Computers in Human Behavior 44 (2015) 305–314

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

The five W's of "bullying" on Twitter: Who, What, Why, Where, and When

Amy Bellmore [a,*], Angela J. Calvin [a], Jun-Ming Xu [b], Xiaojin Zhu [b]

[a] Department of Educational Psychology, University of Wisconsin-Madison, United States
[b] Department of Computer Science, University of Wisconsin-Madison, United States

ARTICLE INFO

Article history:
Available online 12 December 2014

Keywords:
Bullying
Twitter
Social media
Machine learning
Computer science

ABSTRACT

This paper explores the utility of machine learning methods for understanding bullying, a significant social-psychological issue in the United States, through social media data. Machine learning methods were applied to all public mentions of bullying on Twitter between September 1, 2011 and August 31, 2013 to extract the posts that referred to discrete bullying episodes ($N = 9{,}764{,}583$) to address five key questions. Most posts were authored by victims and reporters and referred to general forms of bullying. Posts frequently reflected self-disclosure about personal involvement in bullying. The number of posts that originated from a state was positively associated with the state population size; the timing of the posts reveal that more posts were made on weekdays than on Saturdays and more posts were made during the evening compared to daytime hours. Potential benefits of merging social science and computer science methods to enhance the study of bullying are discussed.

1. Introduction

Bullying is recognized as a serious national health issue within the United States (American Psychological Association, 2004; Committee on Injury, Violence, & and Poison Prevention, 2009; The White House, 2011). This recognition derives from the growing body of research underscoring the wide-ranging harm associated with bullying (Juvonen & Graham, 2014). Bullying is associated with psychological maladjustment (Hawker & Boulton, 2000), physical complaints (Gini & Pozzoli, 2013), and poor functioning in school (Baly, Cornell, & Lovegrove, 2014) and at work (McTernan, Dollard, & LaMontagne, 2013). Given its significance, scholars are charged with identifying factors that influence bullying involvement in a timely fashion. This paper introduces social media as a promising data source for studying bullying and illustrates the applicability of machine learning methods for understanding the large amount of data generated by social media.

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 & Hodges, 2008). Often these surveys are collected only once. 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 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 stands to address each of these limitations, yet it is virtually unexplored. The few exceptions are studies on cyberbullying that target computer science scholars (Bosse & Stam, 2011; Kontostathis, Edwards, & Leatherman, 2010; Latham, Crockett, & Bandar, 2010; Macbeth, Adeyema, Lieberman, & Fry, 2013; Ptaszynski et al., 2010). These works mainly aim at automatically recognizing cyberbullying or hatred speech online, and preventing or reporting them once detected. They do not consider bullying in the physical world or study bullying from a psychological science 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 of youth (Lenhart, Purcell, Smith, & Zickuhr, 2010) and adults (Duggan & Smith, 2013). Social media

* Corresponding author at: University of Wisconsin-Madison, Department of Educational Psychology, 1025 W. Johnson St., Madison, WI 53706, United States.
E-mail address: abellmore@wisc.edu (A. Bellmore).

http://dx.doi.org/10.1016/j.chb.2014.11.052
0747-5632/

enhance relationships (Ellison, Steinfield, & Lampe, 2007) and promote life satisfaction (Oh, Ozkaya, & LaRose, 2014) but are also a context for bullying (Wang, Iannotti, & Nansel, 2009). A meta-analysis of youth cyberbullying research reported that cyberbullying prevalence rates typically range between 10% and 40% of participants (Kowalski, Giumetti, Schroeder, & Lattanner, 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. In this paper, we aim to show that social media is a context for bullying as well as a context for augmenting the study of bullying. Key to this endeavor is that interactions that take place both online and offline might be represented in social media.

Participants of bullying episodes may post text about the experience, leaving valuable, albeit fragmental, traces to be pieced together to understand the episode. Such posts, called "bullying traces," include but far exceed cyberbullying. Most bullying traces are actually online *responses* to traditional bullying or cyberbullying—the actual bullying attack is hidden from view. Bullying traces are difficult to recognize and analyze automatically. They present themselves in diverse ways, from different role perspectives, and usually only contain partial information. Reconstructing episodes from bullying traces can advance our understanding of bullying by providing a real-time, multi-perspective account of discrete experiences about bullying episodes. Shifting the focus to reports about numerous discrete bullying episodes represents an important paradigm shift of the study of bullying that is made possible with the availability of Big Data. As boyd and Crawford (2012) describe, the value of Big Data is not so much in the amount of data that are available but in the quality of the data that allows for making connections between pieces of data, individuals, and individuals in relation to others.

1.1. The present study

In this paper, we introduce the utility of machine learning methods for understanding bullying through social media data. Machine learning, a branch of computer science and statistics, describes a process through which computer systems learn from data (Bishop, 2006; Hastie, Tibshirani, & Friedman, 2009; Wasserman, 2003; Zhu & Goldberg, 2009). It has been used in human-centric applications that rely on large yields of data such as detecting and tracking disease outbreaks, characterizing brain activation patterns yielded from fMRI data, and recognizing speech and handwritten text (Mitchell, 2006).

