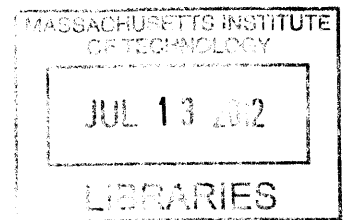


**Ruminati**  
**Modeling the Detection of Textual Cyber-bullying**  
**By**  
**Karthik Dinakar**

**ARCHIVES**



Submitted to the  
Program in Media Arts and Sciences,  
School of Architecture and Planning,  
in partial fulfillment of the requirements of the degree of  
Master of Science  
In  
Media Arts & Sciences  
at the Massachusetts Institute of Technology

June 2012

© 2012 Massachusetts Institute of Technology. All rights reserved.

Author: 

Program in Media Arts and Sciences  
May 11<sup>th</sup>, 2012

Certified By: 

Dr. Henry Lieberman  
Principal Research Scientist, MIT Media Lab

Accepted by: \_\_\_\_\_

Prof. Mitchel Resnick  
Academic Head  
Program in Media Arts and Sciences



# Ruminati

Modeling the Detection of Textual Cyber-bullying

By

Karthik Dinakar

Submitted to the  
Program in Media Arts and Sciences,  
School of Architecture and Planning,  
in partial fulfillment of the requirements of the degree of  
Master of Science  
at the Massachusetts Institute of Technology

June 2012

© 2012 Massachusetts Institute of Technology. All rights reserved.

## Abstract

The scourge of cyber-bullying has received widespread attention at all levels of society including parents, educators, adolescents, social scientists, psychiatrists and policy makers at the highest echelons of power. Cyber-bullying and its complex intermingling with traditional bullying has been shown to have a deeply negative impact on both the bully as well as the victim. We hypothesize that tackling cyber-bullying entails two parts – detection and user-interaction strategies for effective mitigation. In this thesis, we investigate the problem of detecting textual cyber-bullying. A companion thesis by Birago Jones will investigate use-interaction strategies.

In this thesis, we explore mechanisms to tackle the problem of textual cyber-bullying using computational empathy - a combination of detection and intervention techniques informed by scoping the social parameters that underlie the problem as well as a sociolinguistic treatment of the underlying socially mediated communication on the web. We begin by presenting a qualitative analysis of textual cyber-bullying based on data gathered from two major social networking websites and decompose the problem of detection into sub-problems. I then present Ruminati - a society of models of models involving supervised learning, commonsense reasoning and probabilistic topic modeling to tackle each sub-problem.

---

Thesis Supervisor: Dr. Henry Lieberman  
Title: Principal Research Scientist, MIT Media Lab



# Ruminati

Modeling the Detection of Textual Cyber-bullying

By

Karthik Dinakar

The following person served as a reader for this thesis:

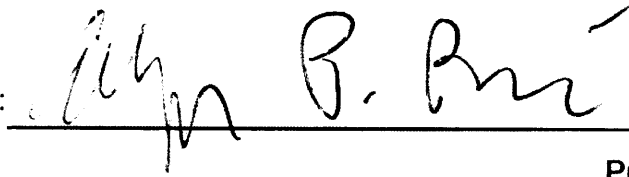
Thesis Reader: *Rosalind V. Picard*  
Prof. Rosalind Picard  
Professor  
Program in Media Arts and Sciences

# Ruminati

Modeling the Detection of Textual Cyber-bullying  
By  
Karthik Dinakar

The following person served as a reader for this thesis:

Thesis Reader:

A handwritten signature in black ink, appearing to read "Carolyn B. Rose", written over a horizontal line.

Prof. Carolyn Rose  
Associate Professor  
Carnegie Mellon University

# Contents

<b>1</b>	<b>INTRODUCTION</b>	<b>14</b>
1.1	Motivation . . . . .	16
1.2	Hypothesis . . . . .	18
1.3	Contributions . . . . .	19
1.4	Organization . . . . .	22
<b>2</b>	<b>BACKGROUND, RELATED WORK &amp; SCOPE</b>	<b>25</b>
2.1	Sociology and psychiatry . . . . .	26
2.2	Computational linguistics & interaction analysis . . . . .	28
2.3	Related applications . . . . .	28
2.4	Problem Definition, Decomposition & Scope . . . . .	29
2.5	Disclaimer . . . . .	32
2.6	Alternate framing of the problem: Discourse Analysis . . . . .	33
<b>3</b>	<b>MODELING EXPLICIT FORMS OF ABUSE</b>	<b>34</b>
3.1	Granularity in the arc of cyber-bullying . . . . .	36
3.2	Corpora . . . . .	36
3.3	Statistical Supervised Machine Learning Techniques . . . . .	40
3.4	Feature Space Design . . . . .	42
3.5	Evaluation & Discussion . . . . .	44
<b>4</b>	<b>COMMONSENSE REASONING: MODELING INDIRECT FORMS OF ABUSE</b>	<b>49</b>
4.1	The Open Mind Commonsense knowledge base . . . . .	50
4.2	The AnalogySpace inference technique . . . . .	53
4.3	The Blending knowledge combination technique . . . . .	54
4.4	The BullySpace knowledge base . . . . .	55

4.5	Cosine similarity of extracted and canonical concepts . . . . .	57
4.6	Evaluation of commonsense reasoning models . . . . .	60
4.7	Error Analysis of the commonsense reasoning model . . . . .	62
<b>5</b>	<b>TEENAGE DRAMA: PROBABILISTIC TOPIC MODELING</b>	<b>64</b>
5.1	Teenage drama and its thematic distributions . . . . .	65
5.2	The MTV teenage stories corpus . . . . .	66
5.3	Extracting high-level themes . . . . .	69
5.4	Thematic breakdown of a story . . . . .	75
5.5	Theme-based story matching . . . . .	76
5.6	Evaluation . . . . .	77
<b>6</b>	<b>CONCLUSION</b>	<b>81</b>
6.1	Comparison: Three pillars of a robust AI architecture . . . . .	82
6.2	Simplifying assumptions and limitations . . . . .	84
6.3	Contributions . . . . .	85
<b>7</b>	<b>FUTURE WORK</b>	<b>87</b>
7.1	A society of models working in tandem to help moderators of so- cial networking websites . . . . .	88
7.2	Social nonparametric and parametric machine learning . . . . .	89
7.3	An ontology of autism in developing children . . . . .	89
7.4	Empathetic Computing . . . . .	90
	<b>REFERENCES</b>	<b>91</b>



## Listing of figures

1.2.1 Hypothesis. . . . .	19
1.3.1 Interaction analysis versus statistical machine learning. . . . .	20
2 Scope . . . . .	26
2 Problem Decomposition . . . . .	36
3.2.1 YouTube Corpora Distribution . . . . .	39
3.5.1 Evaluation of supervised learning models . . . . .	45
4.5.1 Cosine Similarity of Concepts 1 . . . . .	58
4.5.2 Cosine Similarity of Concepts 2 . . . . .	59
4.6.1 Evaluation of the Commonsense reasoning model . . . . .	61
5.2.1 Sample Thinline Story . . . . .	67
5.3.1 Extracting High Level Themes . . . . .	71
5.3.2 Thematic breakdown of a story . . . . .	73
5.3.3 Distribution of themes . . . . .	74
5.3.4 Co-occurring themes . . . . .	75
5.6.1 Evaluation of the story-matching model . . . . .	79

*TO MOM, DAD, SHAMBHAVI & MY GRANDPARENTS*

## Acknowledgments

It was a bright, hot July Saturday, as I zoomed in on MIT's dome from a skywalk on the Prudential Towers building in Boston. I had arrived in the United States only two days earlier. Though the school I was most interested in was in the city of Pittsburgh, my mother and I decided to drop into Boston first, because of my many cousins who live in the area. Since we were to spend almost two weeks in Boston, my cousin had drawn up a most exhaustive itinerary, from learning about the first Pilgrims to visiting MGH and the other two '*gems*' as he called it. No sooner did we get off the Prudential than we pulled up at the Killian Court to get a glimpse of MIT.

As we entered this building, I confessed myself feeling slightly intimidated. There was something really extraordinary about the atmosphere in this campus, some giant magnetic field of fierce intelligence and alpha waves. '*Why didn't you apply to MIT?*' asked my mom as she saw me ogling at a bunch of notice boards. I really didn't know the answer as to why I never really even considered applying to MIT. Back at Yahoo Research in Bangalore, three brainy alums from Carnegie Mellon had drilled it into me that Carnegie Mellon was indeed the best in computer science. '*Maybe I'll come here one day*', I said to my mom as we talked past the math department. I don't know why I said it, but I felt a strange sense of foreboding, a sense that there was something that was resonating that I could not place my finger on.

After two years at my beloved Carnegie Mellon University, I found myself looking at an email from one Linda Peterson, informing me that I had been admitted to Media Laboratory at MIT under Dr. Henry Lieberman. No sooner had I finished

---

reading those short two lines than my mind raced back to that feeling I felt during my first visit to MIT two years ago. And my life has and will never be the same again.

Over the past year and a half, I have frequently been reminded of what an absolutely extraordinary and magical place MIT really is. I have learned more during these past two years than at any other point during my life. I have learned certain life-altering skills that one cannot learn too easily. I have learned the value of having loving parents and a loving sister, as well the joy of enduring, deep friendships. I have learned more about myself. I have learned that the world is strange aggregation of probability distributions. I've seen machine learning and natural language processing through the lens of autism. I've learned how to deal with stress, how to manage expectations, and how to pull oneself out of difficult circumstances.

I am very fiercely proud to be under the tutelage of three extraordinary human beings that are the best researchers in their area in the whole world. I am very fond of each of them. To my advisor **Dr. Henry Lieberman**, I am most grateful, for I would have never set foot on the MIT campus without him. He has helped me refine many an idea, helped me in all of my publications, and been a constant source of realism, balancing my propensity for over exerting myself with a nuanced view of my work.

To work with **Professor Rosalind Picard** has been an ennobling experience. I am inspired by the way she lives, by her rapt dedication to her work, and most of all, for her extraordinary insights into complex and nebulous human phenomenon from autism to visceral affective states. I will never ever forget her class on autism, that will probably be the most illuminating set of insights that I have ever had great fortune of studying.

I will never forget the book on language that **Professor Carolyn Rose** lent me when I was at Carnegie Mellon. Not only did that book probably change my brain, but in accordance with legend has helped me in more ways than one. Her views on Applied Machine Learning have made such an impact on me that it has set in stone a personal philosophy of 'if you can't use it to build something, it's probably not worth doing'. I admire and respect her work immensely, and will forever be grateful for all the mentoring she has given me both during my time at Carnegie Mellon and beyond as well.

I am very privileged to have **Professor Mitchel Resnick** support me in my every

aspect of my stay at the lab - he has shown my great care and affection that can only be compared to the enigmatic Albus Dumbledore.

I love my mom, dad and my sister, for they mean the world to me. If I ever looked into the *mirror of Erised*, I would be looking at a picture of myself with my parents and my sister. They have given me everything I have ever asked for, and they have taught me that life is not Woody Allen country - that your family and faith are the most important, and whatever time you get beyond it, you use it to gain knowledge and contribute to the world.

I have some of the most precious human beings as my friends. I am enormously grateful to **Matthew Thoman** for being a wonderful buddy. I want to thank **Shantnu Chandel, Farrah Joon, Layla Tabatabaei, Birago Jones, Karen Brennan, Ehsan Hoque, Julia Ma** and Dustin Arthur Smith.

I'm very fond of and grateful to the wonderful folks who make the Media Lab go around. To **Linda Peterson**, who has shown me enormous affection, to **Felice Garnder** for her constant encouragement, to **Ellen Hoffman** and **Stacie Slotnick** who never cease to make me smile, and to each of the other staff in the lab who have been so very kind to me - I can, quite literally, call the Media Lab a home of my own.

The rich legacy and history of MIT inspires me. My mentors prod me forward. My well-wishers care deeply about me. I realize that very few are given all these things. To him for whom much is given, much is asked and much is expected. I will leave no stone unturned to do my bit, and do it to the best of my abilities.

*It's nice to be important, but it's more important to be nice.*

Unknown

# 1

## Introduction

**W**HEN thirteen year old Megan Meier persuaded her mother to let her add the handsome Josh Evans on her *myspace* webpage, she seemed to be head over heels for Josh who had added her as a friend on his own webpage. When Tina Meier asked her daughter if she knew Josh, Megan replied that she had never met Josh before, but that he was very good-looking and wanted to add him as a friend. Tina relented, and Megan added Josh as a friend. Just days before her fourteenth birthday, Megan received messages from Josh telling her that she was

not a nice person and that all her friends hate her. Days before her fourteenth birthday, Megan sobbingly told her parents that some of her friends were posting nasty things about her appearance and her weight on the internet. While both her parents tried to convince Megan that she would be fine, they had little idea of the horror that would haunt them the rest of their lives. Just before her actual birthday, Tina Meier found her daughter dead. Megan had hung herself in her bedroom [23]. While Megan's struggles with weight and self-esteem were well-known to both her parents, they were unaware that the entity called Josh Evans, whose final messages to Megan might have triggered the tragedy that followed was in fact a fake account. That fake account was that of Lori Drew, the mother of a friend that lived down the street that Megan had a falling out with. Megan's parents believe that their daughter's death can be attributed to cyber-bullying.

The trial of Lori Drew - the arguments presented by the prosecution, her eventual acquittal and the widespread press attention the story has received all highlight the complexity of this problem. Some have argued that Megan's troubles with anxiety, body-image and depression underlines the need to identify and help young people in distress. Some have called for embedding empathy in the school curriculum, with training for both educators and parents alike. Others have clamored for tougher laws to deter bullies on the web.

While Megan's body-image and self-esteem troubles were definitely factors in this sad story, there have been other cases of teenagers ending their own lives where factors have ranged from homophobic and racial hatred to a host of teenage related dramas all in the form of cyber-bullying. While a concoction of factors might be responsible for Megan's death, there can be no denying that cyber-bullying was a contributing factor in that lethal concoction.

The scourge of bullying in children and adolescents is not a recent phenomenon. Social scientists, psychiatrists and educators have been studying this phenomenon for dozens of years, documenting its prevalence, investigating its harmful developmental affects and evaluating school programs designed to mitigate its effects. What is new however, is the precarious dimension that bullying takes when it is done using digital communication, particularly on social networks. Cyber-bullying is as much a threat to the viability of online social networks for youth today as spam once was to email in the early days of the Internet. While true solutions need to address the distribution of issues involved in this complex problem, few have considered innovative design of social network software as a tool for mitigating this problem. In this thesis, we attempt to make the case that technology can play a vital role in mitigating cyber-bullying.

We begin by providing the motivation behind this work before making a hypothesis on how artificial intelligence and human-computer interaction can help in mitigation. We then provide an overview of the technical approaches while highlighting the main contributions of this work. We conclude by introducing the reader to each of the subsequent chapters and the organization of this thesis.

## 1.1 MOTIVATION

We discuss the motivation behind this thesis in three parts, namely the grave nature of the menace of cyber-bullying, the algorithmic challenges in the fields of machine learning and natural language processing with respect to this problem, as well the dearth of technical solutions to tackle this problem.

Firstly, it is worth removing our academic hats for a while and observing the problem of bullying and cyber-bullying in particular from a purely humanistic per-



spective. No amount of consolation or time can fully heal the broken hearts of a parent whose child's life has either been tragically ended or has been marred because of cyber-bullying as contributing factor. Any damage done, either mentally or physically or any loss of life due to this phenomenon is vexingly mindless and is a scar upon the face of society at large. While the overwhelming portion of this thesis is technical in nature, we wish to underline the severity and gravity of this problem and do not intend to be flippant about it. One of the main motivating factors behind this work is the realization that we as computer scientists can contribute in a meaningful way towards alleviating a very serious social problem, and the dearth of work in the field of computer science in this area affords a unique opportunity to make a seminal contribution.

