (Beyer, Goldstein, Ramakrishnan, & Shaft, 1999).

In constructing a machine learning method, feature selection involves selecting the significant features among the set of all proposed features based on statistical analysis methods (feature selection techniques). Selected features are provided as input to the machine learning algorithms, and this part is important in the learning process. Nevertheless, whether all proposed features together or only the selected features by statistical analysis methods are the most discriminative feature is decided with the improvement in classifier performance. Selecting the most significant features does not always provide improved performance, similar to using the minimum number of features. In certain circumstances, the best classifier performance is found when all proposed features are used. Features that appear irrelevant when used separately may be relevant when utilized in combination (Domingos, 2012), whereas all proposed features work together to provide discriminative features for the classifier. For example, "school" (age 1 feature) could be an insignificant feature by a selection method because it may exist in both classes equally with no discriminative signal. However, if this feature is used along with "stupid," a vulgarity feature, then it shows high probability

toward cyberbullying rather than non-cyberbullying because both features come together in most cyberbullying instances. Adopting the feature-by-feature selection method may show that some features are insignificant, but when these features are used along with others, they provide significant improvement to classifier performance. The importance of feature selection is to remove redundant features. Redundant features are irrelevant features that make determining meaningful patterns challenging. Feature selection algorithms are used for identifying and eliminating noisy or redundant features to reduce the training and executing times (Libbrecht & Noble, 2015). Nevertheless, feature selection does not always guarantee the optimal performance of the classifier.

4.4.2 Effectiveness of the Cyberbullying Detection Method Based on Proposed Features

Given the restrictions in API and the possible ethical/privacy considerations, no public twitter data set (which contains the content of tweets) was available to test the effectiveness of the proposed features. To investigate such effectiveness, two baseline features were created from the extracted data set, namely, bag-of-words and a combination of possible features proposed in previous studies (Chavan & Shylaja, 2015; Dadvar et al., 2012a; Dadvar, Trieschnigg, Ordelman, et al., 2013; Kontostathis et al., 2013; Reynolds et al., 2011) (refer the feature used in these studies in Chapter 2, Table 2.1). These studies are selected for comparison with the proposed features because they comprehensively covered the most commonly used features in the literature. Doing so can result in assertive outcomes. A set of experiments to measure the performance of four classifiers using these two baseline features was run under the four experiment settings. All settings were kept the same as used. Only the proposed features were replaced with baselines features, and then the same experiments were run to find the best setting for each baseline feature. The first baseline feature achieves the best result for NB with information gain (AUC = 0.614), whereas the second baseline feature

achieves the best result for RF using SMOTE alone (AUC = 0.724). The best results obtained from the proposed features with those obtained from two baseline features were compared. Table 4.8 shows the significance of the proposed features.

Table 4.8: Comparison of the AUC Results of the Cyberbullying detection Methods Using Proposed Features and Baselines Features

Features used	AUC under the best setting
Proposed features	0.943
Baseline 1	0.614
Baseline 2	0.724

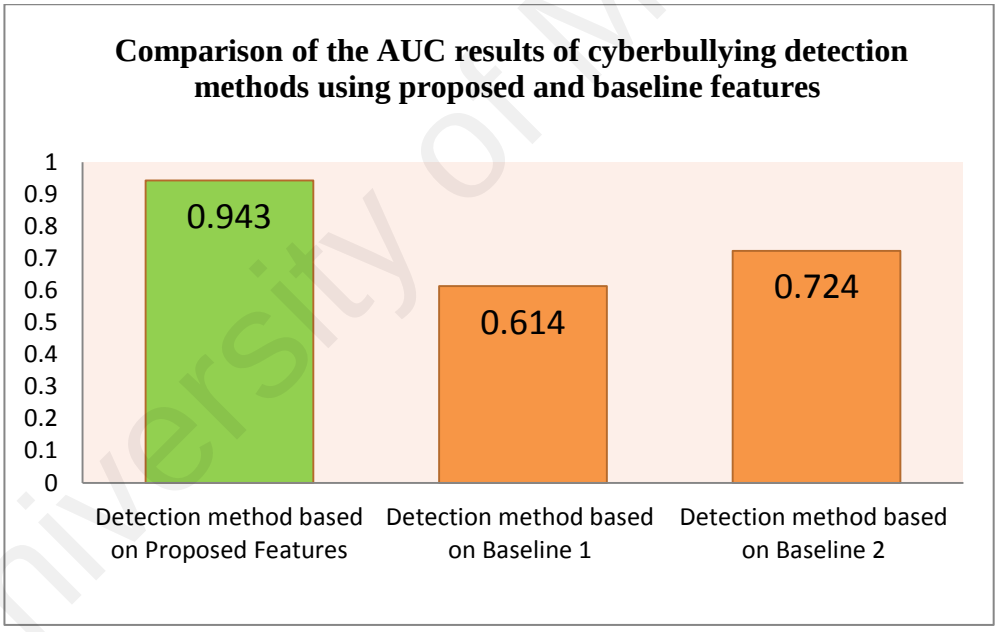

Figure 4.9: Comparison of the AUC Results of the Proposed and Baseline Features under their Best Performance Setting

As discussed in Chapter 2, one of the most important parts of developing an effective method is the proposed set of features in which the machine learning algorithms learn effectively from (learning vectors). However, raw data are not in a format in which machine learning methods can learn from. Features can be created by feature engineering, leading to effective cyberbullying detection method. Simple

representation of the features from textual content, such as bag-of-words (each word in data set is used as feature), may lead to less actual feature representation for training machine learning algorithm. This phenomenon is observed in the results in which the proposed features outperformed bag-of-words. Bag-of-words is governed by searching over a large dictionary to build a set of words from training data to be used as features. Such feature engineering concept imposes significant drawbacks; if none of the words in the training set are included in the testing set, then a divergence will exist between the words in training and testing data sets, leading to low classifier performance. This common issue exists when constructing a machine learning classifier for a field with diverse words to be used such as cyberbullying. Not all words can be significant features, and selected and highly correlated words can be used as significant features to build effective classifier. Therefore, feature-engineering methods, such as bag-of-words fail to performance well. Baseline 2 consists of features from previous studies that aimed to improve cyberbullying detection performance. Although Baseline 2 contains highly significant features for cyberbullying detection (e.g., vulgarity feature, one of the most significant signals to detect cyberbullying occurrences), these features are inadequately comprehensive because they only focus on important features while ignoring others such as words related to age and gender or directly extracted features such as activity and social network features. For example, the word “stupid” alone may not always indicate cyberbullying occurrences, when other features, such as age-related word for age 1 (school), is used also in the sentence, both features exhibit high probabilities for cyberbullying occurrences. Consequently the reasons for success of features proposed in this research is as it proposes a comprehensive set of features compared with those in previous studies, and, thus, the proposed set outperforms Baseline 2.

Therefore, the comparison outcomes show the cyberbullying detection method using the proposed features show to be more effective cyberbullying detection method compared to the methods using the baselines features (see Figure 4.9).

4.5 Conclusion

This chapter discusses the construction of an effective method for detecting cyberbullying. A set of proposed features that uses features from tweets, such as social network, activity, user, and tweet content, is adopted to construct a machine learning classifier for classifying the tweets as either cyberbullying or non-cyberbullying. An extensive set of experiments were run to measure the performance of the four selected classifiers, namely, NB, LIBSVM, RF, and KNN. Three feature selection algorithms were selected, namely, chi-square test, information gain, and Pearson correlation, to determine the most significant feature. Feature analysis algorithms were applied using different feature combinations, and different numbers of features were iteratively selected to determine a combination with a significant discriminative power that can provide an improved result. Oversampling of the minority (abnormal) class and undersampling of the majority (normal) class (SMOTE) were applied along with weight adjusting approaches (cost-sensitive) to handle the imbalanced class distribution in the manually labeled data set. SMOTE improved the overall performance of the classifiers. Given that the manually labeled data set contains imbalanced class distribution, AUC was used as the main performance measure because of its high robustness for evaluating classifiers. The best overall classifiers performance was achieved by using classifiers that use SMOTE to handle the imbalanced data distribution. RF using SMOTE alone showed the best AUC (0.943) and f-measure (0.936). The comparison between the best results from the proposed features and those from the two baseline features emphasize

the effectiveness of the proposed features and, in turn, that of the effectiveness of cyberbullying detection method using proposed features.

CHAPTER 5: EFFECTIVE INFLUENTIAL SPREADERS IDENTIFICATION METHOD

5.1 Introduction

This chapter presents an effective method for identifying influential spreaders in an OSN. The proposed Interaction Weighted K-core Decomposition method (IWK_S) is developed and evaluated.

Influential spreaders identification is an important subject to control the dynamics of information diffusion in OSNs. Targeting these influential spreaders is significant in hindering the diffusion of unwanted elements, such as cyberbullying, rumors, virus, and online negative behavior. As discussed in chapter 2 the previous studies have introduced methods for identifying influential spreaders. The most applicable algorithm for identifying influential spreaders are degree centrality, PageRank, and k-core. Many studies have used PageRank and its extension to identify influential spreaders in OSNs (W. Chen et al., 2012; Ding et al., 2013; Jabeur et al., 2012; Java et al., 2006; Nguyen & Szymanski, 2013; Tunkelang, 2009; J. Weng et al., 2010; Yamaguchi et al., 2010). However, given the difficulty in extracting the complete network of most OSNs because of ethical and technical reasons, this issue has led to the unavailability of the complete OSN structure. PageRank algorithm is not a reliable measurement for OSNs because the entire network data are required (Pei et al., 2014). The measurements given by PageRank applied to random networks are also responsive to perturbations in network topology (Ghoshal & Barabási, 2011). Thus, PageRank is an unreliable ranking measurement for incomplete or noisy networks. PageRank is frequently used to find influential spreaders based on the hypothesis of random spread of information in a network. In practice, however, information spread is not completely grounded to random walks (Goel et al., 2012). This quality can lead to substantial divergence between PageRank and actual outcomes. A comparative study (Pei et al., 2014)

conducted with large data sets from OSNs demonstrated that supreme influential spreaders are determined with k-core. K-core algorithm performs better than degree centrality and PageRank. The algorithm computes the influence of users more capably and identifies the important super-spreaders more accurately than other methods.

K-core is thus considered more suitable for identifying influential spreaders in OSNs than other algorithms in the current literature review. Generally (referee to the issue of k-core in chapter 2), the major limitation of the k-core method is that it deals with unweighted graphs. Nevertheless, most real networks are weighted, and their weights describe significant properties of underlying systems. To overcome the original k-core algorithm issues related to treating all links equally, the number of connected links to the nodes rather than the quality of links among the nodes should be considered. This research attempts to develop a method for OSNs by proposing a novel link-weighting method based on user interaction. User interaction reflects the link strength, which is an important element for measuring the diffusion ability of a user within OSNs (L. Weng et al., 2012). Proposing an effective method for identifying influential spreaders is crucial to restrain the spread of cybercrime. this can be accomplished by either minimizing the cybercrime spread by blocking spread rumor (Kwon et al., 2013; L. Zhao et al., 2011) or preventing cyberbullying (Nahar et al., 2013) or maximizing user awareness by spreading prevention strategies to a large number of users (e.g., making kind words go viral) (Ang, 2016; Patchin & Hinduja, 2013).

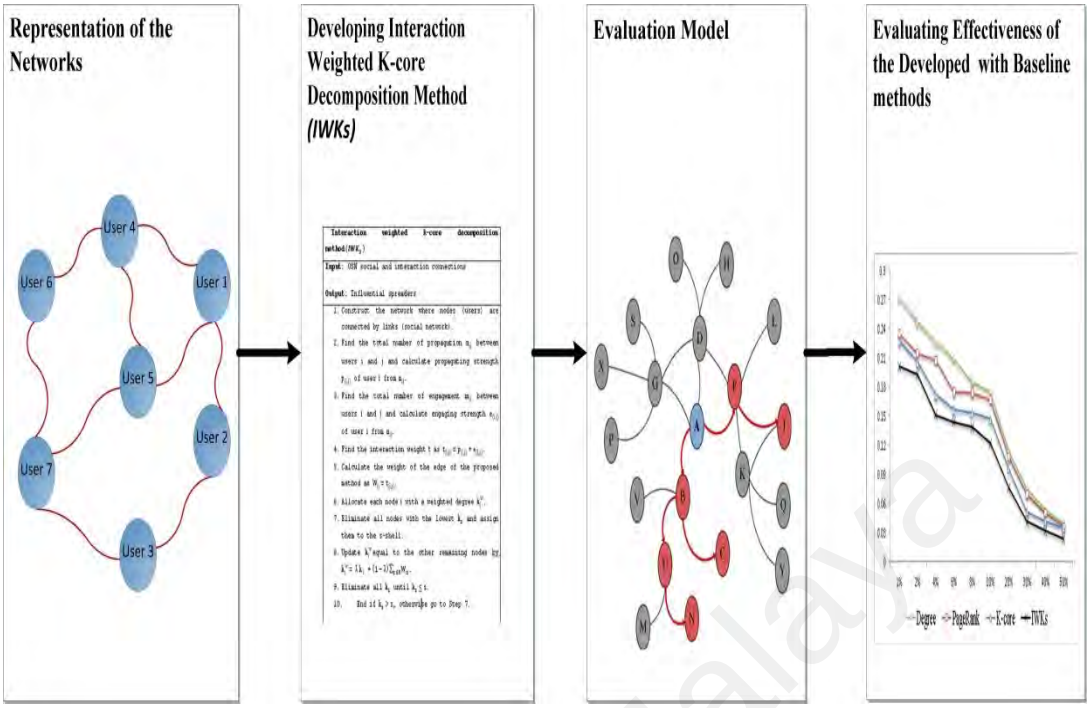

Figure 5.1 : Experimental Processes of Developing and Evaluating the Effective Method for Influential spreaders Identification

Figure 5.1 shows experimental processes of developing and evaluating effective method for influential spreaders identification. The following sections explain each block. Section 5.2 demonstrates the networks representation. Section 5.3 discusses in details describes in detail the developed IWK_S. . Section 5.4 deliberates the evaluation model. Section 5.5 evaluates the effectiveness of the developed method.

5.2 Representation of the Network

Based on data sets mentioned in section 3.3.1 (networks (network 1 , and network 2), social network nodes is used to construct the network, and the number of retweets and mentions corresponding to each user are extracted from the retweet and mention networks. Then, the generated data are used to create weights for the social network.

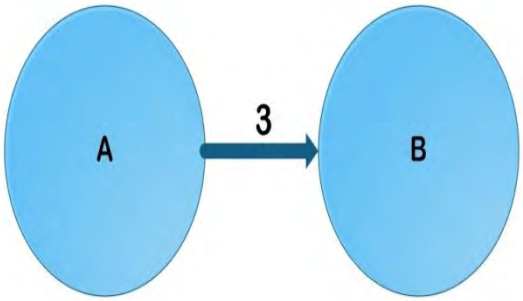

Figure 5.2 : Directed network

For example in Figure 5.2, if user A follows B, A retweets user B, (2 times) and user A mentions user B (1 time). Directed link is created from users A and B with a weight of 3. The whole network is similarly constructed from the network 1 dataset.