Supervised learning is one important machine learning task. Given a set of annotated examples, inputs with their desired outputs, a supervised learning algorithm automatically builds a model from input to output and uses it to predict the outputs for previously unseen inputs. We use supervised learning in this study to recognize relevant posts from a general social media stream and automatically classify them into different pre-defined categories. Specifically, we apply machine learning methods to public mentions of bullying on Twitter between September 1, 2011 and August 31, 2013 to address five fundamental questions. We use a school-year rather than a calendar-year timeline because the predominant context for much bullying research is school.

1.1.1. Question 1: *Who is posting about bullying on Twitter?*

Current perspectives emphasize the group-based nature of bullying and the significance of all social role players in a bullying episode (Salmivalli, Lagerspetz, Björkqvist, Österman, & Kaukiainen, 1996). These include the bully(ies), victim(s), bystander(s) (who saw the event but did not intervene), defender(s) of the victim, assistant(s) to the bully (did not initiate but went along with the bully), and reinforcer(s) (did not directly join in with the bully but encouraged the bully by, for example, laughing). Because bullying involves multiple roles, any individual is susceptible to being impacted by bullying, even those who are not directly involved (Rivers, Poteat, Noret, & Ashurst, 2009). In this paper, we seek to identify the role-players who post about bullying episodes on Twitter, as well as their distribution, to describe who posts about bullying versus who participates in bullying.

1.1.2. Question 2: *What form of bullying is mentioned or used on Twitter?*

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) (Wang et al., 2009). Any form may be represented on Twitter because both online and offline interactions 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.

1.1.3. Question 3: *Why are people posting about bullying on Twitter?*

Twitter can serve as a platform for both sharing information (Java, Song, Finin, & Tseng, 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.

1.1.4. Question 4: *Where are people posting about bullying on Twitter?*

Bullying cross cuts culture and geography (Jimerson, Swearer, & Espelage, 2009). As such, there is no expectation about different prevalence rates in posts about bullying episodes across U.S. states after controlling for population size. Nevertheless, 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.

1.1.5. Question 5: *When are people posting about bullying on Twitter?*

Several timing issues related to bullying episodes are important to know. 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 discrete bullying episodes. To fill this gap, we report the distribution of bullying episodes across two school years to specify what day of the week and what time of day posts occur. We expected to see fewer bullying traces on weekends when individuals are away from school and work contexts.

2. Methods

2.1. Data

Our data are derived from the public Twitter streaming Application Programing Interface (API). We started collecting data in 2011, and the data used in this research was collected between the period of September 1, 2011 and August 31, 2013, which covers two consecutive academic years. Twitter was used for this project because of its broad user base and public nature. The public

streams available through the API are used by industrial developers and academic researchers (see https://dev.twitter.com/docs/streaming-apis). To address our questions, various aspects of the data present within the 140 characters in discrete Twitter posts were coded and analyzed.

The public Twitter streaming API may only provide access to a small sample of all relevant data in some instances and thus pose a potential sampling bias (Morstatter, Pfeffer, Liu, & Carley, 2013). We do not believe our data was subject to this. The potential bias arises because the streaming API requires users to specify a list of keywords and only returns the posts containing these keywords. When the percentage of the posts containing the requested keywords over all posts created in the same time interval does not exceed one percent, Twitter provides all the posts containing the keywords. When the percentage exceeds one percent, Twitter only sends a subset of them. Since our keyword list, which will be discussed in Section 2.2.1, consists of only a few words that are not popularly used in the general Twitter stream, we almost always received all of what we requested from Twitter. In other words, we rarely hit the rate limit in our data collection process.

2.2. Procedure

The methods using social media data were primarily derived from the computer science field.

2.2.1. Step 1: Identify bullying tweets

The first step was to identify bullying traces in English tweets among the massive amounts of tweets produced every day. The proportion of bullying traces to all Twitter posts was expected to be small (Xu, Jun, Zhu, & Bellmore, 2012). This posed a challenge for our human coders to find enough bullying traces without labeling an unreasonable amount of tweets. Thus, we restricted ourselves to an "enriched dataset," which is obtained by collecting tweets using the public Twitter streaming API. To capture a post, specific 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 these keywords related to the term bully: bullied, bully, bullyed, bullying, bullyer, bulling. We also followed several other words that were identified 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 of 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 collected and retained even though the keyword bullies was not initially followed. Whereas this post, "He pushed the kid" would be collected but not retained. In lieu of pre-determining every possible version of the word bully to follow on Twitter, this process was used to maximize the versions that might be used. From this dataset, we further removed re-tweets (the analogue 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. As a result, we also worked to ensure that both humans and computers could recognize true bullying traces within the enriched dataset.