Secondly, the computational detection of cyber-bullying raises unique questions on the many classes of algorithms in the fields of machine learning and natural language processing with respect to the phenomenon of social interaction analysis, especially in the online domain. While statistical supervised and unsupervised methods emphasize generalization, abstraction and an exploitation of patterns in the data, the phenomena of social interaction analysis and sociolinguistics places an emphasis on specificity, uniqueness, presence of affect, ascription to community and the personality of the individuals and their use of language. It would seem that the fields of machine learning and natural language processing have a significant chasm and dissonance with the fields of social interaction analysis and sociolinguistics. Another motivating factor behind this thesis is to investigate how one might plug in the chasm between these seemingly disparate fields - that an effective parameterization approach to exert the full power and weight of statistical machine learning and natural language processing involves

the drawing of relevant parameters from the fields of sociology, psychiatry and sociolinguistics, all three of which have been studying the phenomenon of bullying and meanness for decades.

Thirdly, we found it both surprising and unsettling to find a complete dearth of work in the fields of computational linguistics and human-computer interaction specific to cyber-bullying. While most popular online social networks seem to possess an unrelenting focus on growing the size of their networks, very little if any effort has been expended on addressing cyber-bullying. Even the user-interface interaction paradigms of these sites seem to have been designed to increase the number of users on their websites, with virtually no thought given to developing a reflective user interaction paradigm. Another goal behind this thesis to encourage a new model of community interaction on social networks, which we describe in the next section.

## 1.2 HYPOTHESIS

After an investigation into the state of the art computational algorithms and user-interaction paradigms specific to the phenomenon of social interaction analysis on online social networking websites, we frame a hypothesis as follows: an effective mitigation of the problem of cyber-bullying involves two integral components - computational detection of the phenomenon and the design of reflective user interaction strategies working in tandem. This thesis is primarily concerned with the computational detection part with a sprinkling of reflective user interaction aspects. For an in-depth treatment and elucidation of reflective user-interaction paradigms, the reader is encouraged to refer to a companion thesis titled '*Reflective UI: Facilitating the Mitigation of Cyberbullying*' by Birago Koraya Jones.

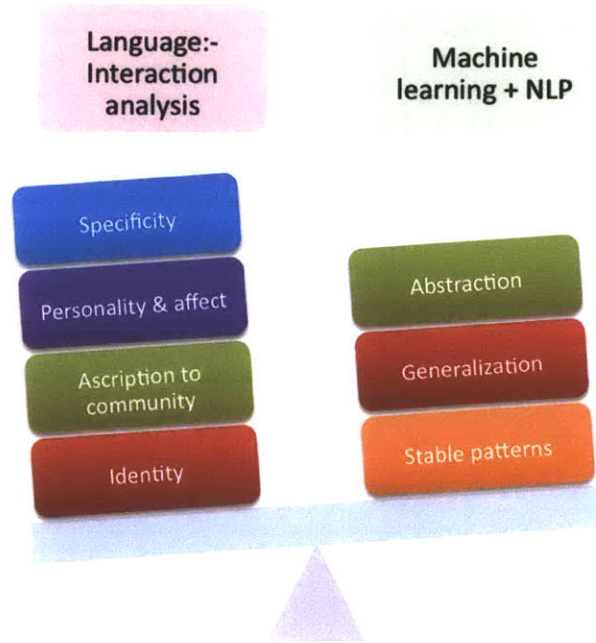


**Figure 1.2.1:** Mitigation involves the combination of detection and reflective user interaction.

For the computational detection part, we examine the state of the art in the fields of statistical supervised and unsupervised learning, commonsense reasoning and probabilistic natural language processing. We hypothesize that the problem of textual cyber-bullying can be decomposed into sub-problems, each of which would need a particular class of algorithms for effective detection. We further hypothesize that within each sub-problem, effective parameterization must involve relevant parameters from allied fields such as social science, psychiatry and sociolinguistics that have scrutinized the problem of bullying for decades. In the next section, we describe the main contributions of this thesis and also briefly address its scope vis-a-vis the complexity involved in the problem.

### 1.3 CONTRIBUTIONS

All of language is meant for communication and should be deemed a social act. The evolution of language concerns three different adaptive systems: individual learning, cultural transmission and biological evolution [10]. The field of sociolinguistics, which dates back to the 1890s, offers a detailed and extensive study of the properties of language. Indeed this field espouses language as a marker of personality, emotional state, social identity, and cognitive styles in addition to properties of ascription to community. The use of language, particularly on the social net-



**Figure 1.3.1:** The seemingly divergent foci of interaction analysis and statistical machine learning.

working websites underlines these characteristics. For example, one would find the language used by a group of people who identify themselves a group, based on gender or sexuality to be different from each other. The use of certain phrases and the choice of certain words seem to carry a benign and even playful connotation when it is within the community, but might be seen as insulting or demeaning if used by a person who does not belong to that group. The discourse on online social networks have within them properties of uniqueness, individuality and the other properties mentioned above [34].

On the other hand, statistical machine learning and natural language processing are adept at finding stable patterns in data. They build off abstraction and strongly emphasize generalization. How then, can one use statistical machine

learning and natural language processing for effective analysis of language? The field of applied computational linguistics has seen great strides in areas such as parsing and machine translation. However the use of these statistical tools in the online domain, particularly in the area of interaction analysis seems to be limited to opinion analysis areas such as sentiment polarity. While the area of sentiment polarity has seen a lot of work, it is difficult to digest the notion that people form blanket opinions on most things such as online reviews and story telling. Finding the middle ground between the two divergent areas of the statistical world and the the world of sociolinguistics is key to applying the former to the later. The main contributions of this work are as follows:

- **Problem decomposition and a society of models** - The problem of textual cyber-bullying can be viewed from three perspectives, namely explicit and blatant forms of abuse, employment of subtlety for indirect forms of abuse, and a mixture of complex issues or teenage dramas that play out on the internet. Most of the explicit forms of abuse generate patterns that are ripe for supervised learning. Indirect ways of insulting another person, although more difficult to detect than explicit forms of abuse, lend themselves to commonsense reasoning. Examining distributions of complex issues surrounding personal stories of bullying are very apt for probabilistic topical models. In this thesis, we follow the aforementioned decomposition of the problem and employ the use of relevant classes of algorithms. A key technical contribution of this work is for a comparative analysis of each of the above three classes of algorithms via-a-vis interaction analysis.
- **Computational empathy** - It is not enough to merely detect candidate instances of cyber-bullying. A key contribution of this thesis is the use of prac-

tical, large scale computational models that power a reflective user interaction paradigm. In this thesis, we examine the critical aspects of approaching the building of statistical models that reap their benefit only when used as one part of a two part paradigm: classes of machine learning models and novel interaction paradigms.

- **Real world deployment** A key contribution of this work is the real world deployment of the above paradigm on a major youth network to help actual teenagers in distress. While the academic publications that have emanated from this work are important, the stakeholders of this project also believe that real-world deployment is an integral aspect. Indeed, the deployment of the the above models is probably the first and only deployment of computational models to help a distressed group of individuals, namely teenagers.

In the next section, we describe the organization of this thesis and guide the reader as to what to expect from each subsequent chapter.

#### 1.4 ORGANIZATION

The organization of this thesis is structured as follows: we begin by providing a concise background and then take the reader through three classes of algorithms, highlighting the strengths and limitations of each and providing a rationale for the next class of methods. We then discuss the real-world deployment of the models before summarizing the contributions of this thesis and providing an outline of future work.

*Chapter 2* provides the necessary background and the related work, both in allied fields such a sociology, psychiatry and sociolinguistics. We also define the

problem space and discuss the rationale behind problem decomposition. We discuss the scope of this thesis given the broad and complex nature of the problem of cyber-bullying.

*Chapter 3* explains in detail the use of supervised learning techniques for explicit forms of abuse. We provide a crisp problem formulation and a description of the corpus used for the experiments. We then discuss the state of the art supervised classification methods including support vector machines. We describe in detail the experiment methodology - with particular emphasis on feature space design. We discuss evaluation with an emphasis on error analysis, thereby underlining the limitations of supervised methods and introduce the reader gently to how these limitations might be overcome.

*Chapter 4* builds on the limitations of supervised learning methods for detecting subtle forms of abuse and introduces the reader to commonsense reasoning. Once again, we give a crisp problem formulation before detailing our approach. We introduce the reader to *ConceptNet* and discuss in detail the inference mechanisms used to detect subtle forms of abuse. We discuss evaluation in the context of the strengths and weaknesses of commonsense reasoning and how and where commonsense reasoning might be used.

*Chapter 5* examines the need to deal with the distribution of complex themes that need to be factored in to help bullied teenagers who can bring themselves to share their personal stories in anticipation of help or anonymous advice. We describe in detail the corpus used for our experiments and proceed to provide a detailed treatment of the application of probabilistic topic models in extracting relevant high-level themes. We discuss how such models can power reflective user interaction, before introducing the reader gently to a real-world deployment.

*Chapter 6* summarizes the main contributions of this work. We compare the algorithmic approaches used in this work against the three yardsticks of representation, inference and learning with respect to interaction analysis and highlight other salient features of this work.

*Chapter 7* provides a glimpse into future work based on this thesis, both theoretical and practical.

-----



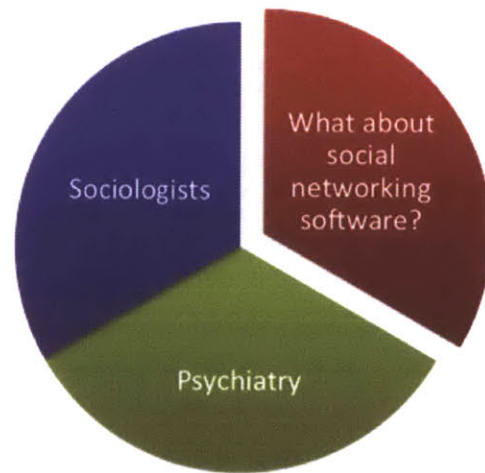
*Empathy is the most revolutionary emotion.*

Gloria Steinem

# 2

## Background, Related Work & Scope

**T**Here is a notion amongst many, that the phenomenon of bullying is a normal rite of passage that marks the transition into adolescence. Believers in this world view espouse that bullying is a normative, yet unfortunate part of growing up and that efforts to rein in bullying behavior amount to alarmist knee-jerk reactions [8]. Yet research from psychiatry and developmental psychology describes in graphic detail the short-term, medium-term and long-term detriments associated with bullying.



**Figure 2:** Social networking software has not been used as an ally in mitigation.

While it is true that bullying is by no means a recent phenomenon, the use of ubiquitous technology such as cellular phones and an ever-increasing penetration rates of online social networks have brought with them new dimensions to this phenomenon. As the real-world story in the introduction section of this thesis shows, cyber-bullying carries with it dimensions that were unthinkable just a decade before. Sociologists such as danah boyd have described the dimensions that make cyber-bullying different from traditional forms of bullying. In the next section, we summarize relevant literature from the perspective of experts, beginning with sociologists that have been studying teenage behavior before discussing related work in computational linguistics and other related applications.

## 2.1 SOCIOLOGY AND PSYCHIATRY

A lot of research in the social sciences has been devoted to understanding the causes of cyber-bullying and the extent of its prevalence, especially for children

and young adults [27]. Research in psychiatry has explored the consequences, both short and long term, that cyber-bullying has on adolescents and school children and ways of addressing it for parents, educators and mental health workers [2]. Such studies, which often involve extensive surveys and interviews, give important pointers to the scope of the problem and in designing awareness campaigns and information toolkits to schools and parents, as well as offering important algorithmic insights to parameterize detection models to catch candidate instances of cyber-bullying. We detail these parameters in each of the subsequent chapters.

Studies have shown that cyber-victimization and cyber-bullying on social networks involving adolescents are strongly associated with psychiatric and psychosomatic problems. A cyber-bully status has been shown to be associated with hyperactivity, conduct problems, low pro-social behavior, frequent smoking and drunkenness and headache, while a cyber-victim status has been shown to include emotional and peer problems, headache, recurrent abdominal pain, sleeping difficulties and not feeling safe at school [40]. Psychiatrists have espoused the need for the induction of strategies to foster cognitive empathy to deal with cyber-bullies as well as cyber-victims [3].

The definition of cyber-bullying has been a topic of consternation and debate among sociologists. While some sociologists hold a view that cyber-bullying is a serious challenge, many others now espouse that increased attention to cyber-bullying actually does more harm than good. The main author of this thesis holds the view that the truth probably lies somewhere between and that this is a topic that is fraught with complex cultural, social and technical issues. That does not however, preclude mobilizing a technical way of mitigating as much of the prob-

lem as possible, even if a true solution probably requires a fundamental change in attitudes.

## 2.2 COMPUTATIONAL LINGUISTICS & INTERACTION ANALYSIS

Machine learning approaches for automated text categorization into predefined labels have witnessed a surge both in terms of applications as well as the methods themselves. Recent machine learning literature has established support-vector machines as one of the most robust methods for text categorization, used widely for email spam filters. The European Union sponsored project PRINCIP has used support vector machines using a bag-of-words approach to classify web pages containing racist text [17]. Indeed, support vector machine was one of our better performing methods for recognizing one of three categories of bullying remarks [13]. Recent work in the NLP community for classification tasks involving online interaction analysis such as identifying fake reviews has shown the effectiveness and importance of drawing intuitive parameters from related domains such as psycholinguistics [32]. In this work, we rely heavily on observations and intuitions from related work in the social sciences and psychology for both problem decomposition as well as feature space design.

## 2.3 RELATED APPLICATIONS

Apart from spam filters, applications that are of a similar nature to this work are in automatic email spam detection and automated ways of detecting fraud and vandalism in Wikipedia [9]. Very few applications have attempted to address the Bullying problem directly with software-based solutions. The FearNot project [47] has explored the use of virtual learning environments to teach 8-12 year old chil-

dren coping strategies for bullying based on synthetic characters. This uses interactive storytelling with animated on-screen characters, where the user gets to play one of the participants in the bullying scenario. The user may select any one of a number of response strategies to a bullying challenge, e.g. fight back, run away, tell a teacher, etc. Though it provides the user with participatory education about the situations, the situations are artificially constructed. They are not part of the users' real lives. It does not make any attempt to analyze or intervene in naturally occurring situations where serious injury might be imminent and might be prevented.

#### 2.4 PROBLEM DEFINITION, DECOMPOSITION & SCOPE

It is important to underline the complexity of the problem of cyber-bullying and carve a crisp problem space that is ripe for the deployment of artificial intelligence and human-computer interaction paradigms. At a fundamental level, bullying amongst the young is influenced by several social and psychiatric factors. If one were to dig deeper into each such factor, it becomes abundantly clear this is a problem that is rooted in societal norms and cultures. For example, consider the set of variables that fall under the larger umbrella of social factors - the tone, tenor and viciousness of adult discourse, at home between parents, as depicted in popular television and movies, between two sides of opposing political and ideological camps all play a role in influencing the developing mindset of young children and of adolescents. Peer pressures, the propensity for depression borne out of self-image issues, varying expressions of angst and worry from violent behavior to a silent accumulation of intense depression are all significant factors. It becomes important to define very clearly what constitutes cyber-bullying, what kind of cyber-bullying

this thesis addresses and what the scope of this work is.

#### 2.4.1 DEFINING CYBER-BULLYING

There are multiple definitions of what constitutes cyber-bullying, even within domains such as sociology, psychiatry and the law. The traditional Olweus definition of bullying stipulates three defining characteristics of bullying [30]:

- Bullying is aggressive behavior that involves unwanted, negative actions.
- Bullying involves a pattern of behavior repeated over time.
- Bullying involves an imbalance of power or strength.

