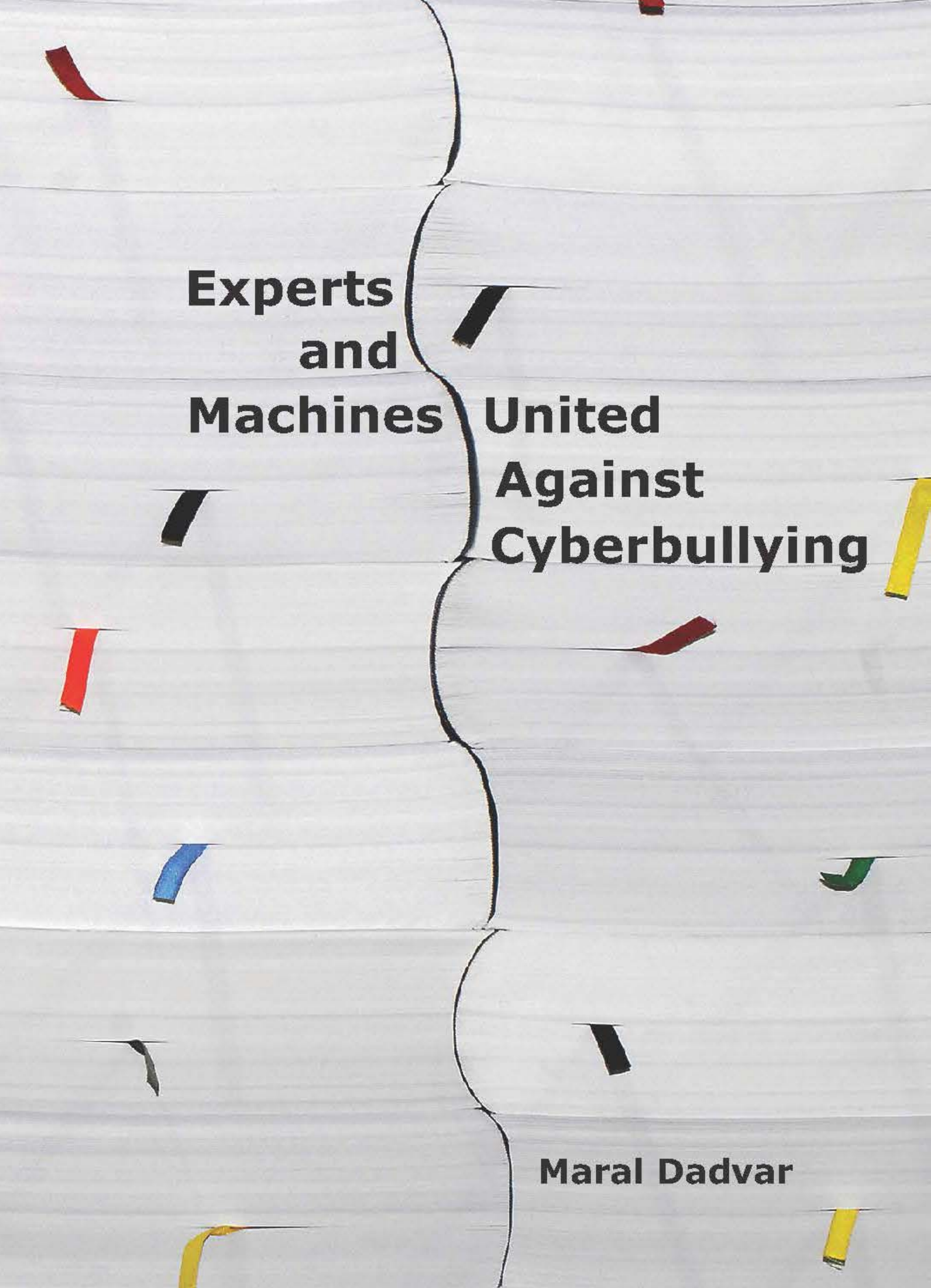
Experts and Machines United Against Cyberbullying
Maral Dadvar

EXPERTS AND MACHINES UNITED AGAINST CYBERBULLYING

Maral Dadvar

PhD dissertation committee:

Chairman and Secretary:

Prof. dr. P. M. G. Apers	University of Twente

Promotor

Prof. dr. F. M. G. de Jong	University of Twente

Members:

Prof. dr. D. J. Pepler	York University
Prof. dr. V. Hoste	Ghent University
Prof. dr. ir. U. Kaymak	Eindhoven University of Technology
Prof. dr. T. W. C. Huibers	University of Twente
Dr. C. H. C. Drossaert	University of Twente

CTIT Ph.D. Thesis Series No. 14-323
Centre for Telematics and Information Technology
P.O. Box 217, 7500 AE, Enschede, The Netherlands.

SIKS Dissertation Series No. 2014-37
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

This work was part of the PuppyIR project, which is supported by a grant of the 7th Framework ICT program (FP7-ICT-2007-3) of the European Union.

ISBN: 978-90-365-3739-1
ISSN: 1381-3617.14-323
DOI: 10.3990/1.9789036537391

EXPERTS AND MACHINES UNITED AGAINST CYBERBULLYING

DISSERTATION

to obtain

the degree of doctor at the University of Twente,

on the authority of the Rector Magnificus,

prof. dr. H. Brinksma,

on account of the decision of the graduation committee,

to be publicly defended

on Friday 12 September 2014 at 14:45 hrs.

by

Maral Dadvar

born on 19 September 1981

in Tehran, Iran.

This dissertation is approved by:

Prof. dr. F. M. G. de Jong (promotor)

Cover Photo: Courtesy of Maryam Zandi ©2007
Cover Design: Benno Masselink
Printed by: ITC Printing Department

Summary

One form of online misbehaviour which has deeply affected society with harmful consequences is known as cyberbullying. Cyberbullying can simply be defined as an intentional act that is conducted through digital technology to hurt someone. Cyberbullying is a widely covered topic in the social sciences. There are many studies in which the problem of cyberbullying has been introduced and its origins and consequences have been explored in detail. There are also studies which have investigated the intervention and prevention strategies and have proposed guidelines for parents and adults in this regard. However, studies on the technical dimensions of this topic are relatively rare. In this research the overall goal was to bridge the gap between social science approaches and technical solutions. In order to be able to suggest solutions that could contribute to minimizing the risk and impact of cyberbullying we have investigated the phenomenon of cyberbullying from different angles. We have thoroughly studied the origin of cyberbullying and its growth over time, as well as the role of technology in the emergence of this type of virtual behaviour and in the potential for reducing the extent of the social concern it raises.

First we introduced a novel outlook towards the cyberbullying phenomenon. We looked into the gradual changes which have occurred in relationships and social communication with the emergence of the Internet. We argued that one should look at virtual environments as virtual communities, because the human needs projected on these environments, the relationships, human concerns and misbehaviour have the same nature as in real-life societies. Therefore, to make virtual communities safe, we need to take safety measures and precautions that are similar to the ones that are common in non-virtual communities. We derived the assumption that if cyberbullying is recognized and treated as a social problem and not just seen as some random mischief conducted by individuals with the use of technology, the methods for handling its consequences are likely to be

more realistic, effective and comprehensive. This part of our study led to the conviction that for combating cyberbullying, behavioural and psychological studies, and the study of technical solutions should go hand in hand.

One of the main limitations that we faced when we started our research was the lack of a comprehensive dataset for cyberbullying studies. We needed a dataset which included real instances of bullying incidents. Moreover, it was essential for our studies to also have the demographic information of the social media users as well as the history of their activities. We started our preliminary experiments using a dataset that was collected from MySpace forums. This dataset did not meet all the requirements for our experiment, namely in terms of size and sufficiency of information. Therefore we developed our own YouTube dataset, with the aim to encompass extensive information about the users and their activities as well as larger numbers of bullying comments. We collected information on user activities and posted textual comments as well as personal and demographic details of the users involved.

Detecting a bullying comment or post at the earliest possible moment in time can substantially decrease the negative effects of cyberbullying incidents.

We started our experiments by showing that besides the conventional features used for text mining methods such as sentiment analysis and specifically bullying detection, more personal features, in this experiment gender, can improve the accuracy of the detection models. As expected the models which were optimized accordingly resulted in a more accurate classification. The improved outcome motivated us to look into other personal features as well, such as age and the writing style of users. By adding more personal information, the previous classification results were outperformed and the detection accuracy enhanced even further.

In the last experiment we made use of experts' knowledge to identify potential bully users in social networks. To better understand and interpret

the intentions underlying the online activities of users of social media, we decided to incorporate human reasoning and knowledge into a bulliness rating system by developing a Multi-Criteria Evaluation System. Moreover, to have more sources of information and to make use of the potential of both human and machine, we designed a hybrid approach, incorporating machine learning models on top of the expert system. The hybrid approach reached an optimum model which outperformed the results obtained from the machine learning models and the expert system individually. Our hybrid model illustrates the added value of integrating technical solutions with insights from the social sciences for the first time.

As argued in this thesis, the integration of social studies into a software-enhanced monitoring workflow could pave the way towards the tackling of this kind of online misbehaviour. The ideas and algorithms proposed for fulfilling this purpose can be a stepping stone for future research in this direction. The work carried out is also a demonstration of the added value of frameworks for text categorization, sentiment mining and user profiling in applications addressing societal issues. This work can be viewed as a contribution to the more general societal challenge of increasing the level of cybersecurity, in particular for the younger generations of social network users. By turning the internet into a safer place for children, the chances increase that they will be able to benefit from the informational richness that it also offers.

Acknowledgements

I have finally reached the end of the road that I started four years and couple of months ago. I knew it would be hard and would need a lot of effort but now I know that it needed encouragement and support most of all. Accomplishment of my PhD was only made possible by the advice and help of my colleagues and by the love and cheer of my family and friends.

I would like to thank my supervisor Franciska de Jong who gave me the opportunity to pursue this PhD. She offered her advice whenever needed and coped with my stubbornness at so many points. Without saying much she could always recognize my weaknesses and problems and solve them by just saying what I needed to hear.

Claudia Hauff deserves a big thanks for being a patient daily supervisor in the first year of my PhD when everything was Greek to me. She introduced me to the world of IR and she was a good friend and colleague since then. Most of all I should thank her for warning me about the tens of rounds of comments that I was going to receive from Franciska on my thesis. It really helped me to know it's not just me! I started a new phase of my research with Roeland Ordelman and he made me to become a stronger person at work. Throughout my PhD, specifically last 2 years, Dolf Trieschnigg's help, support and comments played an important role in finishing my thesis. His critical view on my work and our tough discussions improved my work enormously and made me prepared for the rest to come.

I would like to also thank Jimmy Huang who gave me the chance to spend three fruitful and wonderful months at his group at York University in Canada. I learnt a lot and I met many nice people. During this time I also had the opportunity to meet professor Pepler, professor Connolly and professor Mishna. Their input was of great importance to my work. I would like to also extend my sincere gratitude to professor Hoste, dr

Gutteling and dr Drossaert who were members of my expert panel and allowed me to benefit from their knowledge and experience.

Randy, thanks for being such a nice office mate and friend and keeping up with my mood swings, taking care of my plants (as I would say OUR plants) and trying your best to acquaint me with Dutch culture :D

Our coffee breaks was made even more fun with Andrea stopping by, explaining why life is so hard, talking about his secret connections in town and at some point trying to scare me and leave. Andrea you are a great friend.

Charlotte and Alice are the key persons in our group. I honestly believe without them everyone would get lost. I would like to specially thank them for dealing with all the bureaucracy and being patient and helpful at all time. Lynn I appreciate your time and effort for improving my writing and accepting my last minute requests.

Dear Anton thanks for always being so caring to everyone. I enjoyed our talks about Iran and I hope one day you finally visit there. From the early days of my PhD many kind people, Hayrettin, Andreea, Thijs, Sergio, Khiet, Mariet, Mannes, Hendri, Danny, Betsy, helped me to find my way around. All dear colleagues and friends at HMI, you made my stay more fun and I enjoyed many lunch, borrel, day-out, Christmas-lunch and cakes with you. Thank you all :)

It is hard to express the important role of my family, their love and support in this journey only in couple of words and sentences. It didn't start with my PhD but it started from the very first day of my life. My parents were always there for me and made me believe that I can be whoever that I want to be. My mum thought me to be a tough, strong, independent and intelligent woman and to take risks for getting to extremes. Reminding me that any problem will be either solved or passed, was her magical way of soothing me. My father followed me every single step from the day I left home to make sure I would never feel lonely, and I didn't. He was with me

as if he was getting his own bachelors, masters and PhD degrees. He is the most high-tech father ever, he knows it all; chat, video-chat, Viber, Skype, email and any other technology that links him to me. My precious mum and dad, I love you forever and I hope that my achievements in life have worth the long time that I have been away from you. My beloved Morvarid, your famous encouraging quote "you can do it" kept me going forever. Hearing my nags and in response reminding me that how intelligent and capable I am and nothing worth's my nerves was something that only a kind sister could tell me. Your lovely family, dearest Afshin and my precious little princess Darya, made my life even more beautiful.

My life changed for the best since the day that Aidin came to my life which was with the start of my PhD. He is all I could wish for. He is like a big complete Swiss knife in my life :D He coped with all my ups and very downs in the most loving and kind possible way in the past couple of years. Without his encouragement, advice and love I could never be where I am today.

Maral Dadvar
Enschede, August 2014

Table of Contents

Chapter 1

General Introduction

1.1 Introduction

The emergence of any new technology often imposes enormous changes in human lifestyle, and the invention of the World Wide Web and related technological innovations are no exception. Internet has changed almost all aspects of human life: education, entertainment, politics, relationships and so on. One of the most affected aspects is communication among people. Nowadays friendships and relationships are shaped through a wide array of digital devices. The majority of daily greetings, friendly get-togethers and family chitchats take place from behind a screen. In this thesis we will depict the emergence of a digitalized society in virtual environments: online platforms that facilitate the initiation and maintenance of relationships and interpersonal and community-level communication are shaped in accordance to the new standards for online interaction that have emerged together with the new virtual worlds. However, in spite of all the transitions that mark the genesis of a virtual society, the complexity of human nature has stayed the same, and like in any real-life community, the good and the bad come together. Most of the time people reach out to others for help, love and friendship, but hostility and hatred have also always been part of human culture and they have had determining impact on societal history. Virtual societies are no exception: the offensive wrongdoings and patterns of behaviour driven by the darker sides of human nature can be observed in virtual settings as well. The differences are few and mainly related to the fact that in the latter context the offender is empowered with features that are typical of the virtual world: anonymity of misconduct and impact that expands into the confinement of people's homes.

One form of online misbehaviour which has deeply affected society with harmful consequences is known as cyberbullying. Traditional bullying used to be a demonstration of dominance and consolidation of social status by making use of physical power and creating fear and discomfort for those who were weaker and vulnerable. With the development of online

technology, bullying has also emerged in cybersocieties, but in a new appearance. Cyberbullying can simply be defined as an intentional act that is conducted through digital technology to hurt someone. Unlike traditional bullying, which was inherently limited to streets and school yards, the vast variety of technological devices used in daily lives has brought cyberbullying also into people's homes and bed rooms.

The following posts are copied from social media networks and illustrate the phenomenon of bullying that takes place in cyberspace:

- "*you are ugly and fat. You have no friends and no one will ever love you. Why do you even bother to come to school anymore freak!*"
- "*How does it feel to be the most hated person right now? You are a puke and disgrace to the human race.*"
- "*u r soooo desperate…STOOPID SLAG!!!*"
- "*Looks like yew lost weight, what are yew now 5000 pounds?*"

Cyberbullying is a widely covered topic in the social sciences. There are many studies in which the problem of cyberbullying is introduced and its origins and consequences have been explored in detail (Lamb et al., 2009, Cappadocia et al., 2013). There are also studies which have investigated the intervention and prevention strategies and have proposed guidelines for parents and adults in this regard (Campbell, 2005, Kowalski et al., 2008, Smith et al., 1999, Tokunaga, 2010, Dilmaç and Aydoğan, 2010). However, studies on the technical dimensions of this topic matter are relatively rare. Moreover, for almost all of the few technical studies conducted on cyberbullying (Dinakar et al., 2012, Dinakar et al., 2011, Yin et al., 2009, Reynolds et al., 2011) two common gaps can be observed. First, the approaches proposed for detecting bullying incidents

and taking required actions afterwards, rarely incorporate the findings of the social studies for improving the accuracy of the proposed cyberbullying detection models. Second, solutions presented in state-of-the-art literature are on *detection* of bullying incidents after they have happened and there is hardly any study on *prevention* of cyberbullying by the deployment of computational models.

In this research the overall goal was to bridge the gap between social science approaches and technical solutions. In order to be able to suggest solutions that could contribute to minimizing the risk and impact of cyberbullying we have investigated the phenomenon of cyberbullying from different angles. We have thoroughly studied the origin of cyberbullying and its growth over time, as well as the role of technology in the emergence of this type of virtual behaviour and in the potential for reducing the extent of the social concern it raises.

We also explored the potential for applying methods from the field of information technology and more in particular from the domain of natural language processing and artificial intelligence in the design of measures and solutions for the automatic detection of bullying incidents. Based on the assumption that for the detection of cyberbullying incidents the analysis of textual content posted in online media platforms is one of the challenges, we started our study with an assessment of the applicability of the wide variety of natural language processing methods that have been developed for sentiment analysis and data mining tasks, such as analysing movie reviews or consumers' opinion (Alm et al., 2005, Pang and Lee, 2008, Zhuang et al., 2006). This choice was partly given in by the fact that in the past decade natural language analysis has been expanded to be used for the detection of cybercrimes and supporting law enforcements in combating against terrorism, fraud and cyber-attacks (Hughes et al., 2008, Tsai and Chan, 2007, Chen et al., 2004). Moreover, similar fields of research have found their ways in to artificial intelligence while ago, and specifically expert systems have been used to support police investigations in online crimes (Ratledge and Jacoby, 1989, Brahan et al., 1998).

Based on the insights gained from related studies and methods, we first designed methods based on text mining algorithms and applied them to posts from social media platforms in order to detect bullying incidents. In these methods we integrated profile information of the users in order to take demographic differences into account.

In a second stage we improved the effectiveness of the algorithms by integrating the findings of social studies on cyberbullying. These findings allowed us to develop detection models that incorporate expert knowledge on how to weigh personal characteristics of social networks' users. The models were also used to measure the probability of a user to be a bully in social networks by assigning a *bulliness score* to each user. The higher the score is there is a higher chance that the user is a bully and will conduct further misbehaviours in online environments.

We think that the outcome of our studies contributes to increase the potential of natural language processing and data driven methods to be successfully deployed in the battle against the societal problems of the virtual age and in particular against cyberbullying. This thesis can also be seen as a demonstration of how text mining can be enhanced by the coupling of data-driven machine learning models and knowledge-driven methods.

1.2 Research Motivation

The appearance of novel technology usually comes with excitement and optimism about the advantages that it can bring to human lives and the way it could enhance lifestyles for the better. But after a while often some troubling consequences, predicted or not, also become apparent. The emergence of social networks has enormously affected and changed communication and relationships in society. However, not all the changes

were favourable for the people involved. Cyberbullying is one of the problems which emerged with the growing use of social networks.

There is a variety of online social networks, such as Facebook, Twitter and YouTube, in which mostly teenagers and adolescents are active. Based on a recent annual cyberbullying survey[1] conducted on teenagers and adolescents from UK, the USA, Australia and other countries, 7 out of 10 young people have been victim of cyberbullying. The survey showed that the top three social networks frequently used by Internet users are Facebook (75%), YouTube (66%), and Twitter (43%). These three social networks are also found to be the most common networks for cyberbullying as 54%, 21% and 28% of their users have experienced cyberbullying respectively. Cyberbullying is found to have catastrophic effects upon the self-esteem and social lives of up to 69% of the youngsters. Studies show that youngsters who have experienced traditional bullying or cyberbullying have more suicidal thoughts and are more likely to attempt suicide. There have been several high-profile cases from all over the world involving teenagers taking their own lives in part because of being harassed over the Internet (Hinduja and Patchin, 2010).

All these facts, numbers and sad reports, have raised the question of what suitable solutions there could be for this problem and what is lacking in the existing strategies for dealing with cyberbullying incidents. An obvious idea is to design an alerting system that when integrated, the social networks could detect the bullying incidents with a certain accuracy and could send a warning for the administrators of the networks. Even better, if a system could prevent the bullying incidents from happening in the first place, then the number of people negatively affected by this phenomenon could be decreased to a great degree.

[1] http://www.ditchthelabel.org/annual-cyber-bullying-survey-cyber-bullying-statistics, [Accessed November 2013].

1.3 Research Objectives

The severity of the cyberbullying problem motivated us to dig deeper and to look for means that can overcome the observed shortcomings of the existing solutions (indicated in Section 1.1) and that can help to decrease the negative consequences of cyberbullying in teenagers' and adolescents' lives (pointed at in Section 1.2) by introducing new approaches and techniques that could be deployed in the detection and prevention of bullying incidents. This aim led us to take up the following four main research objectives.

- Objective 1: To present a view on cyberbullying that underlines the kinship with traditional bullying.

  The aim is to illustrate the dynamics in communication and relationships introduced with the emergence of Internet in everyday life. We show that virtual environments represent and act as a society of which participants demonstrate behaviour that is similar to what can be observed in real-life society, and argue that as a consequence the interventions and precautions toward social misbehaviours such as cyberbullying should be similar to the ones that are known to be effective in real-life societies.

- Objective 2: To create a comprehensive dataset to be used in cyberbullying studies.

  One of the main challenges that were faced during this research was lack of suitable and available dataset for research into cyberbullying detection and into digital tools that could contribute to its prevention. The required dataset has to contain a balanced number of bullying and non-bullying comments from a variety of social media platform users. It should include certain types of metadata, such as demographic information for the authors of posts, as well as details on the history of their network activities.

- Objective 3: To improve the accuracy of algorithms for the detection of bullying comments in social networks.

  In the context of this objective, the following two research questions were investigated:

  - Research Question 3.1: Does considering gender information for bullying network users improve the accuracy of cyberbullying incident detection in social networks?
  - Research Question 3.2: Does considering further user profile information for bullying network users, such as age and history of comments, improve the accuracy of cyberbullying incident detection in social networks?

- Objective 4: To design a bulliness likelihood score for identifying potential bullies in social networks.

  The aim is to measure the likeliness of social network users to exhibit bulling behaviour in the future by calculating a bulliness score for each user. Hereafter we refer to this score as bulliness score.

  In the context of this objective, the following two research questions were investigated:

  - Research Question 4.1: How accurately can an expert system assign a bulliness score to a user to represent the level of bulliness of that user?
  - Research Question 4.2: Can an expert system and a system based on machine learning be effectively combined for detecting potential bullies?

Figure 1.1 depicts the way in which the results of this thesis could be integrated as decision support tools for the human agents operating the monitoring/administration environment for social networks. It should be

noted however that the design of such an environment is out of the scope of our study.

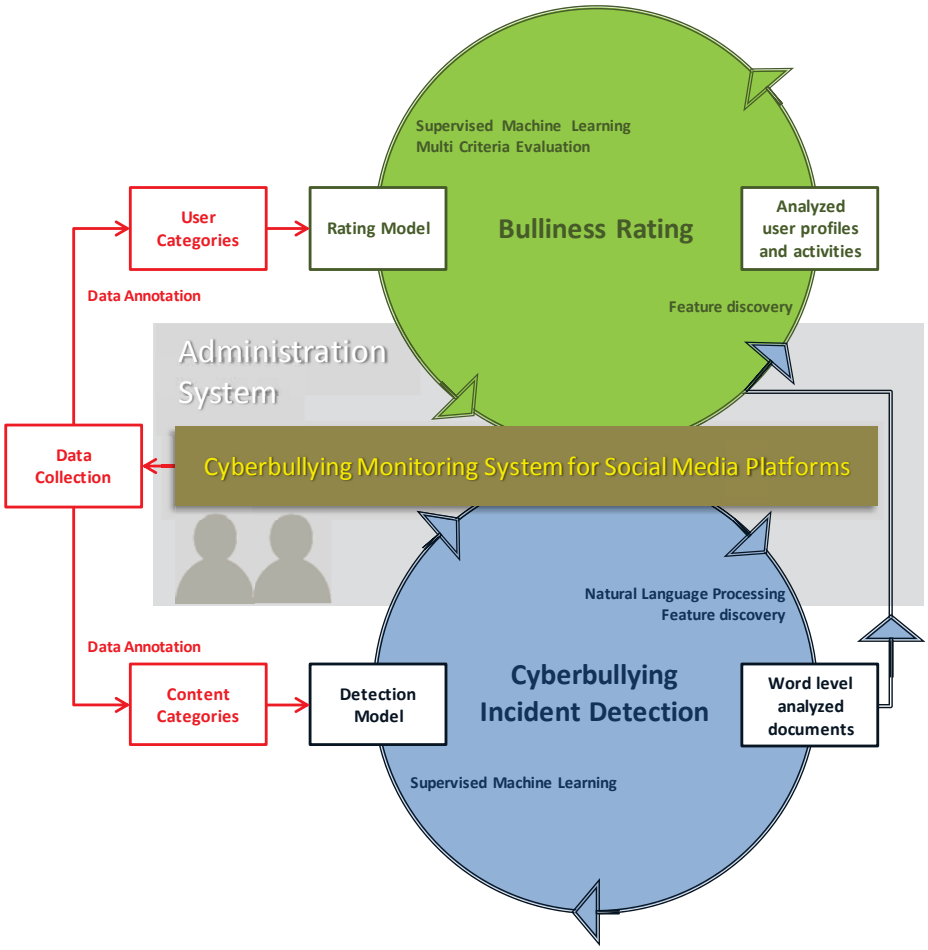

Figure 1.1 Flow diagram representing the way in which the results of this thesis could be integrated in a monitoring environment for social networks (depicted in gray). The parts in red represent the collection and preparation of training data (Objective 2). The part in blue represents the work related to the detection of cyberbullying incidents (Objective 3). The part in green represents the work related to the rating of social media users (Objective 4).

1.4 Structure of the Thesis

The organization of the thesis follows the order of the research objectives formulated above (see Figure 1.2). In Chapter 2, Objective 1 is addressed by describing the transformation of society and lifestyle since the emergence of Internet, and by pointing to the positive and negative effects of the changes it generated for interpersonal relationships. Moreover, it suggests how safety and misbehaviour in virtual communities can be seen as mirroring their counterparts in real life.

The lack of standard datasets for cyberbullying studies is the background of Objective 2 and brought us to develop a dataset to be used in the experiments conducted. Chapter 3 explains the process of data collection as well as the attributes and characteristics of the datasets.

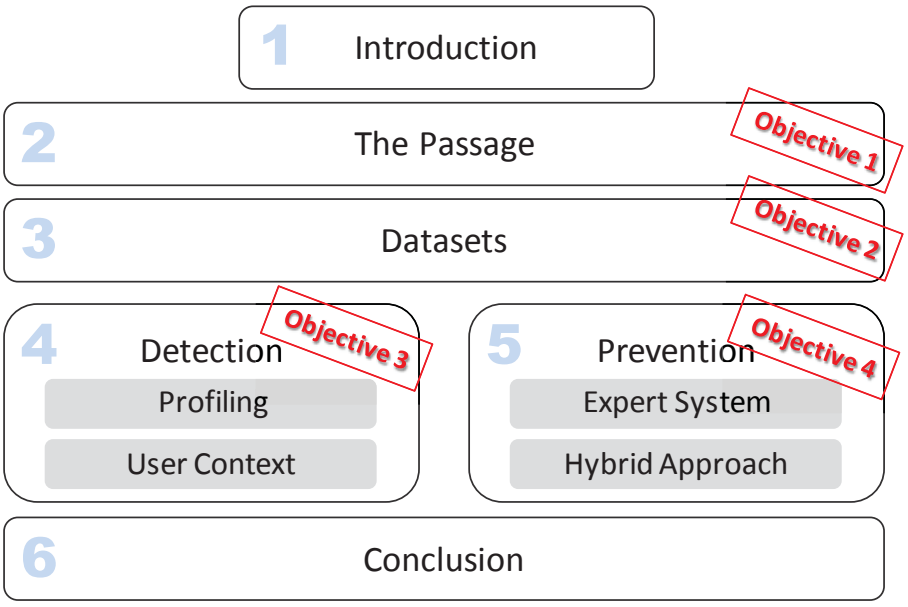

Figure 1.2 Structure of the Dissertation

Objective 3 and the two related research questions regarding detection performance are addressed in Chapter 4. In this chapter we look into the options for the incorporation of personal features of users into the models for the detection of cyberbullying incidents as well as the history of online activities of users for improving the accuracy of cyberbullying detection. The set of research questions related to the concept of a bulliness score which is inherent to Objective 4 is answered in Chapter 5. These research questions are related to a novel approach for discriminating among potential bully and non-bully users by weighing social network users for their likeliness to develop future misconduct. In this chapter we demonstrate that combining the advantages of expert's knowledge and machine learning can improve the discrimination capacity of such scoring mechanisms. Chapter 6 concludes the thesis with a summary of the results and offers suggestions for future research.

Chapter 2

Passage; from Bullying to Cyberbullying

The emergence of the Internet has proven to be a turning point in human culture. Internet has affected relationships, communications and friendships. Initially the depth and impact of the introduction of this technology on society was not as tangible as it currently is. Nowadays Internet has become an inevitable part of our daily lives, intertwined with almost all aspect of human behaviour and literally touching upon the way we interact with the objects and structures that surround us. In this chapter we introduce an outlook on ubiquitous role of Internet and its impact on society that is novel in its focus on social communication patterns that are considered a threat not just in cyberspace but to the society at large. The idea of virtual society was introduced a decade ago and was the basis of several popular games and platforms such as 'Second Life[1]': a virtual society that for many people and organisations became a crucial context for a major part of their activities. What we will address is that in the context of online activities such as entertainment, communication and trading, , not only basic regulations and social conventions can be identified, but that also, like in any society, concerns and risks related to social and criminal misbehaviours emerge that need to be confronted. Here we specifically focus on the problem of cyberbullying. We explain the components and impacts of cyberbullying and we argue how any path towards a solution has to draw upon the social studies' findings as well as upon the potential of digital tools. This chapter addresses Objective 1 of the thesis: introducing a novel outlook towards cyberbullying in virtual society.

[1] http://www.secondlife.com/

2.1 Introduction

The Internet is present in everyday life. On the Internet we search for information, plan trips, order products and read news; we communicate with others by making use of email and chat rooms; we listen to music and watch videos; we meet others, have discussions with others, find friends and fall in love, we get involved in other people's happiness and sadness; we protest, play games and learn; we share ideas; we download software and so on. The internet also affects our mood: we feel connected, happy, loved, lonely, depressed, scared and so forth.

Maybe not willingly, but undoubtedly our lives have become interwoven with Internet. But how has the web transformed our lives? What are the positive and negative effects on the society and on our interpersonal relationships? Have we built a virtual community next to the real one that we are living in? In this virtual community, what are the boundaries and restrictions of relationships? How are safety, privacy and misbehaviours defined and treated?

This chapter contributes to articulating these issues, and to finding answers to these questions. It will explain how the role of internet has changed over time and how this has resulted into new definitions of relationships and communication. The aim of this chapter is to demonstrate how the problems and concerns of the virtual community are similar to those encountered in real-life communities, and to show that to avoid risks and prevent negative consequences it is required to take measures and precautions in ways that are similar to real-life strategies. We specifically describe an old troubling problem, known as bullying, and we explain how it has entered the virtual environments and is now known as cyberbullying. It will be described that the problem originates from and/or mirrors aspects of real-life societal phenomena and human nature, that it requires measures that go beyond the potential of digital tool boxes.