A bullying trace was defined as any mention of bullying within the context of a discrete episode. 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 not evident in the short posts. Moreover, we could not determine whether the episode referred to a single event or a continuous episode across a period of time. We relied entirely on the text from the post, taking it at face value when an author participated in or reported "bully"ing. This could include personal experiences "(ugh u bully me a lot") or reports about specific episodes ("She is the one cyberbullying"), including those that were newsworthy ("5 teens had a 14yo hang herself BC they wouldn't stop bullying her"). Posts that were excluded because they were not defined as a bullying episode were those that clearly copied and pasted a news headline about a bullying episode ("Bully @user Staffers Shove CNN Reporter—to Avoid Answering Questions?"), posts that referred to a bullying episode that may happen in the future ("When school starts, I will bully you"), posts that reflected only an opinion about bullying in general ("Bullying is violence against the weak"), instances 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?"), or instances where a coder recognized the names mentioned in the post as fictional ("Harry Potter stood up to that bully").

With these criteria, human coders labeled 7321 tweets randomly selected from dates August 6, 2011 through August 31, 2011. The interrater agreement for identifying bullying traces from the bullying keyword tweets was calculated based on two coders coding 1000 of the 7321 posts. It was determined to be $\kappa = .83$. Of the 7321 posts, 2102, 28.71% were labeled as bullying traces.

We then applied standard machine learning and natural language processing methods that are well-validated within the computer science field (see Zhu & Goldberg, 2009 for a description of machine learning methods and Xu et al., 2012 for a description of the methods used for this data set). We chose a standard text categorization method, which has been used and tested in many real world applications. First, we built a dictionary, which included all the words that appeared in our corpus, and all pairs of any two consecutive words. We represented each tweet with a vector, which is the number of times each word and word pair in dictionary occurred in a current tweet. Next, we built a text classifier with the human-coded tweets. The classifier learned a score vector for each word in the dictionary, which reflected how likely a tweet belongs to one category (e.g., bullying trace or not a bullying trace) if it contains this word. For the tweets that were not human coded, the score which indicated which category in which the tweet belongs was defined by summing over all the scores of words and word pairs it contained weighted by the number of occurrences. The category with largest scores was the predicted one.

When we applied standard machine learning and natural language processing methods (see Xu et al., 2012) to the tweets from 9/1/11 through 8/31/13 we found that 30.07% (9,764,583) of the 32,477,558 tweets that were captured via the keyword filtering were bullying traces. Moreover, with the training set size of 7321, 86% accuracy in assignment relative to the human coding was achieved, which is higher than the majority class (i.e., no) baseline of 71%. This level of accuracy is similar to the level of agreement achieved by two different human coders, and the proportions of bullying traces and non-bullying traces identified across the two approaches were similar. Therefore, we determined that it is possible to use machine learning to automatically and accurately recognize bullying traces in social media. Table 1 provides the confusion matrix for assignment that illustrates agreement and disagreement based on the human coding and machine learning methods.

2.2.2. Step 2: Code key elements of tweets

The second step was to derive data about key features of the bullying episodes from each social media post. Each tweet identified as a bullying trace was evaluated according to five categories: the bullying role of the author of the post (Who), the form of bullying described in the post (What), the type of bullying post (Why), the geographic location of the source of the post (Where),

Table 1
Confusion matrices demonstrating agreement and disagreement with human coding for category assignment using machine learning methods.

Human-coding	Total	Machine learning predicted as						Accuracy
Binary bullying trace								
		Yes	No					86%
Yes	2102	1555	547					
No	5219	503	4716					
Who: The author's role								
		Accuser	Bully	Defender	Reporter	Victim	Other	70%
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	
What: The form of bullying								
		General	Cyber-bullying	Verbal	Physical			91%
General	1857	1831	20	4	2			
Cyber-bullying	145	68	73	4	0			
Verbal	67	53	12	2	0			
Physical	33	32	0	0	1			
Why: The type of post								
		Accusation	Cyber-bullying	Denial	Report	Self-disclosure		72%
Accusation	316	196	0	2	54	64		
Cyber-bullying	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		

and the time of day and week of the post (When). These were determined using a similar general procedure. Human coders coded the data set of 7321 posts according to predetermined classifications within the bullying role, form of bullying and type of bullying post categories. That coding was then used as training data to train machine learning models. The decisions about what classifications to code for each question were based on the existing social science theoretical foundation about bullying and preliminary analysis of the scope of information present in the tweets. The coding process for each category is discussed in the corresponding Results section. 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. Of the 9,764,583 tweets classified as bullying traces by the classifier, about 2% (191,657) of them contain the 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. The time of day, week, and year of each post were determined from the timestamp that is associated with each tweet. The timestamp is in Coordinated Universal Time (UTC), which may not be users' local time. As a result, we were only able to obtain the local time of geo-tagged tweets.

When we address each of the five questions, we present the data for the 2011–2012 and 2012–2013 years separately and combined to illustrate the consistency in the trends that we identified across the two years under study. 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 across 2012–2013.

