UNDERSTANDING AND FIGHTING BULLYING WITH MACHINE LEARNING

by

Junming Sui

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Computer Sciences)

at the

UNIVERSITY OF WISCONSIN–MADISON

2015

Date of final oral examination: May 21, 2015

The dissertation is approved by the following members of the Final Oral Committee:
Xiaojin (Jerry) Zhu, Associate Professor, Computer Sciences
Amy Bellmore, Associate Professor, Educational Psychology
Mark Craven, Professor, Computer Sciences
Charles R. Dyer, Professor, Computer Sciences
Jude W. Shavlik, Professor, Computer Sciences

To my parents and my wife.

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor Xiaojin (Jerry) Zhu, for being a very supportive and tireless mentor. He selflessly donated a great amount of effort on training me the basic skills for my career. During my first few years working with him, he spent a lot of time with me in reading literature, identifying and analyzing problems, formulating models, debugging programs, and investigating the experiment results. He always found novel connections between different works, insightful advice on how to proceed, and interesting observations from data and results. I benefit a lot from these training and skills in my future career. To improve my English and communication skills, Jerry carefully commented on all my paper drafts, kindly corrected my pronunciations during our conversations and practice talks, encouraged and sponsored me to represent our work in many conferences and seminars, and invited to meet collaborators with diverse backgrounds. I was quite fortunate to have Jerry as my advisor.

I also want to thank my collaborators in educational psychology, Professor Amy Bellmore and her team. Our work is interdisciplinary, which requires knowledge both from machine learning and social sciences. Without their expertise and help, I could not have completed my thesis. They are always passionate and responsive. They provide me the necessary background and research methodology from social science vantage points. A few sections in my thesis directly use materials written by my co-authors to provide a precise and complete documentation of our collaborated work.

I also thank the other members of my dissertation committee, for their support, guidance and expertise to make my thesis stronger. Professor Mark Craven came to my first few practice talks on our work and provided insightful comments on the presentation and future directions. Professor Charles Dyer served on my preliminary examination committee and we worked together on mining image posted in social media. His enthusiasm for research and relentless pursuit of excellence were a great inspiration for me. Professor Jude Shavlik taught me the fundamental knowledge of machine learning, deeper understanding and hands-on

experiences on many algorithms, through his machine learning class. His slides, which have a picture on each page, are still my reference to look up these basics.

I acknowledge my collaborators and co-authors. During our collaboration, I have learned a lot from them. Professor Robert Nowak was in my preliminary examination committee and I am thankful for the experience of working with him for several projects. His solid mathematical background and critical thinking guided our projects to the fruitful directions. Professor Benjamin Recht helped us in solving the optimization problems we met in our project. Professor Benjamin Snyder was in my preliminary examination committee and provided a lot suggestion from natural language processing perspective. We also worked together with Professor Ben Liblit on expanding abbreviations in programs. Professor Chen Yu at Indiana University brought us to an interesting children word learning problem and data source. In addition to these professors, I could not have made my thesis without the help from: Aniruddha Bhargava, Benjamin Burchifiel, Angela J. Calvin, Joshua Dein, Alex Furger, Andrew Goldberg, Megan K. Hines, Hsun-Chih Huang, Kwang-Sung Jun, Christine M. Marsh, Timothy T. Rogers.

I also want to thank my fellow graduate students. It was great to interact with everyone who participated in the various seminars and reading groups, which has been an important channel for me to learn. In alphabetical orders, I want to thank: Scott Alfeld, David Andrzejewski, Manish Bansal, Bryan Gibson, Andrew Glodberg, Kwang-Sung Jun, Sang Kyun Lee, Bo Li, Fengan Li, Ji Liu, Jie Liu, Shike Mei, Nikhil Rao, Ayon Sen, Ameet Soni, Yimin Tan, Ara Vartanian, Jeremy Weiss, Jia Xu, Zhiting Xu and Wei Zhang. I want to thank the team members in educational psychology: Angela C. Calvin, Wei-Ting Chen, Rachael Hansen, Hsun-Chih Huang, Ting-Lan Ma, Felice Resnik and Ji-in You. I want to thank Bryan Gibson and Andrew Goldberg for their constant helps on proof-reading my draft and comments for practice talk. I also want to thank Alex Henna for the help on Twitter data access. In addition to those named above, I thank all my friends who have been great companions to pursue our goals together.

I also want to thank Professor Zhi-Hua Zhou, who introduced me to machine learning research and solidified my interest in pursuing research in this field. I

have benefited from the support and friendship of many other fellow LAMDA team members: Professor Yuan Jiang, Xiangnan Kong, Ju-Hua Hu, Sheng-Jun Huang, Ming Li, Nan Li, Ying-Xin Li, Yu-Feng li, Xiaolin Li, Li-Ping Liu, Xu-Ying Liu, Qi Qian, Qiao-Qiao She, Yu-Yin Sun, Wei Wang, Xinpan Xiao, Miao Xu, Xiaobing Xue, Yang Yu, De-Chuan Zhan, Daoqiang Zhang, Ming-Lin Zhang, Yin Zhang, and Yu Zhang.

I could not have done this without the love and support of my family. Studying aboard means that we could not see each other frequently, but they always support me to pursue my childhood dream, no matter how much we miss each other. They are part of what keeps me going through the challenges in graduate school. To my parents, Guanghua Xu and Xiaofu Liu, thank you for all love and support you provided to me. Thank you to my loving wife, Lingli, for the constant support and encouragement.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Bullying, in both physical and cyber worlds, has been recognized as a serious health issue among adolescents. Given its significance, scholars are charged with identifying factors that influence bullying involvement in a timely fashion. However, previous social studies of bullying are handicapped by data scarcity. The standard psychological science approach to studying bullying is to conduct personal surveys in schools. The sample size is typically in the hundreds, and these surveys are often collected only once. On the other hand, the few computational studies narrowly restrict themselves to cyberbullying, which accounts for only a small fraction of all bullying episodes.

My thesis work shows that social media, with appropriate machine learning and natural language processing techniques, can be a valuable and abundant data source for the study of bullying. Social media afford a context for cyberbullying to take place, and also provide a unique vantage point for scholars interested in studying bullying. We found that participants of a bullying episode (in either physical or cyber venues) often post social media texts about the experience. We collectively call such social media posts **bullying traces**. Bullying traces include but far exceed incidences of cyberbullying. Bullying traces are valuable, albeit fragmented and noisy, data which can be used to reconstruct the underlying episodes.

We build standard machine learning models to automatically recognize bullying traces from social media streams, to identify participants and their roles, to label the forms of bullying episodes, to determine the types of bullying traces, and to summarize the topics in bullying traces. We apply the trained models to bullying traces collected in two consecutive academic years and analyze the dynamics of bullying posts.

We also propose a new model for recovering the spatiotemporal distribution of underlying bullying activities from bullying traces. To recognize the emotions in bullying traces, we design a fast learning procedure to train a classifier without explicitly producing a conventional labeled training dataset. We propose a probabilistic model to assign individual tweets in a user's timeline into their corre-

sponding episodes.

We investigate deletion behaviors and hashtag usages in bullying traces, and correlate them with some important factors. We also identify several differences in microblogs of school-based bullying between Twitter and Weibo, and hypothesize possible explanations for these cultural differences.

To facilitate the research in bullying and social media mining, we make our dataset and software publicly available under Twitter's Terms of Service.

1 INTRODUCTION

Bullying, also called peer victimization, has been recognized as a serious national health issue by the White House (2011), the American Academy of Pediatrics (2009), and the American Psychological Association (2004). This recognition derives from the growing body of research underscoring the wide-ranging harm associated with bullying (Juvonen and Graham, 2014). Bullying is associated with psychological maladjustment (Hawker and Boulton, 2000), physical complaints (Gini and Pozzoli, 2013), and poor functioning in school (Baly et al., 2014) and at work (McTernan et al., 2013).

Given its significance, social scientists are charged with identifying factors that influence bullying involvement in a timely fashion. However, a key challenge is data acquisition and analysis. The standard psychological science approach to studying bullying is to conduct personal surveys in schools (via self, peer, and teacher reports) about general experiences of individuals as victims or perpetrators of bullying (Card and Hodges, 2008). Often these surveys are collected only once. The sample size is typically in the hundreds, and participants typically write 3 to 4 sentences about each bullying episode (Nishina and Bellmore, 2010). When studies are longitudinal, the timeline (usually once or twice a year across several years; e.g., Nylund, Bellmore, Nishina, & Graham (2007); or daily across several weeks; e.g., Nishina & Juvonen (2005)) is imposed by the researcher rather than the phenomenon. Note several limitations of the predominant approach: (1) the sample size is tiny compared to the whole population; (2) school-based experiences of children and adolescents are emphasized over other social contexts and age groups; (3) the assessment is typically done only once, or, when carried out longitudinally, researchers impose a timeline that may be invalid; and (4) the experiences of bullies and victims are examined more frequently than the experiences of other role-players.

The computational study of bullying is largely unexplored, with the exception of a few studies (Lieberman et al., 2011; Dinakar et al., 2011; Ptaszynski et al., 2010; Kontostathis et al., 2010; Bosse and Stam, 2011; Latham et al., 2010; Dinakar et al.,

2012; Macbeth et al., 2013). These works mainly aim at automatically recognizing cyberbullying or hate speech online, and preventing or reporting them once detected. They do not focus on bullying in the physical world or study bullying from a psychological perspective.

Social media are large-scale, near real-time, dynamic data sources that hold promise to enrich the study of bullying. This is due to their properties as data, but more so because social media are an important social context for youth (Lenhart et al., 2010) and adults (Duggan and Smith, 2013). Social media enhance relationships (Ellison et al., 2007) and promote life satisfaction (Oh et al., 2014), but are also a vehicle for bullying (Wang et al., 2009). A meta-analysis of youth cyberbullying research reported that cyberbullying prevalence rates typically range between 10% and 40% of participants (Kowalski et al., 2014). Cyberbullying is often studied in the same way that school-based bullying is studied, via self-reports of frequency of involvement as a perpetrator or victim. As such, it presents a unique context for studying bullying in adolescence–not only does it yield an extremely large amount of data, new data are continuously created.

My thesis work shows that social media, with appropriate machine learning and natural language processing techniques, can be a valuable and abundant data source for the study of bullying. Key to this endeavor is that interactions that take place both online and offline might be represented in social media. We have deployed some off-the-shelf techniques and designed some new models to answer many scientific questions of interest in this area.

1.1 An Introduction to Bullying

One is being bullied or victimized when he or she is exposed repeatedly over time to negative actions on the part of others (Olweus, 1993). Far-reaching and insidious sequelae of bullying include intrapersonal problems (Juvonen and Graham, 2001; Jimerson et al., 2010) and lethal school violence in the most extreme cases (Moore et al., 2003). Youth who experience peer victimization report more symptoms of depression, anxiety, loneliness, and low self-worth compared to their nonvictimized

counterparts (Bellmore et al., 2004; Biggs et al., 2010; Graham et al., 2007; Hawker and Boulton, 2000). Other research suggests that victimized youth have more physical complaints (Fekkes et al., 2006; Nishina and Juvonen, 2005; Gini and Pozzoli, 2009). Victimized youth are absent from school more often and get lower grades than nonvictimized youth (Ladd et al., 1997; Schwartz et al., 2005; Juvonen and Gross, 2008).

Bullying happens traditionally in the physical world and, recently, online as well; the latter is known as cyberbullying (Cassidy et al., 2009; Fredstrom et al., 2011; Wang et al., 2009; Vandebosch and Cleemput, 2009). Bullying usually starts in primary school, peaks in middle school, and lasts well into high school and beyond (Nansel et al., 2001; Smith et al., 1999; Cook et al., 2010). Across a national sample of students in grades 4 through 12, 38% of students reported being bullied by others and 32% reported bullying others (Vaillancourt et al., 2010).

Bullying takes multiple *forms*, most noticeably face-to-face physical (e.g., hitting), verbal (e.g., name-calling), and relational (e.g., exclusion) (Archer and Coyne, 2005; Little et al., 2003; Nylund et al., 2007). Cyberbullying reflects a venue (other than face to face contact) through which verbal and relational forms can occur.

A main reason individuals are targeted with bullying is *perceived differences*, i.e., any characteristic that makes an individual stand out differently from his or her peers. These include race, socio-economic status, gender, sexuality, physical appearance, and behavior.

Participants in a bullying episode take well-defined *roles* (see Figure 1.1). More than one person can have the same role in a bullying episode. Roles include the bully (or bullies), the victims, bystanders (who saw the event but did not intervene), defenders of the victim, assistants to the bully (who did not initiate but went along with the bully), and reinforcers (who did not directly join in with the bully but encouraged the bully by laughing, for example) (Salmivalli, 1999). This recognition that bullying involves multiple roles makes evident the broad-ranging impact of bullying; any child or adolescent is susceptible to participation in bullying, even those who are not directly involved (Janosz et al., 2008; Rivers et al., 2009).

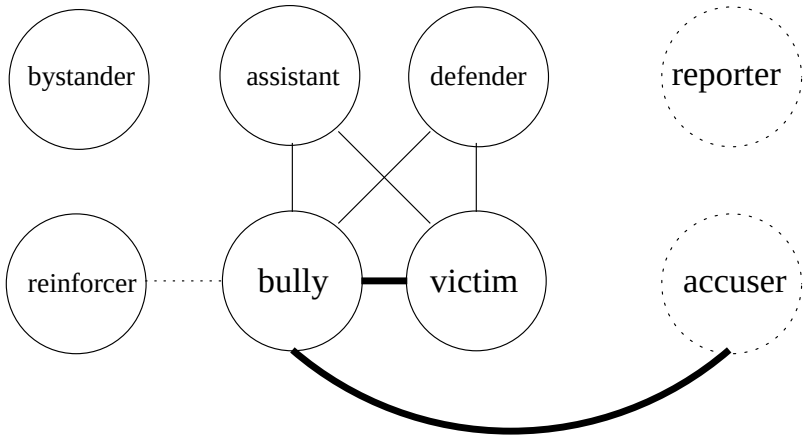


Figure 1.1: The roles in a bullying episode. Solid circles represent traditional roles in social science, while dotted circles are new roles we augmented for social media. The width of the edges represents interaction strength.

1.2 Bullying Traces in Social Media

Twitter presents the opportunity to identify online and offline bullying trends more comprehensively. By some estimates, Twitter currently produces 400 million tweets per day. Many studies use Twitter as a data source to answer social science questions (Lazer et al., 2009; Eisenstein et al., 2010; Gupte et al., 2011).

Participants of a bullying episode (in either physical or cyber venues) often post social media text about the experience. We collectively call such social media posts **bullying traces**. Bullying traces include but far exceed incidences of cyberbullying. Most of them are in fact *responses* to a bullying experience – the actual attack is hidden from view. Among the bullying traces we analyzed, only 0.76% of bullying traces were direct cyberbullying attacks (see Section 2.4 for more details). Bullying traces are valuable, albeit fragmented and noisy, data which we can use to piece together the underlying episodes.

Here are some examples of bullying traces:

- Reporting a bullying episode: *"some tweens got violent on the n train, the one boy got off after blows 2 the chest... Saw him cryin as he walkd away :(bullying not*

cool"

- Accusing someone as a bully: *"@USERNAME i didnt jump around and act like a monkey T_T which of your eye saw that i acted like a monkey :(you're a bully"*
- Revealing self as a victim: *"People bullied me for being fat. 7 years later, I was diagnosed with bulimia. Are you happy now?"*
- Cyber-bullying direct attack: *"Lauren is a fat cow MOO BITCH"*

Bullying traces are abundant. From the publicly available 2011 TREC Microblog track corpus (16 million tweets sampled between January 23rd and February 8th, 2011), we uniformly sampled 990 tweets for manual inspection by five experienced annotators. Of the 990 tweets, the annotators labeled 617 as non-English, 371 as English but not bullying traces, and 2 as English bullying traces. The Maximum Likelihood Estimate of the frequency of English bullying traces, out of all tweets, is $2/990 \approx 0.002$. The exact Binomial 95% confidence interval is (0.0002, 0.0073). This is a tiny fraction. Nonetheless, it represents an abundance of tweets: by some estimates, Twitter produced 250 million tweets per day in late 2011. Even with the lower bound on the confidence interval, it translates into 50,000 English bullying traces per day. The actual number can be much higher.

Bullying traces contain valuable information. For example, Figure 1.2 shows the daily number of bullying traces identified by our classifier, to be discussed in Section 2.1. A weekly pattern was obvious in late August 2011. A small peak was caused by 14-year-old bullying victim Jamey Rodemeyer's suicide on Sept. 18. This was followed by a large peak after Lady Gaga dedicated a song to him on Sept. 24.

1.3 Technical Contributions

Like many complex social issues, effective solutions to bullying go beyond technology alone and require the concerted efforts of parents, educators, and law enforcement. My thesis work focuses on using machine learning and natural language processing techniques to analyze social media posts for bullying. Social

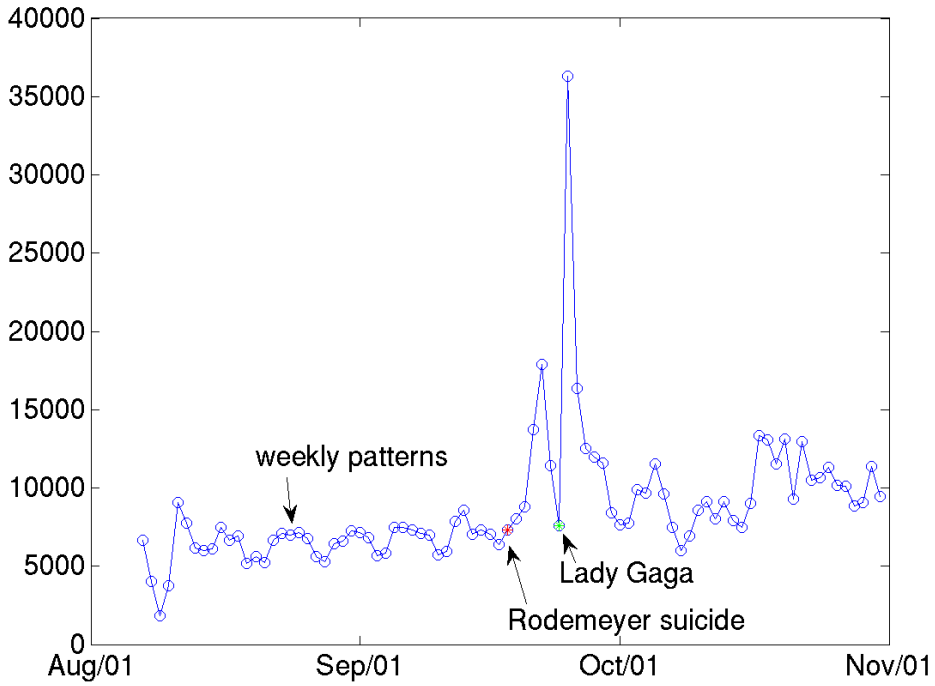


Figure 1.2: Daily count of bullying traces on Twitter identified by our algorithm. Our algorithm only identifies a small fraction of bullying traces – The actual number is much larger.

media posts are often fragmented, noisy and cover only part of a bullying episode. To discover useful information from them, we develop some new models and use some standard approaches.

We introduce the social study of bullying to the machine learning and natural language processing communities by identifying some important scientific questions and formulating them as computational tasks. We build text classifiers with careful feature engineering to recognize bullying traces from social media streams, identifying each author's role in the episode, and understanding the forms and types of the bullying traces. We use role labeling to recognize participants

mentioned in bullying traces and identify their roles. We explore topic modeling methods to understand the topics in bullying traces.

Existing standard models and algorithms are not sufficient for some of our tasks. These new challenges encourage and inspire us to develop new mathematical models and algorithms, which also have broad applications to many domains beyond the study of bullying.

To study the spatiotemporal distribution of bullying episodes and other interesting real-world phenomena, we formulate the problem as a Poisson point process estimation problem. We explicitly incorporate human population bias, time delays and spatial distortions and spatio-temporal regularization into the model. This addresses the problems with simple counting of posts, which is plagued by sample bias, and incomplete data.

Many emotions expressed in bullying traces are not well-studied in the sentiment analysis literature. It poses a challenge to collect enough training data or emotional lexicons. Inspired by "concept labeling" work, we propose a fast training procedure without explicitly producing conventionally labeled training datasets.

To go beyond individual tweets, we look into users' timeline data to obtain more context information about the same incident. We identify the new problem of organizing timelines into conversations and propose a new model based on the distance-dependent Chinese Restaurant Process (Blei and Frazier, 2011).

To discover useful information on bullying from social media, the first step is to identify bullying traces among the massive amounts of social media posts produced every day. We formulate the problem as a binary text classification task (Xu et al., 2012b). With carefully chosen features and classification algorithms, we are able to recognize bullying traces with satisfactory accuracy. To automatically analyze bullying traces, we build text classifiers to recognize an author's role, forms and types of bullying traces. Then, we investigate several important scientific questions with bullying traces (Bellmore et al., 2015), including: who are posting bullying traces? What forms of bullying are mentioned or used on Twitter? Why are people posting about bullying on Twitter? What topics are people posting about bullying?

2.1 Recognizing Bullying Traces

Since bullying traces account for only a tiny fraction of all tweets, a significant challenge is to find enough bullying traces without labeling an enormous number of tweets. For this reason, we restrict ourselves to an "enriched dataset," which is obtained by collecting tweets via the public Twitter streaming API using keywords.

To capture a post, the complete words must be identified and followed on Twitter as opposed to portions of words (e.g., following the word "bull" would not capture the term "bully"). We followed the following keywords related to the term bully: bullied, bully, bullyed, bullying, bullyer, bulling. We included several misspelled keywords as they appear frequently in social media posts (Liu et al., 2012). We also followed several other words that were identified by our annotators as common terms in a content analysis of middle school students' written descriptions of bullying experiences. These keywords were: ignored, pushed, rumors, locker, spread, shoved, rumor, teased, kicked, crying. From all the tweets obtained which contain at least one of these keywords, we then filtered the tweets so that only posts that contained a word starting with "bull" were retained. For example, the post,

"*Bullies pushed the kid*" would be detected and retained even though the keyword bullies was not initially followed. On the other hand, the post, "*He pushed the kid*" would be collected but not retained. In lieu of pre-determining every possible variation of the word bully to follow on Twitter, this process was used to maximize the variations that might be used. From this dataset, we removed re-tweets (the analogy of forwarded emails) by excluding tweets containing the acronym "RT". The enrichment process is meant to retain many first-hand bullying traces at the cost of a selection bias. It is also important to note that this simple keyword filtering is far from perfect: many irrelevant tweets survived and relevant tweets missed.

A bullying trace was defined as any mention of bullying within the context of a specific bullying episode the author was involved in. Note that we did not evaluate the post for compliance with bullying definitions that included notions of power imbalance and repetition (Olweus, 1993) because this information was often not evident in short posts. Moreover, we could not determine whether the episode referred to a single event or a continuous episode over a period of time. We relied entirely on the text from each individual post, taking it at face value when an author participated in or reported "bully"ing. Some positively labeled examples include:

- Personal experiences ("*ugh u bully me a lot*")
- Reports about specific episodes ("*some tweens got violent on the n train, the one boy got off after blows 2 the chest... Saw him cryin as he walkd away :(bullying not cool*")
- Newsworthy posts "(*5 teens had a 14yo hang herself BC they wouldn't stop bullying her*").

Posts were negatively labeled if they were not defined as a bullying episode. Some examples include:

- Posts clearly copied and pasted a news headline about a bullying episode "(*Bully Staffers Shove CNN Reporter-to Avoid Answering Questions?*")

- Posts referred to a bullying episode that may happen in the future "(*When school starts, I will bully you*")
- Posts reflected only an opinion about bullying in general "(*Bullying is violence against the weak*")
- Posts where a behavior may sound like bullying but is not identified as such by the author "(*My friend treats me bad-do you think he is a bully?*")
- Posts where a coder recognized the names mentioned in the post as fictional "(*Harry Potter stood up to that bully*")

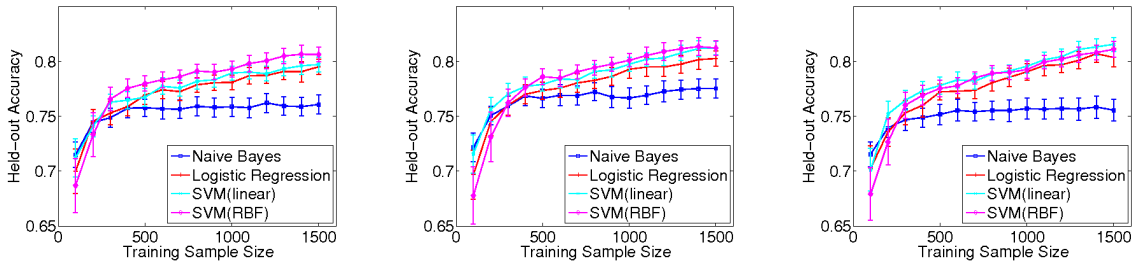
Data

To identifying the true bullying traces from the enriched dataset, we formulate it as a binary text categorization task. Annotators labeled 1,762 tweets sampled uniformly from the enriched dataset on August 6, 2011. Among them, 684 (39%) were labeled as bullying traces. This dataset and its documentation is archived as Bullying Trace Data Set (Version 1) (more details in Appendix A.1).

Methods

Following Settles (2011), these 1,762 tweets were case-folded but without any stemming or stopword removal. Any user mentions preceded by a "@" were replaced by the anonymized user name "@USERNAME". Any URLs starting with "http" were replaced by the token "HTTPLINK". Hashtags (compound words following "#") were not split and were treated as a single token. Emoticons, such as ":)" or ":D", were also included as tokens.

After these preprocessing procedures, we created three different sets of feature representations: unigrams (`1g`), unigrams+bigrams (`1g2g`), and POS-colored unigrams+bigrams (`1g2gPOS`). POS tagging was done with the Stanford CoreNLP package (Toutanova et al., 2003). POS-coloring was done by expanding each token into token:POS.



(a) 1g (b) 1g2g (c) 1g2gPOS

Figure 2.1: Learning curves for different feature sets and classification algorithms on binary bullying trace classification task

We chose four commonly used text classifiers, namely, Naive Bayes, SVM with a linear kernel (`SVM(linear)`), SVM with an RBF kernel (`SVM(RBF)`) and Logistic Regression (equivalent to MaxEnt). We used the WEKA (Hall et al., 2009) implementation for the first three (calling LibSVM (Chang and Lin, 2011) with WEKA's interfaces for SVMs), and the L1General package (Schmidt et al., 2007) for the fourth.

We held out 262 tweets for testing, and systematically varied the training set size among the remaining tweets, from 100 to 1,500 with step-size 100. We tuned all parameters jointly by 5-fold cross validation on the training set with the grid $\{2^{-8}, 2^{-6}, \ldots, 2^{8}\}$. All four text classifiers were trained on the training sets and tested on the test set. The whole procedure was repeated 30 times for each feature representation.

Results

Figure 2.1 reports the held-out set accuracy as the training set size increases. The error bars are ± 1 standard deviation error. With the largest training set size (1,500), the combination of `SVM(linear)` + `1g` achieved an average accuracy of 79.7%. `SVM(linear)` + `1g2g` achieved 81.3%, which is significantly better (t-test, $p = 4 \times 10^{-6}$). It shows that including bigrams can significantly improve the

	Total	Predicted as Yes	Not
Yes	2,102	1,555	547
No	5,219	503	4,716

Table 2.1: Confusion matrix of bullying trace classification

classification performance. `SVM(linear) + 1g2gPOS` achieved 81.6%, though the improvement is not statistically significant ($p = 0.088$), which indicates that POS coloring does not help much on this task. `SVM(RBF)` gives similar performance, Logistic Regression is slightly worse, and Naive Bayes is much worse, for a large range of training set sizes. In summary, `SVM(linear) + 1g2g` is the preferred model because of its accuracy and simplicity. We make it publicly available in our software repository (see Appendix B.1). We also note that these accuracies are much better than the majority class baseline of 61%. On the held-out set, `SVM(linear) + 1g2g` achieved precision P=0.76, recall R=0.79, and F-measure 0.77.

Discussion

As to why the best accuracy is not close to 1, one hypothesis is noisy labels caused by intrinsic disagreement among the annotators. Tweets are short and some are ambiguous. Without prior knowledge about the users and their other tweets, the annotators interpreted the tweets in their own ways. For example, for the very short tweet *"feels like a bully....."* our annotators disagreed on whether it is a bullying trace. Labelers may have different views on these ambiguous tweets, which created noisy bullying trace labels.

Number of Bullying Traces Identified in 2011-2013

Figure 2.1 shows that the learning curves are still increasing, suggesting that better accuracy can be obtained if we annotate more training data. To achieve better performance, our annotators labeled 7,321 tweets (including the 1,762 tweets in the previous experiment), randomly selected from August 6, 2011 through August 31,

	9/1/11 - 8/31/12		9/1/12 - 8/31/13		total	
	Count	%	Count	%	Count	%
Keyword-filtered tweets	12,421,237		20,056,321		32,477,558	
Bullying traces	3,955,458	31.8%	5,809,125	29.0%	9,764,583	30.1%

Table 2.2: The number of bullying traces identified in 2011-2013

2011. The inter-rater agreement for identifying bullying traces from the bullying keyword tweets was calculated based on two annotators coding 1,000 of the 7,321 posts. It was determined to be $\kappa = .83$. Of the 7,321 posts, 2,102, or 28.7%, were labeled as bullying traces. This dataset and its documentation is archived as Bullying Trace Data Set (Version 3) (more details in Appendix A.1) With this larger training set, the accuracy of our text classifier `SVM(linear)` + `1g2g` improved to 86% (see Table 2.1 and Appendix B.1). This level of accuracy is similar to the level of agreement achieved by two different human annotators. Therefore, we determined that it is possible to use machine learning to automatically and accurately recognize bullying traces in social media.

We applied the trained classifier to the enriched dataset collected from September 1, 2011 through August 31, 2013. The result is reported in Table 2.2. We collected 32,477,558 tweets in our enriched dataset via keyword filtering during two consecutive academic years. Among them, we found that 30.1% (9,764,583) were recognized as bullying traces by our text classifier. The proportions of bullying traces and non-bullying traces identified were similar to the human-annotated data in Table 2.1.

The main difference between the two years is that the overall number of bullying traces increased from year 1 to year 2, a trend that likely reflects the increased popularity of Twitter between the two years. There were 3,955,458 bullying traces in 2011-2012 and 5,809,125 in 2012-2013.

2.2 Identifying Participants and Their Roles

Current research emphasizes the group-based nature of bullying and the significance of all social role players in a bullying episode (Salmivalli et al., 1996). Because bullying involves multiple roles, any individual is susceptible to being impacted by bullying, even those who are not directly involved (Rivers et al., 2009). Identifying participants' bullying roles (Figure 1.1) is another important task, which is also a prerequisite for studying how a person's role evolves over time.

As shown and introduced in Figure 1.1, the original six categories we searched for were derived from Salmivalli (1999): bully, victim, bystander, defender, assistant, and reinforcer. For bullying traces in social media, we augmented the traditional role system with two new roles: reporter (i.e., one who shares information about an episode but is not involved in any way, including as a bystander), and accuser (i.e., one who directly accuses someone of a bullying role in the post but it is unclear whether the author is a victim, defender, or some other role). Both roles can be a victim, a defender, or a bystander in the traditional sense – there is just not enough information in the tweet. Accuser (A), bully (B), reporter (R) and victim (V) are the four most frequent roles observed in social media. We merged all remaining roles into a generic category "other" (O) in the following study. Our task is to classify the role (A, B, R, V, O) of the tweet author and any person-mentions in a tweet. For example, AUTHOR$^{(R)}$: "*We*$^{(R)}$ *visited my*$^{(V)}$ *cousin*$^{(V)}$ *today & #Itreallymakesmemad that he*$^{(V)}$ *barely eats bec he*$^{(V)}$ *was bullied . :(I*$^{(R)}$ *wanna kick the crap out of those mean*$^{(B)}$ *kids*$^{(B)}$*.*" Note that the special token "AUTHOR" is introduced to hold the label of the author's role.

Labeling the author's role and other person-mention's role are two different sub-tasks. The former can be formulated as a multi-class text classification task; the latter is better formulated as a sequential tagging task. We will discuss them separately below.

	predicted as				
	A	B	R	V	O
A	33	3	39	10	1
B	5	25	57	11	0
R	15	5	249	27	0
V	1	4	48	109	0
O	1	1	37	3	0

Table 2.3: Confusion matrix of author role classification

Author's Roles

Methods. Our annotators labeled the author's role for each of the 684 positive bullying traces in 1,732 tweets in Bullying Traces Data Set (version 1, Appendix A.1) (296 R, 162 V, 98 B, 86 A, 42 O). We used the same classifiers and features described in Section 2.1. We conducted 10-fold cross validation to evaluate all combinations of classifiers and feature sets. Like before, we tuned all parameters jointly by 5-fold cross validation on the training set with the grid $\{2^{-8}, 2^{-6}, \ldots, 2^{8}\}$.

Results. The best combination was `SVM(linear)` + `1g2g` with cross validation accuracy 61%. Even though it is far from perfect, it is significantly better than the majority class (R) baseline of 43%. It shows that there is signal in the text to infer the authors' roles.

Table 2.3 shows the confusion matrix of the best model. Most R and V authors were correctly recognized, but not B and A. The model misclassified many authors as R. It is possible that the tweets authored by reporters are diverse in topic and style, and overlap with other classes in the feature space.

Discussion. As tweets are short, our feature representation may not be the best for predicting the author's role. Many authors mentioned themselves in the tweets with first-person pronouns, making it advantageous to consider joint classification on the author's role and the person-mention's role. Furthermore, assuming roles change infrequently, it may be helpful to jointly classify many tweets authored by the same person.

	Total	predicted as					
		accuser	bully	defender	reporter	victim	other
accuser	317	212	12	9	56	28	0
bully	303	24	165	4	45	65	0
defender	178	30	6	39	77	26	0
reporter	708	44	19	12	575	58	0
victim	589	19	16	3	63	488	0
other	7	0	0	0	6	1	0

Table 2.4: Confusion matrix of author role classification (six roles)

Who is Posting about Bullying on Twitter?

With the classifier to identify the author's role, we next sought to identify the role-players who post about bullying episodes on Twitter as well as their distribution in order to describe who posts about bullying versus who participates in bullying. To achieve more accurate results, out annotators labeled the role of the author of every post identified as a bullying trace in the training set of 7,321 posts randomly selected from dates August 6, 2011 through August 31, 2011. This dataset and its documentation is archived as Bullying Trace Data Set (Version 3, more details in Appendix A.1). The inter-rater agreement for these nine categories was calculated based on two coders coding 1,000 of the 7,321 posts. It was determined to be $\kappa = .79$. The human-coded data, presented in Table 2.4, revealed that among the 2,102 bullying traces, reporters (708 (33.68%) bullying trace authors) and victims (589 (28.02%) of bullying trace authors) were the two most frequent types of authors of bullying posts.

Because only a very small number of assistants and reinforcers were identified, we classified these groups together into an "other" category. We trained an author role classifier `SVM(linear)` + `1g2g` as in the previous section. With this larger training set, it was able to reliably distinguish defenders from other roles, and achieved 70% cross validation accuracy. The classifier is archived in our repository (see Appendix B.2).

	Human-Code		9/1/11 - 8/31/12		9/1/12 - 8/31/13		total	
	Count	%	Count	%	Count	%	Count	%
accuser	317	15.1%	662,880	16.8%	983,801	16.9%	1,646,681	16.9%
bully	303	14.4%	496,039	12.5%	685,269	11.8%	1,181,308	12.1%
defender	178	8.5%	99,322	2.5%	145,900	2.5%	245,222	2.5%
reporter	708	33.7%	1,337,205	33.8%	1,838,376	31.7%	3,175,581	32.5%
victim	589	28.0%	1,360,001	34.4%	2,155,759	37.1%	3,515,760	36.0%
other	7	0.3%	11	0.0%	20	0.0%	31	0.0%

Table 2.5: Distribution of human-coded and machine learning identified author roles

To analyze who posted across all bullying traces during the 2011-2013 school years, we applied the Author's Role Classifier to the Bullying Traces in Two Academic Years data set (see Appendix A.2). Table 2.4 shows the confusion matrix that illustrates agreement and disagreement between the human-coded bullying role and the predicted bullying role when we used machine learning methods on the data.

The classifier found a similar distribution as our coded data with victims (36.0%, n = 3,515,760) and reporters (32.5%, n = 3,175,581) being identified as authors of bullying traces most frequently. Table 2.5 contains the distributions across roles for both the human-coded and machine-coded data for each school year independently as well as across the two-year time span.

Person-Mention Roles

Many users mentioned in tweet text are also participants involved in a bullying episode. This sub-task labels each person-mention with a bullying role. It uses Named Entity Recognition (NER) (Finkel et al., 2005; Ratinov and Roth, 2009; Ritter et al., 2011) as a subroutine to identify named person entities, though we are also interested in unnamed persons such as "my teacher" and pronouns. It is related to Semantic Role Labeling (SRL) (Gildea and Jurafsky, 2002; Punyakanok et al., 2008) but differs critically in that our roles are not tied to specific verb predicates.