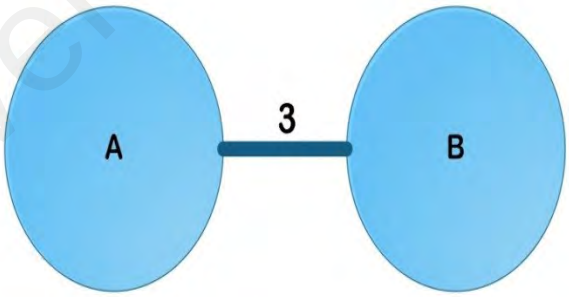

Figure 5.3: Undirected network

For example in Figure 5.3, if users A and B follow each other, they retweet each other (2 times), and mention each other (1 time). The network is constructed in such

way so that an undirected link is created between users A and B with a weight of 3. The whole network is similarly constructed from the network 2 dataset.

5.3 Developing Interaction Weighted K-core Decomposition Method

Original k-core ranking is based on the k-shell decomposition of a network. Each node is assigned a k-shell number, k_s , that is, the order of the shell to which it belongs. In the k-shell decomposition, all nodes with degree $k = 1$ are removed, and pruning processes continue until no node with $k = 1$ remains. Similarly, the pruning processes are applied to the next k-shells. This process continues until the k-core of the network is found (Batagelj & Zaversnik, 2003).

The calculation of the degree of node i in the unweighted k-core can be defined as follows:

$$k_i = \sum_j^N d_{ij}, \tag{5.1}$$

where k_i is the node degree of i, and j is the number of nodes connecting to i. The value of d_{ij} is defined as 1 if node i is connected to node j and 0 otherwise. The degree of a user in OSNs directly indicates the size of the audience for this user.

In this research, the IWK_S is proposed based on the quantity of interactions among users in OSNs. Therefore, the weight of edge of the proposed method is defined as follows:

$$W_{ij} = t_{(i,j)}, \tag{5.2}$$

where t is the interaction weight between nodes.

Interaction t is calculated using two important interaction factors in OSNs that are given as follows:

$$t = p_{(i,j)} + e_{(i,j)}, \tag{5.3}$$

where p is calculated as

$$p_{(i)} = \sum_j^N n_{ij}. \tag{5.4}$$

$p_{(i)}$ represents the propagating strength of the node and measures the propagation of the content. It also indicates the ability of that user to generate content with pass-along value. This quality can be measured using OSN features, which describe the propagation of content, such as retweet in Twitter and share in Facebook. n_{ij} represents the total number of propagated content between i and j. In a directed network, propagating strength is calculated as the total number of propagated content of i by j. In an undirected network, propagating strength is calculated as the total number of propagated content in which i and j propagate each other.

e is calculated as follows:

$$e_{(i)} = \sum_j^N m_{ij}. \tag{5.5}$$

$e_{(i)}$ is the engaging strength of i, and it measures user engagement in conversations. This value indicates the importance of users if they are engaged by others in most conversations, such as mentioning users on Twitter or tagging them on Facebook. m_{ij} represents the total number of engagements between i and j. In a directed network, engaging strength $e_{(i)}$ is calculated as the total number that i was engaged in a conversations by j. In a undirected network, engaging strength $e_{(i)}$ is calculated as the total number that i and j engaged each other in conversations.

Each node is allocated a weighted degree using the following relationship:

$$k_i^w = \lambda \boldsymbol{k_i} + (1-\lambda) \sum_{j \in R} W_{ij,} \quad (5.6)$$

Where R is a set of neighboring nodes of i, and λ is a tunable parameter between 0 and 1. In the present study, set $\lambda = 0.5$, which calculates the link-interaction weights and degree equally. Weighted degrees may not be long integers as they are rounded off to the nearest integer. After preparation, the proposed method applies the same pruning routine as that of the original method.

The detailed decomposition is explained using the following pseudo code (Figure 5.4):-

Developing Interaction Weighted K-core Decomposition Method ($\boldsymbol{IWK_S}$)
Input: OSN social and interaction connections **Output:** Influential spreaders

1. Construct the network where nodes (users) are connected by links (social network).
2. Find the total number of propagation n_{ij} between users i and j and calculate propagating strength $p_{(i,j)}$ of user i from n_{ij}.
3. Find the total number of engagement m_{ij} between users i and j and calculate engaging strength $e_{(i,j)}$ of user i from m_{ij}.
4. Find the interaction weight t as $t_{(i,j)} = p_{(i,j)} + e_{(i,j)}$.
5. Calculate the weight of the edge of the proposed method as $W_{ij} = t_{(i,j)}$.
6. Allocate each node i with a weighted degree k_i^w.
7. Eliminate all nodes with the lowest k_s and assign them to the s-shell.
8. Update k_i^w equal to the other remaining nodes by $k_i^w = \lambda k_i + (1-\lambda)\sum_{s \in R} W_{is}$.
9. Eliminate all k_s until $k_s \leq s$.
10. End if $k_s > s$, otherwise go to Step 7.

Figure 5.4 : Pseudo Code Interaction weighted k-core decomposition method

5.3.1 Difference between the Original K-core and Developed method

The main difference between the k-core and the proposed method lies in calculating the degree of nodes in the network. The original k-core just calculates the degree based on the number of users a social network has (no weight is given to the link; consequently, the degree is only based on the links). Meanwhile, the proposed method calculates the degree of nodes in a network based on the social network and the total interaction between the users (the propagation strength and engagement strength). Consequently, the pruning processes will differ, and the output results of these methods will significantly differ.

To illustrate the function of IWK_S and its difference from the original K-core, the following example is considered:

Table 5.1: Exemplary network

NETWORK EDGE	PROPAGATION $(p_{(i)})$	ENGAGEMENT $(e_{(i)})$
(1,2)	4	3
(1,4)	2	1
(1,7)	1	1
(2,3)	1	1
(2,4)	2	1
(2,7)	2	1
(3,4)	1	1
(3,6)	2	1
(4,6)	1	0
(5,6)	1	1

For example a network as shown in Table 5.1 which consists of 7 users with social connections between these users. The interaction between them represented using propagation strength (i.e. retweet in Twitter) and engagement strength (i.e. such as mention in Twitter).

From Table 5.1, the interaction weight (Total of propagation strength and engagement strength) can be calculated as shown in Table 5.2.

Table 5.2: Network weight

NETWORK EDGE	TOTAL WEIGHT
(1,2)	7
(1,4)	3
(1,7)	2
(2,3)	2
(2,4)	3
(2,7)	3
(3,4)	2
(3,6)	3
(4,6)	1
(5,6)	2

Table 5.2. Shows the social network links between the 7 users as well as the interaction weight between them.

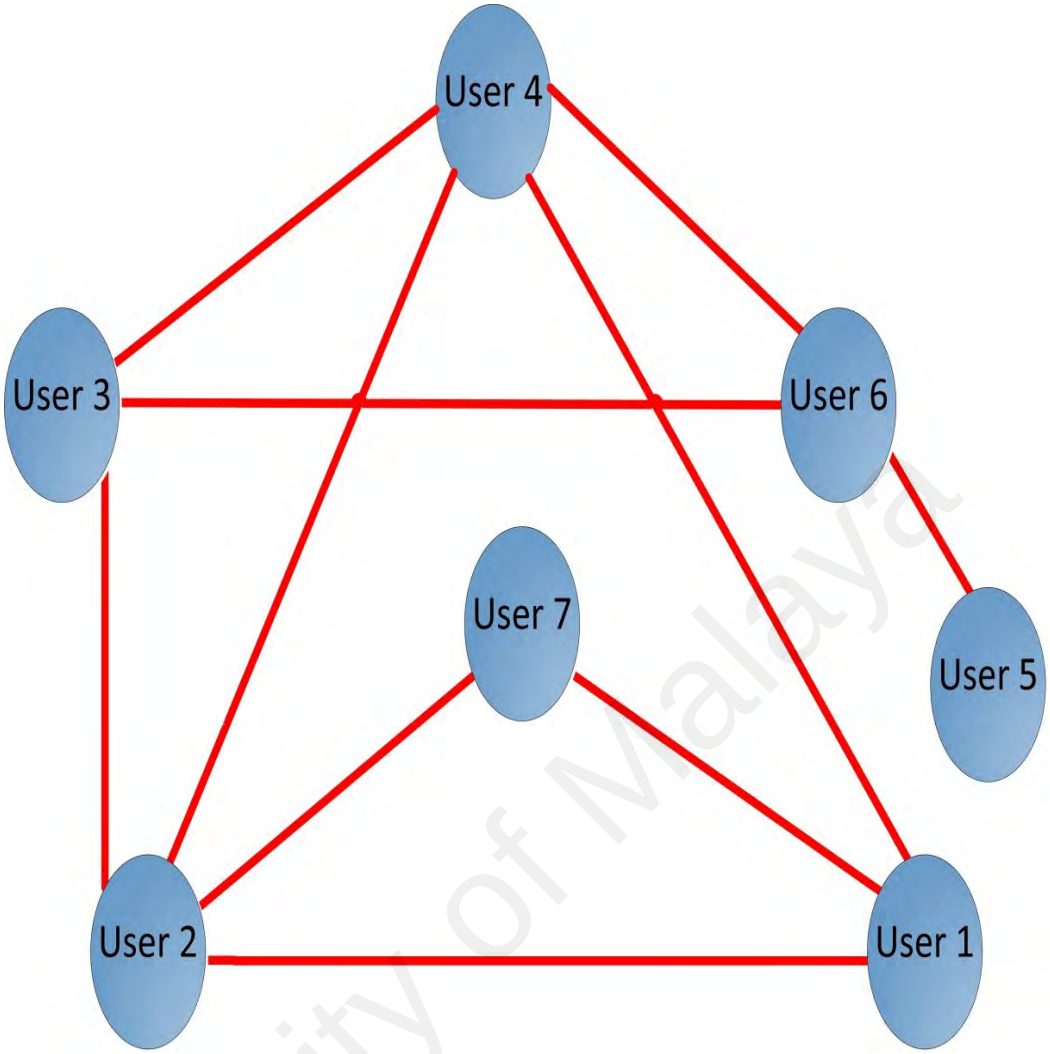

Figure 5.5: Unweighted network

Using network edge only (social network between users) as shown in Table 1, the unweight network is first considered as shown in Figure 5.5. In Figure 5.5 only the social network is used to create the network between the users and no weight is considered consequently the links between users is treated equally in calculating the influential spreaders regardless the link weight.

The weight of the network using interactions is calculated in Table 5.2. is used to create the interaction based on the weighted network as shown in Figure 5.6.

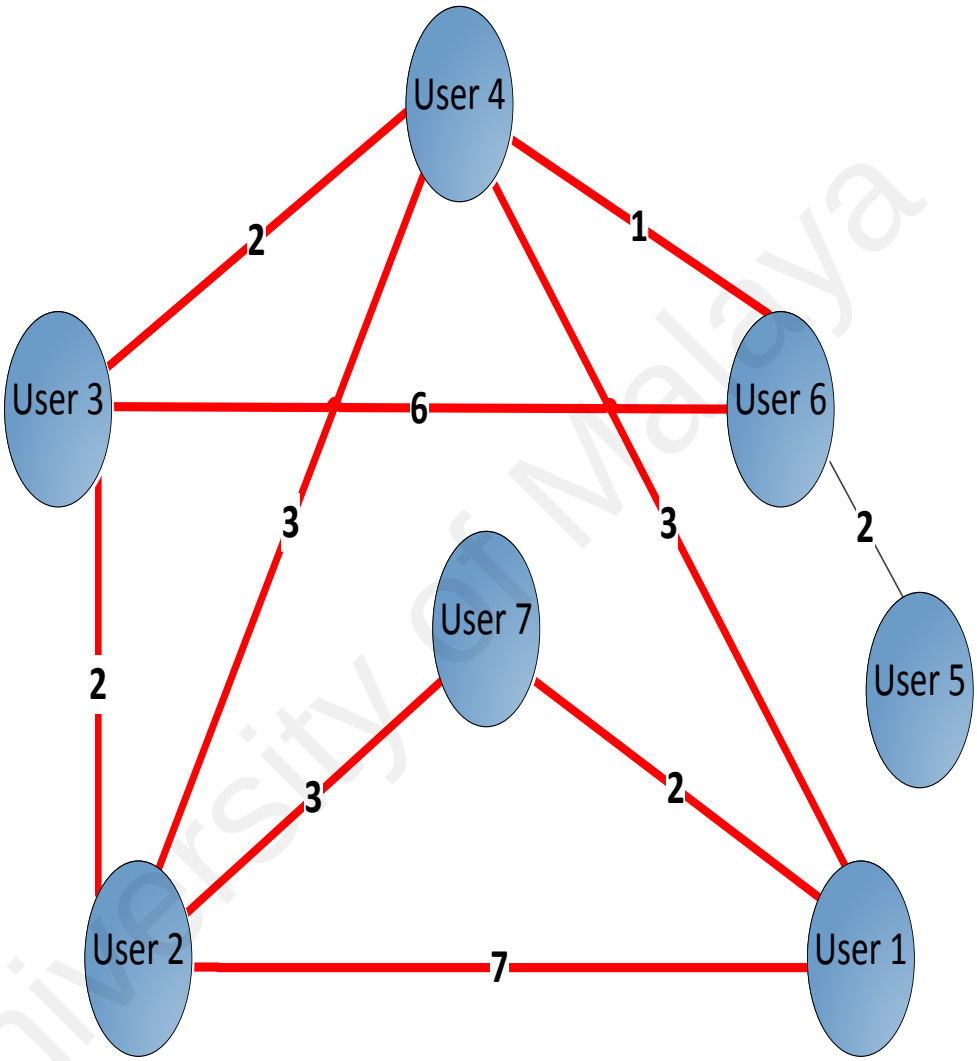

Figure 5.6: Total interaction based on weighted network

In Figure 5.6 interaction weighted network is created. This network is constructed based on interactions social network connections between the users and weighted using interactions between the users.

Applying the original k-core method as described in chapter 2 on above example obtains the following:

$K_S =$ level 1 : user (5)

$K_S =$ level 2: users $(1, 2, 3, 4, 6, 7)$

Applying the developed IWK_S method as described above method obtains the following results:

$K_S =$ level 1 : user (5)

$K_S =$ level 2: users $(3, 6, 7)$

$K_S =$ level 3 : users $(1, 2, 4)$

These three level of IWK_S are obtained as follows

1. Computed the weighted degree of each node using the formula 5.6 and setting lambda as 0.5.

Node Number	1	2	3	4	5	6	7
Weighted Degree	7	9	5	6	1	4	3

2. Removing node 5 as it has the minimum value. Hence, core of 5: 1

3. Updating weighted degree of node 5's neighbors.

    Node: 6

    Weighted Degree: 3

4. Removing node 7 next. Hence, core of 7: 3

5. Updating weighted degree of node 7's neighbors.

    Node: 1

    Weighted Degree: 6

Node: 2

Weighted Degree: 7

6. Removing node 6 next. Hence, core of 6: 3

7. Updating weighted degree of node 6's neighbors.

Node: 4

Weighted Degree: 5

Node: 3

Weighted Degree: 3

8. Removing node 3 next. Hence, core of 3: 3

9. Updating weighted degree of node 3's neighbors.

Node: 4

Weighted Degree: 4

Node: 2

Weighted Degree: 6

10. Removing node 4 next. Hence, core of 4: 4

11. Updating weighted degree of node 4's neighbors.

Node: 2

Weighted Degree: 4

Node: 1

Weighted Degree: 4

12. Removing node 2 next. Hence, core of 2: 4

13. Removing node 1 next. Hence, core of 1: 4

Final Result:

Node Number	1	2	3	4	5	6	7
Core Value	4	4	3	4	1	3	3

The above example shows that the nodes of the network are assigned to only two k-core levels by the original k-core. By contrast, the IWK_S assigns the network nodes divided into three levels. The proposed method clearly assigns the nodes to a larger number of levels compared with the original k-core because it considers both the degree and the interaction weight between nodes. Consequently, for a large network, the proposed method assigns the spreaders to more level unlike the original K-core and better distinguishes the influential spreaders. This example shows the difference working principle of original k-core and developed methods. However, the effectiveness of the proposed method compared to the baselines method (degree centrality, PageRank and original k-core) on large network datasets is evaluated in section 5.5.