3. Results

3.1. Who is posting about bullying on Twitter?

Human coders coded the role of the author of every post identified as a bullying trace in the training set of 7321 posts. The original six categories we searched for were derived from Salmivalli (1999): bully, victim, bystander, defender, assistant, and reinforcer. We also created two new roles that are relevant for social media posts: reporter (shares information about an episode but is not involved in any way, including as a bystander), and accuser (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). For example, AUTHOR(reporter): *"We visited my cousin today & #Itreallymakesmemad that he barely eats bec he was bullied."* The interrater agreement for these 9 categories was calculated based on 2 coders coding 1000 of the 7321 posts in the training set. It was determined to be $\kappa = .79$. The human-coded data, presented in Table 1, revealed that reporters (33.68% of bullying trace authors) and victims (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.

To analyze who posted across all bullying traces identified across the 2011–2013 school years, we trained an author role classifier with support vector machine, a standard machine learning tool. The classifier achieved 70% cross validation accuracy, which is far from perfect but better than the majority class (i.e., reporter) baseline of 33.68%. Table 1 shows the confusion matrix that illustrates agreement and disagreement between human-coded bullying role and predicted bullying role when we use machine learning methods on the data. The classifier found a similar distribution as our coded data with victims (36.01%, $n = 3{,}515{,}760$) and reporters (32.52%, $n = 3{,}175{,}581$) being identified as authors of bullying traces most frequently. Table 2 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.

3.2. What forms of bullying are mentioned or used on Twitter?

Two human coders coded the form of bullying mentioned in every post identified as a bullying trace in the training set of 7321 posts. The categories were general (no information is

Table 2
Distribution of human-coded and machine-learning identified bullying trace categories.

	Human-coded data		Tweets 9/1/11-8/31/12		Tweets 9/1/12-8/31/13		Tweets 9/1/11-8/31/13	
	Count	%	Count	%	Count	%	Count	%
Total tweets	7321		12,421,237		20,056,321		32,477,558	
Bullying traces	2102	28.71	3,955,458	31.84	5,809,125	28.96	9,764,583	30.07
Who: The author's role								
Accuser	317	15.08	662,880	16.76	983,801	16.94	1,646,681	16.86
Bully	303	14.41	496,039	12.54	685,269	11.80	1,181,308	12.10
Defender	178	8.47	99,322	2.51	145,900	2.51	245,222	2.51
Reporter	708	33.68	1,337,205	33.81	1,838,376	31.65	3,175,581	32.52
Victim	589	28.02	1,360,001	34.38	2,155,759	37.11	3,515,760	36.01
Other	7	0.33	11	0.00	20	0.00	31	0.00
What: The form of bullying								
General	1857	88.34	3,765,015	95.19	5,531,636	95.22	9,296,651	95.21
Cyberbullying	145	6.90	164,866	4.17	239,517	4.12	404,383	4.14
Verbal	67	3.19	20,403	0.52	30,931	0.53	51,334	0.53
Physical	33	1.57	5174	0.13	7041	0.12	12,215	0.13
Why: The type of post								
Accusation	316	15.03	595,383	15.05	887,508	15.28	1,482,891	15.19
Cyberbullying	16	0.76	14	0.00	28	0.00	42	0.00
Denial	128	6.09	79,630	2.01	106,298	1.83	185,928	1.90
Report	709	33.73	1,175,234	29.71	1,614,037	27.78	2,789,271	28.57
Self-disclosure	933	44.39	2,105,197	53.22	3,201,254	55.11	5,306,451	54.34

provided to indicate a form), cyberbullying, physical, verbal, relational, and property damage (see Wang et al., 2009 for sample behaviors that correspond with these categories). The interrater agreement across two human coders for these seven categories was $\kappa = .77$.

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 classifications of property damage and relational. After doing this, the classifier achieved 70% accuracy (see Table 1). Across all bullying traces in 2011–2013, the classifier found that posts about general forms of bullying were by far the most common—95.21% ($n = 9{,}296{,}651$) of the bullying traces. Cyberbullying posts comprised the next most frequent form (4.14%, $n = 404{,}383$ posts). See the middle panel of Table 2 for the counts of forms of bullying as labeled by human coders and machine learning methods.

3.3. Why are people posting about bullying on Twitter?

The initial categories of why people post about bullying episodes were determined inductively through preliminary coding and discussions of all bullying traces. The different types of bullying traces identified were:

- Reports: Posts that described a bullying episode someone knows about, "*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, "*@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, "*People bullied me for being fat. 7 years later, I was diagnosed with bulimia.*"
- Denials: Posts where the author denied a bullying role, "*@USER lol I'm not a bully man*"
- Cyberbullying: Posts that were direct attacks from a bully to a victim. "*@USER really I am just cyberbullying you right now*").

The interrater agreement for these five categories based on two human coders was $\kappa = .76$. To analyze the distribution across all bullying traces in 2011–2013, we trained the type (i.e., why people post) classifier with support vector machine. The accuracy of this classification was 72% (see Table 1). The classifier found that self-disclosure posts (54.34%, $n = 5{,}306{,}451$) were most common followed by reports (28.57%, $n = 2{,}789{,}271$), accusations (15.19%, $n = 1{,}482{,}891$) and denials (1.90%, $n = 185{,}928$). See the bottom panel of Table 2.