Methods. Our annotators labeled each token in 684 bullying traces in Bullying Traces Data Set (version 1, see Appendix A.1) with the tags A, B, R, V, O and N for not-a-person. There were 11,751 tokens in total. Similar to the sequential tagging formulation (Màrquez et al., 2005; Liu et al., 2010), we trained a linear CRF to label each token in the tweet with the CRF++ package (http://crfpp.sourceforge.net/).

As standard in linear CRFs, we used pairwise label features $f(y_{i-1}, y_i)$ and input features $f(y_i, \mathbf{w})$, where f's are binary indicator functions on the values of their arguments and $\mathbf{w}$ is the text. In the following, we introduce our input features using the example tweet *"@USERNAME i'll tell vinny you bullied me."* with the current token w_i ="vinny":

(i) The token, lemma, and POS tag of the five tokens around position i. For example, $f_{bully, w_{i-1}=tell}(y_i, \mathbf{w})$ will be 1 if the current token has label $y_i =$ "bully" and $w_{i-1} =$ "tell". Similarly, $f_{victim, POS_{i+2}=VBD}(y_i, \mathbf{w})$ will be 1 if $y_i =$ "victim" and the POS of w_{i+2} is VBD.

(ii) The NER tag of w_i.

(iii) Whether w_i is a person-mention. This is a Boolean feature that is true if w_i is tagged as PERSON by NER, or if $POS_i = pronoun$ (excluding "it"), or if w_i is @USERNAME. For example, this feature is true on "vinny" because it is tagged as PERSON by NER.

(iv) The relevant verb v_i of w_i, v_i's lemma, POS, and the combination of v_i with the lemma/POS of w_i. The relevant verb v_i of w_i is defined by the semantic dependency between w_i and the verb, if one exists. Otherwise, v_i is the closest verb to w_i. For example, the relevant verb of $w_i =$ "vinny" is $v_i =$ "tell" because "vinny" is found as the object of "tell" by dependency parsing.

(v) The distance, relative position (left or right), and dependency type between v_i and w_i. For example, the distance between "vinny" and its relevant verb "tell" is 1. "vinny" is on the right and is the object of "tell".

The lemma, POS tags, NER tags and dependency relationship were obtained using Stanford CoreNLP.

As a baseline, we trained `SVM(linear)` with the same input features as CRF. Classification was done individually on each token. We randomly split the 684

	Accuracy	Precision	Recall	F-1
CRF	0.87	0.53	0.42	0.47
SVM	0.85	0.42	0.31	0.36

Table 2.6: Cross validation result of person-mention roles

tweets into 10 folds and conducted cross validation based on this split. For CRF, we trained on the tweets in the training set with their labels, and tested the model on those in the test set. For SVM, we trained and tested at the token level in the corresponding sets.

Results. Table 2.6 reports the cross validation accuracy, precision, recall and F-1 measure. *Accuracy* measures the percentage of tokens correctly assigned the groundtruth labels, including N (not-a-person) tokens. *Precision* measures the fraction of correctly labeled person-mention tokens over all tokens that are not N according to the algorithm. *Recall* measures the fraction of correctly labeled person-mention tokens over all tokens that are not N according to the groundtruth. *F-1* is the harmonic mean of precision and recall. Linear CRF achieved an accuracy of 87%, which is higher than the baseline of majority class predictor (N, 0.80) (t-test, $p = 10^{-10}$). However, the precision and recall is low, potentially because the tweets are short and noisy. CRF outperformed SVM in all measures, showing the value of joint classification.

Discussion. Table 2.7 shows the confusion matrix of person-mention role labeling by a linear CRF. There are several reasons for these mistakes. First, words like "teacher," "sister," or "girl" were missed by our person mention feature (iii). Second, the NER tagger was trained on formal English, which is a mismatch for the informal tweets, leading to NER errors. Third, noisy labeling continues to affect accuracy. For example, some annotators considered "other people" as an entity and labeled both tokens as person mentions; others labeled "people" only.

In general, bullying role labeling may be improved by jointly considering multiple tweets at the episode level. Co-reference resolution should improve the performance as well.

	predicted as					
	A	B	R	V	O	N
A	0	4	5	10	0	4
B	0	406	13	125	103	302
R	0	28	31	67	0	13
V	0	142	28	380	43	202
O	0	112	4	42	156	86
N	0	78	4	41	16	9,306

Table 2.7: Confusion matrix of person-mention roles by CRF

	Total	predicted as			
		general	cyberbullying	verbal	physical
general	1,857	1,831	20	4	2
cyberbullying	145	68	73	4	0
verbal	67	53	12	2	0
physical	33	32	0	0	1

Table 2.8: Confusion matrix of the forms of bullying traces

	Human-Code		9/1/11 - 8/31/12		9/1/12 - 8/31/13		total	
	Count	%	Count	%	Count	%	Count	%
general	1,857	88.3%	3,765,015	95.2%	5,531,636	95.2%	9,296,651	95.2%
cyberbullying	145	6.9%	164,866	4.2%	239,517	4.1%	404,383	4.1%
verbal	67	3.2%	20,403	0.5%	30,931	0.5%	51,334	0.5%
physical	33	1.6%	5,174	0.1%	7,041	0.1%	12,215	0.1%

Table 2.9: Distribution of human-coded and machine learning identified bullying forms

2.3 Understanding What Forms of Bullying are Mentioned or Used in Bullying Traces

Bullying takes multiple forms, most noticeably face-to-face physical (e.g., hitting), verbal (e.g., name-calling), relational (e.g., exclusion), and cyber (e.g., hacking) (Archer and Coyne, 2005; Little et al., 2003; Nylund et al., 2007; Wang et al.,

2009). Any form may be represented on Twitter because interactions, in both physical world and online, can be mentioned. We aimed to identify the distribution of bullying forms in mentions of bullying on Twitter to establish whether these forms are distinguishable in social media posts, and, if they are, which forms are most prevalent.

Two annotators labeled the form of bullying mentioned in every post identified as a bullying trace in the training set of 7,321 posts in the Bullying Trace Data Set (Version 3, see Appendix A.1). The categories were general (no information was provided to indicate a form), cyberbullying, physical, verbal, relational, and property damage (see (Wang et al., 2009) for sample behaviors that correspond to these categories). The inter-rater agreement across two human coders for these seven categories was $\kappa = .77$.

With the labeled training data, we built a text classifier to predict the forms of bullying traces (see Appendix B.3) with the annotated data. Due to the small number of examples for some of the categories, the classifiers were not able to recognize them correctly when we applied machine learning methods. As a result, we removed the categories of property damage and relational. After doing this, the classifier achieved 70% accuracy (see Table 2.8).

Across all bullying traces in 2011-2013 in the Bullying Traces in Two Academic Years data set (see Appendix A.2), the classifier found that posts about general forms of bullying were by far the most common–95.2% (n = 9,296,651) of the bullying traces. Cyberbullying posts comprised the next most frequent form (4.1%, n = 404,383 posts). See the middle panel in Table 2.9 for the counts across labelings by human coders and the Bullying Form Classifier (see Appendix B.3).

2.4 Understanding Why Users Post Bullying Traces

Twitter can serve as a platform for both sharing information (Java et al., 2007) and making connections with others (Chen, 2011). Both of these are relevant to understanding why people might post about bullying. Any author might post for any reason–victims may seek social support through reporting, defenders may

offer support, and bullies may aggress against others. To understand the different functions that bullying posts might serve, we identified different categories of bullying posts (e.g., self-disclosure) and reported their distribution.

The initial categories of why people post about bullying episodes were determined by our annotators after preliminary coding and discussions of all bullying traces in the Bullying Traces Data Set (Version 1, see Appendix A.1). The different types of bullying traces identified were:

- Reports: Posts that described a bullying episode someone knows about, e.g., "*some tweens got violent on the n train, the one boy got off after blows 2 the chest.... Saw him cryin as he walkd away :(bullying not cool.*"
- Accusations: Posts that accused someone as the bully in an episode, e.g., "*@USER i didnt jump around and act like a monkey T T which of your eye saw that i acted like a monkey :(you're a bully.*"
- Self-Disclosures: Posts that revealed the author himself/herself as the bully, victim, defender, bystander, assistant, or reinforcer, e.g., "*People bullied me for being fat. 7 years later, I was diagnosed with bulimia.*"
- Denials: Posts where the author denied a bullying role, e.g., "*@USER lol I'm not a bully man*"
- Cyberbullying: Posts that were direct attacks from a bully to a victim. For example, "*@USER really I am just cyberbullying you right now.*"

Two annotators labeled the types of bullying trace in Bullying Trace Data Set (Version 3, see Appendix A.1). The inter-rater agreement for these five categories based on two human coders was $\kappa = .76$. We trained a Bullying Trace Type Classifier (i.e., why people post) classifier with `SVM(linear)` + `1g2g` (see Appendix B.5). The accuracy of this classification was 72% (see Table 2.10 for the confusion matrix). To analyze the distribution across all bullying traces in 2011-2013, we applied the Bullying Trace Type Classifier to the Bullying Traces in Two Academic Years data set (Appendix A.2). The classifier found that self-disclosure posts (54.3%, n = 5,306,451)

	Total	predicted as				
		accusation	cyberbullying	denial	report	self-disclosure
accusation	316	196	0	2	54	64
cyberbullying	16	4	0	0	7	5
denial	128	10	0	32	9	77
report	709	36	0	0	538	135
self-disclosure	933	48	0	3	132	750

Table 2.10: Confusion matrix of the types of bullying traces

	Human-Code		9/1/11 - 8/31/12		9/1/12 - 8/31/13		total	
	Count	%	Count	%	Count	%	Count	%
accusation	316	15.0%	595,383	15.1%	887,508	15.3%	1,482,891	15.2%
cyberbullying	16	0.8%	14	0.0%	28	0.0%	42	0.0%
denial	128	6.1%	79,630	2.0%	106,298	1.8%	185,928	1.9%
report	709	33.7%	1,175,234	29.7%	1,614,037	27.8%	2,789,271	28.6%
self-disclosure	933	44.4%	2,105,197	53.2%	3,201,254	55.1%	5,306,451	54.3%

Table 2.11: Distribution of human-coded and machine learning identified bullying trace types

were most common followed by reports (28.6%, n = 2,789,271), accusations (15.2%, n = 1,482,891) and denials (1.9%, n = 185,928). See Table 2.11 for more details.

2.5 Understanding the Topics in Bullying Traces

Besides these quantitative studies of bullying traces, understanding what topics users are talking about in bullying traces is also helpful for social scientists. Given the large volume of bullying traces, methods for automatically analyzing what people are talking about are needed.

Methods. Latent topic models allow us to extract the main topics in bullying traces to facilitate understanding. We used latent Dirichlet allocation (LDA) (Blei et al., 2003) as our exploratory tool. Specifically, we ran a collapsed Gibbs sampling implementation of LDA (Griffiths and Steyvers, 2004).

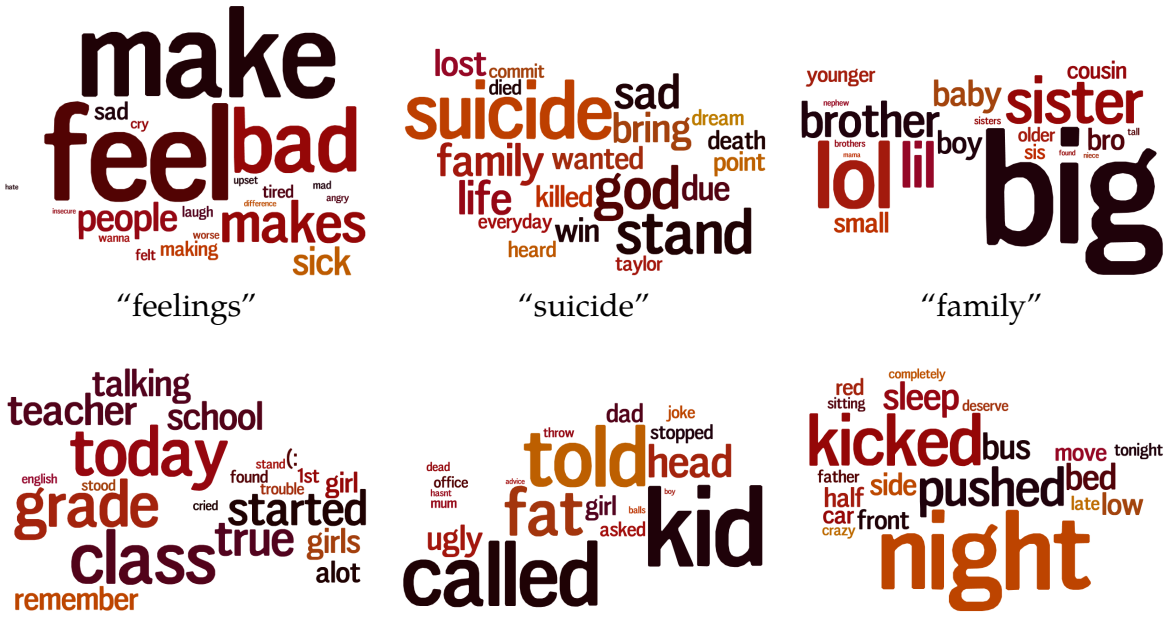
The corpus consisted of 188,000 enriched tweets from August 21 to September 17, 2011 that were classified as bullying traces by our Binary Bullying Trace Classifier (see Appendix B.1). The dataset and its documentation is archived as the Topics in Bullying Traces (see Appendix A.3). We performed stopword removal and further removed word types occurring less than 7 times, resulting in a vocabulary of size about 12,000. We set the number of topics to 50, the Dirichlet parameter for word multinomials to $\beta = 0.01$, the Dirichlet parameter for document topic multinomial to $\alpha = 1$, and ran Gibbs sampling for 10,000 iterations.

Results. Space precludes a complete list of topics. Figure 2.2 shows six selected topics discovered by LDA. Recall that each topic in LDA is a multinomial distribution over the vocabulary. The figure shows each topic's top 20 words with size proportional to $\mathsf{P}(\text{word} \mid \text{topic})$. The topic names were manually assigned.

These topics contain semantically coherent words relevant to bullying: (feelings) how people feel about bullying; (suicide) discussions of suicide events; (family) sibling names probably used in a good buddy sense; (school) the school environment where bullying commonly occurs; (verbal bullying) derogatory words such as fat and ugly; (physical bullying) actions such as kicking and pushing.

We also ran a variational inference implementation of LDA (Blei et al., 2003). The results were similar, so we omit discussion of them here.

Discussion. Some recovered topics, including the ones shown here, provide valuable insights into bullying traces. However, not all topics are interpretable to social scientists. It may be helpful to allow scientists the ability to combine their domain knowledge with latent topic modeling, thus arriving at more useful topics. For example, the scientists can formulate their knowledge in First-Order Logic, which can then be combined with LDA with stochastic optimization (Andrzejewski et al., 2011).



"feelings" "suicide" "family"

"school" "verbal bullying" "physical bullying"

Figure 2.2: Selected topics discovered by latent Dirichlet allocation.

3 SPATIOTEMPORAL DISTRIBUTION OF BULLYING TRACES

Several spatial and timing issues related to bullying episodes are important to know. For example, social scientists want to know if the prevalence rates of bullying episodes are different across different cultures and geography. From longitudinal research, we know when students are most likely to identify as victims across multiple school years (Nylund et al., 2007). We know little about the timing of discrete bullying episodes. Monitoring the spatiotemporal variations of bullying episodes is also important to evaluate the effectiveness of special campaigns or policies, proving feedback to educators and policy makers.

Besides prevalence rates of bullying episodes, many real-world phenomena of interest are spatiotemporal in nature as well. Examples include wildlife mortality, algal blooms, hail damage, and seismic intensity. The signal can be characterized by a real-valued intensity function $\mathbf{f} \in \mathbb{R}_{\geqslant 0}$, where the value $f_{s,t}$ quantifies the prevalence of the phenomenon at location s and time t. Direct sensing of $\mathbf{f}$ using instruments is often difficult and expensive. Social media offers a unique sensing opportunity for such spatiotemporal signals, where users serve the role of "sensors" by posting their experiences of a target phenomenon. For instance, social media users readily post their encounters with dead animals: *"I saw a dead crow on its back in the middle of the road."*

There are at least three challenges faced when using human social media users as sensors:

1. Social media sources are not always reliable and consistent, due to factors including the vagaries of language and the psychology of users. This makes identifying topics of interest and labeling social media posts extremely challenging.

2. Social media users are not under our control. In most cases, users cannot be directed or focused or maneuvered as we wish. The distribution of human users (our sensors) depends on many factors unrelated to the sensing task itself.

3. Location and time stamps associated with social media posts may be erroneous or missing. Most posts do not include GPS coordinates, and self-reported locations can be inaccurate or false. Furthermore, there can be random delays between an event of interest and the time of the social media post related to the event.

Most prior work in social media event analysis has focused on the first challenge. Sophisticated natural language processing techniques have been used to identify social media posts relevant to a topic of interest (Yang et al., 1998; Becker et al., 2011; Sakaki et al., 2010) and machine learning tools have been proposed to discover popular or emerging topics in social media (Allan, 2002; Mei et al., 2006; Yin et al., 2011). We discuss the related work in detail in Section 3.3.

In this chapter, we focus on the latter two challenges. We are interested in a specific topic or target phenomenon of interest that is given, and we assume that we are also given a (perhaps imperfect) method, such as our Binary Bullying Trace Classifier. The main concerns of this work are to deal with the highly non-uniform distribution of human users (sensors), which affects our capabilities for sensing natural phenomena, and to cope with the uncertainties in the location and time stamps associated with related social media posts. The main contribution is a methodology for deriving accurate spatiotemporal maps of the target phenomenon in light of these two challenges.

We first analyzed the raw counts of GPS-tagged bullying traces identified in the academic years 2011-2013 as an empirical exploration study of spatiotemporal distribution of bullying episodes (Bellmore et al., 2015). However, as mentioned, counting is plagued by sample bias, incomplete data, and, paradoxically, data scarcity. We formulate signal recovery as a Poisson point process estimation problem, which can be used for other applications as well. We propose Socioscope (Xu et al., 2012a, 2013a) in Section 3.3, which explicitly incorporates human population bias, time delays and spatial distortions, and spatiotemporal regularization into the model to address noisy count issues. The code of Socioscope is archived as Socioscope in our code repository (see Appendix B.4).

3.1 Where are People Posting about Bullying on Twitter?

Bullying cross-cuts culture and geography (Jimerson et al., 2010). A real-time social media source such as Twitter may reveal temporary hot spots of bullying post activity. These might reflect different authors posting about a single newsworthy bullying episode or different authors posting about different episodes all occurring in one locale. To understand how bullying is represented across the United States, we identified the location of origin of bullying posts and reported their prevalence relative to the size of the population of their origin.

The geographic location of the source of the post was determined from posts in which users enabled the Twitter option to provide the Global Positioning System (GPS) coordinates of their location within their posted tweet. We applied our Binary Bullying Traces Classifier (see Appendix B.1) to the Bullying Traces in Two Academic Years dataset (see Appendix A.2). Of 9,764,583 bullying traces identified in the academic years 2011-2013, about 2% (191,657) of them contain GPS coordinates. We used a reverse geocoding database (http://www.datasciencetoolkit.org) to obtain the state names and determined that 105,655 originated in the United States.

To understand the origins of bullying traces, we estimated the number of bullying traces per capita for the 50 US states and Washington DC identified through posts that contained GPS information. Figure 3.1 shows the state names listed in alphabetical order, their population based on the 2010 census, the number of bullying traces, and the per capita number of bullying traces for the years 2011-2012, 2012-2013, and total. We also present the ranking of each state based on their population size and their per capita volume of bullying traces. The five states with the largest number of bullying traces per capita are Delaware, Washington DC, Maryland, Ohio, and Rhode Island for the period 2011-2013. Spearman rank order correlations reveal a positive association between rankings of states based on population size and rankings of states based on the number of bullying traces per capita, r_s (51) = .38, p = .006 in 2011-2012 and r_s (51) = .30, p = .033 in 2012-2013. These values reflect moderate effect sizes (Cohen, 1998).

US State	Population size (census 2010)	Population rank	2011–2012			2012–2013			2011–2013		
			Number of GPS bullying traces	Number of bullying traces per capita	Per capita bullying traces rank	Number of GPS bullying traces	Number of bullying traces per capita	Per capita bullying traces rank	Number of GPS bullying traces	Number of bullying traces per capita	Per capita bullying traces rank
Alabama	4,779,736	23	522	1.09E−04	16	1158	2.42E−04	23	1680	3.51E−04	23
Alaska	710,231	47	42	5.91E−05	34	123	1.73E−04	37	165	2.32E−04	37
Arizona	6,392,017	16	368	5.76E−05	35	1497	2.34E−04	28	1865	2.92E−04	28
Arkansas	2,915,918	32	266	9.12E−05	27	424	1.45E−04	44	690	2.37E−04	44
California	37,253,956	1	3204	8.60E−05	28	8708	2.34E−04	28	11,912	3.20E−04	30
Colorado	5,029,196	22	247	4.91E−05	42	758	1.51E−04	43	1005	2.00E−04	43
Connecticut	3,574,097	29	551	1.54E−04	6	1009	2.82E−04	10	1560	4.36E−04	10
Delaware	897,934	45	195	2.17E−04	2	391	4.35E−04	1	586	6.53E−04	1
District of Columbia	601,723	50	209	3.47E−04	1	162	2.69E−04	14	371	6.17E−04	14
Florida	18,801,310	4	1781	9.47E−05	23	4099	2.18E−04	32	5880	3.13E−04	32
Georgia	9,687,653	9	1682	1.74E−04	4	2621	2.71E−04	13	4303	4.44E−04	13
Hawaii	1,360,301	40	73	5.37E−05	39	156	1.15E−04	47	229	1.68E−04	47
Idaho	1,567,582	39	44	2.81E−05	50	144	9.19E−05	49	188	1.20E−04	49
Illinois	12,830,632	5	1205	9.39E−05	24	3003	2.34E−04	28	4208	3.28E−04	29
Indiana	6,483,802	15	594	9.16E−05	26	1596	2.46E−04	20	2190	3.38E−04	20
Iowa	3,046,355	30	225	7.39E−05	29	742	2.44E−04	21	967	3.17E−04	21
Kansas	2,853,118	33	207	7.26E−05	31	798	2.80E−04	12	1005	3.52E−04	12
Kentucky	4,339,367	26	406	9.36E−05	25	1026	2.36E−04	26	1432	3.30E−04	26
Louisiana	4,533,372	25	709	1.56E−04	5	939	2.07E−04	33	1648	3.64E−04	33
Maine	1,328,361	41	67	5.04E−05	41	181	1.36E−04	46	248	1.87E−04	46
Maryland	5,773,552	19	1138	1.97E−04	3	1672	2.90E−04	9	2810	4.87E−04	9
Massachusetts	6,547,629	14	746	1.14E−04	14	1987	3.03E−04	7	2733	4.17E−04	7
Michigan	9,883,640	8	1184	1.20E−04	12	2953	2.99E−04	8	4137	4.19E−04	8
Minnesota	5,303,925	21	301	5.68E−05	37	1277	2.41E−04	24	1578	2.98E−04	24
Mississippi	2,967,297	31	360	1.21E−04	10	469	1.58E−04	42	829	2.79E−04	42
Missouri	5,988,927	18	436	7.28E−05	30	1151	1.92E−04	36	1587	2.65E−04	36
Montana	989,415	44	19	1.92E−05	51	50	5.05E−05	51	69	6.97E−05	51
Nebraska	1,826,341	38	123	6.73E−05	32	459	2.51E−04	19	582	3.19E−04	19
Nevada	2,700,551	35	299	1.11E−04	15	938	3.47E−04	2	1237	4.58E−04	2
New Hampshire	1,316,470	42	59	4.48E−05	44	184	1.40E−04	45	243	1.85E−04	45
New Jersey	8,791,894	11	1046	1.19E−04	13	2748	3.13E−04	5	3794	4.32E−04	5
New Mexico	2,059,179	36	106	5.15E−05	40	334	1.62E−04	40	440	2.14E−04	40
New York	19,378,102	3	1945	1.00E−04	22	4285	2.21E−04	31	6230	3.21E−04	31
North Carolina	9,535,483	10	1003	1.05E−04	20	2408	2.53E−04	18	3411	3.58E−04	18
North Dakota	672,591	48	26	3.87E−05	48	178	2.65E−04	15	204	3.03E−04	15
Ohio	11,536,504	7	1422	1.23E−04	9	3984	3.45E−04	3	5406	4.69E−04	3
Oklahoma	3,751,351	28	215	5.73E−05	36	884	2.36E−04	26	1099	2.93E−04	27
Oregon	3,831,074	27	175	4.57E−05	43	662	1.73E−04	37	837	2.18E−04	38
Pennsylvania	12,702,379	6	1344	1.06E−04	19	3091	2.43E−04	22	4435	3.49E−04	22
Rhode Island	1,052,567	43	136	1.29E−04	7	349	3.32E−04	4	485	4.61E−04	4
South Carolina	4,625,364	24	571	1.23E−04	8	1405	3.04E−04	6	1976	4.27E−04	6
South Dakota	814,180	46	32	3.93E−05	47	131	1.61E−04	41	163	2.00E−04	41
Tennessee	6,346,105	17	679	1.07E−04	17	1230	1.94E−04	35	1909	3.01E−04	35
Texas	25,145,561	2	2606	1.04E−04	21	7068	2.81E−04	11	9674	3.85E−04	11
Utah	2,763,885	34	113	4.09E−05	46	664	2.40E−04	25	777	2.81E−04	25
Vermont	625,741	49	26	4.16E−05	45	62	9.91E−05	48	88	1.41E−04	48
Virginia	8,001,024	12	960	1.20E−04	11	2051	2.56E−04	17	3011	3.76E−04	17
Washington	6,724,540	13	362	5.38E−05	38	1351	2.01E−04	34	1713	2.55E−04	34
West Virginia	1,852,994	37	198	1.07E−04	18	481	2.60E−04	16	679	3.66E−04	16
Wisconsin	5,686,986	20	367	6.45E−05	33	955	1.68E−04	39	1322	2.32E−04	39
Wyoming	563,626	51	20	3.55E−05	49	45	7.98E−05	50	65	1.15E−04	50

Note. Population data from http://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_population.

Figure 3.1: State population size, number of GPS bullying traces, per capita bullying traces, and population ranks and per capita bullying traces rank for 50 states and the District of Columbia between September 1, 2011 and August 31, 2013.

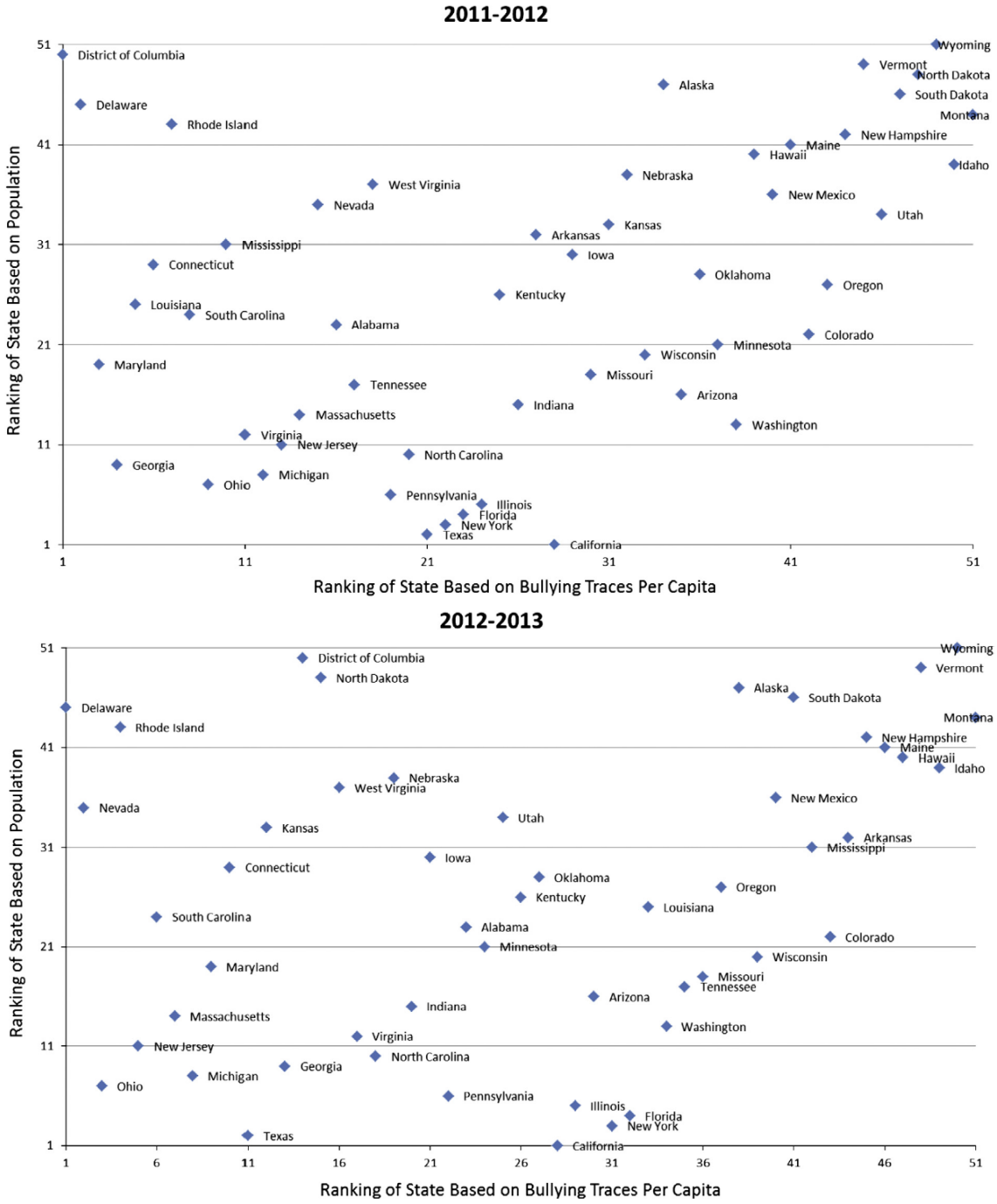


Figure 3.2: Association between ranking of number of bullying traces per capita (1= largest; 51 = smallest) and population size (1 = largest; 51 = smallest) for 50 states and the District of Columbia in 2011-2012 and 2012-2013.

The imperfect association between population size ranking and the number of bullying traces per capita ranking is illustrated in Figure 3.2. In both years, Delaware, Rhode Island, and the District of Columbia emerge as outliers by ranking high on bullying traces relative to their small populations. We inspected all of the bullying traces identified in these states across the study period and found no evidence of any irregularities such as many posts over a short period of time or many posts referring to the same high profile bullying episode in any location that would explain their outlier status. In the future work, we plan to investigate the correlations between bullying trace counts with other factors, such as per capita income, to see if there are strong correlated factors with bullying.

3.2 When are People Posting about Bullying on Twitter?

Several timing issues related to bullying episodes are important to know. In this section, we report on the distribution of bullying episodes across two school years to determine what days of the week and what times of the day posts occur. We expected to see fewer bullying traces on weekends when individuals are away from school and work contexts.

To understand the distribution of bullying traces across time, we evaluated which day of the week and what time of day bullying traces occur most frequently. The timestamp is in Coordinated Universal Time (UTC), which may not be the user's local time. Establishing the location of each post was necessary to appropriately determine the time. In this section, we use geo-tagged tweets to get accurate information of local times. Again, we used the bullying traces identified by our Binary Bullying Traces Classifier (see Appendix B.1) in Bullying Traces in Two Academic Years dataset gt(see Appendix A.2). We calculated the timing from the bullying traces in one east coast state, New York, and from one west coast state, California, for both years 2011-2013 under investigation. We chose these two states because they had the largest number of GPS-tagged bullying traces and contained

	Day of the week							Total
	Mon	Tues	Wed	Thurs	Fri	Sat	Sun	
2011–2012								
New York								
Actual number	287	314	304	294	245	207	294	1945
Expected number	277.86	277.86	277.86	277.86	277.86	277.86	277.86	
Standardized residual	0.55	2.17	1.57	0.97	−1.97	−4.25	0.97	
California								
Actual number	497	518	472	483	451	324	459	3204
Expected number	457.71	457.71	457.71	457.71	457.71	457.71	457.71	
Standardized residual	1.84	2.82	0.67	1.18	−0.31	−6.25	0.06	
2012–2013								
New York								
Actual number	669	659	626	703	548	476	604	4285
Expected number	621.14	621.14	621.14	621.14	621.14	621.14	621.14	
Standardized residual	1.92	1.52	0.20	3.28	−2.93	−5.82	−0.69	
California								
Actual number	1400	1363	1439	1322	1068	909	1207	8708
Expected number	1244.00	1244.00	1244.00	1244.00	1244.00	1244.00	1244.00	
Standardized residual	4.42	3.37	5.53	2.21	−4.99	−9.50	−1.05	

Note. Cell z scores that exceed or fall below ±1.96 are significant at $p < .05$.

Figure 3.3: Number of bullying traces on each day of the week in New York and California in 2011-2012 and 2012-2013.

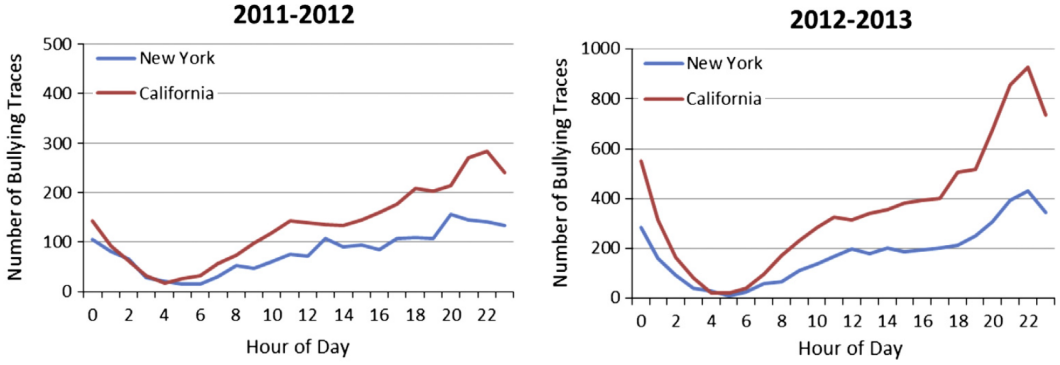


Figure 3.4: Number of bullying traces for each hour of the day from September 1, 2011 through August 31, 2012 (left) and September 1, 2012 through August 31, 2013 (right) that originated in New York and California

a single time-zone. We converted the time in the tweet to the local time in that location and counted the number of bullying traces created in each hour-of-the-day and day-of-the-week. We used chi-square tests to evaluate whether the distribution of bullying traces was statistically uniform across the days of the week in either time period in both New York and California. The chi-square analyses indicated that the distribution was not uniform across days of the week in both New York, $\chi^2(6) = 31.29$, $p < .001$ in 2011-2012 and $\chi^2(6) = 59.78$, $p < .001$ in 2012-2013, and California $\chi^2(6) = 52.32$, $p < .001$ in 2011-2012 and $\chi^2(6) = 182.62$, $p < .001$ in 2012-2013. The effect size for each test was small (Cohen's w = .13 and .12 for New York in 2011-2012 and 2012-2013 and Cohen's w = .13 and .14 for California in 2011-2012 and 2012-2013). The standardized residuals indicate that Saturdays consistently contained fewer posts about bullying episodes than expected if the posts were distributed evenly across all days in each location and in both years. Figure 3.3 contains the actual counts, expected counts, and standardized residuals for each day of the week. With respect to the time of day that bullying traces are posted, Figure 3.4 illustrates that there is a diurnal pattern such that most posts occur in waking periods (especially during the evening hours) in both locations in both years.

3.3 The Socioscope: A Spatiotemporal Model of Social Media

The Proposed Model

We propose Socioscope, a probabilistic model that robustly recovers spatiotemporal signals from social media data. Formally, the signal can be characterized by a real-valued intensity function $\mathbf{f} \in \mathbb{R}_{\geqslant 0}$, where the value $f_{s,t}$ quantifies the prevalence of the phenomenon at location s and time t. Consider $\mathbf{f}$ defined on discrete spatiotemporal bins. For example, bin (s, t) could be a U.S. state s on day t, or a county s in hour t. Once we identify the target social media, we can obtain $x_{s,t}$, the

count of target social media posts within that bin. The task is to estimate $f_{s,t}$ from $x_{s,t}$. A commonly-used estimate is $\widehat{f_{s,t}} = x_{s,t}$ itself. This estimate can be justified as the maximum likelihood estimate of a Poisson model $\mathbf{x} \sim \text{Poisson}(\mathbf{f})$. This idea underlies several emerging systems such as earthquake damage monitoring from Twitter (Earle et al., 2010). However, this estimate is unsatisfactory since the counts $x_{s,t}$ can be *noisy*: as mentioned before, the estimate ignores population bias – more target posts are generated when and where there are more social media users; the location of a target post is frequently inaccurate or missing, making it difficult to assign to the correct bin; and target posts can be sparse even though the total volume of social media is huge. Socioscope addresses these issues.

For notational simplicity, we often denote our signal of interest by a vector $\mathbf{f} = (f_1, \ldots, f_n)^\top \in \mathbb{R}^n_{\geqslant 0}$, where f_j is a non-negative target phenomenon intensity in *source bin* $j = 1 \ldots n$. We will use a wildlife example throughout the section. In this example, a source bin is a spatiotemporal unit such as "California, day 1," and f_j is the squirrel activity level in that unit. The mapping between index j and the aforementioned (s, t) is one-one and will be clear from context.