5.4 Evaluation Model

Evaluating the methods for identifying influential spreaders is important to ensure the effectiveness of any methods compared with other methods. The proposed method is evaluated using the real dynamics of information diffusion in a real-world social network (Pei et al., 2014). Content is generated by one or few independent sources. Then, users propagate this information and refer to the source. Accordingly, that information is propagated to their followers. This process is usually observed in many diffusion networks (Kleinberg, 2007; Watts, 2002). Considering this process, diffusion link initiated from each node (user) i in the Twitter network is followed. The first-layer nodes that have diffused node i's information to their followers are also recognized. The dissemination links initiated from these nodes are followed until a complete diffusion cascading is recovered, similar to the concept proposed in a previous study (Pei et al., 2014). The subsequent set of users signifies the region of influence for user i. The

influence of user i to the diffusion is measured as the number of the users in the region of influence, and this quantity is denoted as M_i. A breadth-first search is used to track the diffusion links by layers. To remove the effect of loops, only newly covered users are placed in the search queue from one layer to the next layer (Pei et al., 2014). Finally, M_i becomes the overall influence for all the posts of node i. The overall spreading efficiency of each user is calculated, and a ranking list of the users (nodes) is generated (Pei et al., 2014). For example, Twitter users can retweet the posts by other users, and these users can refer to the original post through the retweeting feature. To obtain the diffusion network, retweet relations are extracted from the tweets. A retweet (RT @username) corresponds to post propagated from the main source to other users. A user can retweet other users from second level (follower of the user's follower) without having a following relation, and the post reaches them through first-level followers. Such kind of activity cascades information among different levels, thereby creating region of influence. Although the diffused post may be affected as it propagates among the users in the region of influence, the first user is treated to be in authority to the complete cascading (Pei et al., 2014). Accordingly, the information spread can be directly followed from one user to another. A breadth-first search is used to track the diffusion links by levels. For example, if user B retweets a post by user A, then the content diffuses from A to B; if user C propagates a content of user A that has been already retweeted by user B, then the content diffuses from A to C through B. As a result, the information is cascaded from users to other users at different levels. In this way, the diffusion network representing diffusion of information in Twitter is obtained (Pei et al., 2014). Figure 5.7 demonstrates that post diffusion initiates from source user A to six other users (users who propagated the content of user A at different levels). The search escalates through three levels, and the influence region of A is $M_A = 6$.

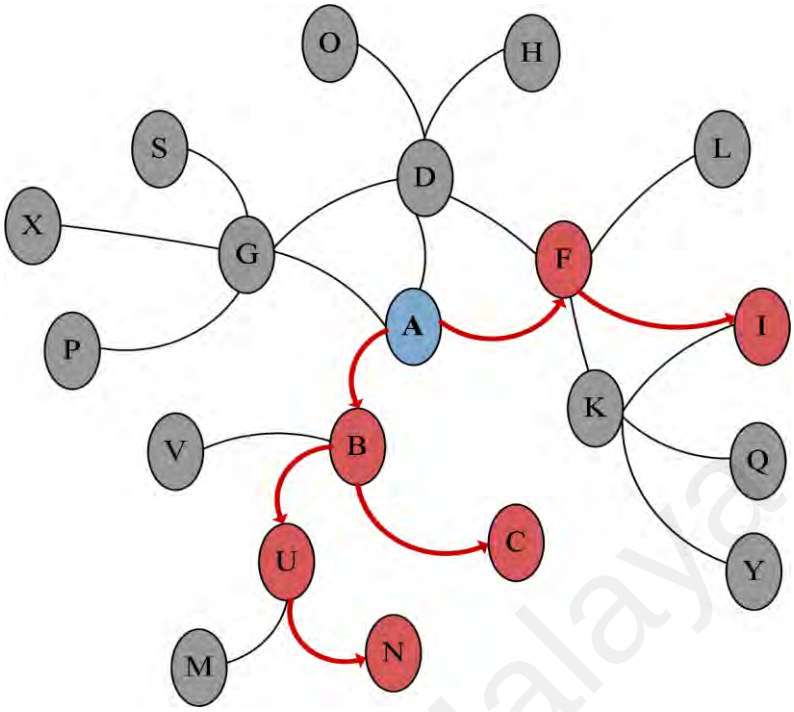

Figure 5.7: Influence of nodes from source node *A*

To obtain the diffusion graph for the data sets used in this study, the retweet network related to all the users in the social network is utilized. The retweet network is the best illustrative network that can describe content propagation (De Domenico et al., 2013). In the retweet network, if user B retweets a tweet of user A, information spreads from A to B, thus creating a diffusion link from A to$B$. In this way, the diffusion of the networks is created. The overall spreading efficiency of each node user is calculated, and a ranking list of users is generated.

5.5 Effectiveness of the Developed Method

The effectiveness of a user initiated from i is quantified through the amount of users in the influence region. The overall spreading effectiveness of each node user is calculated, and a ranking list of users is generated using evaluation model discussed in above section. Effectiveness of a user initiated from i is calculated through the amount of users in the influence region, and this quantity is denoted as M_i. To evaluate which algorithm is more accurate for calculating the diffusion capability of nodes, degree

centrality, PageRank, original K-core, and developed IWK_S are compared by calculating their respective imprecision functions, ϵ_k, ϵ_{PR} , ϵ_{k_s}, and ϵ_{IWk_s} as proposed in a previous study (Kitsak et al., 2010). The imprecision function of degree (ϵ_k) is calculated as

$$\epsilon_k(p) = 1 - \frac{M_k(p)}{M_{eff}(p)}, \quad (5.7)$$

Similarly, the imprecision functions of PageRank, original k-core, and IWK_S (ϵ_{PR}, ϵ_{k_s}, and ϵ_{IWk}, respectively) are calculated as follows:

$$\epsilon_{PR}(p) = 1 - \frac{M_{PR}(p)}{M_{eff}(p)}, \quad (5.8)$$

$$\epsilon_{k_s}(p) = 1 - \frac{M_{k_S}(p)}{M_{eff}(p)}, \quad (5.9)$$

$$\epsilon_{IWk_s}(p) = 1 - \frac{M_{IWk}(p)}{M_{eff}(p)} \quad (5.10)$$

where p is the fraction of network size $N(p \in [1,0])$, $M_{(k)(pr)(k_s)(IWk_s)}(p)$ is the average spreading effectiveness of top fraction of network nodes with the highest (degree centrality, PageRank, k-core, and IWK_S) values, and $M_{eff}(p)$ is the average spreading effectiveness of top fraction of network nodes with the largest spreading effectiveness calculated using diffusion graph. The more accurate the algorithm, the

smaller the imprecision function (ϵ) value. Imprecision function values close to 0 indicate high diffusion effectiveness given that the users selected are mainly those who contribute the most to information dissemination.

The imprecision functions of the two networks are presented in Figures 5.8 and 5.9. The imprecision function of these methods for the top network fractions is compared. The top 10% are the most important users from many applications. When a network must be immunized, the detection method is implemented at the minimum number of nodes, and the comparison is extended to the top 50% to make results conclusive. Hence, (1%, 2%, 4%, 6%, 10%, 20%, 30%, 40%, and 50%) top users are considered.

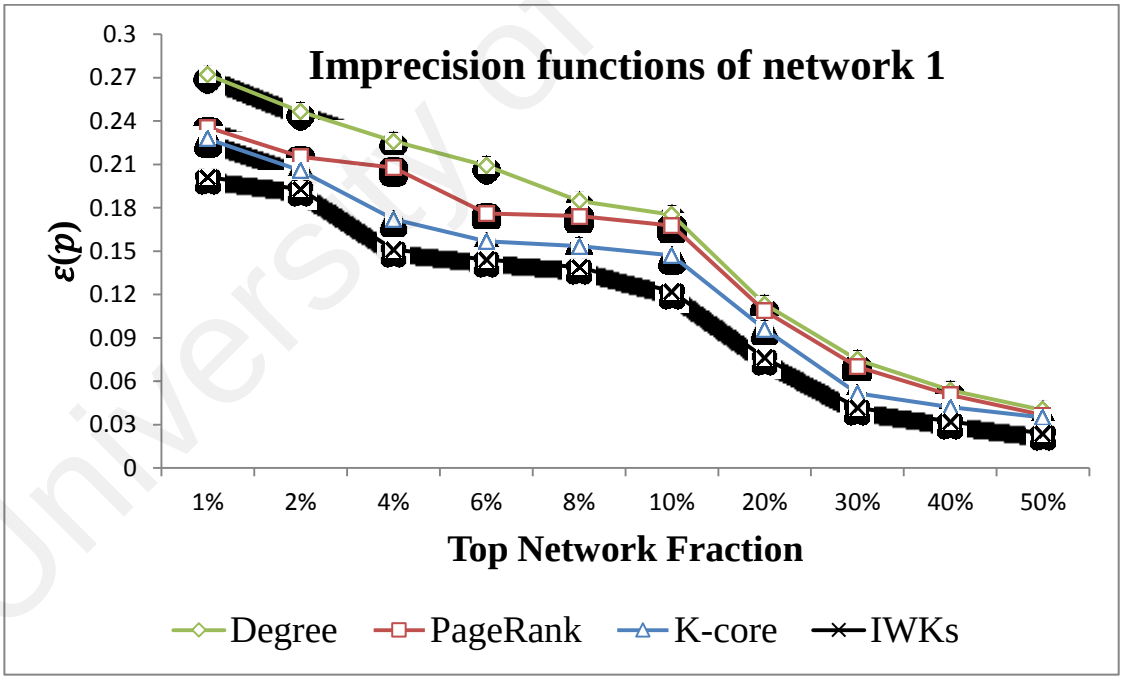

Figure 5.8: Imprecision functions of degree centrality, PageRank, k-core, and IWK_S for network 1

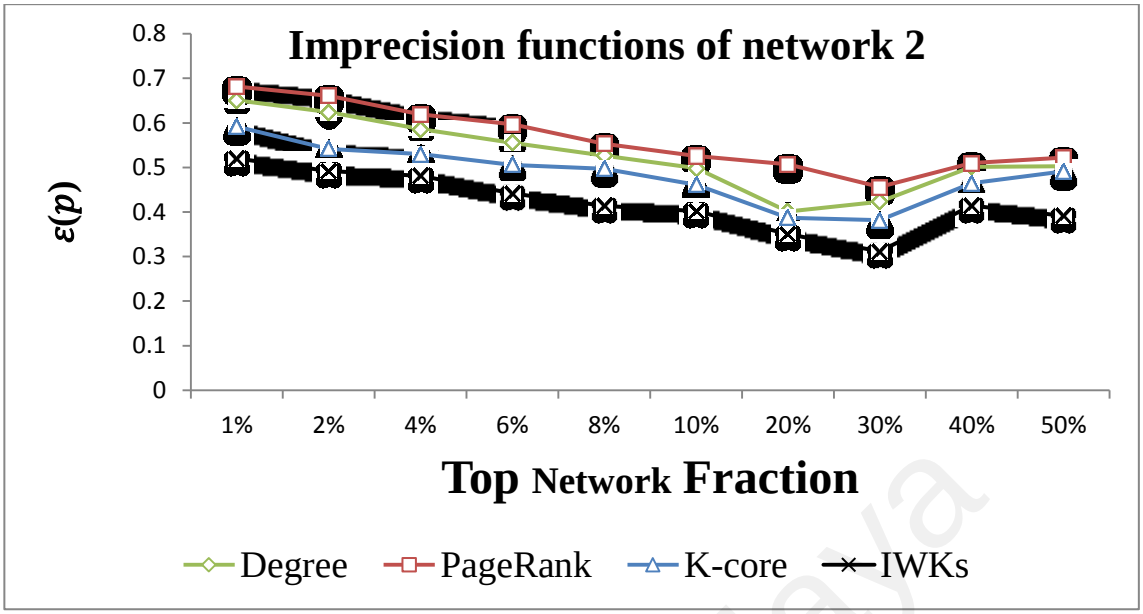

Figure 5.9: Imprecision functions of degree centrality, PageRank, k-core, and IWK_S for network 2

Figures 5.8 and 5.9 (Imprecision functions of degree centrality, PageRank, k-core, and IWK_S for network 1 and network 2 respectively) show that in all cases, the imprecision of IWK_S is lower than those of other algorithms are. This finding indicates that the IWK_S can identify the spreading effectiveness of nodes better than degree centrality, PageRank, and original k-core. The IWK_S improves the identification accuracy by considering the amount of interaction among nodes and quantifying the spreading efficiency of an individual more effectively.

The IWK_s can effectively quantify the diffusion effectiveness well. However, which algorithm can better detect separate influential spreaders remains uncertain. Therefore, recognition rate $r(f)$ is used to evaluate the performance of each algorithm in identifying the influential spreaders as proposed by Pei et al. (Pei et al., 2014). Recognition rate $r(f)$ is calculated as follows:

$$r(f) = \frac{|I_f \cap P_f|,}{|I_f|} \qquad (5.11)$$

where I_f and P_f pertain to the ranking lists in the top f fraction obtained by tracking diffusion links in real spreading dynamics (node influence) and obtained by algorithms (degree centrality, PageRank, K-core, and IWK_s), respectively. The top network fractions (1%, 2%, 4%, 6%, 10%, 20%, 30%, 40%, and 50%) are compared. Figures 5.10 and 5.11 illustrate the recognition rate and indicate that IWK_s obtains the largest recognition rate among degree centrality, PageRank, and k-core.