3.4. Where do posts about bullying originate from?

To understand the origins of bullying traces, we estimated the amount of bullying traces per capita for the 50 US states and Washington DC identified through posts that contained GPS information. Table 3 presents the state names listed in alphabetical order, their population size 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 2011–2013. 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 amount of bullying traces per capita are Delaware, Washington DC, Maryland, Ohio, and Rhode Island across 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, 1988).

The imperfect association between population-size ranking and the number of bullying traces per capita ranking is illustrated in Fig. 1. 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.

3.5. When do posts about bullying occur?

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. Establishing the location of each post was necessary to appropriately determine the time. We calculated

Table 3
State population size, number of GPS bullying traces, per capita bullying traces, and population rank and per capita bullying traces rank for 50 states and the District of Columbia between September 1, 2011 and August 31, 2013.

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.

the timing from the bullying traces in one east coast state, New York, and from a west coast state, California, for both years under investigation. We chose these two states because they had the largest number of GPS-tagged bullying traces and contained 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 is statistically uniform across the days of the week in either time period in both New York and California. The chi-square analyses indicate that the distribution was not uniform across days of the week in both New York, χ^2 (6) = 31.29, $p < .001$ in 2011–2012 and χ^2 (6) = 59.78, $p < .001$ in 2012–2013, and California χ^2 (6) = 52.32, $p < .001$ in 2011–2012 and χ^2 (6) = 182.62,

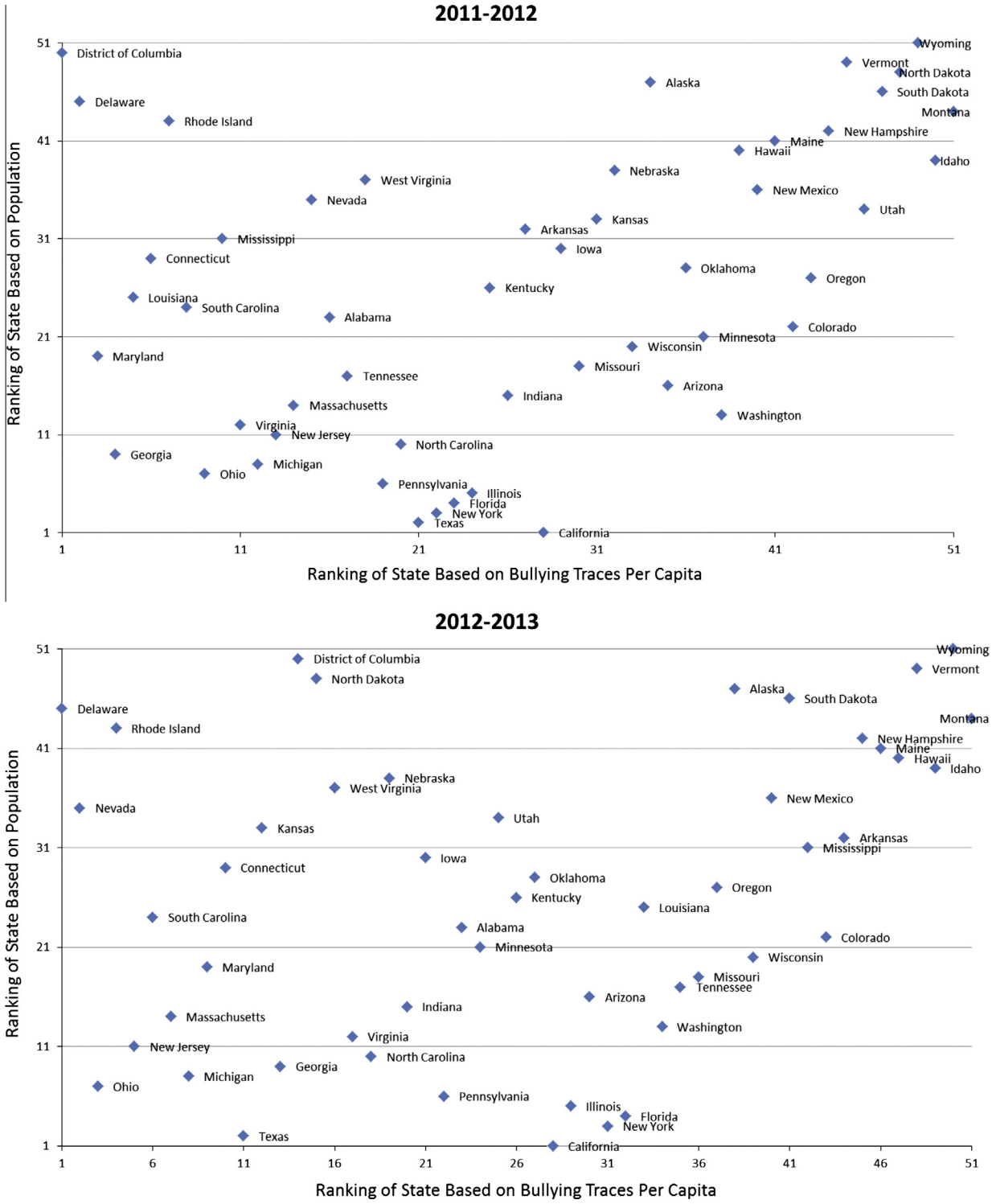

Fig. 1. Association between ranking of number of bullying traces per capita (1 = largest; 51 = smallest) and population size (1 = largest; 51 = smallest) for 50 states and the District of Columbia from September 1, 2011 through August 31, 2012 (top) and September 1, 2012 through August 31, 2013 (bottom).