Correcting Human Population Bias

For now, assume each target post comes with precise location and time metadata. This allows us to count x_j, the number of target posts in bin j. Given x_j, it is tempting to use the maximum likelihood estimate $\widehat{f_j} = x_j$ which assumes a simple Poisson model $x_j \sim \text{Poisson}(f_j)$. However, this model is too naive: Even if $f_j = f_k$, e.g., the level of squirrel activities is the same in two bins, we would expect $x_j > x_k$ if there are more people in bin j than in bin k, simply because more people see the same group of squirrels.

To account for this population bias, we define an "active social media user population intensity" (informally called "human population" below) $\mathbf{g} = (g_1, \ldots, g_n)^\top \in \mathbb{R}^n_{\geqslant 0}$. Let z_j be the count of *all* social media posts in bin j, the vast majority of which are not about the target phenomenon. We assume $z_j \sim \text{Poisson}(g_j)$. Since typically $z_j \gg 0$, the maximum likelihood estimate $\widehat{g_j} = z_j$ is reasonable.

Importantly, we then posit the Poisson model

$$x_j \sim \text{Poisson}(\eta(f_j, g_j)). \tag{3.1}$$

The intensity is defined by a *link function* $\eta(f_j, g_j)$. In this section, we simply define $\eta(f_j, g_j) = f_j \cdot g_j$ but note that other more sophisticated link functions can be learned from data. Given x_j and z_j, one can then estimate f_j with the plug-in estimator $\widehat{f_j} = x_j/z_j$.

Handling Noisy and Incomplete Data

This would be the end of the story if we could reliably assign each post to a source bin. Unfortunately, this is often not the case for social media. In this section, we focus on the problem of spatial uncertainty due to noisy or incomplete social media data. A prime example of spatial uncertainty is the lack of location metadata in posts from Twitter (called tweets).[1] In recent data we collected, only 3% of tweets contain the latitude and longitude at which they were created. Another 47% contain a valid user self-declared location in his or her profile (e.g., "New York, NY"). However, such a location does not automatically change when the user travels and thus may not be the true location at which a tweet is posted. The remaining 50% do not contain location at all. Clearly, we cannot reliably assign the latter two kinds of tweets to a spatiotemporal source bin. [2]

To address this issue, we borrow an idea from Positron Emission Tomography (Vardi et al., 1985). In particular, we define m *detector bins*, which are conceptually distinct from the n source bins. The idea is that an event originating in some

[1] It may be possible to recover occasional location information from the tweet text itself instead of the metadata, but the problem still exists.

[2] Another kind of spatiotemporal uncertainty exists in social media even when the local and time metadata of every post is known: social media users may not immediately post right at the spot where a target phenomenon happens. Instead, there usually is an unknown time delay and spatial shift between the phenomenon and the post generation. For example, one may not post a squirrel encounter on the road until she arrives at home later; the local and time metadata only reflects tweet-generation at home. This type of spatiotemporal uncertainty can be addressed by the same source-detector transition model.

source bin goes through a transition process and ends up in one of the detector bins, where it is detected. This transition is modeled by an $m \times n$ matrix $\mathbf{P} = \{P_{ij}\}$ where

$$P_{ij} = \Pr(\text{detector } i \mid \text{source } j). \tag{3.2}$$

$\mathbf{P}$ is column stochastic: $\sum_{i=1}^{m} P_{ij} = 1, \forall j$. We defer the discussion of our specific $\mathbf{P}$ to a case study, but we mention that it is possible to reliably estimate $\mathbf{P}$ directly from social media data (more on this later). Recall the target post intensity at source bin j is $\eta(f_j, g_j)$. We use the transition matrix to define the target post intensity h_i (note that h_i can itself be viewed as a link function $\tilde{\eta}(\mathbf{f}, \mathbf{g})$) at detector bin i:

$$h_i = \sum_{j=1}^{n} P_{ij} \eta(f_j, g_j). \tag{3.3}$$

For the spatial uncertainty that we consider, we create three kinds of detector bins. For a source bin j such as "California, day 1," the first kind collects target posts on day 1 whose latitude and longitude metadata is in California. The second kind collects target posts on day 1 without latitude and longitude metadata, but whose user self-declared profile location is in California. The third kind collects target posts on day 1 without any location information. Note the third kind of detector bin is shared by all other source bins for day 1, such as "Nevada, day 1," too. Consequently, if we had $n = 50T$ source bins corresponding to the 50 US states over T days, there would be $m = (2 \times 50 + 1)T$ detector bins.

Critically, our observed target counts $\mathbf{x}$ are now with respect to the m detector bins instead of the n source bins: $\mathbf{x} = (x_1, \ldots, x_m)^\top$. We will also denote the count sub-vector for the first kind of detector bins by $\mathbf{x}^{(1)}$, the second kind $\mathbf{x}^{(2)}$, and the third kind $\mathbf{x}^{(3)}$. The same is true for the overall counts $\mathbf{z}$. A trivial approach is to only utilize $\mathbf{x}^{(1)}$ and $\mathbf{z}^{(1)}$ to arrive at the plug-in estimator

$$\widehat{f_j} = x_j^{(1)} / z_j^{(1)}. \tag{3.4}$$

As we will show, we can obtain a better estimator by incorporating noisy data $\mathbf{x}^{(2)}$

and incomplete data $\mathbf{x}^{(3)}$. $\mathbf{z}^{(1)}$ is sufficiently large so we will simply ignore $\mathbf{z}^{(2)}$ and $\mathbf{z}^{(3)}$.

Socioscope: Penalized Poisson Likelihood Model

We observe target post counts $\mathbf{x} = (x_1, \ldots, x_m)$ in the detector bins. These are modeled as independent Poisson-distributed random variables:

$$x_i \sim \text{Poisson}(h_i), \text{ for } i = 1 \ldots m. \tag{3.5}$$

The log likelihood factors are

$$\ell(\mathbf{f}) = \log \prod_{i=1}^{m} \frac{h_i^{x_i} e^{-h_i}}{x_i!} = \sum_{i=1}^{m} (x_i \log h_i - h_i) + c, \tag{3.6}$$

where c is a constant. In (3.6) we treat g as given.

Target posts may be scarce in some detector bins. Indeed, we often have zero target posts for the wildlife case study to be discussed later. This problem can be mitigated by the fact that many real-world phenomena are spatiotemporally smooth, i.e., "neighboring" source bins in space or time tend to have similar intensity. We therefore adopt a penalized likelihood approach by constructing a graph-based regularizer. The undirected graph is constructed so that the nodes are the source bins. Let $\mathbf{W}$ be the $n \times n$ symmetric non-negative weight matrix. The edge weights are such that W_{jk} is large if j and k correspond to neighboring bins in space and time. Since $\mathbf{W}$ is domain specific, we defer its construction to the case study.

Before discussing the regularizer, we need to perform a change of variables. Poisson intensity $\mathbf{f}$ is non-negative, necessitating a constrained optimization problem. It is more convenient to work with an unconstrained problem. To this end, we work with the exponential family natural parameters of Poisson. Specifically, let

$$\theta_j = \log f_j, \quad \psi_j = \log g_j. \tag{3.7}$$

Our specific link function becomes $\eta(\theta_j, \psi_j) = e^{\theta_j + \psi_j}$. The detector bin intensities

become $h_i = \sum_{j=1}^{n} P_{ij}\eta(\theta_j, \psi_j)$.

Our graph-based regularizer applies to θ directly:

$$\Omega(\theta) = \frac{1}{2}\theta^\top \mathbf{L}\theta, \tag{3.8}$$

where $\mathbf{L}$ is the combinatorial graph Laplacian (Chung, 1997): $\mathbf{L} = \mathbf{D} - \mathbf{W}$, and $\mathbf{D}$ is the diagonal degree matrix with $D_{jj} = \sum_{k=1}^{n} W_{jk}$.

Finally, Socioscope is the following penalized likelihood optimization problem:

$$\min_{\theta \in \mathbb{R}^n} -\sum_{i=1}^{m} (x_i \log h_i - h_i) + \lambda\Omega(\theta), \tag{3.9}$$

where λ is a positive regularization weight.

Optimization

We solve the Socioscope optimization problem (3.9) with BFGS, a quasi-Newton method (Nocedal and Wright, 1999). The gradient can be computed by

$$\nabla = \lambda\mathbf{L}\theta - \mathbf{H}\mathbf{P}^\top(\mathbf{r} - 1), \tag{3.10}$$

where $\mathbf{r} = (r_1 \dots r_m)$ is a ratio vector with $r_i = x_i/h_i$, and $\mathbf{H}$ is a diagonal matrix with $H_{jj} = \eta(\theta_j, \psi_j)$.

We initialize θ with the following heuristic. Given counts $\mathbf{x}$ and the transition matrix P, we compute the least-squared projection η_0 to $\|\mathbf{x} - \mathbf{P}\eta_0\|_2$. This projection is easy to compute. However, η_0 may contain negative components not suitable for Poisson intensity. We force positivity by setting $\eta_0 \leftarrow \max(10^{-4}, \eta_0)$ element-wise, where the floor 10^{-4} ensures that $\log \eta_0 > -\infty$. From the definition, $\eta(\theta, \psi) = \exp(\theta + \psi)$, we then obtain the initial parameter

$$\theta_0 = \log \eta_0 - \psi. \tag{3.11}$$

Our optimization is efficient: problems with more than one thousand variables

(n) are solved in about 15 seconds with fminunc() in Matlab.

Parameter Tuning

The choice of the regularization parameter λ has a profound effect on the smoothness of the estimates. It may be possible to select these parameters based on prior knowledge in certain problems, but for our experiments we select these parameters using a cross-validation (CV) procedure, which gives us a fully data-driven and objective approach to regularization.

CV is quite simple to implement in the Poisson setting. A hold-out set of data can be constructed by simply sub-sampling events from the total observation uniformly at random. This produces a partial data set of a subset of the counts that follows precisely the same distribution as the whole set, modulo a decrease in the total intensity per the level of subsampling. The complement of the hold-out set is what remains of the full dataset, and we will call this the training set. The hold-out set is taken to be a specific fraction of the total. For theoretical reasons beyond the scope of this work, we do not recommend leave-one-out CV (Van Der Laan and Dudoit, 2003; Cornec, 2010).

CV is implemented by generating a number of random splits of this type (we can generate as many as we wish), and for each split we run the optimization algorithm above on the training set for a range of values of λ. Then compute the (unregularized) value of the log-likelihood on the hold-out set. This provides us with an estimate of the log-likelihood for each setting of λ. We then select the setting that maximizes the estimated log-likelihood.

Theoretical Considerations

The natural measure of signal-to-noise in this problem is the number of counts in each bin. The higher the counts, the more stable and "less noisy" our estimators will be. Indeed, if we directly observe $x_i \sim \text{Poisson}(h_i)$, then the normalized error $\mathbf{E}[(\frac{x_i - h_i}{h_i})^2] = h_i^{-1} \approx x_i^{-1}$. So larger counts, due to larger underlying intensities, lead to small errors on a relative scale. However, the accuracy of our recovery also

depends on the regularity of the underlying function f. If it is very smooth, for example a constant function, then the error would be inversely proportional to the total number of counts, not the number in each individual bin. This is because in the extreme smooth case, f is determined by a single constant.

To give some insight into dependence of the estimate on the total number of counts, suppose that f is the underlying continuous intensity function of interest. Furthermore, let f be a Hölder α-smooth function. The parameter α is related to the number of continuous derivatives f has. Larger values of α correspond to smoother functions. Such a model is reasonable for the application at hand, as discussed in our motivation for regularization above. We recall the following minimax lower bound, which follows from the results in (Donoho et al., 1996; Willett and Nowak, 2007).

Theorem 3.1. *Let f be a Hölder α-smooth d-dimensional intensity function and suppose we observe N events from the distribution Poisson(f). Then there exists a constant $C_\alpha > 0$ such that*

$$\inf_{\widehat{f}} \sup_{f} \frac{\mathbf{E}[\|\widehat{f} - f\|_1^2]}{\|f\|_1^2} \geqslant C_\alpha N^{\frac{-2\alpha}{2\alpha+d}},$$

where the infimum is over all possible estimators.

The error is measured with the 1-norm, rather than two norm, which is a more appropriate and natural norm in density and intensity estimation. The theorem tells us that no estimator can achieve a faster rate of error decay than the bound above. There exist many types of estimators that nearly achieve this bound (e.g., to within a log factor), and with more work it is possible to show that our regularized estimators, with adaptively chosen bin sizes and appropriate regularization parameter settings, could also nearly achieve this rate. For the purposes of this discussion, the lower bound, which certainly applies to our situation, will suffice.

For example, consider just two spatial dimensions ($d = 2$) and $\alpha = 1$ which corresponds to Lipschitz smooth functions, a very mild regularity assumption. Then the bound says that the error is proportional to $N^{-1/2}$. This gives useful insight into the minimal data requirements of our methods. It tells us, for example,

that if we want to reduce the error of the estimator by a factor of say 2, then the total number of counts must be increased by a factor of 4. If the smoothness α is very large, then doubling the counts can halve the error. The message is simple: More events and higher counts will provide more accurate estimates.

Related Work

To our knowledge, there is no prior work that focuses on robust signal recovery from social media (i.e., the "second stage" as mentioned at the beginning of this chapter). However, there has been considerable related work on the first stage of identifying target social media posts, which we summarize below.

Topic detection and tracking (TDT) aims at identifying emerging topics from text streams and grouping documents based on their topics. The early work in this direction began with news text streamed from newswire services and transcribed from other media (Allan, 2002). Recent research focused on user-generated content on the web and on the spatiotemporal variation of topics. Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is a popular unsupervised method to detect topics. Mei *et al.* (2006) extended LDA by taking spatiotemporal context into account to identify subtopics from weblogs. They analyzed the spatio-temporal pattern of topic θ by $\Pr(\mathtt{time}|\theta, \mathtt{location})$ and $\Pr(\mathtt{location}|\theta, \mathtt{time})$, and showed that documents created from the same spatiotemporal context tend to share topics. In the same spirit, Yin *et al.* (2011) studied GPS-associated documents, whose coordinates are generated by Gaussian Mixture Model in their generative framework. Cataldi *et al.* (2010) proposed a *feature-pivot* method. They first identified keywords whose occurrences dramatically increase in a specified time interval and then connected the keywords to detect emerging topics. Besides text, social network structure also provides important information for detecting community-based topics and user interests.

Event detection is highly related to TDT. Yang *et al.* (1998) used a clustering algorithm to identify events from news streams. Others tried to distinguish posts related to real world events from posts about non-events, such as describing daily

life or emotions (Becker et al., 2011). Real world events were also detected in Flickr photos with meta information and Twitter. Other researchers were interested in events with special characteristics, such as controversial events and local events. Sakaki *et al.* (2010) monitored Twitter to detect real-time events such as earthquakes and hurricanes.

Another line of related work used social media as a data source to answer scientific questions (Lazer et al., 2009). Most previous work studied questions in linguistic, sociology and human interactions. For example, Eisenstein *et al.* (2010) studied the geographic linguistic variation with geotagged social media. Gupte *et al.* (2011) studied social hierarchy and stratification in online social network.

As stated earlier, Socioscope differs from past work in its focus on robust signal recovery on predefined target phenomena. The target posts may be generated at a very low, though sustained, rate, and are corrupted by noise. The above approaches are unlikely to estimate the underlying intensity accurately.

A Synthetic Experiment

We start with a synthetic experiment whose known ground-truth intensity $\mathbf{f}$ allows us to quantitatively evaluate the effectiveness of Socioscope. The synthetic experiment matches the case study in the next section. There are 48 US continental states plus Washington DC, and $T = 24$ hours. This leads to a total of $n = 1,176$ source bins, and $m = (2 \times 49 + 1)T = 2,376$ detector bins. The transition matrix $\mathbf{P}$ is the same as in the case study, to be discussed later. The overall counts $\mathbf{z}$ are obtained from actual Twitter data and $\widehat{\mathbf{g}} = \mathbf{z}^{(1)}$.

We design the ground-truth target signal $\mathbf{f}$ to be temporally constant but spatially varying. Figure 3.5(a) shows the ground-truth $\mathbf{f}$ spatially. It is a mixture of two Gaussian distributions discretized at the state level. The modes are in Washington and New York, respectively. From $\mathbf{P}$, $\mathbf{f}$ and $\mathbf{g}$, we generate the observed target post counts for each detector bin by a Poisson random number generator: $x_i \sim \text{Poisson}(\sum_{j=1}^{n} P_{i,j} f_j g_j)$, $i = 1 \dots m$. The sum of counts in $\mathbf{x}^{(1)}$ is 56, in $\mathbf{x}^{(2)}$ 1,106, and in $\mathbf{x}^{(3)}$ 1,030.

(i)	scaled $\mathbf{x}^{(1)}$	14.11
(ii)	scaled $\mathbf{x}^{(1)}/\mathbf{z}^{(1)}$	46.73
(iii)	Socioscope with $\mathbf{x}^{(1)}$	0.17
(iv)	Socioscope with $\mathbf{x}^{(1)} + \mathbf{x}^{(2)}$	1.83
(v)	Socioscope with $\mathbf{x}^{(1)}$, $\mathbf{x}^{(2)}$	0.16
(vi)	**Socioscope with $\mathbf{x}^{(1)}$, $\mathbf{x}^{(2)}$, $\mathbf{x}^{(3)}$**	**0.12**

Table 3.1: Relative error of different estimators

Given $\mathbf{x}, \mathbf{P}, \mathbf{g}$, we compare the relative error $\|\mathbf{f} - \hat{\mathbf{f}}\|^2/\|\mathbf{f}\|^2$ of several estimators in Table 3.1:

(i) $\hat{\mathbf{f}} = \mathbf{x}^{(1)}/(\epsilon_1 \sum \mathbf{z}^{(1)})$, where ϵ_1 is the fraction of tweets with precise location stamp (discussed later in the case study). Scaling matches it to the other estimators. Figure 3.5(b) shows this simple estimator, aggregated spatially. It is a poor estimator: besides being non-smooth, it contains 32 "holes" (states with zero intensity, colored in blue) due to data scarcity.

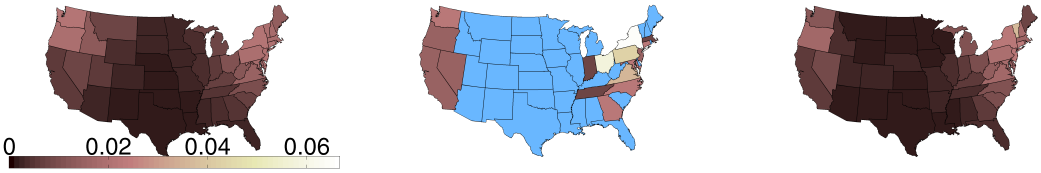
(ii) $\hat{\mathbf{f}} = \mathbf{x}_j^{(1)}/(\epsilon_1 \mathbf{z}_j^{(1)})$ which naively corrects the population bias as discussed in (3.4). It is even worse than the simple estimator, because naive bin-wise correction magnifies the variance in sparse $\mathbf{x}^{(1)}$.

(iii) Socioscope with $\mathbf{x}^{(1)}$ only. This simulates the practice of discarding noisy or incomplete data, but regularizing for smoothness. The relative error was reduced dramatically.

(iv) Same as (iii) but replace the values of $\mathbf{x}^{(1)}$ with $\mathbf{x}^{(1)} + \mathbf{x}^{(2)}$. This simulates the practice of ignoring the noise in $\mathbf{x}^{(2)}$ and pretending it is precise. The result is worse than (iii), indicating that simply including noisy data may hurt the estimation.

(v) Socioscope with $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ separately, where $\mathbf{x}^{(2)}$ is treated as noisy by $\mathbf{P}$. It reduces the relative error further, and demonstrates the benefits of treating noisy data specially.

(vi) Socioscope with the full $\mathbf{x}$. It achieves the lowest relative error among all methods, and is the closest to the ground truth (Figure 3.5(c)). Compared to (v), this demonstrates that even counts $\mathbf{x}^{(3)}$ without location can also help us recover $\mathbf{f}$ better.



(a) ground-truth $\mathbf{f}$ (b) scaled $\mathbf{x}^{(1)}$ (c) Socioscope

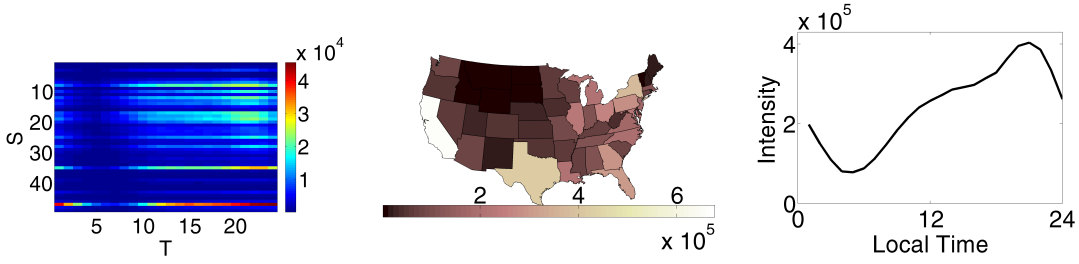
Figure 3.5: The synthetic experiment

Case Study: Roadkill

Before we apply Socioscope to bullying, we want to first apply it to a task where the signal is likely to be strong and interpretable.

We were unaware of public benchmark data sets to test robust signal recovery from social media. Several social media datasets were released recently, such as the ICWSM data challenges and the TREC microblog track. These datasets were intended to study trending "hot topics" such as the Arab Spring, Olympic Games, or presidential elections. They are not suitable for low intensity sustained target phenomena, which is the focus of our approach. In particular, these datasets do not contain ground-truth spatiotemporal intensities and thus are not appropriate testbeds for the problems we are trying to address. Instead, we report on a real-world case study on the spatiotemporal intensity of roadkill for several common wildlife species from Twitter posts.

The study of roadkill has value in ecology, conservation, and transportation safety. The target phenomenon consists of roadkill events for a specific species within the continental United States during the period September 22–November 30, 2011. Our spatiotemporal source bins are state$\times$hour-of-day. Let s index the 48 continental US states plus District of Columbia. We aggregate the 10-week study period into the 24 hours in a day. The target counts $\mathbf{x}$ are still sparse even with aggregation: for example, most state-hour combinations have zero counts for armadillo and the largest count in $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ is 3. Therefore, recovering the



(a) $\widehat{\mathbf{g}}$ (b) spatial (c) temporal

Figure 3.6: Human population intensity $\widehat{\mathbf{g}}$.

underlying signal $\mathbf{f}$ remains a challenge. Let t index the hours from 1 to 24. This results in $|s| = 49, |t| = 24, n = |s||t| = 1,176, m = (2|s|+1)|t| = 2,376$. We will often index source or detector bins by the subscript (s,t), in addition to i or j, below. The translation should be obvious.

Data Preparation

We chose Twitter as our data source because public tweets can be easily collected through its APIs. All tweets include time metadata. However, most tweets do not contain location metadata, as discussed earlier.

Overall Counts $\mathbf{z}^{(1)}$ and Human Population Intensity $\mathbf{g}$.

To obtain the overall counts $\mathbf{z}$, we collected tweets through the Twitter stream API using bounding boxes covering the continental U.S.. The API supplied a subsample of *all* tweets (not just target posts) with geo-tags. Therefore, all these tweets included precise latitude and longitude on where they were created. Through a reverse geocoding database (`http://www.datasciencetoolkit.org`), we mapped the coordinates to a US state. There are a large number of such tweets. Counting the number of tweets in each state-hour bin gave us $\mathbf{z}^{(1)}$, from which $\mathbf{g}$ is estimated.

Figure 3.6 shows the estimated $\widehat{\mathbf{g}}$. The x-axis is hour of day and the y-axis is the states, ordered by longitude from east (top) to west (bottom). Although $\widehat{\mathbf{g}}$ in this matrix form contains full information, it can be hard to interpret. Therefore, we visualize aggregated results as well: First, we aggregate out time in $\widehat{\mathbf{g}}$: for each state

s, we compute $\sum_{t=1}^{24} \widehat{g_{s,t}}$ and show the resulting intensity maps in Figure 3.6(b). Second, we aggregate out state in $\widehat{\mathbf{g}}$: for each hour of day t, we compute $\sum_{s=1}^{49} \widehat{g_{s,t}}$ and show the daily curve in Figure 3.6(c). From these two plots, we clearly see that human population intensity varies greatly both spatially and temporally.

Identifying Target Posts to Obtain Counts x.

To produce the target counts $\mathbf{x}$, we need to first identify target posts describing roadkill events. Although not part of Socioscope, we detail this preprocessing step here for reproducibility.

In step 1, we collected tweets using a keyword API. Each tweet must contain the wildlife name (e.g., "squirrel(s)") *and* the phrase "ran over". We obtained 5,857 squirrel tweets, 325 chipmunk tweets, 180 opossum tweets and 159 armadillo tweets during the study period. However, many such tweets did not actually describe roadkill events. For example, *"I almost ran over an armadillo on my longboard, luckily my cat-like reflexes saved me."* Clearly, the author did not kill the armadillo.

In step 2, we built a binary text classifier to identify target posts. Following (Settles, 2011), the tweets were case-folded without any stemming or stopword removal. Any user mentions preceded by a "@" were replaced by the anonymized user name "@USERNAME". Any URLs staring with "http" were replaced by the token "HTTPLINK". Hashtags (compound words following "#") were not split and were treated as a single token. Emoticons, such as ":)" or ":D", were also included as tokens. Each tweet is then represented by a feature vector consisting of unigram and bigram counts. If any unigram or bigram included animal names, we added an additional feature by replacing the animal name with the generic token "ANIMAL". For example, we would created an extra feature "over ANIMAL" for the bigram "over raccoon". The training data consists of 1,450 manually labeled tweets in August 2011 (i.e., *outside* our study period). These training tweets contain hundreds of animal species, not just the target species. The binary label is whether the tweet is a true first-hand roadkill experience. We trained a linear Support Vector Machine (SVM). The CV accuracy was nearly 90%. We then applied this SVM to classify tweets surviving step 1. Those tweets receiving a positive label were treated as target posts.

In step 3, we produce $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}$ counts. Because these target tweets were collected by the keyword API, the nature of the Twitter API means that most do not contain precise location information. As mentioned earlier, only 3% of them contain coordinates. We processed this 3% using the same reverse geocoding database to map them to a US state s, and place them in the $x^{(1)}_{s,t}$ detection bins. 47% of the target posts did not contain coordinates but can be mapped to a US state from the user's self-declared profile location. These are placed in the $x^{(2)}_{s,t}$ detection bins. The remaining 50% contained no location metadata, and were placed in the $x^{(3)}_{t}$ detection bins. [3]

Constructing the Transition Matrix P. In this study, **P** characterizes the fraction of tweets that were actually generated in source bin (s,t) and end up in the three detector bins: precise location $st^{(1)}$, potentially noisy location $st^{(2)}$, and missing location $t^{(3)}$. We define **P** as follows:

$P_{(s,t)^{(1)},(s,t)} = 0.03$, and $P_{(r,t)^{(1)},(s,t)} = 0$ for $\forall r \neq s$ to reflect the fact that we know precisely 3% of the target posts' location.

$P_{(r,t)^{(2)},(s,t)} = 0.47 M_{r,s}$ for all r, s. M is a 49×49 "mis-self-declare" matrix. $M_{r,s}$ is the probability that a user self-declares in her profile that she is in state r, but her post is in fact generated in state s. We estimated **M** from a separate large set of tweets with both coordinates and self-declared profile locations. The **M** matrix is asymmetric and interesting in its own right: many posts self-declared in California or New York were actually produced all over the country; many self-declared in Washington DC were actually produced in Maryland or Virgina; more posts self-declare Wisconsin but were actually in Illinois than the other way around.

$P_{t^{(3)},(s,t)} = 0.50$. This aggregates tweets with missing information into the third kind of detector bins.

Specifying the Graph Regularizer. Our graph has two kinds of edges. Temporal edges connect source bins with the same state and adjacent hours by weight w_t. Spatial edges connect source bins with the same hour and adjacent states by weight w_s. The regularization weight λ was absorbed into w_t and w_s. We tuned

[3] There were actually only a fraction of all tweets without location which came from all over the world. We estimated this US/World fraction using **z**.

the weights w_t and w_s with CV on the 2D grid $\{10^{-3}, 10^{-2.5}, \ldots, 10^{3}\}^2$.

Results

We present results on four animals: armadillos, chipmunks, squirrels, and opossums. Perhaps surprisingly, precise roadkill intensities for these animals are apparently unknown to science (This serves as a good example of the value Socioscope may provide to wildlife scientists). Instead, domain experts were only able to provide a range map of each animal; see the left column in Figure 3.7. These maps indicate presence/absence only, and were extracted from NatureServe (Patterson et al., 2007). In addition, the experts defined armadillo and opossum as nocturnal, chipmunk as diurnal, and squirrels as both crepuscular (active primarily during twilight) and diurnal. Due to the lack of quantitative ground-truth, our comparison will necessarily be qualitative in nature.

Socioscope provides sensible estimates on these animals. For example, Figure 3.8(a) shows counts $\mathbf{x}^{(1)} + \mathbf{x}^{(2)}$ for chipmunks which is very sparse (the largest count in any bin is 3), and Figure 3.8(b) the Socioscope estimate $\widehat{\mathbf{f}}$. The axes are the same as in Figure 3.6(a). In addition, we present the state-by-state intensity maps in the middle column of Figure 3.7 by aggregating $\widehat{\mathbf{f}}$ spatially. The Socioscope results match the range maps well for all animals. The right column in Figure 3.7 shows the daily animal activities by aggregating $\widehat{\mathbf{f}}$ temporally. These curves match the animals' diurnal patterns well, too.

The Socioscope estimates are superior to the baseline methods in Table 3.1. Due to space limits we only present two examples on chipmunks, but note that similar observations exist for all animals. The baseline estimator of simply scaling $\mathbf{x}^{(1)} + \mathbf{x}^{(2)}$ produced the temporal and spatial aggregates in Figure 3.9(a,b). Compared to Figure 3.7(b, right), the temporal curve has a spurious peak around 4-5pm. The spatial map contains spurious intensity in California and Texas, states outside the chipmunk range as shown in Figure 3.7(b, left). Both are produced by population bias when and where there were strong background social media activities (see Figure 3.6(b,c)). In addition, the spatial map contains 27 "holes" (states with zero

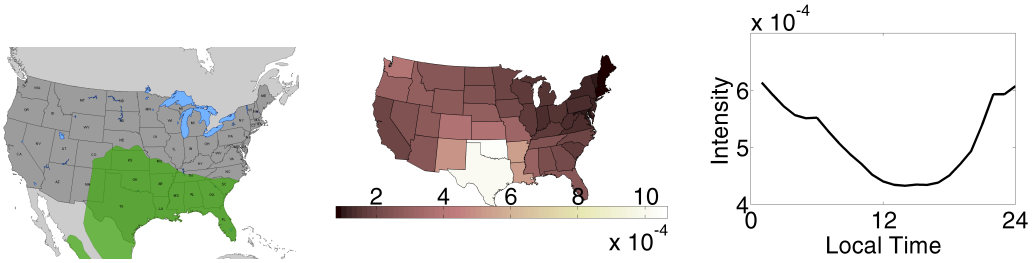
intensity, colored in blue) due to data scarcity. In contrast, Socioscope's estimates in Figure 3.7 avoid this problem by regularization. Another baseline estimator $(\mathbf{x}^{(1)} + \mathbf{x}^{(2)})/\mathbf{z}^{(1)}$ is shown in Figure 3.9(c). Although corrected for population bias, this estimator lacks the transition model and regularization. It does not address data scarcity either.

Applying Socioscope to Bullying Hashtags

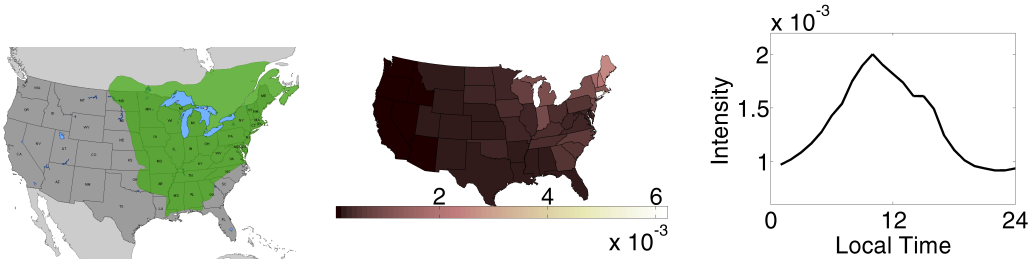
In Section 3.1 and Section 3.2 we studied the spatiotemporal distribution of bullying traces, in which social media users post about their personal experiences about bullying. Besides bullying traces, social media users talk about bullying in many other ways, for example, recent news stories, general opinions about bullying, raising awareness of anti-bullying campaigns, and so on. In this section, we use Socioscope to study the spatiotemporal variations of some special topics about bullying in social media.

Data Preparation

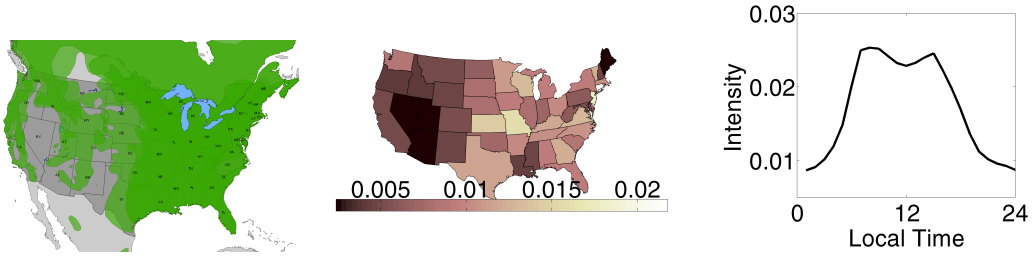
Instead of training a text classifier to recognize target posts, we use hashtags as an indicator for relevance. Hashtags are keywords or acronyms that are prefixed with a # symbol that are annotated within tweets to indicate markers of topic. In Chapter 5, we will investigate the hashtag usages in bullying posts in more detail. We collected data from the public Twitter streaming API between January 1, 2012 and December 31, 2012. We captured tweets that contained at least one of the following keywords: "bully," "bullied," and "bullying" through the Twitter streaming API. We did not remove retweets or apply the Binary Bullying Traces Classifier to the data, as retweets and general comments are useful to understand trending topics and opinions. We case-folded all hashtags (i.e., we replaced upper case letters with lower case ones) to merge different variations of the same hashtag into a single hashtag. For example, "#RIPAmandaTodd", "#ripAmandaTodd", "#ripamandatodd" were all transformed to "#ripamandatodd." So, each tweet in our dataset contains at least one of bullying related keywords and at least one



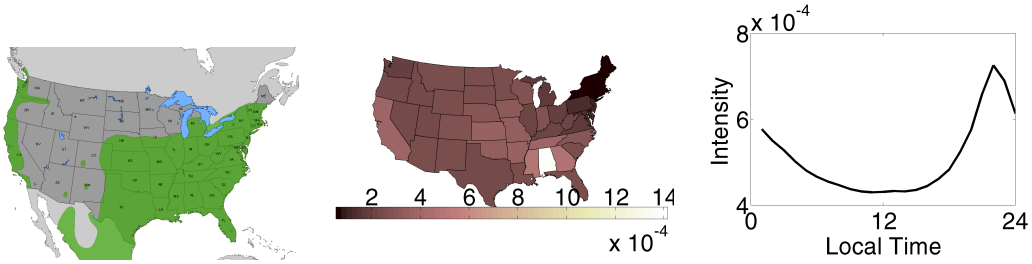
(a) armadillo (Dasypus novemcinctus)



(b) chipmunk (Tamias striatus)

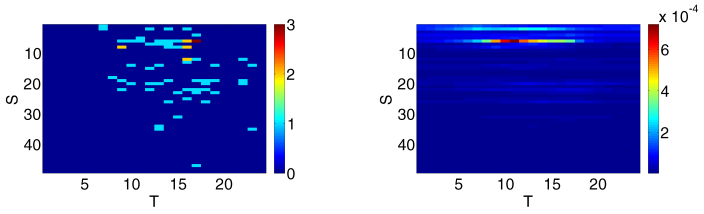


(c) squirrel (Sciurus carolinensis and several others)



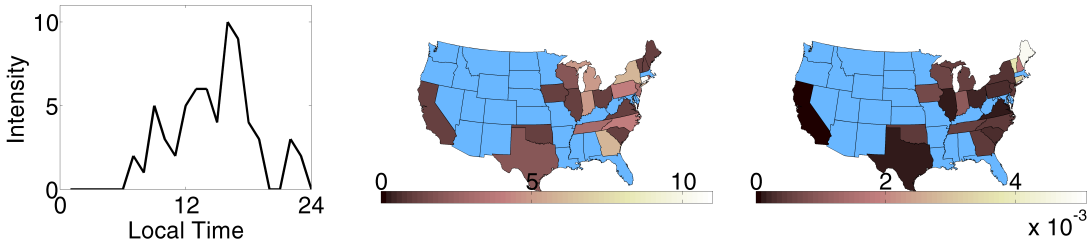
(d) opossum (Didelphis virginiana)

Figure 3.7: Socioscope estimates match animal habits well. (Left) range map from NatureServe, (Middle) Socioscope $\widehat{\mathbf{f}}$ aggregated spatially, (Right) $\widehat{\mathbf{f}}$ aggregated temporally.