Recent work by sociologists have provided a much broader definition of bullying, in the context of teenage drama, where they blur the boundaries between what adults view as bullying and what young adolescents view as drama which is nonetheless as potent as anything else in the damage that it can cause [25].

Cyber-bullying involves a distribution of digital harassment techniques, not limited to but involving the following: uploading of pictures or photos to embarrass a victim, stealing or hacking of personal information such as passwords and user meta information, sending or posting of abusive or damaging messages on social networking websites or through SMS text messages, sexting, making a fake account of an individual on a social network etc. In this thesis, we limit our work to modeling the detection of textual cyber-bullying: both explicit forms of abuse, implicit or indirect ways of abusing another person, and personal recollections of drama-related angst by teenagers. We adopt both the Olweus definition as well as the more recent definition in relation with teenage drama.

#### 2.4.2 DECOMPOSITION

Modeling the detection of textual cyber-bullying can be imagined as a four-pronged problem. There exist three ways of abusing another individual through the use of language. First, the hurling of explicit insults - expletives and language loaded with contextual features, pertaining to sex, sexuality, age, intelligence, race and appearance. Second, the indirect framing of phrases designed to malign or insult another individual. For example, consider the sentence *'You must be having cheeseburgers for dinner every night'*. This sentence is a clever way of insulting the appearance of an individual without the employment of explicit words like *'fat'*. Third, the use of positive sarcasm in an inappropriate context. For example, consider the comment *'Your hair looks lovely'* hurled at a bald man. Fourthly, first-hand personal recollections of bullying experiences carry with them an entire distribution of drama-related themes, ranging from digital abuse to dating and relationship issues.

In this thesis, we decompose the phenomenon of cyber-bullying into the aforementioned four subcategories and apply relevant classes of statistical machine learning and natural language processing methods to detect them. We stitch together relevant variables as espoused by sociology and psychology literature for each subcategory for the purpose of parameterization.

#### 2.4.3 SCOPE

As mentioned above, true solutions to alleviate the problem of cyber-bullying requires a fundamental restructuring of mindsets and cultural change on a gargantuan scale. The purpose of this thesis is to underline the importance of treating

technology as an ally in mitigating its effects. Major social networks like Facebook and Formspring that have tens to hundreds of millions of users often allow users to '*flag*' a given post or comment as inappropriate. Yet, the huge volume of incoming flagged instances, often with a large number of false positives makes moderation and monitoring of the '*social health*' of the websites an intractable task without the use of hundreds of community moderators.

The computational ways to detect textual cyber-bullying can serve as a means for a moderation team to prioritize an queue of flagged instances. Teenagers expressing recollections of distressing events can be directed to targeted help or shown messages that might alleviate their plight. The scope of this thesis includes finding specific scenarios where an embedding of artificial intelligence can foster empathy for distressed teenagers, as well as allowing social networking websites to design their own strategies for mitigation upon detecting serious cases of cyber-bullying.

## 2.5 *DISCLAIMER*

This thesis draws heavily from research in allied fields of sociology and psychiatry. Our main purpose is to integrate their postulations towards detecting cyber-bullying. We do not validate any assertions or stereotypes related to characteristics such as sexuality, gender, race & culture, intelligence or any other societal classifications. Our only aim is to exert the full power and weight of statistical machine learning and human-computer interaction towards mitigating this very serious societal menace.



## 2.6 ALTERNATE FRAMING OF THE PROBLEM: DISCOURSE ANALYSIS

A central tenet in this thesis is the focus on the content of a message. An alternate framing of the problem is one where one looks at not just the content of a message, but what kind of responses it precipitates. A line of work championed by Carolyn Rose et.al [26] attempts to find the authoritativeness of individuals in a discourse by using sociolinguistic frameworks such as the *Negotiation Framework* from the field of systemic functional linguistics [24]. This approach of trying to decipher shifts in the authoritativeness of participants in a discourse is equally important.

The interpretation of the social significance of message or a post is highly contextual and some have argued that it would be impossible for outsiders to view something as bullying without this context. In this thesis, we make a simplifying assumption that this interpretation is possible for a vast majority of cyber-bullying instances. The main author of the thesis acknowledges that this assumption is a rather big one. But a practical system designed to computationally detect such instances demand a relaxation of assumptions. It is our view that an even more robust system to detect cyber-bullying needs to combine discourse analysis and aspects of the social graph along with the approaches outlined in this thesis.

-----

*The tongue is like a sharp knife. It can kill without drawing  
blood*

Buddha

# 3

## Modeling explicit forms of abuse

In this chapter, we focus on the detection of textual cyber-bullying, which is one of the main forms of cyber-bullying. We use a corpus of comments from YouTube videos involving sensitive topics relating to aspects that are either immutable or perceived to be immutable, which make them both personal and sensitive. We preprocess the data, subjecting it to standard operations of removal of stop words and stemming, before annotating it to assigning respective labels to each comment. We perform two experiments:

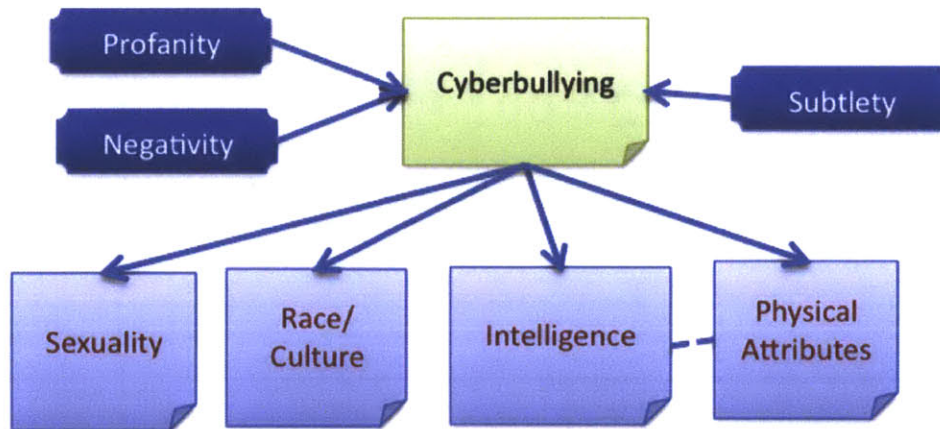
1. training binary classifiers to ascertain if an instance can be classified into one of several sensitive topics or not and
2. multi-class classifiers to classify an instance from a set of sensitive topics.

Within social networking websites, it is a common practice for people to make comments and post messages. When a comment or a message tends to involve sensitive topics that may be personal to an individual or a specific group of people, it becomes a worthy candidate that may qualify as cyber-bullying.

In addition, if the same comment also has a negative connotation and contains profane words, the combination of rudeness and talking of sensitive, personal topics can be extremely hurtful.

For most children in middle school and young adults, the sensitive list of topics often assume one of the following: physical appearance, sexuality, race & culture and intelligence. When a comment or a message on a social networking website involving these sensitive topics are made with rudeness and profanity, and in front of the victim's network of friends, it can be very hurtful. In fact, repeated posting of such messages can lead to the victim internalizing what the bully is saying, which can be harmful to the well-being of the victim.

The problem of detecting hurtful messages on social networking sites can viewed as the following: classifying messages as speaking on sensitive topics and detecting negativity and profanity. The problem then lends itself into a bag-of-words driven text classification experiment.



**Figure 2:** Combination of profanity, negativity & subtlety over sensitive topics.

### 3.1 GRANULARITY IN THE ARC OF CYBER-BULLYING

Social scientists investigating this menace describe the goals of cyber bullies to harm, disrepute or embarrass a victim through 'repeated' acts such as the posting of inappropriate text messages. As such, the arc of textual cyber-bullying consists of a sequence of messages targeting a victim by a lone perpetrator or by a group of individuals.

Exploiting this level of granularity by detecting individual messages that might eventually lead to a tragic outcome assumes importance for two reasons: the design of pre-emptive and reactive intervention mechanisms. In this chapter, we focus on detecting such individual messages.

### 3.2 CORPORA

The dataset for this study was obtained by scraping the social networking site *www.youtube.com* for comments posted on videos. Though YouTube gives the owner of a video the right to remove offensive comments from his or her video,

a big chunk of viewer comments on YouTube are not moderated. Videos on controversial topics are often a rich source for objectionable and rude comments.

Most comments on YouTube can be described as stand-alone, with users expressing opinions about the subject and content of the video. While some of the comments were made as responses to previously posted ones, there were no clear patterns of dialogue in the corpus. As such, we therefore treat each comment as stand-alone, with no conversational features.

Using the YouTube PHP API, we scraped roughly a thousand comments each from controversial videos surrounding sexuality, race & culture and intelligence. We were constrained by the limitations posed by YouTube of being able to download an upper limit of up to a 1000 comments per video. The total number of comments downloaded overall greater was than 50,000.

The downloaded comments were grouped into clusters of physical appearance, sexuality, race & culture and intelligence. A set of 1500 comments from each cluster was then hand annotated (as described in the annotation section below) to make sure that they had the right labels assigned to them. Those comments that were not related to the cluster (for e.g., the comment 'Lol, I think that is so funny') were given a neutral label 'none'. Each cluster had a few comments that belonged to other clusters too, so that the clusters were not mutually exclusive.

### 3.2.1 DATA PREPROCESSING

Each dataset was subjected to three operations: removal of stop-words, stemming and removal of unimportant sequence of characters. Sequences of characters such as '@someuser', 'lollllll', 'hahahahaha', etc., were expunged from the datasets.

### 3.2.2 ANNOTATION

The comments downloaded from all the videos were arranged in a randomized order prior to annotation. Two annotators of whom one was an educator who works with middle school children, and the other a graduate student specializing in prevention science annotated each comment along the lines of three labels defined as follows:

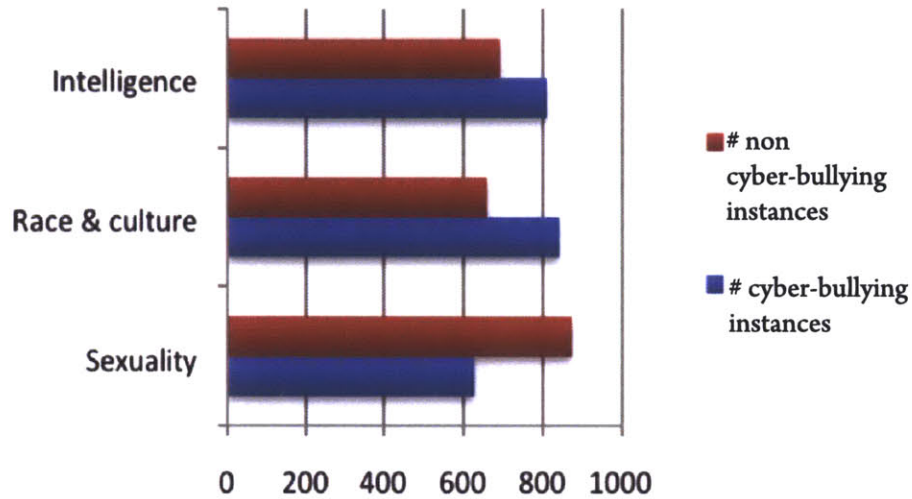
- **Sexuality:** Negative comments involving attacks on sexual minorities and sexist attacks on women.
- **Race & Culture:** Attacks bordering on racial minorities (e.g., African-American, Hispanic and Asian) and cultures (e.g., Jewish, Catholic and Asian traditions) including unacceptable descriptions pertaining to race and stereotypical mocking of cultural traditions.
- **Intelligence:** Comments attacking the intelligence and mental capacities of an individual.

### 3.2.3 INTER-RATER AGREEMENT

Annotated comments with an inter-rater agreement of Cohen's kappa  $\geq 0.4$  were selected and grouped under each label until 1500 comments were available for each label as follows:

$$\kappa = \frac{p_o - p_e}{1 - p_e},$$

where where  $p_o$  is the observed proportion of agreement and  $p_e$  is the proportion of agreement expected by chance.



**Figure 3.2.1:** Distribution of categories after annotation. A given comment tests positive against a category if the annotators agree that it is an instance of bullying with regard to that category.

Bullying Category	# positive instances	# negative instances
Sexuality	627	873
Race & Culture	841	841
Intelligence	809	691

**Table 3.2.1:** Comments downloaded were annotated and grouped by human coders into three categories of 1500 instances each under sexuality, race & culture and intelligence. 627, 841 and 809 instances were found be cyber-bullying instances with labels related to sexuality, race & culture and intelligence respectively.

### 3.3 STATISTICAL SUPERVISED MACHINE LEARNING TECHNIQUES

Interaction analysis on social networks is a complex phenomenon to model mathematically. The field of sociolinguistics that has been studying interaction analysis argues vigorously that the use of language between individuals in a social setting is parameterized by a rich set of characteristics, including identity, ascription to a particular community, personality and affect. They argue that it is specificity and uniqueness that matter the most for effective interaction analysis. But machine learning techniques for language are often reductionist approaches that place a heavy emphasis on abstraction, generalization and stable patterns in the data. Finding a balance between these paradigms is crucial for analyzing discourse on social networks, highlighting the importance of effective feature space design. Indeed, recent work in the computational discourse analysis community has seen the incorporation of principles from sociolinguistics for analyzing discourse [24].

Our approach towards using statistical supervised machine learning is to show its strengths and weaknesses in detecting cyber-bullying. Since explicit verbal abuse involves the use of stereotypical slang and profanity as recurring patterns, those aspects lend themselves nicely to supervised learning algorithms. We also hypothesize that instances of cyber-bullying where the abuse is more indirect, and which does not involve the use of profanity or stereotypical words are likely to be misclassified by supervised learning methods.

We adopt a bag-of-words supervised machine learning classification approach to identifying the sensitive theme for a given comment. We divide the YouTube corpus into 50% training, 30% validation and 20% test data. We choose three types of supervised learning algorithms in addition to Naive Bayes from toolkit WEKA [18], a rule-based learner, a tree-based learner and support-vector machines with



default parameters, described briefly as follows:

1. Repeated Incremental Pruning to Produce Error Reduction, more commonly known as JRip, is a propositional rule learner proposed by Cohen et.al [11]. It is a two-step process to incrementally learn rules (grow and prune) and then optimize them. This algorithm constructs a set of rules to cover all the positive instances in the dataset (those with cyber-bullying labels related to sexuality or race and culture or intelligence) and has been shown to perform efficiently on large, noisy datasets for the purpose of text classification[38].
2. J48 is a popular decision tree based classifier based on the C4.5 method proposed by Ross Quinlan [36]. It uses the property of information gain or entropy to build and split nodes of the decision tree to best represent the training data and the feature vector. Despite its high temporal complexity, J48's performance for classifying text has shown to produce good results [16].
3. Support-vector machines (SVM) [12] are a class of powerful methods for classification tasks, involving the construction of hyper-planes that at the largest distance to the nearest training points. Several papers reference support-vector machines as the state of the art method for text classification [45], [37], [14]. We use a nonlinear poly-2 kernel [20] to train our classifiers, as preliminary experiments with a linear kernel did not yield statistically significant differences with a poly-2 kernel, which has also been a finding in some recent empirical evaluation of SVM kernels [15]].

In the first experiment, binary classifiers using the above were trained on each of the three datasets for each of the three labels: sexuality, race & culture, and

intelligence to predict if a given instance belonged to a label or not. In the second experiment, the three datasets were combined to form a new dataset for the purpose of training a multi-class classifier using the aforementioned methods. The feature space, which we describe in the next section, was built in an iterative manner, using data from the validation set in increments of 50 instances to avoid the common pitfall of over-fitting.