$p < .001$ in 2012–2013. The effect size for each test is 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. Table 4 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, Fig. 2 illustrates that there is diurnal pattern such that most posts occur in waking periods (especially during the evening hours) in both locations in both years.

4. Discussion

In this paper we merged social science and computer science to address five fundamental questions about bullying by examining all bullying traces captured with bullying keywords that occurred on Twitter over the school years of 2011–2013. The machine learning methods provided a unique perspective on how discrete bullying episodes are represented in social media. Victims and reporters were identified as role players WHO post about bullying most frequently. Most posts did not indicate WHAT form of bullying was used in each episode. A salient reason for WHY authors posted was to disclose about their involvement in a bullying episode. Exploration of WHERE bullying traces originated indicated a moderate association between the rankings of the size of a state and where the largest number of bullying traces originated. The timing of the bullying traces shows that posting is not uniform across the week and that it peaks in frequency during the evenings, possibly WHEN individuals are awake, yet not in school or work contexts.

In addition to uncovering new answers to questions about bullying, this study showcases the potential of new data sources. Among the strengths of the approach relative to traditional social science methods is that it utilized large volumes of data (i.e., millions of posts) collected from any English-language public tweets. Additionally, each post can refer to anyone, anywhere: Twitter users and/or non-users and school and/or non-school settings. The posts also allow for a broader representation of all role players in bullying episodes, including reporters who had not been identified in previous work.

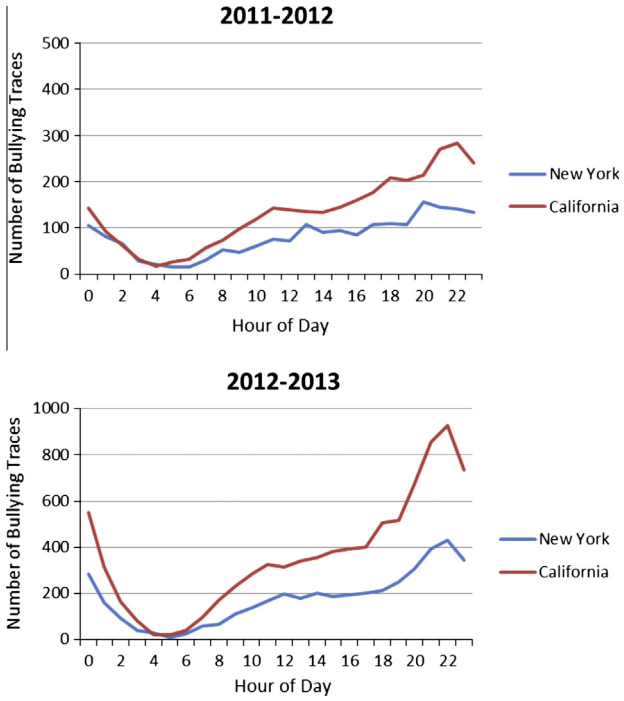

Fig. 2. Number of bullying traces for each hour of the day from September 1, 2011 through August 31, 2012 (top) and September 1, 2012 through August 31, 2013 (bottom) that originated in New York and California.

The study illustrates that social media data can be interpreted by human coders and situated within a broader research literature and accurately analyzed by machine learning methods. The key is to identify discrete, meaningful categories from the limited text present within the posts. In the future, more complex machine learning models can be applied that will consider the context information available in related posts such as a reply chain of tweets. Such context can be very helpful in disambiguating the short tweet posts, and machine learning models can benefit to achieve higher accuracies. These models will also allow for the construction of partial social networks based on interactions between users.

Table 4
Number of bullying traces on each day of the week in New York and California in 2011–2012 and 2012–2013.

	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 $p < .05$.

4.1. Limitations

While it is clear that this approach can address old questions in new ways and even generate new questions, there are clear limitations with the methodology. Chief among these is the reliance on a set of keywords. The approach used here is able to collect a substantial number of bullying traces, but does not capture all bullying-related tweets. A next step could 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. Even with this approach, it will be difficult to capture all bullying episodes because they can be represented in so many different ways. Further, to avoid sampling bias with the use of the Twitter API that arises when researcher-requested keywords produce a large number of posts that exceed the rate limits set by Twitter, it is essential that choosing keywords is balanced with avoiding this potential bias. In all, because the approach relies so strongly on keyword selection, it is critical that researchers attend to the limits of the questions that can be addressed given the keywords that they choose.