2.2 Citizens of Information Universe

Nowadays the Internet is an inevitable part of the majority of people's life in developed and developing countries and is the main source of information mainly used for entertainment, education, communication and other social activities. However, the amount of time spent on the Internet and the extent to which it is ubiquitous in everyday life differs cross countries and societies. The amount of internet use depends on the social background of the users and there are several other factors that have an effect on it, such as economics, system functionality, privacy regulations and most importantly age and education (Nie and Erbring, 2000, Välimäki, 2012).

The International Telecommunication Union reports over 2.7 billion people are using the Internet worldwide (ITU, 2013). In the developing countries, 31% of the population is online, compared with 77% in the developed countries. Europe is the region with the highest Internet penetration rate in the world (75%), followed by the Americas (61%). Studies by Nie et al. (2000) and Välimäki (2012) show that the highest rate of Internet use (91%) exists among 16-24 years old individuals, compared with a 40% rate among users above 60 years old. The studies also illustrate that a college education increases internet access by over 40% compared to the figures for the least educated individuals. Knowing all these facts and figures raises questions about the transition of the habits of all these people: from writing letters, talking in the streets and playing in the school yards, to using their computers to do all these.

As said, Internet has influenced and modified almost all personal and social aspects of life: communication, education as well as health, economy, politics and democracy. This chapter specifically focuses on the changes in personal and social relationship and communication as a highly affected aspect. We are interested to know when these changes were for better or for worse, and how we can overcome some of the negative consequences that resulted from these changes. In the following sections we will briefly

explain the transformation of social media's position in human life and how this transformation has resulted into a notion of real-life community in cyberspace, often referred to as 'virtual community'.

2.2.1 Dynamics in the Appreciation of Social Media

Since online communication technologies such as email and chat rooms became popular in the 1990s, the formation of friendships, relationships and communication has started to change. Face-to-face conversations and hangouts with friends and family partly shifted to online communication with faceless strangers in chat rooms. Since the start of this revolution, there have been debates about its overall positive or negative effects and consequences.

At first, with low-level one-to-one online communication, it was assumed that the Internet motivates adolescents to form superficial online relationships with strangers that are not as meaningful as their real-world relationships, and that time spent with online strangers occurs at the expense of time spent within existing relationships (Nie, 2001). Several studies in the early years of the Internet, conducted among adolescents and adults, demonstrated the negative consequences of Internet use on social well-being and involvement. For example a study by Kraut et al. (1998) showed that Internet use reduced adolescents' social connectedness with a period of 1 year (Kraut et al., 1998). In addition, Nie et al. (2000) demonstrated that adults who spent more time on the Internet spent less time with friends. Finally, Mesch (Mesch, 2001) found that adolescents who had fewer friends were more likely to be Internet users.

However, as communication technologies improved and developed into higher-level social media, the overall believes on their negative effects also changed. Early online communication used to take place between strangers in chats rooms, but in recent years new technologies such as Instant Messaging and social networking sites such as Facebook, encourage

communication with existing friends. Recent Internet studies have demonstrated that adolescents' online communication stimulates, rather than reduces, social connectedness. For example, in a 2-year follow-up study based on their initial studies on Internet impact, Kraut et al. (2002) found that Internet use improved social connectedness and well-being (Kraut et al., 2002). Several other recent studies have demonstrated significantly positive relationships between online communication and adolescents' social connectedness (Bessière et al., 2008) (Valkenburg and Peter, 2007). In another study (Peter et al., 2005), the motives for online communication are investigated and the findings indicate that adolescents who are introvert and have difficulties to interact, are strongly motivated to communicate online to compensate for lacking social skills. This increases their chances of making friends online. Social networks facilitate sharing personal information (or self-disclosure) which is an important aspect of relationship development both online and offline (Steijn and Schouten, 2013). Self-disclosure can lead to more closeness, intimacy and more trust between partners as well as to the development of new relationships (Sheldon, 2009, Steijn and Schouten, 2013, Park et al., 2011).

Obviously, the effects of Internet on relationships and communication cannot be generalized (Ruggiero, 2000). People's use of media and their effects may differ from what the media's objective would suggest. People's motives for use of Internet can determine its consequences on their relationships. However, one thing that it is widely agreed upon is that people log on to newsgroups and social networks for the same reason they might hang out at a bar or a school yard corner or at the coffee machine at work; they have either something to say or an ear to lend to those who do (Porter, 1996). The Internet provides each individual user an opportunity to speak and to portray their self or to construct an identity (Porter, 1996). This empowerment and the ability to connect to other people encourage a sense of community.

2.2.2 From Social Inhibition to Social Empowerment

The conceptual space in which online communication occurs is often referred to as "*cyberspace*" (Porter, 1996). In cyberspace a form of virtual presence can be established as a result of individual electronic interactions not being restricted by traditional boundaries of time and space; this electronic interactions is the basis of what is commonly referred to as *"virtual community"* (Porter, 1996). In an earlier study in 1993, Howard Rheingold has defined the concept of virtual community as "social aggregations that emerge from the Net when enough people carry on discussions long enough, with sufficient human feeling, to form webs of personal relationships in cyberspace" (Rheingold, 1993).

Communication among people is not the only thing that happens in virtual communities. These communities deal with matters that are related to needs and interests of human nature as well as their problems and concerns. The whole spectrum of interpersonal dynamics is adjusted to the special conditions of virtual communities; there are unique virtual indications of respect, love and bounding alongside indications of harassment, violence and hostility. For example these emotions and intentions can be expressed through small icons known as "emoticons" or through an extra exclamation mark at the end of a sentence or even by not reacting to an online post.

Like in any community, in a virtual community a variety of crimes, threats and misbehaviours take place that should be taken care of taking into account the way in which their nature has been adapted to the digital settings. Therefore, in a cybercommunity, some form of cybersecurity is required to protect us from cybercrimes (von Solms and van Niekerk, 2013). The International Telecommunication Union defines[1] cybersecurity as the collection of tools, policies, security concepts, security safeguards,

[1] http://www.itu.int/en/ITU-T/studygroups/com17/Pages/cybersecurity.aspx [Accessed August 2013]

guidelines, risk management approaches, actions, training, best practices, assurance and technologies that can be used to protect the cyber environment and organization, and the users' assets. These tools and policies are selected to be put into action according to the crime that they are used for. Like traditional crime, cybercrime has different facets and it occurs in a wide variety of scenarios and environments. With the growth and improvement of technology, the design and severity of the cybercrimes also changed. Definitions of cybercrime differ depending on the people involved (victim, protector and bystander), and have evolved with the evolution of computer-related crimes (Gordon and Ford, 2006). The United Nations Manual on the Prevention and Control of Computer Related Crime on 1995 used the word cybercrime to refer to offences ranging from fraud and forgery to unauthorized access of online information. [United Nations: The United Nations manual on the prevention and control of computer related crime, 1995, supra note 41, paragraphs 20 to 73 in International Review of Criminal Policy, pp. 43–44 (1995)]. As the availability of online information and data sources improved, the definition was also modified to include criminal activity against data and copyright infringement (Krone, 2005, Zeviar-Geese, 1997). With the appearance of social networks and online communication, more recent studies (von Solms and van Niekerk, 2013) suggest a broader definition, including activities such as unauthorized access as well as child pornography, cyberterrorism, fraud and cyberbullying.

The definitions suggest that all crimes are disturbances that need to be tackled and stopped. Although it is not possible to address them all at once and each of them is in itself a broad topic to be studied and investigated, they provide important background information for the problem that will be addressed in the rest of this chapter: cyberbullying. Cyberbullying is a growing and troubling issue which has mostly targeted the young generation. Although bullying also happens among adults and at work places, we focus on cyberbullying among teenagers and adolescents as they

are more vulnerable towards adversities and effects of Internet and social networks (Allison and Schultz, 2001).

In the coming sections of this chapter, we present the definition of cyberbullying and will see how bullying has transformed over time from physical bullying into cyberbullying. We also explore the consequences and threats of this problem as well as the measures that could be taken from social, technical and legal perspectives.

2.3 Cyberbullying; Bullying in the Internet Yard

Bullying is usually defined as a subcategory of aggressive behaviour (Smith et al., 1999). It is characterized by repetition over time and an imbalance of power between bully and victim (Smith and Sharp, 1994). In the 1980s bullying was mostly seen as direct face-to-face physical (such as hitting) and verbal (such as teasing) attacks (Slonje and Smith, 2008). During 1990s the scope of bullying has been broadened to also include indirect aggression, such as spreading rumours, and relational aggression, for example by damaging someone's relationships (Bj rkqvist et al., 1992). In recent years, with the development of technologies and growth of Internet use, a new form of bullying has emerged, called cyberbullying.

Cyberbullying is a general term that also refers to similar constructs such as online bullying and Internet harassment. There are different categories of common cyberbullying (Willard, 2007, Beran and Li, 2008):

- Flaming: Sending rude and vulgar messages to a group or person.
- Outing: Posting private information (picture, phone number,…) or manipulated/photo-shopped personal materials of an individual without her or his consent.
- Harassment: Repeatedly sending insulting messages or emails to a person.

- Exclusion: Excluding someone from participating in an online group.
- Impersonation: Pretending to be someone else and sending out materials on her or his behalf.
- Cyberstalking: Terrorising someone by sending threatening and intimidating messages.
- Denigration: Spreading online gossips about a person.

There is a certain lack of conceptual clarity in the definition of cyberbullying and the distinction among different types of cyberbullying is often vague (Vandebosch and Van Cleemput, 2008). Several definitions of cyberbullying are suggested in the literature and all of them somehow refer to an aggressive and harmful act which is conducted through an electronic device. However, these definitions can be distinguished through their details, such as those who are involved in the incident (groups and individuals) and requirements for being deliberate and repeated overtime (Tokunaga, 2010). Table 2.1 presents some of the definitions of cyberbullying suggested in the literature. However, Dehue and colleagues (Dehue et al., 2008) suggest that a situation must meet three conditions to be considered as cyberbullying; the act should be intentional, be repeated over time and should involve psychological torment.

Cyberbullying can happen through different modalities. It can happen through posting nasty videos about someone or publicly uploading private pictures without having the consent of their owner. Cyberbullying through text is one of the most common mediums, in which vulgar comments are posted and threatening and foul messages are sent to the victim.

In this research we prefer the definition given by Smith and colleagues (Smith et al., 2008) because it thoroughly encompasses all aspects of cyberbullying. They define cyberbullying as “an aggressive, intentional act

carried out by a group or individual, using electronic forms of contact (e.g. email and chat rooms), repeatedly or over time, against a victim who cannot easily defend him or her-self".

Table 2.1 Definition of cyberbullying in several studies

Literature	Definition
Smith et al. (2008)	*An aggressive, intentional act carried out by a **group** or **individual**, using electronic forms of contact, **repeatedly** or **over time** against a victim who cannot easily defend him or herself.*
Tokunaga (2010)	*Cyberbullying is any behaviour performed through electronic or digital media by **individuals** or **groups** that **repeatedly** communicates hostile or aggressive messages intended to inflict harm or discomfort on others.*
Patchin and Hinduja (2006)	*Wilful and **repeated** harm inflicted through the medium of electronic text.*
Juvoven and Gross (2008)	*The use of the Internet or other digital communication devices to insult or threaten someone.*

However, this definition has aspects which cannot be fully covered when it is considered in experimental settings for studying cyberbullying from a technical perspective, e.g., developing algorithms and tools that can automatically detect and remove bullying posts, or trigger some kind of an administrator's follow-up action in response to online bullying incidents. For example, the repetitiveness of the act cannot always be determined, as part of incidents may happen in private conversations which are not accessible. Moreover, the balance of the power between the victim and the bully cannot be easily verified by just analysing the content of the bullying

incidents. Therefore, in our studies we look into aggressive, intentional act carried out by an individual, through textual content, against a victim.

2.3.1 Components of Cyberbullying

Cyberbullying consists of several components. These components affect how the bullying takes place and consequently the studies conducted on cyberbullying differ depending on the components involved. The components under study should be clarified and carefully selected to make sure that their differences are taken into consideration and the proposed approaches match the nature of each component. The components are depicted schematically in Figure 2.1.

- The fundamental component is the people, called ***actors***, involved in the incident. The actors can be grouped into the following three categories:
  - Bully: the person who intentionally uses obscenity, threat or aggression to impose domination or cause fear and distress in others.
  - Victim: the person who is targeted by the bully. Victims cannot easily defend themselves and are usually vulnerable to the imbalance of power between them and the bully.
  - Bystander: the person who witnesses the incident but is not directly involved in the process. The bystanders can provide support for the victim by posting positive feedbacks for the victim and reacting against the bullies. They can also escalate the distress caused by the bullies, by supporting their actions.
- The ***platform*** in which cyberbullying takes place is another influential component in the process and therefore it should also be taken into consideration in the studies. Online social networks are the main communication platforms. An online social network is a web-based

platform to build social relations among people with similar interests and activities. Social networks introduce each of their members through her/his personal page (profile) which mostly contains personal information and interests of the user. Networks also provide means for users to interact over the Internet, for example through e-mail and instant messaging. Social network sites are varied and they offer different activities such as photo and video sharing, posting comments and following the activities of others in the network. In some cases, part of the dynamics comes from the presence of a monitoring function that could help to discourage bullying behaviour.

– Another component is the ***content*** and the ***modality*** through which the bullying takes place. As explained earlier, cyberbullying can happen through videos, pictures as well as through posting hurtful and offensive textual contents.

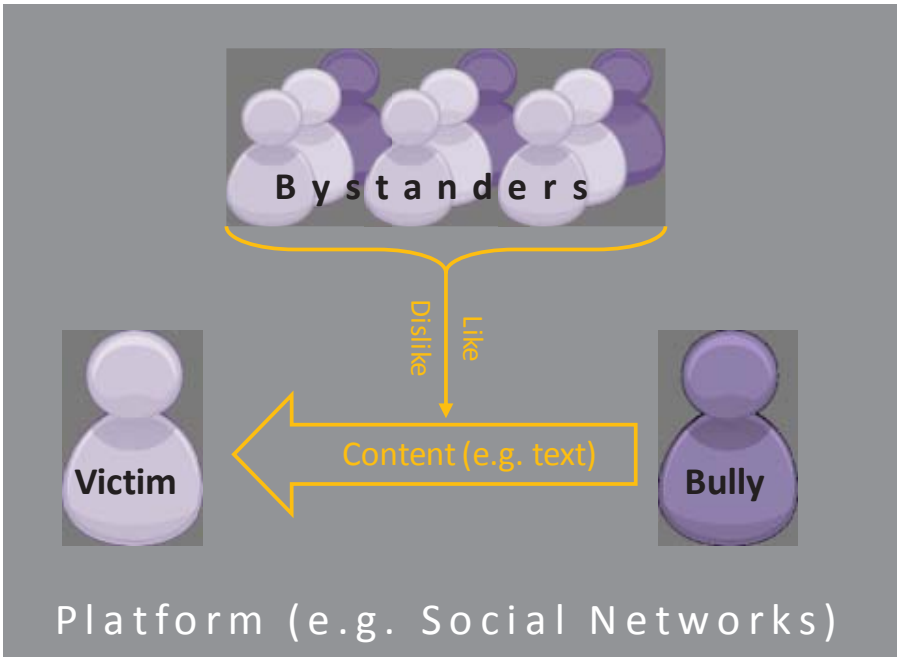

Figure 2.1 Schematic representation of components of cyberbullying

2.3.2 Impact of Cyberbullying

Studies show that in European countries about 18% of the children have been involved in cyberbullying via Internet or mobile phones (Hasebrink et al., 2008). A survey conducted in Britain shows that 25% of adolescents between 11 to 19 years old, have experienced cyberbullying (National Children's Home, 2002). The National Crime Prevention Council reported[1] in 2011 that cyberbullying is a problem that affects almost half of all American teens.

The consequences of cyberbullying are similar to traditional bullying, and have been shown to include depression, low self-esteem and in cases even ending up to suicide attempts (Campbell, 2005, Dehue et al., 2008, Patchin and Hinduja, 2006, Bannink et al., 2014, Smith et al., 2008, Bucchianeri et al., 2014). However, in some cases the consequences of cyberbullying can be more severe and longer lasting due to some specific characteristics of cyberbullying. Cyberbullying can be undertaken 24 hours a day, every day of the week, and unlike traditional bullying, it is independent of place and location (Shariff and Patchin, 2009). Moreover, online bullies can stay anonymous (Kowalski and Limber, 2007, Shariff, 2008, Ybarra and Mitchell, 2004) and being bullied by an unknown person can be more distressing than being bullied by someone familiar (Kowalski et al., 2012). Furthermore, anonymity triggers cyberbullying behaviour for people that would not bully face-to-face (Campbell, 2005).

Online materials spread very fast and in couple of minutes thousands of Internet users can see whatever that goes online (Shariff, 2008, Kowalski and Limber, 2007). There is also the persistency and durability of online materials and the power of the written word (Campbell, 2005). In the case of cyberbullying through text, the targeted victim and bystanders can read what the bully has said over and over again, and also in the case of images the hurtful content can stay online for a long period of time and if tagged

[1] http://www.ncpc.org/ [Accessed July 2011]

with the name or other personal features of the victim it will keep showing up in the results of searches.

2.3.3 Phases of Cyberbullying

In traditional bullying the moment at which the bullying takes place can be clearly recognized. The kicking, cursing and biting are evident indicators that signal the moment of bullying. Therefore, the social studies on the bullying problem can easily be divided into those which propose preventive training and awareness raising programmes for the stages before the bullying happens, and those which provide support and guidance for the consequences of bullying after an incident. Unlike for what is the case in traditional bullying, it is very difficult to determine the exact moment in which cyberbullying takes place. Therefore, in technical studies on cyberbullying such divisions have not been considered. However we consider the availability of a conceptual framework in relation to which we can discuss the various components of cyberbullying and the measures proposed an essential condition for a clear presentation of the various dimensions of the study. Therefore, following traditional bullying studies, we propose a framework for discussing the phenomenon of cyberbullying and suggest to split up the problems, possible solutions and precautions related to cyberbullying according to the two main phases of the entire chain of activity and reaction: the pre-bullying phase and the post-bullying phase. The studies we will present will mostly deal with each phase separately. In the study of measures addressing the pre-bullying phase the main concentration is on prevention strategies while in the study of measures addressing the post-bullying phase the focus is on the detection of bullying incidents after they have happened. Computational models for the detection of risky user profiles typically require information on previous cyberbullying incidents. Note that in order to come up with alerts suggesting action that could be taken to stop or decrease future harmful acts by a bully, the pre-bullying models need input from the models for the

detection of cyberbullying incidents, which are applied in the post-bullying phases.

Table 2.2 Cyberbullying components in pre- and post-bullying phases and the actions that could be triggered by the prediction modules proposed.

		Pre-Bullying	Bullying	**Post-Bullying**
Actors	***Bully***	To be monitored		To be identified / To be warned or to be excluded from the network
	Victim	To be trained To be educated		To be identified / To receive support
	Bystanders	To be alerted To be monitored		To be alerted To be monitored
Platform		Exclusion of risky user profiles		Identification of bullies and victims. Follow-up actions, e.g., organizing help after incident, alerting of bystanders, removing offensive
Content		Previously analysed content to be used to identify risky user profiles		Bullying content to be detected, offensive content to be deleted

Ultimately we envisage a monitoring framework that integrates element from the models that capture and weigh the signals picked up from what is going on in the various phases in order to alert the social media stewards that an intervention may be needed. For clarification, Table 2.2 illustrates the status of the components distinguished in each phase.

2.4 Confronting Cyberbullying

In general the cyberbullying problem can be approached from two perspectives, social and technical. Consequently the ingredients for policies and strategies for tackling this problem would stem from these domains of studies. In the following sections we present a series of studies and solutions conducted on cyberbullying from both social and technical perspectives.

2.4.1 Social Solutions

Many social and psychological studies (Dilmac, 2009, Rivers and Noret, 2010, Tokunaga, 2010, Mesch, 2009) are dedicated to cyberbullying problem and both pre- and post- bullying phases are thoroughly addressed in these studies. The severity of the problem has brought many countries and research institutes to work together and to share expertise on cyberbullying specifically in educational settings, coping with negative consequences and enhancing positive uses of new technologies and moving towards a common set of guidelines. An example is "COST Action IS0801 Cyberbullying"[1] running from 2008 till 2012 with partners from 28 countries. The main objectives of this Action were: sharing expertise and measurement techniques across researchers, as well as sharing of input from outside the research community, specifically from legal experts. Another goal of this Action was to distribute the nationally published

[1] https://sites.google.com/site/costis0801/ [Accessed July 2012]

guidelines and recommended coping strategies in different countries and to move towards a common set of guidelines applicable for the European Community. And finally the goal was to increase awareness about the cyberbullying problem.

Social studies on the prevention of bullying (pre-bullying phase), show that there are several ways to reduce the incidence of bullying in schools (Campbell, 2005, Smith and Ananiadou, 2003, Olweus, 2013). One of the first steps in any prevention program is to make people aware of the problem (Besag, 1989). Teachers, parents and youngsters need to be made aware of cyberbullying in particular as well as bullying in general. For this purpose many online portals have been developed across nations which provide awareness about cyberbullying and educate their audience about coping strategies and provide information about things that should be done to help the victims and prevent future harms. Table 2.3 illustrates few examples.

Another step is education. Adults should become acquainted with the existing technologies and online environments, to be able to provide the necessary guidance for the youngsters. Teenagers and adolescents should also be educated about the effects and consequences of bullying as well as coping strategies. In the same fashion that adults supervision of youngsters' activities in the playground may decreases the incidence of face-to-face bullying (Smith and Shu, 2000), online activities of adolescents have to be monitored and supervised. The monitoring can be done both by the adults supervising the online activities of youngsters at home and school, and by the administrators of the online communities, websites and forums. Regarding post-bullying phase, several studies have been conducted to provide coping strategies and solutions for the victims of cyberbullying to overcome its negative social and emotional effects (Machmutow et al., 2012, Perren et al., 2012).

Table 2.3 Examples of online portals provides education and awareness against cyberbullying.

Country	Portal name	About the portal	URL
Canada	PREV Net	Promoting relationships and eliminating violence network.	prevnet.ca
European Union	Insafe	Promoting safe, responsible use of the Internet and mobile devices to young people.	insafecommunity.saferinternet.org
Canada	Stop Cyberbullying	First cyberbullying prevention program in North America.	stopcyberbullying.org
USA	Cyberbullying Research Centre	Dedicated to providing up-to-date information about the nature, extent, causes, and consequences of cyberbullying among adolescents	cyberbullying.us
The Netherlands	Mijnkindonline	Educating adults and children about Internet use and safety	mijnkindonline.nl
USA	NoBullying	Bringing innovative, sustainable solutions to bullying and harassment in schools.	nobully.co
United Kingdom	The Cybersmile Foundation	A cyberbullying charity committed to tackling all forms of online bullying and hate campaigns.	cybersmile.org

There are studies regarding the role of different parties (victim, bully and bystanders) involved in cyberbullying and they show that cyberbullying is a social problem and needs to be solved in a social context (Campbell, 2005). Therefore, it is not sufficient to deal with this problem individually and to concentrate on a single online activity. But we should consider

users' overall behaviour as someone's who is part of a society. Bystanders for example, play an important role in bullying incidents. Therefore, to make use of their position it is necessary to create empathy in youngsters, so that the bystanders speak out against bullies (Noble, 2003, Holfeld, 2014).

2.4.2 Technical Solutions

On the other hand, we should not overlook the significant impact of technical solutions in overcoming the problem of cyberbullying. With the increase of the number of reports on troubling consequences of bullying on youth, the number of studies and other materials dedicated to cyberbullying in online environments has increased (Figure 2.2).

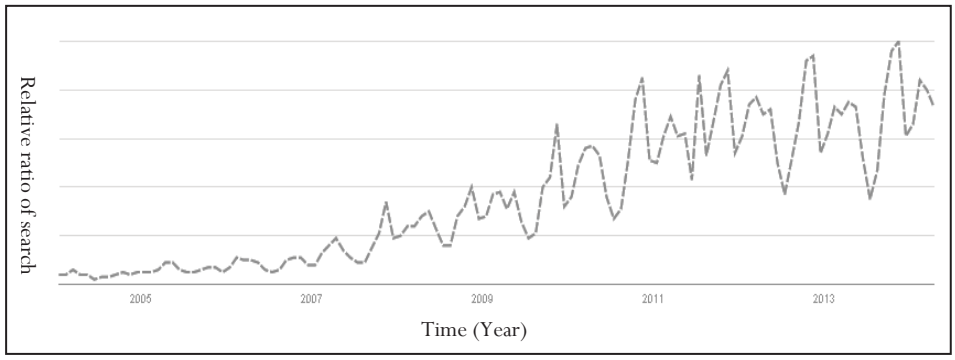

Figure 2.2 Increasing number of online reports, studies and other materials on cyberbullying since 2004. The graph reflects the ratio of searches that have been done for the topic of Cyberbullying, relative to the total number of searches done on Google over time. It does not represent absolute search volume numbers, because the data is normalized and presented on a scale from 0-100. Each point on the graph is divided by the highest point, or 100. Source: Google trends.

In recent years several studies have been dedicated to developing tools and solutions to deal with cyberbullying. An example is the AMiCA[1] project with the purpose of identifying possibly threatening situations on social networks. One of the critical situations investigated in this project is cyberbullying. Another project which focuses on relevant factors in governing social behaviour in online environments and looks into different kinds of interventions, technological as well as social and legal, is "Empowering and protecting children and adolescents against cyberbullying"[2]. The objective of this project is to recognize the possibilities for protection of individuals against online misbehaviours through different kinds of regulatory modalities.

For pre-cyberbullying tools, there are a wide range of software designed for parents and adults to control the online activities of children, for example Norton Online Family[3], Windows Live Family Safety[4], AVG Family Safety[5] and more. These software packages are sensitive towards certain words in the content of emails, messages or links sent or received by the children. When such words appear, the software either automatically blocks the content or alerts the parents. This type of monitoring software is considered to be preventive since the systems work based on the assumption that users will change their behaviour if they know their activities are being watched. However, a recent study found that the user monitoring software does not correlate with less cyberbullying victimization (Mesch, 2009).

There are also studies conducted into solutions to post-cyberbullying stage. Most of the studies up until now, have looked into automatic detection of

[1] http://www.amicaproject.be/ [Accessed June 2014]
[2] http://www.nwo.nl/onderzoek-en-resultaten/onderzoeksprojecten/18/2300154018.html [Accessed June 2014]
[3] https://onlinefamily.norton.com/familysafety/basicpremium.fs [Accessed August 2013]
[4] http://windows.microsoft.com/en-gb/windows-live/essentials-other#essentials=overviewother [Accessed August 2013]
[5] https://eshop.avg.com/us-en/cart?ECID=af%3Acj%3Atl%3Aus-ish [Accessed August 2013]

cyberbullying incidents which have happened through harassing comments and posts in social networks (Dinakar et al., 2011, Yin et al., 2009). There are also tools which have tried to flag the users which have posted the hurtful messages as well the messages themselves (Chen et al., 2012). Moreover, recently programs have been designed to help the victims or potential victims, after they have been cyberbullied. The victims can communicate with these programs through the designed interfaces (van der Zwaan et al., 2010, Jacobs et al., 2014), or the programs provide an intelligent agent that engage youngsters by using different emotional strategies, including emotional support by expressing empathy and encouraging them to take active steps to improve the troubling situation (Heylen, 2009, Adam, 2009). In most of the social networks, such as Facebook, the victims can also report the harassing and hurtful messages and ask the administrators to remove the content or block the offender. However, these types of interventions and supports need the user's initiative and the victim or bystanders should be aware of such support systems and know how they function. They also need to have the courage and strength of using them. Besides automatic detection and monitoring systems, experts in the field of cyberbullying highly recommend follow-up strategies that should focus on preventing future cyberbullying and empowering the parties involved (Van Royen et al., 2014).

2.5 The Gap

Although many studies and researches are dedicated to tackling the cyberbullying problem, there are still shortcomings in this area which need to be addressed in order to reach the ultimate goal, which is to wipe cyberbullying out for good, or more realistically, to minimize its sad and negative consequences.

A weakness which we have observed in the studies conducted on cyberbullying from the social perspective as well as the technical perspective, is that both of these fields of research have neglected the benefit of integration of the other field's findings in their studies. The social studies have purely dived into psychological, behavioural and personal reasons and causes of online misbehaviours and consequently their proposed solutions fail to incorporate the technical attributes and feasibilities of Internet and social networks. Computational facilities such as automatic detection of bullying incidents, identifying potential bullies by automatic screening of user profiles and other alerting functions, can enhance the implementation and the achievement of the proposed behavioural solutions, such as supervision and monitoring. On the other hand, the majority of approaches that are based on technical functionalities have overlooked the subtle yet important points which are highlighted in the social studies. For instance, the technical solutions are mainly generic and work the same for everyone, irrespective of the personal characteristics of individuals and the differences in the way that people bully in different social groups. Another shortcoming of most technical studies on cyberbullying is that they have mainly concentrated on detection of bullying incidents after they happened, while there is no attention for the possibility of tools contributing to preventing the bullying incidents and stopping the potential bullies from harming others.

Cyberbullying is a dynamic multidimensional problem which should be tackled from different aspects. The problem is deep rooted in the complexity of the human mind and it transforms in parallel with technological innovations that can be put to use for yet another type of bullying behaviour. Therefore, it might be unrealistic to aim for the day that the thought of bullying others does not cross someone's mind, but we can think of solutions that restrict the power of those with ill intentions to act upon their thoughts, provide Internet users with tools to protect themselves and make social networks a safer place for teenagers to mingle. Cyberbullying is a social dilemma that raises debates regarding potentially

valuable learning experiences gained through cyberbullying incidents (Shariff, 2008). Moreover, there are concerns regarding invading privacy of adolescents and limiting their right to freedom of expression. However, these concerns should be balanced against the benefits of Internet safety. In some cases protecting children's mental health might be of higher importance than protecting their privacy and freedom of expression. Cyberbullying is a major concern of this era and when the creation of appropriate instruments is feasible, online communities should be equipped with sound and effective tools that may enable the society at large to leave this passage behind.

2.6 Proposed Solutions

The overview presented in this chapter gave an insight into what needs to be explored and studied to fulfil some of the shortcomings in regard to cyberbullying. It is the aim of our research to contribute to the safety of youngsters by paying attention to their individual personalities and characteristics while trying to detect, and even further, to prevent bullying incidents. Meanwhile, we had this hypothesis that we should look into the cyberspace as a real society in transition which is dealing with ever transforming social concerns and misbehaviours, and we concluded that, cyberbullying, as one of these online misbehaviours, should be approached from different point of views and perspectives, just as is common for many other societal problems. Therefore we have not restricted our computational approaches to the detection of a single incident of misbehaviour and took a broader angle for combating the problem by also looking into the behavioural trend and history of activity of users and by incorporating their personality and other personal characteristics. In more general terms, we propose a multidisciplinary approach in which the findings and knowledge of the social and behavioural studies are integrated with technical algorithms and techniques.

We specially focus on two task domains. The first is automatic detection of cyberbullying in social networks using demographic information of the user population to be applied in the post-bullying phase represented in Table 2.2. In our approach we differentiate the way in which different age and gender groups bully other users in their network and these differences are taken into account to improve the detection accuracy of bullying content. This will be the topic of Chapter 4. For the second task domain we make use of human knowledge to identify Internet users with the potential of being a bully user. As explained in Chapter 5, we incorporate experts' knowledge from different related areas of research to identify online behaviours that can lead to harmful misbehaviour in future. It will be explained how each Internet user can be evaluated and assigned a score. This score represents whether a user can potentially be a threat to the cybersociety and can be applied in a way that is comparable to how all kinds of indicators are used in the context of background checks in real life. The proposed approach provides a novel element in the spectrum of preventive measures in the battle against cyberbullying.