(a) $\mathbf{x}^{(1)} + \mathbf{x}^{(2)}$ (b) Socioscope $\widehat{\mathbf{f}}$

Figure 3.8: Raw counts and Socioscope $\widehat{\mathbf{f}}$ for chipmunks



(a) $\mathbf{x}^{(1)} + \mathbf{x}^{(2)}$ (b) $\mathbf{x}^{(1)} + \mathbf{x}^{(2)}$ (c) $(\mathbf{x}^{(1)} + \mathbf{x}^{(2)})/\mathbf{z}^{(1)}$

Figure 3.9: Examples of inferior baseline estimators. In all plots, states with zero counts are colored in blue.

hashtag. This dataset is archived as Hashtags in the Bullying Traces Data Set (see Appendix A.6), and more details are discussed in Chapter 5.

We selected four hashtags from different categories defined in Chapter 5 to capture different usage scenarios. "#bullying" is a general term to refer to bullying and the most frequently used hashtag in our dataset. Users may use this term in many different ways in their tweets related to bullying. "#oomf" is an Everyday Twitter Trend term, as many social media users use it to refer "one of my follower." It is not directly related to bullying, and users may use it everyday. "#spiritday" is a hashtag related to a campaign that occurred on October 19, 2012 and refers to GLAAD's (Gay and Lesbian Alliance Against Defamation) anti-bullying campaign. "#ripamandatodd" is a hashtag referring to the suicide of Amanda Todd, a 15 year-old, who was a victim of cyberbullying.

For each hashtag, we consider all tweets containing the hashtag (after case-

folding) in our dataset collected by keyword filtering as target posts. For #bullying and #oomf, as users use them everyday, we used a longer study period, January 1 - 31, 2012, to see if we can identify some interesting patterns. For #spiritday and #ripamandatodd, as all hashtag usages are limited to a relative short time window, we chose a two-week study period, October 10-23, 2012. We chose different time windows, which may also discover different patterns between different times in the year. Our spatiotemporal source bins are state $\times$ day. Let s index the 48 continental US states plus District of Columbia. Let t index the days in our study period.

Overall Counts z and Social Media Population Intensity g.

Due to differences in population and hashtag usage across states and days, the raw counts may not reflect how much social media users engaged into a topic in each spatiotemporal bin. Therefore, we still need to control these potential biases. Instead of using all GPS tagged posts from a separate random tweet stream, we used all tweets collected in our dataset, in which each tweet contained at least one bullying keyword and one hashtag. So, the overall counts reflect the engagement of bullying topics and usage of hashtags.

We collected 332,350 tweets with at least one bullying keyword and one hashtag, during January 1 - 31, 2012. Among them, we have 1,423 tweets with GPS coordinates, and 62,028 tweets with identifiable U.S. states information. We collected 462,175 tweets during October 10 - 23, 2012, among which, 2,255 had GPS coordinates, and 90,022 had U.S. state information. Since the number of GPS-tagged tweets are not sufficient, we used all these tweets together to estimate social media population intensity. We used the same "mis-self-declare" matrix in our roadkill study and Socioscope to recover $\mathbf{g}$. The same parameter setting and tuning procedures were used.

Figure 3.10 shows the raw counts of posts with location information and the estimated results $\widehat{\mathbf{g}}$ by Socioscope. The recovered $\widehat{\mathbf{g}}$ looks similar to the raw counts $\mathbf{z}^{(1)} + \mathbf{z}^{(2)}$, as the counts are relatively large compared to the ones in the roadkill study. Figure 3.11 shows the aggregated results of $\widehat{\mathbf{g}}$. The spatial distribution seems relatively stable in two different study periods. High population states, such as California, New York and Texas, have higher intensities. The same trends appear

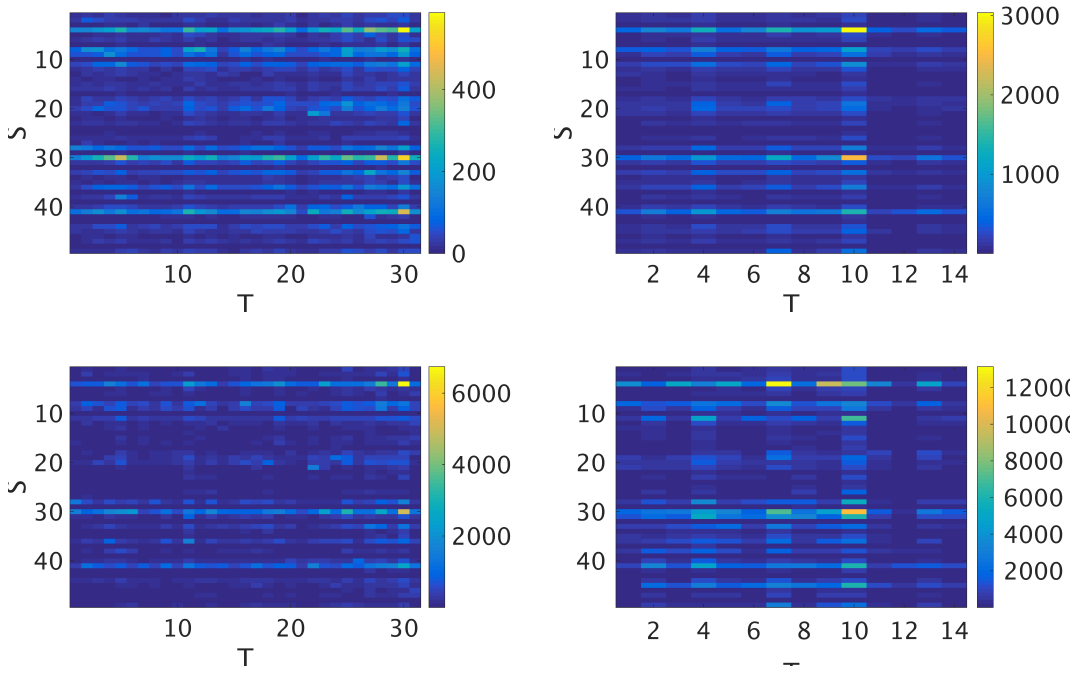
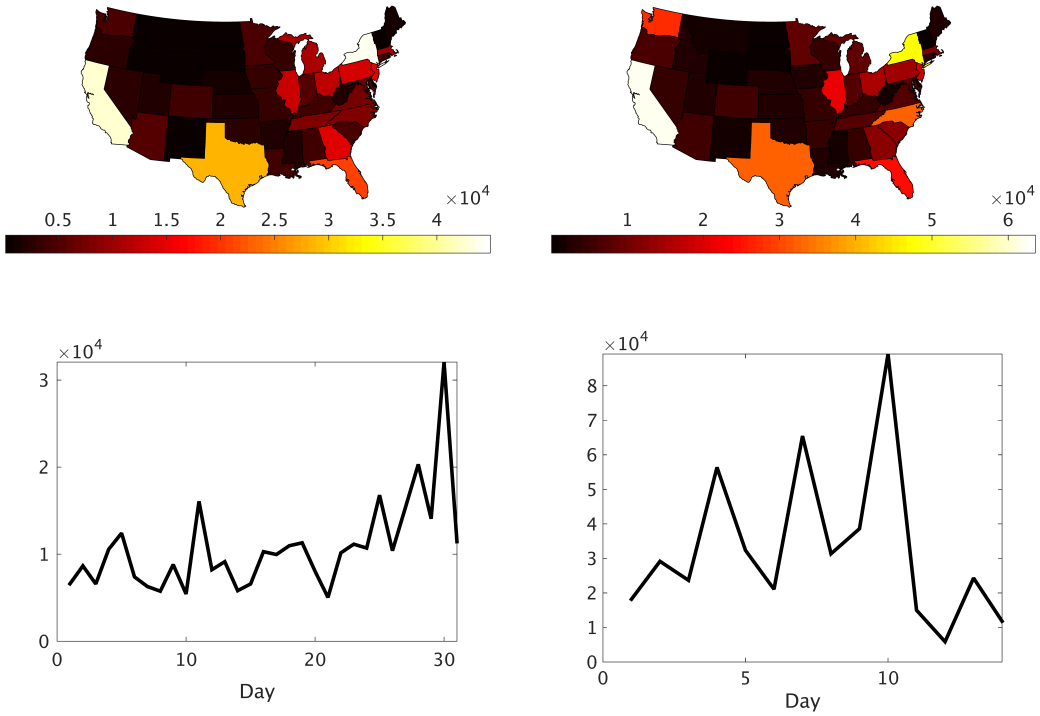


Figure 3.10: Raw counts $\mathbf{z}^{(1)} + \mathbf{z}^{(2)}$ and Socioscope $\widehat{\mathbf{g}}$ for bullying hashtags. The top row shows the raw counts for tweets with GPS coordinates or identifiable U.S. state information $\mathbf{z}^{(1)} + \mathbf{z}^{(2)}$. The bottom row shows the recovered estimation of $\widehat{\mathbf{g}}$ by Socioscope.

in both study periods. There seem to be more temporal variations. The weekly pattern is not obvious, but we can see local peaks appear repeatedly over a few days. At the end of January 2012, there was a huge jump in the number of tweets with bullying keyword and hashtags. Therefore, we should take variations of background population into accounts.

Obtaining Target Counts x and Confusion Matrix P.

We use the hashtag as an indicator for target posts on one topic. This is clean and does not need any supervision to build the classifiers. We produce $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}$ counts by direct counting. For #bullying, we collected 19,424 target posts and among them, 75 tweets are with GPS coordinates, 5,757 with state information.

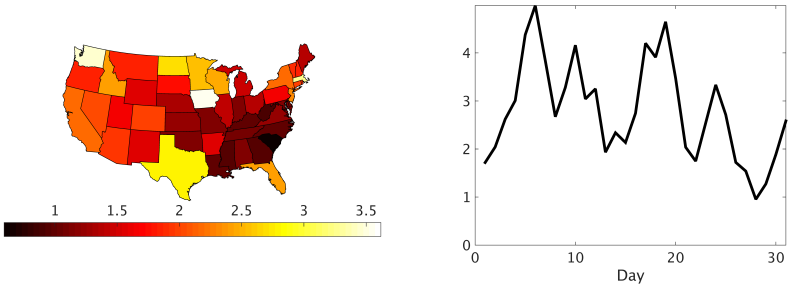


January 1-31, 2012 October 10-23, 2012

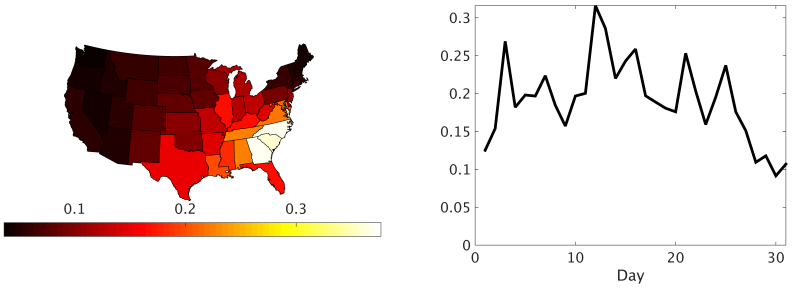
Figure 3.11: (Top) Socioscope $\widehat{\mathbf{g}}$ aggregated spatially, (Bottom) $\widehat{\mathbf{g}}$ aggregated temporally for bullying hashtag.

For #oomf, we collected 1,250 target posts in total, 19 GPS-tagged tweets and 281 tweets with user reported location. For #spiritday, we collected 66,854 target posts in total, 206 GPS-tagged tweets and 22,569 tweets with user reported location. For #ripamandatodd, we collected 45,186 target posts in total, 127 GPS-tagged tweets and 5,632 tweets with user reported location.

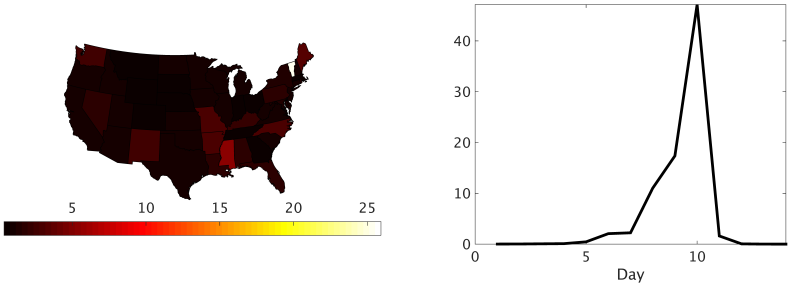
We use the same "mis-self-declare" matrix in our roadkill study to construct the confusion matrix $\mathbf{P}$. However, we use the actual proportions of tweets with GPS tags, tweets with user-reported locations and tweets without location information to set the actual weight for different parts of $\mathbf{P}$. For example, for hashtag #bullying, we set $P_{(s,t)^{(1)},(s,t)} = 79/19424 = 0.004$, $P_{(r,t)^{(2)},(s,t)} = 0.296M_{r,s}$, and $P_{t^{(3)},(s,t)} = 0.7$.



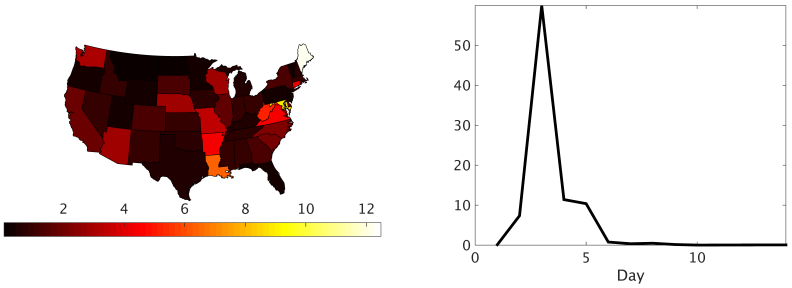
#bullying



#oomf



#spiritday



#ripamandatodd

Figure 3.12: Socioscope estimates $\widehat{\mathbf{f}}$ for bullying hashtags. (Left) Socioscope $\widehat{\mathbf{f}}$ aggregated spatially, (Right) $\widehat{\mathbf{f}}$ aggregated temporally.

Result.

Figure 3.12 shows the aggregated results of $\widehat{\mathbf{f}}$ estimated by Socioscope. #bullying was widely used across the U.S. with similar intensity. #oomf seems to be more popular in the southeaster U.S. than other regions. We have extremely high intensity in New Hampshire for #spiritday. But it is not a local event in New Hampshire. Gay and Lesbian Alliance Against Defamation Organization, based in New York, worked to promote this event, and we also received target posts from other states. #ripamandatodd also received attention from all over the country.

In Figure 3.10, we observe a huge jump at the end of January, but we do not see the same trend for #bullying and #ommf in Figure 3.12. Some special events/games have occurred at that time to raise the overall counts, but lowered the relative usages for these two general hashtags. On the other hand, #spiritday and #ripamandatodd show very high peaks in a short time periods, but nearly zero counts on other days, as they correspond to a special campaign and news story that happened just once in 2012. However, these two curves are different. #spiritday has a longer rising period before it peaked and then dropped dramatically. It is possible that organizers used social media to promote the awareness of the campaign several days ahead the event. So more and more users were involved in the topic. After the event, only a few users continued the discussion. However, for #ripamandatodd, the peak appears earlier and drops relatively slower. Most users were aware of the news on the first few days, and posted or retweeted using the hashtag. But it lasted longer and people continued discussing that for several days.

Discussion

Using social media as a data source for spatiotemporal signal recovery is an emerging area. Socioscope represents a first step toward this goal. There are many open questions:

1. We treated target posts as certain. In reality, a natural language processing system can often supply a confidence, e.g., Logistic Regression. For example, a tweet might be deemed to be a target post only with probability 0.8. It will be

interesting to study ways to incorporate such confidence into our framework.

2. The temporal delay and spatial displacement between the target event and the generation of a post is commonplace, as discussed in footnote 2. Estimating an appropriate transition matrix **P** from social media data so that Socioscope can handle such "point spread functions" remains future work.

3. It might be necessary to include psychological factors to better model human "sensors." For instance, as we will shown in Chapter 6, students in the western society and China have different school bullying behaviors and different biases on how to post them.

4. Instead of discretizing space and time into bins, one may adopt a spatial point process model to learn a continuous intensity function instead (Møller and Waagepetersen, 2004).

Addressing these considerations will further improve Socioscope.

4 EMOTIONS IN BULLYING TRACES

Sentiment analysis on bullying traces is of significant importance. Victims usually experience negative emotions such as depression, anxiety and loneliness. In extreme cases such emotions are more violent or even suicidal, for example,

> "I'm tired of all this bullying. I could never stand up for myself & sometimes I just want to kill myself."

Detecting at-risk individuals via sentiment analysis enables potential interventions. In addition, social scientists are interested in sentiment analysis on bullying traces to understand participants' motivations.

There are a wide range of emotions expressed in bullying traces. After manually inspecting a number of bullying traces in Twitter, our domain experts identified seven most common emotions:

1. Anger: *"He is always laughing at me because he is a bully damnit! #Ashley"*
2. Embarrassment: *"@USER everyone is bullying me because I couldn't find the word peach in a crossword. It's 1am"*
3. Empathy: *"@USER I'm sorry you get bullied. I'm really surprised at how many people this has happened to. #bulliesSuck "*
4. Fear: *"i was being bullied and i didn't want to go to school really i would throw fits everymorning and i hope that michel sees this"*
5. Pride: *"Everyone on this earth is a bully , except me . Because I'm perfect. #jillism"*
6. Relief: *"@USER I was rambling and then... I cried. Like, CRIED. He was touched! APC helped me thru the teasing and bullying man..."*
7. Sadness: *"things were bad when I was younger I got bullied so much because of my disabilites I don't want the same thing happening to my brother."*

This list is by no means comprehensive. Other emotions or mixtures of several basic emotions may also appear in bullying traces. These seven categories are the most common ones based on our current study. Also note that due to the length limits (140 characters), an individual tweet may be only a sentence in a conversation thread. Therefore, the majority of bullying traces in Twitter cannot be associated with definite emotions.

Orthogonal to these emotions, we observed that many bullying traces were written jokingly. One example of a teasing post is *"@USERNAME lol stop being a cyber bully lol :p."* Teasing may indicate the lack of severity of a bullying episode; It may also be a manifest of coping strategies in bullying victims. Therefore, there is considerable interest among social scientists to understand teasing in bullying traces.

Besides these emotions expressed in the text, we found that social media users who post bullying related tweets may later experience regret. After they post bullying related tweets, they may be aware of the potential risks, such as re-victimization, embarrassment, and social ostracization. They do not post a new post to express their regret. Instead, they may delete their original posts.

In this chapter, we first build a standard text classifier to recognized teasing in bullying traces (Xu et al., 2012b). Then we propose a fast machine learning procedure for the sentiment analysis in bullying traces (Xu et al., 2012c), as many categories are not well-studies in the community. To study the regret in bullying traces (Xu et al., 2013b), we construct a corpus of bullying tweets and periodically check the existence of each tweet in order to infer if and when it becomes deleted. We then conduct exploratory analysis in order to isolate factors associated with deleted posts. Finally, we propose the construction of a regrettable posts predictor to warn users if a tweet might cause regret.

4.1 Teasing in Bullying Traces

As it is of interest to social scientists, we first investigate teasing in bullying traces. We formulated it as a binary classification problem, similar to classic positive/neg-

	predicted as	
	Tease	Not
Tease	52	47
Not	26	559

Table 4.1: Confusion matrix of teasing classification

ative sentiment classification (Pang and Lee, 2004). Our annotators labeled each of the 684 bullying traces in Bullying Traces Data Set (version 1, Appendix A.1) as teasing (99) or not (585). We used the same feature representations, classifiers and parameter tuning as in Section 2.1 and 10-fold cross validation procedure. The classifier is archived as Teasing Bullying Trace Classifier (see Appendix B.6).

The best cross validation accuracy of 89% is obtained by `SVM(linear) + 1g2g`. This is significantly better than the majority class (not-teasing) baseline of 86% (t-test, $p = 10^{-33}$). It shows that even simple features and off-the-shelf classifier can detect some signal in the text. However, the accuracy is not high. Table 4.1 shows the confusion matrix. About half of the tease examples were misclassified. We found several possible explanations. First, teasing is not always accompanied by joking emoticons or tokens like "LOL," "lmao," "haha." For example, *"I may bully you but I love you lots. Just like jelly tots!"* and *"Been bullied into watching a scary film, I love my friends!"* Such teasing sentiment requires deeper NLP or much larger training sets. Second, tweets containing those joking emoticons and tokens are not necessarily teasing. For example, *"This Year I'm Standing Up For The Kids That Are Being Bullied All Over The Nation :) ."* Third, the joking tokens have diverse spellings. For example, "lol" was spelled as "loll," "lolol," "lollll," "loool," "LOOOOOOOOOOL"; "haha" was spelled as "HAHAHAHA," "Hahaha," "Bwahahaha," "ahahahah," "hahah."

Specialized word normalization for social media text may significantly improve performance. For example, word lengthening can be identified and used as cues for teasing (Brody and Diakopoulos, 2011). Teasing is diverse in its form and content. Our training set is perhaps too small. Borrowing training data from other corpora, such as one-liner jokes (Mihalcea and Strapparava, 2005), may be helpful.

4.2 A Fast Machine Learning Procedure for Sentiment Analysis on Bullying

As shown above, some emotions involved in bullying traces have not been well studied in sentiment analysis, for example, *embarrassment* and *relief*. To make the problem worse, manually labeling a large amount of training tweets is difficult and time consuming even for our domain experts.

Recognizing these challenges, we use a fast training procedure for sentiment analysis. Our goal is supervised learning, specifically classifying a tweet into one of the predefined emotion categories. However, we require no explicit labeled training data on tweets. Instead, we will rely on "distantly labeled data" (to be made clear next) that are much easier to obtain. We point out upfront that it will be difficult to assess the accuracy of the resulting classifier, since we do not have an in-domain labeled dataset. However, our observations point to a useful classifier. Coupled with the ease of training and its applicability to other emotions and domains, our procedure is still attractive.

Relations to Prior Work

Most sentiment analysis work focused on the overall polarity of a document: positive, negative or neutral (Pang and Lee, 2008; Liu and Zhang, 2012). A few works considered several basic emotions at a finer level and created emotional lexicons for each category (Strapparava and Valitutti, 2004). Recently, sentiment analysis on social media (Yang et al., 2007), especially Twitter (Pak and Paroubek, 2010), has been receiving increasing attention. Cambria *et al.* (2010) proposed a sentiment analysis approach to identify malicious posts from social media. Our domain of bullying is fresh with very few existing resources. In addition, although bullying traces are abundant, only a small fraction of them are associated with strong emotions. It poses challenges to obtain enough training examples for all the emotion categories, especially the rare and non-standard ones.

Our approach is inspired by the "concept labeling" work of Chenthamarakshan

et al. (2011) to minimize the supervision effort in constructing text classification models. In their system, instead of labeling a set of training examples, experts annotate how "concepts" are related to the target class. We push this idea further where neither labeled examples nor labeled concepts are necessary for building the emotion classifiers.

Our procedure consists of two steps. The first step is in the same spirit as the dictionary-based sentiment lexicon generation method (Hu and Liu, 2004), which exploits synonym structure of a dictionary to bootstrap the sentiment lexicon. Our second step is similar to the idea of corpus-based sentiment lexicon generation method (Hatzivassiloglou and McKeown, 1997; Kanayama and Nasukawa, 2006), which uses a domain corpus to extend sentiment lexicon by sentence structure or sentiment consistency assumption. As tweets are very short – usually a few sentences – the sentiment is usually consistent within a tweet.

Task Description

We obtain bullying traces identified by our Binary Bullying Trace Classifier. We want to recognize the emotion involved in each bullying trace. We define eight emotion classes: *anger*, *embarrassment*, *empathy*, *fear*, *pride*, *relief*, *sadness*, and *other*. The last class captures bullying traces without obvious emotion or not one of the seven emotions. Thus, our task is to build an eight-class Bullying Trace Emotion Classifier (see Appendix B.7) with little supervision.

Fast Learning

Our learning procedure includes four steps: (1) collecting seed words, (2) collecting online documents, (3) creating feature extractors, and (4) building a text classifier. None of the steps requires explicit labeling a corpus.

Collecting Seed Words. We start by collecting seed words S_e which are related to each emotion e (except for the *other* category). Lexicons exist for certain emotions such as *anger* and *sadness* but not all (Strapparava and Valitutti, 2004). As we want to

handle the non-standard emotion categories, we create such lists from two general resources which are available for all emotions:

1. Many websites provide synonym dictionary service.[1] We look up the category name of emotion e such as "anger" and add all its synonyms to S_e^{SYN}.

2. We search for the category name of emotion e in WordNet (Miller, 1995; Fellbaum, 1988), and add all words appearing in the synsets to S_e^{WN}. In addition, we also include all words in synsets listed as their "derivationally related form" and their "similar to," "full troponym" or "full hyponyms" sets depending on the part of speech (adjectives, verbs, or nouns).

By doing so, we collect two seed word lists S_e^{SYN} and S_e^{WN} for each emotion e. This step took less than half an hour manually. Note that it does not require any human judgments and can be implemented automatically if preferred.

Collecting Online Documents. We can broaden the coverage of the keywords by collecting documents containing them. We invoked Twitter search API to query each keyword and retrieve up to 100 recent tweets per query. We queried each word in S_e^{SYN} and S_e^{WN} separately, and obtained two tweet corpora T_e^{SYN} and T_e^{WN}. Obviously, other search services can be employed, too. Given the seed word list, this step can be automated without any human intervention.

Creating Feature Extractors. We perform stopword removal and stemming on T_e^{SYN} and T_e^{WN} as in Section 2.1. Our stopword list is based on the SMART system (Salton, 1971), augmented with domain specific stopwords such as "bully," "bullying," "bullied," "bullies," "@USER," "ref" and some punctuations. We then represent each tweet in T_e^{SYN}, T_e^{WN} by unigrams and bigrams features. We count the occurrences of each feature collectively within T_e^{SYN} or T_e^{WN} and remove features appearing less than five times. We define a vocabulary as the union of seed words in $S^{SYN} \cup S^{WN}$ and the remained features in $T^{SYN} \cup T^{WN}$.

[1] `http://www.synonyms.net/synonym`
`http://dico.isc.cnrs.fr/dico/en/search`
`http://dictionary.reference.com/`

With the vocabulary, we represent S_e^{SYN} as a feature vector v_e^{SYN} where the elements are the counts. We normalize the vector so that it has norm 1. Do the same for S_e^{WN}, T_e^{SYN}, and T_e^{WN} separately to obtain v_e^{WN}, v_e^{SYN}, and v_e^{WN}. Here we treat each of T_e^{SYN} and T_e^{WN} as a single large document. Furthermore, we treat the union $S_e^{SYN} \cup S_e^{WN} \cup T_e^{SYN} \cup T_e^{WN}$ as single document and compute its feature vector v_e^{all}. Thus, for each emotion e we have five feature vectors $\{v_e^{SYN}, v_e^{WN}, v_e^{SYN}, v_e^{WN}, v_e^{all}\}$. In total, we have 35 such feature vectors for the seven emotions.

We use these 35 vectors as feature extractors. Given a test document we apply the same text processing and represent it as feature vector d. We then compute the inner product

$$d^\top v$$

against each of the 35 feature extractors v above and obtain a 35-dimensional vector x. Clearly, no supervision from human is needed in this step either.

Building Bullying Trace Emotion Classifier This is the step where traditionally labeled bullying tweets are needed. Instead, we use easy-to-obtain distantly-labeled data. Though our domain is tweets, we train a text classifier on Wikipedia pages. Wikipedia API supports downloading pages matching a title or category name query. For each word in $S_e^{SYN} \cup S_e^{WN}$, we collect the retrieved Wikipedia pages. [2] Each such page is automatically labeled with emotion e. We therefore automatically constructed a labeled Wikipedia corpus with 964 pages.

We run each Wikipedia page in this corpus through our feature extractors to represent the page as a 35-dimensional vector. We train a standard seven-class SVM (note we do not model the "other" class yet) on the Wikipedia corpus. We compared linear and RBF kernels, tuned SVM regularization parameter C and γ in the RBF kernel function ($\exp(-\gamma\|x-y\|^2)$) in the grid $\{10^{-3}, 10^{-2}, \ldots, 10^3\}$ with 10-fold cross validation. The best model is obtained with RBF kernel, $C = 1000$ and $\gamma = 0.1$. On the Wikipedia corpus, it achieves a CV error of 15%. Its confusion matrix is shown in Table 4.2.

[2] It is important to note that the nature of the Wikipedia API means that the pages do not necessarily contain the query keywords, which enables us to learn something more than keyword matching.

Table 4.2: Confusion matrix of the seven-class SVM on the Wikipedia corpus

	predicted as						
	ang.	emb.	emp.	fear	pri.	rel.	sad.
ang.	112	0	0	9	0	2	3
emb.	0	21	0	3	0	2	1
emp.	1	0	7	7	0	1	2
fear	3	1	0	381	1	23	4
pri.	0	0	0	4	23	0	1
rel.	2	1	0	42	0	198	3
sad.	4	0	2	14	0	4	82

Model Evaluation and Usage

To understand the performance of the trained SVM, we compare it against three baseline methods. Note the comparison is based on the Wikipedia corpus, not the Twitter domain where we have no labeled data. Using the SVM on Twitter will be discussed at the end of this section.

The three baseline methods are:

1. S^{SYN}. For test document d, we compute the inner product $d^\top v_e^{SYN}$ for each emotion e and predict the class with the maximum value:

$$e^* = \arg\max_e d^\top v_e^{SYN}.$$

   Ties are randomly broken.

2. S^{WN}. Same as above, but use the WordNet keywords:

$$e^* = \arg\max_e d^\top v_e^{WN}.$$

   Both baselines are related to simple keyword matching.

3. Majority. All five feature extractors make their own predictions as above, and there is a majority vote among the five for the final decision. Again, ties are randomly broken.

Table 4.3: Cross validation error of different methods

Fast Training SVM	S^{SYN}	S^{WN}	Majority
0.15	0.31	0.43	0.42

Table 4.3 shows the cross validation error of these methods. The proposed fast training SVM achieves the lowest error.

However, the above results were all on the Wikipedia corpus. Recall that our test domain is Twitter, for which we do not have labeled data. When a test tweet comes, we first convert it into the 35-dimensional vector via the feature extractors and apply the trained SVM. We set a threshold τ on the margin output from SVM, whenever the largest margin is lower than τ, we predict it as *other*. Otherwise, we predict the label with the largest margin. The threshold τ is set manually by controlling the positive rate at 5% on a separate random tweet data set.

4.3 Emotion Distribution in Bullying Traces

We apply the Bullying Trace Emotion Classifier (see Appendix B.7) to 3,001,427 bullying traces from August 5, 2011 to April 12, 2012 (about eight months). The dataset and its documentation is archived as Bullying Trace Emotion data set (see Appendix A.4). Figure 4.1 shows the number of daily bullying traces in each emotion categories. Overall, the number of bullying traces is increasing because of growing social media usages. All emotion curves have the similar shape but different offset (note the y-axis is in log scale), indicating that the fraction of different emotions remain stable in the study period. The curves show a weekly (7-day) pattern, which we hypothesize is due to fewer direct interactions among students during the weekends. The few spikes are caused by celebrity events related to bullying which generated a large number of tweets. In what follows, we remove the few spike days since they are outliers.

We aggregate the counts over the study period for each category and compute their fractions over all bullying traces. Figure 4.2(a) shows that most (94%) bullying traces are not associated with obvious emotions, which matches our observations

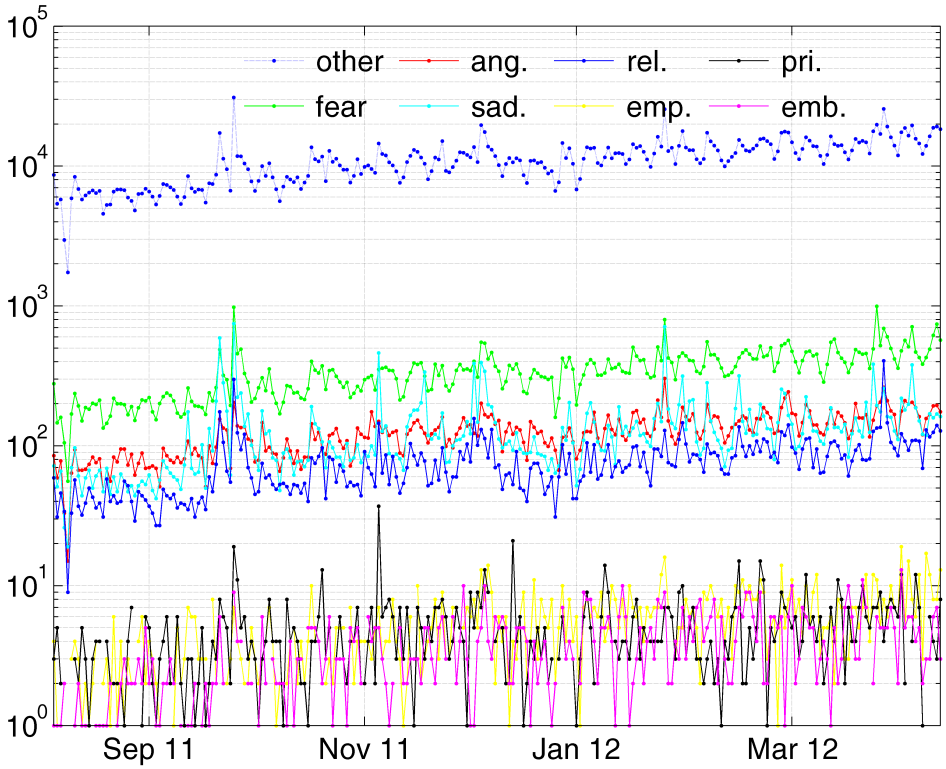
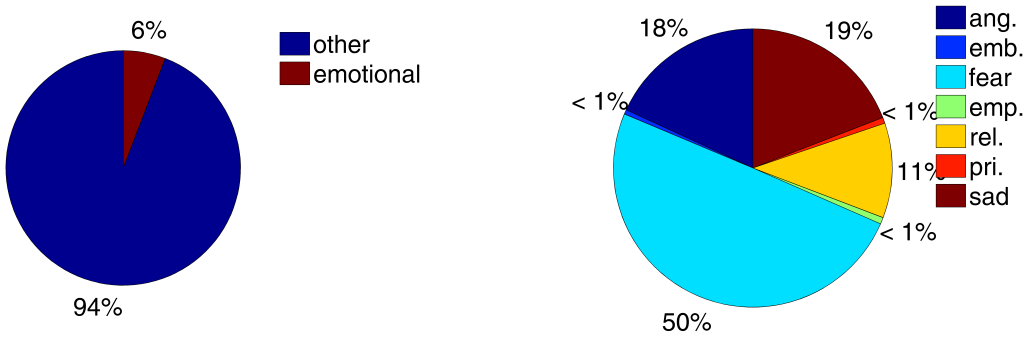


Figure 4.1: The daily counts of bullying traces in different emotion categories from August 5, 2011 to April 12, 2012.

from manual inspection. Figure 4.2(b) presents a break down of the 6% emotional bullying traces. Half of them contain *fear*, followed by *sadness*, *anger* and *relief*. *Embarrassment*, *empathy* and *pride* are virtually abscent. This also highlights the data skewness issue if the human annotators were to manually label bullying traces.

Recall that participants in a bullying episode take several well-defined roles. We hypothesize that different roles may express different emotions. We apply author-role classifier in Section 2.2 to the bullying traces, therefore, each tweet is



(a) emotional or not (b) fraction within emotional

Figure 4.2: Fraction of emotion categories

associated with an author role label by this classifier. Figure 4.3 shows the fraction of emotions for each author's role. We assume that authors of one role generate tweets in one emotion with probability p. The bars show the MLE estimations of p and the error bars indicate the Binomial 95% confidence intervals. Compared to other roles, accusers seem to express more fear but less anger. Reporter and victims seem to experience more sadness and relief than other roles. However, these observations should be taken with a grain of salt: The emotion in a bullying trace may not be the author's own feelings. It is possible that the authors sometimes discuss other participants' emotions. Our emotion classifier is not capable of distinguishing emotions of the author vs. of other people. A detailed analysis with deeper natural language processing remains future work. In addition, we have noticed that accusers often express fear jokingly (i.e., teasing; see below). For example, *"@USER lol really?! I'm so scared!! I hope I am not verbally beaten. You cyber bully ;),"* *"@USER you are such a bully!!!haha & im sooooo scared if him.lol."* This might help explain why accusers seems to have more fear.