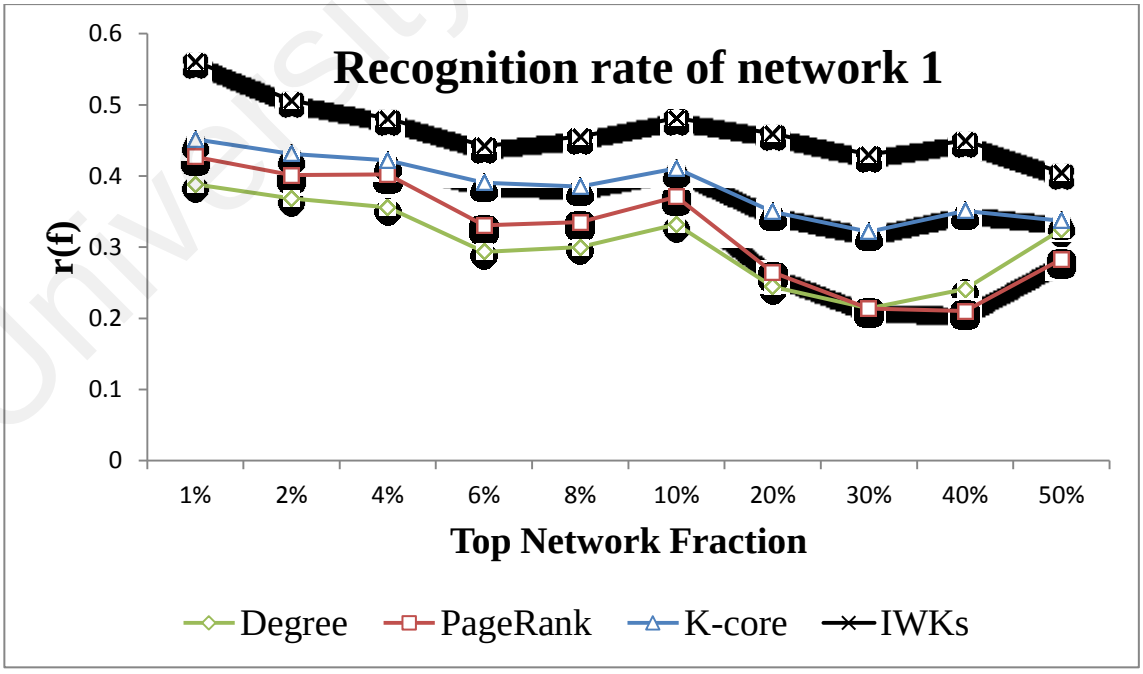

Figure 5.10: Recognition rate *r(f)* for network 1

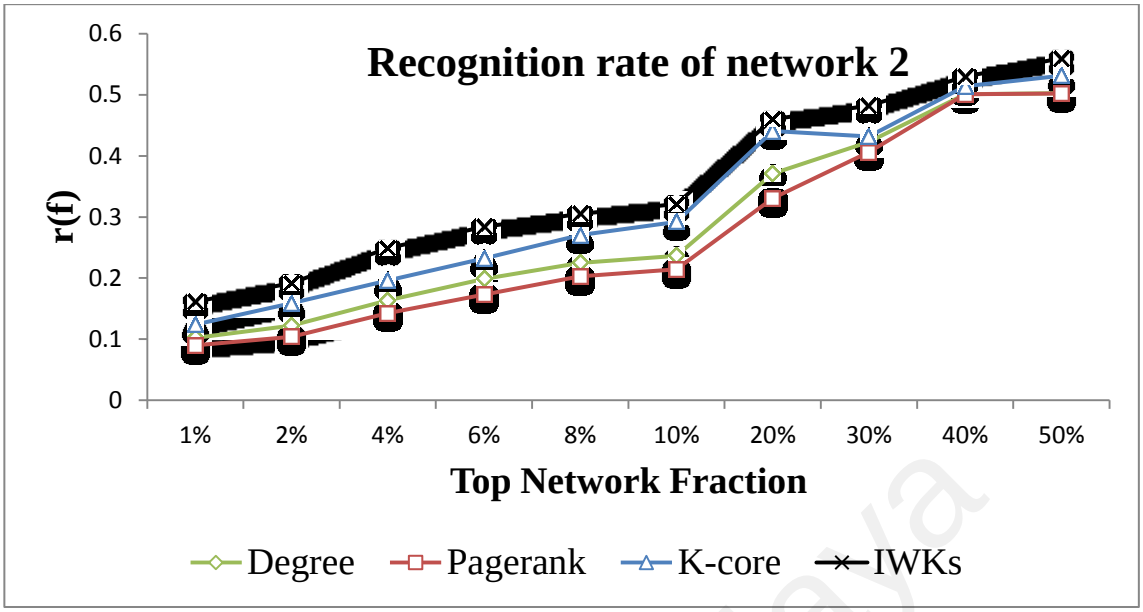

Figure 5.11: Recognition rate *r(f)* for network 2

Figures 5.10 and 5.11 (Recognition rate of degree centrality, PageRank, k-core, and $\boldsymbol{IWK_S}$ for network 1 and network 2 respectively) show that, In both networks, the $\boldsymbol{IWK_S}$ can recognize the influential spreaders more effectively than degree centrality, PageRank, and k-core. The better performance of $\boldsymbol{IWK_S}$ in a directed network results from user interaction that represents the spread of information better than the interaction among users in an undirected network.

As reported by the experimental results, considering the interaction among users with the weighted k-core processes increases the accuracy of identifying the most influential spreaders. The IWK_S is more accurate in calculating node influence than the original K-core. This result can be explained by the weighted k-core biasing the pruning process based on the degrees of the nodes as well as the interaction of the nodes with other nodes. $\boldsymbol{IWK_S}$ eliminates the original k-core limitation, which considers all links of users equally whether active or not. Although the experiment does not investigate whether these links play important roles in spreading information, the results show the relationship between the interaction factors and the spreading behavior of OSNs.

From the experiment results, PageRank is unsuccessful in detecting influential spreaders compared with the proposed method. Both networks present partial network, and PageRank is responsive to changes in network representation, rendering it unreliable for incomplete or noisy networks (Ghoshal & Barabási, 2011). The complete OSN structure is unavailable because of the inherent limitations of OSNs caused by API restrictions, user privacy, and network dynamics. Therefore, PageRank is an unreliable method for an OSN with such characteristics. The success of PageRank in web network is due to the unpremeditated result of the scale-free nature of the web graph (Ghoshal & Barabási, 2011). If the web graph were an exponential network, the ranking generated by PageRank would have been unreliable given the incompleteness of the web graph (Ghoshal & Barabási, 2011)

Methods such as degree centrality usually highly rank rich-club hubs (Morone & Makse, 2015). In complex networks, the most connected nodes are typically considered responsible for the largest information dissemination and are viewed as the most influential nodes (Albert et al., 2000).. However, reasonable situations exist in which the influential spreaders do not correspond to the most highly connected users (Kitsak et al., 2010). The failure of degree centrality in this experiment compared to with the proposed method is because the local features of nodes (number of connections) are not constantly represented by the spreading efficiency of nodes in the network. The location of users within the network along with the diffusion efficiency of their connected users plays a major role in the dissemination of information in OSNs. These factors cannot be captured by degree centrality, which simply represents the local connection features of users.

K-core measures the spreading efficiency of users more effectively than degree centrality and PageRank. K-core defines the most influential nodes as those that are

located within the core of the network, and they can be successfully identified by the k-core decomposition method. Comparing original k-core with degree centrality and PageRank reconfirms that k-core performs better. This result is consistent with those of previous studies (Kitsak et al., 2010; Pei et al., 2014).

The limitations related to the k-core decomposition, such as considering the links equally regardless strength of links between the users, result in the original k-core falling behind the proposed method. The most influential spreaders in the network are connected to many users with low k_s values who are removed at the beginning. Therefore, k-core cannot detect these users. These influential spreaders are detected when the interaction between the users are considered to bias the pruning processes of k-core.

5.6 Conclusion

This study develops the IWK_S by introducing a novel-weighting scheme that uses the quantity of interaction among users. The aim of this weighting technique is grounded on the observation that the interaction among users exhibit significant features that can measure the spreading efficiency of a user in OSNs (L. Weng et al., 2012). To evaluate the effectiveness of the proposed method in the spreading capability of nodes, the degree centrality, PageRank, original K-core, and IWK_S are compared by calculating their respective imprecision functions, ϵ_k, ϵ_{PR} , ϵ_{k_s}, and ϵ_{IWk_s} as proposed by Kitsak et al. (Kitsak et al., 2010). Recognition rate $r(f)$ is also calculated for all methods in two networks to verify the performance of each algorithm in recognizing influential spreaders. The developed method achieves the best performance in both networks. The IWK_S performs better than other methods in identifying the most influential spreaders and in quantifying the spreading effectiveness of nodes.

CHAPTER 6: CONCLUSION

6.1 Reappraisal of the Research Objectives

This chapter presents the conclusions of this thesis and discusses the potential future directions. This thesis is concluded by revisiting the research objectives presented in Chapter 1 and describing how they are achieved. The contributions of this thesis are also presented, along with the discussion on the limitations and future research directions.

This thesis intends to accomplish four objectives. In this section, these objectives are revisited to discuss how they are achieved. The objective and how they achieved are schematically mapped in Figure 6.1.

The first objective is to propose a set of significant features, which can provide discriminative power to improve classifier performance for detecting cyberbullying in OSNs. To achieve this objective, a set of features extracted based on network information, activity information, user information, and tweet content are presented, as discussed in Chapter 4 (Section 4.3). These features are extracted and used as inputs for different machine learning classification algorithms.

The second objective is to construct an effective method for detecting cyberbullying in OSNs based on the proposed features. To achieve this objective, an extensive set of experiments are ran to construct cyberbullying detection method based on the proposed features under four different experiment setups using four classifiers (i.e., NB, LIBSVM, RF, and KNN) to select best setting for the proposed features. All four classifiers are tested in four different settings, namely, basic classifiers, classifiers with feature selection techniques, classifiers with SMOTE alone and with feature selection techniques, and classifiers with cost-sensitive alone and with feature selection technique. .The constructed cyberbullying detection method based on the proposed

features obtained an area under the receiver operating characteristic (ROC) curve (AUC) of 0.943 and an f-measure of 0.936 using RF with SMOTE.

The second objective is exploratory (investigative) objective. Applying machine learning for any field is not direct task; it requires deep investigating of the stability of machine learning algorithm also understanding the data nature and evaluation metric in order to build reliable detection based on machine learning methods. This task is where the efforts of most researcher of machine learning are spent, (to explore and investigate). The achievement of this objective (objective 2) started with extensive literature review to understand the previous research and previous machine learning algorithms applied to detect cyberbullying, secondly after robust literature review, the most commonly effective method from the literature are selected to be investigated. Thirdly understanding the nature the data in the research is compulsory to setup the experiment and to avoid the overfitting of the data. Lastly, the evaluation metric should be carefully selected in order to have reliable results. The whole process differ from field to field consequently, the contribution of this objective (objective 2) is the investigation of these algorithms and techniques for cyberbullying detection using the proposed features from objective 1 to construct effective method for cyberbullying detection.

The third objective is to develop effective method for identifying influential spreaders in OSNs. To achieve this objective, the IWK_S is developed. This method proposes a novel link-weighting method based on user interaction. The developed method is compared with degree centrality, PageRank, and original K-core using large real networks from Twitter.

The fourth objective is to evaluate the effectiveness of above methods by comparing them with the baseline methods using real data sets. This objective is divided into two

parts: to evaluate the effectiveness of the proposed method for detecting cyberbullying based on proposed features and that of the proposed method for identifying influential spreaders in OSNs. In order to achieve the first part, results obtained using the proposed features are compared with the results obtained from two baseline features using real data set from Twitter. The comparison outcomes using AUC as the evaluation metric confirmed the effectiveness of the constructing cyberbullying detection method based on proposed features compared with constructing cyberbullying detection methods based on baseline features. Comparison outcomes show that the proposed features has provided approximately 22% improvement in AUC values of cyberbullying detection method compared to cyberbullying detection method using baselines features (refer Table 4.8). To achieve the second part, the developed methodIWK_S, is evaluated using a ranking list obtained by tracking diffusion links in the real dynamics of information spread explained in Chapter 5 (Section 5.5) and using real data set from Twitter. The Imprecision function and recognition rate of four methods (IWK_S, degree centrality, PageRank, and original k-core) are compared for the top network fractions (1%, 2%, 4%, 6%, 10%, 20%, 30%, 40%, and 50%). The results verify the effectiveness of the developed method for identifying influential spreaders in OSNs compared with degree centrality, PageRank, and original k-core.

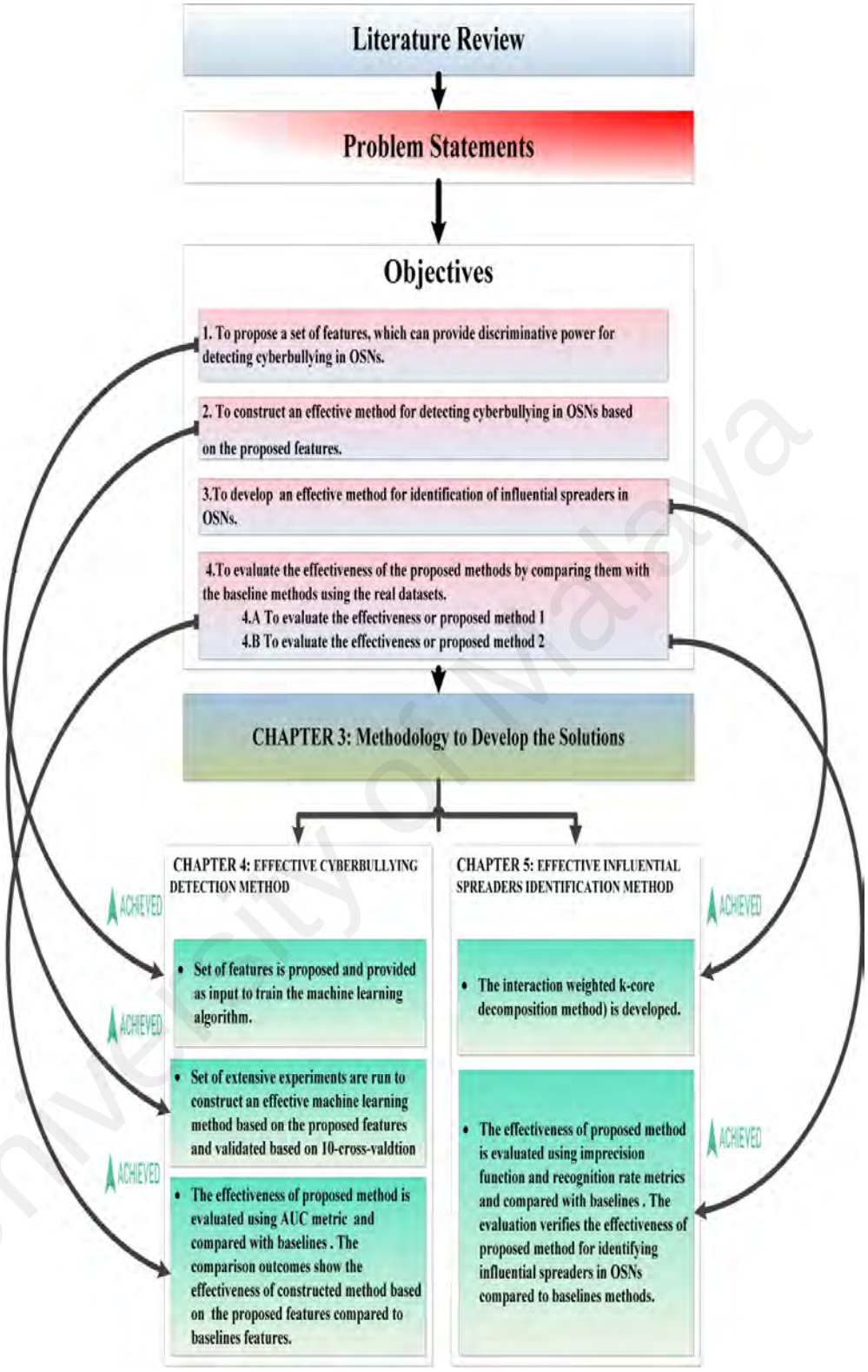

Figure 6.1: Schematic mapping of the objectives

6.2 Contributions of the Research

This research contributes to the body of knowledge in the following aspects:

- **Comprehensive literature review:** The conducted literature review exposes the limitations of the existing techniques related to cyberbullying detection and influential spreader identification. Content analysis methods for detecting cyberbullying in OSNs are comprehensively reviewed and compared, and related issues are identified and investigated. Network analysis-based methods are also comprehensively reviewed and compared with social network analysis methods for identifying influential spreaders. Issues arising from these reviewed studied are identified and investigated.
- **Effective method was constructed for detecting cyberbullying:** This research contributes to the body of knowledge by proposing a set of features extracted as input of machine learning algorithms for detecting cyberbullying. Results indicate that the constructed cyberbullying detection method based on the proposed features provides a feasible solution to detecting cyberbullying in an OSN.
- **Effective method was developed for identifying influential spreaders:** This research contributes to the body of knowledge by proposing the IWK_S drawn on a novel link-weighting method based on user interaction for the effective identification of influential spreaders in an OSN.