Once used, the instances from the validation set were discarded and not used again to ensure as little over-fitting as possible. The trained models were washed over data from the test set for an evaluation. The kappa statistic, a measure of the reliability of a classifier, which takes into account agreement of a result by chance as well as the F-1 scores were used to gauge the performance of the methods. 10-fold cross validation was applied for training, validation and testing for both the experiments. Measuring just the accuracy of a classifier can be deceptive. Depending on the application, the overall classification rate may not be sufficient if one, or more, of the classes fail in prediction. Cohen's kappa coefficient is a statistical measure of inter-rater agreement for qualitative items [48]. It is generally thought to be a more robust measure than simple percent agreement calculation, since it takes into account the agreement occurring by chance.

### 3.4 FEATURE SPACE DESIGN

The feature space design for the two experiments can be categorized into two kinds: general features that are common for all three labels and specific features for the detection of each label. The intuition behind this is as follows: negativity and profanity appear across many instances of cyber-bullying, irrespective of the subject or label that can be assigned to an instance. Specific features can then be

used to predict the label or the subject (sexuality, race & culture and intelligence).

1. **General features** - The general features consist of TF-IDF (term frequency, inverse-document frequency) weighted unigrams, the Ortony lexicon of words denoting negative connotation, a list of profane words and frequently occurring stereotypical words for each label.
2. **TF-IDF** - The TF-IDF (term frequency times inverse document frequency) is a measure of the importance of a word in a document within a collection of documents, thereby taking into account the frequency of occurrence of a word in the entire corpus as a whole and within each document.
3. **Ortony Lexicon for negative affect** -The Ortony lexicon [31] (containing a list of words in English that denotes the affect) was stripped of the positive words, thereby building a list of words denoting a negative connotation. The intuition behind adding this lexicon as unigrams into the feature set is that not every rude comment necessarily contains profanity and personal topics involving negativity are equally potent in terms of being hurtful.
4. **Part-of-speech tags**- Part-of-speech tags for bigrams, namely, *PRP\_VBP*, *JJ\_DT* and *VB\_PRP* were added to detect commonly occurring bigram pairs in the training data for positive examples, such 'you are', '\*\*\*\* yourself' and so on.
5. **Label Specific Features** -For each label, label specific unigrams and bigrams were added into the feature space that was commonly observed in the training data. The label specific unigrams and bigrams include frequently used forms of verbal abuse as well as widely used stereotypical utterances For ex-

ample, the words 'fruity' and 'queer' are two unigram features for the label sexuality, because of their use for hurtful abuse of LGBT individuals.

We discuss an evaluation of the aforementioned supervised learning methods in the next section. Our hypothesis is that supervised learning methods generally fare well when it comes to detecting explicit forms of verbal abuse owing to the presence of stable patterns. We anticipate in our error analysis that instances of cyber-bullying that are indirect and which do not involve the use of explicit language, of which there aren't enough training samples, are likely to be misclassified by the models. In the next chapter we discuss the need for using commonsense knowledge reasoning to detect instances of cyber-bullying that could not be caught using the aforementioned conventional supervised learning methods.

### 3.5 EVALUATION & DISCUSSION

The statistical models discussed above were evaluated against 200 unseen instances for each classifier. The labels assigned by the models were compared against the labels that were assigned to the instances during annotation. The accuracy and kappa values of the classifiers are in the tables below.

To avoid lexical overlap, the 200 instances for each label were derived from video comments that were not part of the original training and validation data. Prior work on the assessment of classifiers suggests that accuracy alone is an insufficient metric to gauge reliability. The kappa statistic  $\kappa$  (Cohen's kappa), which takes into account agreement by chance, has been argued as a more reliable metric in conjunction with accuracy [7]. We evaluate each classifier in terms of the accuracy, the F1-score, as well the kappa statistic.

Overall, multi-class classifiers underperformed compared to binary classifiers.

Multi-class Naive Bayes for the merged dataset had lower accuracy, F1 and kappa scores compared to individual binary classifiers for all three labels. For JRip, although F1 scores for race and sexuality binary classifiers were slightly better than the multi-class classifier, accuracy levels were lower. In terms of support-vector machines, although there was no significant difference in terms of accuracy, F1 scores and the kappa statistics for each of the binary classifiers were better than the multi-class classifier.

#### 3.5.1 ERROR ANALYSIS OF THE SUPERVISED LEARNING MODELS

As we hypothesized, an error analysis on the results reveals that instances of bullying that are apparent and blatant are simple to model because of their stable, repetitive patterns. Such instances either contains commonly used forms of abuse or profanity, or expressions denoting a negative tone. For example, consider the following instances:

```
u1 as long as fags don t bother me let them do what they want
```

```
u2 hey we didnt kill all of them, some are still alive today.  
And at least we didnt enslave them like we did the monkeys,  
because that would have been more humiliating
```

---

**Figure 3.5.1 (following page):** Multi-class classifiers for the merged dataset Binary classifiers trained for individual labels fare better than multi-class classifiers trained for all the labels. JRip gives the best performance in terms of accuracy, whereas SMO is the most reliable as measured by the kappa statistic.

Binary Classifiers

	Naïve Bayes			Rule-based JRip			Tree-based J48			SMO (SVM)		
	Accuracy	F1	Kappa	Accuracy	F1	Kappa	Accuracy	F1	Kappa	Accuracy	F1	Kappa
<b>Sexuality</b>	66%	0.67	0.65	<b>80%</b>	0.76	0.59	63%	0.57	0.57	67%	0.77	<b>0.79</b>
<b>Race</b>	66%	0.52	0.78	<b>68%</b>	0.55	0.78	63%	0.48	0.65	67%	0.63	<b>0.71</b>
<b>Intelligence</b>	72%	0.46	0.46	<b>70%</b>	0.51	0.51	70%	0.51	0.56	72%	0.58	<b>0.77</b>

JRip's performance as measured by accuracy was better than the other three approaches

Support-vector machines were the best in terms of F1 scores and kappa values

Multi-class classifiers

<b>Mixture</b>	63%	0.57	0.44	63%	0.60	0.50	61%	0.58	0.45	66%	0.63	0.65
----------------	-----	------	------	-----	------	------	-----	------	------	-----	------	------

A merged set of instances from the three clusters of sexuality, race and intelligence

Both the instances shown above (the first pertaining to sexuality and the second pertaining to race) contain unigrams and expressions that lend them to be positively classified by the models. Instances such as the ones shown above, which contain lexical and syntactic patterns of abuse, lend themselves to supervised learning for effective detection. However, the learning models misclassified instances that do not contain these patterns and those that require at least some semantic reasoning. For example, consider the following instances:

```
u3 they make beautiful girls, especially the one in the green
top
```

```
u4 she will be good at pressing my shirt
```

In the first instance, which was posted on a video of a middle school skit by a group of boys, the bully is trying to ascribe female characteristics to a male individual in the video. The instance has no negativity or profanity, but implicitly tries to insult the victim by speculating about his sexual orientation. 'Tops' and 'beautiful' are concepts that are more associated with girls rather than boys, and hence if attributed to the wrong gender, can be very hurtful. In the second instance, a bully exploits the common sexist stereotype that pressing clothes is an act reserved primarily for women. The learning models misclassified these two instances, as it would need to have some background knowledge about the stereotypes and social constructs and reason with it. In the next section, we discuss our work with supervised learning models in the context of related approaches to sentiment analysis.

### 3.5.2 DISCUSSION

Prior research in sentiment analysis has focused on sentiment polarity for opinion analysis for movie and product reviews [33]. However, the nature of interpersonal and group interaction on social networks is different from sentiment polarity of reviews from two perspectives and hence difficult to compare. First, interaction of social networks (like Formspring) as a sociolinguistic phenomenon is more targeted towards a specific audience (an individual or a group of individuals), whilst movie and product reviews are intended for a larger, more general audience. Second an analysis of discourse on social networks involves deeper attributes such as identity, ascription to a particular community, personality and affect, which is more than just sentiment polarity of movies or product reviews where there is a prior acknowledgement of the domain under scrutiny (a movie or a particular product).

Recent work with affect recognition in text has attempted a fine-grain compositional approach to gauging emotions [29]. While we did not adopt a finer granularity approach towards gauging emotion, we emphasize that our focus was on overcoming the limitations of supervised learning methods in catching indirect, subtle forms of abuses using social constructs which requires reasoning along relevant dimensions (such as gender roles).

In the next chapter, we discuss a class of algorithms for leapfrogging over the limitations of the statistical supervised techniques to deal with indirect forms of abuse.

-----



*There are a sort of men whose visages do cream and mantle like  
a tanding pond*

Shakespeare

# 4

## Commonsense reasoning: Modeling indirect forms of abuse

Prior to our error analysis, we anticipated that instances of cyber-bullying that are indirect and which do not involve the use of explicit language, of which there aren't enough training samples, are likely to be misclassified by the models. Indeed, an error analysis of the supervised learning methods in the previous chapter showed us two things: of the strengths of these methods in catching explicit forms of abuse and their limitations with regard to more subtle cases of bullying

that require a deeper level of reasoning. In this chapter we discuss the need for using commonsense knowledge reasoning to detect instances of cyber-bullying that could not be caught using the aforementioned conventional supervised learning methods.

#### 4.1 THE OPEN MIND COMMONSENSE KNOWLEDGE BASE

When we reason about the world, we are using our knowledge of what is expected, to react to and anticipate situations. As discussed before, traditional supervised learning techniques tend to rely on explicit word associations that are present in text. Using common sense can help provide information — about people’s goals and emotions and object’s properties and relations — that can help disambiguate and contextualize language.

The goal of the Open Mind Common Sense (OMCS) [39] project is to provide intuition to AI systems and applications by giving them access to a broad collection of basic knowledge, along with the computational tools to work with it. This knowledge helps applications to understand the way objects relate to each other in the world, people’s goals when they go about their daily lives, and the emotional content of events or situations. OMCS has been collecting common sense statements from volunteers on the Internet since 1999. At the time of this research, we have collected tens of millions of pieces of English language common sense data from crowd sourcing, integrating other resources, and the Semantic Web.

This knowledge allows us to understand hidden meaning implied by comments, and to realize when others are making comments designed to make us feel like our behavior is outside of the *‘normal’*. When we communicate with each other, we rely on our background knowledge to understand the meanings in conversation. This

follows from the maxim of pragmatics that people avoid stating information that the speaker considers obvious to the listener.

Common sense allows us a window to what the average person knows about a concept or topic. This allows us to look for stereotypical knowledge, especially about sexuality and gender roles. OMCS knows that a girl is capable of doing housework, holding puppies, wearing bows in their hair, and babysitting and that a boy is capable of crying wolf, bagging leaves, wrestling, playing video games, and shouting loudly. More direct clues can be found in the gender associations of certain words: For example, OMCS associates dresses and cosmetics more strongly with girls. We emphasize that it is not our intention to validate or approve of any of these stereotypes, but only to use such stereotypical assertions for detection of subtle, indirect forms of verbal abuse.

For the knowledge we collect to become computationally useful, it has to be transformed from natural language into more structured forms that emphasize the contextual connections between different concepts. ConceptNet represents the information in the OMCS corpus as a directed graph [22]. The nodes of this graph are concepts, and its labeled edges are assertions of common sense that connect two concepts.

Concepts represent aspects of the world as people would talk about them in natural language, and they specifically correspond to normalized forms of selected constituents of common sense statements entered in natural language. This research uses ConceptNet 4, in which one of twenty-one different relations connects two concepts, forming an assertion. Each assertion has a notation of whether or not the relationship is considered to be negated (polarity), and a score representing the public's general opinion on whether the predicate is true or not. For exam-

ple, the assertion '*A skirt is a form of female attire*' connects the 'skirt' and 'form of female attire' nodes with the 'IsA' relation.

ConceptNet can also be represented as a matrix where the rows are concepts in the graph. The columns represent graph 'features' or combinations of relation edges and target concepts. Features can be thought of as properties that the object might have such as '*made of metal*' or '*used for flying*'. This network of concepts, connected by one of about twenty relations such as '*IsA*', '*PartOf*', or '*UsedFor*', are labeled as expressing positive or negative information using a polarity flag. The relations are based on the most common types of knowledge entered into the OMCS database, both through free text entry and semi-structured entry. For the assertion '*A beard is part of a male's face*', for instance, the two concepts are '*beard*' and '*male*', the relation is '*IsA*', and the polarity is positive. For the assertion '*People don't want to be hurt*', the concepts are '*person*' and '*hurt*', the relation is '*Desires*', and the polarity is negative.

Each concept can then be associated with a vector in the space of possible features. The values of this vector are positive for features that produce an assertion of positive polarity when combined with that concept, negative for features that produce an assertion of negative polarity, and zero when nothing is known about the assertion formed by combining that concept with that assertion. As an example, the feature vector for '*blouse*' could have +1 in the position for '*is part of a female attire*', +1 for '*is worn by girls*', and +1 for '*is worn by women*'. These vectors together form a matrix whose rows are concepts, whose columns are features, and whose values indicate truth values of assertions. The degree of similarity between two concepts, then, is the dot product between their rows in the concept/feature matrix. This representation is discussed in detail by Speer & Havasi et.al [42].

#### 4.2 THE ANALOGYSPACE INFERENCE TECHNIQUE

In order to reason over this data set, we needed to develop an algorithm that was both noise resistant and which took advantage of patterns inherent in how we see the world. When we determine if an object is animate, for example, we may look at the properties of that object. Does it move on its own? Is it fuzzy? Or made of metal? Is it a common pet? We also think about what objects are most similar to it. Does it look like a rabbit? Or a robot? Is it a concrete object like a pony or an immaterial quantity such as happiness?

Each question you might ask about a concept can be thought of as a '*dimension*' of a concept space. Then, answering a question such as where does an object lie along the '*animate vs. inanimate*' dimension, can be thought of as reducing the dimensions of the space from every question you might ask, to just the question of interest; that is, projecting the concept onto that one dimension. We therefore use mathematical methods for dimensionality reduction, such as singular value decomposition (SVD) [42] to reduce the dimensionality of the concept-feature matrix. This determines the principal components, or axes, which contain the salient aspects of the knowledge, and which can be used to organize it in a multi-dimensional vector space. The resulting space can be used to determine the semantic similarity using linear operations over vectors representing concepts in the semantic space. Concepts close together in the space are treated as similar; these are also more likely to combine to form a valid inference.

Let us call the matrix whose rows are concepts, whose columns are features, and whose values indicate truth values of assertions as  $A$ . This matrix  $A$  can be factored into an orthonormal matrix  $U$ , a diagonal matrix  $\Sigma$ , and an orthonormal matrix  $V^T$ , so that  $A = U\Sigma V^T$ . The singular values are ordered from largest to

smallest, where the larger values correspond to the vectors in  $U$  and  $V$  that are more significant components of the initial  $A$  matrix. We discard all but the first  $k$  components - the principal components of  $A$  - resulting in the smaller matrices  $U_k$ ,  $\Sigma_k$ , and  $V_k^T$ . The components that are discarded represent relatively small variations in the data, and the principal components form a good approximation to the original data. This truncated SVD represents the approximation  $A$ , such that  $A_k = U_k \Sigma_k V_k^T$ . As  $\text{AnalogySpace}$  is an orthogonal transformation of the original concept and feature spaces, dot products in  $\text{AnalogySpace}$  approximate dot products in the original spaces. This fact can be used to compute similarity between concepts or between features in  $\text{AnalogySpace}$ .

#### 4.3 THE BLENDING KNOWLEDGE COMBINATION TECHNIQUE

While it is useful to use common sense to acquire more common sense, we benefit more when we use these techniques to learn from multiple data sets. Blending [19] is a technique that performs inference over multiple sources of data simultaneously by taking advantage of the overlap between them. Two matrices are combined using a blending factor and then a SVD is taken over both data sets. Blending can be used to incorporate other kinds of information, such as information about stereotypes, into a common sense matrix to create a space more suited for a particular application.