It is interesting to see if there is any differences in terms of emotion between teasing and non-teasing bullying traces. In Figure 4.4, we observe that teasing

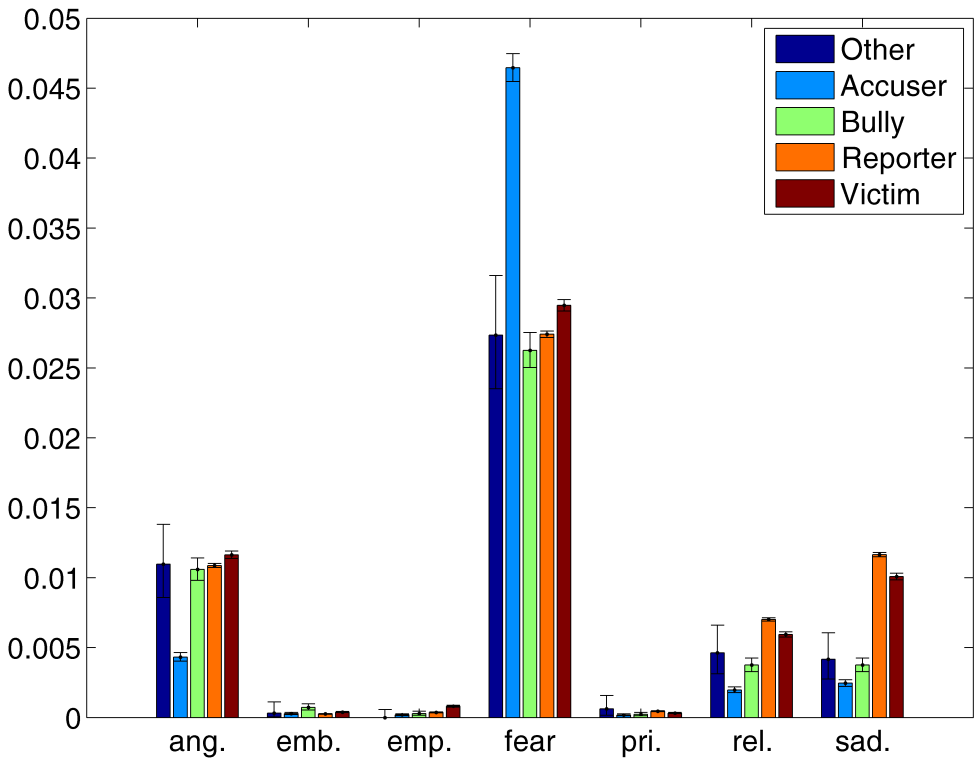


Figure 4.3: The fraction of emotions by author's role.

bullying traces contain less *sadness* and *relief*. This seems reasonable, as in general these emotions are expressed more seriously rather than jokingly. On the other hand, teasing bullying traces contain more *fear*. We speculate that people may pretend to be afraid of a bully even though in reality they are not. For example,*"@USER I'm so scared haha there's like ten girls then like 30 lads! I'm gonna get so bullied#boohoo," "@USER eh ya!!!! sometimes i very scared to approach them, like i want to bully them like that LOL HAHAHA.."*

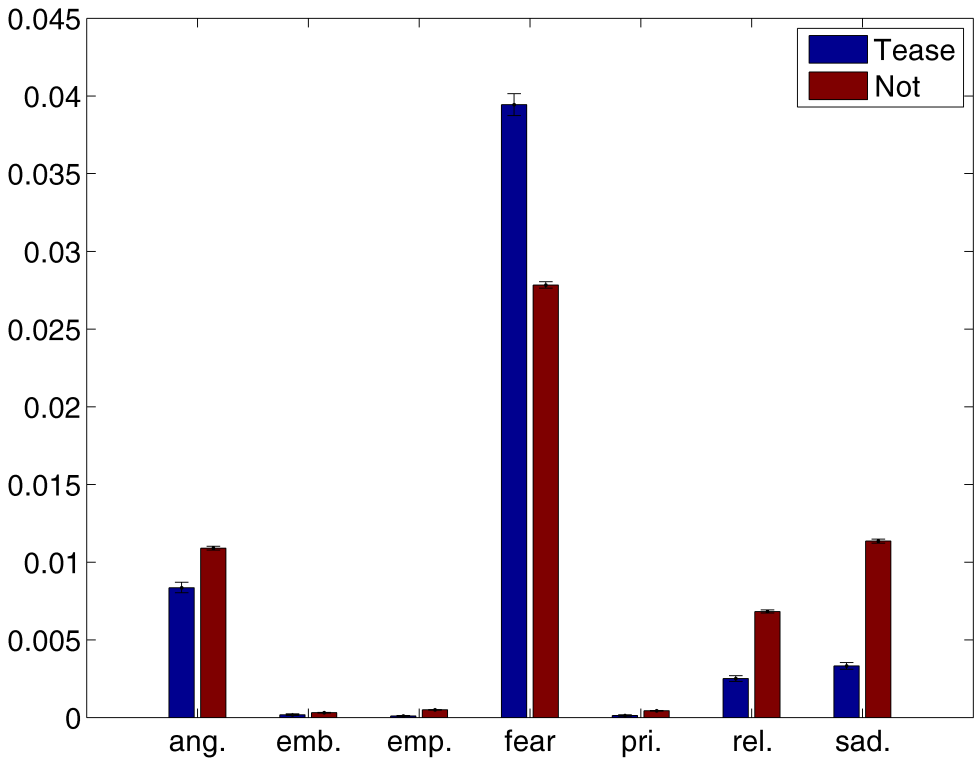


Figure 4.4: The fractions of emotions by teasing or not.

4.4 Regrets in Bullying Traces

A large body of literature suggests that participants in bullying events, including victims, bullies, and witnesses, are likely to report psychological adjustment problems Jimerson et al. (2010). One potential source of therapy for these issues can be self-disclosure of the experience to an adult or friend Mishna and Alaggia (2005); existing research suggests that victims who seek advice and help from others report less maladjustment than victims who do not Shelley and Craig (2010).

Disclosure of bullying experiences through social media may be a particularly effective mechanism for participants seeking support because social media has the

potential to reach large audiences and because participants may feel less inhibition when sharing private information in an online setting Walther (1996). Furthermore, there is evidence that online communication stimulates self-disclosure, which leads to higher quality social relationships and increased well-being Valkenburg and Peter (2009).

Online disclosure may also present risks for those involved in bullying however, such as re-victimization, embarrassment, and social ostracization. Evidence exists that some individuals may react to these risks retroactively, by deleting their social media posts (Child et al., 2011; Christofides et al., 2009). Several relevant motives have been found to be associated with deleting posted information, including conflict management, safety, fear of retribution, impression management, and emotional regulation (Child et al., 2012).

To better understand, and possibly prevent, user regret after posting bullying related tweets, we collect bullying traces as in Section 2.1 and perform regular status checks to determine if and when tweets become inaccessible. While a tweet becoming inaccessible does not guarantee it has been deleted, we attempt to leverage http response codes to rule out other common causes of inaccessibility. Speculating that regret may be a major cause of deletion, we first conduct exploratory analysis on this corpus and then report the results of an off-the-shelf regret predictor.

Data Collection

We obtain bullying traces as in Section 2.1. Each identified trace contains at least one bullying related keyword and passes a bullying-or-not text classifier.

Our data was collected in realtime using the Twitter streaming API; once a tweet is collected, we query its url (`https://twitter.com/USERID/status/TWEETID`) at regular intervals and infer its status from the resulting http response code. We interpret an HTTP 200 response as an indication a tweet still exists and an HTTP 404 response, which indicates the tweet is unavailable, as indicating deletion. A user changing their privacy settings can also result in an HTTP 403 response; we do not consider this to be a deletion. Other response codes, which appear quite rarely,

are treated as anomalies and ignored. All non HTTP 200 responses are retried twice to ensure they are not transient oddities.

To determine when a tweet is deleted, we attempted to access each tweet at time points $T_i = 5 \times 4^i$ minutes for $i = 0, 1 \ldots 7$ after the creation time. These roughly correspond to periods of 5 minutes, 20 minutes, 1.5 hours, 6 hours, 1 day, 4 days, 2 weeks, and 2 months. While we assume that user deletion is the main cause of a tweet becoming unavailable, other causes are possible such as the censorship of illegal contents by Twitter (Twitter, 2012).

Our sample data was collected from July 31 through October 31, 2012 and contains 522,984 bullying traces. Because of intermittent network and computer issues, several multiple day data gaps exist in the data. To combat this, we filter our data to include only tweets of unambiguous status. If any check within the 20480 minutes (about two weeks) interval returns an HTTP 404 code, the tweet is no longer accessible and we consider it *deleted*. If the 20480 minute or 81920 minute check returns an HTTP 200 response, that tweet is still accessible and we consider it *surviving*. The union of the surviving and deleted groups formed our cleaned dataset, containing 311,237 tweets in total.

This dataset and its documentation is archived as Bullying Trace Regret data set (see Appendix A.5).

Exploratory Data Analysis

A user's decision to delete a bullying trace may be the result of many factors which we would like to isolate and understand. In this section we will examine several such possible factors.

Word Usage

Our dataset contains 331,070 distinct words and we are interested in isolating those with a significantly higher presence among either deleted or surviving tweets. We

define the odds ratio of a word w

$$r(w) = \frac{P(w \mid \text{deleted})}{P(w \mid \text{surviving})},$$

where $P(w \mid \text{deleted})$ is the probability of word w occurring in a deleted tweet, and $P(w \mid \text{surviving})$ is the probability of w appearing in a surviving tweet. In order to ensure stability in the probability estimation, we only considered words appearing at least 50 times in either the surviving or deleted corpora.

Following (Bamman et al., 2012), we qualitatively analyzed words with extreme values of $r(w)$, and found some interesting trends. There was a significant tendency for "joking" words to occur with $r(w) < 0.5$; examples include "xd," "haha," and "hahaha." Joking words occur less frequently in deleted tweets than surviving ones. On the other end of the spectrum, there were no joking words with $r(w) > 2$. What we found instead were words such as "rip," "fat," "kill," and "suicide." While it is relatively clear that joking is less likely to occur in deleted tweets, there was less of a trend among words appearing more frequently in deleted tweets.

Surviving Time

Let N be the total number of tweets in our corpus, and $D(T_i)$ be the number of tweets that were first detected as deleted at minute T_i after creation. Note that $D(T_i)$ is not cumulative over time: it includes only deletions that occurred in the time interval $(T_{i-1}, T_i]$. Then we may define the deletion rate at time T_i as

$$R_T(T_i) = \frac{D(T_i)}{N(T_i - T_{i-1})}.$$

In other words, $R_T(t)$ is the fraction of tweets that are deleted during the one minute period $(t, t+1)$.

We plot R_T vs. t using logarithmic scales on both axes in Figure 4.5 and the result is a quite strong linear trend. Fitting the plot with a linear regression, we

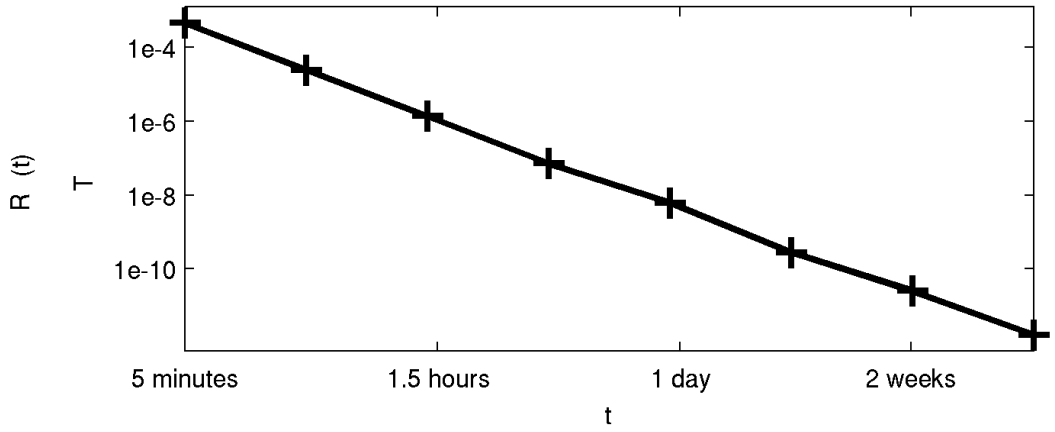


Figure 4.5: Deletion rate decays over time.

derive an inverse relationship between R_T and t of the form

$$R_T(t) \propto 1/t.$$

This result makes sense; the social effects of a particular bullying tweet may decay over time, making regret less of a factor. Furthermore, the author may assume an older tweet has already been seen, rendering deletion ineffective. Additionally, because the drop off in deletion rate is so extreme, we are able to safely exclude deletions occurring after two weeks from our filtered dataset without introducing a significant amount of noise. Finally, $\sum_{t=0}^{\infty} R_T(t)$ gives the overall fraction of deletion, which in our case is around 4%.

Location and Hour of Creations

Some bullying traces contain location meta-data in the form of GPS coordinates or a user-created profile string. We employed a reverse geocoding database (`http://www.datasciencetoolkit.org`) and a rule-based string matching method to map these tweets to their origins (at the state level; only for tweets within the USA). This also allowed us to convert creation timestamps from UTC to local time by mapping

user location to timezone. Because many users don't share their location, we were only able to successfully map 85,465 bullying traces to a US state s, and local hour of day h. Among these traces, 3,484 were deleted which translates to an overall deletion rate of about 4%.

Let $N(s,h)$ be the count of bullying traces created in state s and hour h. Aggregating these counts temporally yields $N_S(s) = \sum_h N(s,h)$, while aggregating spatially produces $N_H(h) = \sum_s N(s,h)$. Similarly, we can define $D(s,h)$, $D_S(s)$ and $D_H(h)$ as the corresponding counts of deleted traces. We can now compute the deletion rate

$$R_H(h) = \frac{D_H(h)}{N_H(h)}, \text{ and } R_S(s) = \frac{D_S(s)}{N_S(s)}.$$

The top row of Figure 4.6 shows $N_H(h)$, $D_H(h)$, and $R_H(h)$. We find that $N_H(h)$ and $D_H(h)$ peak in the evening, indicating social media users are generally more active at that time. The peak of $R_H(h)$ appears at late night and, while there are multiple potential causes for this, we hypothesize that users may fail to fully evaluate the consequences of their posts when tired. The bottom row of Figure 4.6 shows $N_S(s)$, $D_S(S)$, and $R_S(s)$. The plot of $N_S(s)$ shows that bullying traces are more likely to originate in California, Texas or New York which is the result of a population effect. Importantly however, the deletion rate $R_S(s)$ is not affected by population bias and we see, as expected, that spatial differences in $R_S(s)$ are small. We performed χ^2-test to see if a state's deletion rate is significantly different from the national average. We chose the significance level at 0.05 and used Bonferroni correction for multiple testing. Only four states have significantly different deletion rates from the average: Arizona (6.3%, $p = 5.9 \times 10^{-5}$), California (5.2%, $p = 2.7 \times 10^{-7}$), Maryland (1.9%, $p = 2.3 \times 10^{-5}$), and Oklahoma (7.1%, $p = 3.5 \times 10^{-5}$).

Author's Role

Participants in a bullying episode assume well-defined roles which dramatically affect the viewpoint of the author describing the event. We used our Author's Role Classifier (Version 1, see Appendix B.2), to label each bullying trace in the cleaned corpus by author role: *Accuser*, *Bully*, *Reporter*, *Victim* or *Other*.

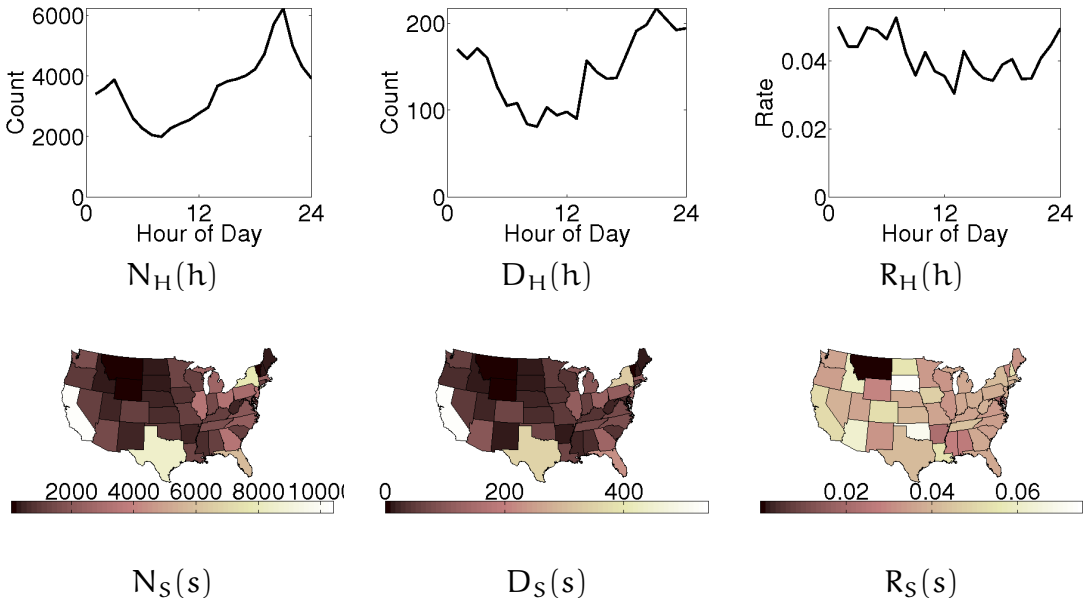


Figure 4.6: Counts and deletion rates of geo-tagged bullying traces.

	Deleted	Total	P(deleted | Role)
Accuser	2,541	50,088	5.1%
Bully	1,792	30,123	6.0%
Reporter	11,370	147,164	7.7%
Victim	6,497	83,412	7.8%
Other	41	450	9.1%

Table 4.4: Counts and deletion rate for different roles.

Table 4.4 shows that compared to tweets produced by bullies, victims create more bullying traces, possibly due to an increased need for social support on the part of the victim. More importantly, P(deleted | victim) is higher than P(deleted | bully), a statistically significant difference in a two-proportion z-test. Possibly, victims are more sensitive to their audience's reaction than bullies.

	Deleted	Total	P(deleted | Teasing?)
Yes	858	22,876	3.8%
Not	21,383	288,361	7.4%

Table 4.5: Counts and deletion rate for teasing or not.

Teasing

Many bullying traces are written jokingly. We applied our Teasing Bullying Trace Classifier (see Appendix B.6) to the cleaned corpus to identify teasing posts.

Table 4.5 shows that P(deletion | Teasing) is much lower than P(deletion | Not Teasing) and the difference is statistically significant in a two-proportion z-test. It seems plausible that authors are less likely to regret teasing posts because they are less controversial and have less potential to generate negative audience reactions. This also corroborates our findings in word usage that joking words are less frequent in deleted tweets.

Predicting Regrettable Tweets

Once a bullying tweet is published and seen by others, the ensuing effects are often impossible to undo. Since ill-thought-out posts may cause unexpectedly negative consequences to an author's reputation, relationship, and career (Wang et al., 2011), it would be helpful if a system could warn users before a potentially regrettable tweet is posted. One straightforward approach is to formulate the task as a binary text categorization problem, and build a Bullying Trace Regret Classifier (see Appendix B.8).

We use the cleaned dataset, in which each tweet is known to be surviving or deleted after 20,480 minutes (about two weeks). Since this dataset contains 22,241 deleted tweets, we randomly sub-sampled the surviving tweets down to 22,241 to force our deleted and surviving datasets to be of equal size. Consequentially, the baseline accuracy of the classifier is 0.5. While this does make the problem artificially easier, our initial goal was to test for the presence of a signal in the data.

We then followed the preprocessing procedure in Section 2.1, performing case-folding, anonymization, and tokenization, treating URLs, emoticons and hashtags specially. We also chose the unigrams+bigrams feature representation, only keeping tokens appearing at least 15 times in the corpus.

We chose to employ a linear SVM implemented in LIBLINEAR (Fan et al., 2008) due to its efficiency on this large sparse text categorization task and a 10-fold cross validation was conducted to evaluate its performance. Within the first fold, we use an inner 5-fold cross validation on the training portion to tune the regularization parameter on the grid $\{2^{-10}, 2^{-9}, \ldots, 1\}$; the selected parameter is then fixed for all the remaining folds.

The resulting cross validation accuracy was 0.607 with a standard deviation of 0.012. While it is statistically significantly better than the random-guessing baseline accuracy of 0.5 with a p-value of 5.15×10^{-10}, this accuracy is nevertheless too low to be useful in a practical system. One possibility is that the tweet text contains very limited information for predicting inaccessibility; a user's decision to delete a tweet potentially depends on many other factors, such as the conversation context and the characteristics of the author and audience.

In the spirit of exploring additional informative features for deletion prediction, we also used Teasing Bullying Trace Classifier and Author's Role Classifier, and appended the predicted teasing, and author role labels to our feature vector. This augmented feature representation achieved a cross validation accuracy of 0.606, with standard deviation 0.007; not statistically significantly different from the text-only feature representation. While it seems that a signal does exist, leveraging it usefully in real world scenarios may prove challenging due to the highly-skewed nature of the data.

Discussion

There have been several recent works examining causes of deletion in social media. Wang *et al.* (2011) qualitatively investigated regret associated with users' posts on social networking sites and identified several possible causes of regret. Bamman

et al. (2012) focused on censorship-related deletion of social media posts, identifying a set of sensitive terms related to message deletion through a statistical analysis and spatial variation of deletion rate.

Assuming that deletion in social media is indicative of regret, we studied regret in a bullying context by analyzing deletion trends in bullying related tweets. Through our analysis, we were able to isolate several factors related to deletion, including word usage, surviving time, and author role. We used these factors to build a regret predictor which achieved statistically significant results on this very noisy data. In the future, we plan to explore more factors to better understand deletion behavior and regret, including users' recent posts, historical behavior, and other statistics related to their specific social network.

5 HASHTAGS USAGE IN BULLYING TRACES

Hashtags are keywords or acronyms that are prefixed with a # symbol that are annotated within tweets to indicated markers of topic. Hashtagging was introduced around February 2008 for users to tag their content for retrieval (Huang et al., 2010). Tagging had already been used on Del.icio.us (online web bookmarking site) and blogging platforms for bloggers to retrieve post topics for themselves and for their readers. However, now hashtagging is used more as a search term to filter out certain tweets and to elevate certain topics on Twitter. When hashtags were first used back in 2008, it was uncommon for tweets to have more than one hashtag. Now hashtags are used within the phrase or sentence of a tweet and provide a user with more chances for recognition the more hashtags they use (Huang et al., 2010).

Twitter users can use a keyword to gather instant updates of an event or a conversation, such as the protests in Egypt in 2011 (Papacharissi and de Fatima Oliveira, 2012) or The Wall Street occupy movement (Gleason, 2013), sometimes faster than the news media can (see (Mitchell and Guskin, 2013) for more Twitter event reports). Twitter users cannot only obtain breaking news from elite news affiliates, but also from individuals who are witnessing the event in real-time. Twitter users contribute to the conversation of an event by including an event's hashtag with their tweet. The more retweets and mentions a Twitter user receives, the more that user and her or his tweet is recognized. Hashtags can also bring people together and establish a sense of solidarity in times of crisis (Papacharissi and de Fatima Oliveira, 2012). Hashtags not only disseminate information to interested Twitter users, but the users of hashtags may also have an impact on how events unfold (Kirkland, 2014).

Bullying represents a phenomenon that is discussed in relation to tragic, high profile events. It also may be discussed on a more day-to-day basis as a function of individuals' personal experiences or via organizations that are working to lessen the negative impact of bullying. As a result, bullying is expected to have relevance for a broad range of hashtag uses. In this chapter, we seek to understand the bullying topics that Twitter users posted about across 2012 by studying which hashtags were

employed and how they were utilized (Calvin et al., 2015). The identified hashtags, annotations and documentations are archived as Hashtags in Bullying Traces Data Set (see Appendix A.6).

5.1 Identification of the Hashtags

Our first goal was to identify the hashtags most frequently associated with Twitter posts that use bullying keywords in 2012. We expected that this would yield information about important events related to bullying and also general public opinions about bullying.

We collect data from the public Twitter streaming API between the period of January 1, 2012 and December 31, 2012. We captured tweets in 2012 that contained at least one of the following keywords, "bully," "bullied," and "bullying" through the Twitter streaming API. Unless we hit the maximum allowable tweets in a given day, we received all posts that satisfied this condition for 2012. In total, we collected 25,370,824 tweets. From this collection of bullying tweets, we extracted tweets that contained hashtags by searching for any strings which started with "#" and consisted of only alphanumeric characters (e.g., "#bully"). We case-folded all hashtags (i.e., we replaced upper case letters with lower case ones) to merge different variations of the same hashtag into a single hashtag. For example, "#StopBullying", "#stopBullying", "#STOPBullying" were all transformed to "#stopbullying." After case-folding, we ended up with 552,831 distinct hashtags in total and 9,815,715 out of the 25,370,824 tweets contained at least one hashtag. We counted the number of tweets in which each distinct hashtag appeared. Then we sorted all of the hashtags by the number of tweets in which they appeared to identify the top 500 most popular hashtags that appear in Twitter posts with the keywords "bully," "bullied," and "bullying."

Table 5.1: Top 500 hashtags used in tweets that contained bullying keywords collected between January 1, 2012 and December 31, 2012

#bullying	#share	#obese
#bully	#bullies	#istopbullying
#stopbullying	#bullyingé	#impactlive
#bullymovie	#education	#moms
#spiritday	#uk	#edchat
#ripamandatodd	#h	#39
#lgbt	#cyberbullyingiswrong	#cdnpoli
#bullied	#ripdemizadirectioner	#wheniwaslittle
#teamfollowback	#1dfact	#gop
#ripboybeliebermartin	#capricorn	#suicide
#rt	#ifihadthepower	#truth
#libra	#kids	#twitition
#xfactor	#obesity	#ripamanda
#gangupforgood	#health	#sad
#ff	#amandatodd	#school
#antibullying	#growup	#love
#tcot	#mentionke	#banbeliebers
#staystrong	#stopthebullying	#peopleshouldstop
#stop	#1dfacts	#autism
#cyberbullying	#china	#nowplaying
#inmiddleschool	#retweet	#japan
#oomf	#schoolmemories	#justsaying
#noh8	#basketballwives	#debate
#news	#itwasnevercool	#badluckbrian
#p2	#parents	#twitter
#in2013nomore	#fuerapameladiaz	#stopcyberbullies
#mentionto	#india	#confessionnight
#aries	#stopcyberbullying	#stompoutbullying
#1	#turtle	#thingsthatwegottastop
#usa	#10basicfactsaboutme	#notcool
#ripcharlotte	#gay	#bulldog
#whydopeoplethinkitsokayto	#weight	#antibullyingweek
#fb	#dosomething	#weasagenerationneedtoletgoof
#jj	#mexico	#voteobama
#cnn	#taurus	#bbw
#bbuk	#lol	#cyberbully
#riseabovehate	#eh	#parejaperfecta
#in6thgrade	#ihatepeoplewho	#speakup
#parenting	#thingspeoplehavetostopdoing	#imsickof
#canada	#cosasquenotienenqueexistir	#cbc
#bulling	#obama	#stophatestartlove
#rip	#smh	#2
#endbullying	#ripkevin	#kindle
#romney	#teachers	#realtalk
#np	#fact	#10thingsaboutmyself
#thatdoesnotmakeyoucool	#fat	#jessierocks

Continued on next page

#standup
#brave
#auspol
#eating
#r
#itgetsbetter
#bornbravebus
#childabuse
#sorrynotsorry
#itsnotokayto
#nw
#dads
#no
#wwe
#bullyingisnothot
#20thingsidontlike
#respect
#students
#tna
#fitness
#youth
#dog
#children
#nobullying
#anti-bullying
#fcsqueamo
#imagine
#ripblake
#beatbullying
#facts
#st
#signon
#sto
#primaryschoolmemories
#crying
#believe
#ctv
#racism
#3
#teaparty
#raw
#prevention
#mtv
#mannnmovement
#loveislouder
#rhonj

#debates
#in8thgrade
#saynotobullying
#hypocrite
#nobodylikes
#facebook
#30thingsaboutme
#s
#ripmegan
#schools
#porn
#wtf
#pitbull
#fridayfightagainst
#bgc9
#glee
#5
#citepessoasbonitasdotwitter
#stopthehate
#ripsparksforselena
#teens
#whydopeoplethink
#aquarius
#help
#bull
#4
#beastar
#tweet4taiji
#amand
#demifacts
#osmelhorestwittersde2012
#t
#cbb
#kingshit
#whydopeople
#7
#bitch
#citeosfcsmaistopsdotwitter
#peopleneedtostop
#stoprush
#nhl
#eventful
#standuptobullying
#mannntctw
#100thingsaboutme
#thingsicantstand

#savealife
#bullyingisforlosers
#dt
#pathetic
#neverwouldiever
#greysonfact
#karma
#whydoyouthinkitsokayto
#stopbullyingnow
#100thingsihate
#englishbulldog
#típicodegordos
#pinkshirtday
#mean
#iwanttopunchpeoplewho
#anonymous
#nfl
#respectbeliebers
#wheniwasakid
#10
#wshh
#please
#8
#follow
#tomyfuturekids
#dancemoms
#abuse
#wheniwasalittlekid
#inhighschool
#stopbulling
#7pessoasqueeupegaria
#mikebully
#newbedon
#suicideawareness
#family
#12
#thingsthatbotherme
#sosad
#vpdebate
#music
#fail
#ows
#6
#leaveitin2012
#bbc
#bb14

Continued on next page

#bieberfact
#tlot
#teamup
#edtech
#wlcauthor
#france
#howimetmybestfriend
#ifitwasuptome
#stupidthingspeopledo
#27
#video
#inelementaryschool
#honestyhour
#suicideawarenessday
#b
#onpoli
#13
#overth
#koreanindo
#elementaryschoolmemories
#ontario
#mentalhealth
#yolo
#1dnews
#ripvanessa
#obama2012
#ple
#backinelementaryschool
#fcqueeusemprequisseramiga
#in7thgrade
#ripastronautchelsea
#depression
#bbau
#btgstrongkids
#dumbfanficmoments
#amalayer
#rude
#debate2012
#middleschoolmemories
#wheniwasyounger
#notfair
#iwillneverunderstand
#harper
#ripjesuseduardobelieber
#justsayin
#ripericmonster
#peoplewhowerebulliedbutnowsuccessful
#socialmedia
#nomorebullies
#hypocrites
#15
#safety
#change
#toronto
#7pessoaslindas
#biggest
#shinedown
#openfollow
#cyber
#equality
#citepessoaslindas
#teen
#9
#hr
#againstbullying
#20factsaboutme
#amanda
#bgc
#15factsaboutme
#southpark
#11
#qanda
#nba
#1lawan1
#itdoesnotmakeyoucool
#sex
#jodieagainstbullying
#bullyinghurts
#asdtanya
#not
#damai
#weliveinageneration where
#operationpurplesky
#takeastand
#asshole
#14
#hope
#ripfatma
#antibully
#itshouldntbethatway
#wstopbullying
#glbt
#bullyingneedstostop
#beliebers
#media
#rememberthatkidinschool
#xatiada
#germany
#17
#purple
#swiftfact
#truestory
#igotnorespectforyou
#mittromney
#sadtweet
#harassment
#hate
#wow
#childtrafficking
#nomasbullying
#itkillsmewhenpeople
#ripcourtney
#tomyfutureson
#pets

Continued on next page

#highschoolmemories
#youtube
#ufg2012
#us
#pickone
#16
#soda
#bullterrier
#bbcqt
#soundcloud
#stopbu
#loveisallweneed
#liamfact
#middlefingerup
#confissoesdamadrugada
#abc
#ihopesomeday
#teamcashl3wis
#fuckyou
#subtweet
#friend
#29
#ifitwereuptome
#rn20
#shinedown2012
#mlb
#politics
#victim
#cbs
#cite10avatareslindos
#ripswiftystephanie
#mental
#p21
#seriously
#thingsthatgetmeupset
#dogs
#tuite10pessoasbonitas
#p2b
#5thingsicantstand
#24
#30secondmom
#israel
#1dfamily
#teammindless
#oprah
#stopbullyin
#life
#bcpoli
#xxx
#stophate
#liamfacts
#ladygaga
#avenidabrasil
#i
#teach
#karenklein
#workplace
#bullyinges
#21
#whatwereyouthinking
#whatif
#demifact
#dnc2012
#backwheniwasakid
#meanie
#puppy
#23
#panorama
#thegroovyproject
#tfb
#pls
#fml
#twograves
#makeachange
#tvo
#hot
#getalife
#stfu
#gaza
#20
#angka8
#peace
#bul
#bigmarch2012
#22
#w
#pussy
#admita
#sorry
#topsecretreidfact
#notalone
#xfactorusa
#10thingsthatgetsonmynerves
#scorpio
#cite15pessoasbonitas
#iran
#thataintcool
#you
#followme
#25
#tbay
#mhsm
#support
#icantstandpeople
#ronpaul
#serialkiller
#banter
#secondaryschoolconfessions
#18

We limited our study to top 500 most popular hashtags because these are most likely to have been used by many users and may therefore be a better representation of hashtag use associated with bullying. Five hundred is only about 0.1% of all unique hashtags (552,831 in total), but they cover 50% of tweets about bullying with hashtags. The number of tweets with one of the top 500 hashtags is 4,931,270. We have 9,815,715 tweets with any hashtags. So $4{,}931{,}270/9{,}815{,}715 \simeq 0.50$.

The 500 hashtags that appeared most frequently are reported in Table 5.1 Among the top 500, the most common hashtag was #bullying, which appeared 354,128 times and the least common was #18, which appeared 1,491 times. The identified hashtags, together with their features and annotations below are archived as Hashtags in Bullying Traces Data Set (see Appendix A.6).

5.2 Categories of Hashtags

Our second goal was to identify the different types of bullying hashtags that are used by evaluating their different intents. In doing so, we learn whether hashtags are used differently despite focusing on the shared topic of bullying. Some users may seek to raise awareness about specific bullying episodes. For example, in response to Jamey Rodemeyer's death, Lady Gaga asked her Twitter followers to trend "#MakeaLawforJamey." Other users may be posting general messages about bullying ("*#bullying is wrong*") or supporting specific causes such as ("*Stand against bullying! Wear purple and make your profile pic purple for #SpiritDay*").

We annotated all of the top 500 hashtags to discern whether they could be categorized. We followed an inductive approach to generate the categories (Miles and Huberman, 1994). A team of four bullying scholars independently read through all of the hashtags and noted their first impressions of what they thought the hashtag referred to. Next, they evaluated the context in which each hashtag was used through a random sample of 500 tweets that contained the hashtag. For example, whether the hashtag was used as a word within a sentence such as "*#InMiddleSchool I was bullied a lot because I was different*" or whether the hashtag was used separately such as "*I'm taking a stand. Tired of me along with many other people getting bullied.*

#antibullying" was evaluated. At this stage, consistencies in how the hashtag was used across tweets was also considered (e.g., The hashtag was always used to promote a cause). Finally, to further identity the category of the hashtag, the coders searched for the hashtag among all current Twitter posts to see whether it was still being used as of December 2013, used http://www.tagdef.com to identify the meaning of the hashtag to discern acronyms and slang terms, and searched for the hashtag using Google. After independently generating and assigning a category, the scholars met as a group to review the entire list of 500 hashtags to reach agreement on their category assignment.

Each of the top 500 hashtags were evaluated and assigned to a category. Eight categories were identified.

- The General Bullying category ($n = 44$) included hashtags that contained terms that included the word bullying in some way (e.g., #bully, #stopbullying).
- The Campaign category ($n = 32$) included hashtags that were associated with campaigns such as #spiritday, a day that occurred on October 19, 2012 and refers to GLAAD's (Gay and Lesbian Alliance Against Defamation) antibullying campaign, in which supporters for LGBT who are bullied wear purple or tint their profile pictures or logos in purple to spread awareness.
- The Suicide/Death category ($n = 19$) contained hashtags with death-related terms (e.g., #suicide) or references to deaths or suicides that resulted from bullying (e.g., #ripamandatodd). Amanda Todd, a 15 year-old from British Columbia, posted a YouTube video involving handwritten flashcards that described how she was blackmailed into exposing herself to an unknown individual via webcam, and the photos later went viral to Facebook. About five weeks after the video was posted, Amanda committed suicide, and her story was used as a demonstration of the seriousness consequences of cyberbullying.
- The Bullying Terms category ($n = 45$) included hashtags that referred to variables or factors that have been studied within psychology in association with bullying (e.g., #abuse, #glbt).

- The Media category (n = 33) included hashtags that referred to TV shows (#xfactorusa), movies (#bullymovie), or content providers (#bbc).
- The Everyday Twitter Trend category (n = 63) contained terms that are used in everyday online exchanges that are not connected to bullying (e.g., #lol, #smh).
- The Fill-in-the-Blank/Game category (n = 74) contained hashtags that are designed as templates for users to copy and use while adding on a response of their own. These usually take the form of phrases such as #backinelementaryschool and #whydopeoplethinkitsokay.
- The remaining hashtags (n = 190) were assigned to the Other category. This included a wide variety of terms such as #libra and #usa.

5.3 Characteristics of Hashtags and Hashtag Categories

Our third goal was to describe the features of tweets associated with the hashtag categories. The eight features that were analyzed for each tweet were derived directly from data present within each post that Twitter makes available through its streaming Application Programming Interface (API). We chose eight features that we believed would demonstrate differences in various aspects of the level of recognition of each individual hashtag. We conducted a one-way ANOVA comparing hashtag categories for features one through six. We report the hashtag category mean values and standard deviations along with indicators for where significant group differences exists in Table 5.2.

The Number of Tweets

First, we counted the number of tweets in which each hashtag appeared as this is an indicator of its overall popularity.