All proposed methods in this thesis are published in reputable journals. (Refer to pg. 177 for list of publications)

6.3 Limitation and Future Research Directions

6.3.1 Human Data Characteristics

Individual behavior and mood are an affective state that is important for physical and emotional wellbeing, creativity, and decision making (Golder & Macy, 2011; Ruths & Pfeffer, 2014). Using Twitter API, a study (Golder & Macy, 2011) analyzed the effects of hourly, daily, and seasonal changes at individual levels. The authors (Golder &

Macy, 2011) investigated positive affect (PA), such as enthusiasm, delight, and activeness, and negative affect (NA), such as distress, fear, anger, guilt, and disgust. The positive affect changes with seasonal variation (Golder & Macy, 2011). Similarly, an interesting future research area is investigating how the seasonal variation of users' mood and psychological condition during a year can affect the language used to exhibit cyberbullying behavior and if this change can occur, how can it affect the accuracy of machine learning detection. Consequently in future Collecting long-term data will enable machine learning algorithms to be trained using various human behavior data with the psychological conditions of different users to better detect cyberbullying behavior in OSNs .

6.3.2 Language Dynamics

Language is quickly changing, particularly among young generation. New slang is regularly integrated into the language culture. Therefore, researchers are invited to propose dynamic algorithms to detect new slang and abbreviations related to cyberbullying behavior in OSNs and keep updating training processes of machine learning algorithms using these newly introduced words.

6.3.3 Detection of Cyberbullying Severity

The level of cyberbullying severity should be determined. The effect of cyberbullying is proportional to its severity and spread. Detecting different levels of cyberbullying severity does not only require machine learning understanding but also comprehensive investigation to define and categorize the level of cyberbullying severity from social and psychological perceptions. Efforts from different disciplines are required to define and identify the levels of severity and then introduce related factors that can be converted into features to build multiclassifier machine learning for

classifying cyberbullying severity into different levels as oppose to binary classifier that only detects whether an instance is cyberbullying or not.

6.3.4 Unsupervised Machine Learning and Deep Learning

Human learning is essentially unsupervised. The structure of the world is discovered by observing it and not by being told the name of every objective. Nevertheless, unsupervised machine learning has been overshadowed by the successes of supervised learning (LeCun, Bengio, & Hinton, 2015). This gap in the literature may be because nearly all current studies rely on manually labeled data as input to the supervised algorithm for classifying the classes, and, thus, finding patterns between the two classes using unsupervised grouping remains difficult. Intensive research is required to develop unsupervised algorithm that can detect effective patterns from data. In addition, deep learning has recently attracted attention of many researchers in different fields. Natural language understanding is a new area in which deep learning is poised to make a large effect over the next few years (LeCun et al., 2015).

6.3.5 Multilayer Network

In this study, social network is used as a single network and weighted with retweet and mention information of the same user ID in social network. However, users in OSNs can communicate and interact without having a social link between them. Thus, some links existing in an interaction network may not exist on other interaction networks. This phenomenon leads to the partial representation of the entire interaction among users. Future works must intensively analyze the multilayer network to comprehensively understand the communication and interaction among users and develop accurate algorithms for identifying information diffusion and influential spreaders.

6.3.6 Understanding the Role of Influential Spreaders in OSNs

The current literature of influential spreaders are based on the assumption that targeting the most influential spreaders is a key factor in accelerating information spread and slowing down misinformation spread. The common characteristic of these identified influential spreaders is their strong connection. However, the spread of information can be derived not only by influential spreaders but also by a critical mass of easily influenced individuals. Therefore, further investigation is required to gain an increased understanding of the role of each user in a network, the characteristics of individual users, and the interplay between weakly connected users and influential spreaders in OSNs.

6.3.7 Network data availability

Most previous studies have analyzed OSN graphs with only partial network data. Given the privacy regulations imposed by OSNs, obtaining data from the entire network is difficult; crawling millions of users from an OSN is challenging as well [154]. Most OSNs are active, in which the users are free to customize their profiles and pages with high flexibility, thereby complicating the design; thus, developing crawlers that can efficiently handle these dynamic complex networks is difficult [155]. OSNs allow users to adjust their privacy settings, and some users prefer to keep their profiles private to be seen only by their friends; therefore, such profiles become black holes for crawlers. Most previous studies do not explain how this limitation affects their observations and results [155].

For a future OSN crawler to be effective, it must consider four issues, namely, crawler efficiency, which can be measured by how rapidly the node and links are visited; the bias of some crawling algorithms (i.e., breadth-first-search crawling algorithms) toward large-scale networks [156]; the influence of black holes on the

crawling process [155]; and the distinct properties of OSNs despite their provision of similar services.

6.3.8 Connection diversity

As discussed in section in literature, the explicit and implicit connections in OSNs induce connection diversity. Furthermore, relationship strength is one of the most important factors that affect information diffusion and consequently, the level of influence [157]. The relationship strength widely diverges, ranging from strong ties (i.e., best friend) to weak ties (i.e., acquaintances) [158]. However, in OSNs, the lack of knowledge on link strength between the users can result in networks with heterogeneous relationship strengths (e.g., acquaintances and best friends mixed together) [159]. Therefore, the binary relationship (the relationship that describes only if the relationship exists without considering its strength) will generate dubious relationship information representation, and consequently, deceptive identification results. Studies have examined modeling relationship strength in OSNs [50, 157-159]. However, these studies have discussed how tie strength and connection diversity affect information diffusion rather than deeply considering how these two factors affect user influence measurement, as well as how tie strength can be used in influential spreader identification techniques.

In network theory, nodes are commonly assumed to be linked by a single type of static edge that describes the relationship between them, although in numerous circumstances, this hypothesis simplifies the complexity of the network. Ignoring the reality of multiple relationships between users or combining such relationships to a single weighted network alters the topological and dynamic properties of the entire system [160, 161], as well as the importance of the nodes with respect to the entire structure [162, 163]. Consequently, this idea induces the wrong identification of the

most important nodes [161]. Therefore, multiple relationships between users should be considered for accurate identification.

6.3.9 Network evolution

Network anatomy is important to thoroughly analyze a network because the structure of the network affects function [151]. For example, the structure of OSNs, that is, how the users are connected with one another, affects the spread of information. However, one of the most inherent difficulties in understanding the structure of a network is network evolution. OSNs are dynamic and evolve with time axis; nodes and links are created and deleted every minute. Users in OSNs tend to build their online communication network based on several factors, such as mutual acquaintances, proximity, common interests, and their combinations [164]. The analysis of the effect of these factors on an evolving OSN has demonstrated that preferential attachment was able to capture the evolution of the network and that its effect varied based on node age (i.e., user account age) [164]. Investigating the network evolution will help predict the spread of information in advance; therefore, accelerating or slowing down the information as required is possible [165]. However, a deeper understanding is required on how OSNs evolve with the time axis, which factors are responsible for these evolutions, and how these factors can affect the spreader's influence measurement.

6.3.10 Efficiency of Identification Algorithm-related Issues

The efficiency of the applied algorithm is a crucial success factor for its applicability in a real-world context [166]. As discussed in a previous sections, existing studies that aim to identify and target the most influential spreader suffer from either the computational time (e.g., greedy approaches [88, 104, 106, 107]) or the result quality (e.g., heuristic approaches [108-110]). Therefore, the measurement algorithms should be selected based on the application requirements. For example, if an application requires

the rapid identification of influential spreaders in real time and the result quality is less of a priority, then a heuristic algorithm is more suitable than a greedy algorithm.

However, previous studies faced a number of issues related to the efficiency of the influential spreaders measures. Starting with the processing of a large amount of unstructured and incomplete network data subsequently required efficient algorithms compared with traditional social networks. Furthermore, given that these studies used incomplete network data, number efficiency issues would be raised, and proving the efficiency of their approaches would be difficult for researchers [167]. Moreover, given the various characteristics of OSNs and the data collection limitations, a different source of bias will exist in the identification of influential spreaders, such as selection bias and bias by homophily or assortativity in the networks [168]. Therefore, the algorithms from social network analysis or traditional web pages must be optimized to be accurately applied to the OSN context [13], [169]. Consequently, additional investigation is required to overcome these issues and accomplish a better perception in research and practice.

6.3.11 Validation-related Issues

A key drawback of previous studies is that most of the proposed identification algorithms have been validated through the information spread models and not by studying the dynamics of real information spread. These models, such as susceptible-infectious-recovered (SIR) [9], [170], susceptible-infectious-susceptible (SIS) [99], rumor spreading models [100], and random walks for PageRank [171] have actively simulated information spread. Studies concluded that such models failed to generate an accurate diffusion pattern [8, 98]. This conclusion has explained the intensive arguments in previous research regarding the approaches that most effectively measure user influence. These studies yielded inconsistent outcomes according to the specific

model used to illustrate the information diffusion process [8, 98]. The models inspired by the spread of a contagious disease [99] are proposed based on the basic hypothesis of human behavior that could not be illustrative and representative of the real dynamics of information diffusion [8, 98]. Centola *et al* & Singh *et a l*[101, 103] tracked real diffusion processes and demonstrated that the spread of diseases and spread of information differed. Recent research [98] has reported three possible factors affecting the contagion of information, namely, human behavior [172, 173], homophily [174], and social reinforcement [173]. Subsequently, analyzing human behaviors related to information diffusion in OSNs is vital for many applications. Further investigation is required to enhance the understanding of how these factors can affect the results of information diffusion and consequently measure the most influential spreaders. This aspect may result in the proposal of a new model for information diffusion in OSNs, which is more representative and illustrative of the real dynamics of information spread. Furthermore, it may introduce a new generation of information spread models.

6.3.12 User Privacy-related Issues

The privacy rules and regulations of users imposed by OSN operators have raised one of the most vital issues in this area, that is, the lack of benchmark content datasets. Consequently, almost all the studies on the construction of cyberbullying detection methods have used their own collected dataset.

LIST OF PUBLICATIONS, PAPERS PRESENTED AND ACHIEVEMENTS

Full Article Publication (ISI) Articles

1. **Al-garadi, Mohammed Ali**, Kasturi Dewi Varathan, and Sri Devi Ravana. "Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network." **Published in Computers in Human Behavior** 63 (2016): 433-443. **Q1 (Tier 1) Impact Factor (3.4).**
2. **Al-garadi, Mohammed Ali,** Kasturi Dewi Varathan, and Sri Devi Ravana. "Identification of influential spreaders in online social networks using interaction weighted K-core decomposition method." **Published** in **Physica A: Statistical Mechanics and its Applications** 468 (2017): 278-288. **(Tier 1) Impact Factor (2.24).**
3. **Al-garadi, Mohammed Ali**, Kasturi Dewi Varathan, and Sri Devi Ravana. "Analyzing online social network connection for Identification of Influential Users: Survey and Open Research Issues". "**Received revision"** doing the revisions " in **ACM Computing Surveys Q1 Impact Factor (6.7).**
4. **Al-garadi, Mohammed Ali**, et al. "Using online social networks to track a pandemic: A systematic review." **Published** in **JBI** 62 (2016): 1-11. **Best Paper in Volume 62"Editors' Choice", Tier1 Impact Factor (2.7).**
5. **Al-garadi, Mohammed Ali**, et al. "Identifying the influential spreaders in multilayer interactions of online social networks." **Published** in **Journal of Intelligent & Fuzzy Systems 31.5 (2016): 2721-2735**. ." **Impact Factor 1.20.**

Conference paper

1. Mining Social Media for Crime Detection: Review. Proceeding of the 3rd International Conference on Computer Science & Computational Mathematics, pg.48-54

Awards

- **Best of the Best** invention Award for Securing Online Communication:- Cyberbullying Detection in Twitter Network at Eureka Innovation Exhibition UniKL Malaysian and Spanish institute (MIS), 2015**.**

- **Best invention award** of Electrical & Electronics, ICT, Multimedia and Telecommunication Award for Intelligent Cyberguard System On Twitter at Eureka Innovation Exhibition 2015.
- **Gold Award**, for Intelligent Cyberguard System On Twitter at Invention, Innovation & Design Exposition (IIDEX) Malaysia 2015 .
- **Gold Award**, International Young Inventors Award for Securing Online Communication:- Cyberbullying Detection in Twitter Network at IYIA Indonesia 2016.

Patent

- Method and System for detecting spreaders in complex OSNs "submitted for filing".

REFERENCES

Abu-Nimeh, S., Chen, T. M., & Alzubi, O. (2011). Malicious and spam posts in online social networks. *Computer, 44*(9), 23-28.

Adali, S., & Golbeck, J. (2012). *Predicting personality with social behavior.* Paper presented at the Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012).

Aggarwal, A., Rajadesingan, A., & Kumaraguru, P. (2012). *PhishAri: Automatic realtime phishing detection on twitter.* Paper presented at the eCrime Researchers Summit (eCrime), 2012.

Aggarwal, C. C., & Zhai, C. (2012). A survey of text classification algorithms *Mining text data* (pp. 163-222): Springer.

Ahlqvist, T., Bäck, A., Halonen, M., & Heinonen, S. (2008). Social media roadmaps. *Helsinki: Edita Prima Oy*.

Albert, R., & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of modern physics, 74*(1), 47.

Albert, R., Jeong, H., & Barabási, A.-L. (2000). Error and attack tolerance of complex networks. *nature, 406*(6794), 378-382.

AlFalahi, K., Atif, Y., & Abraham, A. (2014). Models of Influence in Online Social Networks. *International Journal of Intelligent Systems, 29*(2), 161-183.

Andrieu, C., De Freitas, N., Doucet, A., & Jordan, M. I. (2003). An introduction to MCMC for machine learning. *Machine learning, 50*(1-2), 5-43.

Ang, R. P. (2016). Cyberbullying: Its Prevention and Intervention Strategies *Child Safety, Welfare and Well-being* (pp. 25-38): Springer.

Anzai, Y. (2012). *Pattern Recognition & Machine Learning*: Elsevier.

Arıcak, O. T. (2009). Psychiatric symptomatology as a predictor of cyberbullying among university students. *Eurasian Journal of Educational Research, 34*(1), 169.

Asur, S., & Huberman, B. A. (2010). *Predicting the future with social media.* Paper presented at the Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on.

Bakshy, E., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). *Everyone's an influencer: quantifying influence on twitter.* Paper presented at the Proceedings of the fourth ACM international conference on Web search and data mining.

Balakrishnan, V. (2015). Cyberbullying among young adults in Malaysia: The roles of gender, age and Internet frequency. *Computers in Human Behavior, 46*, 149-157.

Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *science*, *286*(5439), 509-512.

Basaras, P., Katsaros, D., & Tassiulas, L. (2013). Detecting influential spreaders in complex, dynamic networks. *Computer*, *46*(4), 0024-0029.

Batagelj, V., & Zaversnik, M. (2003). An O (m) algorithm for cores decomposition of networks. *arXiv preprint cs/0310049*.

Bauman, S., Toomey, R. B., & Walker, J. L. (2013). Associations among bullying, cyberbullying, and suicide in high school students. *Journal of adolescence*, *36*(2), 341-350.

Bavelas, A. (1950). Communication patterns in task-oriented groups. *Journal of the acoustical society of America*.

BBC. (2012). Huge rise in social media 'crimes' *http://www.bbc.com/news/uk-20851797*. Retrieved from http://www.bbc.com/news/uk-20851797

Bellmore, A., Calvin, A. J., Xu, J.-M., & Zhu, X. (2015). The five W's of "bullying" on Twitter: Who, What, Why, Where, and When. *Computers in Human Behavior*, *44*, 305-314.

Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson correlation coefficient *Noise reduction in speech processing* (pp. 1-4): Springer.

Beyer, K., Goldstein, J., Ramakrishnan, R., & Shaft, U. (1999). *When is "nearest neighbor" meaningful?* Paper presented at the International conference on database theory.

Bigonha, C., Cardoso, T. N., Moro, M. M., Gonçalves, M. A., & Almeida, V. A. (2012). Sentiment-based influence detection on Twitter. *Journal of the Brazilian Computer Society*, *18*(3), 169-183.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, *3*, 993-1022.

Bora, N., Zaytsev, V., Chang, Y.-H., & Maheswaran, R. (2013). *Gang Networks, Neighborhoods and Holidays: Spatiotemporal Patterns in Social Media.* Paper presented at the Social Computing (SocialCom), 2013 International Conference on.

Borgatti, S. P., & Everett, M. G. (2006). A graph-theoretic perspective on centrality. *Social networks*, *28*(4), 466-484.

Borge-Holthoefer, J., & Moreno, Y. (2012). Absence of influential spreaders in rumor dynamics. *Physical Review E*, *85*(2), 026116.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). *A training algorithm for optimal margin classifiers.* Paper presented at the Proceedings of the fifth annual workshop on Computational learning theory.

Bouguessa, M. (2011). *An unsupervised approach for identifying spammers in social networks.* Paper presented at the Tools with Artificial Intelligence (ICTAI), 2011 23rd IEEE International Conference on.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.

Brin, S., & Page, L. (2012). Reprint of: The anatomy of a large-scale hypertextual web search engine. *Computer networks, 56*(18), 3825-3833.

Buczak, A. L., & Guven, E. (2015). A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications Surveys & Tutorials, 18*(2), 1153-1176.

Budak, C., Agrawal, D., & El Abbadi, A. (2011). *Limiting the spread of misinformation in social networks.* Paper presented at the Proceedings of the 20th international conference on World wide web.

Burger, J. D., Henderson, J., Kim, G., & Zarrella, G. (2011). *Discriminating gender on Twitter.* Paper presented at the Proceedings of the Conference on Empirical Methods in Natural Language Processing.

Calvete, E., Orue, I., Estévez, A., Villardón, L., & Padilla, P. (2010). Cyberbullying in adolescents: Modalities and aggressors' profile. *Computers in Human Behavior, 26*(5), 1128-1135.

Catanese, S., De Meo, P., Ferrara, E., Fiumara, G., & Provetti, A. (2012). Extraction and analysis of facebook friendship relations *Computational Social Networks* (pp. 291-324): Springer.

Centola, D. (2010). The spread of behavior in an online social network experiment. *science, 329*(5996), 1194-1197.

Centola, D., & Macy, M. (2007). Complex contagions and the weakness of long ties1. *American journal of sociology, 113*(3), 702-734.

Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, P. K. (2010). Measuring User Influence in Twitter: The Million Follower Fallacy. *ICWSM, 10*(10-17), 30.

Chai, W., Xu, W., Zuo, M., & Wen, X. *ACQR: A Novel Framework to Identify and Predict Influential Users in Micro-Blogging*.

Chai, W., Xu, W., Zuo, M., & Wen, X. (2013). *ACQR: A Novel Framework to Identify and Predict Influential Users in Micro-Blogging.* Paper presented at the PACIS.

Chatfield, A. T., Reddick, C. G., & Brajawidagda, U. (2015). *Tweeting propaganda, radicalization and recruitment: Islamic state supporters multi-sided twitter networks.* Paper presented at the Proceedings of the 16th Annual International Conference on Digital Government Research.

Chavan, V. S., & Shylaja, S. (2015). *Machine learning approach for detection of cyber-aggressive comments by peers on social media network.* Paper presented at the

Advances in Computing, Communications and Informatics (ICACCI), 2015 International Conference on.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research, 16*(1), 321-357.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research, 16*, 321-357.

Chen, D., Lü, L., Shang, M.-S., Zhang, Y.-C., & Zhou, T. (2012). Identifying influential nodes in complex networks. *Physica A: Statistical Mechanics and its Applications, 391*(4), 1777-1787.

Chen, H., Mckeever, S., & Delany, S. J. (2017). Harnessing the Power of Text Mining for the Detection of Abusive Content in Social Media *Advances in Computational Intelligence Systems* (pp. 187-205): Springer.

Chen, W., Cheng, S., He, X., & Jiang, F. (2012). *Influencerank: An efficient social influence measurement for millions of users in microblog.* Paper presented at the Cloud and Green Computing (CGC), 2012 Second International Conference on.

Chen, Y., Zhou, Y., Zhu, S., & Xu, H. (2012). *Detecting offensive language in social media to protect adolescent online safety.* Paper presented at the Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom).

Cheng, T., & Wicks, T. (2014). Event Detection using Twitter: A Spatio-Temporal Approach. *PloS one, 9*(6), e97807.

Colizza, V., Flammini, A., Serrano, M. A., & Vespignani, A. (2006). Detecting rich-club ordering in complex networks. *Nature Physics, 2*(2), 110-115.

Comin, C. H., & da Fontoura Costa, L. (2011). Identifying the starting point of a spreading process in complex networks. *Physical Review E, 84*(5), 056105.

Connolly, I., & O'Moore, M. (2003). Personality and family relations of children who bully. *Personality and Individual Differences, 35*(3), 559-567.

Corcoran, L., Connolly, I., & O'Moore, M. (2012). Cyberbullying in Irish schools: an investigation of personality and self-concept. *The Irish Journal of Psychology, 33*(4), 153-165.

Cosley, D., Huttenlocher, D. P., Kleinberg, J. M., Lan, X., & Suri, S. (2010). Sequential Influence Models in Social Networks. *ICWSM, 10*, 26.

Cossu, J.-V., Dugué, N., & Labatut, V. (2015). Detecting real-world influence through Twitter. *arXiv preprint arXiv:1506.05903*.

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology, 88*(11), 2783-2792.

Dadvar, M., & De Jong, F. (2012). *Cyberbullying detection: a step toward a safer Internet yard.* Paper presented at the Proceedings of the 21st international conference companion on World Wide Web.

Dadvar, M., de Jong, F., Ordelman, R., & Trieschnigg, R. (2012a). Improved cyberbullying detection using gender information.

Dadvar, M., de Jong, F. M., Ordelman, R., & Trieschnigg, R. (2012b). Improved cyberbullying detection using gender information.

Dadvar, M., Trieschnigg, D., & Jong, F. (2013). Expert knowledge for automatic detection of bullies in social networks.

Dadvar, M., Trieschnigg, D., Ordelman, R., & de Jong, F. (2013). Improving cyberbullying detection with user context *Advances in Information Retrieval* (pp. 693-696): Springer.

Dailymail. (2014). From IHML (I hate my life) to Mos (mum over shoulder): Why this guide to cyber-bullying slang may save your child's life

Retrieved from http://www.dailymail.co.uk/news/article-2673678/Why-guide-cyber-bullying-slang-save-childs-life-From-IHML-I-hate-life-Mos-mum-shoulder.html

De Domenico, M., Lima, A., Mougel, P., & Musolesi, M. (2013). The anatomy of a scientific rumor. *Scientific reports, 3, 2980*. doi:DOI:10.1038/srep02980

Dean, M. (OCTOBER 18, 2012). The Story of Amanda Todd [Online]. Available: http://www.newyorker.com/culture/culture-desk/the-story-of-amanda-todd. *New Yorker*. Retrieved from http://www.newyorker.com/culture/culture-desk/the-story-of-amanda-todd

Deng, Z., Zhu, X., Cheng, D., Zong, M., & Zhang, S. (2016). Efficient kNN classification algorithm for big data. *Neurocomputing, 195*, 143-148.

Dilmac, B. (2009). Psychological Needs as a Predictor of Cyber Bullying: A Preliminary Report on College Students. *Educational Sciences: Theory and Practice, 9*(3), 1307-1325.

Dinakar, K., Jones, B., Havasi, C., Lieberman, H., & Picard, R. (2012). Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS), 2*(3), 18.

Dinakar, K., Reichart, R., & Lieberman, H. (2011). Modeling the detection of Textual Cyberbullying.

Ding, Z.-y., Jia, Y., Zhou, B., Han, Y., He, L., & Zhang, J.-f. (2013). Measuring the spreadability of users in microblogs. *Journal of Zhejiang University SCIENCE C, 14*(9), 701-710.

Doerr, B., Fouz, M., & Friedrich, T. (2012). Why rumors spread so quickly in social networks. *Communications of the ACM, 55*(6), 70-75.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM, 55*(10), 78-87.

Dong, W., Liao, S., Xu, Y., & Feng, X. (2016). Leading Effect of Social Media for Financial Fraud Disclosure: A Text Mining Based Analytics.

Dorogovtsev, S. N., Goltsev, A. V., & Mendes, J. F. F. (2006). K-core organization of complex networks. *Physical review letters, 96*(4), 040601.

Dreiseitl, S., Ohno-Machado, L., Kittler, H., Vinterbo, S., Billhardt, H., & Binder, M. (2001). A comparison of machine learning methods for the diagnosis of pigmented skin lesions. *Journal of biomedical informatics, 34*(1), 28-36.

Duda, R. O., Hart, P. E., & Stork, D. G. (2012). *Pattern classification*: John Wiley & Sons.

Ebel, H., Mielsch, L.-I., & Bornholdt, S. (2002). Scale-free topology of e-mail networks. *Physical Review E, 66*(3), 035103.

Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., . . . Sap, M. (2015). Psychological language on Twitter predicts county-level heart disease mortality. *Psychological science, 26*(2), 159-169.

Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication, 13*(1), 210-230.

Fast, L. A., & Funder, D. C. (2008). Personality as manifest in word use: correlations with self-report, acquaintance report, and behavior. *Journal of personality and social psychology, 94*(2), 334.

Faust, K. (1997). Centrality in affiliation networks. *Social networks, 19*(2), 157-191.

Fawcett, T. (2004). ROC graphs: Notes and practical considerations for researchers. *Machine learning, 31*(1), 1-38.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861-874.

Feng, P. E. B. J. (2011). Measuring user influence on twitter using modified k-shell decomposition.

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems. *J. Mach. Learn. Res, 15*(1), 3133-3181.

Fire, M., Goldschmidt, R., & Elovici, Y. (2014a). Online social networks: threats and solutions. *IEEE Communications Surveys & Tutorials, 16*(4), 2019-2036.

Fire, M., Goldschmidt, R., & Elovici, Y. (2014b). Online Social Networks: Threats and Solutions. *Communications Surveys & Tutorials, IEEE, 16*(4), 2019-2036.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of machine learning research, 3*(Mar), 1289-1305.

Foster, P. (April 2013). 'Bogus' AP tweet about explosion at the White House wipes billions off US markets [Online]. Available: http://www.telegraph.co.uk/finance/markets/10013768/Bogus-AP-tweet-about-explosion-at-the-White-House-wipes-billions-off-US-markets.html *The Telegraph, Finance/Market, Washington*.

Freeman, D. M. (2013). *Using naive bayes to detect spammy names in social networks.* Paper presented at the Proceedings of the 2013 ACM workshop on Artificial intelligence and security.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 35-41.

Freeman, L. C. (1979). Centrality in social networks conceptual clarification. *Social networks, 1*(3), 215-239.

Frommholz, I., al-Khateeb, H. M., Potthast, M., Ghasem, Z., Shukla, M., & Short, E. (2016). On Textual Analysis and Machine Learning for Cyberstalking Detection. *Datenbank-Spektrum*, 1-9.

Galán-García, P., de la Puerta, J. G., Gómez, C. L., Santos, I., & Bringas, P. G. (2014). *Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying.* Paper presented at the International Joint Conference SOCO'13-CISIS'13-ICEUTE'13.

Gao, C., Liu, J., & Zhong, N. (2011). Network immunization and virus propagation in email networks: experimental evaluation and analysis. *Knowledge and information systems, 27*(2), 253-279.

Garas, A., Schweitzer, F., & Havlin, S. (2012). A k-shell decomposition method for weighted networks. *New Journal of Physics, 14*(8), 083030.

García-Recuero, Á. (2016). *Discouraging Abusive Behavior in Privacy-Preserving Online Social Networking Applications.* Paper presented at the Proceedings of the 25th International Conference Companion on World Wide Web.

Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *nature, 521*(7553), 452-459.

Ghoshal, G., & Barabási, A.-L. (2011). Ranking stability and super-stable nodes in complex networks. *Nature communications, 2*, 394.

Gill, A. J., Nowson, S., & Oberlander, J. (2009). *What Are They Blogging About? Personality, Topic and Motivation in Blogs.* Paper presented at the ICWSM.

Goel, S., Watts, D. J., & Goldstein, D. G. (2012). *The structure of online diffusion networks.* Paper presented at the Proceedings of the 13th ACM conference on electronic commerce.

Golbeck, J., Robles, C., Edmondson, M., & Turner, K. (2011). *Predicting personality from twitter.* Paper presented at the Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on.

Golbeck, J., Robles, C., & Turner, K. (2011). *Predicting personality with social media.* Paper presented at the CHI'11 extended abstracts on human factors in computing systems.

Goldenberg, J., Libai, B., & Muller, E. (2001). Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing letters, 12*(3), 211-223.

Golder, S. A., & Macy, M. W. (2011). Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *science, 333*(6051), 1878-1881.

González-Bailón, S., Wang, N., Rivero, A., Borge-Holthoefer, J., & Moreno, Y. (2014). Assessing the bias in samples of large online networks. *Social networks, 38*, 16-27.

Gray, R. M. (1990). Entropy and information *Entropy and Information Theory* (pp. 21-55): Springer.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences, 101*(suppl 1), 5228-5235.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter, 11*(1), 10-18.

Hall, M. A. (1999). *Correlation-based feature selection for machine learning.* The University of Waikato.

Hatzivassiloglou, V., & McKeown, K. R. (1997). *Predicting the semantic orientation of adjectives.* Paper presented at the Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics.

He, H. (2007). Eigenvectors and reconstruction. *the electronic journal of combinatorics, 14*(1), N14.

He, Z., Cai, Z., Yu, J., Wang, X., Sun, Y., & Li, Y. (2016). Cost-efficient strategies for restraining rumor spreading in mobile social networks. *IEEE Transactions on Vehicular Technology*.

Hethcote, H. W. (2000). The mathematics of infectious diseases. *SIAM review, 42*(4), 599-653.

Hinduja, S., & Patchin, J. W. (2010). Bullying, cyberbullying, and suicide. *Archives of suicide research, 14*(3), 206-221.

Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural computation, 18*(7), 1527-1554.

Ho, J. Y., & Dempsey, M. (2010). Viral marketing: Motivations to forward online content. *Journal of Business Research, 63*(9), 1000-1006.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398): John Wiley & Sons.

Hosseini, M., & Tammimy, Z. (2016). Recognizing users gender in social media using linguistic features. *Computers in Human Behavior, 56*, 192-197.

Hosseinmardi, H., Han, R., Lv, Q., Mishra, S., & Ghasemianlangroodi, A. (2014). *Towards understanding cyberbullying behavior in a semi-anonymous social network.* Paper presented at the Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on.

Hosseinmardi, H., Mattson, S. A., Rafiq, R. I., Han, R., Lv, Q., & Mishra, S. (2015). Detection of cyberbullying incidents on the instagram social network. *arXiv preprint arXiv:1503.03909*.

Hsu, C.-W., Chang, C.-C., & Lin, C.-J. (2003). A practical guide to support vector classification.