We can use this technique to create a specific knowledge base to collect knowledge about different types of stereotypes beyond those in the OMCS database. Blending balances the sizes and composition of the knowledge bases such that the small size of such a knowledge base is not overpowered by the (much) larger  $\text{ConceptNet}$ . Additionally, information about implicit stereotypes may bring out

other lightly stereotyped knowledge in the database and allows us to expand the reach of entered stereotypical knowledge. For example, adding OMCS allows us to discover that mascara, not just makeup, is usually associated with girls in the context of fashion.

Common sense can be used to fill in the gaps in other knowledge sources, both structured and unstructured, or can be designed to cover knowledge surrounding a narrow special topic. For example, in his work with SenticNet, Erik Cambria [5] created a specialized knowledge base with information about emotions. That database has been combined with common sense and domain specific texts to create a system that attempts to understand affect in free text [6].

In the following sections, we build a knowledge base to perform commonsense reasoning over a specific slice of cyber-bullying, namely that concerning gay and lesbian issues.

#### 4.4 THE BULLYSPACE KNOWLEDGE BASE

A key ingredient to tackling implicit ways of insulting another person is to transform commonly used stereotypes and social constructs into a knowledge representation. For example consider the following instance from the Formspring corpus:

put on a wig and lipstick and be who you really are

In the above instance, a bully is trying to speculate about or malign the sexuality of a straight, male individual implicitly, by trying to attribute characteristics of the opposite sex. (Of course, in the context of a conversation between openly gay people, such a comment may be completely innocuous.) The underlying social construct here is that, in a default heterosexual context, people don't like to be

attributed with characteristics of the opposite sex. This attribution is made using the common stereotype that wigs and lipstick are for women or for men who want to dress as women.

In this work, we observe the Formspring dataset and build a knowledge base about commonly used stereotypes used to bully individuals based on their sexuality. The representation of this knowledge is in the form of an assertion, connecting two concepts with one of the twenty kinds of relations in ConceptNet. For the above example, the assertions added were as follows:

```
lipstick is used by girls  
lipstick is part of makeup  
makeup is used by girls  
a wig is used by girls  
a toupee is used by men
```

We build a set of more than two hundred assertions based on stereotypes derived from the LGBT related instances in the Formspring database. We emphasize that our aim is not to endorse any of these stereotypes, but merely to detect their use in bullying. We then convert these assertions into a sparse matrix representation of concepts versus relations, in the same manner as ConceptNet. We then use AnalogySpace’s joint inference technique, blending, to merge them together to create a space that is more suited for the purpose of detecting implicit insults concerning LGBT issues. While blending, we give double post-weight to the matrix generated from the set of assertions specifically designed to capture LGBT stereotypes. Once the two matrices have been merged, we then perform an AnalogySpace inference by performing an SVD to reduce the dimensionality of the matrix



by selecting only the top  $k = 100$  set of principal components. We now have the basic machinery required to perform commonsense reasoning.

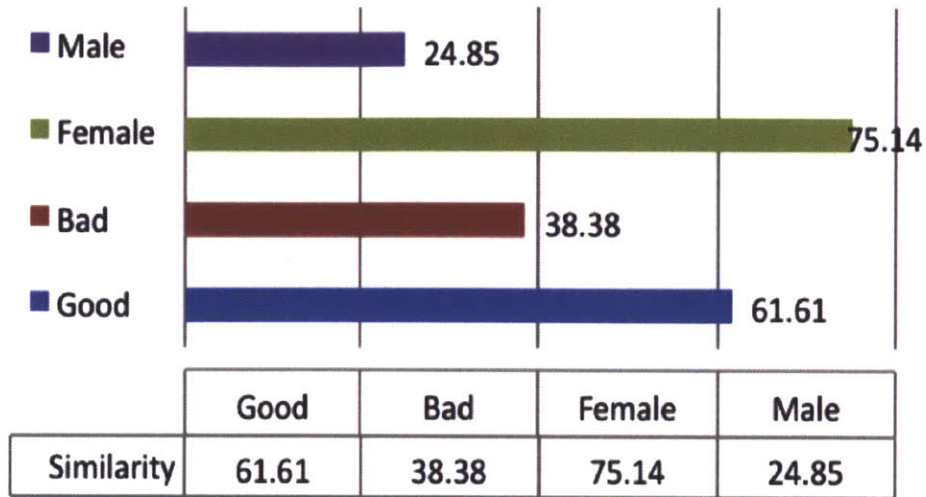
#### 4.5 COSINE SIMILARITY OF EXTRACTED AND CANONICAL CONCEPTS

A given comment is first subjected to an NLP module to perform the standard normalization operations: removing stop-words, and tokenizing the text to have a clear separation of words from punctuation marks. Next, we extract a list of concepts from the normalized text that is also present in the concept axes of the dense matrix derived after performing the SVD, as explained in the previous section.

The next task is to choose a set of canonical concepts for comparison with the concepts that have been extracted from the comment. We select four canonical concepts, namely the affective valences positive and negative, as well as gender, namely male and female. The idea here is to compare each extracted concept for similarity with each of the canonical concepts. This is achieved by performing a dot product over the extracted concept with a canonical concept. After this comparison, we normalize the values derived for each of the canonical concepts to get an overall measure of how similar the given comment is to each of the canonical concepts.

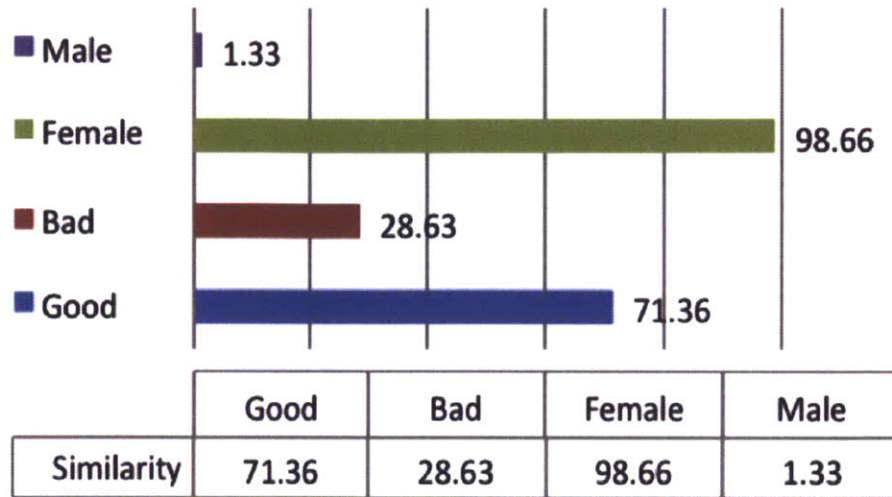
For example, consider the comment '*Did you go shopping yesterday?*'. This comment is subjected to the process described above to yield similarity scores for the canonical concepts of good, bad, boy and girl.

Similarity scores derived from the above examples show that shopping as a concept is more oriented towards girls than boys, and is largely considered as an enjoyable activity rather than a bad one. Based on these similarity scores, it can be inferred that this is a fairly innocent comment.



**Figure 4.5.1:** Results for the comment 'Did you go shopping yesterday?'. Shopping as a concept is more related to females. It is also more considered a good activity, owing to the fact that more users in the ConceptNet database regard it as a good activity rather than bad.

Consider another comment, '*Hey Brandon, you look gorgeous today. What beauty salon did you visit?*'. Although this contains no profanity, it does appear on the face of it, to be a comment more attributable to a girl than a boy.



**Figure 4.5.2:** Results for the comment 'Hey Brendan, you look gorgeous today. What beauty salon did you visit?' is the concept overwhelmingly tending towards a female rather than a male because of the words 'gorgeous' and 'beauty salons'. Even though the lack of profanity or rudeness might give an impression of denoting positive affect, it relates more to females. If this comment was aimed at a boy, it might be an implicit way of accusing the boy of being effeminate, and thus a candidate sentence for cyber-bullying.

An analysis of the comment shows that it is overwhelmingly more similar towards female concepts rather than male. The concepts of 'gorgeous' and 'beauty salons' are those that typically used in reference to girls rather than boys. If this comment was aimed at a boy, it might be an implicit way of accusing the boy of being effeminate, and thus becoming a candidate sentence for cyber-bullying that deserves further scrutiny. Note that 'gorgeous' by itself has a positive connotation, so it would be misinterpreted by something merely looking for positive vs. negative words.

Here, we have focused on LGBT accusations, but in much the same way, domain-specific knowledge about other topics connected with cyber-bullying such as race

and culture, intelligence and physical appearance, social rejection etc can be built. Canonical concepts can be selected for each of the topics in much the same way. For example, for the topic of physical appearance, the concept of 'fat' would be a canonical concept. 'French fries' and 'cheeseburgers' for example, would be closer to the concept of 'fat' than 'salads'.

We discuss the evaluation of both the statistical supervised learning methods and commonsense reasoning in the next section. An error analysis on the supervised learning methods highlights the need for commonsense reasoning. Of course, detection is just the first part of addressing cyber-bullying. In the next part of this chapter, we discuss some approaches for reflective user interaction and intervention to formulate an end-to-end model of tackling textual cyber-bullying from detection to mitigation.

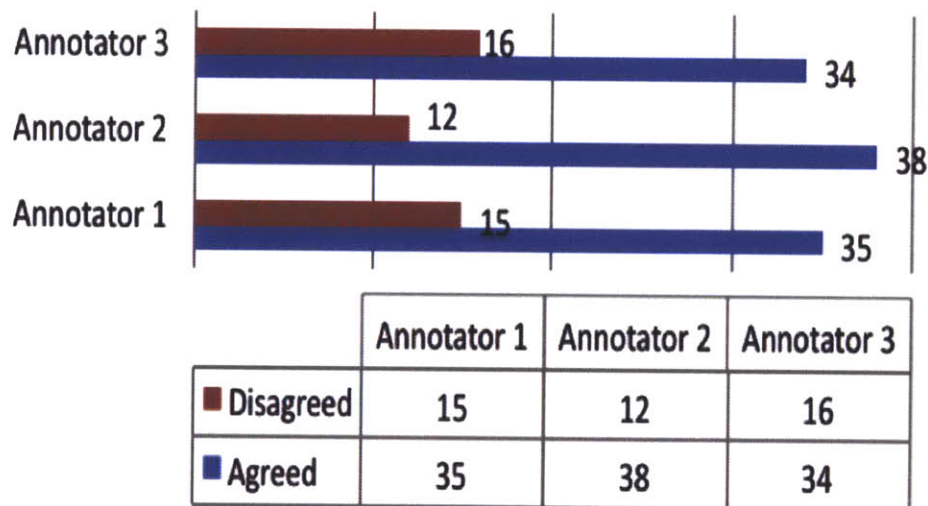
#### 4.6 EVALUATION OF COMMONSENSE REASONING MODELS

For an evaluation of the detection part involving commonsense reasoning, it is essential to have a dataset that contains instances of cyber-bullying that are devoid of profanity and are implicitly crafted to insult or malign a user. To address the specific slice of LGBT bullying in this work, it is essential to have a test dataset that pertains to LGBT bullying as well as some instances that do not pertain to LGBT issues.

We build such a test by performing a filtering operation on the original Formspring dataset as follows: the same set of people who annotated the YouTube corpus were asked to pick instances from the Formspring dataset that satisfied the dual criteria of not having any profanity and implicitly trying to attack, insult or speculate the sexuality of the victim. Of the 61 instances of bullying that were

obtained from the three annotators, 50 instances were made into a test set. It is important to keep in mind that the original Formspring corpus contains instances that have already been flagged as bullying. Hence the annotators were not asked to check if an instance was bullying or not.

Since the goal of this detection approach that we take in this paper is to prioritize reported instances of bullying based on similarity scores, we adopt a similar approach for this test dataset. The test dataset was evaluated with the approach mentioned in section 2 to generate similarity metrics for each instance with the canonical concepts girl and boy. The results were shown to each of the three annotators to check if they agreed with the metrics generated by the commonsense reasoning model. The results are as follows:



**Figure 4.6.1:** Evaluation of the commonsense reasoning model. 50 instances from the Formspring dataset were evaluated by the model to generate similarity scores for the canonical concepts girl and boy. The same three annotators, as shown above, validated the results. All the 50 instances were previously flagged as instances of cyberbullying.

## 4.7 ERROR ANALYSIS OF THE COMMONSENSE REASONING MODEL

An analysis of the instances for which the annotators disagreed with the commonsense reasoning model can be classified into three kinds. The first kind were instances for which the similarity metrics did not make common sense and the second kind were instances for which the annotators did not agree to the scale of the similarity metrics.

Most of the instances for which similarity metrics did not make commonsense were largely due to sparsity of data in the space that was built for performing the SVD. For instance, consider the following example having similarity metrics that did not make common sense:

George Michael or Elton John?

This instance received an extremely high score for the concept 'boy' owing to the names of the individuals mentioned above. However, a deeper analysis shows that the above individuals are celebrity singers who also have one thing in common between them they are both openly gay. The three annotators all agreed that by suggesting that an individual likes these singers, the perpetrator is implicitly trying to speculate or mock their sexuality. To address such instances, one really needs to have more canonical concepts than girl and boy.

Those instances for which the relative scale of the similarity scores was not agreeable to the annotators can be attributed to the scoring process in Concept-Net, which relies on the frequency of an assertion to determine its relative scoring. For example, consider the following instance:

why did you stop wearing makeup?

The above example generated a normalized similarity scores as follows: 60.7% for the concept of girl and 32.2% for the concept of boy. While there can be men who in various roles such as actors routinely wear makeup, makeup is more strongly associated with women than men. This suggests the need for an in-context weighting of assertions. Makeup and costumes, for example are more likely to be associated with individuals in the performing arts, irrespective of their gender.

It is clear from the evaluation that the problem of sparsity, as well as the ability to individually weight an assertion will be vital if this approach is to be implemented in a large-scale user community. One can imagine crowd-sourced collection of such relevant social constructs as commonsense assertions from both users and moderators of social networks.

-----

*Probability is expectation founded upon partial knowledge. A perfect acquaintance with all the circumstances affecting the occurrence of an event would change expectation into certainty, and leave neither room nor demand for a theory of probabilities.*

George Boole

# 5

## Teenage Drama: Probabilistic Topic Modeling

In the previous two chapters, we have explored how conventional machine learning and commonsense reasoning can be used to detect explicit and indirect forms of abuse. There is some debate as to what topics constitute bullying and how distressed teenagers view their problems as opposed to adults. In this chapter, we look at more deeply at addressing this conundrum and attempt to provide an empathetic framework to help distressed teenagers. It is important to understand that while the approaches described in the previous two chapters are geared to-



wards detection of bullying, here we attempt to provide a computational framework to help distressed teenagers.

As stated before studies have shown that cyber-victimization and cyber-bullying on social networks involving adolescents are strongly associated with psychiatric and psychosomatic problems. A cyber-bully status has been shown to be associated with hyperactivity, conduct problems, low pro-social behavior, frequent smoking and drunkenness and headache, while a cyber-victim status has been shown to include emotional and peer problems, headache, recurrent abdominal pain, sleeping difficulties and not feeling safe at school [41] To understand what constitutes adolescent cyber-bullying and cyber-victimization, we draw a distinction between how adults view cyber-bullying versus how teenagers perceive it. The latest research by social scientists has shown what might be perceived as cyber-bullying by adults is often viewed as mere 'drama' by teenagers. [25] Psychiatrists have espoused the need for the induction of strategies to foster cognitive empathy to deal with cyber-bullies as well as cyber-victims. [4].

We synthesize the aforementioned ideas from sociologists and psychiatrists to frame a hypothesis that the tackling of cyber-bullying involves two key components - detection and an in-context reflective user-interaction strategy to encourage empathy on social networks. In this chapter, we use a corpus of real-world stories by distressed teenagers from MTV for two purposes - to understand the distribution of 'drama-like' themes from the collection of stories as well as how that can be used for theme-based story matching that can power a reflective user-interface. We use probabilistic latent Dirichlet allocation and principles from sociolinguistics to draw dominant themes from the corpus and use that to match a new story to a similar story - with a view of trying to show distressed teenagers

that they are not alone in their plight.