Categories	No. of tweets		% of retweets		No. of URLS		No. of hashtags		No. of authors		No. of mentions		% with negative sentiment		% with neutral sentiment		% with positive sentiment	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Campaign ($n = 32$)	6,567.41 $_{b}$	12,603.05	.54 $_{a}$	.24	.45 $_{b}$	.34	1.83 $_{bc}$	1.16	4,651.87 $_{b}$	9,273.23	1.06 $_{a}$	.30	.14 $_{a}$	.19	.78 $_{a}$	.23	.08 $_{a}$	.17
Everyday Twitter trend ($n = 63$)	4,820.60 $_{b}$	6,024.93	.31 $_{bc}$	.20	.27 $_{bc}$	.33	2.30 $_{bc}$	1.85	3,502.95 $_{b}$	3,635.58	.76 $_{bc}$	.43	.18 $_{a}$	.13	.78 $_{a}$	.13	.05 $_{a}$	.03
Fill-in-the-blank/Game ($n = 74$)	4,009.22 $_{b}$	2,885.91	.50 $_{ab}$	.21	.09 $_{c}$	.21	1.15 $_{c}$	.17	3,785.38 $_{b}$	2,741.18	.65 $_{c}$	.17	.17 $_{a}$	.15	.79 $_{a}$	.15	.04 $_{a}$	.05
General bullying ($n = 44$)	24,671.93 $_{a}$	73,535.22	.55 $_{a}$	.29	.33 $_{b}$	.26	1.85 $_{bc}$	.69	16,080.07 $_{a}$	45,842.33	.95 $_{ab}$	.32	.20 $_{a}$	.19	.75 $_{a}$	.19	.04 $_{a}$	.04
Media ($n = 33$)	7,666.73 $_{b}$	14,362.50	.36 $_{bc}$	.27	.30 $_{bc}$	.38	3.49 $_{b}$	4.47	686.24 $_{b}$	11,904.52	.56 $_{c}$	.36	.23 $_{a}$	.20	.73 $_{a}$	.20	.05 $_{a}$	.05
Other ($n = 190$)	3,924.61 $_{b}$	4,124.72	.38 $_{abc}$	.30	.50 $_{b}$	.42	3.48 $_{b}$	4.02	2,400.60 $_{b}$	3,072.62	.83 $_{ab}$	.56	.20 $_{a}$	.22	.76 $_{a}$	.23	.04 $_{a}$	.10
Suicide/Death ($n = 19$)	6,870.05 $_{b}$	9,989.29	.58 $_{a}$	.22	.29 $_{bc}$	.31	2.33 $_{bc}$	2.11	5,537.47 $_{b}$	8,754.76	.76 $_{bc}$	.18	.28 $_{a}$	.20	.65 $_{a}$	.21	.06 $_{a}$	.08
Bullying term ($n = 45$)	5,194.80 $_{b}$	6,684.75	.24 $_{bc}$	.16	.74 $_{a}$	.23	5.62 $_{a}$	3.68	1,993.07 $_{b}$	4,116.40	.54 $_{c}$	.36	.20 $_{a}$	.18	.76 $_{a}$	.17	.04 $_{a}$	.03

Note. RT = retweet (forwarding of another user's tweet); URLs = link to a web address; user mentions = tweets that contain @username. Within column, means with different subscripts are significantly different at $p < .05$ using the Tukey test.

Table 5.2: Means, standard deviations, and mean differences between hashtag categories on tweet features

The average number of tweets containing each hashtag differed between categories, $F(7, 492) = 4.58$, $p < .01$, partial eta squared = .06, such that General Bullying Hashtags appeared in a significantly larger number of tweets than hashtags in any other category.

Fraction of Retweets

Second, we focused on what percentage of the overall number of tweets in which each hashtag appeared were retweets. A retweet may represent a qualitatively different type of post than a user-generated post because the user may not be adding any of her or his own unique content to such posts.

The fraction of retweets to all tweets that contained each hashtag also differed between categories, $F(7, 492) = 10.25$, $p < .01$, partial eta squared = .13. The hashtag categories that contained the highest percentage of retweets were the Suicide, General Bullying, Campaign, Fill-in-the-Blank/Game and Other categories. Only the Suicide, Campaign, and General Bullying hashtags contained a significantly higher percentage of retweets than the Media, Everyday Twitter Trend, and Bullying Term categories.

Number of URLs

Third, we evaluated the number of URLs (i.e., web addresses) included in each tweet with a given hashtag. Because tweets are limited to 140 characters, individuals may use that space to direct users to a web address containing additional information.

The number of URLs associated with each hashtag also differed by hashtag category, $F(7, 492) = 19.37$, $p < .01$, partial eta squared = .22. The Bullying Term category had the highest number of URLs associated with its hashtags and differed from all other categories. Other, Campaign, and General Bullying hashtags had the next highest number of URLS, which was significantly higher than the number used in the Fill-in-the-Blank/Game category. Media, Suicide, and Everyday Twitter Trend hashtags fell in between these groups and did not differ from either.

Number of Hashtags

Fourth, we investigated the number of hashtags included in each tweet with a given hashtag. Hashtags that appear with more other hashtags would be expected to get more attention because users searching for other hashtags may also come across the tweet.

The number of hashtags included in each tweet also differed by hashtag category, $F(7, 492) = 11.27$, $p < .01$, partial eta squared = .14. Bullying Terms had the highest number of hashtags associated with the use of each hashtag within their category and differed from all other categories. Other and Media had the next highest numbers; their number of hashtags was significantly higher than those used in the Fill-in-the-Blank/Game category. Suicide, Everyday Twitter Trend, Campaign, and General Bullying categories did not differ from one another nor from the Other, Media, and Fill-in-the-Blank categories.

Number of Distinct Authors

Fifth, we determined the number of distinct authors who used each hashtag. This count was expected to reveal how widely disseminated the hashtag usage was among the population.

The number of distinct authors who used each hashtag differed across categories, $F(7, 492) = 4.91$, $p < .01$, partial eta squared = .07, such that General Bullying hashtags had a larger number of authors than all other categories of hashtags.

Number of User Mentions

Sixth, we calculated the number of users mentioned in each tweet with a given hashtag. High user mentions was expected to function similarly to high hashtag usage in that it should get an individual tweet or hashtag more attention among other users through endogenous sharing within Twitter (Lehmann et al., 2012).

The number of users mentioned in the same tweet with a hashtag differed across categories, $F(7, 492) = 7.36$, $p < .01$, partial eta squared = .10. Campaigns, General

Bullying, and Other hashtags were posted along with the largest number of user mentions. General Bullying and Other hashtags did not differ from Suicide/Death and Everyday Twitter Trend hashtags, which contained the next largest number of user mentions. Bullying Terms, Media, and Fill-in-the-Blank/Game hashtags were posted along with the smallest number of user mentions; Everyday Twitter trend, Suicide/Death and Other did not differ in their frequency from these terms.

Negative, Neutral, and Positive Sentiments

Seventh, we evaluated the strength of the sentiment associated with each hashtag. In Chapter 4, we study the emotions present within bullying posts on Twitter. Fear, sadness, anger, and relief were found to be the most common emotions present within bullying posts. We included a broader range of posts that contained bullying keywords in this chapter; we expected to capture a wider range of emotions.

To address whether different emotions characterized the tweets, sentiment analysis was computed. Sentiment analysis is widely studied in natural language processing and has been applied in many business and social domains (Liu and Zhang, 2012). Many sentiment analysis algorithms first learn a sentiment lexicon with weights, which are words and phrases commonly used to express positive or negative sentiments, from a coded list and/or existing corpus. The sentiment of a new document is then detected by aggregating the weights of the sentiment lexicon that appears in the document with a set of rules designed from the earlier coded list. Recent algorithms also take informal spelling and emoticons into account to improve the performance on social media posts. SentiStrength (Thelwall et al., 2010) was used as it has been shown to have human-level accuracy for short social media posts in English. This algorithm allowed us to produce the fraction of negative ("*@USER I'm in so much painnnnn. #Bully*"), neutral ("*my lil sister a #BULLY*"), and positive tweets (e.g., "*The movie #bully is extremely inspirational. Strongly suggest.*") associated with each hashtag by assigning a sentiment to each tweet associated with each hashtag.

Once the fraction of negative, positive, and neutral tweets was determined for

each hashtag we conducted a Repeated Measures ANOVA comparing the hashtag categories on these variables. This analysis revealed that the fraction differed between sentiments, $F(2, 491) = 2247.20$, $p < .01$, partial eta squared = .90. Posthoc pairwise comparisons conducted with a Bonferroni correction revealed that neutral tweets ($M = .75$, $SE = .01$) represented a significantly larger fraction of tweets than negative tweets ($M = .20$, $SE = .01$). Positive tweets represented a significantly smaller fraction of tweets than both other sentiments ($M = .05$, $SE = .00$). Neither the between-category main effect nor the category by sentiment interaction revealed significant differences. See Table 5.2 for the mean fraction of tweets for each sentiment that was found for each hashtag category.

Temporal Patterns of Hashtags

Eighth, we tallied the daily number of tweets containing each hashtag across the year. Doing so allowed us to differentiate between hashtags that showed a high percentage of usage for a short period of time, called micro-memes or "bursty" hashtags (Lehmann et al., 2012), and those that have more spread over time and no extreme increases or decreases of use (Huang et al., 2010).

To understand whether temporal patterns differ among the top 500 individual hashtags, we computed the daily number of tweets containing each hashtag for one year, and then ran a Principal Component Analysis (PCA). For each hashtag h, we had a 366-dimension vector $\mathbf{x}_h$, which consists of the daily number of tweets using hashtag h. We plotted them as temporal curves (see Figure 5.2) to study how the numbers vary over the year. The scale of $\mathbf{x}_h$ reflects the popularity of the hashtag h, and we have studied it in the first feature (the number of tweets containing each hashtag) above. Now, we focus on the temporal variations, how the curves relatively change over the year, ignoring the actual scales. The cosine distance between two vectors $\mathbf{x}_{h_1}$ and $\mathbf{x}_{h_2}$ is scale invariance, i.e., the distance does not change with the scale of $\mathbf{x}_{h_1}$ or $\mathbf{x}_{h_2}$. In addition, many hashtags have similar trends, e.g., single peak, but happened at different time points. This indicates they are very similar to each other and so we wanted to further remove this variance in the distance measure.

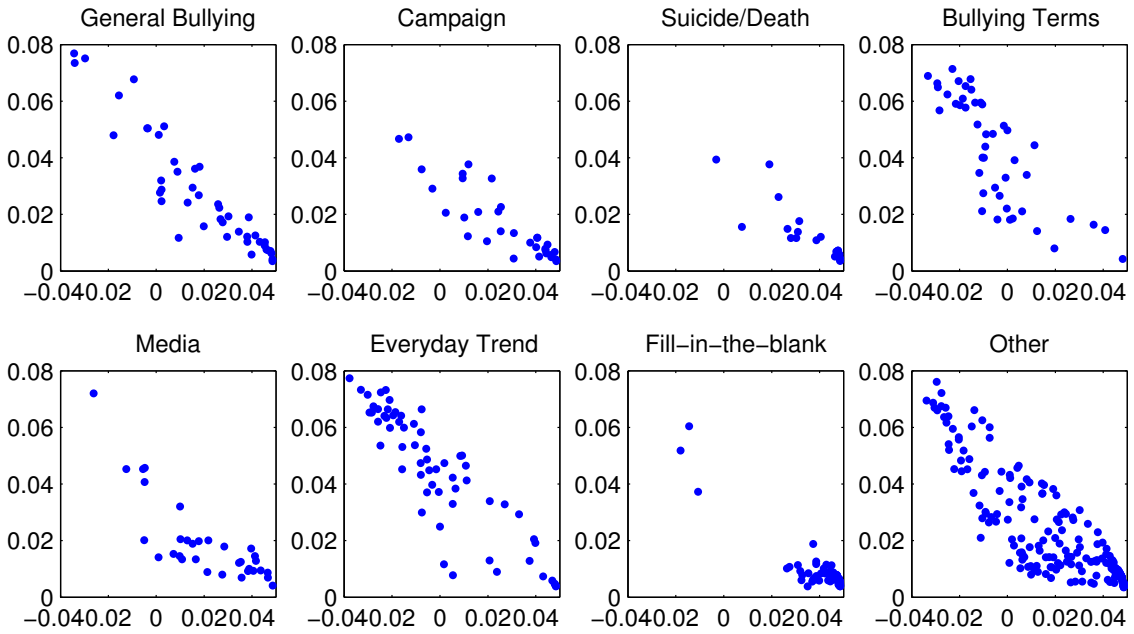


Figure 5.1: Representation of the hashtags within each of eight hashtag categories along Principal Component 1 (x-axis), which is related to the number of major peaks and Principal Component 2 (y-axis), which is related to the relative level of background counts and peaks.

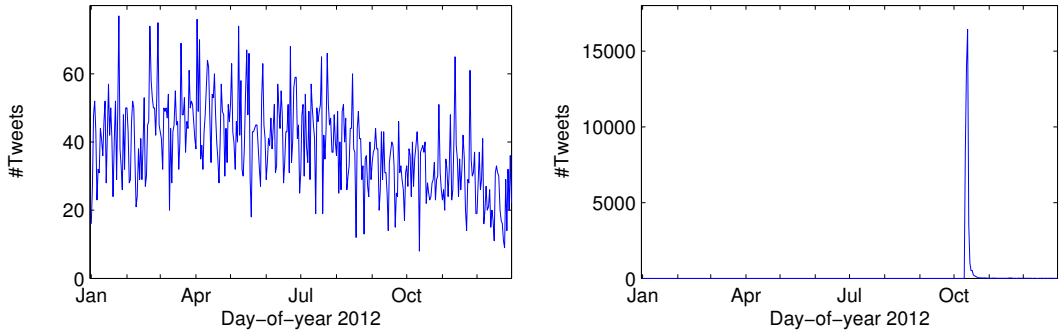


Figure 5.2: Number of bullying keyword tweets that contain #oomf (left) and #ripamandatodd (right) on each day of 2012

Therefore, we defined a shifting function $\tau(\mathbf{x}, t)$ as

$$\tau(\mathbf{x}, t) = [x_{t+1}x_{t+1} \dots x_{366}x_1 \dots x_t]$$

which shifts the first t elements of $\mathbf{x}$ to its end. We define the cosine distance with alignment as

$$d(h_1, h_2) = \min_{t=1,\dots,366} \text{consine\_dist}(\mathbf{x}_{h_1}, \tau(\mathbf{x}_{h_2}, t))$$

Intuitively, the measure tries all possible ways to shift one temporal curve to match the other, and uses the cosine distance between the best match to the other as the distance.

We computed the pairwise distances between all pairs of hashtags and performed PCA to embed the hashtags in two-dimension space. The results are shown in Figure 5.1. Each dot represents one hashtag, and hashtags with similar trends are closer. Each panel shows hashtags of one category, and all the panels have exactly the same ranges. The PCA procedure did not use any category information. We investigated the curves embedded at different positions in the space, and found that the two components identified by PCA are related to the shapes of the curves. Principal Component 1 (x-axis) is related to the number of major peaks, and Principal Component 2 (y-axis) is related to the relative level of background counts and peaks. #oomf (Figure 5.2 (left)) at the upper left corner has the highest background count level; in other words, it has a lot of major peaks with similar scale. At the other end, #ripamandatodd (Figure 5.2 (right)) at the lower right corner have a single peak and zeros for all other days.

The PCA results also show that for some categories, temporal curves of hashtags in the same category tend to be similar. Most temporal curves from Suicide/Death and Fill-in-the-blank are embedded in the right lower corner, because they each have a single peak as the associated events happen only once. Most curves for Campaign and Media are embedded in the right lower corner as well, but spread over more space than the previous two categories, because there are some campaigns and TV shows that run multiple times a year. They may have multiple peaks but not too many. Most curves of Bullying Terms and Everyday Twitter Trends are embedded

in the left upper corner, as most of them are used on a daily basis. The General Bullying and Other category spread over all spaces, as they capture different types of curves discussed above.

5.4 Discussion

We identified the most widely used hashtags associated with Twitter posts from 2012 that used bullying keywords. The results reveal a range of uses for hashtags associated with bullying. The uses ranged from discussing high profile suicides/deaths to discussing current television programs to promoting antibullying campaigns and participating in Twitter culture through games and established hashtags. Importantly, we found several features from the tweets associate with different hashtags, which indeed demonstrated differences between the hashtag categories. The differences between categories show how hashtags associated with bullying can have both an immediate, large-scale influence on a lot of people as well as a more enduring everyday impact.

One consistent difference between hashtag categories was that hashtags within the General Bullying category (e.g., #bully, #bullying) were found to occur within the highest number of tweets and to be associated with the highest number of different authors. Both of these features indicate widespread usage of the hashtag within bullying keyword tweets. The finding that General Bullying hashtags occurred within the highest number of tweets might be explained by the fact that the tweets studied in our work all contained bullying keywords. As a result, the same text, #bully, might serve as both a keyword for collecting the tweet and a hashtag. However, this overlap is not explained by the fact that General Bullying hashtags were used by the highest number of authors. Rather, it may be that authors who hoped to gain attention to their post about bullying chose to include the most generic terms that directly related to bullying.

Hashtags within the Bullying Terms category occurred in tweets with the largest number of URLs and the largest number of hashtags. This group of hashtags, which contained terms such as “#parents,” “#children,” “#school,” may have been used

to share information about a range of topics related to bullying or with key stakeholders. The target audience may have been directed to a website with additional information on these topics. To evaluate if this is the case, the content of the links could be coded in a future analysis.

The hashtags associated with the largest fraction of retweets were General Bullying, Suicide/Death, Campaign, and Fill-in-the-Blank/Game hashtags. This pattern makes sense especially for the Suicide/Death and Campaign hashtags as users may have been re-reporting events or campaigns with their tweets. This sort of activity could benefit organizations promoting their campaigns about bullying. Note that the Campaign hashtags were among the categories that contained the most mentions of other users. This might be the first step in the process–A campaign can mention several users within their tweet and these users will then be notified of their inclusion within the tweet. This may prompt the users to retweet the message. This process is important for nonprofit organizations that use hashtags to relay information, foster online communities with their followers, and promote action by their followers (Lovejoy and Saxton, 2012). Retweets also indicate endogenous influence suggesting that news about bullying events, such as Suicide/Death and Campaign hashtags, are spread largely through users forwarding these events to their followers (Lehmann et al., 2012).

Sentiment analysis revealed differences in which emotions were distributed across all 500 top hashtags associated with bullying keywords, but not between-category differences. In general, most tweets (i.e., 75%) using the hashtags were neutral in their tone. Twenty percent of tweets were identified as negative in tone, and 5% were positive. While the percentage with an emotion may seem low, it is higher than the 6% of bullying episodes found to contain an emotion in Section 4.3. Different methods were used to identify emotions in the two studies, but the finding may suggest that hashtag usage is associated with stronger emotionality.

The temporal analysis illustrated striking differences between hashtag categories. Among the eight categories, we found several different patterns of hashtag use across one year. The daily pattern of hashtag popularity has been described as originating through both exogenous and endogenous propagation (Lehmann

et al., 2012). Exogenous propagation involves a hashtag becoming popular because of an online or offline influence outside of Twitter, such as the media (news or entertainment) or relevant events (elections, sports, etc.). Endogenous reasons for a hashtag becoming popular involve activity within Twitter through either retweets, popular users such as Lady Gaga using the hashtag (Katrandjian, 2011), or the posting of the hashtag on Twitter's top trending topics list. Popular hashtags display four common daily temporal profiles around a peak in which the hashtag is most frequently used: activity before and during the peak (usually involving anticipation for a particular event such as #spiritday), high activity during the peak and after (occurs with unexpected events such as a teen committing suicide), activity symmetrically occurring before and after the peak, and activity occurring on the single day during the peak (Lehmann et al., 2012). High activity during the peak and afterwards characterized how the hashtag #ripamandatodd, a hashtag within the Suicide/Death category, was used. In contrast, more symmetric usage across the year, with no single peak, characterized how the #oomf, an Everyday Trend was used. These trends map on for these individual hashtag examples. However, some hashtag categories showed a lot of differences in their temporal patterns. Future work is needed to evaluate whether important subcategories of hashtags exist that might explain these different temporal patterns.

Limitation

The present study reflects an inquiry into a new area of research on bullying. As such, it faces some unique limitations. It is important to keep in mind that this study examines these hashtags only in relation to tweets that contained bullying keywords. The exact same hashtag may operate very differently when combined with other words. An additional challenge within our work is that many hashtags were coded as "Other" because their function could not be determined from single tweets alone. Some hashtags also were not necessarily hashtags-e.g., #1. Because we relied on a # matching procedure to collect hashtags, we may have included some that were not intended to be used as such. Our use of the Twitter API generates

individual tweets with our criteria (i.e., keywords), rather than full conversations. This makes it difficult to evaluate the context and meaning of hashtag usage without additional exploration. Further, a small proportion of tweets that use hashtags could be considered spam, such that possible users may be trying to promote or advertise a website, video, or business by composing the tweet almost entirely with hashtags to gain attention. Future studies might filter out such tweets as they may skew the data in detrimental ways.

An additional limitation is that we do not know how the hashtag originated or whether the hashtag was introduced through endogenous or exogenous means. For instance, many media outlets now use hashtags to gather real time opinions or for fans, viewers, or readers to follow a particular live event through hashtag searching (e.g., #xfactor). Further resources are needed to identify who first initiated the hashtag, which could give a better idea of the hashtag's purpose. Like the game of telephone, sometimes the message or news is altered from user to user or from retweet to retweet (boyd et al., 2010). Locating the origin of the hashtag will help to track the progression of bullying events, opinions, and trends.

Future Directions

Our work provides a first glance into how bullying is discussed within social media. It illustrates that the online world is an important data source for observing social interactions that take place there. There are numerous directions future work could pursue. It would be interesting to see how conversations on Twitter affect a particular event. In the present study and most other studies, the presumptions is that particular event impacts Twitter activity. However, social media usage itself may have an impact on an event or future events, such that popular hashtags may get more attention (e.g., #bringbackourgirls (Kirkland, 2014)). Another line to pursue could be the association between posting on social media and offline behavior. Sharing publically through hashtags online about one's bullying opinion, bullying events, or promoting a bullying cause may change how users see themselves, make users more accountable for their actions, and pressure users to follow through

on their beliefs (Gonzales and Hancock, 2011). For instance, do users that tweet "Stand up to bullies" or "be a friend to someone who is bullied" actually perform these activities? Future studies connecting social media behavior with overt offline behavior could be beneficial for bullying intervention. Finally, the focus here was at the tweet level. Public social media data also allow researchers to take a user-level perspective. For example, an important next step would be to investigate the network infiltration of bullying keyword hashtags. Power or popularity within the conversation network is achieved through other users retweeting one's tweets, and having a large number of followers. Having one's hashtags used by others may be one way to gain this power. It would be intriguing to examine how different types of hashtags travel through users within social media to identify those who are the most influential on different topics related to bullying. In this way, the influence of social media on bullying could be parlayed for good to get positive messages and information out rather than only serving as conduit for cyberbullying.

6 CULTURE DIFFERENCES IN BULLYING TRACES

Bullying at school is a worldwide health issue among adolescents. During the last decade, several East Asian countries and regions started paying close attention to this problem as well. Researchers have reported bully and victim prevalence rates and forms in Asian countries that are similar to those in western society (Kanetsuna et al., 2006; Schwartz et al., 2001; Wei et al., 2007).

Yet similar prevalence rates do not necessarily mean similar dynamics among bullying participants. East Asian cultures are more prone to emphasize the development of interdependence and the relational self, during which an individual is expected to keep group harmony and align one's own behaviors with others' in the same context, whereas individuals in western countries are more prone to develop a sense of independence and the separate self (Kağitçibaşi, 2007; Lam and Zane, 2004). These cultural differences have implications for the different behaviors of participants in bullying episodes. However, to the best of our knowledge, the study on such differences is largely unexplored.

The widespread usage of social media makes it convenient to collect data from different countries. This facilitates many cultural comparative studies, such as user behaviors (Yang et al., 2011) and emoticon usages (Park et al., 2013). We propose to use social media as an excellent data source for a cultural comparative study on bullying.

In this chapter, we collect a bilingual microblogs corpus on school bullying[1] , including English posts (tweets) from Twitter.com and Chinese posts (weibos) from Weibo.com, to study the differences on school bullying behaviors between western society and China. This dataset and its documentation is archived as Bilingual Bullying Traces Data Set (see Appendix A.7) We investigate the corpus to examine cultural differences in author's role, teasing, temporal dynamics and social process. We also hypothesize possible explanations for these differences.

[1]Our corpus is not aligned, meaning that one language is not the translation of the other. The data is available at http://research.cs.wisc.edu/bullying.

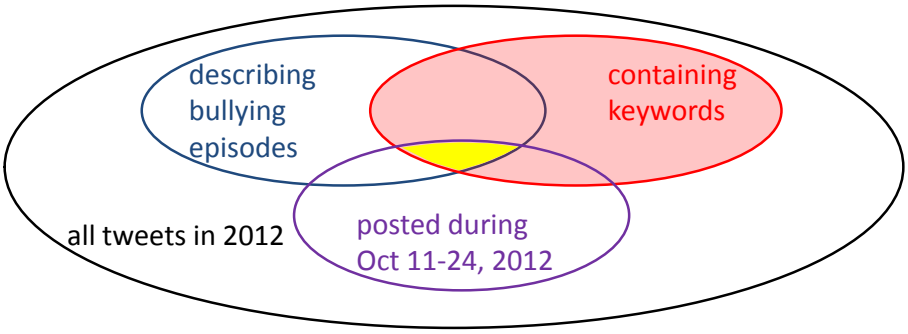


Figure 6.1: Venn diagram of bullying tweets. The temporal analysis is based on the red and yellow set. All other analyses are based on the yellow set only.

6.1 Data Collection

We collected English tweets using the public Twitter Streaming API by tracking bullying related keywords: "bully," "bullied," and "bullying." As our focus is on school bullying posts, we only kept the tweets which further contain at least one of the school-related words: "college," "university," "school," and "class." The filtering is case-insensitive and we included the plural forms of these keywords. We removed retweets by filtering tweets with the token "RT."

We collected Chinese weibos through the keyword search function provided by Weibo.com. Since there is no single term in Chinese that exactly corresponds to the English word bullying, we considered all seven near synonyms suggested in (Smith et al., 2002): 凌辱, 欺负, 欺凌, 欺辱, 欺侮, 欺压, 侮辱. We chose three corresponding school keywords: 学, 校, 班. We required at least one match from each keyword list, with the option "original post only" to exclude posts reposting other weibos.

We collected data in this way for the whole year of 2012. In total, there are 756,449 tweets and 75,044 weibos in our dataset (the red and yellow set in Figure 6.1). As shown in Section 2.1, not all of keyword-filtered posts are *bullying traces*, i.e. posts describing actual school bullying episodes.

To ensure the quality of our result, we conducted our analysis on an annotated

Author's Role	Tweet		Weibo		p-value
Accuser	67	6.0%	11	1.4%	6.5×10^{-7}
Assistant	1	0.1%	1	0.1%	6.3×10^{-1}
Bully	72	6.4%	133	16.4%	3.6×10^{-12}
Defender	26	2.3%	36	4.4%	1.3×10^{-2}
Reinforcer	3	0.3%	1	0.1%	8.5×10^{-1}
Reporter	429	38.3%	296	36.5%	4.6×10^{-1}
Victim	523	46.7%	333	41.1%	1.6×10^{-2}

Table 6.1: Number and percentage of author's role in bullying traces.

subset of the corpus. We selected a study period of October 11-24, 2012, with the consideration of avoiding major vacations and holidays. 45,785 tweets and 3,123 weibos fell in this study period. To reduce the burden of annotation, for each day we randomly subsampled tweets so that it has the same size as all weibos collected on that day. Therefore, our annotators labeled 3123 tweets and 3123 weibos (purple set in Figure 6.1). Among them, 1121 (36%) tweets and 811 (26%) weibos were coded as bullying traces (yellow set in Figure 6.1). One possible explanation for the lower percentage of bullying traces in Weibo is that multiple Chinese bullying keywords have other meanings as well.

6.2 Fewer Victims in Weibo

In Section 2.2, we categorize the author of a bullying trace into several role. We expected the roles to be identical across the two cultures, but hypothesized that their distribution may differ. Therefore, our annotators labeled each author's role of the 1121 tweets and 811 weibos. Table 6.1 shows the number of posts from each role and their percentages. We conducted χ^2-tests to test if the fraction of one category in tweets is significantly different from the one in weibos, and reported the p-value in the table, too. We found that the fractions of bullies and defenders in weibos almost double the ones in tweets. On the other hand, the fractions of accusers and victims in Weibo are significantly lower than the ones in tweets.

	Tweet		Weibo		p-value
Teasing	44	3.9%	70	8.6%	2.3×10^{-5}

Table 6.2: Number and percentage of teasing posts in bullying traces.

The distribution of author roles may reflect cultural differences between the two societies. Asian culture differs from western culture in that it stresses values of interdependence in which the development of relational self is emphasized and group harmony is highly valued over individual independence (Wei et al., 2007). In contrast, western society is conceptualized as a culture of independence in which the independent and separate self is strongly shaped (Kağitçibaşi, 2007). It is possible that youth in the Asian culture, where greater emphasis is on interpersonal relationships, will perceive more social responsibilities for each other in terms of offering help in a peer victimization event. As a result, more youth may be identified as defenders in the Chinese language social media posts.

There were fewer victims identified in the Asian culture. This may be because of the prevalent notion of "saving face" – the confidence and moral values in ego's integrity that an individual must keep (Shi, 2011; Yu, 2003). In contrast to posts generated in tweets, Weibo victims, to save face, may be less likely to post about their own experiences and others may be less likely to post about them. Instead, more people label themselves or act as a bully in weibos.

6.3 More Teasing in Weibo

In Section 4.1, we discusses that some bullying traces are written jokingly, which indicates lower severity of a bullying episode; It may also represent positive social interaction among friends to increase relational bonds. For example, (Tweet)*"Miss them. No, don't think if I miss the school but I miss my friends. I miss the moment when I bullying them ·_· Well, I miss the foods too."*

Due to the different levels of self concerns and face concerns in the cultures, we would expect that Asians are more likely to accept teasing because they tend to think

Author's Role	Tweet		Weibo		p-value
Accuser	11	25%	0	0%	4.6×10^{-5}
Bully	10	23%	39	56%	1.1×10^{-3}
Defender	1	2%	0	0%	8.1×10^{-1}
Reporter	7	16%	10	14%	9.7×10^{-1}
Victim	15	34%	21	30%	4.6×10^{-1}

Table 6.3: Number and percentage of author's role in teasing bullying traces.

affiliation is a positive consequence of teasing with friends (Keltner et al., 2001). Table 6.2 shows the number of teasing posts within the posts coded as bullying traces. Among the annotated bullying traces, 44 (3.9%) tweets and 70 (8.6%) weibos were written jokingly. The fraction of teasing posts is significantly higher in Weibo than in Twitter (p-value 2.3×10^{-5}). More Weibo users talk about bullying as an interaction among friends, instead of a serious issue.

Members of different cultures and backgrounds may tease and perceive teasing in different ways (Campos et al., 2007). Table 6.3 shows the number and percentage of author's roles in teasing bullying traces. Here are some example teasing posts where the author takes the bully role (Weibo, translated)*"I used to bully a nerd boy in my class. In fact, it was wrong when I look back. I am a fool. lol lol"*, the accuser role (Tweet)*"@USER @USER report yall for cyber bullying ! lol"*, and victim role (Tweet)*"@USER lol shut the hell up. You're always bullying me at school in the hallways!"*

More than half of teasing weibos were written by bullies, and the fraction is significantly higher than the one in Twitter. In contrast, we found more teasing tweet from accusers and victims. This result is consistent with our assumption on saving face in the Asian culture. Even in teasing, users tend to act as bullies instead of victims.

6.4 More Weibo Posts in the Evening

Understanding the temporal dynamics of school bullying is important for research and practice. The traditional social science study of bullying relies on personal

	In Semester	Off Semester	Off/In Ratio
Tweet	2194	1770	81%
Weibo	222	157	71%

Table 6.4: Average daily counts of microblogs containing school bullying keywords off/in-semester, and the ratio of these two categories.

surveys in schools. The number of participants and the frequency of such survey are usually low. Therefore, the study of temporal dynamics is handicapped by data scarcity. In contrast, we can collect a large number of school bullying microblogs at near real-time with very high temporal resolution.

Figure 6.2 (left) shows the percentage of microblogs containing school bullying keywords we collected from Twitter and Weibo in each day of 2012. Although these counts include the false positives (non bullying traces), the false positive rate is relatively stable during the study period of October 11-24, 2012. Therefore, the trend of actual bullying traces should be similar to Figure 6.2(left).

We first look into the peaks and valleys. Twitter has several extremely high and narrow peaks, which are usually caused by special events. On the other hand, Weibo has a relatively stable but slowly increasing trend. It is possible that new users kept signing up. Most narrow valleys in both platforms appear during weekends, when students have less direct interactions. The percentages are even lower during major long holidays, as highlighted in Figure 6.2(left).

To quantify the differences between in-semester and off-semester, we computed the average daily counts of microblogs containing school bullying keywords. Most schools in western societies are in-semester during mid-January to mid-June and September to mid-December. Most schools in China are in-semester during mid-February to end of June and September to mid-January. All other days are considered as off-semester. Table 6.4 shows the results. There are more posts with school bullying keywords in-semester as we expected. However, the number of such posts off-semester is far from zero. This shows that a focus on in-semester data collection as is normally done in psychology may be missing the bigger picture, since bullying or discussion thereof are happening off-semester as well.

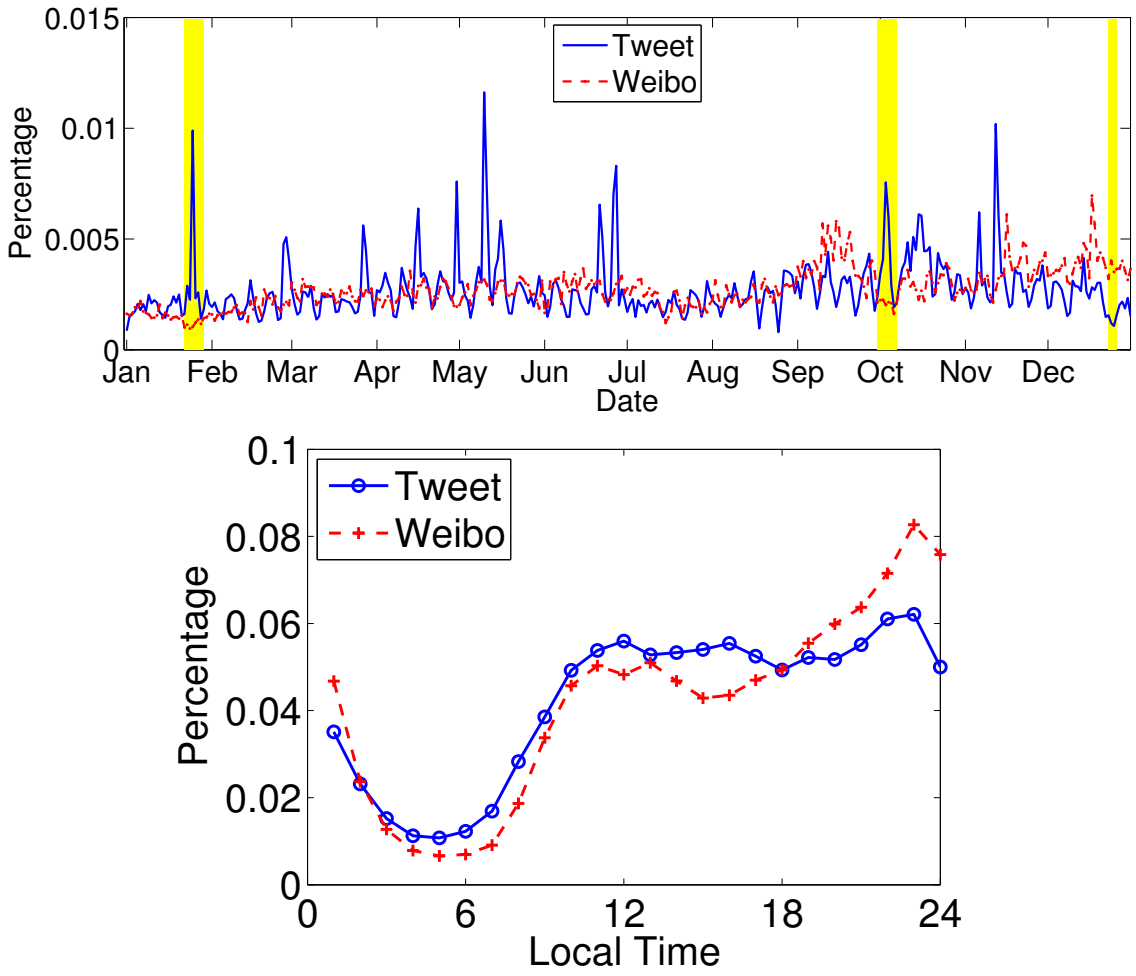


Figure 6.2: (top) The percentage of microblog posts containing school bullying keywords created in each day over the year of 2012. The highlight regions are(from left to right): Chinese New Year, Chinese National Day, and Christmas. (bottom) The percentage of microblogs containing school bullying keywords created in each hour-of-the-day.