Another limitation of the methodology is that a reliance on public tweets, as opposed to studying a sample of users who also fill out other measures, limits information about actual occurrences of bullying within a population as well as key details about the participants. Among teens on Twitter, approximately 64% have public accounts, 24% are private, and 12% do not know if their tweets are public or private (Madden et al., 2013). Private tweets about bullying may qualitatively differ from public ones, such as the proportion of different roles and types of tweets. Privacy of tweets may elicit users to disclose themselves in more stigmatized roles (i.e., bully or assistant) or increase cyberbullying, especially if users are anonymous and/or their followers are strangers (Kowalski et al., 2014; Postmes & Spears, 1998; Whittaker & Kowalski, 2015). Without collecting other measures of bullying occurrences, we cannot calibrate the number of instances reported with social media to actual occurrences. Other basic information about users is also missing. Only 2% of users included their location. Even less information is available about gender, ethnicity, and age because Twitter users do not report these in the same way as is done with other social network sites. Without this basic information, we do not know how representative the sample is of the larger U.S. population.

4.2. Conclusion

This paper demonstrates the promise of using machine learning methodology to study social phenomena on social media. The methodology removes some of the barriers of other methods by studying many people in a dynamic way within a salient real-life environment. The focus on experiences within contexts that are highly relevant to individuals' everyday lives illustrates the strength of the method both for addressing questions of theoretical and practical importance. In sum, we believe that this approach can be added to existing methods that are already in play to study significant social-psychological issues such as bullying.

Acknowledgements

This study was funded by grant IIS-1216758 awarded to Xiaojin Zhu and Amy Bellmore by the National Science Foundation. The funding source had no involvement with data collection, analysis, or write-up.

References

American Psychological Association, Council of Representatives (2004). *APA resolution on bullying among children and youth*. Retrieved from <https://www.apa.org/about/policy/bullying.pdf>.

Baly, M. W., Cornell, D. G., & Lovegrove, P. (2014). A longitudinal investigation of self- and peer reports of bullying victimization across middle school. *Psychology in the Schools, 51*, 217–240. http://dx.doi.org/10.1002/pits.21747.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York, NY: Springer Verlag.

Bosse, T., & Stam, S. (2011). A normative agent system to prevent cyberbullying. *WI-IAT, 2011*, 425–430. http://dx.doi.org/10.1109/WI-IAT.2011.24.

boyd, d., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication, & Society, 15*, 662–679. http://dx.doi.org/10.1080/1369118X.2012.678878.

Card, N. A., & Hodges, E. V. E. (2008). Peer victimization among schoolchildren: Correlations, causes, consequences, and considerations in assessment and intervention. *School Psychology Quarterly, 23*, 451–461. http://dx.doi.org/10.1037/a0012769.

Chen, G. M. (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*, 755–762. http://dx.doi.org/10.1016/j.chb.2010.10.023.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Committee on Injury, Violence, and Poison Prevention (2009). Policy statement – Role of the pediatrician in youth violence prevention. *Pediatrics, 124*, 393–402. http://dx.doi.org/10.1542/peds.2009-0943.

Duggan, M., & Smith, A. (2013). Social media update 2013: 42% of online adults use multiple social networking sites, but Facebook remains the platform of choice. *Pew research center*. Retrieved from <http://www.pewinternet.org/2013/12/30/social-media-update-2013/>.

Ellison, N. B., Steinfield, C., & Lampe, C. (2007). The benefits of Facebook "friends:" Social capital and college students' use of online social network sites. *Journal of Computer-Mediated Communication, 12*, 1143–1168. http://dx.doi.org/10.1111/j.1083-6101.2007.00367.x.

Gini, G., & Pozzoli, T. (2013). Bullied children and psychosomatic problems: A meta-analysis. *Pediatrics, 132*, 720–729. http://dx.doi.org/10.1542/peds.2013-0614.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York, NY: Springer-Verlag.

Hawker, D. S. J., & Boulton, M. J. (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*, 441–455. http://dx.doi.org/10.1111/1469-7610.00629.

Java, A., Song, X., Finin, T., & Tseng, B. (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* (pp. 56–65). ACM. http://dx.doi.org/10.1145/1348549.1348556.

Jimerson, S. R., Swearer, S. M., & Espelage, D. L. (2009). *Handbook of bullying in schools: An international perspective*. New York, NY: Routledge.

Juvonen, J., & Graham, S. (2014). Bullying in schools: The power of bullies and the plight of victims. *Annual Review of Psychology, 65*, 159–185. http://dx.doi.org/10.1146/annurev-psych-010213-115030.

Kontostathis, A., Edwards, L., & Leatherman, A. (2010). Text mining and cybercrime. In M. W. Berry & J. Kogan (Eds.), *Text mining: Applications and theory* (pp. 1–14). Chichester, UK: John Wiley & Sons.