### 5.1 TEENAGE DRAMA AND ITS THEMATIC DISTRIBUTIONS

As we discussed in Chapter 2, the traditional Olweus definition of bullying highlights three aspects of bullying, namely the unwanted and negative aggressive behavior of a bully, its repeating nature, and power imbalances that make self-defense difficult for a victim. If one were to detect textual cyber-bullying under this definition, it becomes essential to detect the underlying topic of discussion that might cause a power imbalance. Previous work in this realm has espoused that topics that are immutable with respect to individual characteristics, such as race, culture, sexuality and physical appearance etc, are helpful to identify power imbalances [50].

But recent research suggests a much broader definition of bullying beyond that of Olweus that has interesting implications for any algorithmic detection of cyber-bullying. Drawing the distinction that teenagers like to refer to adult-labeled bullying as mere drama rather than bullying broadens the scope of what constitutes cyber-bullying. Researchers have shown that drama includes a distribution of aggressive and passive-aggressive behaviors, ranging from posting what teens often refer to as 'inappropriate' videos and photos and the resulting fallout; conflicts that escalate into public standoffs; cries for attention; relationship breakups, makeup's and jealousies etc [25].

This suggests that for effective detection of textual cyber-bullying on social networks one needs to factor in the distribution of these teenage drama themes, beyond classification techniques to label an interaction on a social networking website into just a single label. In the next section, we discuss an approach to extract

these themes from a corpus of real-world stories by distressed teenagers from the popular MTV website *athinline.org*. We begin by describing the sociolinguistic characteristics of the corpus and then describe our approach of extracting themes.

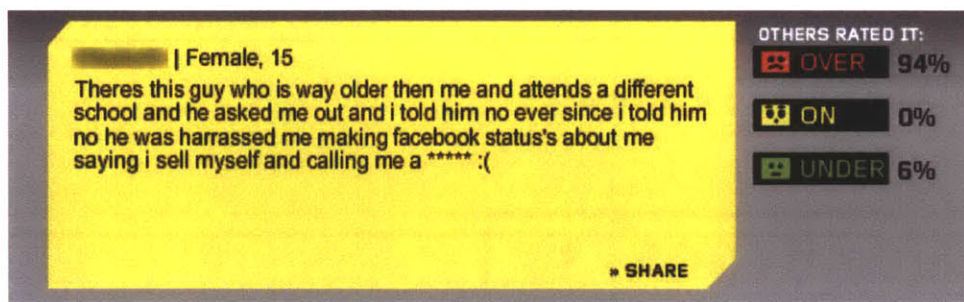
## 5.2 THE MTV TEENAGE STORIES CORPUS

The popular youth culture network MTV's website, *athinline.org* allows adolescents and young adults in distress to share their stories anonymously with a view of getting crowd-sourced feedback and advice. When a teenager posts a story on their website, anyone can read it and vote on its severity in three ways - over the line (severe), on the line (moderate to mild) and under the line (not very serious). Although the site first started as an attempt to help teenagers from digital harassment such as sexting and social network bullying, the range of topics in the stories that teenager talk about on the site have a much broader scope, from dating to very serious cases of physical abuse. The age of those posting stories on the site ranges from 12-24, although more than half of it comes from teenagers.

We analyze a completely anonymized corpus of 5500 personal stories from this website, with data on the votes that each story received. Upon an initial examination of the corpus, most stories contained a set of themes. For example, in Figure 5.2.1, the two dominant themes for the story could be imagined as a combination of 'high school drama' and 'bullying on social networks'.

### 5.2.1 DIGLOSSIA AND HEDGING

A further examination of some of the stories in the corpus revealed interesting sociolinguistic attributes characterizing teenage discourse. Consider the following two stories from the corpus:



**Figure 5.2.1:** A sample story from a distressed teenager on athinline.org describing her negative experience. The reactions from others who viewed this story are shown on the right. 94% of the people who voted think who voted think that this story was over the line , while 6% think that this was under the line and not very serious..

I had this one guy trick me into thinking he was considering" liking me over the summer. I have big boobs and he really wanted to see them online. It was all done over sexting and all that crap. I didn't do it but he was pressuring me.

i have this guy friend that is really nice but hes always asking me to send him sexy pic or pics of my \*\*\*\* or \*\*\*\* n i say no be cause im not a \*\*\*\* so sorry n the next day he doesnt talk to me n calls me a \*\*\*\* not cool right????????????????????????????????

Both of the stories have 'Sending / uploading nude / naked pictures of boyfriend or girlfriend' as their dominant theme though the styles employed by the former is more formal than the later. In the former, there is no mention of the word 'picture', but the story still has the same theme. The two stories highlight diglossia in online teenage discourse [1], a situation where a single community employs two dialects

or styles, which in this case happens to be girls. Similarly, consider the following story by a teenage girl whose severity is rather grave:

```
my bf doesn t luv me anymore. im kinda sucks,  
kinda scared want me life 2 end
```

This story is representative of hedging [49], or softening one’s expression of something acerbic. The two dominant theme in this story appear to be ‘breakup heartache’ and ‘feeling scared’. The example tells us that the absence of keywords such as suicide is not representative of how grave it really is. In fact, there were at least 14 other stories by girls with the same combination of themes that were indicative of suicidal tendencies that were voted as `over the line` by third-party individuals. These kinds of intuitions can serve as a yardstick for prioritizing targeted help for the most severe stories. In the next section, we describe an approach to extract themes from these stories keeping in mind the aforementioned sociolinguistic themes in the corpus.

### 5.3 EXTRACTING HIGH-LEVEL THEMES

To extract high-level themes from the MTV corpus, there are two plausible approaches. The first was to subject the corpus to human annotation by requiring annotators to mark each story with the themes present in it. Whilst this would curate the corpus for supervised hierarchical classification, it would require all the 5500 stories to be annotated by the same set of people for the best results. Instead, we choose a different two-step approach: a) to apply latent Dirichlet Allocation (LDA) to get a set of word clusters and b) interpret these sets word clusters to assign a theme to each individual word cluster using principles from sociolinguistics.

The rationale behind choosing LDA is to avoid manual annotation of every story and to assign theme distributions to each individual story.

We use the following standard LDA model for this purpose as follows [43]: Let  $T$  denote the number of themes in teenage stories, and  $D$  denote the number of stories in your database, and  $N$  the total number set of words in the corpus. Let  $P(z)$  denote the distribution over themes  $z$  in a particular story, and  $P(w|z)$  for the probability mass over word  $w$  given theme  $z$ . We can then write the following distribution of words within a given story:

$$P(w_a) = \sum_{b=1}^T P(w_a|z_a = b)P(z_a = b)$$

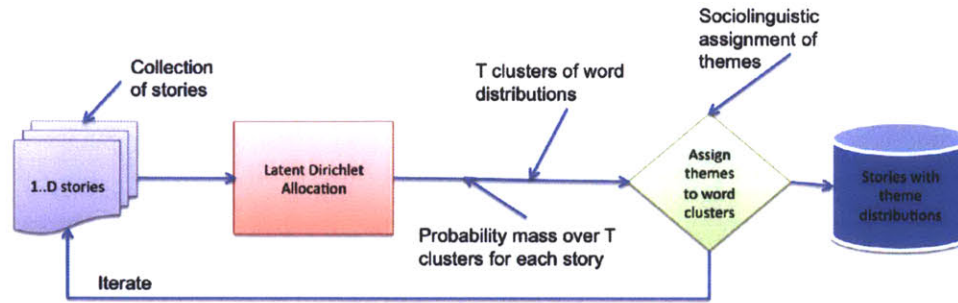
We define parameters  $\varphi$  and  $\theta$  as:

$$\theta^{(d)} = P(z) \text{ and } \varphi^{(b)} = P(w|z = b)$$

We follow the model by Griffiths and Steyvers in drawing a Dirichlet priors  $Dir(\beta)$  and  $Dir(\alpha)$  for both  $\varphi$  and  $\theta$ . We use the Gibbs sampling method [35] and use  $\alpha = 50/T$  and  $\beta = 0.01$  for each iteration of our approach, which we describe below.

### 5.3.1 ASSIGNING THEMES TO WORD CLUSTERS

We begin the process of extracting themes by setting the number of topics  $T$  in increments of 10. The process of determining whether a satisfactory number of themes have emerged from the LDA model is as follows:



**Figure 5.3.1:** Extracting high-level themes from the MTV corpus. The stories are first run through LDA to get clusters of word distribution. Using principles from sociolinguistics, the clusters are interpreted into themes. This process is repeated until the a satisfactory number of themes can be extracted consistent with observations from the latest social science research on teenage drama.

- **Step 1:** Begin with number of topics  $T = 10$ , with hyper-parameter updating every 10 iterations
- **Step 2:** Assign themes to each of the  $T$  word clusters by examining the words under each cluster
- **Step 3:** Repeat until a satisfactory number of well-defined themes have emerged by updating the  $T$  by 10, that is  $T = T + 10$

Since LDA is an inverted generative process that assumes a distribution of topics structure in the way humans write text, the process of assigning themes to word clusters can happen only if the latent space is semantically meaningful and interpretable in the context of addressing teenage drama. Hence, we asked a group of 3 people to examine each cluster of words and use a subset of them to create sentence(s) that denotes the semantic theme that they thought was best for the cluster. The participants were encouraged to use few other words apart from the cluster of words the purpose of creating sentence(s) from each of the clusters to denote a theme.

For example, consider the following word cluster that emerged from the LDA model:

```
bf trust ive cheated times months break yrs past issues  
caught cheat hasnt bi fone multiple numbers forgive  
touch.
```

For the aforementioned cluster of words, the sentences constructed were as follows:

```
I caught my boyfriend cheating on me
```

```
I found her phone and noticed she had a lot  
of messages from numbers on her phone.  
They were from boys in my school.
```

After the above exercise, each of the participants was asked to assign a theme for the sentences. For the example, the theme was 'cheating and trust issues'. This loop of interpreting the LDA results was done until the number of topics  $T$  reached 30.

### 5.3.2 EVALUATION OF THE ASSIGNMENT

The inter-rater agreement or Cohen's kappa values for theme assignment for all the 3 coders was  $> 0.6$  for 25 out of the 30 topics generated. Of the remaining topics, clustered shorthand notations such as 'lol, idk, rofl' etc and four clusters did not yield anything of semantic value.



Theme	% of stories	#	Theme	% of stories
Using naked pictures of girlfriend or boyfriend	7.6%	14	Hookups	3.8%
Duration of a relationship/dating	5.0%	16	Shorthand notations (not a theme)	3.8%
Bullying on social media	4.9%	17	Age & dating	3.7%
Bullying connected with appearance	4.9%	18	One-sided relationships	3.5%
High school & college drama	4.7%	19	Communication gaps & misunderstandings	3.4%
Bullying on email & cell-phone	4.6%	20	Hanging out with friends	3.3%
Involving parents, siblings or spouses	4.4%	21	Jealousy	3.2%
Feeling scared, threatened or worried	4.3%	22	Talking negatively behind one's back	3.2%
Sexual acts, pregnancy	4.1%	23	Anguish & depression	3.1%
Inability to express how you feel	4.0%	24	Ending a relationship	3.1%
First dates & emotions	3.9%	25	Long-term relationships under duress	3.0%
Falling in love	3.8%	26	Post-breakup issues	2.9%
Cheating & trust issues	3.8%			

**Figure 5.3.2:** Table showing the distribution of themes (theme that gets the most probability mass in a story is the most dominant theme for that story) for the 5500 stories in the MTV corpus

### 5.3.3 STORY THEME DISTRIBUTIONS

Based on the LDA output, each story gets a probability mass for each of the 30 themes. We make this into a visualization to show the distribution of the themes present in the corpus. The most prevalent theme present in the corpus was "uploading of naked or nude pictures by boyfriend or girlfriend", with advice sought on "duration of relationships" followed by "bullying on social media" and "bullying connected with appearance". This is well documented in the social science literature investigating the problem of bullying, especially the fact that most bullying with regard to physical appearance tends to emanate from girls bullying girls. [44].

### 5.3.4 CO-OCCURRING THEMES

As previously explained, it becomes important to understand a story from a multiple thematic perspective. We investigate the top 3 themes that each story gets, to see what other themes co-occur with it. By investigating the co-occurrence of a theme in a story, a moderator or the interface could use it for targeted help. For

"okay so i had this boyfriend, and i kissed another guy, but i told him. and we broke up. then months later we got back together and his friend came up to me and kissed me. and then her told my boyfriend, so he broke up with me. but i want him back!"

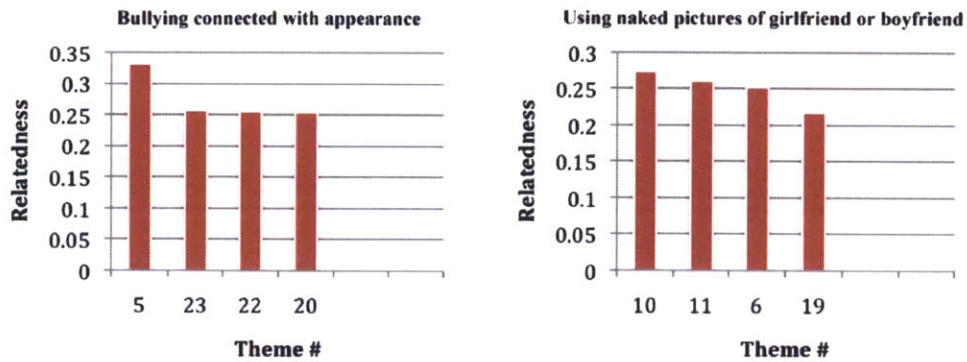


**Figure 5.3.3:** Thematic breakdown of a story whose dominant themes include duration of a relationship and post-breakup issues the top 2 dominant themes. As one can see, the most accurate themes get assigned the most probability mass by the LDA model.

example, most stories with the co-occurring themes 'Breakups, anguish and depression' and 'Feeling scared, threatened and worried' are by girls. By connecting the meta-data about the individual's gender with the co-occurrence of themes, moderators can point the individual for advice specifically tailored for girls dealing with depression connected with breakups, which might be better than generalized advice on depression. The relatedness score to calculate co-occurring stories across the corpus was done as follows for each theme  $A$ : let  $C_{(B,A)}$  be the count of how many times another theme  $B$  co-occurs with  $A$  amongst the top 3 themes for all stories. Let  $C_{(A)}$  denote the total number of times  $A$  occurs in the top 3 themes for all stories. Then the relatedness  $R_{(B,A)}$  is defined by:

$$R_{(B,A)} = \frac{C_{(B,A)}}{C_{(A)}}$$

Figure 5.3.3 shows the top 5 themes that were most related to the themes 'Breakups, anguish and depression' and 'Uploading naked pictures by boyfriend / girlfriend' respectively. As seen in the picture, bullying connected with appear-



**Figure 5.3.4:** Top 4 themes that are most related with a given theme. The relatedness is calculated by how many times another theme co-occurred with the current theme amongst the top 3 themes for all stories. That count was divided by the total number of times the given theme occurred in the top 3 themes for all stories.

ance is the most related topic connected with 'Breakups, anguish and depression'. Since the overwhelming gender talking about breakups in an anguished way were girls, we can begin to connect research from psychiatry that associates appearance as a very important aspect of female adolescent psyche [44]. For the theme of "uploading naked pictures of boyfriend or girlfriend", a related theme is "first dates and emotions" and "bullying via email and cell phones". We wish to state that we are not validating any stereotypes here, but merely connecting the dots from disparate fields of study for the problem of cyber-bullying that might lead to more targeted help for those teenagers that need help the most.