	Family	Friend	Humans
Tweet	0.61	0.25	1.84
Weibo	1.41	0.24	2.10

Table 6.5: Social process scores of bullying traces by LIWC.

It is also interesting to look at the number of posts created in each hour-of-the-day. China uses a single time zone, and timestamps in Weibo are in local time. Twitter users spread cross many time-zones and location information is needed to convert the timestamps to the user's local time. We employed a reverse geocoding database (http://www.datasciencetoolkit.org) and a rule-based string matching method to map tweets to their origins (at the state level; only for tweets within the United States).

Figure 6.2(right) shows the percentage of microblogs containing school bullying keywords created in each hour-of-the-day. For both Twitter and Weibo the percentage is low at late night and in the early morning, and high in the evening. This is the typical diurnal social media usage pattern as we expected. The difference between the two cultures is obvious if we compare two time intervals, afternoon (12:00-18:00) and evening (18:00-24:00). From afternoon to evening, the increment of Weibo is more significant. This difference may be caused by the difference of cellphone usage policies in schools between the two countries. It is also possible that China may have a longer school day than the US (Fuligni and Stevenson, 1995), so Twitter users have more hours in the afternoon when they are free to generate posts.

6.5 Family Mentioned More in Weibo

Social media users are involved in different social groups, families, and friends. We want to see the strength of interactions with different groups when users talk about their bullying experiences. Linguistic Inquiry and Word Count (LIWC) (Tausczik and Pennebaker, 2010) is a text analysis tool, which calculates the degree to which people use different categories of words. We applied LIWC to the annotated bully-

ing tweets and the weibos (translated into English by Google Translate). Google Translate did a reasonable job in word choice, which is sufficient for word counting by LIWC.

Table 6.5 shows the scores of different categories under social process produced by LIWC. Weibo users use more words related to family, as the significance of family in Asian countries is presumed to be higher than in western countries in line with its collectivistic orientation where the group is emphasized over the individual (Triandis, 1995). Chinese parents pay close attention to children's education performance and environment. For example, (Weibo, translated) *"There is one bully in my daughter's class. Several parents complain that he bullies other girls, graping their faces, even pushing them from stairs. My daughter also told me many times. I think naughty is children's nature, but manners are also very important. Parents should not let their children be offensive. They should see a psychiatrist and apologize to other parents."*

6.6 Discussion

Social media provide an excellent data source for comparative study of school bullying in different cultures. We collected and annotated a bilingual microblogs corpus on school bullying consisting of Chinese Weibo and English Twitter posts. We examined the differences in author's role, teasing, temporal dynamics and social process, and proposed possible explanations for several observed differences. There could be alternative causes for our findings as well. For instance, both Twitter and Weibo limit a post to 140 characters, but in English and Chinese, respectively. The information content of a single Weibo post is thus considerably higher than that of a tweet. Such difference may affect our annotator's confidence and hence the labels. In future work we plan to validate these and other hypotheses.

7 SEGMENTING USER'S TIMELINE INTO EPISODES

So far, we have been investigating individual tweets one by one. Since tweets are short and some are ambiguous, it would be helpful to look at it under its context if available. More context about the incidents could be more informative for us to determine if they are bullying traces and other characters of these incidents. We use the count of bullying traces to study the prevalence rates of bullying episodes in previous chapters, with the implicit assumption that individual bullying trace correspond to distinct bullying episodes. However, this may not be always true, as several users may have a conversation containing multiple bullying traces.

To address these issues, we should put the bullying traces in the same episode together. However, given the huge amount of bullying traces, it is not feasible to consider all pairs of bullying traces without explicit connections. Therefore, as a first step towards collectively investigate bullying posts, we focus on users' timelines. The timeline of user u includes (a)tweets created or retweeted by u, (b)replies to any tweet created by u, and (c)retweets of any tweet created by u. This can be easily collected through Twitter Streaming APIs via following a set of users.

In this chapter, we are interested in identifying *name calling episode*, which includes the tweets where name calling happened and all the context tweets about this incident. Name calling could be one form of verbal bullying. We solve this task by two steps. We first segment the timelines into episodes, *i.e.,* assign each tweet an episode id. Tweets with the same episode id are considered to be talking about the same episode. The second step is training a classifier to recognize name calling episodes from the segmentation result. We focus on the first step in this chapter.

7.1 Proposed Model

Notation

For notation simplicity, we describe our model with a single timeline. It is straightforward to extend our model to multiple ones by considering them as independent

realizations of the same random process.

Let the total number of tweets collected in the timeline be N. Denote $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, where $\mathbf{x}_i \in \mathbb{Z}^{*V}$ is the bag-of-words counts for i-th tweet and the vocabulary size is V. For each tweet, we also have its creation time $t_i \in \mathbb{R}_+$ from its timestamp. The tweets are sorted by their creation time, so $t_1 \leqslant t_2 \leqslant \cdots \leqslant t_N$. Let $\mathcal{S}$ be the set of links between tweets. We consider two types of links, $\mathtt{reply}$ and $\mathtt{prev}$. If tweet i replies to tweet j, then $(i, j, \mathtt{reply}) \in \mathcal{S}$. If a tweet replies to a tweet which is not in our dataset or it does not reply to any other tweet, we call it *original tweet*. If tweet i is an original tweet, and tweet j is the most recent original tweet before tweet i, then $(i, j, \mathtt{prev}) \in \mathcal{S}$. Note that the links are directed. For any tweet $i > 1$, there is exactly one link pointing from i. So $\mathcal{S}$ defines the edge set of a tree.

Distance Dependent Chinese Restaurant Process (dist-CRP)

Tweets in a timeline comes sequentially, which is very similar as the customers in Chinese Restaurant Process (CRP). We want to assign the tweets to their corresponding episodes. As a non-parametric model, CRP does not require setting the number of tables. This is very important for our task, as we may not know how many episodes in a timeline. It is not valid to assume tweets are independent or exchangeable and use standard as the generative story for the timeline. Zhu et al. (2005) and Blei and Frazier (2011) defined some variations of CRP, where customers arrive in sequential, and the probability that a new customer sits at the same table with previous customers depends on the history and the distances.

CRP represents the partition of customers with *table assignments*. Each customer is assigned with a table label (cluster index). Dist-CRP (Blei and Frazier, 2011) represents the partition with connections between customers. When a new customer comes, dist-CRP assigns the index of customer with whom the new customer sits. For customer i, c_i is a back pointer to the previous customer with which i shares the same cluster. At the end, if we connect customers by this relationship, customers in the same connected component are assigned to the same cluster. So note that $c_i \in \{1, \ldots, i\}$ in the following section means that tweet i is in the same episode with

tweet c_i; if $c_i = i$, tweet i starts a new episode. c_i is not the cluster assignment.

Let the first tweet starts a new episode $c_1 = 1$. When the i-th tweet arrives (ordered by the creation time), the probabilities that it has the same episode with tweet $k \in \{1, \ldots, i-1\}$ and the probability that it starts a new episode ($k = i$) are defined as,

$$P(c_i = k) = \frac{h(i,k)}{\sum_{k'=1}^{i} h(i,k')}, k \in \{1, 2, \ldots, i\} \tag{7.1}$$

where $h(i,k)$ is a function that reflects the probability that tweet i has the same episode with tweet k, and $h(i,i)$ reflects the probability that tweet i starts a new episode. So we may still have the probability to come back to any old episodes, and the probabilities may decay over the distance between the pair of tweets.

Notice the choice of c_i does not depend on the values of other c_{-i}. The probability is determined by the $h(i,\cdot)$. So the joint distribution of $\mathbf{c} = \{c_1, \ldots, c_N\}$ is

$$P(\mathbf{c}) = \prod_{i=1}^{N} P(c_i \mid h(i,\cdot)) \tag{7.2}$$

Choice on $h(i,k)$

In general, we have M different features to characterize the probability that $c_i = k$, $k \in \{1, 2, \ldots, i\}$ with feature vector $\mathbf{f}_{ik} \in \mathbb{R}_+^M$, and we have a weight vector $\boldsymbol{\beta} \in \mathbb{R}^M$. Note that the feature vector $\mathbf{f}_{ik}$ is defined on a pair of tweets, instead of individual ones. We require that the features take non-negative values, and larger feature values mean that the pair of tweets tend to be in the same episode. Since $\mathbf{f}$ are non-negative, we require $\boldsymbol{\beta} \geqslant 0$ to make sure h is non-negative.

$$h(i,k) = \boldsymbol{\beta}^\top \mathbf{f}_{ik}, \boldsymbol{\beta} \geqslant 0 \tag{7.3}$$

In our tasks, we have four different sources of information that may be related to the probability: difference in creation time, reply relationship, temporal relation-

ships and text similarity. For each pair of tweets i and k, $k \leqslant i$ we have

$$\mathbf{f}_{ik} = \begin{pmatrix} \mathbf{1}[k = i] \\ \frac{\mathbf{1}[k<i]}{|t_i - t_k|+1} \\ \mathbf{1}[(i, k, prev) \in \mathcal{S}] \\ \mathbf{1}[(i, k, reply) \in \mathcal{S}] \\ \mathbf{1}[k < i] \frac{\mathbf{x}_i^\top \mathbf{x}_k}{\|\mathbf{x}_i\| \|\mathbf{x}_k\|} \end{pmatrix} \tag{7.4}$$

Note that we consider the sequential CRP. For any $k > i$, we define the feature vectors $\mathbf{f}_{ik} = 0$, so $h(i, k) = 0$. When the first feature $\mathbf{1}[k = i]$ is one, all other features are 0. This allows us to specify the potential that tweet i starts a new episode, so β_1 serves as the concentration parameter α in dist-CRP.

From c to Cluster Assignment

The distribution of $P(\mathbf{c})$ is well defined by dist-CRP. An assignment of $\mathbf{c}$ for all tweets induces a partition of the tweets. However, the mapping from $\mathbf{c}$ to partition is many-to-one, *i.e.* multiple different values of $\mathbf{c}$ could induce the same partition. In our task, we care more about the partition, instead of how the tweets link to each other. Our labeled data have episode IDs, but no direct labels on $\mathbf{c}$. So we need to fill the gap between $\mathbf{c}$ and the segmentation result.

For a partition $B = \{B_1, \ldots, B_M\}$ of N tweets with M blocks ($M \leqslant N$), we label each block B_m, $m \in \{1, \ldots, M\}$, with the smallest index of the tweets in B_m. Then we represent the partition B with $\mathbf{z} = \{z_1, \ldots, z_N\}$, where z_i is the label of the block in which i-th tweet is assigned to by B. It is obvious that for each B, we will have a unique $\mathbf{z}$; and for any meaningful $\mathbf{z}$, we could recover the partition B.

Since the block label is the smallest index in a block, $z_i \in \{1, \ldots, i\}$ for all $i \in \{1, \ldots, N\}$. However not any sequence $\{z_1, \ldots, z_N\}$ is a valid representation of a partition under our block label rule. For example, $\{1, 1, 2\}$ is not valid, as the label for the block that the third tweet belongs to should be 3. Our model assigns probability zero to these $\mathbf{z}$.

We have defined the distribution of $\mathbf{c}$, and we know that $\mathbf{c}$ induces a unique partition B, and B induces a unique representation $\mathbf{z}$. Now let us consider the distribution of $\mathbf{z}$,

$$P(\mathbf{z}) = \sum_{\mathbf{c} \text{ that forms partition } \mathbf{z}} P(\mathbf{c}) \tag{7.5}$$

As we assume the tweets arrive sequentially, c_i could take any value in $\{1, \ldots, i\}$. If $z_i = i$, the i-th tweet starts a new episode, and c_i has to be i. If $z_i < i$, the i-th tweet belongs to some old episode with previous tweets, and c_i must point to one of the tweets in that episode. Given $\{z_1, \ldots, z_{i-1}\}$, z_i depends only on which tweet c_i points to.

$$\begin{aligned} \mathbf{1}[\mathbf{c} \text{ forms } z_i \mid z_1, \ldots, z_{i-1}] &= \mathbf{1}[c_i \text{ forms } z_i \mid z_1, \ldots, z_{i-1}] \\ &= \mathbf{1}[c_i = z_i = i \text{ or } (c_i < i \text{ and } z_{c_i} = z_i)], \end{aligned} \tag{7.6}$$

and

$$\mathbf{1}[\mathbf{c} \text{ forms } \mathbf{z}] = \prod_{i=1}^{N} \mathbf{1}[c_i = z_i = i \text{ or } (c_i < i \text{ and } z_{c_i} = z_i)]. \tag{7.7}$$

We have

$$\begin{aligned} P(\mathbf{z}) &= \sum_{\mathbf{c}} \mathbf{1}[\mathbf{c} \text{ forms } \mathbf{z}] P(\mathbf{c}) \\ &= \sum_{c_1=1}^{1} \sum_{c_2=1}^{2} \cdots \sum_{c_N=1}^{N} \left(\prod_{i=1}^{N} \mathbf{1}[c_i = z_i = i \text{ or } (c_i < i \text{ and } z_{c_i} = z_i)] P(c_i) \right) \\ &= \prod_{i=1}^{N} \left(\sum_{c_i=1}^{i} \mathbf{1}[c_i = z_i = i \text{ or } (c_i < i \text{ and } z_{c_i} = z_i)] P(c_i) \right) \end{aligned} \tag{7.8}$$

Another way to think about $P(z_i = j \mid z_1, \ldots, z_{i-1}, \mathbf{c})$ is that the cluster assignment only depends on previous cluster assignment and c_i. If $z_i = i$, then c_i has to be i; if $z_i < i$, then c_i points to any tweets in the same cluster will produce z_i. It

does not depends on other c_i or any future cluster assignment.

$$\begin{aligned}
P(z_i = j \mid z_1, \ldots, z_{i-1}) &= \sum_{k=1}^{i} P(z_i = j, c_i = k \mid z_1, \ldots, z_{i-1}) \\
&= \sum_{k=1}^{i} P(z_i = j \mid z_1, \ldots, z_{i-1}, c_i = k) P(c_i = k) \\
&= \sum_{k=1}^{i} \mathbf{1}[k = j = i \text{ or } (k < i \text{ and } z_k = j)] P(c_i = k) \\
&= \sum_{k=1}^{i} \mathbf{1}[k = j = i \text{ or } (k < i \text{ and } z_k = j)] \frac{h(i,k)}{\sum_{k'=1}^{i} h(i,k')} \\
&= \frac{\sum_{k=1}^{i} \mathbf{1}[k = j = i \text{ or } (k < i \text{ and } z_k = j)] h(i,k)}{\sum_{k'=1}^{i} h(i,k')} \\
&= \frac{\mathbf{1}[j = i] h(i,i) + \sum_{k=1}^{i-1} \mathbf{1}[z_k = j] h(i,k)}{\sum_{k'=1}^{i} h(i,k')}
\end{aligned} \tag{7.9}$$

Therefore, by the chain rule, we have the joint distribution of $\mathbf{z}$

$$\begin{aligned}
P(\mathbf{z}) &= \prod_{i=1}^{N} P(z_i \mid z_1, z_2, \ldots, z_{i-1}) \\
&= \prod_{i=1}^{N} \frac{\mathbf{1}[z_i = i] h(i,i) + \sum_{k=1}^{i-1} \mathbf{1}[z_k = z_i] h(i,k)}{\sum_{k'=1}^{i} h(i,k')}
\end{aligned} \tag{7.10}$$

Inference

For inference, we are given $\boldsymbol{\beta}$ and $\mathbf{f}$ for all pairs. So, we have the full distribution of $\mathbf{z}$ defined as above in the conditional form. Since we have included all the information in the function $h(\cdot)$, we do not have a generative model for the features from the distribution of $\mathbf{z}$. Therefore, we need to find the $\mathbf{z}$ with highest probability defined by Eq 7.10.

Sampling from the Conditional Distribution

We could sample $\mathbf{z}$ by following the above conditional probability $P(z_i = j \mid z_1, \ldots, z_{i-1})$ defined in Eq 7.9. If we are interested in finding the $\mathbf{z}$ with the highest probability, we can generate multiple samples, evaluate their probabilities and choose the one with the highest probability. It is easy to generate samples (time complexity $O(n^2)$), but the probability of hitting the mode is small. Another possibility is to sample $\mathbf{c}$, which is much faster. Then we convert $\mathbf{c}$ to $\mathbf{z}$, and see which $\mathbf{z}$ appears most frequently.

Beam Search

Instead of generating a single sample by following $p(z_i \mid z_1, \ldots, z_{i-1})$, we could do beam search to find the best configuration of $z_1, \ldots z_N$. For first few z_i's, we can keep all possible partial paths $z_1, \ldots, z_i$. Then for each path fit in our memory, we compute all possible configurations of $z_1, \ldots, z_i, z_{i+1}$ and their scores. If we have more than R (a constant to make sure all paths in memory) partial paths $z_1, \ldots, z_i, z_{i+1}$, we only keep the top R paths with the highest scores, and discard the others.

Parameter Learning

In our model, the only parameters to estimate is $\boldsymbol{\beta}$. We want to find $\boldsymbol{\beta}$ which maximizes the likelihood of $\mathbf{z}$ observed,

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\beta}) &= \log P(\mathbf{z} \mid \boldsymbol{\beta}) \\
&= \sum_{i=1}^{N} \log P(z_i \mid \boldsymbol{\beta}, z_1, \ldots, z_{i-1}) \\
&= \sum_{i=1}^{N} \log \left(\frac{\mathbf{1}[z_i = i] h(i,i) + \sum_{k=1}^{i-1} \mathbf{1}[z_k = z_i] h(i,k)}{\sum_{k'=1}^{i} h(i,k')} \right) \\
&= \sum_{i=1}^{N} \left[\log \left(\mathbf{1}[z_i = i] h(i,i) + \sum_{k=1}^{i-1} \mathbf{1}[z_k = z_i] h(i,k) \right) - \log \left(\sum_{k'=1}^{i} h(i,k') \right) \right] \\
&= \sum_{i=1}^{N} \left[\log \left(\mathbf{1}[z_i = i] \boldsymbol{\beta}^\top \mathbf{f}_{ii} + \sum_{k=1}^{i-1} \mathbf{1}[z_k = z_i] \boldsymbol{\beta}^\top \mathbf{f}_{ik} \right) - \log \left(\sum_{k'=1}^{i} \boldsymbol{\beta}^\top \mathbf{f}_{ik'} \right) \right]
\end{aligned} \tag{7.11}$$

To maximize this log-likelihood, we find the gradient of $\boldsymbol{\beta}$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{N} \left[\frac{\mathbf{1}[z_i = i]\mathbf{f}_{ii} + \sum_{k=1}^{i-1} \mathbf{1}[z_k = z_i]\mathbf{f}_{ik}}{\mathbf{1}[z_i = i]\boldsymbol{\beta}^\top \mathbf{f}_{ii} + \sum_{k=1}^{i-1} \mathbf{1}[z_k = z_i]\boldsymbol{\beta}^\top \mathbf{f}_{ik}} - \frac{\sum_{k'=1}^{i} \mathbf{f}_{ik'}}{\sum_{k'=1}^{i} \boldsymbol{\beta}^\top \mathbf{f}_{ik'}} \right] \tag{7.12}$$

We can use gradient method to maximize $\mathcal{L}$ with the constraints $\boldsymbol{\beta} \geqslant 0$.

Parameter Learning with Partially Observed z

Social scientists are only interested in name calling episodes. Therefore, they will only annotate z_i for those tweets in name calling episodes. The remaining z_i's are non-interesting, but separate episodes will not be annotated. However, for these unknown tweet i, we know that z_i could not be the same as any other observed cluster indices. So c_i can only point to itself or any other previous k with unknown z_k. Therefore, if z_i is unknown, we denote $z_i = unk$ and the possible value for z_i is $\{1, 2, \ldots, i\} \cup \{unk\}$.

$$P(z_i \mid z_1, \ldots, z_{i-1}) = \frac{\mathbf{1}[z_i \in \{i, unk\}]h(i, i) + \sum_{k=1}^{i-1} \mathbf{1}[z_k = z_i]h(i, k)}{\sum_{k'=1}^{i} h(i, k')} \tag{7.13}$$

Note that when both z_i and z_k are unknown, we consider $\mathbf{1}[z_k = z_i] = 1$.

Similarly, we find the gradient of $\boldsymbol{\beta}$, and estimate the parameter $\boldsymbol{\beta}$ by maximum likelihood via gradient descent.

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{N} \left[\frac{\mathbf{1}[z_i \in \{i, unk\}]\mathbf{f}_{ii} + \sum_{k=1}^{i-1} \mathbf{1}[z_k = z_i]\mathbf{f}_{ik}}{\mathbf{1}[z_i \in \{i, unk\}]\boldsymbol{\beta}^\top \mathbf{f}_{ii} + \sum_{k=1}^{i-1} \mathbf{1}[z_k = z_i]\boldsymbol{\beta}^\top \mathbf{f}_{ik}} - \frac{\sum_{k'=1}^{i} \mathbf{f}_{ik'}}{\sum_{k'=1}^{i} \boldsymbol{\beta}^\top \mathbf{f}_{ik'}} \right] \tag{7.14}$$

7.2 Experiment

Synthetic Dataset

To verify the effectiveness of our proposed model as well as learning and inference procedure, we create a synthetic dataset, where the groundtruth is available to

compare with.

Our model specifies the probabilities of partitions given the feature values and parameters. It is not a fully generative model to generate reply tree structures or tweet content. However, to generate our synthetic dataset, we need these features. Instead of hand crafting a small toy example, we generate our dataset with a probabilistic model, whose parameters is similar to our real data in next section. There are five features in Eq 7.1. To generate these feature values for our synthetic dataset, we need to generate creation time of each tweet, reply structure of timeline, and text similarity among each pair of tweets.

First, we generate tweet creation time with an exponential distribution. If we assume the number of tweets in a time window follows a Poisson distribution, then the gap between two consecutive tweets follows an exponential distribution. Therefore, we let the creation time of the first tweet in timeline to be zero, and for each new tweets, we add a random number sampled from exponential distribution with parameter 5. This is a simplified model, as we do not consider the diurnal patterns of social media users. If timelines span multiple days, it will be clear that user do not post for a few hours during sleep, but post a lot in a short time period. We think the simplified model is still valid, as it allows a large range of gaps.

Second, we need the reply structure for previous original tweet and reply-to features. For each tweet, we first sample a number from a Bernoulli distribution with head probability 0.6. It it is head, the current tweet is an original tweet, and we link it with 'prev' link to most recent original tweet before it. If it is tail, it is a tweet reply to a previous tweet in the timeline. We observed from the real data that about half of reply-to-other tweets replies to the most recent tweet in timeline, and about one quarter replies to the second most recent tweet in timeline, which suggests an exponential decay pattern. Therefore, if the current tweet is replying to other tweet, we sample all previous tweets with exponential decaying weights, which assigns the highest weight to most recent tweet.

To assign the text similarity among each pair of tweet, we use a simple language model to generate tweet text. We choose a vocabulary with size 100. For each tweet, its length is sampled from a Poisson distribution with parameter 15. Then

we sample that number of words from vocabulary with the same probability. Given the generated word counts, we can computer the similarities between each pair.

Parameter Learning

Following above procedures, we have created the unlabeled timelines. We want to show that with reasonable amount of annotation, our learning algorithm could learn the parameter. For purpose, we set the underlying parameter $\boldsymbol{\beta} = [1, 0.1, 0.1, 10, 0.1]$. We know the importance of reply-to links, and set a large weigh to it. We set equal values for other features except the first one, which controls the probability of new episode. We choose the timeline size to be 3000, which is a reasonable number for annotation. $\mathbf{z}$ is sampled as in our model to create clustering.

Our learning algorithm uses feature values $\mathbf{f}$ and labeled $\mathbf{z}$ to learn $\boldsymbol{\beta}$ as in our real task. With this amount of data, our algorithms learn $\widehat{\boldsymbol{\beta}} = [1.0000, 0.0778, 0.0701, 10.0632, 0.1011]$. The estimated result is very similar to the underlying parameters. It assigns the correct magnitude to different features.

Inference

To segment unlabeled timelines, we need to infer the partition with the highest probability given our model and learned parameters. As we mentioned, the search space is factorial in the number tweets in timeline. To make sure that we find the partition with the highest probability, we run global search, which computes the probability for every partition. Due to the huge search space, we were not able to do this with large number of tweets. Therefore, we set the number of tweets in timeline to be 12. Even with this small number of tweets, we have 479,001,600 different assignments of $\mathbf{c}$. We limit our beam size in beam search to be 10, which keeps only a tiny portion of partial paths. In each trial, we randomly generate features and use the true parameter to find the partition with highest probability. We repeated 100 trials, and beam search always finds the correct partitions. The global search takes 54.6 seconds for each trial, but beam search only takes 0.0159 seconds. So

User	#tweets	#episodes	#tweets in name calling episode	average size	median size
1	4,291	18	45	2.50	1
2	1,222	7	9	1.29	1
3	1,119	26	44	1.69	1
4	1,364	1	2	2.00	2
5	5,810	135	4,171	30.90	7
6	3,861	290	1,432	4.94	2
7	2,877	97	176	1.81	1
8	6,069	135	1,753	12.99	9
9	1,746	185	411	2.22	1

Table 7.1: Basic statistics of labeled name calling timeline data.

we believe beach search is efficient and have high probability to discover the best partitions for the given model.

Real Dataset

From April 2013, we have been following 5000 users, who posted most bullying traces during January - March 2013. Our annotators manually choose a few timelines with the hope to identify more name calling incidents in their timelines. For the chosen 9 timelines, the tweets were collected during September 2013. For each timeline, the annotators will try to identify all name calling incident and label all tweets in the same episode with the same episode ID. But they will leave all other tweets in the timeline, which are not related to any name calling episodes, unlabeled. More details on this dataset and its documentation is archived and can be found in Appendix A.8.

It is clear that the timeline is partially labeled as many tweets are not associated with any episode ids. Table 7.1 shows the statistics of our data set. The number of tweets are significantly different in different timelines, as user may post in different level of engagement in that month. The number of name calling episodes in each timeline and the tweets related to name calling episode are significantly different as well. For User 5, we have 135 name calling episode, and more than 70% tweets are

annotated with an episode id. On the other hand, there is only one name calling episode in timeline of User 4, which make it not useful for our experiment. The average size of episode (the number of tweets in one episode) is small for most timelines, except User 5 and User 8. In these two timelines, they constantly discuss some topics over time and many name calling in that topic were assigned with the same ID. The median size of bullying episode shows that most episode only have few tweets with it.

From these statistics, we see that timelines are very different. Therefore, we may need to learn parameters individually. We split each timeline into training and test parts. The training set contains the first 2/3 of tweets in each timeline, and the test set contains the rest 1/3. The goal is to learn the parameter with the training set and test the segmentation performance on the test set.

We choose the measurement commonly used in word segmentation (Peng et al., 2004) and topic segmentation (Cardoso et al., 2013) in natural language processing community. For each gap between tweet, there could be a separator. We measure the precision, recall and F-1 measure of the predicted separators comparing to the labeled data in test set. As our dataset is partially labeled, we do not consider the separators between two unlabeled tweets. However, we do measure the separators between labeled and unlabled tweets, as we know these unlabeled tweets belong to different episodes from the labeled ones.

We choose a simple baseline method, which assigns each reply-tree as a single episode. It doesn't consider time stamp and text features. This method does not need any training, and is a reasonable baseline.

Table 7.2 shows the results of our proposed method and the simple baseline on the test set. We did not include timeline of User 4, as there is only one episode. The baseline method achieves better numbers than our method on average. We notice that baseline method achieves higher performances on timelines of User 5 and User 8. In the timeline of User 5, several topics constantly appear over time. User may talked about multiple times, but the tweets in the same episode were split into multiple disconnected chucks. Over segmenting the timeline does not introduce extra errors. Therefore, baseline method may obtain better result. In the timeline

	Baseline			Proposed		
User	precision	recall	F1	precision	recall	F1
1	0.97	1.00	0.98	0.97	0.93	0.95
2	1.00	1.00	1.00	1.00	1.00	1.00
3	0.83	0.71	0.77	0.75	1.00	0.86
5	0.79	1.00	0.88	1.00	0.55	0.71
6	0.85	1.00	0.92	0.85	0.99	0.92
7	1.00	0.76	0.86	0.91	0.97	0.93
8	0.61	1.00	0.76	1.00	0.39	0.56
9	0.92	0.82	0.87	0.86	0.99	0.92
Average	0.87	0.91	0.88	0.92	0.85	0.86

Table 7.2: Performance of proposed and baseline methods on test set.

of User 8, only a few tweets were replying to other tweets. The user posts multiple short tweets in a short time period about one topic, which shows the time features and previous original tweet link might be important. But this is not always true. Actually, this timeline is also challenging for baseline method. Baseline method receives high recall but low precision, as it over segmented this timeline. Given these observations, we should consider the patterns how user posts and extract more meaningful features to capture these information and make better prediction.

8 CONCLUSION

This thesis has presented how we explore social media, with appropriate machine learning and natural language processing techniques, as a valuable and abundant data source for the study of bullying.

8.1 Summary

As the first step of the study, Chapter 2 demonstrated how we recognize bullying traces from large-scale real-time social media stream, and how we automatically extract basic information about these posts and underlying bullying episodes. Bullying traces account for only a tiny fraction of all social media posts, which poses a significant challenge for our annotators to find enough bullying traces without labeling an unreasonable amount of tweets. We restricted ourselves to an "enriched dataset" obtained by keyword filtering from the public Twitter streaming APIs. To further remove false positives, we explored multiple feature representations and classification algorithms, and SVM(linear) with unigrams+bigrams achieves the best accuracy at 86%. This accuracy is similar to the level of agreement achieved by two different human annotators. This trained Binary Bullying Trace Classifier identified 9,764,583 bullying traces between September 1, 2011 to August 31, 2013. To analyze the large amount of data, we also built machine learning models to identify participants and their roles, to categorize bullying traces by their types and forms of bullying episodes, and to discover the topics users are talking about.

In Chapter 3, we studied the spatiotemporal distribution of bullying traces, as several spatial and timing issues related to bullying episodes are important to know. As an empirical exploration study, we first analyzed raw counts of GPS-tagged bullying traces, whose location and time information can be extracted from meta-data. However, only 2% bullying traces contain GPS coordinates. Most posts do not include GPS coordinates, and self-reported locations can be inaccurate or false. Besides the data scarcity issue, such direct counting method is also plagued

by sample bias and incomplete data. To address these issues, we formulated the task as a Poisson point process estimation problem and propose Socioscope in Section 3.3. It explicitly incorporated human population bias, time delays and spatial distortions, and spatiotemporal regularizations into the model. Socioscope has broad applications where spatiotemporal signals are of interest, such as wildlife mortality, algal blooms, hail damage, and seismic intensity.

We focused on emotions associated with bullying traces in Chapter 4. First, we built a Teasing Bullying Traces Classifier to recognize bullying traces written jokingly, as there is considerable interest among social scientists to understand teasing in bullying traces. After manually inspecting a number of bullying traces, our domain experts identified seven most common emotions: anger, embarrassment, empathy, fear, pride, relief, and sadness. Some emotions have not been well studied in sentiment analysis community. Therefore, it requires manually labeling a large amount of training tweets or emotional lexicons. To address this challenge, we proposed a fast training procedure for sentiment analysis without explicitly producing a conventional labeled training dataset. We applied it to a large amount of bullying traces to study emotion distributions. Last, we investigated the behavior of deleting bullying traces after post. We managed to collect a corpus with deletion information, conduct exploratory analysis and build an off-the-shelf regret predictor.

Hashtags are widely used in Twitter to mark keywords or topics in a Tweet. Therefore, analyzing hashtags associated with public mentions of bullying can help us understand general discussions on bullying. In Chapter 5, we extracted 552,831 distinct hashtags used in tweets with the keywords "bully," "bullied," and "bullying" collected between January 1, 2012 and December 31, 2012. We organized the most frequently used 500 hashtags into eight categories. Hashtag features, including the number of tweets and retweets in which the hashtag appeared, the number of unique authors who used the hashtag, and the numbers of URLs, hashtags, and user mentions are found to be associated with hashtag category membership. Differences in the daily usage of hashtags used with bullying keywords are also identified. The differences show that bullying has both an immediate, large scale

influence as well as a more every day presence.

In Chapter 6, we collected and annotated a bilingual microblogs corpus on school bullying consisting of Chinese Weibo and English Twitter posts. We examined the differences in author's role, teasing, temporal dynamics and social process, and proposed possible explanations for several observed differences. First, we saw a smaller fraction of victim authors in Weibo than in Twitter. We hypothesized that this may be due to Asian culture's emphasis on saving face where it is more of a taboo to be a victim or label someone a victim. Second, we saw different temporal dynamics of school bullying posts due to differences in holidays and length of school days. Finally, bullying posts from Weibo contain more mentions of family than those from Twitter. This may be due to the greater emphasis on family in Asian cultures. There could be alternative causes for our findings as well.

Since individual bullying trace may be fragmental and noisy, we recovered the underlying episodes by piecing together multiple bullying traces about the same episode in Chapter 7. We focuses on user's timelines, which includes the tweets created or retweeted by seed user, replies to any tweet created by seed user, and retweets of any tweet created by seed user. We proposed a probabilistic model to segment timeline into multiple episodes. Based on our observations from data, we allowed the interleaving among episodes and the number of episodes are not pre-fixed. We applied beam search for inferring the episode assignment for new timelines and proposed efficient learning algorithms with partially observed data.

8.2 Future Directions

Our work introduces a novel data source and research approach to bullying study. It also introduces an interesting application to the machine learning, natural language process and social media mining communities. As such, it faces some unique limitations, and much work remains in this new research direction.

From Enriched Dataset to Full Range of Posts

We restrict ourselves to an enriched dataset, which is keyword filtered social media stream. This approach is able to collect a substantial number of bullying traces, but does not capture all bullying related tweets, as users may discuss bullying without directly using the keywords in our list. The keyword filter may introduce some potential bias in our dataset.

Therefore, a future direction is to extend the Binary Bullying Trace Classifier from the "enriched data" to the full range of tweets. There are a few challenges along this direction. First, most researchers do not have access or are able to process the full range of tweet stream. One trade off might be to strategically expand the keyword filtering to include additional words that capture bullying behaviors that might facilitate better representation of different forms of bullying on Twitter. However, even with this approach, it will be difficult to capture all bullying episodes because they can be represented in so many different ways. Furthermore, Twitter sets a rate limit on the percentage of tweets available in public streaming APIs. If researchers requested a long list of keywords or the requested keywords produce a large number of posts, rate limits will be hit and Twitter will sub-sample the resulting tweets. The sub-sampling procedure is not transparent to researchers and may introduce a potential sampling bias (Morstatter et al., 2013). It is essential to avoid this potential bias by carefully choosing keywords.

Second, since bullying traces only account for a small fraction of public social media posts, finding enough bullying trace examples to build classifiers requires a huge amount of annotation effort. One solution is to train the classifier with the enriched dataset as what we did. Clearly, the training set has different distribution from the targeting test set, the full range of tweets. Techniques used for *covariate shift* may be adapted to solve this problem (Blitzer, 2008). Another possible solution is using active learning to find diverse positive examples from the dataset, which can be used to train a classification model.

Social Network Structure

Our work mainly focuses on the content of social media, and investigates individual post separately. In Chapter 7, we collect user's timeline and have all posts in ego-networks centred at the seed users. There, we capture part of the interactions among users. We could push this further to collect data from a subset of social networks, from example, students in a school, as a bullying episode might be discussed in several separated groups. With the social network structure, friendship relation among users, we may piece these posts together to better recover the underlying episodes.

We could look into the friend relationships and interaction patterns among users to study how these factors correlate with bullying. We can conduct longitudinal studies. We observed that some users post bullying traces frequently, which may indicate that they constantly involve in bullying episodes. We would like to follow their posts for a longitudinal study. This helps us to study the evolution of roles in bullying episodes of a user over time. What's more important, we want to identify the individuals at risk. If we could identify them, necessary intervention may prevent some tragedies.

This chapter lists the datasets we have collected and used in our study. To facilitate researches on bullying, machine learning, and natural language processing, we try our best to make our datasets available at http://research.cs.wisc.edu/bullying. However, as per Terms of Services of Twitter, some datasets are not released there.

All tweets are collected through Twitter Streaming APIs. We have been using two different streams to retrieve tweets from Twitter Streaming APIs. The first one is tracking keywords, we have been tracking bullying related keyword list "ignored, pushed, rumors, locker, spread, shoved, rumor, teased, kicked, crying, bullied, bully, bullyed, bullying, bullyer, bulling" since Aug 3, 2011. Since they are not commonly used words in general tweets, we expect to collect almost all the tweets containing these keywords since then.

The second method is to follow users. We identified the top 5000 users, who posted most bullying traces collected by our algorithm described in Section 2.1 between January 1 and March 31, 2013. We have been following these 5000 seed users since April 24, 2013. We are able to collect all of (a) tweets created or retweeted by these seed users, (b) replies to any tweet created by these seed users, and (c) retweets of any tweet created by these seed users.

Most datasets listed below are the subsets of the tweets we collected above. They might be focused on a special time windows, or be annotated for different tasks.

A.1 Bullying Traces Data Set

The tweets were collected from keyword tracking stream. We only kept the tweets with at least one token starting with "bull." We further removed re-tweets by excluding tweets containing the acronym "RT."