Kowalski, R. M., Giumetti, G. W., Schroeder, A. N., & Lattanner, M. R. (2014). Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth. *Psychological Bulletin, 140*, 1073–1137. http://dx.doi.org/10.1037/a0035618.

Latham, A., Crockett, K., & Bandar, Z. (2010). A conversation expert system supporting bullying and harassment policies. In the *2nd ICAART* (pp. 163–168).

Lenhart, A., Purcell, K., Smith, A., & Zickuhr, K. (2010). Social media & mobile Internet use among teens and young adults. *Pew internet & American life project*. Retrieved from <http://pewinternet.org/Reports/2010/Social-Media-and-Young-Adults.aspx>.

Macbeth, J., Adeyema, H., Lieberman, H., & Fry, C. (2013). Script-based story matching for cyberbullying prevention. In *CHI '13 extended abstracts on human factors in computing systems* (pp. 901–906). New York, NY: ACM.

Madden, M., Lenhart, A., Cortesi, S., Gasser, U., Duggan, M., Smith, A., & Beaton, M. (2013). Teens, social media, and privacy. *Pew Internet & American life project*. Retrieved from <http://www.pewinternet.org/files/2013/05/PIP_TeensSocialMediaandPrivacy_PDF.pdf>.

McTernan, W. P., Dollard, M. F., & LaMontagne, A. D. (2013). Depression in the workplace: An economic cost analysis of depression-related productivity loss attributable to job strain and bullying. *Work & Stress, 27*, 321–338. http://dx.doi.org/10.1080/02678373.2013.846948.

Mitchell, T. M. (2006). *The discipline of machine learning*. Carnegie Mellon University, School of Computer Science, Machine Learning Department. Retrieved from <http://www-cgi.cs.cmu.edu/~tom/pubs/MachineLearningTR.pdf>.

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's Firehose. In *Seventh international AAAI conference on weblogs and social media*, Cambridge, Massachusetts.

Nishina, A., & Juvonen, J. (2005). Daily reports of witnessing and experiencing peer harassment in middle school. *Child Development, 76*, 435–450. http://dx.doi.org/10.1111/j.1467-8624.2005.00855.x.

Nylund, K., Bellmore, A., Nishina, A., & Graham, S. (2007). Subtypes, severity, and structural stability of peer victimization: What does latent class analysis say? *Child Development, 78*, 1706–1722. http://dx.doi.org/10.1111/j.1467-8624.2007.01097.x.

Oh, H. J., Ozkaya, E., & LaRose, R. (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. http://dx.doi.org/10.1016/j.chb.2013.07.053.

Olweus, D. (1993). *Bullying at school: What we know and what we can do*. Oxford, UK: Blackwell.

Postmes, T., & Spears, R. (1998). Deindividuation and antinormative behavior: A meta-analysis. *Psychological Bulletin, 123*, 238–259. http://dx.doi.org/10.1037/0033-2909.123.3.238.

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

Rivers, I., Poteat, V. P., Noret, N., & Ashurst, N. (2009). Observing bullying at school: The mental health implications of witness status. *School Psychology Quarterly, 24*, 211–223. http://dx.doi.org/10.1037/a0018164.

Salmivalli, C. (1999). Participant role approach to school bullying: Implications for interventions. *Journal of Adolescence, 22*, 453–459. http://dx.doi.org/10.1006/jado.1999.0239.

Salmivalli, C., Lagerspetz, K., Björkqvist, K., Österman, K., & Kaukiainen, A. (1996). Bullying as a group process: Participant roles and their relations to social status within the group. *Aggressive Behavior, 22*, 1–15. http://dx.doi.org/10.1002/(SICI)1098-2337(1996) 22:1<1::AID-AB1>3.0.CO;2-T.

The White House, Office of the Press Secretary (2011). Background on White House conference on bullying prevention [Press Release]. Retrieved from <http://www.whitehouse.gov/the-press-office/2011/03/10/background-white-house-conference-bullying-prevention>.

Wang, J., Iannotti, R. J., & Nansel, T. R. (2009). School bullying among US adolescents: Physical, verbal, relational and cyber. *Journal of Adolescent Health, 45*, 368–375. http://dx.doi.org/10.1016/j.jadohealth.2009.03.021.

Wasserman, L. (2003). *All of statistics: A concise course in statistical inference*. New York, NY: Springer.

Whittaker, E., & Kowalski, R. M. (2015). Cyberbullying via social media. *Journal of School Violence, 14*, 11–29.

Xu, J.-M., Jun, K.-S., Zhu, X., & Bellmore, A. (2012). Learning from bullying traces in social media. In *Proceedings of the conference of the North American chapter of the association for computational linguistics: human language technologies* (pp. 656–666). Montreal, Quebec, Canada: Association for Computational Linguistics. Retrieved from <http://www.proceedings.com/15547.html>.

Zhu, X., & Goldberg, A. B. (2009). Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning, 3*, 1–130. http://dx.doi.org/10.2200/S00196ED1V01Y200906AIM006.