#### 5.4 THEMATIC BREAKDOWN OF A STORY

Another useful tool to understand each story better would be to visualize the thematic breakdown of each story as given by the LDA. We build an interface that allows aggregate views of all the themes, drill-down into a particular theme, as well

as view details of a single story. Such a breakdown would be useful to a moderator to not only analyze a given story, but also allow visualizations to drill-down into particular subsets of themes, moving seamlessly between views to better understand the dynamics of the stories and to set moderation policies based on current dynamics. For example, stories of teenage girls that have the first two dominant themes as "breakups, anguish and depression" and "feeling scared, threatened or worried" may merit further scrutiny given the serious nature of previous stories with the same dominant themes.

### 5.5 THEME-BASED STORY MATCHING

Research from adolescent psychiatry suggests that distressed teenagers can be helped with targeted, in-context and specific advice about their situation [4]. When a distressed teenager comes to a help site with his or her story, it would be apt to show the individual a similar story encountered by another teenager with the same experiences. What this calls for is story matching based on a distribution of themes rather than by singular labels such as positive or negative. By matching stories based on themes, it can induce a level of reflection on the part of the distressed individual that can cause both self-awareness and a feeling that one is not alone in their plight. Knowing that there are other teenagers who have experienced similar ordeals can go a long way in helping the individual deal with distress. Let  $S$  denote the set of existing stories. We use the following approach to match new stories to existing ones using the following approach:

- **Step 1:** Apply the LDA model to a new story  $A$  to get a probability distribution  $\mathbf{x}$  for  $T$  themes.

- **Step 2:** Select a subset of previously submitted stories  $B \subseteq S$  where the top 5 themes  $T_1$  through  $T_5$  match the top 5 themes of the new story in the same order.
- **Step 3:** Let  $y$  denote the probability distribution for each story in the above subset  $B$ . The degree of similarity of the new story to each story in subset  $B$  is gauged by using the Kullback-Liebler asymmetric divergence metric as follows:

$$KL(x, y) = \sum_{i=1}^T x_i \log \frac{x_i}{y_i}$$

- **Step 4:** Sort the stories in  $B$  in ascending order of their KL divergence. The story with the least KL divergence is most likely to be similar to the new story  $A$ .

## 5.6 EVALUATION

We design an evaluation protocol to test the two crucial aspects of this work - a) the effectiveness of our algorithmic approach of story matching in relation to a conventional technique's baseline, and b) the effectiveness of showing the matched stories to help reflective thinking in distressed teenagers to feel that they're not alone in their plight. We also provide a qualitative evaluation of the usefulness of identifying the distribution of themes by two community moderators of two popular teenage community websites respectively.

- **Control:** We compare our algorithmic approach of story matching against cosine similarity of tf-idf vectors for the new and old stories as control.

- **Participant selection:** We selected a total of 12 participants of which 8 were female. 5 of them were teenagers at the time of the study. 2 of the participants were teachers at a local public school, while 3 were graduate students working with young children. 2 of the participants were graduate students researching machine learning and human-computer interaction respectively.
- **User-study protocol:** The participants were each subjected to an online user-study as follows. First, each participant was asked to view the MTV Athinline website to familiarize themselves with a) the kind of stories and b) the type of linguistic styles employed by teenagers on the site. Second, each participant was asked to enter a new story using a 250 word limit (as in the case of Athinline) on any relevant themes as they deemed appropriate. 5 matched stories of which 3 were retrieved using our algorithmic approach and 2 using the control approach were shown to the participant after each new story.

Each matched story was evaluated with respect to two questions **Q1**: if the retrieved stories matched the story entered by participants in having similar themes and **Q2**: if they could imagine a distressed teenager feeling a little better if they were shown the matched stories - that they were not alone in their plight. Each participant was asked to write a minimum of 3 new stories, thereby evaluating a total of 15 matches. A total of 38 new stories were entered, with a total of 190 matches.

### 5.6.1 RESULTS AND DISCUSSION

Results show strong performance for story matching using our LDA and KL-divergence approach of extracting themes and matching new stories to old ones. Our approach fared better for both Q1 and Q2 against a simple cosine similarity using tf-idf vectors for the story matching.

An error analysis showed that new stories for which the matches were rated as *Strongly Agree* had very clear themes with linguistic styles very similar to the stories in the corpus. Those new stories for which the matched stories that received a *Strongly disagree* or *Disagree* vote did not have clear themes or used a vocabulary that wasn't common in the corpus. For example, consider the following new story: "I wanted to be in the school dance team, but I was not accepted. I think its cause the captains don't like my bf." The above story is an example of a story that didn't have a clear theme relative to those extracted from the corpus. This best match for this story was "a guy at my school let me flirt with him and he knew i liked him and know he wont even talk to me at all cuz we have no classes together is he over the line or not" . Though the algorithm did produce a coarse level match with respect to school and liking a male, the story still received a *Strongly disagree* vote. This calls for a deeper level of reasoning for fine-grain story matching.

At the end of the user-study, participants were positive about an overall feedback, with comments such as "Even when it missed some things, I could see why it was trying to go that way" and "It was really neat stuff". Two moderators of two teenage community-based social networks viewed the distribution of themes from the corpus extremely positively. The moderators likened the distribution of themes, even on a *flagged* set of instances from their websites, as key to under-

	% Strongly Agree		% Agree		% Neither agree nor disagree		% Disagree		% Strongly disagree	
	LDA+	Control	LDA+	Control	LDA+	Control	LDA+	Control	LDA+	Control
Q1	45.0%	0%	22.1%	3%	0%	0%	17%	51%	15.9%	46%
Q2	35.3%	0%	23.0%	8.3%	13.5%	25.6%	13.2%	35.1%	15%	31%

New Story	Matched story
I like wearing makeup. I dont want to be teased for my weight. Some girls hurt me that no one will like me at all. I don't want to be called names about my looks!	Growing up I havent been the skinnest girls. When I first started Middle School people would always make fun of my wieght. Because of that I would always tell my self that I was fat and ugly and i really took it to heart. And it still follows me.

**Figure 5.6.1:** Evaluation table showing the user-study responses to questions Q1 ( The themes of the story presented matched the themes of the story I wrote ) and Q2 ( After reading the presented story, I can imagine that someone in a similar situation would not feel alone ) .An example from our user-study of a new story and it s best match as measured by asymmetric KL-divergence. Bullying associated with physical appearance is overwhelmingly female in the corpus with 'High school & college drama' as strongly related topic

standing the dynamics of negative behavior on their websites.

We concede that the best evaluation of this work would be to test it on a live on the MTV AThinline website, with in-situ distressed teenagers using the interface without knowing the story matching that is happening in the background. In the next chapter, we discuss the strengths and limitations of the methods used in this thesis. We also provide a discussion of alternate ways of framing of the problem of cyber-bullying.

-----



*Commonsense is the realised sense of proportion*

Mahatma Gandhi

# 6

## Conclusion

In this thesis, we have explored three classes of algorithms for the purpose of interaction analysis of teenagers on social media with respect to the phenomenon of cyber-bullying. We began by exploring supervised learning to detection blatant forms of abuse. We highlighted the limitations of supervised learning in addressing more subtle forms of cyber-bullying, which led us to commonsense reasoning models for a deeper level of reasoning. We then proceeded to tackle address a new framing of the problem of cyber-bullying - to address teenage drama and its mani-

	Representation	Inference	Learning
Supervised learning	Weak	Weak	Strong
Commonsense Reasoning	Modest	Modest	Poor
Probabilistic Topic Models	Strong	Strong	Modest

**Table 6.1.1:** Each class of algorithms have various strengths and weaknesses, especially with regard to building practical systems for interaction analysis.

festations by using probabilistic topic models. While the problem formulation and rationale for employing each of the above methods have been addressed in prior chapters, it is worth to examine these methods from a broader perspective.

In this chapter, we provide a comparative analysis of each class of algorithms used in this work. We discuss the many simplifying assumptions we make in this work and also discuss alternative ways of framing the problem. We highlight the main contributions in this work and discuss challenges in deploying this work in the real world.

## 6.1 COMPARISON: THREE PILLARS OF A ROBUST AI ARCHITECTURE

Critical thinkers in the fields of machine learning and natural language processing such as Kevin Murphy [28] and Daphne Koller [21] speak of the three essential pillars of a robust architecture of intelligent systems, namely **representation, inference and learning**. It becomes important to talk about the three classes of methods in this work with respect to these three pillars.

Supervised learning methods allow representation of features that are essentially treated as mutually exclusive at their worst, with modest interaction at their best. For example, during the feature design process of the supervised models, one always observes the training data and adds features usually without thinking too deeply about the interaction of the features with each others. Classifiers like logis-

tic regression allow for some interaction between features to be captured, but the task of feature design is essentially the same. Very little inference of consequence is possible with supervised learning methods, limited to deciphering a reasonable separation of classes. However, the learning in supervised learning methods is robust, especially with online versions, which is a huge advantage when it comes to practical deployment. The success of spam filters underlines the effectiveness of supervised learning methods in addressing explicit tasks. As expected, the supervised learning methods used in this work perform well in detecting explicit forms of abuse.

The ability of commonsense reasoning models for a representation of data that allows for a level of reasoning beyond traditional supervised learning methods is an advantage. The ability for commonsense reasoning models to blend and merge two sets of knowledge sources that have only a thin slice of intersection between them is also an advantage. However, learning in commonsense reasoning models is rather poor. The necessity of hand-crafted assertions in building a knowledge base hampers the growth of the knowledge base. The field of parsing in natural language processing is nowhere near the advent of using a collection of domain-specific documents to automatically fetch a list of assertions. This is a limitation that becomes especially stark when it comes to the actual deployment of commonsense models towards designing a cyber-bully detector.

Probabilistic topic models, like all graphical models are relatively the most robust class of algorithms that allow for effective representation and inference. For example, the representation of LDA makes it easy to think about the dataset as a whole, rather than pigeon-holing certain classes in the data. They also offer a way to leap over the task of manual hand-annotations and as shown in chapter

5, address practical problems such as diglossia and hedging in ways that neither supervised learning methods or commonsense reasoning methods can very easily. The inference part of probabilistic topic models are a double edged sword. While extracting latent topics that are latent in the corpus is definitely a strength, the task of determining the semantic value of an extracted cluster of words requires human involvement. Online versions of topic models that can act on huge quantities of data make the learning aspect of these models modestly strong. At the time of writing this thesis, MTV's *www.athinline.org* was in the process of deploying an online version of LDA as explained in chapter 5.

## 6.2 SIMPLIFYING ASSUMPTIONS AND LIMITATIONS

There is a school of thought within sociolinguists that instances of cyber-bullying are too complex for any kind of automated detection. For example, a line of work championed by Marion Underood [46] examines relational aggression among girls that are so subtle that it is hard for bystanders to detect but nonetheless pose significant harm and danger.

In this thesis we make simplifying assumptions for each class of algorithms as well as take the view that most acts of cyber-bullying entailing textual interaction can be detected. While it can argued that assuming that explicit forms of bullying involves a certain set of sensitive topics might be a simplification, it is nonetheless necessary for practical work. The simplifying assumptions in this work have been made with care and with the purpose of developing a practical, tractable system.

We also emphasize that the phenomenon of cyber-bullying extends beyond textual interaction to a whole host of acts from uploading damaging content to impersonation of identities and hacking of personal information. This work is limited

only to the problem of textual cyber-bullying.

### 6.3 CONTRIBUTIONS

The main contributions of this thesis is three fold. It is essential to understand the practical strengths and weaknesses of machine learning algorithms with respect to the kinds and distribution of data that they can act on. What the distribution of your data looks like is informed by a reasonable level of domain knowledge. This work combines domain expertise from relevant fields with a view of helping in problem decomposition and parameterization of each class of models.

- **Problem decomposition and a society of models** - The problem of textual cyber-bullying can be viewed from three perspectives, namely explicit and blatant forms of abuse, employment of subtlety for indirect forms of abuse, and mixture of complex issues behind personal recollection of distressing stories. Most of the explicit forms of abuse generate patterns that are ripe for supervised learning. Indirect ways of insulting another person, although more difficult to detect than explicit forms of abuse, lend themselves to commonsense reasoning. Examining distributions of complex issues surrounding personal stories of bullying are very apt for probabilistic topical models. In this thesis, we follow the aforementioned decomposition of the problem and employ the use of relevant classes of algorithms. A key technical contribution of this work is for a comparative analysis of each of the above three classes of algorithms via-a-vis interaction analysis.
- **Computational empathy** It is not enough to merely detect candidate instances of cyber-bullying. A key contribution of this thesis the use of prac-

tical, large scale computational models that powers a reflective user interaction paradigm. In this thesis, we examine the critical aspects of approaching the building of statistical models that reap their benefit only when used as one part of a two part paradigm also involving novel interaction paradigms.

- **Real world deployment** A key contribution of this work is the real world deployment of the above paradigm on a major youth network to help actual teenagers in distress. While the academic publications that have emanated from this work is important, the stakeholders of this project also believe that real-world deployment is an integral aspect. Indeed, the deployment of the the above models is probably the first and only deployment of computational models to help a distressed group of individuals, namely teenagers.

The approaches followed in this thesis raises several interesting theoretical and practical questions about applied artificial intelligence. In the next chapter, we discuss the future directions of this work and areas where the paradigms followed in this thesis might apply.

-----

*Imagination is more important than knowledge. For knowledge is limited, whereas imagination embraces the entire world, stimulating progress, giving birth to evolution.*

Albert Einstein

# 7

## Future Work

There *was* a vision among many researchers in the field of artificial intelligence of replicating the human brain. Grandiose movements within the field which rejected statistical methods in search of semantic architectures are well known. Others have adopted a completely statistical approach, leading to dozens of empirical research venues in the fields of machine learning and natural language processing. The main author of this thesis does not ascribe loyalties to either school of thought. While there is a place for every one in the research ecosystem, there is a

growing realization that applied machine learning and natural language processing are a very different beasts than the theoretical side. This thesis raises some interesting questions from both a theoretical as well as a practical perspective. In this chapter, we provide an overview of the future directions that the approaches used in this thesis might take.

### 7.1 A SOCIETY OF MODELS WORKING IN TANDEM TO HELP MODERATORS OF SOCIAL NETWORKING WEBSITES

In this thesis we decompose the problem of cyber-bullying into sub parts and describe classes of algorithms that are ripe for each sub part. Combining the approaches used in this work - namely supervised learning, commonsense reasoning and unsupervised topic modeling to work in tandem would be apt. One could imagine these three approaches as features to a larger logistic regression model that is used to assign severity values to candidate instances of cyber-bullying. Social networking websites are flooded with large amounts of 'user-flagged' instances that are too many for a moderation team that are in the order of millions of instances. Such a logistic regression model could work to algorithmically prioritize the most serious cases of cyber-bullying from this user-flagged set of instances.

A moderation interface that is built to equip the moderator to add temporally important features (certain celebrity memes are used for bullying that are ephemeral in nature) and which provides an affordance for moderator feedback to reinforce the underlying algorithm (to the level of the sub-part, e.g., unsupervised topic models) raises interesting questions from a human computer interaction perspective, because the interface should be more than just user friendly - it must be designed to elicit proper feedback for algorithm as well. This



intersection between applied machine learning and interaction paradigms is an under-investigated area in computer science.

## 7.2 SOCIAL NONPARAMETRIC AND PARAMETRIC MACHINE LEARNING

The state of the art classes of nonparametric and parametric algorithms are currently devoid of human involvement in their loop. In this work, we developed a novel approach of embedding human input and judgment in the loop of latent dirichlet allocation models to extract hidden themes present in a corpus of stories by distressed teenagers.