We randomly sampled 1762 tweets collected on August 6, 2011 and our annotators labeled each tweets with the following informations. They first annotated if it is a bullying trace. If it is, they also annotated the type of bullying traces, the

form of bullying episode, if it is teasing, the author's and person mentions' roles and emotions. We refer to this dataset as version one of Bullying Traces Data Set. There is also a version two which has exactly the same tweets and labels, but the tweets are identified by Twitter ID so researchers can download the actual tweets from Twitter.

To futhure improve the performance of our text classifiers and obtain reliable results, our annotators labeled more tweets. The tweets were filtered with the same procedure. 7321 posts (including the 1762 posts in previous section) were randomly sampled from the tweets collected from dates August 6, 2011 through August 31, 2011. They annotated the tweets in the same way, except that they did not annotate the person mentions' roles. We refer this dataset as version three of Bullying Traces Data Set.

A.2 Bullying Traces in Two Academic Years

The tweets were collected by tracking bullying keywords during September 1, 2011 to August 31, 2013. Then, we filtered with the bull* keywords and removed the tweets containing "RT". 32,477,558 tweets are contained in this data set. We applied our Binary Bullying Trace Classifier to recognize bullying traces. This data set was mainly used in our paper (Bellmore et al., 2015). We don't have annotation on this dataset, but we do have applied multiple classifiers to study the distributions of author's role, the form of bullyings, and the types of bullying traces. The result is also reported in Chapter 2 and Chapter 3.

A.3 Topics in Bullying Traces

The tweets were collected by tracking bullying keywords during August 21, 2011 to September 17, 2011. Then, we filtered with the bull* keywords and removed the tweets containing "RT". We applied our Binary Bullying Trace Classifier to recognize bullying traces. In total, we have 188,908 bullying traces in this dataset.

This data set was mainly used in our paper (Xu et al., 2012b) to study the topics in bullying traces. The result is also reported in Chapter 2.

A.4 Bullying Trace Emotion

This data set contains all the tweets we collected by keyword tracking during August 5, 2011 to April 12, 2012 (about eight months). We filtered with the bull* keywords, removed the tweets containing "RT", and applied our Binary Bullying Trace Classifier to recognize bullying traces. In total, we have 3,001,427 bullying traces in this dataset. This data set was mainly used in our paper (Xu et al., 2012c) to study the emotion distributions in bullying traces. There is no annotation. We applied our Bullying Trace Emotion Classifier to study the emotion distribution. The result is also reported in Chapter 4.

A.5 Bullying Trace Regret

The dataset used in section 4.4 for studying regret in bullying traces. The tweets were collected by tracking bullying keywords from July 31 through October 31, 2012. We applied our Binary Bullying Trace Classifier to recognize bullying traces. Then we regularly check if they have been deleted after they were posted. Please refer Chapter 4.4 for the details on how to collect the data. There is no annotation. But for each tweet, we have multiple tags if it was deleted at different check points.

A.6 Hashtags in Bullying Traces Data Set

The dataset used in Chapter 5 for studying hashtags in bullying traces. The tweets were collected by tracking bullying keywords during January 1, 2012 to December 31, 2012. We further filtered the tweets with "bull*" keywords. We kept the retweets and did not filtered with Binary Bullying Trace Classifier.

We extracted all hashtags and ranked them by their occurrences. Our annotators assigned each hashtag into one of the eight pre-defined categories. For these 500 hashtags, we also extracted many features from the tweets containing each hashtag. See Chapter 5 for details.

A.7 Bilingual Bullying Traces Data Set

The dataset is used in Chapter 6 for studying culture differences in bullying traces. We collected English tweets using the public Twitter Streaming API by tracking bullying related keywords: "bully," "bullied," and "bullying". As our focus is on school bullying posts, we only kept the tweets which further contain at least one of the school-related words: "college," "university," "school," and "class." The filtering is case-insensitive and we included the plural forms of these keywords. We removed retweets by filtering tweets with the token "RT."

We collected Chinese weibos through the keyword search function provided by Weibo.com. Since there is no single term in Chinese that exactly corresponds to the English word bullying, we considered all seven near synonyms suggested in (Smith et al., 2002): 凌辱, 欺负, 欺凌, 欺辱, 欺侮, 欺压, 侮辱. We chose three corresponding school keywords: 学, 校, 班. We required at least one match from each keyword list, with the option "original post only" to exclude posts reposting other weibos.

We collected data in this way for the whole year of 2012. In total, there are 756,449 tweets and 75,044 weibos in our dataset. Our annotator labeled 3123 tweets and 3123 weibos collected during October 11-24, 2012. Among them, 1121 (36%) tweets and 811 (26%) weibos were coded as bullying traces. For all bullying traces, our annotators also label the author's role, teasing.

A.8 Bullying Timeline Data Set

Starting from May 2013, we have been following 5000 users who posted most bullying traces in January - March, 2013. We manually selected a few timelines which tend to contain more name calling episode. Our annotator labeled 9 timelines

collected during September 2013. For each tweet in timeline, they label if it is a name calling tweet, and if it is, assign it to an episode. The dataset was used in Chapter 7.

B CODE REPOSITORY

This chapter lists all the softwares we developed to collect and analyze bullying traces. We will make them publicly available at http://research.cs.wisc.edu/bullying. We believe these will help social scientists to process social media and other researchers for comparison.

B.1 Binary Bullying Trace Classifier

The best classifier introduced in Section 2.1. It uses the unigram + bigram feature representation and support vector machines. We have two versions of the classifier, which were trained on two version of Bullying Traces Data Set (Appendix A.1 first and third versions). The first version achieves 81.3% cross validation accuracy and the second version was improved to 86%. More information about the construction and performance of the classifier were discussed in Chapter 2.1.

B.2 Author's Role Classifier

It uses the unigram + bigram feature representation and support vector machines. We have two versions of the classifier, which were trained on two version of Bullying Traces Data Set (Appendix A.1, first and third versions). The first version achieves cross validation accuracy of 61% on five categories. The second version was improved to 70% on six categories. We have different number of categories in different versions, because with fewer training examples, there were few examples for Defender category. With larger dataset, we were able to reliably identify them. More information about the construction and performance of the classifier were discussed in Chapter 2.2.

B.3 Bullying Form Classifier

It uses the unigram + bigram feature representation and support vector machines. It was trained on Bullying Traces Data Set (Version 2, Appendix A.1). It achieves cross validation accuracy of 70%. More information about the construction and performance of the classifier were discussed in Chapter 2.3.

B.4 Socioscope

The code package includes the implementation of Socioscope and example datasets. It was implemented in Matlab, with the functionality of tuning regularization parameters by cross validation. Besides the implementation, it also includes an example file of how to use the code with an example datasets.

B.5 Bullying Trace Type Classifier

It uses the unigram + bigram feature representation and support vector machines. It was trained on Bullying Traces Data Set (Version 2, Appendix A.1). It achieves cross validation accuracy of 72%. More information about the construction and performance of the classifier were discussed in Chapter 2.4.

B.6 Teasing Bullying Trace Classifier

It uses the unigram + bigram feature representation and support vector machines. It was trained on Bullying Traces Data Set (Version 2, Appendix A.1). It achieves cross validation accuracy of 89%. More information about the construction and performance of the classifier were discussed in Chapter 4.1.

B.7 Bullying Trace Emotion Classifier

Our classifier was trained with minimum supervision for sentiment analysis in Section 4.2. It takes the raw texts of tweet as input. Necessary code of tokenization and stopword removal are also included in the package. After preprocessing, we convert each tweet into a 35 dimension similarity vectors to the feature extractors. The model was build with libsvm and the training data from Wikipedia pages.

B.8 Bullying Trace Regret Classifier

Our classifier was trained to predict deletion in Section 4.4. The model is trained with the cleaned dataset, in which each tweet is known to be surviving or deleted after 20,480 minutes (about two weeks). Since this dataset contains 22,241 deleted tweets, we randomly sub-sampled the surviving tweets down to 22,241 to force our deleted and surviving datasets to be of equal size. We then followed the pre-processing procedure in Section 2.1, performing case-folding, anonymization, and tokenization, treating URLs, emoticons and hashtags specially. We also chose the unigrams+bigrams feature representation, only keeping tokens appearing at least 15 times in the corpus. We chose to employ a linear SVM implemented in LIBLINEAR (Fan et al., 2008) due to its efficiency on this large sparse text categorization task and a 10-fold cross validation was conducted to evaluate its performance. The resulting cross validation accuracy was 0.607 with a standard deviation of 0.012.

REFERENCES

Allan, James. 2002. *Topic Detection and Tracking: Event-Based Information Organization*. Norwell, MA: Kluwer Academic Publishers.

American Psychological Association. 2004. APA resolution on bullying among children and youth. `http://www.apa.org/about/governance/council/policy/bullying.pdf`.

Andrzejewski, David, Xiaojin Zhu, Mark Craven, and Benjamin Recht. 2011. A framework for incorporating general domain knowledge into latent Dirichlet allocation using first-order logic. In *Proceedings of the 22nd international joint conference on artificial intelligence*, vol. 22, 1171.

Archer, John, and Sarah M. Coyne. 2005. An integrated review of indirect, relational, and social aggression. *Personality and Social Psychology Review* 9:212–230.

Baly, Michael W, Dewey G Cornell, and Peter Lovegrove. 2014. A longitudinal investigation of self-and peer reports of bullying victimization across middle school. *Psychology in the Schools* 51(3):217–240.

Bamman, David, Brendan O'Connor, and Noah Smith. 2012. Censorship and deletion practices in Chinese social media. *First Monday* 17(3-5).

Becker, Hila, Naaman Mor, and Luis Gravano. 2011. Beyond trending topics: Real-world event identification on Twitter. In *Proceedings of the 15th international AAAI conference on weblogs and social media*, 438–441.

Bellmore, Amy, Angela J Calvin, Jun-Ming Xu, and Xiaojin Zhu. 2015. The five w's of "bullying" on Twitter: Who, what, why, where, and when. *Computers in Human Behavior* 44:305–314.

Bellmore, Amy D., Melissa R. Witkow, Sandra Graham, and Jaana Juvonen. 2004. Beyond the individual: The impact of ethnic context and classroom behavioral norms on victims' adjustment. *Developmental Psychology* 40:1159–1172.

Biggs, Bridget K., Jennifer Mize Nelson, and Marilyn L. Sampilo. 2010. Peer relations in the anxiety-depression link: Test of a mediation model. *Anxiety, Stress & Coping* 23(4):431–447.

Blei, David M, and Peter I Frazier. 2011. Distance dependent Chinese restaurant processes. *Journal of Machine Learning Research* 12:2461–2488.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Blitzer, John. 2008. Domain adaptation of natural language processing systems. Ph.D. thesis, University of Pennsylvania.

Bosse, Tibor, and Sven Stam. 2011. A normative agent system to prevent cyberbullying. In *Proceedings of the 2011 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology (WI-IAT)*, vol. 2, 425–430. IEEE.

boyd, Danah, Scott Golder, and Gilad Lotan. 2010. Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter. In *Proceedings of the 43rd Hawaii international conference on system sciences (HICSS)*, 1–10. IEEE.

Brody, Samuel, and Nicholas Diakopoulos. 2011. Cooooooooooooooollllllllllll-lll!!!!!!!!!!!!!!: using word lengthening to detect sentiment in microblogs. In *Proceedings of the conference on empirical methods in natural language processing*, 562–570. Association for Computational Linguistics.

Calvin, Angela J, Amy Bellmore, Jun-Ming Xu, and Xiaojin Zhu. 2015. #bully: Uses of hashtags in posts about bullying on Twitter. *Journal of School Violence* 14(1): 133–153.

Cambria, Erik, Praphul Chandra, Avinash Sharma, and Amir Hussain. 2010. Do not feel the trolls. In *Proceedings of the 3rd international workshop on social data on the web*.

Campos, Belinda, Dacher Keltner, Jennifer M Beck, Gian C Gonzaga, and Oliver P John. 2007. Culture and teasing: The relational benefits of reduced desire for positive self-differentiation. *Personality and Social Psychology Bulletin* 33(1):3–16.

Card, Noel A, and Ernest VE Hodges. 2008. Peer victimization among schoolchildren: Correlations, causes, consequences, and considerations in assessment and intervention. *School Psychology Quarterly* 23(4):451–461.

Cardoso, Paula C.F., Maite Taboada, and Thiago A. S. Pardo. 2013. On the contribution of discourse structure to topic segmentation. In *Proceedings of the special interest group on discourse and dialogue (SIGDIAL)*, 92–96.

Cassidy, Wanda, Margaret Jackson, and Karen N. Brown. 2009. Sticks and stones can break my bones, but how can pixels hurt me? Students' experiences with cyber-bullying. *School Psychology International* 30(4):383–402.

Cataldi, Mario, Luigi Di Caro, and Claudio Schifanella. 2010. Emerging topic detection on Twitter based on temporal and social terms evaluation. In *Proceedings of the 10th international workshop on multimedia data mining*, 4:1–4:10.

Chang, Chih-Chung, and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2:27:1–27:27.

Chen, Gina Masullo. 2011. Tweet this: A uses and gratifications perspective on how active Twitter use gratifies a need to connect with others. *Computers in Human Behavior* 27(2):755–762.

Chenthamarakshan, Vijil, Prem Melville, Vikas Sindhwani, and Richard D Lawrence. 2011. Concept labeling: Building text classifiers with minimal supervision. In *Proceedings of the 22nd international joint conference on artificial intelligence*, 1225–1230.

Child, Jeffrey T., Paul M. Haridakis, and Sandra Petronio. 2012. Blogging privacy rule orientations, privacy management, and content deletion practices: The vari-

ability of online privacy management activity at different stages of social media use. *Computers in Human Behavior* 28(5):1859 – 1872.

Child, Jeffrey T, Sandra Petronio, Esther A Agyeman-Budu, and David A Westermann. 2011. Blog scrubbing: Exploring triggers that change privacy rules. *Computers in Human Behavior* 27(5):2017–2027.

Christofides, Emily, Amy Muise, and Serge Desmarais. 2009. Information disclosure and control on Facebook: Are they two sides of the same coin or two different processes? *CyberPsychology & Behavior* 12(3):341–345.

Chung, Fan R. K. 1997. *Spectral Graph Theory*. Regional Conference Series in Mathematics, Providence, RI: American Mathematical Society.

Cohen, Jacob. 1998. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Erlbaum.

Cook, Clayton R., Kirk R. Williams, Nancy G. Guerra, Tia E. Kim, and Shelly Sadek. 2010. Predictors of bullying and victimization in childhood and adolescence: A meta-analytic investigation. *School Psychology Quarterly* 25(2):65–83.

Cornec, Matthieu. 2010. Concentration inequalities of the cross-validation estimate for stable predictors. *Arxiv preprint arXiv:1011.5133*.

Dinakar, Karthik, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. 2012. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TIIS)* 2(3):18.

Dinakar, Karthik, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of textual cyberbullying. In *International conference on weblog and social media - social mobile web workshop*. Barcelona, Spain.

Donoho, David L, Iain M Johnstone, Gérard Kerkyacharian, and Dominique Picard. 1996. Density estimation by wavelet thresholding. *The Annals of Statistics* 24:508–539.

Duggan, Maeve, and Aaron Smith. 2013. Social media update 2013: 42% of online adults use multiple networking sites, but Facebook remains the platform of choice. *Pewinternet. org*.

Earle, Paul, Michelle Guy, Richard Buckmaster, Chris Ostrum, Scott Horvath, and Amy Vaughan. 2010. OMG earthquake! Can Twitter improve earthquake response? *Seismological Research Letters* 81(2):246–251.

Eisenstein, Jacob, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, 1277–1287.

Ellison, Nicole B, Charles Steinfield, and Cliff Lampe. 2007. The benefits of Facebook "friends:" Social capital and college students' use of online social network sites. *Journal of Computer-Mediated Communication* 12(4):1143–1168.

Fan, Rong-En, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research* 9:1871–1874.

Fekkes, Minne, Frans I.M. Pijpers, A. Miranda Fredriks, Ton Vogels, and S. Pauline Verloove-Vanhorick. 2006. Do bullied children get ill, or do ill children get bullied? A prospective cohort study on the relationship between bullying and health-related symptoms. *Pediatrics* 117:1568–1574.

Fellbaum, Christiane. 1988. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Finkel, Jenny Rose, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, 363–370. Association for Computational Linguistics.

Fredstrom, Bridget K., Ryan E. Adams, and Rich Gilman. 2011. Electronic and school-based victimization: Unique contexts for adjustment difficulties during adolescence. *Journal of Youth and Adolescence* 40(4):405–415.

Fuligni, Andrew J, and Harold W Stevenson. 1995. Time use and mathematics achievement among American, Chinese, and Japanese high school students. *Child Development* 66(3):830–842.

Gildea, Daniel, and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics* 28(3):245–288.

Gini, Gianluca, and Tiziana Pozzoli. 2009. Association between bullying and psychosomatic problems: A meta-analysis. *Pediatrics* 123(3):1059–1065.

———. 2013. Bullied children and psychosomatic problems: A meta-analysis. *Pediatrics* 132(4):720–729.

Gleason, Benjamin. 2013. #Occupy Wall Street: Exploring informal learning about a social movement on Twitter. *American Behavioral Scientist* 57:966–982.

Gonzales, Amy L, and Jeffrey T Hancock. 2011. Mirror, mirror on my Facebook wall: Effects of exposure to Facebook on self-esteem. *Cyberpsychology, Behavior, and Social Networking* 14(1-2):79–83.

Graham, Sandra, Amy Bellmore, and Jaana Juvonen. 2007. Peer victimization in middle school: When self- and peer views diverge. In *Bullying, Victimization, and Peer Harassment: A Handbook of Prevention and Intervention*, ed. Joseph E. Zins, Maurice J. Elias, and Charles A. Maher, 121–141. New York, NY: Haworth Press.

Griffiths, Thomas L, and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences* 101(suppl 1):5228–5235.

Gupte, Mangesh, Pravin Shankar, Jing Li, Shanmugauelayut Muthukrishnan, and Liviu Iftode. 2011. Finding hierarchy in directed online social networks. In *Proceedings of the 20th international conference on world wide web*, 557–566.

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter* 11:10–18.

Hatzivassiloglou, Vasileios, and Kathleen R McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the European chapter of the association for computational linguistics*, 174–181. Association for Computational Linguistics.

Hawker, David S. J., and Michael J. Boulton. 2000. Twenty years' research on peer victimization and psychosocial maladjustment: A meta-analytic review of cross-sectional studies. *Journal of Child Psychology And Psychiatry* 41(4):441–455.

Hu, Minqing, and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD international conference on knowledge discovery and data mining*, 168–177. ACM.

Huang, Jeff, Katherine M Thornton, and Efthimis N Efthimiadis. 2010. Conversational tagging in Twitter. In *Proceedings of the 21st ACM conference on hypertext and hypermedia*, 173–178. ACM.

Janosz, Michel, Isabelle Archambault, Linda S. Pagani, Sophie Pascal, Alexandre J.S. Morin, and François Bowen. 2008. Are there detrimental effects of witnessing school violence in early adolescence? *Journal of Adolescent Health* 43(6): 600–608.

Java, Akshay, Xiaodan Song, Tim Finin, and Belle Tseng. 2007. Why we Twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on web mining and social network analysis*, 56–65. WebKDD/SNA-KDD '07, New York, NY, USA: ACM.

Jimerson, Shane R., Susan M. Swearer, and Dorothy L. Espelage. 2010. *Handbook of Bullying in Schools: An International Perspective*. New York, NY: Routledge/Taylor & Francis Group.

Juvonen, Jaana, and Sandra Graham. 2001. *Peer Harassment in School: The Plight of the Vulnerable and Victimized*. New York, NY: Guilford Press.

———. 2014. Bullying in schools: The power of bullies and the plight of victims. *Annual Review of Psychology* 65:159–185.

Juvonen, Jaana, and Elisheva F. Gross. 2008. Extending the school grounds? – Bullying experiences in cyberspace. *Journal of School Health* 78:496–505.

Kağitçibaşi, Çiğdem. 2007. *Family, Self, and Human Development across Cultures: Theories and Applications*. 2nd ed. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Kanayama, Hiroshi, and Tetsuya Nasukawa. 2006. Fully automatic lexicon expansion for domain-oriented sentiment analysis. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, 355–363. Association for Computational Linguistics.

Kanetsuna, Tomoyuki, Peter K Smith, and Yohji Morita. 2006. Coping with bullying at school: Children's recommended strategies and attitudes to school-based interventions in England and Japan. *Aggressive Behavior* 32(6):570–580.

Katrandjian, Olivia. 2011. Jamey Rodemeyer suicide: Lady Gaga pays tribute to bullying victim. *ABC News* 25.

Keltner, Dacher, Lisa Capps, Ann M Kring, Randall C Young, and Erin A Heerey. 2001. Just teasing: A conceptual analysis and empirical review. *Psychological Bulletin* 127(2):229.

Kirkland, Pamela. 2014. Can Twitter activism #bringbackourgirls? The Washington Post. Retrieved from www.washingtonpost.com/blogs/she-the-people/wp/2014/07/23/can-twitter-activism-bringbackourgirls/.

Kontostathis, April, Lynne Edwards, and Amanda Leatherman. 2010. Text mining and cybercrime. In *Text Mining: Applications and Theory*, ed. Michael W. Berry and Jacob Kogan. Chichester, UK: John Wiley & Sons, Ltd.

Kowalski, Robin M, Gary W Giumetti, Amber N Schroeder, and Micah R Lattanner. 2014. Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth. *Psychological Bulletin* 140:1073–1137.

Ladd, Gary W., Becky J. Kochenderfer, and Cynthia C. Coleman. 1997. Classroom peer acceptance, friendship, and victimization: Distinct relational systems that contribute uniquely to children's school adjustment? *Child Development* 68:1181–1197.

Lam, Amy G, and Nolan WS Zane. 2004. Ethnic differences in coping with interpersonal stressors: A test of self-construals as cultural mediators. *Journal of Cross-Cultural Psychology* 35(4):446–459.

Latham, Annabel, Keeley Crockett, and Zuhair Bandar. 2010. A conversational expert system supporting bullying and harassment policies. In *Proceedings of the 2nd international conference on agents and artificial intelligence*, 163–168.

Lazer, David, Alex (Sandy) Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. 2009. Life in the network: The coming age of computational social science. *Science* 323(5915):721–723.

Lehmann, Janette, Bruno Gonçalves, José J Ramasco, and Ciro Cattuto. 2012. Dynamical classes of collective attention in Twitter. In *Proceedings of the 21st international conference on world wide web*, 251–260. ACM.

Lenhart, Amanda, Kristen Purcell, Aaron Smith, and Kathryn Zickuhr. 2010. Social media & mobile internet use among teens and young adults. *Pew Internet & American Life Project*.

Lieberman, Henry, Karthik Dinakar, and Birago Jones. 2011. Let's gang up on cyberbullying. *Computer* 44:93–96.

Little, Todd D., Christopher C. Henrich, Stephanie M. Jones, and Patricia H. Hawley. 2003. Disentangling the "whys" from the "whats" of aggressive behavior. *International Journal of Behavioral Development* 27:122–133.

Liu, Bing, and Lei Zhang. 2012. A survey of opinion mining and sentiment analysis. In *Mining Text Data*, ed. Charu C. Aggarwal and ChengXiang Zhai, 415–463. Springer US.

Liu, Fei, Fuliang Weng, and Xiao Jiang. 2012. A broad-coverage normalization system for social media language. In *Proceedings of the 50th annual meeting of the association for computational linguistics: Long papers-volume 1*, 1035–1044. Association for Computational Linguistics.

Liu, Xiaohua, Kuan Li, Bo Han, Ming Zhou, Long Jiang, Zhongyang Xiong, and Changning Huang. 2010. Semantic role labeling for news tweets. In *Proceedings of the 23rd international conference on computational linguistics*, 698–706. Association for Computational Linguistics.

Lovejoy, Kristen, and Gregory D Saxton. 2012. Information, community, and action: How nonprofit organizations use social media. *Journal of Computer-Mediated Communication* 17(3):337–353.

Macbeth, Jamie, Hanna Adeyema, Henry Lieberman, and Christopher Fry. 2013. Script-based story matching for cyberbullying prevention. In *CHI'13 extended abstracts on human factors in computing systems*, 901–906. ACM.

Màrquez, Lluís, Pere Comas, Jesús Giménez, and Neus Catala. 2005. Semantic role labeling as sequential tagging. In *Proceedings of the 9th conference on computational natural language learning*, 193–196. Association for Computational Linguistics.

McTernan, Wesley P, Maureen F Dollard, and Anthony D LaMontagne. 2013. Depression in the workplace: An economic cost analysis of depression-related productivity loss attributable to job strain and bullying. *Work & Stress* 27(4):321–338.

Mei, Qiaozhu, Chao Liu, Hang Su, and ChengXiang Zhai. 2006. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *Proceedings of the 15th international conference on world wide web*, 533–542.

Mihalcea, Rada, and Carlo Strapparava. 2005. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, 531–538. Association for Computational Linguistics.

Miles, Matthew B., and A. Michael Huberman. 1994. *Qualitative Data Analysis: An Expanded Sourcebook*. Sage.

Miller, George A. 1995. Wordnet: a lexical database for english. *Communications of the ACM* 38(11):39–41.

Mishna, Faye, and Ramona Alaggia. 2005. Weighing the risks: A child's decision to disclose peer victimization. *Children & Schools* 27(4):217–226.

Mitchell, Amy, and Emily Guskin. 2013. Twitter news consumers: Young, mobile and educated. *Pew Research Journalism Project, Nov* 4.

Møller, Jesper, and Rasmus Plenge Waagepetersen. 2004. *Statistical Inference and Simulation for Spatial Point Processes*. Monographs on statistics and applied probability, Boca Raton, FL: Chapman & Hall/CRC.

Moore, Mark H., Carol V. Petrie, Anthony A. Braga, and Brenda L. McLaughlin. 2003. *Deadly Lessons: Understanding Lethal School Violence*. Washington, DC: The National Academies Press.

Morstatter, Fred, Jürgen Pfeffer, Huan Liu, and Kathleen M Carley. 2013. Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose. In *Seventh international AAAI conference on weblogs and social media*.

Nansel, Tonja R., Mary Overpeck, Ramani S. Pilla, W. June Ruan, Bruce Simons-Morton, and Peter Scheidt. 2001. Bullying behaviors among US youth: Prevalence

and association with psychosocial adjustment. *Journal of American Medical Association* 285(16):2094–2100.

Nishina, Adrienne, and Amy D. Bellmore. 2010. When might aggression, victimization, and conflict matter most?: Contextual considerations. *Journal of Early Adolescence* 5–26.

Nishina, Adrienne, and Jaana Juvonen. 2005. Daily reports of witnessing and experiencing peer harassment in middle school. *Child Development* 76:435–450.

Nocedal, Jorge, and Stephen J Wright. 1999. *Numerical Optimization*. Springer series in operations research, New York, NY: Springer.

Nylund, Karen, Amy Bellmore, Adrienne Nishina, and Sandra Graham. 2007. Subtypes, severity, and structural stability of peer victimization: What does latent class analysis say? *Child Development* 78:1706–1722.

Oh, Hyun Jung, Elif Ozkaya, and Robert LaRose. 2014. How does online social networking enhance life satisfaction? The relationships among online supportive interaction, affect, perceived social support, sense of community, and life satisfaction. *Computers in Human Behavior* 30:69–78.

Olweus, Dan. 1993. *Bullying at School: What We Know and What We Can Do*. Oxford, UK: Blackwell.

Pak, Alexander, and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. *Computer* 1320–1326.

Pang, Bo, and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on association for computational linguistics*, 271–278. Association for Computational Linguistics.

———. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* 2(1-2):1–135.

Papacharissi, Zizi, and Maria de Fatima Oliveira. 2012. Affective news and networked publics: The rhythms of news storytelling on #egypt. *Journal of Communication* 62(2):266–282.

Park, Jaram, Vladimir Barash, Clay Fink, and Meeyoung Cha. 2013. Emoticon style: Interpreting differences in emoticons across cultures. In *Proceedings of the 7th international AAAI conference on weblogs and social data*, 466–475. Boston, MA.

Patterson, Bruce D., Gerardo Ceballos, Wes Sechrest, Marcelo F. Tognelli, Thomas Brooks, Lucía Luna, Pablo Ortega, Irma Salazar, and Bruce E. Young. 2007. Digital distribution maps of the mammals of the western hemisphere, version 3.0. Tech. Rep., NatureServe, Arlington, VA.

Peng, Fuchun, Fangfang Feng, and Andrew McCallum. 2004. Chinese segmentation and new word detection using conditional random fields. In *Proceedings of the 20th international conference on computational linguistics*, 562. Association for Computational Linguistics.

Ptaszynski, Michal, Pawel Dybala, Tatsuaki Matsuba, Fumito Masui, Rafal Rzepka, and Kenji Araki. 2010. Machine learning and affect analysis against cyber-bullying. In *Proceedings of the 36th annual convention of the society for the study of artificial intelligence and the simulation of behaviour (AISB)*, 7–16.

Punyakanok, Vasin, Dan Roth, and Wen-tau Yih. 2008. The importance of syntactic parsing and inference in semantic role labeling. *Computational Linguistics* 34(2): 257–287.

Ratinov, Lev, and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the 13th conference on computational natural language learning*, 147–155. Association for Computational Linguistics.

Ritter, Alan, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Proceedings of the conference on empirical methods in natural language processing*, 1524–1534. Association for Computational Linguistics.

Rivers, Ian, V. Paul Poteat, Nathalie Noret, and Nigel Ashurst. 2009. Observing bullying at school: The mental health implications of witness status. *School Psychology Quarterly* 24(4):211–223.

Sakaki, Takeshi, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on world wide web*, 851–860. WWW '10.

Salmivalli, Christina. 1999. Participant role approach to school bullying: Implications for intervention. *Journal of Adolescence* 22(4):453–459.

Salmivalli, Christina, Kirsti Lagerspetz, Kaj Björkqvist, Karin Österman, and Ari Kaukiainen. 1996. Bullying as a group process: Participant roles and their relations to social status within the group. *Aggressive Behavior* 22(1):1–15.

Salton, Gerard. 1971. *The SMART retrieval system–experiments in automatic document processing*. Englewood Cliffs, NJ: Prentice-Hall.

Schmidt, Mark W., Glenn Fung, and Rómer Rosales. 2007. Fast optimization methods for l1 regularization: A comparative study and two new approaches. In *Proceedings of the 18th European conference on machine learning*, 286–297. Springer.

Schwartz, David, Lei Chang, and JoAnn M Farver. 2001. Correlates of victimization in Chinese children's peer groups. *Developmental Psychology* 37(4):520–532.

Schwartz, David, Andrea Hopmeyer Gorman, Jonathan Nakamoto, and Robin L. Toblin. 2005. Victimization in the peer group and children's academic functioning. *Journal of Educational Psychology* 87:425–435.

Settles, Burr. 2011. Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. In *Proceedings of the conference on empirical methods in natural language processing*, 1467–1478. Association for Computational Linguistics.

Shelley, Danielle, and Wendy M Craig. 2010. Attributions and coping styles in reducing victimization. *Canadian Journal of School Psychology* 25(1):84–100.

Shi, Xingsong. 2011. The impact of face on Chinese students' simulated negotiation practices with Americans. *Language and Intercultural Communication* 11(1):26–40.

Smith, Peter K, Helen Cowie, Ragnar F Olafsson, and Andy PD Liefooghe. 2002. Definitions of bullying: A comparison of terms used, and age and gender differences, in a fourteen–country international comparison. *Child Development* 73(4): 1119–1133.

Smith, Peter K., Kirsten C. Madsen, and Janet C. Moody. 1999. What causes the age decline in reports of being bullied at school? Towards a developmental analysis of risks of being bullied. *Educational Research* 41(3):267–285.

Strapparava, Carlo, and Alessandro Valitutti. 2004. Wordnet-affect: an affective extension of WordNet. In *Proceedings of the 4th international conference on language and resources and evaluation (LREC)*, vol. 4, 1083–1086.

Tausczik, Yla R, and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology* 29(1):24–54.

The American Academy of Pediatrics. 2009. Policy statement–role of the pediatrician in youth violence prevention. *Pediatrics* 124(1):393–402.

The White House. 2011. Background on White House conference on bullying prevention. Http://www.whitehouse.gov/the-press-office/2011/03/10/background-white-house-conference-bullying-prevention.

Thelwall, Mike, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology* 61(12):2544–2558.

Toutanova, Kristina, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 conference of the north American chapter of the association for com-*

putational linguistics on human language technology-volume 1, 173–180. Association for Computational Linguistics.

Triandis, Harry Charalambos. 1995. *Individualism & Collectivism*. Boulder, CO: Westview Press.

Twitter. 2012. The Twitter rules. `http://support.twitter.com/articles/18311-the-twitter-rules`.

Vaillancourt, Tracy, Vi Trinh, Patricia McDougall, Eric Duku, Lesley Cunningham, Charles Cunningham, Shelley Hymel, and Kathy Short. 2010. Optimizing population screening of bullying in school-aged children. *Journal of School Violence* 9: 233–250.

Valkenburg, Patti M, and Jochen Peter. 2009. Social consequences of the internet for adolescents: A decade of research. *Current Directions in Psychological Science* 18(1):1–5.

Van Der Laan, Mark J., and Sandrine Dudoit. 2003. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. *U.C. Berkeley Division of Biostatistics Working Paper Series* 130–236.

Vandebosch, Heidi, and Katrien Van Cleemput. 2009. Cyberbullying among youngsters: Profiles of bullies and victims. *New Media & Society* 11(8):1349–1371.

Vardi, Y., L. A. Shepp, and L. Kaufman. 1985. A statistical model for positron emission tomography. *Journal of the American Statistical Association* 80(389):8–37.

Walther, Joseph B. 1996. Computer-mediated communication impersonal, interpersonal, and hyperpersonal interaction. *Communication Research* 23(1):3–43.

Wang, Jing, Ronald J. Iannotti, and Tonja R. Nansel. 2009. School bullying among adolescents in the United States: Physical, verbal, relational, and cyber. *Journal of Adolescent Health* 45(4):368–375.

Wang, Yang, Gregory Norcie, Saranga Komanduri, Alessandro Acquisti, Pedro Giovanni Leon, and Lorrie Faith Cranor. 2011. "I regretted the minute I pressed share": A qualitative study of regrets on Facebook. In *Proceedings of the seventh symposium on usable privacy and security*, 10:1–10:16. SOUPS '11, ACM.

Wei, Hsi-sheng, Melissa Jonson-Reid, and Hui-ling Tsao. 2007. Bullying and victimization among Taiwanese 7th graders: A multi-method assessment. *School Psychology International* 28(4):479–500.

Willett, Rebecca M, and Robert D Nowak. 2007. Multiscale Poisson intensity and density estimation. *IEEE Transactions on Information Theory* 53(9):3171–3187.

Xu, Jun-Ming, Aniruddha Bhargava, Robert Nowak, and Xiaojin Zhu. 2012a. Socioscope: Spatio-temporal signal recovery from social media. In *European conference on machine learning and principles and practice of knowledge discovery in databases*, 644–659. Bristol, UK: Springer.

———. 2013a. Socioscope: Spatio-temporal signal recovery from social media (extended abstract). In *Proceedings of the 23rd international joint conference on artificial intelligence (IJCAI)*, 3096–3100. Beijing, China.

Xu, Jun-Ming, Benjamin Burchfiel, Xiaojin Zhu, and Amy Bellmore. 2013b. An examination of regret in bullying tweets. In *Proceedings of the 2013 conference of the north American chapter of the association for computational linguistics: Human language technologies (NAACL-HLT)*, 697–702. Atlanta, GA: Association for Computational Linguistics.

Xu, Jun-Ming, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. 2012b. Learning from bullying traces in social media. In *Proceedings of the 2012 conference of the north American chapter of the association for computational linguistics: Human language technologies (NAACL-HLT)*, 656–666. Montréal, Canada: Association for Computational Linguistics.

Xu, Jun-Ming, Xiaojin Zhu, and Amy Bellmore. 2012c. Fast learning for sentiment analysis on bullying. In *Proceedings of the first international workshop on issues of sentiment discovery and opinion mining*, 10. Beijing, China: ACM.

Yang, Changhua, Kevin Hsin-Yih Lin, and Hsin-Hsi Chen. 2007. Emotion classification using web blog corpora. In *IEEE/WIC/ACM international conference on web intelligence*, 275–278. IEEE.

Yang, Jiang, Meredith Ringel Morris, Jaime Teevan, Lada A Adamic, and Mark S Ackerman. 2011. Culture matters: A survey study of social Q&A behavior. In *Proceedings of the 5th international AAAI conference on weblogs and social data*, 409–416. Barcelona, Spain.

Yang, Yiming, Tom Pierce, and Jaime Carbonell. 1998. A study of retrospective and on-line event detection. In *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval*, 28–36.

Yin, Zhijun, Liangliang Cao, Jiawei Han, Chengxiang Zhai, and Thomas Huang. 2011. Geographical topic discovery and comparison. In *Proceedings of the 20th international conference on world wide web*, 247–256.

Yu, Ming-chung. 2003. On the universality of face: Evidence from Chinese compliment response behavior. *Journal of Pragmatics* 35(10):1679–1710.

Zhu, Xiaojin, Zoubin Ghahramani, and John Lafferty. 2005. Time-sensitive Dirichlet process mixture models. Tech. Rep. CMU-CALD-05-104, School of Computer Sciences, Carnegie Mellon University, Pittsburgh, PA.