Chapter 3

Datasets

A major gap that we encountered in designing experiments on the cyberbullying problem was the absence of a publicly available dataset reflecting the nature of bullying on the Internet. There were several requirements for a dataset that could be used throughout out project. It had to contain real cases of bullying posts, authored by network users from different age and gender groups as well as from different backgrounds. Moreover the dataset had to encompass a variety of online activities that users could conduct in a social network. These activities would represent their interests. In this chapter we introduce the dataset that we developed to include the attributes required for our experiments. We selected YouTube as the source platform for our dataset. This chapter addresses Objective 2 of our research: to create a comprehensive dataset to be used in cyberbullying studies.

3.1 Introduction

An important limitation in studies on cyberbullying is the lack of appropriate datasets. At the start of our research there was no standard labelled dataset available fully encompassing all the attributes required for the development and testing of tools for detection of cyberbullying incidents and actors in social media platforms. Therefore an important part of our project was the design and creation of a suitable corpus.

Depending on the purpose of the study, there are certain properties that a corpus should meet. Representativeness, balance and availability are examples of properties considered to be essential in the design of a dataset (Nguyen et al., 2012, Xiao, 2010, McEnery, 2001, Kruskal and Mosteller, 1979).

A dataset to be used in our experiments on cyberbullying not only should consist of textual comments or discussion logs, but should also cover the following properties:

- Representativeness: the content of the dataset should contain material which mirrors the subject of the study. In the case of cyberbullying studies, the dataset should represent the communication behaviour within an online community and contain bullying incidents and profile information of the actors involved. The dataset represent misbehaviours as they happen in online environments and should reflect the nature of bullying on the Internet.

- Availability: this parameter is an important concern especially for cyberbullying studies. Most of the bullying incidents happen in private online environments such as chat logs that are not accessible to the public. Moreover, for ethical and/or legal reasons many social networks do not allow their contents to be used by others. For example Facebook can be an excellent platform for designing a dataset, but the content is not publicly available.

- Heterogeneity of users: to reflect the fact that the users of social networks are a heterogeneous group, the dataset should contain users from different age and gender groups as well as from different backgrounds and with a variety of interests. Moreover the texts should be written in a variety of writing styles, as style is representing the personality and intentions of a writer.
- Balance: in social networks the ratio of bullying comments is usually lower in comparison to non-bullying comments. The dataset should therefore contain enough bullying comments in order to represent the variation of bullying styles. One of the factors determining the minimum required number of comments in order to have a balanced dataset is the intended methodology to be used to analyse the dataset (Xiao, 2010).

There are several reasons that could explain the lack of suitable dataset in cyberbullying studies. One of the reasons might be the privacy issue in online environments. Most of the services provided by social networks through which bullying happens, such as messages or chat logs, are not publicly accessible. Therefore, it is challenging to select an appropriate platform as a source for the development of a dataset. The privacy constraints impose additional restrictions on the suitability of the already limited number of existing datasets, and developers are not allowed to share them with other researchers. Apart from availability, an important concern is the creation of annotation layers. Depending on the purpose of the study, the given instructions and definition for labelling the dataset may differ. Therefore, we should make sure that the dataset is labelled according to the definitions proposed in the experiment in which the dataset is being used.

For our studies a dataset should contain a balanced number of bullying and non-bullying comments. Moreover, the dataset has to include information on the personal background of the users, such as age and gender, as well as the frequency of their online activities, such as commenting and

subscription. This type of information was not available. It was therefore decided to develop a dataset geared towards the attributes required for our experiments.

We selected YouTube as the source platform for developing our dataset. YouTube is a video sharing website which offers its users a variety of online activities. YouTube users can also have their personal profile which includes personal information and interests. In order to have a better understanding of the requirements for a dataset and also to have a baseline for our initial experiments, we decided to use an existing dataset before creating the YouTube dataset. For this purpose we selected a dataset containing MySpace discussion logs. However, as it will be explained in following sections, the MySpace dataset did not include all the required attributes for a study on cyberbullying study and was a confirmation of the necessity of developing the new YouTube dataset.

The part which is addressed in this chapter is highlighted in Figure 3.1 which is based on the flow diagram introduced in Chapter 1 (Figure 1.1).

The next section explains the MySpace datasets. In this section the attributes and characteristics of the dataset are fully described. The YouTube dataset is introduced in Section 3.3 which extensively explains the collecting and sampling process of the dataset as well as the annotation layer and attributes of the dataset. Conclusion and recommendations wrap up this chapter.

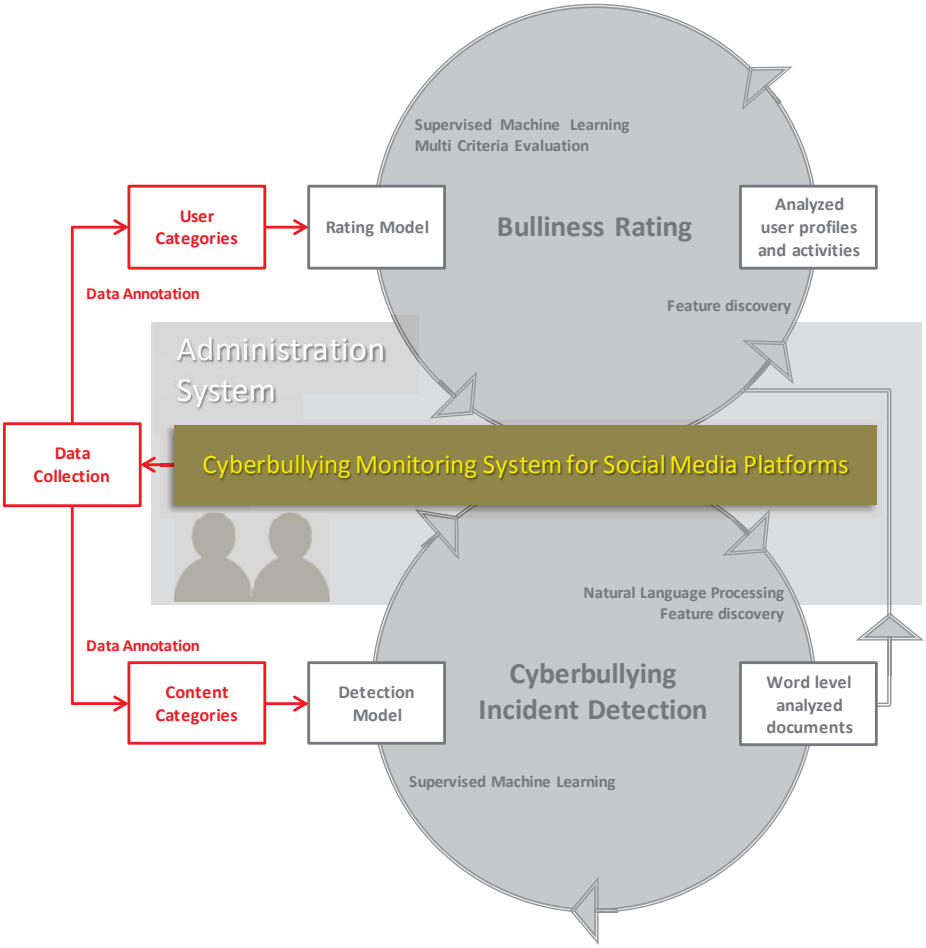

Figure 3.1 The highlighted parts of the flow diagram are addressed in this chapter: data collection process and preparing the dataset for model training**.**

3.2 MySpace

Foundation Barcelona Media provided five datasets for analysis in the Content Analysis for the Web 2.0 (CAW 2.0) workshop, with the general purpose of social media studies, on 2009[1] where one of the objectives of the workshop was misbehaviour detection and addressing the problems of detecting inappropriate activity in a virtual community. One of the dataset was MySpace. MySpace is a social networking site which offers its users the opportunity to participate in forum discussions about predefined topics. We chose MySpace as the dataset to use in our experiment because besides harassing comments, it also included some personal information of the users such as their gender. MySpace is considered a discussion-style community. In discussion-style communities, various discussion topics (threads) are offered and there are multiple posts for each topic. Users are free to start a new discussion or participate in an existing one by adding posts to it (Yin et al., 2009).

3.2.1 Attributes and Factual Statistics

The MySpace dataset consists of about 380,000 posts to 16346 threads dealing either with one of the following topics: campus life, news and politics, and movies. For more information about the sampling process and technical specifications see CAW 2.0 website. The discussion within a thread is related to a predefined topic. Information available for each post includes the user id of the author of the post, the content of the post, and the time of publication. In this dataset gender information of the authors is also available.

[1] Foundation Barcelona Media (FBM). Caw 2.0 training datasets. http://caw2.barcelonamedia.org/?page_id=98 [Accessed April 2012]

3.2.2 Annotation

The MySpace dataset provided by the workshop was not labelled, which required us to create annotations in order to be able to train our models. To develop a training dataset we randomly selected 1587 threads which in total included 9018 posts. We asked three students to manually label the posts as bullying or non-bullying. A post was judged as bullying or non-bullying based on the content of the post and the definition introduced in Chapter 2: an aggressive, intentional act carried out by an individual, through textual content, against a victim. A total of 311(3.3%) of the posts were labelled as bullying, which means 175 (11%) of the threads contained bullying posts. The proportions of bullying and non-bullying posts and threads are shown in Figure 3.2.

In this dataset the ratio of male users is higher than female users. Overall, 42% (n=3792) of the posts are written by female and 58% (n=5226) by male authors. The proportions of posts written by male and female users are illustrated in Figure 3.3. From the total of 311 bullying posts, 63 (20%) of them was written by females and the rest was written by males. Moreover, as illustrated on Figure 3.4, in the threads which contain bullying comments, the ratio of male users (83%) to female users (17%) is much higher compared to non-bullying threads.

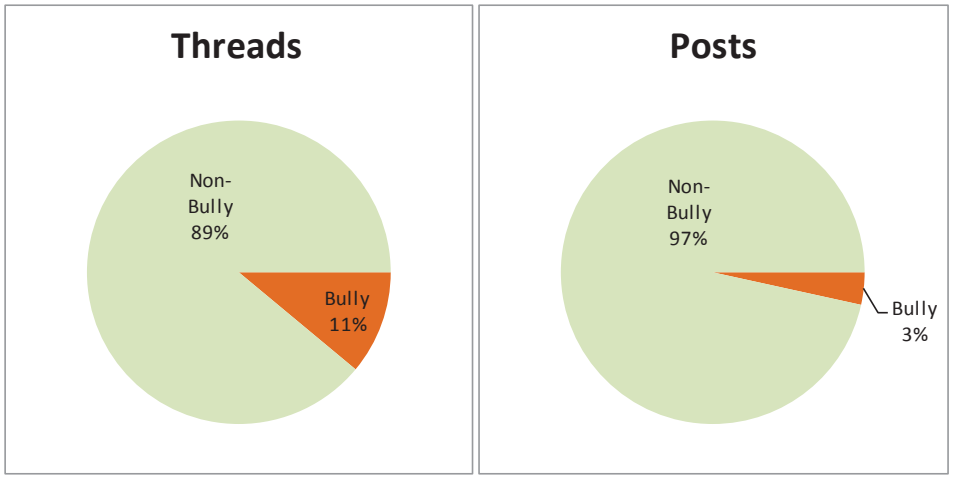

Figure 3.2 Left: the ratio of threads containing bullying posts to threads without bullying posts. Right: The ration of bullying posts to non-bullying posts.

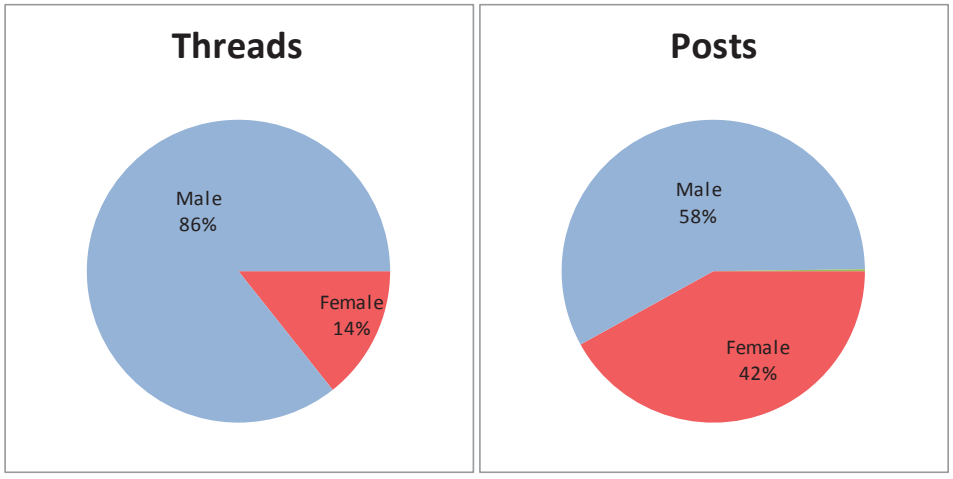

Figure 3.3 The ratio of posts written by male and female users

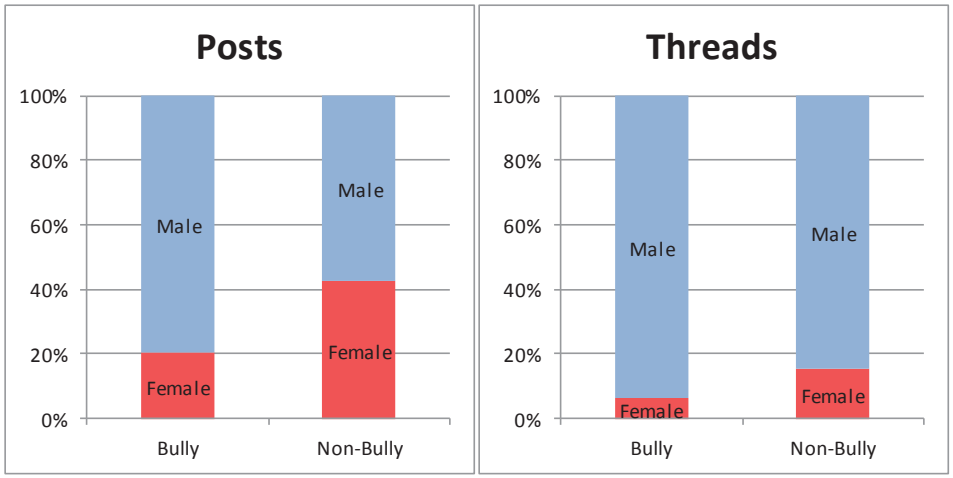

Figure 3.4 Left: The ratio of bully and non-bully posts written by male and female users. Right: The ration male and female users in bullying and non-bullying threads

3.2.3 Inter-annotator Agreement

All comments were labelled by the three students as explained in the previous section. The inter-annotator agreement was 95%, but to have a more robust agreement calculation, we also calculated the Cohen's kappa coefficient.

Cohen's kappa measures the agreement between two annotators who each classify *N* items into *C* mutually exclusive categories. The kappa statistic was first mentioned by Galton (1892) (Galton, 1892).

The equation for kappa is:

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)},$$

where Pr(*a*) is the relative observed agreement among raters, and Pr(*e*) is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly saying each category. If the raters are in complete agreement then $\kappa = 1$. If there is no agreement among the raters other than what would be expected by chance (as defined by Pr(*e*)), $\kappa = 0$. In our study the kappa = 0.79 which is a satisfactory agreement among annotators (Fleiss et al., 2013).

The MySpace dataset satisfied the requirements of our preliminary experiments specially because it contains demographic information for the users, but it could not be employed in the extended experiments. This was mainly due to the fact that it failed to encompass more informative attributes, such as the history of online activities of the users as well as the low ratio of bullying comments. To overcome these limitations we decided to develop a dataset which resolves those shortcomings to a great extent. The next section explains the collection of this dataset.

3.3 YouTube

YouTube is the world's largest user-generated content video system (Cha et al., 2007); 60 hours of video are uploaded every minute, and over 4 billion videos are viewed every day (YouTube press statistics, May 2013). It is localized in 39 countries and YouTube is considered to be a sample of the general internet population in terms of the audience demographics (Cha et al., 2007). There is a broad audience from different age and gender groups. This makes the network comparable to real life situations and therefore suitable as a source for investigating the interaction among users. YouTube users can carry out different actions within the network. For example they can post comments on videos and respond to other users' comments, subscribe to users' channels (personal page of the users that shows their activities), upload videos and, like and dislike other users' actions. As YouTube can be considered representative of the real world, it

also demonstrates some forms of social misbehaviours. Cyberbullying is one of the common and repetitious incidents that have been reported. The broad scope of audience, content of the videos and other users' comments trigger the bullies to disturb and victimize their targets through posting harassing comments in cyberspace. In follow are examples of harassing comments posted on YouTube videos:

"hunny just quit plz cuz i think you r the worst"

"God u suck dam stop singing u fucking ugly ass potato"

"even if you die no one would notice, i promise "

"Ofcourse a stupid fat girl like u can't do any better"

The effect of cumulative and long lasting comments intensifies the bullying acts. Despite the fact that the owners of YouTube videos have the possibility to remove offensive comments from the site, most of the comments are not moderated. Therefore, YouTube comments can be considered an appropriate datasets for cyberbullying studies. In order to satisfy the requirements of our experiments, the social network to be selected as the source of our dataset, had to offer its members a variety of online activities as well as their personal information. Moreover, the social network had to be a platform prone to bullying activities while representing the general online population. YouTube users can carry out different actions, for example they can post comments on videos and respond to other users' comments, subscribe to users' channels (personal page of the users that shows their activities), upload videos and, like and dislike other users' actions.

3.3.1 Sampling

For the purpose of our research, we selected videos using a selection of query terms. To increase the ratio of bullying incidents and variety of bully

users in the dataset, we searched YouTube for topics sensitive to cyberbullying, such as talent, physical appearance, and sexuality. Note that our annotated dataset may not be representative of the total YouTube user community, since we sampled the users from the videos with more bully-sensitive topics. We determined the users who commented on the top three YouTube videos for search terms such as "*cover song*" and "*funny fat people*". We used YouTube API to extract the publicly available information of the users. We removed the comments which had no author information or for which this information was not publicly available.

First, we collected 6000 comments posted by 3470 users along with their profile information. We also captured profile information of the users, such as their age and the date they signed up. We did not store private profiles or private information from the profiles such as the zip code. Second, we extended our dataset by also collecting a log of the users' history of activities for a period of 4 months from April to June 2012. The extended YouTube dataset consists of more than 54,000 comments.

3.3.2 Annotation

For the evaluation process of our experiments, two graduate students were employed to independently annotate the comments as bullying or non-bullying based on the definition of cyberbullying provided earlier. First the primary 6000 comments were labelled, and those that both annotators had labelled as bullying were marked as bullying comments. Disagreements were resolved by the decision of a third annotator (inter-annotator agreement = 93%). In total 17% (n= 1020) of the comments is labelled as a bullying comment.

To annotate the extended dataset, we compiled all the comments posted by every user, and created a dataset consisting of all the users name, history of their comments, as well as their profile information. Then we asked the same students to independently annotate the users as bully or non-bully

based on the same definition of cyberbullying. We assumed that a user is a bully, if there were at least one bullying comment in his/her history. Disagreements were also resolved by the decision of a third annotator (inter-annotator agreement = 91%). In total 12% of the users (n= 416) has at least one bullying comment in their history and were labelled as a bully user.

3.3.3 Attributes and Statistics

On average there are 15.43 comments per user (StDev = 10.7, Median = 14) and the average length of a comment is 12 words. Comparing attributes of bully users with non-bully users revealed that there is a significant difference ($P<0.01$) in number of comments written by users of each class. Bully users on average had 19 comments whereas non-bully users were less active with 15 comments on average. The density of bully users with 3 and more comments is almost stable, while this trend is decreasing for non-bully users. The average age of the users is 24 with 2.5 years of membership duration. While 38.2% of the users have uploaded fewer than 10 videos, 1.3% has uploaded more than 100. About one third of the users have no subscriptions while 56% has fewer than 20, and 1% more than 500 subscriptions. Figure 3.5 illustrates the density graph for membership duration, length of comments, number of comments and age for bully and non-bully users in our dataset.

3.4 Conclusion

The YouTube dataset developed for the purposes of our research includes a vast range of information and includes the properties introduced at the beginning of the chapter. The dataset is collected from a platform which its content is publicly available. It represents real cases of cyberbullying which

are conducted by variety of bully users. The heterogeneity of users is well reflected. YouTube has similar number of male and female users (YouTube press statistics, May 2013) and their age ranges from 13 to above 80. The interests and intentions of the users, the history of their activities as well as personal and demographic details and the content of the comments are well covered.

This dataset could be expanded by adding the social graph information of the network as well as including relational status among the users. This information could enable us to also take the relationships among the users into account. Incorporation of these features can be an interesting line of future work which is addressed in Chapter 6.

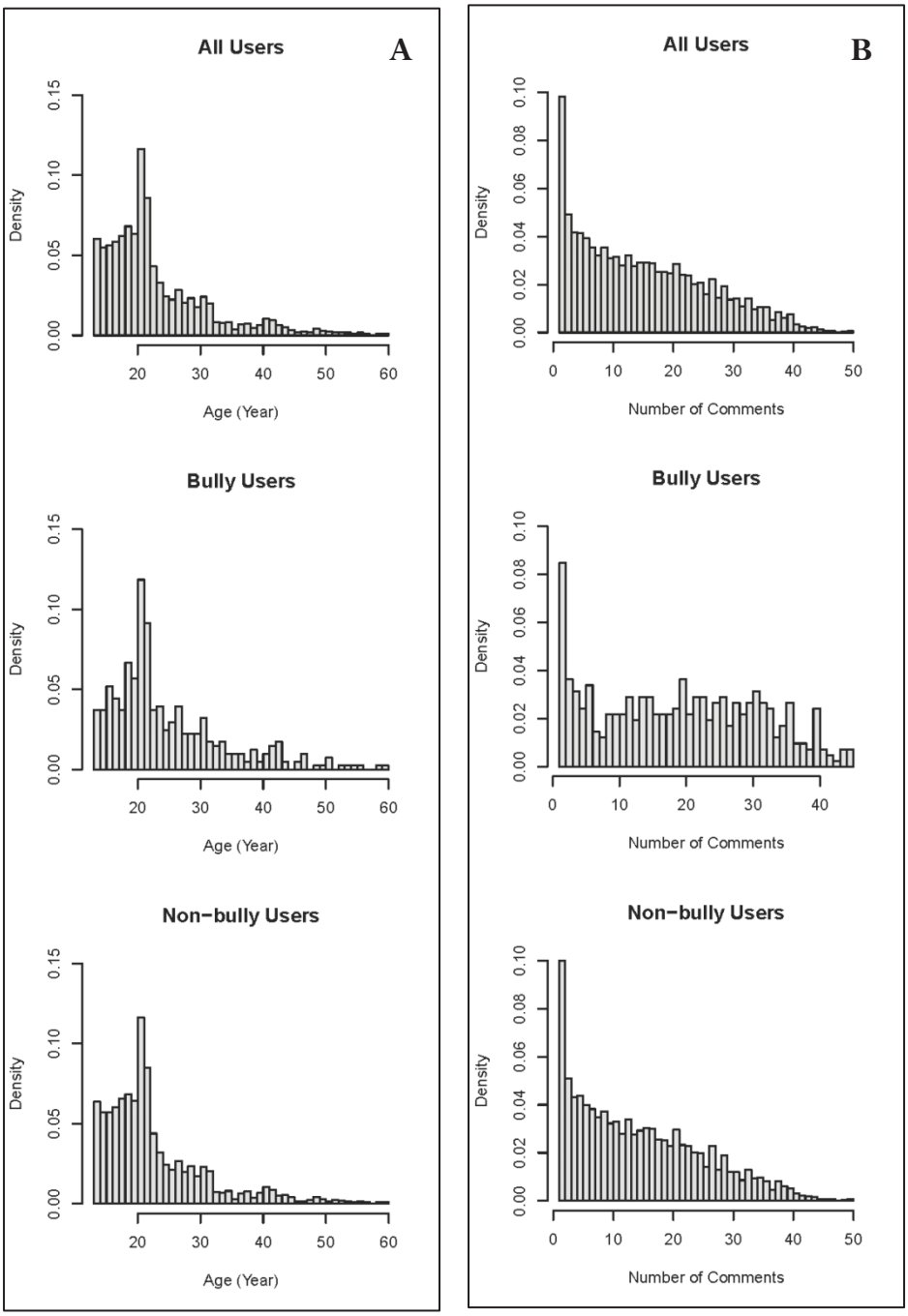

Figure 3.5 Age of the users (A), Number of comments (B), Membership (in years) (C) and Length of comments (in words) (D).

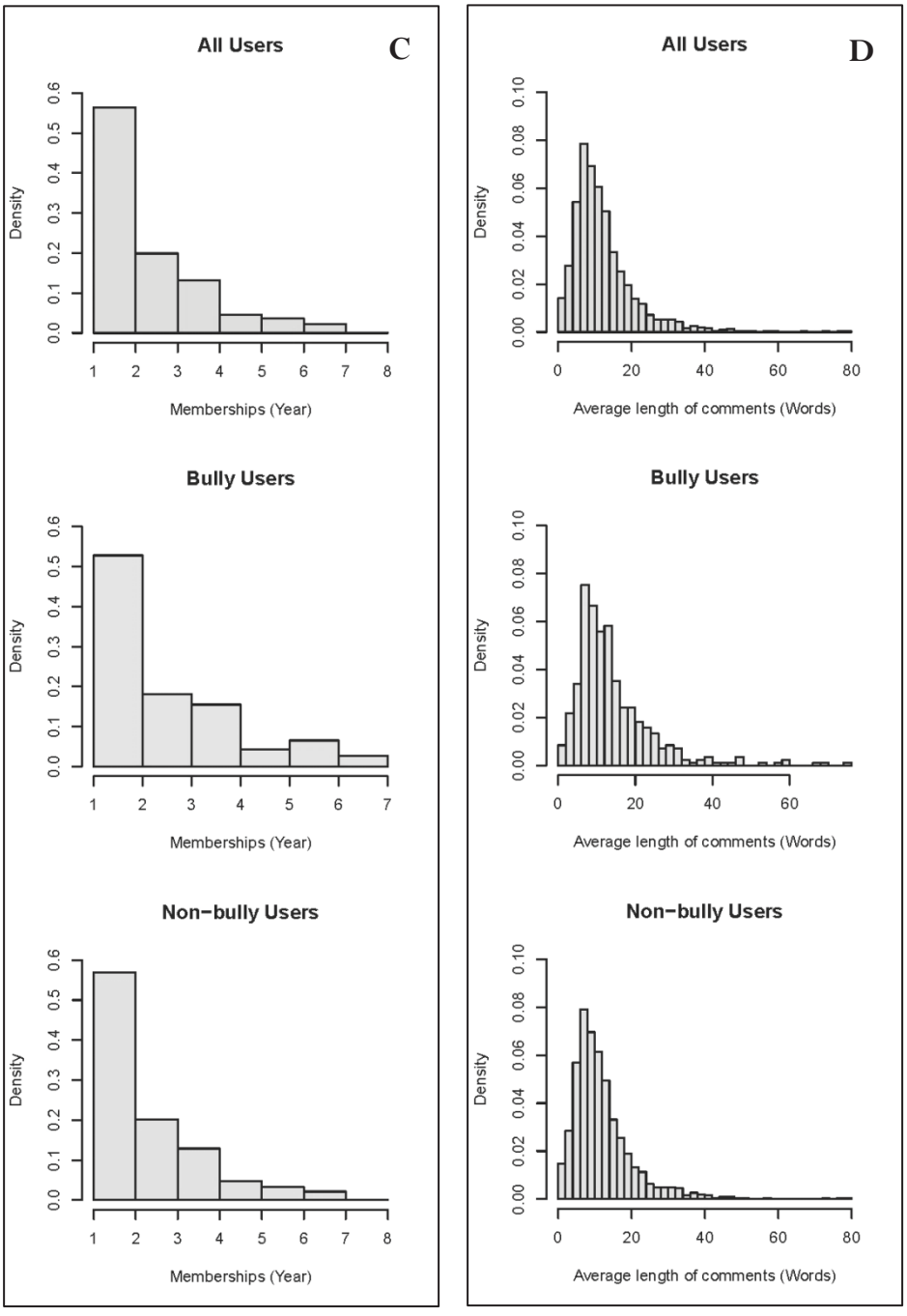

Figure 3.5 (continue)

Chapter 4

Cyberbullying Detection

Parts of this chapter also appear in:

- M. Dadvar, R.B. Trieschnigg, R.J.F. Ordelman and F.M.G. de Jong, *Improving cyberbullying detection with user context*. In Proceedings of the 35th European Conference on IR Research, ECIR 2013, Lecture Notes in Computer Science, volume 7814, Springer Verlag, Berlin, pp. 693-696, 2013
- M. Dadvar and F.M.G. de Jong, *Cyberbullying Detection; a Step toward a Safer Internet Yard*. In Proceedings of the 21st International World Wide Web Conference, WWW 2012 - PhD-Symposium, ACM, New York, pp. 121-125, 2012
- M. Dadvar, F.M.G. de Jong, R.J.F. Ordelman and R.B. Trieschnigg, *Improved cyberbullying detection using gender information*. In Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012), University of Ghent , Ghent, pp. 23-26, 2012
- M. Dadvar, R.J.F. Ordelman, F.M.G. de Jong and R.B. Trieschnigg, *Towards User Modelling in the Combat Against Cyberbullying*. In Proceedings of the17th International Conference on Applications of Natural Language to Information Systems, NLDB 2012, Lecture Notes in Computer Science , volume 7337, Springer Verlag, Berlin, pp. 277-283, 2012

Cyberbullying is a social problem which its solution partly relies on accurate and on-time recognition of the manifestations and the sources of this unwanted form of online behaviour. Detection of the bullying incidents, which according to the introduced framework in Chapter 2 falls in the post-bullying phase, may provide the instrument for taking the required actions addressing the consequences of the incidents. In this chapter we show that incorporation of personal characteristics of users in an automatic cyberbullying detection system improves the detection accuracy of the system. For the detection experiment we make use of the datasets explained in the previous chapter. This chapter focuses on Objective 3, to improve the accuracy of algorithms for the detection of bullying comments in social networks, and will address the following research questions: Does considering gender information of bullying users improve the accuracy of cyberbullying detection in social networks? Does considering further user profile information for bullying network users, such as age and history of comments, improve the accuracy of cyberbullying incident detection in social networks?

4.1 Introduction

As explained in Chapter 2, along with all the changes in communication methods and nature of relationships with the introduction of technology, friendships have also changed extensively. Previously, a friend was someone whom you had met at least a couple of times and with whom you shared many memories. Only close friends knew about our private life, family pictures and personal feelings. With the emergence of social networks, all this has changed. One may have hundreds of friends in an online social network, without ever having met them. Plenty of personal information is accessible for a wide range of people who might not be trustworthy.

These modifications and transformations of relationships and communication methods put the old social problem of bullying behaviour into a new format commonly referred to as cyberbullying.

As explained earlier, cyberbullying can have more thorough and longer lasting consequences due to its nature; hurtful material is available online for a long time and there is a broad audience that can witness it. Cyberbullying can happen through all sorts of technological devices and social media platforms, and at any time of the day. On top of the distress and sadness that is caused by bullying, the continuity of the assaults makes the impact even more unbearable. In order to inform responsible authorities or adults about bullying incidents and to allow them to stop the harassment and/or to provide required support for the victims, cyberbullying incidents have to be detected.