Huang, Q., Singh, V. K., & Atrey, P. K. (2014). *Cyber Bullying Detection Using Social and Textual Analysis.* Paper presented at the Proceedings of the 3rd International Workshop on Socially-Aware Multimedia.

Ipeirotis, P. (2010). Mechanical turk: Now with 40.92% spam. *Behind Enemy Lines blog*.

Jabeur, L. B., Tamine, L., & Boughanem, M. (2012). *Active microbloggers: identifying influencers, leaders and discussers in microblogging networks.* Paper presented at the String Processing and Information Retrieval.

Japkowicz, N., & Shah, M. (2011). *Evaluating learning algorithms: a classification perspective*: Cambridge University Press.

Java, A., Kolari, P., Finin, T., & Oates, T. (2006). *Modeling the spread of influence on the blogosphere.* Paper presented at the Proceedings of the 15th international world wide web conference.

Jeong, S. Y., Koh, Y. S., & Dobbie, G. (2016). *Phishing Detection on Twitter Streams.* Paper presented at the Pacific-Asia Conference on Knowledge Discovery and Data Mining.

Jiang, J., Wilson, C., Wang, X., Sha, W., Huang, P., Dai, Y., & Zhao, B. Y. (2013). Understanding latent interactions in online social networks. *ACM Transactions on the Web (TWEB), 7*(4), 18.

Joachims, T. (1998). *Text categorization with support vector machines: Learning with many relevant features.* Paper presented at the European conference on machine learning.

Kansara, K. B., & Shekokar, N. M. (2015). A Framework for Cyberbullying Detection in Social Network.

Kim, E. S., & Han, S. S. (2009). *An analytical way to find influencers on social networks and validate their effects in disseminating social games.* Paper presented at the Social Network Analysis and Mining, 2009. ASONAM'09. International Conference on Advances in.

Kitsak, M., Gallos, L. K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H. E., & Makse, H. A. (2010). Identification of influential spreaders in complex networks. *Nature Physics, 6*(11), 888-893.

Klausen, J. (2015). Tweeting the Jihad: Social media networks of Western foreign fighters in Syria and Iraq. *Studies in Conflict & Terrorism, 38*(1), 1-22.

Kleinberg, J. (2007). Cascading behavior in networks: Algorithmic and economic issues. *Algorithmic game theory, 24*, 613-632.

Kohavi, R. (1995). *A study of cross-validation and bootstrap for accuracy estimation and model selection.* Paper presented at the Ijcai.

Kontostathis, A., Reynolds, K., Garron, A., & Edwards, L. (2013). *Detecting cyberbullying: query terms and techniques.* Paper presented at the Proceedings of the 5th annual acm web science conference.

Korde, V., & Mahender, C. N. (2012). Text classification and classifiers: A survey. *International Journal of Artificial Intelligence & Applications, 3*(2), 85.

Kovács, I. A., & Barabási, A.-L. (2015). Network science: Destruction perfected. *nature, 524*(7563), 38-39.

Kowalski, R. M., Giumetti, G. W., Schroeder, A. N., & Lattanner, M. R. (2014). Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth. *Psychological bulletin, 140*(4), 1073.

Kowalski, R. M., Giumetti, G. W., Schroeder, A. N., & Reese, H. H. (2012). Cyber bullying among college students: Evidence from multiple domains of college life. *Cutting-edge Technologies in Higher Education, 5*, 293-321.

Kowalski, R. M., Limber, S., Limber, S. P., & Agatston, P. W. (2012). *Cyberbullying: Bullying in the digital age*: John Wiley & Sons.

Kowalski, R. M., & Limber, S. P. (2013). Psychological, physical, and academic correlates of cyberbullying and traditional bullying. *Journal of Adolescent Health, 53*(1), S13-S20.

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). *What is Twitter, a social network or a news media?* Paper presented at the Proceedings of the 19th international conference on World wide web.

Kwon, S., Cha, M., Jung, K., Chen, W., & Wang, Y. (2013). *Prominent features of rumor propagation in online social media.* Paper presented at the Data Mining (ICDM), 2013 IEEE 13th International Conference on.

Langley, P., & Simon, H. A. (1995). Applications of machine learning and rule induction. *Communications of the ACM, 38*(11), 54-64.

Lauw, H., Shafer, J. C., Agrawal, R., & Ntoulas, A. (2010). Homophily in the digital world: A LiveJournal case study. *Internet Computing, IEEE, 14*(2), 15-23.

Lauw, H. W., Shafer, J. C., Agrawal, R., & Ntoulas, A. (2010). Homophily in the digital world: A LiveJournal case study. *Internet Computing, IEEE, 14*(2), 15-23.

Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., . . . Gutmann, M. (2009). Life in the network: the coming age of computational social science. *Science (New York, NY), 323*(5915), 721.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature, 521*(7553), 436-444.

Lee, K., Mahmud, J., Chen, J., Zhou, M., & Nichols, J. (2014). *Who will retweet thisfi.* Paper presented at the Proceedings of the 19th international conference on Intelligent User Interfaces.

Lee, K., Mahmud, J., Chen, J., Zhou, M., & Nichols, J. (2015). Who will retweet this? detecting strangers from twitter to retweet information. *ACM Transactions on Intelligent Systems and Technology (TIST), 6*(3), 31.

Leskovec, J., Adamic, L. A., & Huberman, B. A. (2007). The dynamics of viral marketing. *ACM Transactions on the Web (TWEB), 1*(1), 5.

Leskovec, J., Huttenlocher, D., & Kleinberg, J. (2010). *Predicting positive and negative links in online social networks.* Paper presented at the Proceedings of the 19th international conference on World wide web.

Li, H., Bhowmick, S. S., Sun, A., & Cui, J. (2015). Conformity-aware influence maximization in online social networks. *The VLDB Journal—The International Journal on Very Large Data Bases, 24*(1), 117-141.

Li, L., Sun, M., & Liu, Z. (2014). *Discriminating gender on Chinese microblog: A study of online behaviour, writing style and preferred vocabulary.* Paper presented at the Natural Computation (ICNC), 2014 10th International Conference on.

Li, Q. (2007). New bottle but old wine: A research of cyberbullying in schools. *Computers in Human Behavior, 23*(4), 1777-1791.

Li, Q., Zhou, T., Lü, L., & Chen, D. (2014). Identifying influential spreaders by weighted LeaderRank. *Physica A: Statistical Mechanics and its Applications, 404*, 47-55.

Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics, 16*(6), 321-332.

Liu, J.-G., Ren, Z.-M., & Guo, Q. (2013). Ranking the spreading influence in complex networks. *Physica A: Statistical Mechanics and its Applications, 392*(18), 4154-4159.

Liu, N., Li, L., Xu, G., & Yang, Z. (2014). *Identifying Domain-Dependent Influential Microblog Users: A Post-Feature Based Approach.* Paper presented at the Twenty-Eighth AAAI Conference on Artificial Intelligence.

Liu, S., Zhang, J., & Xiang, Y. (2016). *Statistical Detection of Online Drifting Twitter Spam: Invited Paper.* Paper presented at the Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security.

Liu, W., & Ruths, D. (2013). *What's in a Name? Using First Names as Features for Gender Inference in Twitter*.

Liu, X.-Y., & Zhou, Z.-H. (2006). *The influence of class imbalance on cost-sensitive learning: An empirical study.* Paper presented at the Data Mining, 2006. ICDM'06. Sixth International Conference on.

Liu, Y.-Y., Slotine, J.-J., & Barabási, A.-L. (2011). Controllability of complex networks. *nature, 473*(7346), 167-173.

Liu, Y., Kliman-Silver, C., & Mislove, A. (2014). *The tweets they are a-changin': Evolution of twitter users and behavior.* Paper presented at the International AAAI Conference on Weblogs and Social Media (ICWSM).

Liu, Y., Tang, M., Zhou, T., & Do, Y. (2015). Core-like groups result in invalidation of identifying super-spreader by k-shell decomposition. *Scientific reports, 5*.

Livingstone, S., Haddon, L., Görzig, A., & Ólafsson, K. (2011). Technical report and user guide: The 2010 EU kids online survey.

Lü, L., Chen, D., Ren, X.-L., Zhang, Q.-M., Zhang, Y.-C., & Zhou, T. (2016). Vital nodes identification in complex networks. *Physics reports, 650*, 1-63.

Lü, L., Zhang, Y.-C., Yeung, C. H., & Zhou, T. (2011). Leaders in social networks, the delicious case. *PloS one, 6*(6), e21202.

Macià, N., Bernadó-Mansilla, E., Orriols-Puig, A., & Ho, T. K. (2013). Learner excellence biased by data set selection: A case for data characterisation and artificial data sets. *Pattern Recognition, 46*(3), 1054-1066.

Magro, M. J. (2012). A review of social media use in e-government. *Administrative Sciences, 2*(2), 148-161.

Mahmud, J., Zhou, M. X., Megiddo, N., Nichols, J., & Drews, C. (2013). *Recommending targeted strangers from whom to solicit information on social media.* Paper presented at the Proceedings of the 2013 international conference on Intelligent user interfaces.

Mairesse, F., & Walker, M. Computational Models of Personality Recognition through Language.

Mangaonkar, A., Hayrapetian, A., & Raje, R. (2015). *Collaborative detection of cyberbullying behavior in Twitter data.* Paper presented at the 2015 IEEE International Conference on Electro/Information Technology (EIT).

Margono, H., Yi, X., & Raikundalia, G. K. (2014). *Mining Indonesian cyber bullying patterns in social networks.* Paper presented at the Proceedings of the Thirty-Seventh Australasian Computer Science Conference-Volume 147.

McCallum, A., & Nigam, K. (1998). *A comparison of event models for naive bayes text classification.* Paper presented at the AAAI-98 workshop on learning for text categorization.

Mei, Y., Zhong, Y., & Yang, J. (2015). *Finding and Analyzing Principal Features for Measuring User Influence on Twitter.* Paper presented at the Big Data Computing Service and Applications (BigDataService), 2015 IEEE First International Conference on.

Miller, Z., Dickinson, B., Deitrick, W., Hu, W., & Wang, A. H. (2014). Twitter spammer detection using data stream clustering. *Information Sciences, 260*, 64-73.

Miller, Z., Dickinson, B., & Hu, W. (2012). Gender prediction on Twitter using stream algorithms with N-gram character features.

Min, B., Liljeros, F., & Makse, H. A. (2015). Finding influential spreaders from human activity beyond network location. *PloS one, 10*(8), e0136831.

Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., & Bhattacharjee, B. (2007). *Measurement and analysis of online social networks.* Paper presented at the Proceedings of the 7th ACM SIGCOMM conference on Internet measurement.

Morone, F., & Makse, H. A. (2015). Influence maximization in complex networks through optimal percolation. *nature*.

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. *arXiv preprint arXiv:1306.5204*.

Nadali, S., Murad, M. A. A., Sharef, N. M., Mustapha, A., & Shojaee, S. (2013). *A review of cyberbullying detection: An overview.* Paper presented at the Intelligent Systems Design and Applications (ISDA), 2013 13th International Conference on.

Nahar, V., Li, X., & Pang, C. (2013). An effective approach for cyberbullying detection. *Communications in Information Science and Management Engineering, 3*(5), 238.

Nakano, T., Suda, T., Okaie, Y., & Moore, M. J. (2016). *Analysis of Cyber Aggression and Cyber-Bullying in Social Networking.* Paper presented at the 2016 IEEE Tenth International Conference on Semantic Computing (ICSC).

Nalini, K., & Sheela, L. J. (2015). *Classification of Tweets Using Text Classifier to Detect Cyber Bullying.* Paper presented at the Emerging ICT for Bridging the Future-Proceedings of the 49th Annual Convention of the Computer Society of India CSI Volume 2.

Navarro, J. N., & Jasinski, J. L. (2012). Going cyber: Using routine activities theory to predict cyberbullying experiences. *Sociological Spectrum, 32*(1), 81-94.

Nepusz, T., & Vicsek, T. (2012). Controlling edge dynamics in complex networks. *Nature Physics, 8*(7), 568-573.

Ngai, E. W., Moon, K.-l. K., Lam, S., Chin, E. S., & Tao, S. S. (2015). Social media models, technologies, and applications: an academic review and case study. *Industrial Management & Data Systems, 115*(5), 769-802.

Nguyen, D.-P., Gravel, R., Trieschnigg, R., & Meder, T. (2013). " How old do you think I am?" A study of language and age in Twitter.

Nguyen, D., Gravel, R., Trieschnigg, D., & Meder, T. (2013). *" How Old Do You Think I Am?"; A Study of Language and Age in Twitter.* Paper presented at the Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media.

Nguyen, T. H., & Szymanski, B. K. (2013). *Social ranking techniques for the web.* Paper presented at the Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on.

Ni, L. M., Tan, H., & Xiao, J. (2016). Rethinking big data in a networked world. *Frontiers of Computer Science, 10*(6), 965-967.

O'Keeffe, G. S., & Clarke-Pearson, K. (2011). The impact of social media on children, adolescents, and families. *Pediatrics, 127*(4), 800-804.

Oh, O., Kwon, K. H., & Rao, H. R. (2010). *An Exploration of Social Media in Extreme Events: Rumor Theory and Twitter during the Haiti Earthquake 2010.* Paper presented at the ICIS.

Pastor-Satorras, R., & Vespignani, A. (2001). Epidemic spreading in scale-free networks. *Physical review letters, 86*(14), 3200.

Pastor-Satorras, R., & Vespignani, A. (2002). Immunization of complex networks. *Physical Review E, 65*(3), 036104.

Patchin, J., & Hinduja, S. (2013). *Words Wound: Delete cyberbullying and make kindness go viral*: Free Spirit Publishing.

Pearce, M. (Sep. 2013). Florida girl, 12, found dead after bullies said 'kill yourself [Online]. Available: http://articles.latimes.com/2013/sep/12/nation/la-na-nn-florida-cyberbullying-20130912 *Los Angeles Times, LosAngeles, CA, USA*.

Peersman, C., Daelemans, W., & Van Vaerenbergh, L. (2011). *Predicting age and gender in online social networks.* Paper presented at the Proceedings of the 3rd international workshop on Search and mining user-generated contents.

Pei, S., & Makse, H. A. (2013). Spreading dynamics in complex networks. *Journal of Statistical Mechanics: Theory and Experiment, 2013*(12), P12002.

Pei, S., Muchnik, L., Andrade Jr, J. S., Zheng, Z., & Makse, H. A. (2014). Searching for superspreaders of information in real-world social media. *Scientific reports, 4*.

Pei, S., Muchnik, L., Tang, S., Zheng, Z., & Makse, H. A. (2015a). Exploring the complex pattern of information spreading in online blog communities.

Pei, S., Muchnik, L., Tang, S., Zheng, Z., & Makse, H. A. (2015b). Exploring the Complex Pattern of Information Spreading in Online Blog Communities. *PloS one, 10*(5).

Pennacchiotti, M., & Popescu, A.-M. (2011). A Machine Learning Approach to Twitter User Classification. *ICWSM, 11*, 281-288.

Peterson, J. K., & Densley, J. (2016). Is Social Media a Gang? Toward a Selection, Facilitation, or Enhancement Explanation of Cyber Violence. *Aggression and Violent Behavior*.

Preotiuc-Pietro, D., Eichstaedt, J., Park, G., Sap, M., Smith, L., Tobolsky, V., . . . Ungar, L. (2015). The role of personality, age and gender in tweeting about mental illnesses. *NAACL HLT 2015*, 21.