Given the effectiveness of that approach, we are led to believe very strongly that there is a middle ground between pure machine learning and pure crowd sourcing that the concept of *human in the loop* computation can extend to entire classes of algorithms in machine learning and natural language processing. There seems to be a very powerful middle ground in between pure machine learning and pure crowdsourcing. Embedding human input in the inference loop of LDA (sampling or variational) for example, is an interesting area to examine. Moreover such a paradigm allows us to statistically learn more about each human performing the judgement task, which is also very interesting to look at.

## 7.3 AN ONTOLOGY OF AUTISM IN DEVELOPING CHILDREN

Many websites and help forums allow parents to talk about symptoms and behaviors of children that may have a propensity for autism. The writers of this thesis believe that autism is an umbrella term that is used for a wide spectrum of conditions ranging from dietary and gastrointestinal problems to sensory integration difficulties to social interaction difficulties. It is strange and almost shocking that

there appears no ontology of what the autism spectrum might look like.

Adopting the same paradigm used for the MTV deployment on websites and forums for parents of children with autism will serve a dual purpose: to provide an empathetic affordance for concerned parents and to build an ontology of the various traits and conditions that they talk about. This kind of an ontology might be of value for researchers in the field of autism.

#### 7.4 EMPATHETIC COMPUTING

Current paradigms in human-computer interaction revolve around understanding the end-user's needs. Contextual inquiry and contextual design, participatory design, heuristic evaluations etc., are all paradigms designed to elicit the user's needs. However, there are users such as distressed teens and anxious parents, hypochondriacs etc., where the primary focus is not what the user needs, but deals with the user's affective state.

Employing machine learning and interaction paradigms to provide an empathetic affordance to users is a research area that currently does not exist in the the IUI community. A future direction of this thesis would be to lay broad-based principles of what that kind of a paradigm should involve.

In short, we remain very excited about the future directions that this thesis has both directly and inadvertently sparked!

-----

## References

- [1] *The handbook of sociolinguistics*. Blackwell Publishers, Oxford, UK Cambridge, Mass, 1998. ISBN 9780631211938.
- [2] *Cyberbullying prevention and response expert perspectives*. Routledge, New York, 2012. ISBN 0415892376.
- [3] Rebecca Ang and Dion Goh. Cyberbullying among adolescents: The role of affective and cognitive empathy, and gender. *Child Psychiatry and Human Development*, 41:387–397, 2010. ISSN 0009-398X. URL <http://dx.doi.org/10.1007/s10578-010-0176-3>. 10.1007/s10578-010-0176-3.
- [4] Rebecca Ang and Dion Goh. Cyberbullying among adolescents: The role of affective and cognitive empathy, and gender. *Child Psychiatry & Human Development*, 41:387–397, 2010. ISSN 0009-398X. URL <http://dx.doi.org/10.1007/s10578-010-0176-3>. 10.1007/s10578-010-0176-3.
- [5] Erik Cambria, Amir Hussain, Catherine Havasi, and Chris Eckl. Sentic computing: Exploitation of common sense for the development of emotion-sensitive systems. In Anna Esposito, Nick Campbell, Carl Vogel, Amir Hussain, and Anton Nijholt, editors, *Development of Multimodal Interfaces: Active Listening and Synchrony*, volume 5967 of *Lecture Notes in Computer Science*, pages 148–156. Springer Berlin / Heidelberg, 2010. ISBN 978-3-642-12396-2.
- [6] Erik Cambria, Robert Speer, Catherine Havasi, and Amir Hussain. Senticnet: A publicly available semantic resource for opinion mining, 2010. URL <https://www.aaai.org/ocs/index.php/FSS/FSS10/paper/view/2216>.
- [7] Jean Carletta. Assessing agreement on classification tasks: the kappa statistic. *Comput. Linguist.*, 22(2):249–254, June 1996. ISSN 0891-2017. URL <http://dl.acm.org/citation.cfm?id=230386.230390>.
- [8] Bob Carr. "anti-bullying" law nonsense. 2011. URL <http://blogs.ajc.com/bob-barr-blog/2011/09/02/anti-bullying-law-nonsense/>.

- [9] Si-Chi Chin, W. Nick Street, Padmini Srinivasan, and David Eichmann. Detecting wikipedia vandalism with active learning and statistical language models. In *Proceedings of the 4th workshop on Information credibility, WICOW '10*, pages 3–10, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-940-4. doi: 10.1145/1772938.1772942. URL <http://doi.acm.org/10.1145/1772938.1772942>.
- [10] Morten H. Christiansen and Simon Kirby. Language evolution: consensus and controversies. *Trends in Cognitive Sciences*, 7(7):300 – 307, 2003. ISSN 1364-6613. doi: 10.1016/S1364-6613(03)00136-0. URL <http://www.sciencedirect.com/science/article/pii/S1364661303001360>.
- [11] William W. Cohen and Yoram Singer. A simple, fast, and effective rule learner. In *Proceedings of the sixteenth national conference on Artificial intelligence and the eleventh Innovative applications of artificial intelligence conference innovative applications of artificial intelligence, AAAI '99/IAAI '99*, pages 335–342, Menlo Park, CA, USA, 1999. American Association for Artificial Intelligence. ISBN 0-262-51106-1. URL <http://dl.acm.org/citation.cfm?id=315149.315320>.
- [12] Corinna Cortes and Vladimir Vapnik. Support-vector networks. In *Machine Learning*, pages 273–297, 1995.
- [13] Karthik Dinakar, Roi Reichart, and Henry Lieberman. Modeling the detection of textual cyberbullying, 2011. URL <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/3841>.
- [14] Evgeniy Gavrilovich and Shaul Markovitch. Text categorization with many redundant features: using aggressive feature selection to make svms competitive with c4.5. In *Proceedings of the twenty-first international conference on Machine learning, ICML '04*, pages 41–, New York, NY, USA, 2004. ACM. ISBN 1-58113-838-5. doi: 10.1145/1015330.1015388. URL <http://doi.acm.org/10.1145/1015330.1015388>.
- [15] Ya Gao and Shiliang Sun. An empirical evaluation of linear and nonlinear kernels for text classification using support vector machines. In *Fuzzy Systems and Knowledge Discovery (FSKD), 2010 Seventh International Conference on*, volume 4, pages 1502 –1505, aug. 2010. doi: 10.1109/FSKD.2010.5569327.
- [16] Teresa Gonçalves and Paulo Quaresma. A preliminary approach to the multi-label classification problem of portuguese juridical documents. In Fernando Pires and Salvador Abreu, editors, *Progress in Artificial Intelligence*, volume 2902 of *Lecture Notes in Computer Science*, pages 435–444. Springer Berlin / Heidelberg, 2003. ISBN 978-3-540-20589-0.

- [17] Edel Greevy and Alan F. Smeaton. Classifying racist texts using a support vector machine. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 468–469, New York, NY, USA, 2004. ACM. ISBN 1-58113-881-4. doi: 10.1145/1008992.1009074. URL <http://doi.acm.org/10.1145/1008992.1009074>.
- [18] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November 2009. ISSN 1931-0145. doi: 10.1145/1656274.1656278. URL <http://doi.acm.org/10.1145/1656274.1656278>.
- [19] Catherine Havasi, Robert Speer, James Pustejovsky, and Henry Lieberman. Digital intuition: Applying common sense using dimensionality reduction. *IEEE Intelligent Systems*, 24(4):24–35, July 2009. ISSN 1541-1672. doi: 10.1109/MIS.2009.72. URL <http://dx.doi.org/10.1109/MIS.2009.72>.
- [20] Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features, 1998.
- [21] Daphne Koller. *Probabilistic graphical models : principles and techniques*. MIT Press, Cambridge, MA, 2009. ISBN 0262013193.
- [22] H. Liu and P. Singh. Conceptnet - a practical commonsense reasoning toolkit. *BT Technology Journal*, 22(4):211–226, October 2004. ISSN 1358-3948. doi: 10.1023/B:BTTJ.0000047600.45421.6d. URL <http://dx.doi.org/10.1023/B:BTTJ.0000047600.45421.6d>.
- [23] Christopher Maag. A hoax turned fatal draws anger but no charges. 2007. URL [http://www.nytimes.com/2007/11/28/us/28hoax.html?\\_r=2&oref=slogin](http://www.nytimes.com/2007/11/28/us/28hoax.html?_r=2&oref=slogin).
- [24] Martin. *Working with discourse : meaning beyond the clause*. Continuum, London New York, 2003. ISBN 0826455085.
- [25] Alice E. Marwick and danah Boyd. The Drama! Teen Conflict, Gossip, and Bullying in Networked Publics. *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society, September 2011, 2011*.
- [26] Elijah Mayfield and Carolyn Penstein Rosé. Recognizing authority in dialogue with an integer linear programming constrained model. In *Proceedings*

- of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1018–1026, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-932432-87-9. URL <http://dl.acm.org/citation.cfm?id=2002472.2002600>.
- [27] Faye Mishna, Michael Saini, and Steven Solomon. Ongoing and online: Children and youth's perceptions of cyber bullying. *Children and Youth Services Review*, 31(12):1222 – 1228, 2009. ISSN 0190-7409. doi: 10.1016/j.childyouth.2009.05.004. URL <http://www.sciencedirect.com/science/article/pii/S0190740909001200>.
- [28] Kevin Patrick Murphy. *Dynamic bayesian networks: Representation, inference and learning*, 2002.
- [29] Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. Compositionality principle in recognition of fine-grained emotions from text, 2009. URL <http://aaai.org/ocs/index.php/ICWSM/09/paper/view/197/525>.
- [30] DAN OLWEUS. Bullying at school: Knowledge base and an effective intervention program. *Annals of the New York Academy of Sciences*, 794(1): 265–276, 1996. ISSN 1749-6632. doi: 10.1111/j.1749-6632.1996.tb32527.x. URL <http://dx.doi.org/10.1111/j.1749-6632.1996.tb32527.x>.
- [31] Andrew Ortony, Gerald L. Clore, and Mark A. Foss. The referential structure of the affective lexicon. *Cognitive Science*, 11(3):341–364, 1987. ISSN 1551-6709. doi: 10.1207/s15516709cog1103\_4. URL [http://dx.doi.org/10.1207/s15516709cog1103\\_4](http://dx.doi.org/10.1207/s15516709cog1103_4).
- [32] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 309–319, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-932432-87-9. URL <http://dl.acm.org/citation.cfm?id=2002472.2002512>.
- [33] Bo Pang. *Opinion mining and sentiment analysis*. Now Publishers, Hanover, MA, 2008. ISBN 1601981503.
- [34] JW Pennebaker, MR Mehl, and KG Niederhoffer. Psychological aspects of natural language use: Our words, our selves. *ANNUAL REVIEW OF PSYCHOLOGY*, 54:547–577, 2003. ISSN 0066-4308. doi: {10.1146/annurev.psych.54.101601.145041}.

- [35] Ian Porteous, David Newman, Alexander Ihler, Arthur Asuncion, Padhraic Smyth, and Max Welling. Fast collapsed gibbs sampling for latent dirichlet allocation. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '08*, pages 569–577, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-193-4. doi: 10.1145/1401890.1401960. URL <http://doi.acm.org/10.1145/1401890.1401960>.
- [36] J Quinlan. *C4.5 : programs for machine learning*. Morgan Kaufmann Publishers, San Mateo, Calif, 1993. ISBN 1558602380.
- [37] Monica Rogati and Yiming Yang. High-performing feature selection for text classification. In *Proceedings of the eleventh international conference on Information and knowledge management, CIKM '02*, pages 659–661, New York, NY, USA, 2002. ACM. ISBN 1-58113-492-4. doi: 10.1145/584792.584911. URL <http://doi.acm.org/10.1145/584792.584911>.
- [38] M. Sasaki and K. Kita. Rule-based text categorization using hierarchical categories. In *Systems, Man, and Cybernetics, 1998. 1998 IEEE International Conference on*, volume 3, pages 2827–2830 vol.3, oct 1998. doi: 10.1109/ICSMC.1998.725090.
- [39] Push Singh, Thomas Lin, Erik T. Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. Open mind common sense: Knowledge acquisition from the general public. In *On the Move to Meaningful Internet Systems, 2002 - DOA/CoopIS/ODBASE 2002 Confederated International Conferences DOA, CoopIS and ODBASE 2002*, pages 1223–1237, London, UK, UK, 2002. Springer-Verlag. ISBN 3-540-00106-9. URL <http://dl.acm.org/citation.cfm?id=646748.701499>.
- [40] Andre Sourander, Anat Brunstein Klomek, Maria Ikonen, Jarna Lindroos, Terhi Luntamo, Merja Koskelainen, Terja Ristkari, and Hans Helenius. Psychosocial risk factors associated with cyberbullying among adolescents: A population-based study. *Arch Gen Psychiatry*, 67(7):720–728, 2010. doi: 10.1001/archgenpsychiatry.2010.79. URL <http://archpsyc.ama-assn.org/cgi/content/abstract/67/7/720>.
- [41] Andre Sourander, Anat Brunstein Klomek, Maria Ikonen, Jarna Lindroos, Terhi Luntamo, Merja Koskelainen, Terja Ristkari, and Hans Helenius. Psychosocial risk factors associated with cyberbullying among adolescents: A population-based study. *Arch Gen Psychiatry*, 67(7):720–728, 2010. doi: 10.1001/archgenpsychiatry.2010.79. URL <http://archpsyc.ama-assn.org/cgi/content/abstract/67/7/720>.

- [42] Robert Speer, Catherine Havasi, and Henry Lieberman. Analogyspace: reducing the dimensionality of common sense knowledge. In *Proceedings of the 23rd national conference on Artificial intelligence - Volume 1, AAAI'08*, pages 548–553. AAAI Press, 2008. ISBN 978-1-57735-368-3. URL <http://dl.acm.org/citation.cfm?id=1619995.1620084>.
- [43] M. Steyvers and T. Griffiths. *Latent Semantic Analysis: A Road to Meaning*, chapter Probabilistic topic models. Laurence Erlbaum, 2007.
- [44] Marika Tiggemann and Jessica Miller. The internet and adolescent girls weight satisfaction and drive for thinness. *Sex Roles*, 63:79–90, 2010. ISSN 0360-0025. URL <http://dx.doi.org/10.1007/s11199-010-9789-z>. 10.1007/s11199-010-9789-z.
- [45] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. In *JOURNAL OF MACHINE LEARNING RESEARCH*, pages 999–1006, 2000.
- [46] Marion Underwood. *Social aggression among girls*. Guilford Press, New York, 2003. ISBN 1572308656.
- [47] Marco Vala, Pedro Sequeira, Ana Paiva, and Ruth Aylett. Fearnot! demo: a virtual environment with synthetic characters to help bullying. In *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems, AAMAS '07*, pages 271:1–271:2, New York, NY, USA, 2007. ACM. ISBN 978-81-904262-7-5. doi: 10.1145/1329125.1329452. URL <http://doi.acm.org/10.1145/1329125.1329452>.
- [48] S.M. Vieira, U. Kaymak, and J.M.C. Sousa. Cohen's kappa coefficient as a performance measure for feature selection. In *Fuzzy Systems (FUZZ), 2010 IEEE International Conference on*, pages 1–8, july 2010. doi: 10.1109/FUZZY.2010.5584447.
- [49] Ronald Wardhaugh. *An introduction to sociolinguistics*. Blackwell Pub, Malden, MA, 2002. ISBN 0631225404.
- [50] Janis Wolak, Kimberly J. Mitchell, and David Finkelhor. Does online harassment constitute bullying? an exploration of online harassment by known peers and online-only contacts. *Journal of Adolescent Health*, 41(6, Supplement):S51–S58, 2007. ISSN 1054-139X. doi: 10.1016/j.jadohealth.2007.08.019. URL <http://www.sciencedirect.com/science/article/pii/S1054139X07003631>. <ce:title>Youth Violence and Electronic Media: Similar Behaviors, Different Venues?</ce:title>.