In cyberbullying detection the focus is on comments and posts which may contain vulgar and bullying content. In terms of the framework described in Chapter 2, cyberbullying detection falls into the post-bullying phase as it deals with incidents right after they have happened and after the harassing posts have been put online. The detection is to be considered a stepping stone towards an intervention; the aim is to take the necessary actions,

either removing the harassing content or provide the required support for the victim, after a bullying incident has been detected.

Most of the forums, especially those which are commonly used by younger teenagers, have safety centres (e.g. Facebook Family Safety Centre[1], YouTube Safety Center[2], or Twitter Safety and Security[3]) that support users and monitor the conversations and activities upon user's requests. An effective cyberbullying detection system in a social network can be used as a tool to support and facilitate the monitoring task of the online environments. However, the high volume of entries in these forums makes it impossible for moderators to keep an eye on everything that happens online. A system that gives warnings in case of an offensive post or activity would help the moderator to focus only on those cases and take the required actions in the quickest possible time (e.g. blocking the bully user's account). See Chapter 2 for more information.

In recent years various studies have looked into automatic detection of cyberbullying incidents (see section 3.2) however, there are still shortcomings in their proposed approaches. An important detail that most studies fail to contemplate is the personal differences that exist among individuals that act as bully. These studies have suggested generic algorithms for the detection of bullying posts by people with different characteristics from different age or gender categories, while the writing style and communication approach of individuals is greatly influenced by their personality and varies in different age and gender groups.

This chapter will describe our contribution to the improvements of the accuracy of automatic detection of cyberbullying in social networks. We will show that the accuracy of an automatic cyberbullying detection system can be improved by:

[1] https://www.facebook.com/safety [Accessed May 2014]
[2] http://www.youtube.com/yt/policyandsafety/safety.html [Accessed May 2014]
[3] https://support.twitter.com/groups/57-safety-security [Accessed May 2014]

- incorporation of gender information of the bullying actors,
- incorporation of the user's context, i.e. the user's history of comments and a wider set of personal characteristics of the users, such as age.

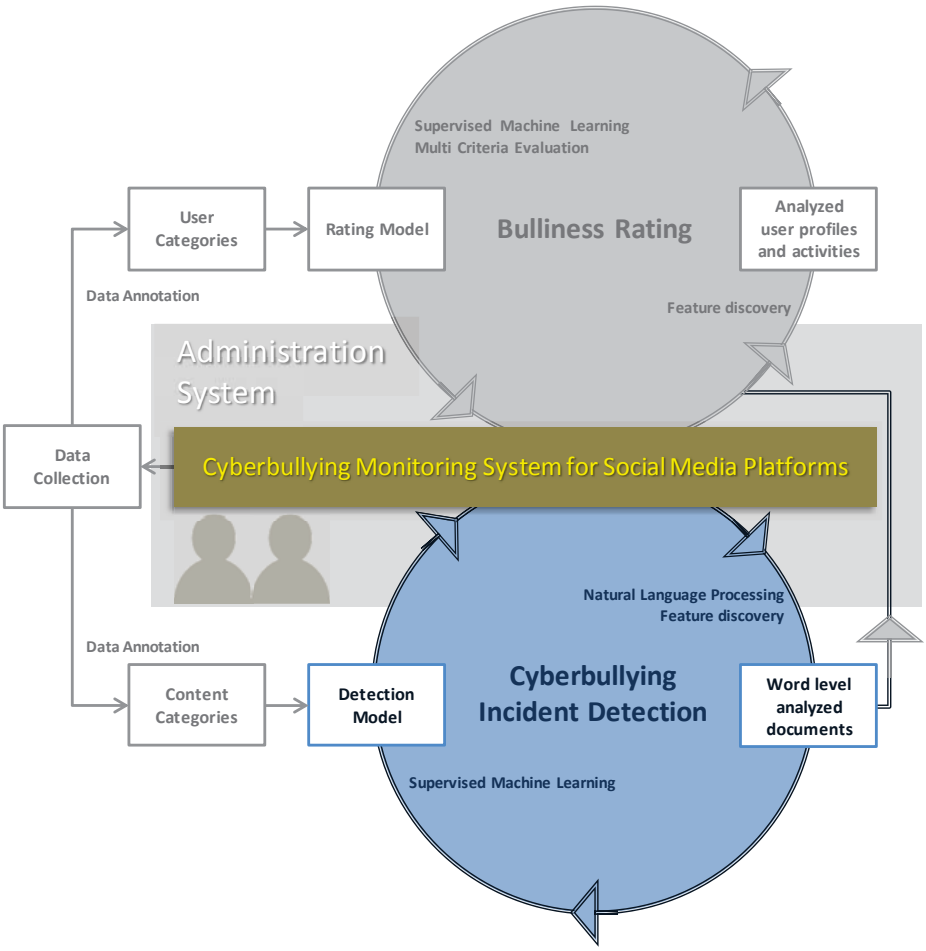

Figure 4.1 The highlighted parts of the flow diagram are addressed in this chapter: detection of bullying comments collected from the social network.

The part which is addressed in this chapter is highlighted in Figure 4.1 which is based on the flow diagram introduced in Chapter 1 (Figure 1.1). The work that we present is organized as follows: first in Section 4.2 we will provide an overview of previous studies conducted on cyberbullying detection. Our first innovation, incorporation of gender information, is introduced in Section 4.3. The datasets which have been used in this experiment as well as the experimental setup and results are also addressed. Our work is extended by making use of user context which a detailed explanation of the procedure and the outcome is given in Section 4.4. The chapter concludes in Section 4.5.

4.2 State-of-the-art in Cyberbullying Detection

There are several fields of research that are related to the detection of online crimes and misbehaviours. Most studies are based on text mining paradigms that deal with the detection of related troubling social behaviour shaped through online language, such as identifying online sexual predators and paedophiles (Kontostathis, 2009, McGhee et al., 2011, Pendar, 2007), detection of destructive article revisions, so-called vandalism detection (Smets et al., 2008, Potthast et al., 2008), spam detection (Tan et al., 2010, Sahami et al., 1998, Castillo et al., 2007), detection of internet abuse and cyber-terrorism (Simanjuntak and Ipung, 2010) and last but not least the studies conducted on offensive language use in social media (Chen et al., 2012).

The related studies provide some inspiration for cyberbullying detection as they all deal with forms of opinion mining tasks which make use of text analysis algorithms which can be also used in analysing bullying comments. However, their approaches are not fully suitable for the detection of bullying incidents. In most of the studies mentioned, the topics that are searched for can be distinguished through clear and predefined sets of

words, and specifying sets of rules to encompass the potential occurrences is more straightforward. For instance, the main difference between a spam message/email and an offensive one is that the former is sent identically to large numbers of people and is usually about a different topic than the topic of discussion. Spam messages are mostly commercial advertisements about a product or a service and therefore they are easier to be distinguished from the rest of the texts. On the other hand, offensive messages are personalized and are often a continuation of the previous conversation and the diversity expands up to the diversity of the human mind.

Research specifically focusing on technical solutions for cyberbullying detection has been scarce. One of the influential parameters on scarcity of technical studies for this topic might be the absence of appropriate datasets for developing and testing detection tools. As explained in Chapter 3, unlike for most of the other sentiment and text analysis fields, at the onset of the work on this thesis no dataset was available specifically developed for cyberbullying studies that cover all the relevant aspects. Not having a common dataset has also made it difficult to compare the final findings of the different studies. With the increase of the number of reports on troubling consequences of bullying on youths, several studies have been dedicated to detection of cyberbullying in online environments.

Yin and colleagues (Yin et al., 2009) used a supervised learning approach for detecting harassment in social networks. They used local, textual, and contextual features of documents to train a support vector machine classifier for a corpus of online posts. As the local features authors used each distinct term as one feature and calculated a Term Frequency-Inverse Document Frequency (TF-IDF) (Baeza-Yates and Ribeiro-Neto, 1999) value for each feature. They used frequency of profanities, second person pronouns and all the other pronouns in the posts. The contextual features were used to look at the context of the posts. For this purpose each post was compared to its neighboring posts to check their similarities. The authors assumed that harassing posts would be different from their neighboring posts. In their study only the content of the posts was used to

determine whether a message was harassing or not, and the characteristics of the author of the posts were not considered. They have used the combination of the features and their results show improvements over their three baselines in which word N-grams (N=1, 2 and 3), foul language and TF-IDF weightings were used individually as a feature to train a classifier. In another study with the same dataset as Yin et al. (2009) an attempt was made to identify clusters containing cyberbullying using a rule based on a dictionary of key words (Bayzick et al., 2011). The overall accuracy was 58.63%. The authors suggest that their rules should be refined to decrease the number of false alarms. In a more recent study on cyberbullying detection (Dinakar et al., 2011), a range of binary classifiers was used to classify an instance into a bullying sensitive or non-sensitive topic. Sensitive topics are usually related to race, culture, sexuality and intelligence. These topics are sensitive as they pertain to aspects that people cannot change about themselves. Moreover, the authors used multiclass classifiers to classify bullying sensitive topics. The classifiers were applied on a manually labelled corpus of YouTube comments. The authors treated each comment on its own and did not consider other aspects of the problem such as the pragmatics of dialogue and conversation. The findings showed that binary classifiers for the detection of textual cyberbullying can outperform multiclass classifiers. In another interesting study by Lieberman and colleagues (2011), in addition to machine learning classifiers, the authors used a commonsense knowledgebase with associated reasoning techniques. They collected about one million sentences describing everyday life that provide the kind of background knowledge that Artificial Intelligence programs need to simulate the informal reasoning that people do, rather than reasoning with mathematical precision. Moreover, in this study the authors had designed an interface that in case of detecting a potential bullying message, a message would appear on the screen encouraging the user to carefully reconsider their behavior and choice (Lieberman et al., 2011). Chen and colleges (2012) proposed the use of a lexical syntactic feature-based approach to detect the level of offensiveness in the comments

and potentially offensive users (Chen et al., 2012). They also considered the writing style of the users by checking offensive words used as nouns, verbs, adjectives, or adverbs for identification of the potential offensive users. We will return to the identification of the potential offensive users in Chapter 5.

Most of these studies are limited to the detection of bullying comment based on just the content of the comments. These studies fail to incorporate the personal characteristics of the users and they do not consider the context and the differences that exist in the manner of bullying across different age and gender groups. In this chapter we will explain that personal characteristics and context of use are important aspects in the data sources to be mined that can lead to improved incident detection accuracy.

4.3 The Impact of Gender Information on Detection Performance

We hypothesize that the incorporation of information on the gender of actors involved in a bullying incident, in the models for cyberbullying detection, alongside the content of their conversations, will improve the accuracy of cyberbullying detection. Social studies show that there are differences between males and females in the way they bully each other. Females tend to use relational styles of aggression, such as excluding someone from a group and ganging up against them and more implicit hostility, for example:

"No one would ever notice your absence.",

"You look too easy",

whereas males use more threatening expressions and explicit profane words (Chisholm, 2006):

"You fucking retard just die",

"Fags like you make me sick".

Argamon and colleagues (Argamon et al., 2003) found that females use more pronouns (e.g. "I", "you", "she") and males use more noun specifiers (e.g. "a", "the", "that"). These findings motivated our study of the effect of gender-specific language features on the detection of cyberbullying in social networks.

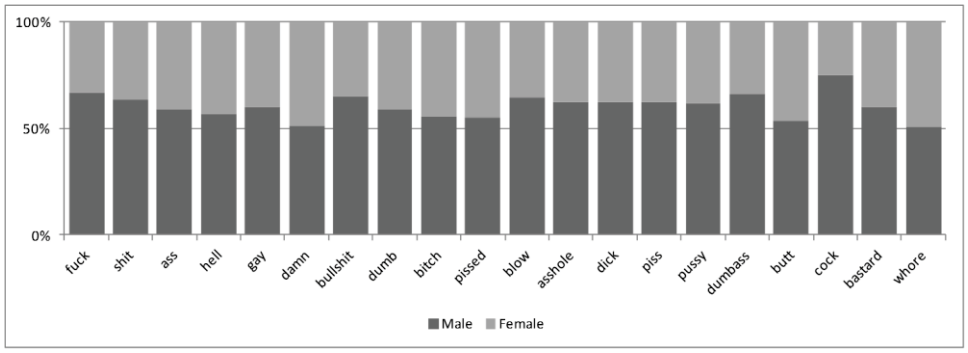

Figure 4.2 Top 20 frequently used foul words in the dataset. Each column represents the percentage of foul word used by female (light grey) and male (dark grey) users.

To illustrate the difference in offensive language use between genders, prior to start of our experiment, we first analysed the use of foul words in 100,000 randomly selected posts from the dataset. To do so, we compared the most frequently used foul words by each gender group using a Wilcoxon signed-rank test. A Wilcoxon signed-rank test is a non-parametric statistical hypothesis test used when comparing two related samples on a single sample to assess whether their population mean ranks

differ. As it was determined by the test, male and female authors used significantly ($p < 0.05$) different frequencies of foul words in their posts. The results are illustrated in Figure 4.2. An interesting point which can be inferred from the outcome is that female users tend to use less explicit profanities in compared with male users.

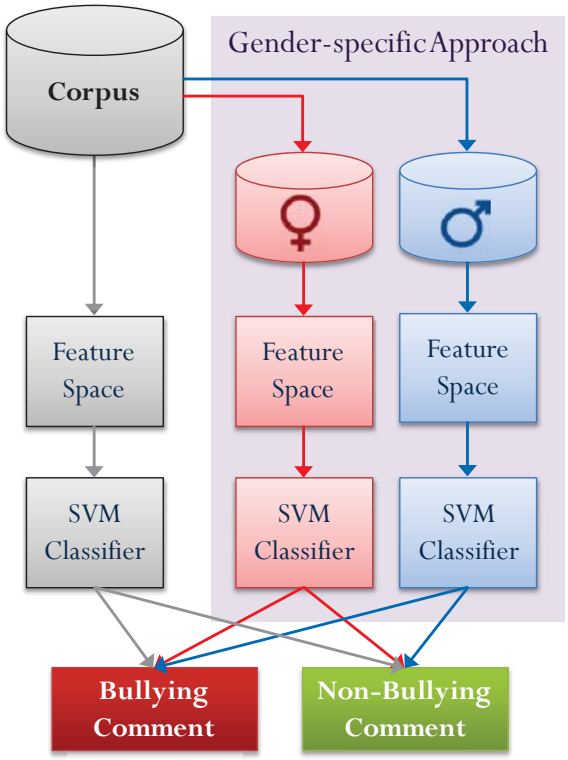

Figure 4.3 Conceptual framework of the proposed gender-specific approach versus conventional cyberbullying detection approach

4.3.1 Methods and Materials

In this experiment we employed a supervised learning approach to detect bullying comments in MySpace posts. Studies show that Support Vector Machines (SVM) classifier is one of most common classifier for text

classification tasks and it outperforms other methods over variety of different learning tasks (Joachims, 1998, Burges, 1998). Therefore we chose SVM (Vapnik, 1998, Cristianini and Shawe-Taylor, 2000) as the basis for the classification. We constructed an SVM classifier using WEKA (Hall et al., 2009) and we selected features that could represent the differences among the bullying style of different gender groups and indicate their writing structures. Figure 4.3 shows the conceptual model of the approach taken for detection of bullying incidents.

Dataset

In this part of our experiment we used MySpace posts as our dataset. As explained in Chapter 3, MySpace is a popular social networking site which offers its registered users the opportunity to participate in discussions about predefined topics. Our dataset contains 8,938 comments posted by 5,173 male and 3,765 female users. In total 311 (3.3%) of the comments were labelled as bullying. The proportion of bullying comments in male users (4.8%) was more than two times of the female users (1.7%).

Feature Space

To train the machine learning model for classification of bullying comments, we used four features that are commonly used for harassment detection (Yin et al., 2009, Dinakar et al., 2011). All these features are extracted from the content of the posts.

- *Profane words*[1], including their abbreviations and acronyms. For this feature, all the profane words of each post were treated as a single term and the ratio of profane words in the post was calculated. This

[1] Obtained from http://www.noswearing.com/dictionary

number was then normalized by dividing it by the post length. Since foul language appears sparsely in the dataset, the grouping strategy can reinforce the effect of this feature. Profane words are the most explicit indicators of possible harassments.

- *Second person pronouns* (e.g. 'you', 'yours'). All the second person pronouns were treated as a single word in each post and their ratio was calculated. Personal pronouns are frequently used in harassing posts. In sentences with structures such as {"second person pronoun" + "profanity"} or {"second person pronouns" + "verb" + "profanity"}, the pronoun can be another pointer to harassment occurrences. Although this feature is similar to the next feature, we separated it because we believe that second person pronouns are more important than other personal pronouns for detecting hostile sentences as they might indicate addressing someone specific.
- *All the other pronouns* (e.g. 'I', 'her' and 'him'). All the personal pronouns other than second person pronouns were treated as a single word in each post and their ratio was calculated.
- *TFIDF value of the words in each post*. We used each distinct term as one feature and calculated a TFIDF value for each of them.

Evaluation Metrics

Precision and recall are the basic measures commonly used in evaluating search and classification strategies. The two measures are sometimes used together, known as F_1-measure, to provide a single measurement for a system. Precision is the fraction of retrieved documents that are truly bullying, while recall is the fraction of bullying comments that are correctly identified. F_1-measure is the harmonic mean of precision and recall and is a measure that combines the two. We also calculated F_2-measure that weights recall twice as much as precision. We used *K*-fold cross-validation

to estimate how accurately our classifiers have performed. In *K*-fold cross-validation, the dataset is split into *K* partitions where the model is trained based on $K - 1$ partitions and the K^{th} partition functions as test set. This process is repeated *K* times. In our study $K = 10$.

4.3.2 Experimental Setup

We trained an SVM classifier using the features mentioned earlier to classify comments as bullying and non-bullying. We first trained our model on a dataset which was consisting of posts written by both male and female authors. The results of this run were used as the baseline. In the next step, we split the dataset into two groups based on the gender of the authors. The classifier was again trained on each dataset separately. The final results were calculated based on the proportion of each gender group in the whole corpus (34% female, and 66% male).To evaluate the classification accuracy we used 10-fold cross validation and calculated the corresponding evaluation metrics.

4.3.3 Results

Our first baseline (B1) based on TFIDF weighting of the words has resulted in precision and recall values of 0.32 and 0.30 respectively. In comparison to the previous studies Yin and colleges (2009) (B2), the recall value has improved 29% while the precision value has decreased by 9%. The difference in the outcomes can be due to the different training data which has been used in the studies. The results are illustrated in Table 4.1.

In the next step to train our classifier, we added the other three features (i.e. profanities, second person pronouns and other pronouns) to the TFIDF weighting used in the baseline. The features were extracted from the comments regardless of the gender of the authors. In comparison to the baseline (B1) there is a slight improvement in recall and precision; the

values have improved about 3%. The results indicate that the added features did not play a significant role in improvement of the outcome. In gender-specific approach the results are based on the classifiers trained on female-specific and male-specific models. As the results illustrate the recall has improved 12% in comparison to the non-gender-specific as well as the precision with 3% improvement. As hypothesized, incorporation of gender information of the users significantly improved the overall accuracy of the classifier.

Table 4.1 The accuracy measures of basic and gender-specific approaches for cyberbullying detection in the MySpace corpus

	TFIDF	*Foul Language*	*2nd Person Pron.*	*All other Pron.*	**Precision**	**Recall**	**F_1-meaure**	**F_2-meaure**
B 1	■				0.32	0.30	0.31	0.31
B 2 *	■				0.35	0.21	0.27	0.23
Non-gender-specific	■	■	■	■	0.33	0.31	0.32	0.32
Gender-specific	╳	╳	╳	╳	0.35	0.34	**0.35**	0.35
Female-specific	╲	╲	■	■	**0.41**	0.28	0.34	0.30
Male-specific	╱	╱	■	■	0.31	**0.38**	0.35	**0.37**

* (Yin et al., 2009)

Considering this kind of algorithm would mostly be applied in forums and social networks, and also considering the fact that it would affect the decision of the administrators, we believe that it is more important to first put the emphasize on developing an algorithm that results in a high recall.

The results show that for gender-specific models the F_2 measure improved to 0.35 compared to 0.32 for non-gender-specific model. To have a better insight on the male- and female-specific classifiers we also analysed them individually. The female-specific model has the highest precision in comparison to male-specific and gender-specific models, 24% and 17% respectively. On the other hand, the male-specific model resulted in the highest recall in comparison to female-specific and gender-specific models, 26% and 8% respectively.

4.3.4 Discussion

This experiment showed that information on the author of a post, such as gender, can be leveraged to improve the detection of misbehaviour in online social networks. This approach integrated the social studies' findings with technical algorithms, to emphasize the importance of considering personal characteristics and differences across individuals and resulted into a higher precision on detection of bully comments in female users and a higher recall for male users. One reason for this contrast can be the difference in usage of foul words by girls and boys and the way in which they bully each other. Girls tend to use less explicit profanities, and express more indirect negative and excluding attitude in their sentences while boys use explicit and vulgar language for bullying others. Another reason for the difference in performance of male- versus female-specific models can be the small size of the female-authored training dataset. In our training dataset there were only 64 bullying comments with female authors which is a small number for training the classifier. We expect better performance when we have a larger and more balanced dataset.

In the task of cyberbullying detection, it is essential to make sure that all the bullying comments are identified, meaning to have a high recall. It is also necessary to try not to wrongly flag the non-bully cases, which requires a high precision. Although it may need more time and effort from

an administrator's side to filter out the false alarms, it is better to make sure that as many as possible bullying comments are detected even if some non-bullying ones are also included. Once we have reached a certain level of recall, we can concentrate on improving the precision of the model as well. The improved results of this experiment motivated us to look into other personal characteristics of the users and into the incorporation of user context features for the further enhancement of our bullying detection model. The following section will explain the details of the next part of our experiment.

4.4 The Impact of User Context Features on Detection Performance

The studies conducted on cyberbullying are limited to content of a single comment posted by a user and they fail to incorporate other personal and contextual elements that can be informative for an automatic bullying detection model. Social networks store the history of activities of the users in their personal profile. This history also includes all the comments which have been posted over time by the user. Analysis of all the comments posted by a user can give a better understanding about the personality and nature of that person rather than judging only based on one comment. Moreover, the profiles also provide personal information about the users such as age, which can be an added value in identifying the intentions of the users in their writings.

In this experiment we show that incorporation of user context, such as users' comments histories, as well as characteristics such as age, into a detection model can improve the performance of the detection model.

4.4.1 Methods and Materials

In this part of our experiment we approach cyberbullying detection as a supervised classification task. We constructed an SVM (Vapnik, 1998, Cristianini and Shawe-Taylor, 2000) classifier using three incremental feature sets. These features represent the differences among the personality and writing style of the users.

Dataset

YouTube is the world's largest user-generated content site and its broad scope in terms of audience, videos, and users' comments make it a platform that is eligible for bullying and therefore an appropriate platform for collecting datasets for cyberbullying studies.

The attributes and characteristics of the YouTube dataset used in this experiment are thoroughly explained in Section 3.3.

Feature Space

The following three feature sets were used to train the classifier model. Part of these feature sets are similar to ones explained in section 4.3.1 plus some additional features. These features are extracted from the content of the comments posted by the users as well as user profile information.

- *Set 1: Content-based features.* These features are based on the contents of the comments itself. The following features are included: *(1)* The number of profane words in the comment, based on a dictionary [1] of profanities, normalized by the total number of words in the comment. The dictionary consists of 414 profane words including

[1] http://www.noswearing.com/dictionary [Accessed September 2012]

acronyms and abbreviation of the words. The majority of the words are adjectives and nouns. *(2)* To detect the comments which are personal and targeting a specific person, we included the normalized number of first and second person pronouns in the comment, based on a list of pronouns. *(3)* Profanity windows of different sizes (2 to 5 words) were chosen. These are Boolean features which indicate whether a second person pronoun is followed by a profane word within the size of the window.*(4)* To capture explicit emotions, the number of emoticons was counted and normalized by the number of words. And finally *(5)* to capture "shouting" and aggression in comments, the ratio of capital letters in a comment was computed.

- *Set 2: Cyberbullying features.* The second set of features is more specific and aims at identifying bullying topics such as minority races, religions and physical characteristics by making use of words that are commonly used to address these topics. It consists of: *(1)* the (normalized) number of cyberbullying words, based on a manually compiled dictionary. Cyberbullying words are the words that are commonly used by bullies against their victims to refer to sensitive topics earlier mentioned. Examples of these words are: "fat" and "negro". *(2)* The length of the comment can be another identifier of bullying occurrence, as it is observed that bullying comments are usually shorter than other comments (Yin et al., 2009).

- *Set 3: User-based features*. To be able to exploit information about the background of the users in the detection process, we looked at the *(1)* history of users' activities; the content-based features (Set 1) were extracted from the users' history of comments to see whether there is a pattern of offensive language use. As type of words and language structures may vary in different ages, we also considered the *(2)* age of the users as a feature.

4.4.2 Experimental Setup

We used the three incremental feature sets for training a Support Vector Machine to classify comments as bullying or non-bullying. As a baseline we only used content-based features (further referred to as Set 1). For Set 2 we included the cyberbullying features and for Set 3 also the user-based features were used. As a pre-processing step, stop-word removal and stemming were applied. We used 10-fold cross validation to evaluate the performance of our model with precision, recall and F-measure.

4.4.3 Results

As the results of our experiment, listed in Table 2 indicate, the detection performance improved when the bullying specific features were added to the model. Further improvements were observed when context features were also included in the feature set. For Set 1, a bag of profane words, pronoun-profanity windows, and second person pronouns' frequency were the main contributing features. Capital letters and emoticons however, did not add significant contributions. This could indicate that in the YouTube dataset, bullying comments do not necessarily contain more capital letters or emoticons in comparison to non-bullying comments.

Adding Set 2 features significantly improved both precision and recall (two sample t-test, $p < 0.01$). From Set 2 the length feature did not have any significant contribution, while the bag of profane words including bullying words contributed the most. Further analyses indicated that the most effective words for classification were vulgar words that refer to race and to sexuality.

As we hypothesized, the incorporation of users' profile information further improved (two sample t-test, $p < 0.05$) the precision and the recall to 77% and 55% respectively. As the classification was not just based on one comment and one instance of profanity use, the non-bullying cases were identified more accurately which lead to higher precision. The recall was

also improved which can be due to more accurate detection of implicit bullying comments by using the background of their authors. Implicit bullying comments are those which do not contain any explicit profane word and the harassment is indirectly and for example through sarcasm. The number of profanities in the history of each user had a major contribution, and the age feature had contributed but not as much as expected in the classification of bullying comments.

Table 4.2 Summary of the experiment results

	Content-based	Cyberbullying	User-based	Exclude number of profanities in user's history	Exclude number of profanities	Exclude pronoun-profanity window	**Precision**	**Recall**	**F_1-meaure**	**F_2-meaure**
Set 1	■						0.72	0.45	0.55	0.49
Set 2	■	■					0.75	0.51	0.60	0.54
Set 3	■	■	■				0.77	**0.55**	**0.64**	**0.58**
Set 3_1	■	■	■	■			0.76	0.52	0.62	0.55
Set 3_2	■	■	■		■		**0.78**	0.54	0.63	0.57
Set 3_3	■	■	■			■	0.76	**0.55**	0.63	**0.58**

4.4.4 Discussion

In this experiment we investigated the effect of content-based features and users' personal information on the improvement of cyberbullying detection. Our results showed that incorporation of context in the form of users' activity histories improves cyberbullying detection accuracy. The

feature sets were used incrementally for training the classifier model. The low recall of the first feature set can be explained by the occurrence of bullying comments without explicit profanities and by implicit bullying through sarcasm, or comments addressing sensitive topics using other words than profanities. However, adding cyberbullying features improved both precision and recall. Moreover, for the features in Set 3, the age feature did not contribute as it was expected. This might be due to the fact that many users do not indicate their real personal information.

This work could be extended to develop models that detect expressions involving sarcasm or implicit harassment. In future studies, further user features extracted from their network activities can also be taken into account. These types of features can provide further information about the interests and personal characteristics of the users.

4.5 Conclusion

In this chapter we demonstrated that the accuracy of an automatic cyberbullying incident detection system can be improved by incorporation of gender information of the users into the detection system, as well as by taking the user's context, such as history of online activities and personal characteristics, into account. The proposed approaches can also be used in other social networks. Our approaches are in principle language-independent and adaptable to other languages and the only required modification is the dictionaries to make sure that the profanities and words addressing bullying sensitive topics in the target language are included in the dictionary.

Profile information is not always stated in accordance to the 'actual' facts and figures for a user. For example, to meet the minimum legal age for signing up in the social networks or pretending to be younger or older, users may enter a false number as their age. The same goes for other

personal features such as gender. Therefore it might be beneficial to employ prediction routines, such as age prediction algorithms, prior to using profile information to improve the reliability of the classifiers.

The improvements resulting from the incorporation of personal features in our experiments are an indication of the importance of the information conveyed through personality and characteristics of the users in the prediction and detection of their intentions and behaviour. This outcome is a motivation to look into other sources of information. In the next chapter we will demonstrate how the adoption of a multidisciplinary approach can contribute to the battle against cyberbullying: integration of insights from the social sciences on what distinguishes a potential bully from the average social media user seems to be a useful approach in designing the outlines of a framework for preventing bullies to become effective. Adding a human touch into the technical workflow for the detection of online bullying incidents seems to pay off.

Chapter 5

Bulliness Score

Parts of this chapter also appear in:

- M. Dadvar, R.B. Trieschnigg and F.M.G. de Jong Expert knowledge for automatic detection of bullies in social networks. In 25th Benelux Conference on Artificial Intelligence, BNAIC 2013, TU Delft , Delft, pp. 57-64, 2013
- M. Dadvar, R.B. Trieschnigg and F.M.G. de Jong Experts and Machines Against Bullies: A Hybrid Approach to Detect Cyberbullies. In 27th Canadian Conference on Artificial Intelligence, University of Waterloo, Montréal, Canada, 2014

Automatic detection of bullying incidents in social networks is a crucial functionality for an integrated approach towards tackling cyberbullying and dealing with its harmful consequences. Accurate and quick detection of bullying comments can result in a timely reaction of the responsible parties for removal of the content or other responses. However, one step ahead efforts can be made to prevent bullying incidents. In this chapter we introduce an approach that is focussing on the recognition of user profiles that are likely to manifest themselves as bullies and that could be integrated in a monitoring platform. The study conducted into this type of preventive analysis is based on data for bully users who post bullying comments on YouTube. We measure their level of bulliness and assign a bulliness score which represents their likeliness of future misconducts. The bulliness score is based on the analysis of traces of intentions and personality features in online user activities, as well as on personal and demographic information available for users. For this purpose we use an expert system, three machine learning models and we also introduce a hybrid modelling approach which makes use of both expert knowledge and machine leaning models. This chapter addresses Objective 4, the design of a bulliness score for identifying potential bullies in social networks, and provides answers to the following research questions: How accurately can an expert system assign a bulliness score to a user to represent the level of bulliness of that user? Can an expert system and a system based on machine learning be effectively combined for detecting potential bullies?