Prieto, V. M., Matos, S., Alvarez, M., Cacheda, F., & Oliveira, J. L. (2014). Twitter: a good place to detect health conditions. *PloS one, 9*(1), e86191.

Provost, F. J., & Fawcett, T. (1997). *Analysis and visualization of classifier performance: comparison under imprecise class and cost distributions.* Paper presented at the KDD.

Ptaszynski, M., Dybala, P., Matsuba, T., Masui, F., Rzepka, R., & Araki, K. (2010). Machine learning and affect analysis against cyber-bullying. *the 36th AISB*, 7-16.

Qabajeh, I., & Thabtah, F. (2014). *An Experimental Study for Assessing Email Classification Attributes Using Feature Selection Methods.* Paper presented at the Advanced Computer Science Applications and Technologies (ACSAT), 2014 3rd International Conference on.

Quan, H., Wu, J., & Shi, J. (2011). Online social networks & social network services: A technical survey. *Pervasive Communication Handbook. CRC*, 4.

Quercia, D., Kosinski, M., Stillwell, D., & Crowcroft, J. (2011). *Our Twitter profiles, our selves: Predicting personality with Twitter.* Paper presented at the Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on.

Räbiger, S., & Spiliopoulou, M. (2015). A framework for validating the merit of properties that predict the influence of a twitter user. *Expert Systems with Applications, 42*(5), 2824-2834.

Rahman, M. S., Huang, T.-K., Madhyastha, H. V., & Faloutsos, M. (2012). *FRAppE: detecting malicious facebook applications.* Paper presented at the Proceedings of the 8th international conference on Emerging networking experiments and technologies.

Raisi, E., & Huang, B. (2016). Cyberbullying Identification Using Participant-Vocabulary Consistency. *arXiv preprint arXiv:1606.08084*.

Rangel, F., & Rosso, P. (2013). Use of language and author profiling: Identification of gender and age. *Natural Language Processing and Cognitive Science, 177*.

Rao, D., Yarowsky, D., Shreevats, A., & Gupta, M. (2010). *Classifying latent user attributes in twitter.* Paper presented at the Proceedings of the 2nd international workshop on Search and mining user-generated contents.

Ratkiewicz, J., Conover, M., Meiss, M., Gonçalves, B., Flammini, A., & Menczer, F. (2011). *Detecting and Tracking Political Abuse in Social Media.* Paper presented at the ICWSM.

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation *Encyclopedia of database systems* (pp. 532-538): Springer.

Reynolds, K., Kontostathis, A., & Edwards, L. (2011). *Using machine learning to detect cyberbullying.* Paper presented at the Machine Learning and Applications and Workshops (ICMLA), 2011 10th International Conference on.

Romero, D. M., Galuba, W., Asur, S., & Huberman, B. A. (2011). Influence and passivity in social media *Machine learning and knowledge discovery in databases* (pp. 18-33): Springer.

Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *science, 346*(6213), 1063-1064.

Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. *Systems, Man and Cybernetics, IEEE Transactions on, 21*(3), 660-674.

Sampasa-Kanyinga, H., Roumeliotis, P., & Xu, H. (2014). Associations between cyberbullying and school bullying victimization and suicidal ideation, plans and attempts among Canadian schoolchildren. *PloS one, 9*(7), e102145.

Sanchez, H., & Kumar, S. (2011). Twitter bullying detection. *UCSC ISM245 Data Mining course report*.

Santosh, K., Bansal, R., Shekhar, M., & Varma, V. (2013). Author profiling: Predicting age and gender from blogs. *Notebook for PAN at CLEF*, 119-124.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., . . . Seligman, M. E. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one, 8*(9), e73791.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR), 34*(1), 1-47.

Shekokar, N. M., & Kansara, K. B. (2016). *Security against sybil attack in social network.* Paper presented at the Information Communication and Embedded Systems (ICICES), 2016 International Conference on.

Silva, A., Guimarães, S., Meira Jr, W., & Zaki, M. (2013). *ProfileRank: finding relevant content and influential users based on information diffusion.* Paper presented at the Proceedings of the 7th Workshop on Social Network Mining and Analysis.

Singh, P., Sreenivasan, S., Szymanski, B. K., & Korniss, G. (2013). Threshold-limited spreading in social networks with multiple initiators. *Scientific reports, 3*.

Slonje, R., & Smith, P. K. (2008). Cyberbullying: Another main type of bullying? *Scandinavian journal of psychology, 49*(2), 147-154.

Slonje, R., Smith, P. K., & Frisén, A. (2013). The nature of cyberbullying, and strategies for prevention. *Computers in Human Behavior, 29*(1), 26-32.

Sood, S. O., Antin, J., & Churchill, E. F. (2012). *Using Crowdsourcing to Improve Profanity Detection.* Paper presented at the AAAI Spring Symposium: Wisdom of the Crowd.

Soucy, P., & Mineau, G. W. (2001). *A simple KNN algorithm for text categorization.* Paper presented at the Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on.

Sourander, A., Klomek, A. B., Ikonen, M., Lindroos, J., Luntamo, T., Koskelainen, M., . . . Helenius, H. (2010). Psychosocial risk factors associated with cyberbullying among adolescents: A population-based study. *Archives of general psychiatry, 67*(7), 720-728.

Squicciarini, A., Rajtmajer, S., Liu, Y., & Griffin, C. (2015). *Identification and characterization of cyberbullying dynamics in an online social network.* Paper presented at the Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015.

Sticca, F., & Perren, S. (2013). Is cyberbullying worse than traditional bullying? Examining the differential roles of medium, publicity, and anonymity for the perceived severity of bullying. *Journal of youth and adolescence*, *42*(5), 739-750.

Sun, J., & Tang, J. (2011). A survey of models and algorithms for social influence analysis *Social network data analytics* (pp. 177-214): Springer.

Talebi, M., & Kose, C. (2013). *Identifying gender, age and Education level by analyzing comments on Facebook.* Paper presented at the Signal Processing and Communications Applications Conference (SIU), 2013 21st.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, *29*(1), 24-54.

Tokunaga, R. S. (2010). Following you home from school: A critical review and synthesis of research on cyberbullying victimization. *Computers in Human Behavior*, *26*(3), 277-287.

Tong, R. M. (2001). *An operational system for detecting and tracking opinions in on-line discussion.* Paper presented at the Working Notes of the ACM SIGIR 2001 Workshop on Operational Text Classification.

Tong, S., & Koller, D. (2001). Support vector machine active learning with applications to text classification. *Journal of machine learning research*, *2*(Nov), 45-66.

Tsugawa, S., & Ohsaki, H. (2015). *Negative Messages Spread Rapidly and Widely on Social Media.* Paper presented at the Proceedings of the 2015 ACM on Conference on Online Social Networks.

Tunkelang, D. (2009). A twitter analog to PageRank. *http://thenoisychannel.com/2009/01/13/a-twitter-analog-to-PageRank*.

Turney, P. D. (2002). *Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews.* Paper presented at the Proceedings of the 40th annual meeting on association for computational linguistics.

Van Hee, C., Lefever, E., Verhoeven, B., Mennes, J., Desmet, B., De Pauw, G., . . . Hoste, V. (2015). *Detection and fine-grained classification of cyberbullying events.* Paper presented at the International Conference Recent Advances in Natural Language Processing (RANLP).

Van Royen, K., Poels, K., Daelemans, W., & Vandebosch, H. (2015). Automatic monitoring of cyberbullying on social networking sites: From technological feasibility to desirability. *Telematics and Informatics*, *32*(1), 89-97.

Van Royen, K., Poels, K., & Vandebosch, H. (2016). Harmonizing freedom and protection: Adolescents' voices on automatic monitoring of social networking sites. *Children and Youth Services Review*, *64*, 35-41.

Vandebosch, H., & Van Cleemput, K. (2009). Cyberbullying among youngsters: Profiles of bullies and victims. *New media & society, 11*(8), 1349-1371.

Vanschoren, J., Blockeel, H., Pfahringer, B., & Holmes, G. (2012). Experiment databases. *Machine learning, 87*(2), 127-158.

Vapnik, V. (2013). *The nature of statistical learning theory*: Springer Science & Business Media.

Vayena, E., Salathé, M., Madoff, L. C., Brownstein, J. S., & Bourne, P. E. (2015). Ethical challenges of big data in public health. *PLoS Comput Biol, 11*(2), e1003904.

Wang, A. H. (2010). *Don't follow me: Spam detection in twitter.* Paper presented at the Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on.

Wang, W., Chen, L., Thirunarayan, K., & Sheth, A. P. (2014). *Cursing in english on twitter.* Paper presented at the Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing.

Watters, P. A., & Phair, N. (2012). Detecting illicit drugs on social media using automated social media intelligence analysis (ASMIA) *Cyberspace safety and security* (pp. 66-76): Springer.

Watts, D. J. (2002). A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences, 99*(9), 5766-5771.

Watts, D. J., Peretti, J., & Frumin, M. (2007). *Viral marketing for the real world*: Harvard Business School Pub.

Wei, B., Liu, J., Wei, D., Gao, C., & Deng, Y. (2015). Weighted k-shell decomposition for complex networks based on potential edge weights. *Physica A: Statistical Mechanics and its Applications, 420*, 277-283.

Weinberg, T. (2009). *The new community rules: Marketing on the social web*: O'Reilly Sebastopol, CA.

Weir, G. R., Toolan, F., & Smeed, D. (2011). The threats of social networking: Old wine in new bottles? *Information Security Technical Report, 16*(2), 38-43.

Wen, S., Jiang, J., Xiang, Y., Yu, S., Zhou, W., & Jia, W. (2014a). To shut them up or to clarify: Restraining the spread of rumors in online social networks. *IEEE Transactions on Parallel and Distributed Systems, 25*(12), 3306-3316.

Wen, S., Jiang, J., Xiang, Y., Yu, S., Zhou, W., & Jia, W. (2014b). To Shut Them Up or to Clarify: Restraining the Spread of Rumors in Online Social Networks. *Parallel and Distributed Systems, IEEE Transactions on, 25*(12), 3306-3316.

Weng, J., Lim, E.-P., Jiang, J., & He, Q. (2010). *Twitterrank: finding topic-sensitive influential twitterers.* Paper presented at the Proceedings of the third ACM international conference on Web search and data mining.

Weng, L. Learning Human Dynamics with Big Data from Online Social Networks.

Weng, L. (2014). *Information diffusion on online social networks.* Citeseer.

Weng, L., Flammini, A., Vespignani, A., & Menczer, F. (2012). Competition among memes in a world with limited attention. *Scientific reports, 2*.

Weng, L., Menczer, F., & Ahn, Y.-Y. (2013). Virality prediction and community structure in social networks. *Scientific reports, 33, 2522*.

Whittaker, E., & Kowalski, R. M. (2015). Cyberbullying Via Social Media. *Journal of School Violence, 14*(1), 11-29.

Williams, K. R., & Guerra, N. G. (2007). Prevalence and predictors of internet bullying. *Journal of Adolescent Health, 41*(6), S14-S21.

Wolke, D., Lee, K., & Guy, A. (2017). Cyberbullying: a storm in a teacup? *European Child & Adolescent Psychiatry*, 1-10.

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE transactions on evolutionary computation, 1*(1), 67-82.

Wu, S. (2013). *The Dynamics Of Information Diffusion On On-Line Social Networks.* Cornell University.

Wu, S., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). *Who says what to whom on twitter.* Paper presented at the Proceedings of the 20th international conference on World wide web.

Wu, S., Tan, C., Kleinberg, J. M., & Macy, M. W. (2011). *Does Bad News Go Away Faster?* Paper presented at the ICWSM.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., . . . Philip, S. Y. (2008). Top 10 algorithms in data mining. *Knowledge and information systems, 14*(1), 1-37.

Xiang, G., Fan, B., Wang, L., Hong, J., & Rose, C. (2012). *Detecting offensive tweets via topical feature discovery over a large scale twitter corpus.* Paper presented at the Proceedings of the 21st ACM international conference on Information and knowledge management.

Xiao, C., Zhang, Y., Zeng, X., & Wu, Y. (2013). Predicting user influence in social media. *Journal of Networks, 8*(11), 2649-2655.

Xu, J.-M., Jun, K.-S., Zhu, X., & Bellmore, A. (2012). *Learning from bullying traces in social media.* Paper presented at the Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies.

Yamaguchi, Y., Takahashi, T., Amagasa, T., & Kitagawa, H. (2010). Turank: Twitter user ranking based on user-tweet graph analysis *Web Information Systems Engineering–WISE 2010* (pp. 240-253): Springer.

Yan, G., Chen, G., Eidenbenz, S., & Li, N. (2011). *Malware propagation in online social networks: nature, dynamics, and defense implications.* Paper presented at the Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security.

Yang, C., Harkreader, R., Zhang, J., Shin, S., & Gu, G. (2012). *Analyzing spammers' social networks for fun and profit: a case study of cyber criminal ecosystem on twitter.* Paper presented at the Proceedings of the 21st international conference on World Wide Web.

Yang, W., Wang, H., & Yao, Y. (2015). An immunization strategy for social network worms based on network vertex influence. *Communications, China, 12*(7), 154-166.

Yang, Y., & Pedersen, J. O. *A comparative study on feature selection in text categorization*.

Yang, Y., & Pedersen, J. O. (1997). *A comparative study on feature selection in text categorization.* Paper presented at the Icml.

Yardi, S., Romero, D., & Schoenebeck, G. (2009). Detecting spam in a twitter network. *First Monday, 15*(1).

Yin, D., Xue, Z., Hong, L., Davison, B. D., Kontostathis, A., & Edwards, L. (2009). Detection of harassment on web 2.0. *Proceedings of the Content Analysis in the WEB, 2*.

Yin, Z., & Zhang, Y. (2012). *Measuring pair-wise social influence in microblog.* Paper presented at the Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom).

Zanin, M., Papo, D., Sousa, P. A., Menasalvas, E., Nicchi, A., Kubik, E., & Boccaletti, S. (2016). Combining complex networks and data mining: why and how. *Physics reports, 635*, 1-44.

Zeng, A., & Zhang, C.-J. (2013). Ranking spreaders by decomposing complex networks. *Physics Letters A, 377*(14), 1031-1035.

Zhang, H. (2004a). The optimality of naive Bayes. *A A, 1*(2), 3.

Zhang, H. (2004b). The optimality of naive Bayes.

Zhang, Z.-K., Liu, C., Zhan, X.-X., Lu, X., Zhang, C.-X., & Zhang, Y.-C. (2016). Dynamics of information diffusion and its applications on complex networks. *Physics reports, 651*, 1-34.

Zhao, L., Wang, Q., Cheng, J., Chen, Y., Wang, J., & Huang, W. (2011). Rumor spreading model with consideration of forgetting mechanism: A case of online blogging LiveJournal. *Physica A: Statistical Mechanics and its Applications, 390*(13), 2619-2625.

Zhao, R., Zhou, A., & Mao, K. (2016). *Automatic detection of cyberbullying on social networks based on bullying features.* Paper presented at the Proceedings of the 17th International Conference on Distributed Computing and Networking.

Zheng, Z., Wu, X., & Srihari, R. (2004). Feature selection for text categorization on imbalanced data. *ACM Sigkdd Explorations Newsletter, 6*(1), 80-89.

Zhu, Z. (2013). Discovering the influential users oriented to viral marketing based on online social networks. *Physica A: Statistical Mechanics and its Applications, 392*(16), 3459-3469.

Zou, C. C., Towsley, D., & Gong, W. (2007). Modeling and simulation study of the propagation and defense of internet e-mail worms. *Dependable and Secure Computing, IEEE Transactions on, 4*(2), 105-118.