5.1 Introduction

As described in Chapter 2, cyberbullying is a growing concern in online environments and the existing solutions are still inadequate for addressing this phenomenon which inherently is primarily a social problem. Most existing technical studies on cyberbullying, concentrate on the detection of bullying incidents in textual comments. These studies apply conventional sentiment analysis techniques trying to identify bullying incidents that happen in online environments (Dinakar et al., 2011). As explained in Chapter 2 and Chapter 4, for improving the bullying incident detection accuracy we used personal information of the users, such as age and gender. However, methods and procedures for the prevention of cyberbullying could benefit from gaining additional information about the sources of bullying incidents, namely the bullies in social networks. To enable networks to eliminate or constrain potential bullies, it would be crucial to have instruments to identify threatening users in social networks and to apply procedures to block their harmful activities before they can further hurt any more people. For a preventive approach personal and behavioural information about the users needs to be collected and aggregated in addition to the information elements which can be extracted by a computational framework as described in Chapter 4 from their online activities. It requires elements that are close to human nature and mind set as well as techniques which can reach and analyse more complicated sides of human mind, in line with the claim we made in Chapter 2 that bullying is an old phenomenon with roots in community life and human mind set in the pre-cyber era that is now occurring in cyber-communities as well and therefore its solution also needs the same human touch that has been part of the fight against bullying in non-virtual contexts. Examples of approaches that have proven effective in real-life situations can be inspirations for virtual environment approaches. Let's take an example from another domain: car driving safety. In many countries there is a "penalty point" system in car driving regulations which adds points to drivers who violate

driving regulations. If the number of points exceeds a certain limit, the driver licence will be revoked. Such a system prevents drivers from misconducts and stops the drivers who are a threat to society and themselves. We envisage a monitoring function based on a similar system to prevent potential bully users from further harming others. Our monitoring system will assign a score to each user which is based on the frequency of their online misbehaviours. This score represents the likelihood of the users to be a bully and conduct future misbehaviours. Therefore the score can be used to monitor or stop the high risk users. Moreover, similar to the penalty point system in driving regulations, being privately aware of the scores may warn and encourage the users to not get engaged in further violations and the system will act as a preventive tool.

Classification of user behaviour and/or user groups in online social networks have been used in various applications, such as improving advertisement recommendations in online social networks (Maia et al., 2008, Wilson et al., 2009). It has been shown that for user behaviour characterization, individual attributes of the users do not provide sufficient evidence and attributes which represent contextual parameters such as social interactions and the history of online user activities are also required for an adequate performance of user classification techniques (Maia et al., 2008). In this chapter we integrate similar contextual features to determine a user's bulliness score.

Given that cyberbullying is a multi- dimensional problem, several aspects need to be covered while investigating the problem. Cyberbullying can be influenced by online activities and environmental characteristics (for example, the degree to which a person is active in social networks) as well as personal features of the users. Furthermore, we believe that understanding the intention of a user who posts a comment can be a useful hint in identifying bullying cases. Understanding the underlying intention of a user in a comment is a challenging task, as intentions are implicit and hidden behind words. For instance, the sentence "*I know you are really trying!*" can be interpreted as an encouraging sentence while the intention

of the author is more of the opposite. Therefore, in the design of a preventive functionality we aimed to take all these elements into account. We have investigated two different designs implemented in two experiments. Intentions, mind set and feelings of the human beings can be best perceived and interpreted by humans. Therefore in our first experiment we decided to make use of an expert system that is based on human knowledge of the features that distinguish bullies from non-aggressive users. We then introduce a novel hybrid approach in which machine learning models and and an expert system are effectively combined. By the trust commonly put in human knowledge, this may bring a human touch to a problem that is rooted in human nature. The main contributions described in this chapter are:

- the introduction of the concept of a "bulliness score" representing the likeliness of a user being a bully,
- the novel application of an existing expert system framework for assigning the bulliness score,
- the novel application of existing machine learning models for assigning the bulliness score, and
- the development of a *hybrid* system which combines an expert system with various machine learning models for assigning the bulliness score.

The part which is addressed in this chapter is highlighted in Figure 5.1 which is based on the flow diagram introduced in Chapter 1 (Figure 1.1). In Section 5.2 we briefly present the dataset used in the experiments presented in this chapter. We discuss the experiment conducted in order to optimize the design of a preventive solution by making use of experts' knowledge in Section 5.3. This will be followed by a detailed overview of the expert knowledge elicitation and its application for detecting bully users. In the next experiment (section 5.4) we employed machine learning

methods to score potential bully users using contextual and personal features. To further develop this experiment we also proposed a hybrid approach as a solution to the bullying problem. In this section it will be explained how combining the machine learning models with expert knowledge can result in a powerful system for the rating of potential bully users. The conclusions in Section 5.5 will round off this chapter.

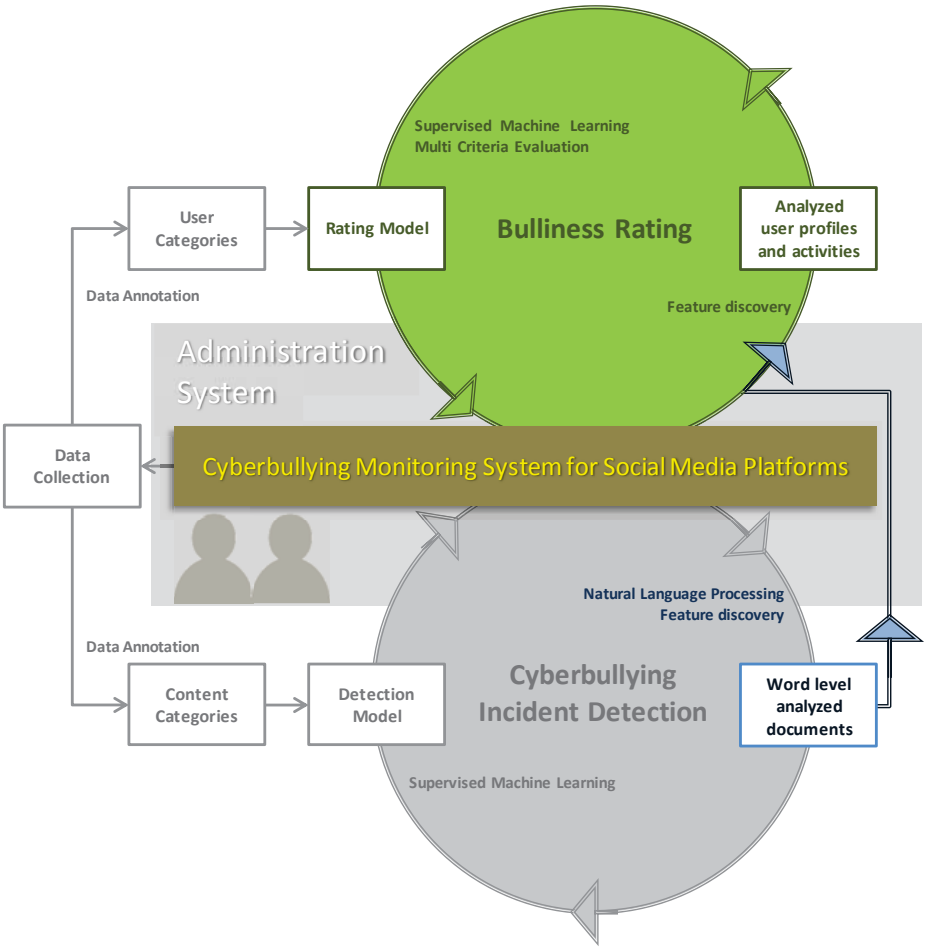

Figure 5.1 The highlighted parts of the flow diagram are addressed in this chapter; rating for bulliness of the social network users based on their online activities and personal information.

5.2 From Detection to Prevention; Motivation and Related Work

As mentioned earlier, most of the studies on technical approaches addressing the problem of cyberbullying have focused on developing solutions and algorithms to detect cyberbullying incidents that occur in online environments through vulgar posts and comments. Most of these studies develop machine learning algorithms which use features that are extracted from the content of the comments without incorporating personal information of the users (Dinakar et al., 2012, Reynolds et al., 2011). For instance, in a study on cyberbullying (Dinakar et al., 2012), authors have used a support vector machine to classify bullying comments found on YouTube. They trained the classifier using content features, such as frequency of profanities in comments and topics sensitive to bullying. In another experiment, Reynolds et al. (2011) have compared different machine learning methods, such as a decision tree and support vector machine, to select the best classification model for cyberbullying detection. Also in this study, the features used for training the machine learning models were content-driven. Cyberbullying detection should not only be based on content features but also on contextual details and information about the individual users. In our text mining approach of cyberbullying detection (Chapter 4), we also took personal information of the post authors into account, such as age and gender, which was shown to improve the accuracy of detecting bullying comments. The result of this work suggests that there is added value in personal information and characteristics of users for the modelling of cyberbullying and that it could be beneficial to investigate the incorporation of more personal features.

Even though there is a certain overlap in the tasks of detection of bullying posts and identification of potential bullies in social networks, they require a different approach. Detection of bullying comments as approached in this thesis is a text mining task while identification of users who bully is approached as a more complex task that requires information elements

which represent contextual characteristics and personality of the users. Few studies have focused on the detection of bully users. The use of the Lexical Syntactic Feature architecture to detect offensive content and identify potential offensive users has been introduced in (Chen et al., 2012). In particular, the authors incorporated users' writing style, structure and specific cyberbullying content as features to predict the users' potentiality to post offensive content. Pazienza and Tudorache (2011) proposed the incorporation of user profiling features to detect aggressive discussions (Pazienza and Tudorache, 2011). They used users' online histories of presence (duration of activity) and conversations to predict whether or not users' future posts will be offensive. Both of the mentioned studies point out interesting directions for the incorporation of user information in detecting offensive contents that could help to improve the detection rate. However, more refined user information with not only personal characteristics such as age, but also users' network activities (such as number of uploads in a social network) and details on communication patterns have not been included.

All the existing studies on cyberbullying prevention are based on plain machine learning methods. They are purely data-driven and they do not make use of human knowledge and reasoning. A solution for integrating expert's knowledge and human reasoning into the process of monitoring cyberbullying is the incorporation of a kind of expert system known as Multi-Criteria Evaluation System (MCES). A MCES provides the possibility to combine different sources of information, such as the knowledge from multiple experts, in order to make a decision among alternative options. MCESes are applied in a variety of evaluation and decision making research fields (Shee and Wang, 2008, Jiang and Eastman, 2000) and are used for decision-making in business, industry, and finance (Figueira et al., 2005). Expert systems have been used in several studies in cybersecurity, such as for the detection and prevention of cyber-attacks. Cyber-attack is the term referring to any kind of offensive activity that targets computer systems, networks or personal computers and it can

range from installing spyware to attempts to destroy the infrastructure of a country. Rule-based expert systems have been used for recognition of attack signatures. Signature-recognition techniques store signatures of known attacks, and match the new suspicious observed behaviors with the previous known signatures. The system will signal a cyber-attack when there is a match (Ye et al., 2001). Moreover expert systems have been used for the detection of network intrusions (Bauer and Koblentz, 1988, Denning, 1987, Bass, 2000). Most current information retrieval models determine document relevance by computing a single score which aggregates values of some attributes (Farah and Vanderpooten, 2006). However recent studies in information retrieval (Farah and Vanderpooten, 2006) and the measurement of document content reliability (Bong et al., 2012), show performance improvement when several sources of information are combined. To our knowledge this is the first time that a MCES is used for cyberbullying detection in social networks.

5.3 Expert Knowledge for Automatic Rating of Bully Users

To classify and rate users in social networks according to the likeliness that they will exhibit bullying behaviour, not only the content of posted comments need to be processed, but also more subtle features of the users and their personality can be of added value. There is ample information available, both implicitly and explicitly. Examples of explicit information are the details on age, gender. Examples of implicit information are the signals and traces in the language used or intangible things such as intentions and personality that can be reflected in a variety of ways, such as the choice of username on the network, or the number of comments and uploaded videos which can indicate the level of activity and popularity. These human characteristics can be sensed, yet it is not straightforward how to capture them in the form of features that could be fed to machine

learning models. Among the alternatives that are less focussed on feature-based modelling are the deductive frameworks that can offer an analysis that goes beyond activity patterns extracted from content features.

For the development of a system for the rating of potential bully users we hypothesized that for the capturing of implicit features the integration of expert knowledge could be effective and we have investigated the effectiveness of MCES as a candidate approach that integrates human reasoning and experts' opinions in a deductive way. With an MCES model the complex combination of personal and contextual factors that determine whether a user of social media will exhibit bullying behaviour can be captured based on expert knowledge. In the study described in this section, the resulting MCES classification is used to assign a score to YouTube users expressing the likeliness of future bullying behaviour. The performance of the classification approach was evaluated using manually annotated dataset.

5.3.1 Multi-Criteria Evaluation System (MCES)

As indicated in Section 5.3, MCES refers to a framework for setting and combining a variety of criteria derived from the features and their corresponding rules. In this study the rules are set by a group of experts in the area of cyberbullying with a background in psychology and social studies. We used an MCES to combine the criteria for rating potential bully users based on the knowledge of experts to calculate the level of bulliness of a user based on the user features. An MCES is not limited to the knowledge-driven criteria (i.e. the criteria that are generated using experts' knowledge), and can be also applied to combine criteria from other sources of information. In this section we only used knowledge-driven criteria to reach a final decision on the bulliness score of a user, however in the section 5.5 we also benefited from criteria derived from user features. The working of our proposed MCES is depicted in the flow diagram in Figure 5.2.

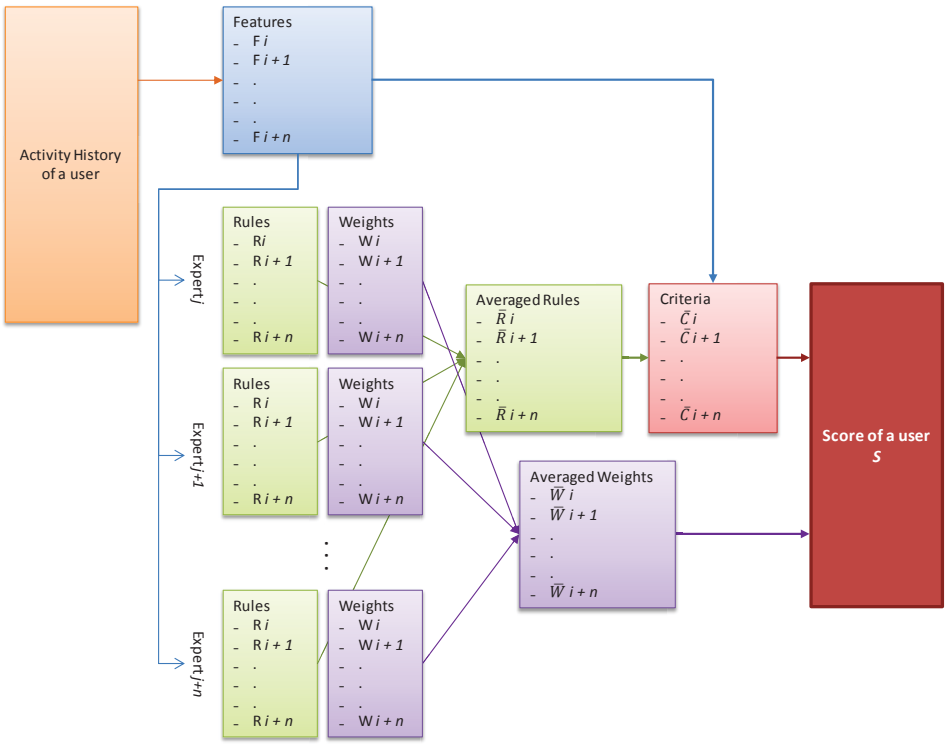

Figure 5.2 Flow diagram of the proposed multi-criteria evaluation system. Firstly the features are extracted from YouTube and experts define the rules that are applicable to each feature and weigh the features based on their importance. A final score is calculated for each user based on the values assigned for each criterion.

Feature values reflect the characteristics of users. For example the number of profanity words used in the comments partially represents the characteristics of a user. Considering another feather about the age or the gender of the user, would improve the representation, distinguishing between a teenage boy and a middle-aged lady using profanities in their comments. To be able to combine two features we need to generate a corresponding numerical proxy for each feature. This numerical proxy is called a *criterion*. The absolute values of each feature (F_i) correspond to the

numerical values for each criterion (C_i). To assign such correspondences we asked experts ($E_{j\ j=1\ldots m}$) to set a rule (R_i) for each of the features. The expert assigns her/his knowledge in form of likelihood (P) given the feature. We set four linguistic probabilities (Bárdossy and Fodor, 2004, Xu et al., 2003a). "very unlikely", "unlikely", "likely", and "very likely" in order to subsequently derive a numerical value to represent the likelihood (Figure 5.3). We then applied the rule (R_i) on the corresponding feature (F_i) to calculate the value for the criterion. Experts were also asked to assign weights ($W_{i\ i=1\ldots n}$) to the criteria ($C_{i\ i=1\ldots n}$) corresponding to its relative importance. The final score for each user is calculated by taking the weighted average of the criteria.

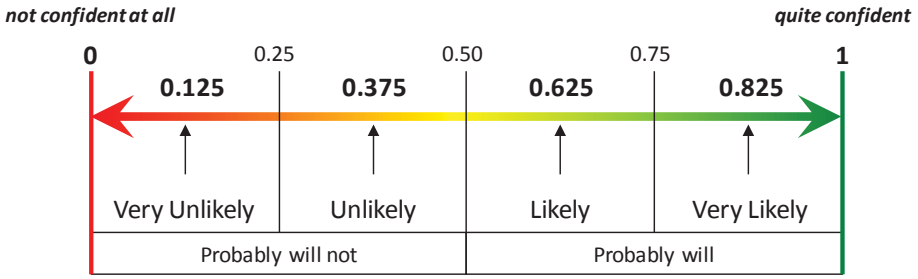

Figure 5.3 Word-to-probability relationship adopted from (Xu et al., 2003b)

Example box 1:
To set a criterion for the age of users (F_1), we provided two experts (E_1) and (E_2) with F_1.

The first expert (E_1) sets two rules for the given Feature:

- $R_{E1,F1\text{-}1}$ = if the age of the user is between 13 and 16 years old, then it is "very likely" that the user is a bully
- $R_{E1,F1\text{-}2}$ = if the age of the user is higher than 16 years old, then it is "less likely" that the user is a bully

And the second expert (E_2) sets the following rules:

- $R_{E2,F1\text{-}1}$ = if the age of the user is between 13 and 16 years old, then it is "likely" that the user is a bully
- $R_{E2,F1\text{-}2}$ = if the age of the user is higher than 16 years old, then it is "less likely" that the user is a bully

Both of the experts have assigned weights to the feature;

- $W_{E1,F1} = 4$ (out of 4), and $W_{E2,F1} = 3$ (out of 4)

We used the linguistic probability values (see figure 5.2), in order to subsequently derive numerical values to represent criteria;

- $C_{E1,F1\text{-}1} = 0.875$, $C_{E1,F1\text{-}2} = 0.375$, $C_{E2,F1\text{-}1} = 0.625$ & $C_{E2,F1\text{-}2} = 0.375$

We also asked experts (E_1) and (E_2) to set a criterion for profanity in the username (F_2):

- $C_{E1,F2\text{-}yes} = 0.625$, $C_{E1,F2\text{-}no} = 0.375$, $C_{E2,F2\text{-}yes} = 0.625$, $C_{E2,F2\text{-}no} = 0.125$, $W_{E1,F2} = 3$, & $W_{E2,F2} = 2$

Using equations (1) and (2) we combine and calculated the average of the each criterion and the corresponding weights assigned by expert panel:

$$\bar{C}_1 = \frac{1}{n}\sum_{i=1}^{n} C_i \qquad (1)$$

$$\bar{W}_1 = \frac{1}{n}\sum_{i=1}^{n} W_i \qquad (2)$$

The bulliness score of each user is then calculated by taking the weighted average of the criterion:

$$\bar{C}_1 = \frac{1}{n}\sum_{i=1}^{n} C_i \qquad (3)$$

Example box 2:

Applying equation (1) and (2) we have

- $\bar{C}_{F1-1} = \frac{0.875+0.625}{2} = 0.750$, $\bar{C}_{F1-2} = \frac{0.375+0.375}{2} = 0.375$
- $\bar{C}_{F2-yes} = \frac{0.625+0.625}{2} = 0.625$, $\bar{C}_{F2-no} = \frac{0.375+0.125}{2} = 0.250$
- $\bar{W}_{F1} = \frac{4+3}{2} = 3.5$, $\bar{W}_{F2} = \frac{3+2}{2} = 2.5$

So, for a user of 15 years old and profanity in the username the final score would be:

$$S = \frac{(\bar{C}_{F1-1} \times \bar{W}_{F1}) + (\bar{C}_{F2-yes} \times \bar{W}_{F2})}{\bar{W}_{F1} + \bar{W}_{F2}} = \frac{2.625 + 1.5625}{6} = 0.69$$

Or for a user of 45 years old and no profanity in the username the final score would be:

$$S = \frac{(\bar{C}_{F1-2} \times \bar{W}_{F1}) + (\bar{C}_{F2-no} \times \bar{W}_{F2})}{\bar{W}_{F1} + \bar{W}_{F2}} = \frac{1.312 + 0.625}{6} = 0.32$$

5.3.2 Experimental Setup

We conducted an experiment in which we used the MCES to assign a bulliness score to the YouTube users in our dataset. The knowledge and experience of experts were used to determine the predictive value of

certain online activities (such as posting comments and uploading videos) and user profiles in their behaviour. For example, according to our experts, online activities can reveal how introvert or extravert a user is, or how active or passive he or she is in online discussions. In the following sections the elements which form the criteria of an MCES are described. Moreover, the size and composition of the expert panel and knowledge elicitation procedure are extensively explained.

Dataset

For the purpose of our experiments we needed a dataset based on a social network platform in which personal information for the users was accessible; the users would have to be involved in variety of online activities in the network. Moreover, the social network to be chosen had to be a platform that exhibits bullying behaviour while also representing the general online population. As explained in Chapter 3, YouTube is one of the social networks which meets these requirements.

Other social networks, such as Facebook, in which bullying takes place could provide alternative datasets. However there is a very limited public access to Facebook users' information. Moreover, based on 2013 cyberbullying report[1] YouTube stands on the second place in bullying rate compared to other social networks which makes it an eligible choice for a dataset to be used in cyberbullying studies. Our dataset contains 3,825 users with a total of 54,050 comments. For more details on the dataset see Section 3.3.

[1] http://www.ditchthelabel.org/annual-cyber-bullying-survey-cyber-bullying-statistics, [Accessed November 2013]

Feature Space

Based on a literature review on cyberbullying and social behaviour factors (Maia et al., 2008, Wilson et al., 2009), and consultation of domain experts, we compiled a set of 13 features in three categories to identify bullying users. These features were used to generate the criteria of the MCES. The selection of the features was limited to what is technically possible to extract from YouTube. The features are grouped in three categories, representing the characteristics, actions and behaviour of the users, respectively (also see Table 5.1).

- The **user features** consist of the personal and demographic information derived from the users' profiles: *(F1)* The age of the users, divided in 5 age groups: below 15, 15-19, 20-24, 25-29, and above 30 years old. The categories correspond to the official age categories in North America[1] so that it can be better adopted with the findings of social studies using the same division. As explained in Chapter 4, cyberbullying differs across different age categories. Frequency of bullying incidents as well as choice of words and language structures change in different age groups. The youngest age at which users can sign up in YouTube is 13 years old and the categories correspond to educational level. We assume that the provided information in the user profile is correct, but we are aware of the fact that this might not be the case (see discussion in Chapter 6). *(F2)* The membership duration of the users, divided into 3 groups: less than 1 year, 1- 3 years and above 3 years.
- The **content features** are derived from the content of the user comments and pertain to the writing structure and usage of specific words. (*F3*) *Number of profane words* in the comment based on a dictionary of profanities (Dadvar et al., 2013), normalized by the total number of words in the comment. The dictionary consists of

[1] http://www.statcan.gc.ca/concepts/definitions/age2-eng.htm, [Accessed November 2013]

414 profane words including acronyms and abbreviation of the words. The majority of the words are adjectives and nouns. To identify frequent bullying topics such as minority races, religions and physical characteristics the manually compiled set of cyberbullying words were also added to the dictionary. (*F4*) *Length of the comments*, which is relevant information as bullying comments are typically short (Yin et al., 2009). To detect the comments which are personal and targeting a specific person, we included the normalized number of (*F5*) *first person pronouns (For example, I and my)* and (*F6*) *second person pronouns (for example, you and yours)* in the comment (see 4.3.1.2). (*F7*) *Usernames containing profanities*; YouTube users can choose their username to be their real name and/or surname or can choose any other aliases and combinations of symbols and words. We believe that it is more likely that users with bad intentions would hide their real identity. We used the same profanity dictionary as in F1, plus a list of most common punctuation marks which can be used in the usernames; (*F8*) *Non-standard spelling* of the words in the users' comments. This includes misspellings (e.g. 'funy' instead of 'funny'), or informal short forms of the words that are used in online chats and posts (e.g. 'brb' which means 'be right back').

– The **activity features** help to determine how active the user is in the online environment. One of the common activities of the users is to upload videos. A user can also post comments on uploaded videos and respond to other users' comments. Most of the YouTube users have a public channel, in which they upload their videos and in which their activities such as posted comments can be viewed. Users can subscribe to others channels and follow the activities of the owner of the channel if they find it interesting. In this feature set we consider (*F9*) *number of uploads*, (*F10*) *number of subscriptions* and (*F11*) *number of posted comments*.

Table 5.1 The summary of feature sets and the units in which they have been presented

Feature Set	Feature Name		Unit	Details
User features	F1	Age	Categorial	Categories; below 15,15-19,20-24,25-30, and above 30 years old.
	F2	Membership duration	Categorial	Categories; less than 1 year, 1-3 years, More than 3 years.
Content features	F3	Profanities and bullying sensitive topics	Numerical	Average per comment in YouTube: 1.2 %
	F4	Length of the comments	Numerical	Average in Youtube: 12 words
	F5	First person pronouns	Numerical	Average per comment in YouTube: 2.2 %
	F6	Second person pronouns	Numerical	Average per comment in YouTube: 2.3 %
	F7	Profane words in the username	Boolean	True, False
	F8	Non-standard spellings	Numerical	Average per comment in YouTube: 21.5 %
Activity features	F9	Number of uploads	Numerical	Average in YouTube: 4.56
	F10	Number of subscriptions	Numerical	Average in YouTube: 23
	F11	Number of posted comments	Numerical	Average in YouTube: 14.4

Expert Panel

To gather knowledge and opinion of experts on the elements that covey information about YouTube users' characteristics and behaviour, a panel of twelve experts in the area of cyberbullying was convened. The experts had

a background in psychology, social studies and communication sciences. The majority of the panel works on cyberbullying causes, effects and solutions from social and psychological perspectives. A smaller number works on social behaviour, psychology and communication studies. During a preliminary meeting the purpose of this questionnaire was explained. The experts completed the questionnaire individually. The questionnaire required approximately 20 minutes to be completed. It took about three weeks to receive all the responses from the expert panel. The outcome of the questionnaire was a set of rules and weights set by the expert corresponding to the given features.

Expert Knowledge Elicitation

Experts were provided with an online questionnaire, consisting of 22 factual questions. To avoid ambiguities, each question also provided a brief definition of the concepts addressed. For each of the features experts were asked to express their opinion on the likelihood that a bully user belongs to a certain category relevant for that feature. For example, "What is the likelihood that a bully user belongs to the following age categories?" where the age categories are given in the question. In these type of questions, experts could express their opinion through a four-point scale answering options (Xu et al., 2003a, Bárdossy and Fodor, 2004); 'Unlikely', 'Less likely', 'Likely' and 'Very likely' corresponding to values 0.125, 0.375, 0.625 and 0.875 respectively (see Figure 5.2). We also added the 'I don't know' choice to the available options. Experts could provide comments at the end of each question. To understand how informative and helpful the features are in the determination of personality and potential behaviour of a user we also asked experts to weigh the features. The experts' could choose from 4 values: not informative, partially informative, informative and very informative. The main point that experts had commented on in the questionnaire was the importance of combining one or more criteria. For example: "if age is above 30, the number of profanities would have a

lower likelihood than otherwise". Therefore we added two combined criteria to those explained earlier. The combined criteria were based on the expert's comments: age and profanity (C1), and age and misspellings (C2). The questionnaire is presented in Appendix I.

Evaluation of MCES Performance

To measure the agreement among the expert panel, the assignments were analysed in terms of overall disagreement among experts. For each expert, we compared the value that an expert had assigned to each criterion, to the 'median value ± 1' of that criterion assigned by all experts. If the assigned value was out of this range, it was considered as a different opinion and therefore a disagreement. The final disagreement rate was calculated by taking the ratio of the total number of disagreements to the total number of opinions expressed by each expert on all the criteria.

To evaluate the performance of the proposed MCES approach, we assessed the discrimination capacity of our model to rate potential bully users using the independently manually labelled dataset. We measured the area under receiver operation characteristic curve (ROC) to quantify discrimination power of the model. A ROC curve plots "sensitivity" values (true positive fraction) on the y-axis against "1–specificity" values (false positive fraction) for all thresholds on the x-axis (Fielding and Bell, 1997). The area under such a curve (AUC) is a threshold-independent metric and provides a single measure of the performance of the model. AUC scores vary from 0 to 1. AUC values of less than 0.5 indicate discrimination worse than chance; a score of 0.5 implies random predictive discrimination; and a score of 1 indicates perfect discrimination.

5.3.3 Results

To implement the MCES we first averaged the weights and the criterion corresponding to each feature using the output of the questionnaire. The experts' agreement rate regarding the assigned likelihood values was 95%.

Figure 5.4 illustrates the relative importance of each feature based on the weights that were assigned to them by the experts. According to the weights, profanities and bullying sensitive topics in the history of a user's comments constitute the most informative feature (average weight equals 3.6). In other words, this feature most strongly predicts future bullying behaviour. The second and third most informative features are the inclusion of profanities in usernames (average weight equals 3.2) and age (average weight equals 3) respectively. The least informative feature is the number of non-standard spellings in the history of users' comments (average weight equals 1.7).

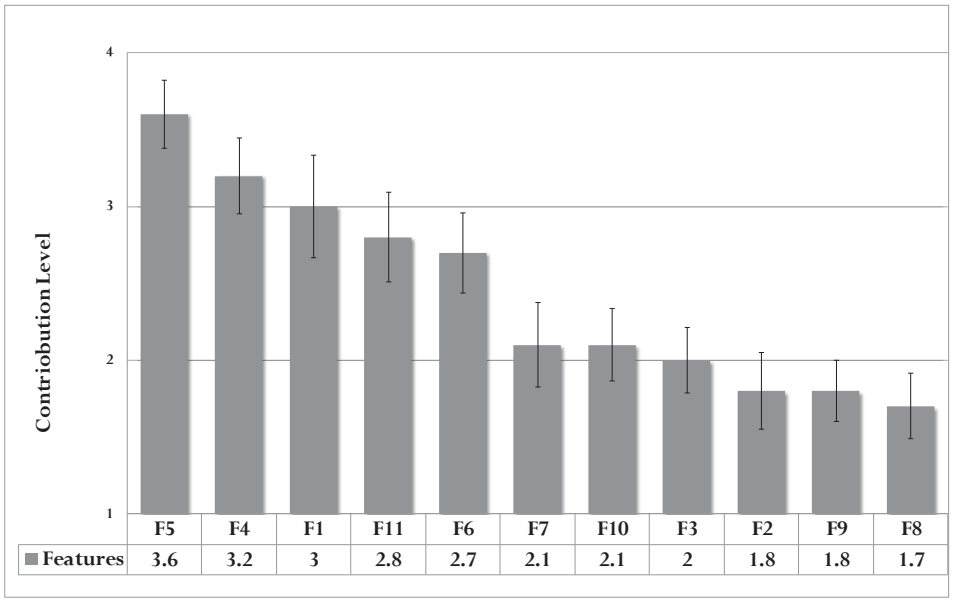

Figure 5.4 Feature weights indicated by experts (feature codes corresponds to table 5.1)

Table 5.2 Results of experts' knowledge elicitation. The Likelihood column illustrates the likelihood that a bully belongs to each of the feature relevant to the specified criteria. The features' ids are identical to Table 5.1. The Std column shows the standard deviation of each category.

Feature ID	Rule		Likelihood	Std
F1	R1-1	IF F1 < 15 THEN	0.725	0.242
F1	R1-2	IF 15 =< F1 < 20 THEN	0.597	0.232
F1	R1-3	IF 20 =< F1 < 25 THEN	0.431	0.167
F1	R1-4	IF 25 =< F1 < 30 THEN	0.344	0.088
F1	R1-5	IF F1 > 30 THEN	0.268	0.134
F2	R2-1	IF F2 < 1 THEN	0.525	0.224
F2	R2-2	IF 1 < F2 < 3 THEN	0.475	0.224
F2	R2-3	IF F2 > 3 THEN	0.275	0.137
F3	R3-1	IF F3 < Average THEN	0.688	0.222
F3	R3-2	IF F3 > Average THEN	0.375	0.000
F4	R4-1	IF F4 = True THEN	0.700	0.237
F4	R4-2	IF F4 = False THEN	0.225	0.129
F5	R5-2	IF F5 < Average THEN	0.375	0.000
F5	R5-2	IF F5 > Average THEN	0.688	0.222
F6	R6-1	IF F6 < Average THEN	0.339	0.173
F6	R6-2	IF F6 > Average THEN	0.732	0.197
F7	R7-1	IF F7 < Average THEN	0.375	0.000
F7	R7-2	IF F7 > Average THEN	0.375	0.000
F8	R8-1	IF F8 < Average THEN	0.375	0.000
F8	R8-2	IF F8 > Average THEN	0.486	0.182
F9	R9-1	IF F9 < Average THEN	0.500	0.231
F9	R9-2	IF F9 > Average THEN	0.375	0.125
F10	R10-1	IF F10 < Average THEN	0.458	0.204
F10	R10-2	IF F10 > Average THEN	0.417	0.102
F11	R11-1	IF F11 < Average THEN	0.236	0.132
F11	R11-2	IF F11 > Average THEN	0.725	0.211
F1 / F5	Rx1-5	IF F5 > Average AND F1 > 30 THEN	0.675	NA
F1 / F8	Rx8-2	IF F1 < 15 AND F8 > Average THEN	0.125	NA

An average likelihood was also calculated and then assigned to the subcategories in each feature set. For some features we should have compared the user's feature value with the average value of the feature in the YouTube community. How high or low the value for a certain feature is, was measured in comparison to the average value of that feature in YouTube. The experts' choices on the likelihoods, was taken to correspond to values of each choice.

The user group in the age range between 13 and 15 years is indicated to be most likely to contain bullies. The age category above 30 years old is ranked as corresponding to the lowest bulliness likelihood. Experts indicate that a typical bully has a membership period shorter than 1 year. The outcome of the questionnaire also indicate that it is more likely that bully users have a high ratio of second person pronouns in their comments as well as profane words in their usernames. In the Content features set these latter two features are ranked as being the highest predictor of bullying behaviour. Moreover for potential bullies the likelihood of writing short and right-to-the-point comments is higher than the likelihood of lengthy ones. In the Activity features set, a high number of posting comments is considered to be the strongest indicator of bulliness in comparison to the other features. The number of uploads comes in the second place in this set. Table 5.2 illustrates the detailed results and the likelihood of each feature.

Table 5.3 The performance of the MCES using different settings to discriminate potential bully users. Score range represents the minimum and the maximum bulliness score assigned to a user.

MCES setting	AUC	Score Range
Non-weighted Criteria	0.71	0.33 – 0.65
Weighted Criteria	0.72	0.29 - 0.71
Weighted Criteria + Combined	0.74	0.29 - 0.75

5.3.4 Discussion

In this experiment we built an MCES to assign a bulliness score to YouTube users. We employed experts' knowledge to set the criteria and to define the rules and the corresponding weights to be used in the MCES. Evaluating the results of our model using the annotated YouTube dataset (introduced in Chapter 3) revealed that the bulliness score can discriminate among users with a bullying history and those who had not been engaged in hurtful interactions. Our approach is based on human experts who provided the knowledge underlying the rule-based rating scheme. It is flexible towards inconsistencies among different sources of information and can be easily fine-tuned by adding specific criteria to identify forms of human behaviour that are hard to capture. According to the experts' opinion, the importance of features differs and experts have assigned different weights for each criterion. However our results show that the discrimination capacity of the models has slightly improved after considering the weights of the features. This can be due to the disagreement among experts on the weights though within common boundaries, i.e. >80% adjacent likelihood classes- Since we had a relatively small number of experts, this disagreement may have neutralized the effect that weights should have had on the results. Having a larger expert panel with more experts may diminish the effect of the above mentioned disagreement. The advantage of our approach is that the questionnaire can be easily updated and adapted to new online environments by adding the extra features and there is no need to develop any training data to train a new model. We also demonstrated another advantage of the proposed approach; since we can set the criteria according to experts' experiences, we can easily carry out and fine-tune the studies from their perspective. As was suggested by the experts, combining the criteria can lead to more accurate and meaningful information about users. Therefore coming up with more combined criteria, or providing opportunities for experts to make more complex criteria may improve the performance of the models. For example, one expert argued that it is not sufficient to only know how

active a user is at a specific moment in time, but we have to study its frequency and changes over time, as the level of activities or harassing behaviour may vary in time.

5.4 A Hybrid Approach for Automatic Rating of Bully Users

For the second experiment presented in this chapter, the aim was to improve the MCES approach by combining it with machine learning models. The main limitation of the expert system approach is that it cannot make use of abilities that machine learning approaches have as it fails to incorporate the complex yet informative textual patterns. To complement what we started in the previous section we propose an approach that uses the potential of machine learning together with experts' knowledge for determining the level of bulliness of users in YouTube. Each of the two approaches has its weaknesses when used individually. Machine learning models can analyse and extract complex textual patterns that cannot be captured by criteria in an expert system. Expert knowledge, on the other hand, could overcome some limitations of machine learning approaches, such as the biases and noise in the training data. Experts can rely on their judgment and knowledge to come up with rules that generalize better to unseen data. Words can be put together in a sentence in many different ways, conveying various intentions. This makes it hard for data-driven methods to generalize based on the information that they have already encountered.

In this section we demonstrate that a hybrid detection approach based on a machine learning model and an expert system can yield results for bully rating that outperform the individual approaches. We combine machine learning and expert systems into a hybrid system in two ways:

- using the outcome of the expert system as an extra feature for training the machine learning model (H1), and
- using the results of the machine learning model as a new criterion for the expert system (H2).

In Section 5.5 it is described how we trained three supervised classifiers to identify bully users based on machine learning models. In Section 5.4.2 the experimental settings and the three features sets used to train the classifiers are explained.

5.4.1 Hybrid Approach

We experimented with three well-known supervised machine learning methods, which learn from pre-labelled training data: a Naive Bayes classifier (Lewis, 1998), a classifier based on decision trees (C4.5) (Joachims, 1998) and Support Vector Machines (SVM) (Vapnik, 1998, Cristianini and Shawe-Taylor, 2000) with a linear kernel (Witten et al., 2011). For training the classifiers the implementation available in WEKA 3 was used (Hall et al., 2009).

We combine the MCES and the machine learning methods into a hybrid classifier in two ways, as illustrated in Figure 5.5. In the first setting (H1) a hybrid system is formed by adding the following features to the machine learning classifier: (1) the results of the MCES, (2) the features' categories that were used in the expert system as new set of features, and (3) the combined features (C1 and C2). This means, that the classifier will be re-trained with the previous feature sets plus the new ones. In the second setting (H2) the classification obtained from the machine learner is used as an additional criterion in the MCES. As was done previously for MCES, we again assigned equal weights to all the criteria used in the system, including the machine learner criterion.

5.4.2 Experimental Setup

In the following sections we will explain the features that are used for training the machine learning models. The evaluation method used for evaluation of the performance of the two hybrid settings will be also described.

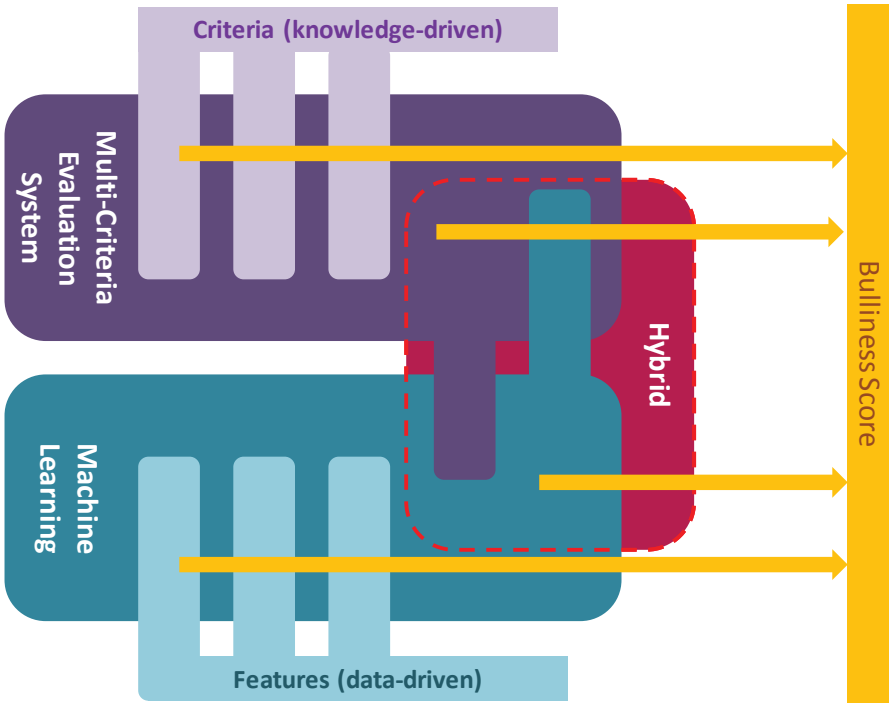

Figure 5.5 Conceptual representation of the hybrid approach for bulliness score

Feature Space

To make the machine learning model comparable to the expert system model, we used the same set of features that is used in the expert system for the implementation of the machine learning models. (Cf. Section 5.4.2 where we introduced three feature sets: user features, content features and

activity features.) We also added several features which are only interpretable by the machine, introduced earlier in Chapter 4. For example, to capture shouting in comments, the ratio of capital letters in a comment was computed (*F12*). To capture explicit emotions, the number of emoticons was counted and normalized by the number of words (*F13*). Profanity windows of different sizes (2 to 5 words) were chosen (*F14*). These are Boolean features that indicate whether a second person pronoun is followed by a profane word within the size of the window. The term frequency–inverse document frequency (TFIDF) value of the words was also computed (*F15*). The combined criteria (C1 and C2) and categories for age and membership duration were excluded from the feature space.

Evaluation

To evaluate and compare the performance of the approaches, we evaluated the discrimination capacity of our approaches by analysing its receiver operation characteristic (ROC) curves. ROC curves are extensively explained in Section 5.4.2. To evaluate the performance of the machine learning classifiers we used 10-fold cross-validation to measure the ROC value. As a baseline, we trained an SVM classifier (SVM_B) using the content features that are extracted only from the body of the comments (i.e. F3-F6).

5.4.3 Results

According to our findings regarding the performance of the machine learning classifiers, the decision tree classifier performed the worst, followed by the SVM classifier. The Naive Bayes classifier with discrimination capacity of 0.66 outperformed the other two algorithms. The contribution of each feature was assessed by excluding it from the feature sets. The results revealed that the number of profane words,

second person pronouns and pronoun-profanity windows, were the strongest contributing features. Capital letters and emoticons however gave only a minor contribution. We used the MCES with the best performance (see section 5.3 for more details) as an extra feature for the machine learning models to the build first hybrid approach (H1). The results of hybrid approach H1 show improvements in the discrimination capacity of all machine learning methods. The new feature was not very informative for the decision tree algorithm. Although the SVM gained the highest improvement, the Naïve Bayes classifier still outperformed in discrimination capacity. The overall improvements in all three machine learning approaches in the first setting (H1) was significant (two sample t-test, $P<0.05$).

Table 5.4 The performance of models to discrimination the potential bully users, comparison of machine learning methods, Expert System and two hybrid approaches. Score range represents the minimum and the maximum bulliness score assigned to a user.

	Approach	AUC	Score Range
Baseline	SVM_B	0.57	0.44 – 0.54
Machine Learning	Naive Bayes	0.66	0.39 - 0.55
	Decision Tree	0.52	0.44 - 0.52
	SVM	0.59	0.41 - 0.58
Expert System	MCES	0.74	0.29 – 0.75
Hybrid	H1 (Naive Bayes +	0.73	0.35 – 0.65
	H1 (Decision Tree +	0.55	0.43 – 0.55
	H1 (SVM + MCES)	0.69	0.40 – 0.61
	H2 (MCES + Naive	0.76	0.30- 0.75

For the second setting (H2), we selected the outcome of the best machine learning approach (Naive Bayes) and added it to the criteria set of the MCES. The discrimination capacity of the MCES was improved to 0.76 and outperformed the other models. Further investigation revealed that although the discrimination capacity of the MCES has improved significantly, the range of bulliness scores decreased. This means that most of the scores accumulated around the moderate scores (i.e. 0.4 – 0.6). However, the ideal situation would be to have users with bullying incidents in the higher scores (i.e. 0.7 – 0.9) and non-bully users in the lower scores (i.e. 0.3- 0.1). Table 5.4 presents the AUC value in each approach and the ROC curves are illustrated in Figure 5.6.

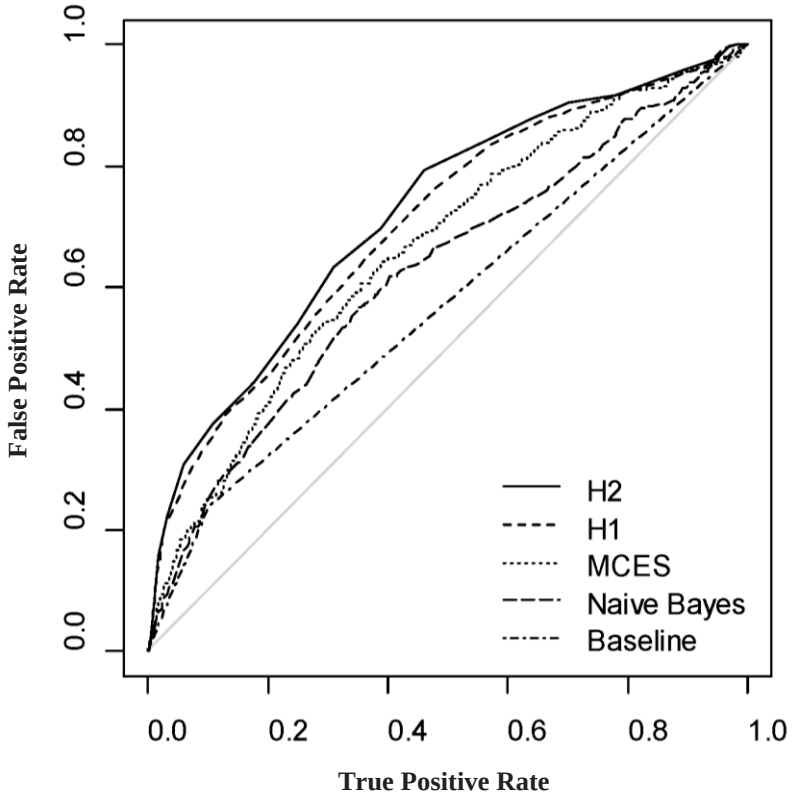

Figure 5.6 Comparison of receiver operating characteristic curves

5.4.4 Discussion

The results show that the performance of models to discriminate potential bully users improved when coupled; either by adding the outcome of MCES as a feature to machine learners, or considering the result of a machine learner as a criterion in the hybrid MCES.

Among machine learners in this experiment, Naive Bayes outperformed the other two algorithms. At the moment we cannot provide a clear explanation for this result. However, from the perspective of computer science it would be interesting to investigate what characteristics of cyberbullying made the naïve base outperform other classifiers. Moreover, the results show that the discrimination capacity of MCES outperforms the machine learning models. A possible explanation for the observed differences between the machine learning methods and the expert system is the sensitivity of the machine learning methods for class skew, which was quite high in our dataset (12% bullying, 88% non-bullying).

For cases like cyberbullying incidents that reflect users' mind set and that are highly influenced by human characteristics, incorporating human reasoning can result in a better coverage of the users' intentions and subtle indications of agitation can be included as important sources of information about the users. Since the MCES we used was an absolutely deductive approach, it was not affected by the quality of a training set. Experts express their opinion without being biased by a particular dataset, therefore the multi-criteria system using their knowledge assigned more reliable score to bully users, compare to machine learning, which were mainly distracted by the data itself. On the other hand, the superiority of machine learning model over expert system is its capability of analysing complex patterns which cannot be easily expressed as assessable criteria by experts, such as TFIDF values or occurrences of pronoun-profanity combinations. We distinguished between two settings for combining the MCES models with a machine learning model. We compared the results for the two hybrid approaches H1 and H2 in order to investigate which one

would results in a more accurate discrimination of bullies; using the MCES output as a new feature in machine learning methods or the machine learners as a new source of knowledge for the MCES. We demonstrated that for the H2 setting, i.e. coupling the machine learning results to the expert system as a new criterion, the results improved the MCES-only setting and outperformed all the other system variants considered.

5.5 Conclusion

In this chapter we proposed and tested two approaches for bully user rating. This functionality could be deployed for the tackling of cyberbullying in YouTube. The proposed approaches, a certain expert system (MCES) and a hybrid system combining an expert system and machine learning models, may help to identify social network users who may act in a hurtful way. With this functionality network administrators and moderators are better equipped for the timely stopping of potential bullies from causing any further harm.

We combined the potential of MCES as a deductive method, with the inductive approach underlying machine learning algorithms to improve the reliability of the bulliness scores in comparison to the MCES-only method described in Section 5.4. Based on a comparison of the sores with a manually annotated set consisting of YouTube data, we demonstrated that a hybrid approach performs better in comparison to each of the individual approaches. Our approach can be further investigated in three main directions. Machine learning algorithms need to be fine-tuned to cope with the characteristics of cyberbullying datasets. Demonstration of the preliminary outcome of machine learning approach, to the experts, might facilitate the knowledge elicitation procedure and the expression of implicit knowledge. Spatial features such as location of the users as well as temporal features might improve the accuracy of the score. The proposed

approaches can be used in other social networks as well. Depending on the network under study and the design of the platform, the activity features may vary. For example if the approach would be applied on data from Twitter, the number of followers could be selected as one of the activity features. Our approaches are in principle language-independent and adaptable to other languages just by making the required modifications in the dictionaries.

Chapter 6

Conclusion

6.1 Introduction

In this thesis we presented a multi-perspective study on cyberbullying in social networks. Our first step was to understand the source and nature of the problem as a social phenomenon in order to identify the aspects for which measures could be developed to reduce the volume and the impact of the problems caused by it. An important insight from this study was that in spite of the fact that the origin of the problem of cyberbullying roots into the complexities of human mind and darker sides of human beings, its solution also depends on having a good understanding of human characteristics and mind set. Either to detect a bullying incident which has happened or to identify people who are capable of online aggression, we need to know the factors that distinguish bullying cases and bully users from the others. It also became obvious that we needed to have a clear definition of the phenomenon called bullying, as not any profanity expressed via social media can be considered as a bullying case. Friends may use more informal language among themselves and use slang or foul words just as a sign of their close relationships.

To increase the understanding of the context of cyberbullying and to make it easier to present our views, we came up with a framework to talk about cyberbullying. We defined two phases, pre-bulling and post-bullying. For the post-bullying phase we looked into the elements which can be influential in detecting bullying incidents after they have happened. For the pre-bullying phase we proposed a workflow for gaining knowledge about the intention of the bullies and for identifying the users of social media who potentially will develop misbehaviours. In our investigations, we have put the focus on two aspects in particular: (1) detecting textual bullying comments that are posted by bullies and (2) prevention of further bullying incidents by identifying the potential bully users. Our main goal with designing tools for the detection of cyberbullying incidents was to improve and optimize the few existing detection algorithms. We learned that incorporation of findings from other fields of research, such as social,

psychological and behavioural studies, can have added value and provide further beneficial information. Therefore we integrated personal information from the profiles of users, such as details on age and gender, into the detection models, in order to improve the accuracy of the algorithm for the detection of bullying incidents.

Obviously it would be even better if we could help users of online platforms not to go through this devastating experience in the first place. Therefore, instead of only focusing on the detection of bullying incidents after they have taken place, we also dedicated a large amount of our studies to preventive approaches for cyberbullying. Particularly, we investigated the identification of attributes in writings and online activities of the users, which convey information regarding their intentions and characteristics. We used expert knowledge to analyze these attributes and to produce a score which represents the level of bulliness for individual user and shows the potential for future misbehaviour.

In this final chapter, we conclude by summarizing the objectives and research questions introduced in Chapter 1, and the answers provided in chapters 2 to 5. We also present a discussion of the limitations of our approach, and suggest directions for future works.

6.2 Revisiting Research Objectives

In the coming sections we will revisit the research objectives and questions which have been investigated in the course of our project. In order to summarize the contributions of this thesis to the societal problem of cyberbullying and to the field of computer science, the answers to the research questions and outcomes of our experiments will be outlined here too. The objectives of our research can be listed as follow:

Objective1: To present a view on cyberbullying that underlines the kinship with traditional bullying.

Objective2: To create a comprehensive dataset suited as a basis for experimental cyberbullying studies.

Objective3: To improve the accuracy of algorithms for the detection of bullying comments in social networks.

Objective 4: To design a bulliness score for identifying potential bullies in social networks

6.2.1 A Novel Outlook Towards Cyberbullying in Virtual Societies (Obj. 1)

In Chapter 2, we introduced a novel outlook towards the cyberbullying phenomenon. We looked into the gradual changes which have occurred in relationships and social communication with the emergence of the Internet. We analyzed some of the positive and negative effects that this transition has had on human life and society. We argued that one should look at virtual environments as virtual communities, because the human needs projected on these environments, the relationships, human concerns and misbehaviours have the same nature as in real-life societies. Therefore, to make virtual communities safe, we need to take safety measures and precautions that are similar to the ones that are common in non-virtual communities.

In essence, cyberbullying as one of the troubling misbehaviours in virtual environments is the transformed technologized version of the traditional bullying that teenagers and adolescents have been struggling with for ages. In fact, we are dealing with an old social problem which has adapted itself to new social conditions. We derived the assumption that cyberbullying is recognized and treated as a social problem and not just seen as some random mischief conducted by individuals with the use of technology, the methods for handling its consequences are likely to be more realistic, effective and comprehensive.

The main limitation of existing studies on cyberbullying is that they have approached this problem only from one perspective at a time, either social or technical. But in order to tackle this problem, behavioural and psychological studies, and the study of technical solutions should go hand in hand.

The issue of how to interpret the transition of traditional bullying into cyberbullying and its roots in societal and psychological constants is thoroughly addressed in Chapter 2. This part of our study led to the conviction that for combating cyberbullying a broader angle than a merely technical one should be taken. This brought us to the approaches proposed in Chapter 4 and Chapter 5 which at the time of conducting the experiments were the first attempts to take user characteristics into account for the detection of bullying incidents. We combined technical advances with findings from the social sciences to reach milestones in filling the gaps in studies on combating cyberbullying.

6.2.2 A Comprehensive Dataset for Cyberbullying Studies (Obj. 2)

One of the primary limitations that we faced when we started our research was the lack of a comprehensive dataset for cyberbullying studies. As explained in chapter 3, we needed a dataset which include real instances of bullying incidents. Moreover, it was essential for our studies to also have information about the people who generated and posted the bullying comments. Therefore we needed our dataset to have the demographic information of the social media users as well as the history of their activities.

Absence of a suitable dataset is a common problem in related studies on cyberbullying. Lack of common data or a benchmark dataset including all

the information required for investigations into patterns of cyberbullying, made it hard to compare the findings of the various studies and approaches.

In the absence of a suitable dataset we started our preliminary experiments using a dataset that only partially met the requirements. This dataset was collected from MySpace forums and was provided by Content Analysis for the Web 2.0 workshop on 2009. The details and attributes of this dataset are thoroughly explained in Chapter 3. The MySpace dataset did not meet all the requirements for our experiment, namely in terms of size (i.e. very few proportion of bullying posts), and sufficiency of information (i.e. very limited information on users' activities). Therefore we developed our own dataset, with the aim to encompass extensive information about the users and their activities as well as larger number of bullying comments. Recently the dataset has been made available to other researchers in this rapidly growing field of study.

We chose YouTube as the platform for our study since it ranked as second among the social networks regarding the frequency of cyberbullying incidents. Moreover, it offers its members a variety of options for online activities and/or communication. We collected information on user activities and posted textual comments for 3825 users for a period of four months, as well as personal and demographic details of the users involved. All the collected information was publicly accessible and scraped directly from YouTube. The dataset meets the requirements of size and balance inherent to the type of study conducted. Detailed information about the process of data collection as well as the quality, statistics and characteristics of the dataset has been explained in detail in Chapter 3.

6.2.3 Improved Cyberbullying Detection Accuracy (Obj. 3)

Detecting a bullying comment or post at the earliest possible moment in time and taking the required measures for removing the harmful content or reporting it to the responsible authorities can substantially decrease the

negative effects of cyberbullying incidents on the victims involved. The questions posed as part of the third research objective were aimed at finding ways for improving methods for automatic cyberbullying detection:

- Does considering gender information of bullying users improve the accuracy of cyberbullying detection in social networks?
- Does considering user profile information of bullying users improve the accuracy of cyberbullying detection in social networks?

These questions were answered by experimentally investigating the effect of incorporating gender information of users on the improvement of the accuracy of cyberbullying detection as well as the influence of considering user profile information of the users on improving the accuracy of cyberbullying detection.

In Chapter 4 we were able to show that besides the conventional features used for text mining methods such as sentiment analysis and specifically bullying detection, more personal features can improve the accuracy of the detection models. The main auxiliary features that we considered in our classification models were the personal information of the users retrieved from their profiles. We took the differences that exist in the way that boys and girls bully into account, as exemplified by the use of profanities and other choice of wordings. As expected the models which were optimized accordingly resulted in a more accurate classification. The improved outcome motivated us to look into other personal features as well, such as age and the writing style of users. By adding more personal information, the previous classification results were outperformed and the detection accuracy enhanced even further.

The personal information retrieved from the users' profiles is all provided by the users themselves. There are several reasons why users may not enter their information correctly. For example most of the social networks have a minimum age requirement for membership which causes users to state their ages falsely; another example is the predators who state their gender

differently for deceiving their targets. Therefore, there is the possibility of having noise in the data. From earlier studies it was known that age and gender classification algorithms can verify the correctness of the information provided by users and that the use of such algorithms might be a useful preliminary step to improve the effectiveness and performance of cyberbullying detection algorithms.

Having access to temporal and geographical information of the bullying events, such as the times that a user has posted comments over a period of time, or the location from which the comments have been posted, provides unique features which are specific to each comment. These features can reveal extra information about users' behaviour. Information about the moment in time at which a comment has been posted may indicate at what time of the day users are most busy and bullying behaviour takes place. For example, if the morning is the busiest time, it means that many bullying cases happen at schools and probably through school's computers. If it's midnight it can be inferred that bullies are more active from home. The geographical information can be used in the same manner. Depending on the geographical boundaries, countries with the highest rate of bullying comments can be identified, as well as city areas that are more prone to aggressive behaviour. The patterns that can be derived from these features can be used to point out the more alarming times and locations and consequently to increase the sensitivity of the monitoring systems during that period.

6.2.4 Bulliness Score for Social Network Users (Obj. 4)

The intentions and personality of social networks users can be inferred from their online activities and previous conducts. This information about users can be used to assign each user a score which represent their level of bulliness and the probability of future hurtful acts. The bulliness score we proposed is calculated based on personal characteristics and online

behaviour of users on social media platforms weighted in accordance to the insights collected among a panel of experts. This score represents the probability that a bully will be causing further harm to other users. Assigning a bulliness score to users of social media platforms and thereby predicting their potential for being a bully (Objective 4), is an instrument that can provide input for a preventive protocol or monitoring system that can stop bullies from misbehaviours. Designing a workflow for such an approach would fill the space for interventions in the pre-bullying phase. Two questions related to Objective 4 of our research were posed and addressed in Chapter 5:

- How accurately can an expert system assign a bulliness score to a user to represent the level of bulliness of that user?
- Can an expert system and a system based on machine learning be effectively combined for detecting potential bullies?

To better understand and interpret the intentions underlying online activities of users of social media, we decided to incorporate human reasoning and knowledge into a bulliness rating system by developing a Multi-Criteria Evaluation System. A wide range of expert knowledge, experience and opinions were deployed to support the in-depth analysis of users' behaviour. From the experts' analysis, a series of behavioural patterns emerged that could be summarized as a set of rules. The rules were applied against each user's profile to calculate a score for each individual user.

To have more sources of information and to make use of the potential of both human and machine, we designed a hybrid approach, incorporating machine learning models on top of the expert system. As a preparatory step we calculated the discrimination capacity of the machine learning models as a second baseline. For the hybrid approach we reached an optimum model which outperformed the results obtained from the

machine learning models and the expert system individually. The hybrid models improved 10% on average, in comparison to expert system and machine learning models individually, so our hybrid model illustrates the added value of integrating technical solutions with insights from the social sciences. As discussed earlier, cyberbullying takes place through technological devices, but its causes and nature is close to the essence of the human mind and culture. An approach based on a combination of technical capabilities and the understanding of human behaviour can yield a more effective solution.

We studied the history of activities and behaviour of users in a four-month time frame. Our observations confirmed the intuitive expectation that the length of the time frame is playing an important role in determining the characteristics and mind set of a user. A short period of user activities may therefore not fully represent all aspects of their personality and interests. The shorter the time frame, the more probable that we face a temporary shift in the user's mind set. Tracing the behaviour of a user over a longer period of time would result in a more accurate analysis of the characteristics. In a longer time frame more information can be collected about a user and the judgment won't be based on a short snap of the activities but based on a representative sample of behaviours.

The approach advocated here makes use of information that is collected from a specific social network: YouTube, and in particular: the textual posts from this social media channel. The choice of activity features for the training of the models is of course constrained by the activities that are available in that specific network. Therefore our models are adaptable to any social network by only modifying the activity features which are applicable to that specific network. For example, posting comments is an activity which is possible in almost all the social networks; YouTube, Twitter, Facebook and so forth. Thus, all the features selected from this type of activity, such as the number of profanities used in the comments, length of the comments and etc. are relevant for capturing the patterns in these networks. On the other hand, also some activities are unique for a

certain network, such as subscribing to a personal channel on YouTube or re-twitting a post on Twitter. Depending on the network under study, the activity features should be adjusted to and updated with these unique features and the models should be trained accordingly.

6.3 Future Research and Application

A number of future research avenues have become evident in the course of our research. Throughout this thesis we elaborated at several occasions the reasons why the cyberbullying phenomenon should be considered societal misbehaviour rather than a personal act taken by individuals. As in the real world, the consequences and effects of misconduct in the virtual world can be traced in various societal contexts and the victims may react in different ways and through variety of mediums. For example, when children are bitten at school, they may go to their friends to talk about it or they may write something about it in their diaries. In the case of cyberbullying the equivalent could be a social media chat box, or a digital notebook. If we could have access to this information, we can gain a better understanding of how a child bullied in cyberspace has been affected and how he or she is handling it. The most crucial effects and impact of a bullying incident may not be apparent in the environment in which bullying has happened, but the reaction to the incident may be traceable in another online environment.

All existing studies on cyberbullying have investigated the causes and effects of bullying in a particular environment without considering the possible further reactions of the individuals involved in other social networks. Nowadays, most of the people who are familiar with Internet and social networks are active in several networks at a same time and have personal profiles in each of them. If for instance someone gets bullied on YouTube, the reactions and emotions may be expressed on Twitter and

victims may reveal their feelings and state of the mind through a tweet to their friends or by posting a status on their Facebook profile. Given the multiplatform context of virtual lives, one particular direction to be explored in the future could be cross-system user modelling. Identifying users via interaction over the web is a newly emerging field of work. While providing profile information for social networks or browsing the web, users leave large number of traces. This distributed user data can be used as a source of information for systems that provide personalized services for their users or need to find more information about their users (Abel et al., 2010). Connecting data from different sources has been used for different purposes, such as standardization of APIs (e.g. OpenSocial[1]) and personalization (Carmagnola et al., 2009). The aggregation of users' profiles information and activities from different social networks can provide comprehensive and accurate information about the state of the mind of a user. We believe that studying the social connections of an individual user across different networks might provide a deeper understanding of the situation and consequently offer insight in how to organize support in an optimal manner.

Another important consideration is that not all aggression or use of foul language leads to a bullying case. Getting victimized and feeling threatened is closely dependent on the personality and characteristics of the person involved. A person who is more sensitive and vulnerable may feel bullied, threatened and depressed by the same sentences that do not affect and cause any hurtful feelings in someone with a less sensitive personality. Therefore, even if a sentence contains harassing words and is intended to bully someone, it does not necessarily mean that the other party will feel offended or victimized. The information to be gathered through cross-system users' profile analysis may also shed light on how to predict the impact of the bullying incident on the targeted person in a refined way. Moreover, as mentioned earlier, vulgar language is commonly used among

[1] http://code.google.com/apis/opensocial/

young generation as an indication of friendships and many profanities are used sarcastically. For example the following sentence: "*I hate your guts*" can be interpreted in two ways: the hurtful way, which is expression of hate towards someone, or the funky way, which is expression of liking someone in a cool way. Therefore, as another extension to our current research, identification of sarcastic sentences can be suggested.

A future research track can also be to study the alternatives for acting upon the bulliness scores resulting from the approach proposed in Chapter 5. As explained earlier, the bulliness score indicates the likelihood of a user to conduct bullying behaviour. It is important to study the optimal way in which this information can be put into use and investigate the options for reacting and measures towards bullies. Furthermore, to put the bulliness scores into use, it is required to investigate a threshold which can best distinguish the bully and non-bully users in a social network. This threshold may differ depending on the platform and the target group under study.

Several existing internet safety technologies were introduced in Chapter 2, such as filtering and monitoring software, as well as applications for reporting and blocking undesirable contents. These technologies search for webpages with inappropriate content, conversations with harassing language or undesirable communications in social networks and forums. Choosing the best intervention policy needs further investigation and should be studied from a multidisciplinary perspective including social and psychological angles.

Another observation made throughout our research in Chapter 4, is that besides the bully, other actors involved in cyberbullying or related phenomena also play a very important role. In Chapter 2 we described the vital role that bystanders play in the bullying process. We have observed cases that although a user had been repeatedly bullied and targeted with harassing comments, but the supportive and encouraging comments of bystanders have neutralized the hurtful and negative effect of the hurtful comments. This may also go the other way around: when bystanders

support and '*like*' the harassing comments posted by bullies, they amplify the upsetting impact for victims of those comments. Therefore, follow-up research can be to study cyberbullying by zooming in on victims and investigation of public effect and role of bystanders.

6.4 Concluding Remarks

The work presented here on how to confront the phenomenon of cyberbullying exemplifies the potential added value of taking a multidisciplinary perspective.

Bullying is an old social phenomenon that is rooted in human nature. Cyberbullying is a more recent variant conducted using digital infrastructure. As argued in this thesis, the integration of social studies into a software-enhanced monitoring workflow could pave the way towards the tackling of this kind of online misbehaviour. The ideas and algorithms proposed for fulfilling this purpose can be a stepping stone for future research in this direction.

The work carried out is also a demonstration of the added value of frameworks for text categorization, sentiment mining and user profiling in applications addressing societal issues.

Finally the work reported can be viewed as a contribution to the more general societal challenge of increasing the level of cybersecurity, in particular for the younger generations of social network users. By turning the internet into a safer place for children, the chances increase that they will be able to benefit from the informational richness that it also offers.

Appendix

Online Questionnaire Used for Expert Knowledge Elicitation

This version is presented in rich-text format as an appendix of the fifth chapter of this dissertation. Some part of the information acquired from this questionnaire was not used in the manuscript and was aimed for future studies.

Dear expert,
My name is Maral Dadvar. I am a PhD student at the Human Media Interaction Group of the University of Twente working under the supervision of Professor Franciska de Jong.
My research is on Automatic Cyberbullying Detection. In the past few years I have been trying to improve cyberbullying detection algorithms by integrating the outcome of social studies on cyberbullying with technical solutions.
For the last chapter of my PhD thesis, my aim is to define and determine online behaviours and features which could be modelled in order to enable us to automatically identify the actors who are involved in cyberbullying and specifically those who might be a threat as a bully.
For this purpose, I am studying social networks, in this experiment YouTube, which are widely used for entertainment and communication. YouTube is the world's largest user-generated content site and its broad scope in terms of audience, videos, and users' comments make it a platform that is prone to bullying activities. In our current study we want to identify the users who are more likely to be a bully by analysing their history of activities and personal information.
We selected a set of features such as number of comments and age of YouTube users. By user we mean someone who is active in YouTube and uses this network for communication, entertainment and other purposes. A bully is a user with misbehaviour in this network in the form of posting threatening, vulgar or hateful comments targeted other users.
The aim of this questionnaire is to collect your opinion on selecting and weighting the features and parameters that may be informative in understanding and identifying user's behaviours in online environments.
The questionnaire has a multiple choice format. However there is also a text box at the end of each section where you can enter your general comment/advise. Please feel free to write me any comments you have in mind.
There are 21 questions which should take 10 to 15 minutes in total to complete. Please keep in mind that it is not possible to save and resume the questionnaire. It should be completed in one attempt.

You will be asked about your opinion on characteristics of a bully using the given features. For example: what is the likelihood that a bully user belongs to a certain age group?
In such a case you can answer this question using one of the values on a four-point scale: 'Unlikely', 'Less likely', 'Likely' and 'Very likely'. You can also select the 'I don't know' option. Questions marked with symbol * require an answer and others are optional.
Many thanks for taking the time to participate in this survey. Your answers are of great value to our research. For any further information or feedback please do not hesitate to contact me.

With kind regards
Maral Dadvar

1. **Cyberbullying may differ across different age categories. What is the likelihood that a BULLY user belongs to the following age categories?**

   [unlikely, less likely, likely, very likely, I do not know]

   - 13 – 16 years
   - 17 – 19 years
   - 20 – 25 year
   - 25 – 30 years
   - Above 30 years

   If you have any comments regarding this feature please specify below.

   [text box]

2. **Do you agree with the above age ranges? If you think they should be set differently, please specify below.**

   [text box]

3. **YouTube users can choose their Username to be their real name and/or surname or can choose any other aliases and combinations of symbols and words.**

   [unlikely, less likely, likely, very likely, I do not know]

   - What is the likelihood that BULLY users choose their real name/surname as their username?
   - What is the likelihood that BULLY users choose an alias as their username and mask their real name?
   - What is the likelihood that BULLY users have a profane word in their username?

If you have any comments regarding this feature please specify below.

[text box]

4. **YouTube users may vary in their activity on the network and can be active in different ways such as, uploading videos, posting comments, and like or dislike others videos and comments. What is the likelihood that a BULLY user is relatively ACTIVE in YouTube?**

   [unlikely, less likely, likely, very likely, I do not know]

5. **One of the common activities of the users is to UPLOAD videos. These videos can be home videos provided by users themselves, or videos made by others. Each user can post COMMENTS on uploaded videos as well as on other users' comments. Users can also show their opinion about others videos or comments by pressing the LIKE / DISLIKE buttons. Most of the YouTube users have a public channel, in which they upload their videos and in which their activities such as posted comments can be viewed. Other users can SUBSCRIBE to that channel and follow the activities of the owner of the channel if they find it interesting.**

   [unlikely, less likely, likely, very likely, I do not know]

   - What is the likelihood that a BULLY user is relatively ACTIVE in Upload/Post comments/Like/Dislike/Subscribe?
   - What is the likelihood that a BULLY user is relatively INACTIVE in Upload/Post comments/Like/Dislike/Subscribe?

   If you have any comments regarding this feature please specify below.

   [text box]

6. **One of the characteristics of a Youtube profile is its sign up date, which tells how long a user has been a member of the community. What is the likelihood that a BULLY user belongs to the following membership periods?**

   [unlikely, less likely, likely, very likely, I do not know]

   - Less than 1 year
   - 1 – 3 years
   - More than 3 years

   If you have any comments regarding this feature please specify below.

   [text box]

7. **Do you think the membership periods categories are appropriate in the context of cyberbullying? If you think they should be modified please specify your suggestions below.**

   [text box]

8. **The writing style of users may differ according to their characteristics and state of mind. The tendency to write short and right to-the-point statements or more lengthy sentences is one of these differences.**

   [unlikely, less likely, likely, very likely, I do not know]

   - What is the likelihood that a BULLY user writes relatively SHORT comments?
   - What is the likelihood that a BULLY user writes relatively LENGTHY comments?

   If you have any comments regarding this feature please specify below.

   [text box]

9. **Posted comments may contain strong language, such as profanities.**

   [unlikely, less likely, likely, very likely, I do not know]

   - What is the likelihood that the ratio of profanity use of a BULLY user, is relatively HIGH in the context of YouTube?
   - What is the likelihood that the ratio of profanity use of a BULLY user, is relatively LOW in the context of YouTube?

   If you have any comments regarding this feature please specify below.

   [text box]

10. **Do you have any comments on whether it can be determined if a profane word is used for insult or just as an informal language between friends?**

    [text box]

11. **Another linguistic component which can be an indicator of personality and intentions of a user is the use of pronouns in the comments. Second person pronouns such as, "You" and "Your", and first person pronouns such as, "I" and "My".**

    [unlikely, less likely, likely, very likely, I do not know]

    - What is the likelihood that the ratio of SECOND person pronouns use of a BULLY user is HIGH in the context of YouTube?

- What is the likelihood that the ratio of SECOND person pronouns use of a BULLY user is LOW in the context of YouTube?
- What is the likelihood that the ratio of FIRST person pronouns use of a BULLY user is HIGH in the context of YouTube?
- What is the likelihood that the ratio of FIRST person pronouns use of a BULLY user is LOW in the context of YouTube?

If you have any comments regarding this feature please specify below

[text box]

12. **The comments may be posted at different times during the day: morning, afternoon, evening and night. What is the likelihood that a BULLY user is engaged in an online activity at the following times of the day?**

[unlikely, less likely, likely, very likely, I do not know]

- Morning
- Afternoon
- Evening
- Night

If you have any comments regarding this feature please specify below

[text box]

13. **Education, family and security are examples of environmental and personal characteristics influential on cyberbullying which are not consistent across countries. How would you rate the following regions with respect to frequency of reported cyberbullying incidents?**

[not frequent, frequent, very frequent, I do not know]

- North America
- South America
- Europe
- Asia
- Africa
- Australia and New Zeeland

If you have any comments regarding this feature please specify below

[text box]

14. In the users' comments words may be spelled in non-standard ways, which may include wrong spellings (e.g. 'funy' instead of 'funny'), or informal short forms of the words which are used in online chats and posts (e.g. 'brb' which means 'be right back').

[unlikely, less likely, likely, very likely, I do not know]

- What is the likelihood that the ratio of wrong spellings of a BULLY user is relatively HIGH in the context of YouTube?
- What is the likelihood that the ratio of wrong spellings of a BULLY user is relatively LOW in the context of YouTube?
- What is the likelihood that the ratio of informal short forms of the words used by a BULLY user is relatively HIGH in the context of YouTube?
- What is the likelihood that the ratio of informal short forms of the words used by a BULLY user is relatively LOW in the context of YouTube?

If you have any comments regarding this feature please specify below

[text box]

15. Boys and girls vary in the way they bully and the style and word usage is different among them. To what extent do you agree with the following statements?

[disagree, partially agree, agree, totally agree, I do not know]

- Boys bully more than girls.
- Girls bully more than boys.
- Boys bully boys more often than they bully girls.
- Boys bully girls more often than they bully boys.
- Girls bully girls more often than they bully boys.
- Girls bully boys more often than they bully girls.

If you have any comments regarding this feature please specify below

[text box]

Please rank the following personal and public features of a user, based on how informative and helpful they are in the determination of personality and potential behaviour of that user.

16. User features: These features are the personal and demographic information of the users and personal parameters.

[not informative, partially informative, informative, very informative]

- Age
- Duration of user membership
- Country of residence of the user
- Gender

Please indicate if there are any other parameters which can be informative for the determination of a user's personality?

[text box]

17. **Content features: These features are derived from the content of user's comments. This category pertains to the writing structure and usage frequency of specific words.**

    [not informative, partially informative, informative, very informative]

    - Profane words in the username
    - Length of the user comments
    - Profanity use in the user's comments
    - Second person pronouns (e.g. "You","Your") used in the user's comments
    - First person pronouns (e.g. "I","My") used in the user's comments
    - Wrong spellings in the user's comments

    Please indicate if there are any other parameters which can be informative for the determination of a user's personality?

    [text box]

18. **Environment features: These features relate to the network that is the context of the research, here YouTube. These features can represent the level of activity and popularity of the user in the network.**

    [not informative, partially informative, informative, very informative]

    - Activity of a user in terms of uploading videos
    - Activity of a user in terms of posting comments
    - Activity of a user in terms of subscribing to the other users' profiles
    - Date of the comments
    - Time of the comments (morning, afternoon, night)
    - Activity of a user in terms of "like" posts (videos or comments) of the other users

- Activity of a user in terms of "dislike" posts (videos or comments) of the other users
- Percentage of comments directed towards a specific user

Please indicate if there are any other parameters which can be informative for the determination of a user's personality?

[text box]

19. **This is the last page of this questionnaire. If you have any comments/suggestions regarding features which you think might be informative to decide about the personality and intentions of a YouTube user please indicate below.**

    [text box]

20. **All your responses will be confidential. However, I would like to acknowledge your collaboration in my thesis and future publications. Do you agree that I mention your name and affiliation in the acknowledgement or would you prefer to stay anonymous?**

    [Keep me anonymous, I agree to be mentioned in the acknowledgment of the research deliverables]

21. **Your name and affiliation**

    [text box]

Bibliography

Abel, F., Henze, N., Herder, E. & Krause, D. Linkage, aggregation, alignment and enrichment of public user profiles with Mypes. *In* Proceedings of the 6th International Conference on Semantic Systems, 2010. ACM, 1-8.

Adam, C. P.Ye.Reasoning about emotions in an engaging interactive toy. *In:* Proceedings of the 8th International Conference on Autonomous agents and Multiagent Systems (AAMAS 2009), 2009. 31-32.

Allison, B. N. & Schultz, J. B. 2001. Interpersonal identity formation during early adolescence. *Adolescence,* 36**,** 509-523.

Alm, C. O., Roth, D. & Sproat, R. Emotions from text: machine learning for text-based emotion prediction. *In:* Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, 2005. Association for Computational Linguistics, 579-586.

Argamon, S., Koppel, M., Fine, J. & Shimoni, A. R. 2003. Gender, genre, and writing style in formal written texts. *Text - Interdisciplinary Journal for the Study of Discourse.,* 23**,** 321-346.

Baeza-Yates, R. & Ribeiro-Neto, B. 1999. *Modern information retrieval*, ACM press New York.

Bannink, R., Broeren, S., Van De Looij–Jansen, P. M., De Waart, F. G. & Raat, H. 2014. Cyber and Traditional Bullying Victimization as a Risk Factor for Mental Health Problems and Suicidal Ideation in Adolescents. *PloS one,* 9**,** e94026.

Bárdossy, G. & Fodor, J. 2004. *Evaluation of uncertainties and risks in geology: new mathematical approaches for their handling*, Springer.

Bass, T. 2000. Intrusion detection systems and multisensor data fusion. *Communications of the ACM,* 43**,** 99-105.

Bauer, D. S. & Koblentz, M. E. NIDX-an expert system for real-time network intrusion detection. *In:* Computer Networking Symposium, 1988., Proceedings of the, 1988. IEEE, 98-106.

Bayzick, J., Kontostathis, A. & Edwards, L. 2011. Detecting the Presence of Cyberbullying Using Computer Software.

Beran, T. & Li, Q. 2008. The relationship between cyberbullying and school bullying. *The Journal of Student Wellbeing,* 1**,** 16-33.

Besag, V. E. 1989. Bullies and victims in schools. A guide to understanding and management.

Bessière, K., Kiesler, S., Kraut, R. & Boneva, B. S. 2008. Effects of Internet use and social resources on changes in depression. *Information, Community & Society,* 11**,** 47-70.

Bj Rkqvist, K., Lagerspetz, K. M. & Kaukiainen, A. 1992. Do girls manipulate and boys fight? Developmental trends in regard to direct and indirect aggression. *Aggressive behavior,* 18**,** 117-127.

Bong, C. W., Holtby, D. W. & Ng, K. S. Fuzzy Multicriteria Decision Analysis for Measurement of Document Content Reliability. *In:* Computational Intelligence and Design (ISCID), 2012 Fifth International Symposium on, 2012. IEEE, 303-306.

Brahan, J. W., Lam, K. P., Chan, H. & Leung, W. 1998. AICAMS: artificial intelligence crime analysis and management system. *Knowledge-Based Systems,* 11**,** 355-361.

Bucchianeri, M. M., Eisenberg, M. E., Wall, M. M., Piran, N. & Neumark-Sztainer, D. 2014. Multiple types of harassment: Associations with emotional well-being and unhealthy behaviors in adolescents. *Journal of Adolescent health,* 54**,** 724-729.

Burges, C. J. 1998. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery,* 2**,** 121-167.

Campbell, M. A. 2005. Cyber bullying: An old problem in a new guise? *Australian Journal of Guidance and Counselling,* 15**,** 68-76.

Cappadocia, M. C., Craig, W. M. & Pepler, D. 2013. Cyberbullying Prevalence, Stability, and Risk Factors During Adolescence. *Canadian Journal of School Psychology,* 28**,** 171-192.

Carmagnola, F., Osborne, F., Torre, I. & White, B. Cross-Systems Identification of Users in the Social Web. *In*, 2009. 129-134.

Castillo, C., Donato, D., Gionis, A., Murdock, V. & Silvestri, F. Know your neighbors: Web spam detection using the web topology. *In:* Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, 2007. ACM, 423-430.

Cha, M., Kwak, H., Rodriguez, P., Ahn, Y.-Y. & Moon, S. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. *In:* Proceedings of the 7th ACM SIGCOMM conference on Internet measurement, 2007. ACM, 1-14.

Chen, H., Chung, W., Xu, J. J., Wang, G., Qin, Y. & Chau, M. 2004. Crime data mining: a general framework and some examples. *Computer,* 37**,** 50-56.

Chen, Y., Zhou, Y., Zhu, S. & Xu, H. Detecting Offensive Language in Social Media to Protect Adolescent Online Safety. *In:* Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom), 2012. IEEE, 71-80.

Chisholm, J. F. 2006. Cyberspace violence against girls and adolescent females. *Annals of the New York Academy of Sciences,* 1087**,** 74-89.

Cristianini, N. & Shawe-Taylor, J. 2000. *An introduction to support vector machines and other kernel-based learning methods*, Cambridge university press.

Dadvar, M., Trieschnigg, D., Ordelman, R. & De Jong, F. 2013. Improving cyberbullying detection with user context. *Advances in Information Retrieval.* Springer.

Dehue, F., Bolman, C. & Völlink, T. 2008. Cyberbullying: Youngsters' experiences and parental perception. *CyberPsychology & Behavior,* 11**,** 217-223.

Denning, D. E. 1987. An intrusion-detection model. *Software Engineering, IEEE Transactions on***,** 222-232.

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

Dilmaç, B. & Aydoğan, D. 2010. Parental attitudes as a predictor of cyber bullying among primary school children. *International Journal of Psychological and Brain Sciences,* 5**,** 649-653.

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**,** 18.

Dinakar, K., Reichart, R. & Lieberman, H. 2011. Modeling the Detection of Textual Cyberbullying. *Social Mobile Web Workshop at International Conference on Weblog and Social Media*

Farah, M. & Vanderpooten, D. A multiple criteria approach for information retrieval. *In:* String Processing and Information Retrieval, 2006. Springer, 242-254.

Fielding, A. H. & Bell, J. F. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental conservation,* 24**,** 38-49.

Figueira, J., Greco, S. & Ehrgott, M. 2005. *Multiple criteria decision analysis: state of the art surveys*, Springer.

Fleiss, J. L., Levin, B. & Paik, M. C. 2013. *Statistical methods for rates and proportions*, John Wiley & Sons.

Galton, F. 1892. *Finger prints*, Macmillan and Company.

Gordon, S. & Ford, R. 2006. On the definition and classification of cybercrime. *Journal in Computer Virology,* 2**,** 13-20.

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**,** 10-18.

Hasebrink, U., Livingstone, S. & Haddon, L. 2008. Comparing children's online opportunities and risks across Europe: Cross-national comparisons for EU Kids Online. *status: published*.

Heylen, D. Sensitive empathic agents. *In:* Proceedings of the 8th International Conference on Autonomous agents and Multiagent Systems (AAMAS 2009), 2009. 9-12.

Holfeld, B. 2014. Perceptions and attributions of bystanders to cyber bullying. *Computers in Human Behavior,* 38**,** 1-7.

Hughes, D., Rayson, P., Walkerdine, J., Lee, K., Greenwood, P., Rashid, A., May-Chahal, C. & Brennan, M. 2008. Supporting law enforcement in digital communities through natural language analysis. *Computational Forensics.* Springer.

ITU 2013. ICT Data and Statistics Division, The World in 2013, ICT Facts and Figures.

Jacobs, N. C., Völlink, T., Dehue, F. & Lechner, L. 2014. Online Pestkoppenstoppen: systematic and theory-based development of a web-based tailored intervention for adolescent cyberbully victims to combat and prevent cyberbullying. *BMC public health,* 14**,** 396.

Jiang, H. & Eastman, J. R. 2000. Application of fuzzy measures in multi-criteria evaluation in GIS. *International Journal of Geographical Information Science,* 14**,** 173-184.

Joachims, T. 1998. *Text categorization with support vector machines: Learning with many relevant features*, Springer.

Juvonen, J. & Gross, E. F. 2008. Extending the school grounds?—Bullying experiences in cyberspace. *Journal of School health,* 78**,** 496-505.

Kontostathis, A. ChatCoder: Toward the tracking and categorization of internet predators. *In*, 2009. Citeseer.

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. 2007. Electronic bullying among middle school students. *Journal of Adolescent health,* 41**,** S22-S30.

Kowalski, R. M., Limber, S. P. & Agatston, P. W. 2008. *Cyber bullying: Bullying in the digital age*, Blackwell Publishing.

Kraut, R., Kiesler, S., Boneva, B., Cummings, J., Helgeson, V. & Crawford, A. 2002. Internet paradox revisited. *Journal of social issues,* 58**,** 49-74.

Kraut, R., Patterson, M., Lundmark, V., Kiesler, S., Mukophadhyay, T. & Scherlis, W. 1998. Internet paradox: A social technology that reduces social involvement and psychological well-being? *American psychologist,* 53**,** 1017.

Krone, T. 2005. Concepts and terms. *High tech crime brief*.

Kruskal, W. & Mosteller, F. 1979. Representative sampling, II: Scientific literature, excluding statistics. *International Statistical Review/Revue Internationale de Statistique***,** 111-127.

Lamb, J., Pepler, D. J. & Craig, W. 2009. Approach to bullying and victimization. *Canadian Family Physician,* 55**,** 356-360.

Lewis, D. D. 1998. Naive (Bayes) at forty: The independence assumption in information retrieval. *Machine learning: ECML-98*. Springer.

Lieberman, H., Dinakar, K. & Jones, B. 2011. Let's Gang Up on Cyberbullying. *Computer,* 44**,** 93-96.

Machmutow, K., Perren, S., Sticca, F. & Alsaker, F. D. 2012. Peer victimisation and depressive symptoms: can specific coping strategies buffer the negative impact of cybervictimisation? *Emotional and Behavioural Difficulties,* 17**,** 403-420.

Maia, M., Almeida, J. & Almeida, V. Identifying user behavior in online social networks. *In:* Proceedings of the 1st workshop on Social network systems, 2008. ACM, 1-6.

Mcenery, T. 2001. *Corpus linguistics: An introduction*, Edinburgh University Press.

Mcghee, I., Bayzick, J., Kontostathis, A., Edwards, L., Mcbride, A. & Jakubowski, E. 2011. Learning to identify Internet sexual predation. *International Journal of Electronic Commerce,* 15**,** 103-122.

Mesch, G. S. 2001. Social relationships and Internet use among adolescents in Israel. *Social Science Quarterly,* 82**,** 329-339.

Mesch, G. S. 2009. Parental mediation, online activities, and cyberbullying. *CyberPsychology & Behavior,* 12**,** 387-393.

Nguyen, D., Demeester, T., Trieschnigg, D. & Hiemstra, D. Federated search in the wild: the combined power of over a hundred search engines. *In:* Proceedings of the 21st ACM international conference on Information and knowledge management, 2012. ACM, 1874-1878.

Nie, N. H. 2001. Sociability, interpersonal relations, and the internet reconciling conflicting findings. *American behavioral scientist,* 45**,** 420-435.

Nie, N. H. & Erbring, L. 2000. Internet and society. *Stanford Institute for the Quantitative Study of Society*.

Noble, T. 2003. Nobody left to hate. *EQ Australia,* 4**,** 8-9.

Olweus, D. 2013. School bullying: Development and some important challenges. *Annual review of clinical psychology,* 9**,** 751-780.

Pang, B. & Lee, L. 2008. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval,* 2**,** 1-135.

Park, N., Jin, B. & Annie Jin, S.-A. 2011. Effects of self-disclosure on relational intimacy in Facebook. *Computers in Human Behavior,* 27**,** 1974-1983.

Patchin, J. W. & Hinduja, S. 2006. Bullies move beyond the schoolyard a preliminary look at cyberbullying. *Youth violence and juvenile justice,* 4**,** 148-169.

Pazienza, M. T. & Tudorache, A. G. 2011. Interdisciplinary contributions to flame modeling. *AI* IA 2011: Artificial Intelligence Around Man and Beyond.* Springer.

Pendar, N. Toward spotting the pedophile telling victim from predator in text chats. *In:* Semantic Computing, 2007. ICSC 2007. International Conference on, 2007. IEEE, 235-241.

Perren, S., Corcoran, L., Cowie, H. & Dehue, F. 2012. Coping with Cyberbullying:ASystematic Literature Review. *Final'Report of'the'COST'IS'0801*.

Peter, J., Valkenburg, P. M. & Schouten, A. P. 2005. Developing a model of adolescent friendship formation on the Internet. *CyberPsychology & Behavior,* 8**,** 423-430.

Porter, D. 1996. *Internet culture*, Routledge.

Potthast, M., Stein, B. & Gerling, R. 2008. Automatic vandalism detection in Wikipedia. *Advances in Information Retrieval.* Springer.

Ratledge, E. C. & Jacoby, J. E. 1989. *Handbook on artificial intelligence and expert systems in law enforcement*, Greenwood Publishing Group Inc.

Reynolds, K., Kontostathis, A. & Edwards, L. Using machine learning to detect cyberbullying. *In:* Machine Learning and Applications and Workshops (ICMLA), 2011 10th International Conference on, 2011. IEEE, 241-244.

Rheingold, H. 1993. *The virtual community: Finding commection in a computerized world*, Addison-Wesley Longman Publishing Co., Inc.

Rivers, I. & Noret, N. 2010. 'I h8 u': findings from a five-year study of text and email bullying. *British Educational Research Journal,* 36**,** 643-671.

Ruggiero, T. E. 2000. Uses and gratifications theory in the 21st century. *Mass communication & society,* 3**,** 3-37.

Sahami, M., Dumais, S., Heckerman, D. & Horvitz, E. A Bayesian approach to filtering junk e-mail. *In:* Learning for Text Categorization: Papers from the 1998 workshop, 1998. 98-105.

Shariff, S. 2008. *Cyber-bullying: Issues and solutions for the school, the classroom and the home*, Routledge.

Shariff, S. & Patchin, J. W. 2009. *Confronting cyber-bullying*, Cambridge University Press.

Shee, D. Y. & Wang, Y.-S. 2008. Multi-criteria evaluation of the web-based e-learning system: A methodology based on learner satisfaction and its applications. *Computers & Education,* 50**,** 894-905.

Sheldon, P. 2009. I'll poke you. You'll poke me! Self-disclosure, social attraction, predictability and trust as important predictors of

Facebook relationships. *Journal of Psychosocial Research on Cyberspace,* 3.

Simanjuntak, D. A. & Ipung, H. P. Text Classification Techniques Used to Faciliate Cyber Terrorism Investigation. *In*, 2010. IEEE, 198-200.

Slonje, R. & Smith, P. K. 2008. Cyberbullying: Another main type of bullying? *Scandinavian journal of psychology,* 49**,** 147-154.

Smets, K., Goethals, B. & Verdonk, B. Automatic vandalism detection in Wikipedia: Towards a machine learning approach. *In*, 2008. 43–48.

Smith, P. K. & Ananiadou, K. 2003. The nature of school bullying and the effectiveness of school-based interventions. *Journal of Applied Psychoanalytic Studies,* 5**,** 189-209.

Smith, P. K., Mahdavi, J., Carvalho, M., Fisher, S., Russell, S. & Tippett, N. 2008. Cyberbullying: its nature and impact in secondary school pupils. *Journal of Child Psychology and Psychiatry,* 49**,** 376-385.

Smith, P. K., Morita, Y. E., Junger-Tas, J. E., Olweus, D. E., Catalano, R. F. & Slee, P. E. 1999. *The nature of school bullying: A cross-national perspective*, Taylor & Frances/Routledge.

Smith, P. K. & Sharp, S. 1994. The problem of school bullying. *School bullying: Insights and perspectives***,** 1-19.

Smith, P. K. & Shu, S. 2000. What Good Schools can Do About Bullying Findings from a Survey in English Schools After a Decade of Research and Action. *Childhood,* 7**,** 193-212.

Steijn, W. M. & Schouten, A. P. 2013. Information Sharing and Relationships on Social Networking Sites. *Cyberpsychology, Behavior, and Social Networking*.

Tan, P. N., Chen, F. & Jain, A. Information assurance: Detection of web spam attacks in social media. *In*, 2010.

Tokunaga, R. S. 2010. Following you home from school: A critical review and synthesis of research on cyberbullying victimization. *Computers in Human Behavior,* 26**,** 277-287.

Tsai, F. S. & Chan, K. L. 2007. Detecting cyber security threats in weblogs using probabilistic models. *Intelligence and Security Informatics.* Springer.

Välimäki, M. 2012. A Report on Nationally Published Guidelines in Different Countries. *International Conference of Cyberbullying COST IS0801, Paris, France*.

Valkenburg, P. M. & Peter, J. 2007. Preadolescents' and adolescents' online communication and their closeness to friends. *Developmental psychology,* 43**,** 267.

Van Der Zwaan, J., Dignum, V. & Jonker, C. 2010. Simulating peer support for victims of cyberbullying. *health promotion,* 8**,** 18.

Van Royen, K., Poels, K., Daelemans, W. & Vandebosch, H. 2014. Automatic monitoring of cyberbullying on social networking sites: From technological feasibility to desirability. *Telematics and Informatics*.

Vandebosch, H. & Van Cleemput, K. 2008. Defining cyberbullying: A qualitative research into the perceptions of youngsters. *CyberPsychology & Behavior,* 11**,** 499-503.

Vapnik, V. 1998. Statistical learning theory. 1998. Wiley, New York.

Von Solms, R. & Van Niekerk, J. 2013. From Information Security to Cyber Security. *Computers & Security*.

Willard, N. E. 2007. *Cyberbullying and cyberthreats: Responding to the challenge of online social aggression, threats, and distress*, Research Press.

Wilson, C., Boe, B., Sala, A., Puttaswamy, K. P. & Zhao, B. Y. User interactions in social networks and their implications. *In:* Proceedings of the 4th ACM European conference on Computer systems, 2009. Acm, 205-218.

Witten, I. H., Frank, E. & Hall, M. A. 2011. *Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques*, Elsevier.

Xiao, R. 2010. Corpus creation. *Handbook of Natural Language Processing (2n Revised edition)***,** 147-165.

Xu, Z., Khoshgoftaar, T. M. & Allen, E. B. 2003a. Application of fuzzy expert systems in assessing operational risk of software. *Information and Software Technology,* 45**,** 373-388.

Xu, Z. W., Khoshgoftaar, T. M. & Allen, E. B. 2003b. Application of fuzzy expert systems in assessing operational risk of software. *Information and Software Technology,* 45**,** 373-388.

Ybarra, M. L. & Mitchell, K. J. 2004. Youth engaging in online harassment: Associations with caregiver–child relationships, Internet use, and personal characteristics. *Journal of adolescence,* 27**,** 319-336.

Ye, N., Giordano, J. & Feldman, J. 2001. A process control approach to cyber attack detection. *Communications of the ACM,* 44**,** 76-82.

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 2.0 (CAW2.0) Workshop at WWW2009, Madrid, Spain* 2.

Youtube Press Statistics May 2013. www.youtube.com/t/press_statistics.

Zeviar-Geese, G. 1997. State of the Law on Cyberjursidiction and Cybercrime on the Internet, The. *Gonz. J. Int'l L.,* 1**,** 119.

Zhuang, L., Jing, F. & Zhu, X.-Y. Movie review mining and summarization. *In:* Proceedings of the 15th ACM international conference on Information and knowledge management, 2006. ACM, 43-50.

SIKS Dissertation Series (2009-2014)

Since 1998, all dissertations written by PhD students who have conducted their research under auspices of a senior research fellow of the SIKS research school are listed at www.siks.nl/dissertations.php.

2014-37 Maral Dadvar (UT), Experts and Machines United Against Cyberbullying
2014-36 Joos Buijs (TUE), Flexible Evolutionary Algorithms for Mining Structured Process Models
2014-35 Joost van Oijen (UU), Cognitive Agents in Virtual Worlds: A Middleware Design Approach
2014-34 Christina Manteli (VU), The Effect of Governance in Global Software Development: Analyzing Transactive Memory Systems
2014-33 Tesfa Tegegne Asfaw (RUN), Service Discovery in eHealth
2014-32 Naser Ayat (UVA), On Entity Resolution in Probabilistic Data
2014-31 Leo van Moergestel (UU), Agent Technology in Agile Multiparallel Manufacturing and Product Support
2014-30 Peter de Kock Berenschot (UvT), Anticipating Criminal Behaviour
2014-29 Jaap Kabbedijk (UU), Variability in Multi-Tenant Enterprise Software
2014-28 Anna Chmielowiec (VU), Decentralized k-Clique Matching
2014-27 Rui Jorge Almeida (EUR), Conditional Density Models Integrating Fuzzy and Probabilistic Representations of Uncertainty2014-26 Tim Baarslag (TUD), What to Bid and When to Stop
2014-25 Martijn Lappenschaar (RUN), New network models for the analysis of disease interaction
2014-24 Davide Ceolin (VU), Trusting Semi-structured Web Data
2014-23 Eleftherios Sidirourgos (UvA/CWI), Space Efficient Indexes for the Big Data Era
2014-22 Marieke Peeters (UU), Personalized Educational Games - Developing agent-supported scenario-based training
2014-21 Kassidy Clark (TUD), Negotiation and Monitoring in Open Environments
2014-20 Mena Habib (UT), Named Entity Extraction and Disambiguation for Informal Text: The Missing Link
2014-19 Vincius Ramos (TUE), Adaptive Hypermedia Courses: Qualitative and Quantitative Evaluation and Tool Support
2014-18 Mattijs Ghijsen (VU), Methods and Models for the Design and Study of Dynamic Agent Organizations
2014-17 Kathrin Dentler (VU), Computing healthcare quality indicators automatically: Secondary Use of Patient Data and Semantic Interoperability
2014-16 Krystyna Milian (VU), Supporting trial recruitment and design by automatically interpreting eligibility criteria
2014-15 Natalya Mogles (VU), Agent-Based Analysis and Support of Human Functioning in Complex Socio-Technical Systems: Applications in Safety and Healthcare
2014-14 Yangyang Shi (TUD), Language Models With Meta-information
2014-13 Arlette van Wissen (VU), Agent-Based Support for Behavior Change: Models and Applications in Health and Safety Domains
2014-12 Willem van Willigen (VU), Look Ma, No Hands: Aspects of Autonomous Vehicle Control
2014-11 Janneke van der Zwaan (TUD), An Empathic Virtual Buddy for Social Support
2014-10 Ivan Salvador Razo Zapata (VU), Service Value Networks

2014-09 Philip Jackson (UvT), Toward Human-Level Artificial Intelligence: Representation and Computation of Meaning in Natural Language
2014-08 Samur Araujo (TUD), Data Integration over Distributed and Heterogeneous Data Endpoints
2014-07 Arya Adriansyah (TUE), Aligning Observed and Modeled Behavior
2014-06 Damian Tamburri (VU), Supporting Networked Software Development
2014-05 Jurriaan van Reijsen (UU), Knowledge Perspectives on Advancing Dynamic Capability
2014-04 Hanna Jochmann-Mannak (UT), Websites for children: search strategies and interface design - Three studies on children's search performance and evaluation
2014-03 Sergio Raul Duarte Torres (UT), Information Retrieval for Children: Search Behavior and Solutions
2014-02 Fiona Tuliyano (RUN), Combining System Dynamics with a Domain Modeling Method
2014-01 Nicola Barile (UU), Studies in Learning Monotone Models from Data
2013-43 Marc Bron (UVA), Exploration and Contextualization through Interaction and Concepts
2013-42 Léon Planken (TUD), Algorithms for Simple Temporal Reasoning
2013-41 Jochem Liem (UVA), Supporting the Conceptual Modelling of Dynamic Systems: A Knowledge Engineering Perspective on Qualitative Reasoning
2013-40 Pim Nijssen (UM), Monte-Carlo Tree Search for Multi-Player Games
2013-39 Joop de Jong (TUD), A Method for Enterprise Ontology based Design of Enterprise Information Systems
2013-38 Eelco den Heijer (VU), Autonomous Evolutionary Art
2013-37 Dirk Börner (OUN), Ambient Learning Displays
2013-36 Than Lam Hoang (TUe), Pattern Mining in Data Streams
2013-35 Abdallah El Ali (UvA), Minimal Mobile Human Computer Interaction
2013-34 Kien Tjin-Kam-Jet (UT), Distributed Deep Web Search
2013-33 Qi Gao (TUD), User Modeling and Personalization in the Microblogging Sphere
2013-32 Kamakshi Rajagopal (OUN), Networking For Learning; The role of Networking in a Lifelong Learner's Professional Development
2013-31 Dinh Khoa Nguyen (UvT), Blueprint Model and Language for Engineering Cloud Applications
2013-30 Joyce Nakatumba (TUE), Resource-Aware Business Process Management: Analysis and Support
2013-29 Iwan de Kok (UT), Listening Heads
2013-28 Frans van der Sluis (UT), When Complexity becomes Interesting: An Inquiry into the Information eXperience
2013-27 Mohammad Huq (UT), Inference-based Framework Managing Data Provenance
2013-26 Alireza Zarghami (UT), Architectural Support for Dynamic Homecare Service Provisioning
2013-25 Agnieszka Anna Latoszek-Berendsen (UM), Intention-based Decision Support. A new way of representing and implementing clinical guidelines in a Decision Support System
2013-24 Haitham Bou Ammar (UM), Automated Transfer in Reinforcement Learning
2013-23 Patricio de Alencar Silva(UvT), Value Activity Monitoring
2013-22 Tom Claassen (RUN), Causal Discovery and Logic
2013-21 Sander Wubben (UvT), Text-to-text generation by monolingual machine translation
2013-20 Katja Hofmann (UvA), Fast and Reliable Online Learning to Rank for Information Retrieval
2013-19 Renze Steenhuizen (TUD), Coordinated Multi-Agent Planning and Scheduling

2013-18 Jeroen Janssens (UvT), Outlier Selection and One-Class Classification
2013-17 Koen Kok (VU), The PowerMatcher: Smart Coordination for the Smart Electricity Grid
2013-16 Eric Kok (UU), Exploring the practical benefits of argumentation in multi-agent deliberation
2013-15 Daniel Hennes (UM), Multiagent Learning - Dynamic Games and Applications
2013-14 Jafar Tanha (UVA), Ensemble Approaches to Semi-Supervised Learning Learning
2013-13 Mohammad Safiri(UT), Service Tailoring: User-centric creation of integrated IT-based homecare services to support independent living of elderly
2013-12 Marian Razavian(VU), Knowledge-driven Migration to Services
2013-11 Evangelos Pournaras(TUD), Multi-level Reconfigurable Self-organization in Overlay Services
2013-10 Jeewanie Jayasinghe Arachchige(UvT), A Unified Modeling Framework for Service Design.
2013-09 Fabio Gori (RUN), Metagenomic Data Analysis: Computational Methods and Applications
2013-08 Robbert-Jan Merk(VU), Making enemies: cognitive modeling for opponent agents in fighter pilot simulators
2013-07 Giel van Lankveld (UvT), Quantifying Individual Player Differences
2013-06 Romulo Goncalves(CWI), The Data Cyclotron: Juggling Data and Queries for a Data Warehouse Audience
2013-05 Dulce Pumareja (UT), Groupware Requirements Evolutions Patterns
2013-04 Chetan Yadati(TUD), Coordinating autonomous planning and scheduling
2013-03 Szymon Klarman (VU), Reasoning with Contexts in Description Logics
2013-02 Erietta Liarou (CWI), MonetDB/DataCell: Leveraging the Column-store Database Technology for Efficient and Scalable Stream Processing
2013-01 Viorel Milea (EUR), News Analytics for Financial Decision Support
2012-51 Jeroen de Jong (TUD), Heuristics in Dynamic Sceduling; a practical framework with a case study in elevator dispatching
2012-50 Steven van Kervel (TUD), Ontologogy driven Enterprise Information Systems Engineering
2012-49 Michael Kaisers (UM), Learning against Learning - Evolutionary dynamics of reinforcement learning algorithms in strategic interactions
2012-48 Jorn Bakker (TUE), Handling Abrupt Changes in Evolving Time-series Data
2012-47 Manos Tsagkias (UVA), Mining Social Media: Tracking Content and Predicting Behavior
2012-46 Simon Carter (UVA), Exploration and Exploitation of Multilingual Data for Statistical Machine Translation
2012-45 Benedikt Kratz (UvT), A Model and Language for Business-aware Transactions
2012-44 Anna Tordai (VU), On Combining Alignment Techniques
2012-42 Dominique Verpoorten (OU), Reflection Amplifiers in self-regulated Learning
2012-41 Sebastian Kelle (OU), Game Design Patterns for Learning
2012-40 Agus Gunawan (UvT), Information Access for SMEs in Indonesia
2012-39 Hassan Fatemi (UT), Risk-aware design of value and coordination networks
2012-38 Selmar Smit (VU), Parameter Tuning and Scientific Testing in Evolutionary Algorithms
2012-37 Agnes Nakakawa (RUN), A Collaboration Process for Enterprise Architecture Creation
2012-36 Denis Ssebugwawo (RUN), Analysis and Evaluation of Collaborative Modeling Processes
2012-35 Evert Haasdijk (VU), Never Too Old To Learn -- On-line Evolution of Controllers in Swarm- and Modular Robotics
2012-34 Pavol Jancura (RUN), Evolutionary analysis in PPI networks and applications

2012-33 Rory Sie (OUN), Coalitions in Cooperation Networks (COCOON)
2012-32 Wietske Visser (TUD), Qualitative multi-criteria preference representation and reasoning
2012-31 Emily Bagarukayo (RUN), A Learning by Construction Approach for Higher Order Cognitive Skills Improvement, Building Capacity and Infrastructure
2012-30 Alina Pommeranz (TUD), Designing Human-Centered Systems for Reflective Decision Making
2012-29 Almer Tigelaar (UT), Peer-to-Peer Information Retrieval
2012-28 Nancy Pascall (UvT), Engendering Technology Empowering Women
2012-27 Hayrettin Gurkok (UT), Mind the Sheep! User Experience Evaluation & Brain-Computer Interface Games
2012-26 Emile de Maat (UVA), Making Sense of Legal Text
2012-25 Silja Eckartz (UT), Managing the Business Case Development in Inter-Organizational IT Projects: A Methodology and its Application
2012-24 Laurens van der Werff (UT), Evaluation of Noisy Transcripts for Spoken Document Retrieval
2012-23 Christian Muehl (UT), Toward Affective Brain-Computer Interfaces: Exploring the Neurophysiology of Affect during Human Media Interaction
2012-22 Thijs Vis (UvT), Intelligence, politie en veiligheidsdienst: verenigbare grootheden?
2012-21 Roberto Cornacchia (TUD), Querying Sparse Matrices for Information Retrieval
2012-20 Ali Bahramisharif (RUN), Covert Visual Spatial Attention, a Robust Paradigm for Brain-Computer Interfacing
2012-19 Helen Schonenberg (TUE), What's Next? Operational Support for Business Process Execution
2012-18 Eltjo Poort (VU), Improving Solution Architecting Practices
2012-17 Amal Elgammal (UvT), Towards a Comprehensive Framework for Business Process Compliance
2012-16 Fiemke Both (VU), Helping people by understanding them - Ambient Agents supporting task execution and depression treatment
2012-15 Natalie van der Wal (VU), Social Agents. Agent-Based Modelling of Integrated Internal and Social Dynamics of Cognitive and Affective Processes.
2012-14 Evgeny Knutov(TUE), Generic Adaptation Framework for Unifying Adaptive Web-based Systems
2012-13 Suleman Shahid (UvT), Fun and Face: Exploring non-verbal expressions of emotion during playful interactions
2012-12 Kees van der Sluijs (TUE), Model Driven Design and Data Integration in Semantic Web Information Systems
2012-11 J.C.B. Rantham Prabhakara (TUE), Process Mining in the Large: Preprocessing, Discovery, and Diagnostics
2012-10 David Smits (TUE), Towards a Generic Distributed Adaptive Hypermedia Environment
2012-09 Ricardo Neisse (UT), Trust and Privacy Management Support for Context-Aware Service Platforms
2012-08 Gerben de Vries (UVA), Kernel Methods for Vessel Trajectories
2012-07 Rianne van Lambalgen (VU), When the Going Gets Tough: Exploring Agent-based Models of Human Performance under Demanding Conditions
2012-06 Wolfgang Reinhardt (OU), Awareness Support for Knowledge Workers in Research Networks
2012-05 Marijn Plomp (UU), Maturing Interorganisational Information Systems
2012-04 Jurriaan Souer (UU), Development of Content Management System-based Web Applications
2012-03 Adam Vanya (VU), Supporting Architecture Evolution by Mining Software Repositories

2012-02 Muhammad Umair(VU), Adaptivity, emotion, and Rationality in Human and Ambient Agent Models

2012-01 Terry Kakeeto (UvT), Relationship Marketing for SMEs in Uganda

2011-49 Andreea Niculescu (UT), Conversational interfaces for task-oriented spoken dialogues: design aspects influencing interaction quality

2011-48 Mark Ter Maat (UT), Response Selection and Turn-taking for a Sensitive Artificial Listening Agent

2011-47 Azizi Bin Ab Aziz(VU), Exploring Computational Models for Intelligent Support of Persons with Depression

2011-46 Beibei Hu (TUD), Towards Contextualized Information Delivery: A Rule-based Architecture for the Domain of Mobile Police Work

2011-45 Herman Stehouwer (UvT), Statistical Language Models for Alternative Sequence Selection

2011-44 Boris Reuderink (UT), Robust Brain-Computer Interfaces

2011-43 Henk van der Schuur (UU), Process Improvement through Software Operation Knowledge

2011-42 Michal Sindlar (UU), Explaining Behavior through Mental State Attribution

2011-41 Luan Ibraimi (UT), Cryptographically Enforced Distributed Data Access Control

2011-40 Viktor Clerc (VU), Architectural Knowledge Management in Global Software Development

2011-39 Joost Westra (UU), Organizing Adaptation using Agents in Serious Games

2011-38 Nyree Lemmens (UM), Bee-inspired Distributed Optimization

2011-37 Adriana Burlutiu (RUN), Machine Learning for Pairwise Data, Applications for Preference Learning and Supervised Network Inference

2011-36 Erik van der Spek (UU), Experiments in serious game design: a cognitive approach

2011-35 Maaike Harbers (UU), Explaining Agent Behavior in Virtual Training

2011-34 Paolo Turrini (UU), Strategic Reasoning in Interdependence: Logical and Game-theoretical Investigations

2011-33 Tom van der Weide (UU), Arguing to Motivate Decisions

2011-32 Nees-Jan van Eck (EUR), Methodological Advances in Bibliometric Mapping of Science

2011-31 Ludo Waltman (EUR), Computational and Game-Theoretic Approaches for Modeling Bounded Rationality

2011-30 Egon van den Broek (UT), Affective Signal Processing (ASP): Unraveling the mystery of emotions

2011-29 Faisal Kamiran (TUE), Discrimination-aware Classification

2011-28 Rianne Kaptein(UVA), Effective Focused Retrieval by Exploiting Query Context and Document Structure

2011-27 Aniel Bhulai (VU), Dynamic website optimization through autonomous management of design patterns

2011-26 Matthijs Aart Pontier (VU), Virtual Agents for Human Communication - Emotion Regulation and Involvement-Distance Trade-Offs in Embodied Conversational Agents and Robots

2011-25 Syed Waqar ul Qounain Jaffry (VU)), Analysis and Validation of Models for Trust Dynamics

2011-24 Herwin van Welbergen (UT), Behavior Generation for Interpersonal Coordination with Virtual Humans On Specifying, Scheduling and Realizing Multimodal Virtual Human Behavior

2011-23 Wouter Weerkamp (UVA), Finding People and their Utterances in Social Media

2011-22 Junte Zhang (UVA), System Evaluation of Archival Description and Access

2011-21 Linda Terlouw (TUD), Modularization and Specification of Service-Oriented Systems

2011-20 Qing Gu (VU), Guiding service-oriented software engineering - A view-based approach

2011-19 Ellen Rusman (OU), The Mind ' s Eye on Personal Profiles
2011-18 Mark Ponsen (UM), Strategic Decision-Making in complex games
2011-17 Jiyin He (UVA), Exploring Topic Structure: Coherence, Diversity and Relatedness
2011-16 Maarten Schadd (UM), Selective Search in Games of Different Complexity
2011-15 Marijn Koolen (UvA), The Meaning of Structure: the Value of Link Evidence for Information Retrieval
2011-14 Milan Lovric (EUR), Behavioral Finance and Agent-Based Artificial Markets
2011-13 Xiaoyu Mao (UvT), Airport under Control. Multiagent Scheduling for Airport Ground Handling
2011-12 Carmen Bratosin (TUE), Grid Architecture for Distributed Process Mining
2011-11 Dhaval Vyas (UT), Designing for Awareness: An Experience-focused HCI Perspective
2011-10 Bart Bogaert (UvT), Cloud Content Contention
2011-09 Tim de Jong (OU), Contextualised Mobile Media for Learning
2011-08 Nieske Vergunst (UU), BDI-based Generation of Robust Task-Oriented Dialogues
2011-07 Yujia Cao (UT), Multimodal Information Presentation for High Load Human Computer Interaction
2011-06 Yiwen Wang (TUE), Semantically-Enhanced Recommendations in Cultural Heritage
2011-05 Base van der Raadt (VU), Enterprise Architecture Coming of Age - Increasing the Performance of an Emerging Discipline.
2011-04 Hado van Hasselt (UU), Insights in Reinforcement Learning; Formal analysis and empirical evaluation of temporal-difference
2011-03 Jan Martijn van der Werf (TUE), Compositional Design and Verification of Component-Based Information Systems
2011-02 Nick Tinnemeier(UU), Organizing Agent Organizations. Syntax and Operational Semantics of an Organization-Oriented Programming Language
2011-01 Botond Cseke (RUN), Variational Algorithms for Bayesian Inference in Latent Gaussian Models
2010-53 Edgar Meij (UVA), Combining Concepts and Language Models for Information Access
2010-52 Peter-Paul van Maanen (VU), Adaptive Support for Human-Computer Teams: Exploring the Use of Cognitive Models of Trust and Attention
2010-51 Alia Khairia Amin (CWI), Understanding and supporting information seeking tasks in multiple sources
2010-50 Bouke Huurnink (UVA), Search in Audiovisual Broadcast Archives
2010-49 Jahn-Takeshi Saito (UM), Solving difficult game positions
2010-47 Chen Li (UT), Mining Process Model Variants: Challenges, Techniques, Examples
2010-46 Vincent Pijpers (VU), e3alignment: Exploring Inter-Organizational Business-ICT Alignment
2010-45 Vasilios Andrikopoulos (UvT), A theory and model for the evolution of software services
2010-44 Pieter Bellekens (TUE), An Approach towards Context-sensitive and User-adapted Access to Heterogeneous Data Sources, Illustrated in the Television Domain
2010-43 Peter van Kranenburg (UU), A Computational Approach to Content-Based Retrieval of Folk Song Melodies
2010-42 Sybren de Kinderen (VU), Needs-driven service bundling in a multi-supplier setting - the computational e3-service approach
2010-41 Guillaume Chaslot (UM), Monte-Carlo Tree Search
2010-40 Mark van Assem (VU), Converting and Integrating Vocabularies for the Semantic Web

2010-39 Ghazanfar Farooq Siddiqui (VU), Integrative modeling of emotions in virtual agents

2010-38 Dirk Fahland (TUE), From Scenarios to components

2010-37 Niels Lohmann (TUE), Correctness of services and their composition

2010-36 Jose Janssen (OU), Paving the Way for Lifelong Learning; Facilitating competence development through a learning path specification

2010-35 Dolf Trieschnigg (UT), Proof of Concept: Concept-based Biomedical Information Retrieval

2010-34 Teduh Dirgahayu (UT), Interaction Design in Service Compositions

2010-33 Robin Aly (UT), Modeling Representation Uncertainty in Concept-Based Multimedia Retrieval

2010-32 Marcel Hiel (UvT), An Adaptive Service Oriented Architecture: Automatically solving Interoperability Problems

2010-31 Victor de Boer (UVA), Ontology Enrichment from Heterogeneous Sources on the Web

2010-30 Marieke van Erp (UvT), Accessing Natural History - Discoveries in data cleaning, structuring, and retrieval

2010-29 Stratos Idreos(CWI), Database Cracking: Towards Auto-tuning Database Kernels

2010-28 Arne Koopman (UU), Characteristic Relational Patterns

2010-27 Marten Voulon (UL), Automatisch contracteren

2010-26 Ying Zhang (CWI), XRPC: Efficient Distributed Query Processing on Heterogeneous XQuery Engines

2010-25 Zulfiqar Ali Memon (VU), Modelling Human-Awareness for Ambient Agents: A Human Mindreading Perspective

2010-24 Dmytro Tykhonov , Designing Generic and Efficient Negotiation Strategies

2010-23 Bas Steunebrink (UU), The Logical Structure of Emotions

2010-22 Michiel Hildebrand (CWI), End-user Support for Access to\\ Heterogeneous Linked Data

2010-21 Harold van Heerde (UT), Privacy-aware data management by means of data degradation

2010-20 Ivo Swartjes (UT), Whose Story Is It Anyway? How Improv Informs Agency and Authorship of Emergent Narrative

2010-19 Henriette Cramer (UvA), People's Responses to Autonomous and Adaptive Systems

2010-18 Charlotte Gerritsen (VU), Caught in the Act: Investigating Crime by Agent-Based Simulation

2010-17 Spyros Kotoulas (VU), Scalable Discovery of Networked Resources: Algorithms, Infrastructure, Applications

2010-16 Sicco Verwer (TUD), Efficient Identification of Timed Automata, theory and practice

2010-15 Lianne Bodenstaff (UT), Managing Dependency Relations in Inter-Organizational Models

2010-14 Sander van Splunter (VU), Automated Web Service Reconfiguration

2010-13 Gianluigi Folino (RUN), High Performance Data Mining using Bio-inspired techniques

2010-12 Susan van den Braak (UU), Sensemaking software for crime analysis

2010-11 Adriaan Ter Mors (TUD), The world according to MARP: Multi-Agent Route Planning

2010-10 Rebecca Ong (UL), Mobile Communication and Protection of Children

2010-09 Hugo Kielman (UL), A Politiele gegevensverwerking en Privacy, Naar een effectieve waarborging

2010-08 Krzysztof Siewicz (UL), Towards an Improved Regulatory Framework of Free Software. Protecting user freedoms in a world of software communities and eGovernments

2010-07 Wim Fikkert (UT), Gesture interaction at a Distance

2010-06 Sander Bakkes (UvT), Rapid Adaptation of Video Game AI
2010-05 Claudia Hauff (UT), Predicting the Effectiveness of Queries and Retrieval Systems
2010-04 Olga Kulyk (UT), Do You Know What I Know? Situational Awareness of Co-located Teams in Multidisplay Environments
2010-03 Joost Geurts (CWI), A Document Engineering Model and Processing Framework for Multimedia documents
2010-02 Ingo Wassink (UT), Work flows in Life Science
2010-01 Matthijs van Leeuwen (UU), Patterns that Matter
2009-46 Loredana Afanasiev (UvA), Querying XML: Benchmarks and Recursion
2009-45 Jilles Vreeken (UU), Making Pattern Mining Useful
2009-44 Roberto Santana Tapia (UT), Assessing Business-IT Alignment in Networked Organizations
2009-43 Virginia Nunes Leal Franqueira (UT), Finding Multi-step Attacks in Computer Networks using Heuristic Search and Mobile Ambients
2009-42 Toine Bogers (UvT), Recommender Systems for Social Bookmarking
2009-41 Igor Berezhnyy (UvT), Digital Analysis of Paintings
2009-40 Stephan Raaijmakers (UvT), Multinomial Language Learning: Investigations into the Geometry of Language
2009-39 Christian Stahl (TUE, Humboldt-Universitaet zu Berlin), Service Substitution -- A Behavioral Approach Based on Petri Nets
2009-38 Riina Vuorikari (OU), Tags and self-organisation: a metadata ecology for learning resources in a multilingual context
2009-37 Hendrik Drachsler (OUN), Navigation Support for Learners in Informal Learning Networks
2009-36 Marco Kalz (OUN), Placement Support for Learners in Learning Networks
2009-35 Wouter Koelewijn (UL), Privacy en Politiegegevens; Over geautomatiseerde normatieve informatie-uitwisseling
2009-34 Inge van de Weerd (UU), Advancing in Software Product Management: An Incremental Method Engineering Approach
2009-33 Khiet Truong (UT), How Does Real Affect Affect Affect Recognition In Speech?
2009-32 Rik Farenhorst (VU) and Remco de Boer (VU), Architectural Knowledge Management: Supporting Architects and Auditors
2009-31 Sofiya Katrenko (UVA), A Closer Look at Learning Relations from Text
2009-30 Marcin Zukowski (CWI), Balancing vectorized query execution with bandwidth-optimized storage
2009-29 Stanislav Pokraev (UT), Model-Driven Semantic Integration of Service-Oriented Applications
2009-28 Sander Evers (UT), Sensor Data Management with Probabilistic Models
2009-27 Christian Glahn (OU), Contextual Support of social Engagement and Reflection on the Web
2009-26 Fernando Koch (UU), An Agent-Based Model for the Development of Intelligent Mobile Services
2009-25 Alex van Ballegooij (CWI), "RAM: Array Database Management through Relational Mapping"
2009-24 Annerieke Heuvelink (VUA), Cognitive Models for Training Simulations
2009-23 Peter Hofgesang (VU), Modelling Web Usage in a Changing Environment
2009-22 Pavel Serdyukov (UT), Search For Expertise: Going beyond direct evidence
2009-21 Stijn Vanderlooy (UM), Ranking and Reliable Classification
2009-20 Bob van der Vecht (UU), Adjustable Autonomy: Controling Influences on Decision Making

2009-19 Valentin Robu (CWI), Modeling Preferences, Strategic Reasoning and Collaboration in Agent-Mediated Electronic Markets

2009-18 Fabian Groffen (CWI), Armada, An Evolving Database System

2009-17 Laurens van der Maaten (UvT), Feature Extraction from Visual Data

2009-16 Fritz Reul (UvT), New Architectures in Computer Chess

2009-15 Rinke Hoekstra (UVA), Ontology Representation - Design Patterns and Ontologies that Make Sense

2009-14 Maksym Korotkiy (VU), From ontology-enabled services to service-enabled ontologies (making ontologies work in e-science with ONTO-SOA)

2009-13 Steven de Jong (UM), Fairness in Multi-Agent Systems

2009-12 Peter Massuthe (TUE, Humboldt-Universitaet zu Berlin), Operating Guidelines for Services

2009-11 Alexander Boer (UVA), Legal Theory, Sources of Law & the Semantic Web

2009-10 Jan Wielemaker (UVA), Logic programming for knowledge-intensive interactive applications

2009-09 Benjamin Kanagwa (RUN), Design, Discovery and Construction of Service-oriented Systems

2009-08 Volker Nannen (VU), Evolutionary Agent-Based Policy Analysis in Dynamic Environments

2009-07 Ronald Poppe (UT), Discriminative Vision-Based Recovery and Recognition of Human Motion

2009-06 Muhammad Subianto (UU), Understanding Classification

2009-05 Sietse Overbeek (RUN), Bridging Supply and Demand for Knowledge Intensive Tasks - Based on Knowledge, Cognition, and Quality

2009-04 Josephine Nabukenya (RUN), Improving the Quality of Organisational Policy Making using Collaboration Engineering

2009-03 Hans Stol (UvT), A Framework for Evidence-based Policy Making Using IT

2009-02 Willem Robert van Hage (VU), Evaluating Ontology-Alignment Techniques

2009-01 Rasa Jurgelenaite (RUN), Symmetric Causal Independence Models

EXPERTS AND MACHINES UNITED AGAINST CYBERBULLYING

Maral Dadvar

ISBN: 978-90-365-3739-1
ISSN: 1381-3617.14-323
DOI: 10.3990/1.9789036537391