Learning from Bullying Traces in Social Media

Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu
Department of Computer Sciences
University of Wisconsin-Madison
Madison, WI 53706, USA
{xujm,deltakam,jerryzhu}@cs.wisc.edu

Amy Bellmore
Department of Educational Psychology
University of Wisconsin-Madison
Madison, WI 53706, USA
abellmore@wisc.edu

Abstract

We introduce the social study of bullying to the NLP community. Bullying, in both physical and cyber worlds (the latter known as cyberbullying), has been recognized as a serious national health issue among adolescents. However, previous social studies of bullying are handicapped by data scarcity, while the few computational studies narrowly restrict themselves to cyberbullying which accounts for only a small fraction of all bullying episodes. Our main contribution is to present evidence that social media, with appropriate natural language processing techniques, can be a valuable and abundant data source for the study of bullying in both worlds. We identify several key problems in using such data sources and formulate them as NLP tasks, including text classification, role labeling, sentiment analysis, and topic modeling. Since this is an introductory paper, we present baseline results on these tasks using off-the-shelf NLP solutions, and encourage the NLP community to contribute better models in the future.

1 Introduction to Bullying

Bullying, also called peer victimization, has been recognized as a serious national health issue by the White House (The White House, 2011), the American Academy of Pediatrics (The American Academy of Pediatrics, 2009), and the American Psychological Association (American Psychological Association, 2004). 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, Swearer, and Espelage, 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, Nelson, and Sampilo, 2010; Graham, Bellmore, and Juvonen, 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, Kochenderfer, and Coleman, 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, Jackson, and Brown, 2009; Fredstrom, Adams, and Gilman, 2011; Wang, Iannotti, and Nansel, 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, Madsen, and Moody, 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).

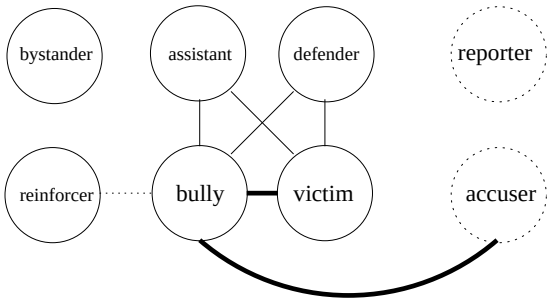


Figure 1: The roles in a bullying episode. Solid circles represent traditional roles in social science, while dotted circles are new roles we augmented for social media. The width of the edges represents interaction strength.

1.1 The Structure of a Bullying Episode

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 behaviors.

Participants in a bullying episode take well-defined *roles* (see Figure 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).

1.2 Some Scientific Questions NLP can Answer

Like many complex social issues, effective solutions to bullying go beyond technology alone and require the concerted efforts of parents, educators, and law enforcement. To guide these efforts it is paramount to study the dynamics of bullying. Such study critically depends on text in the form of self-report social study surveys and electronic communication among participants. Such text is often fragmental, noisy, and covers only part of a bullying episode from a specific role's perspective. As such, the NLP community can help answer a host of scientific questions: Which pieces of text refer to the same underlying bullying episode? What is the form, reason, location, time, etc. of a bullying episode? Who are the participants of each episode, and what are their roles? How does a person's role evolve over time? This paper presents our initial investigation on some of these questions, while leaving others to future research by the NLP community.

1.3 Limitations of the State-of-the-Art

The social science study of bullying has a long history. However, a fundamental problem there is data acquisition. The standard approach is to conduct time-consuming personal surveys in schools. The sample size is typically in the hundreds, and participants typically write 3 to 4 sentences about each bullying episode (Nishina and Bellmore, 2010). Such a small corpus fails to assess the true frequency of bullying over the population, and cannot determine the evolution of roles. The computational study of bullying is largely unexplored, with the exception of a few studies on cyberbullying (Lieberman, Dinakar, and Jones, 2011; Dinakar, Reichart, and Lieberman, 2011; Ptaszynski et al., 2010; Kontostathis, Edwards, and Leatherman, 2010; Bosse and Stam, 2011; Latham, Crockett, and Bandar, 2010). These studies did not consider the much more frequent bullying episodes in the physical world.

2 Bullying Traces in Social Media

The main contribution of the present paper is not on novel algorithms, but rather on presenting evidence that social media data and off-the-shelf NLP tools can be an effective combination for the study of bullying. 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. Bullying traces are valuable, albeit fragmental and noisy, data which we can use to piece together the underlying episodes.

In the rest of the paper, we focus on publicly available Twitter "tweets," though our methods apply readily to other social media services, too. 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 (not the authors of the present paper). 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 produces 250 million tweets per day in late 2011. Even with the lower bound in 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 2 shows the daily number of bullying traces identified by our classifier, to be discussed in section 3. A weekly pattern was obvious in late August. 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.

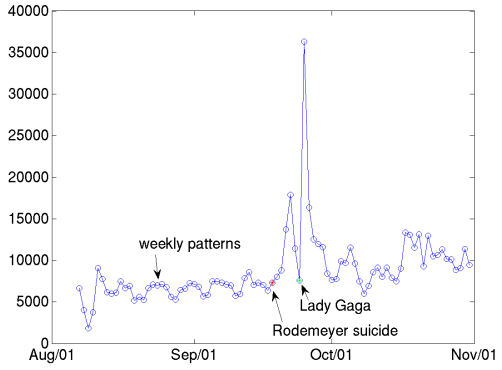


Figure 2: Temporal variation of bullying traces

In the following sections, we identify several key problems in using social media for the study of bullying. We formulate each key problem as an NLP task. We then present standard off-the-shelf NLP approaches to establish baseline performances. Since bullying traces account for only a tiny fraction of all tweets, it posed a significant challenge for our annotators to find enough bullying traces without labeling an unreasonable amount of tweets. For this reason, in the rest of the paper we restrict ourselves to an "enriched dataset." This enriched dataset is obtained by collecting tweets using the public Twitter streaming API, such that each tweet contains at least one of the following keywords: "bully, bullied, bullying." 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.

3 NLP Task A: Text Categorization

One important task is to distinguish bullying traces from other social media posts. Our enriched dataset, generated by simple keyword filtering, still contains many irrelevant tweets. For example, *"Forced veganism by removing a persons choice is just another form of bullying"* is not a bullying trace, since it does

not describe a bullying episode. Our task is to distinguish posts like this from true bullying traces such as those mentioned in the previous section. We formulate it as a binary text categorization task.

Methods. The same annotators who labeled the TREC corpus labeled 1762 tweets sampled uniformly from the enriched dataset on August 6, 2011. Among them, 684 (39%) were labeled as bullying traces.

Following (Settles, 2011), these 1762 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.

We chose four commonly used text classifiers, namely, Naive Bayes, SVM with linear kernel (`SVM(linear)`), SVM with 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, Fung, and Rosales, 2007) for the fourth.

We held out 262 tweets for test, and systematically varied training set size among the remaining tweets, from 100 to 1500 with the 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 the 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.

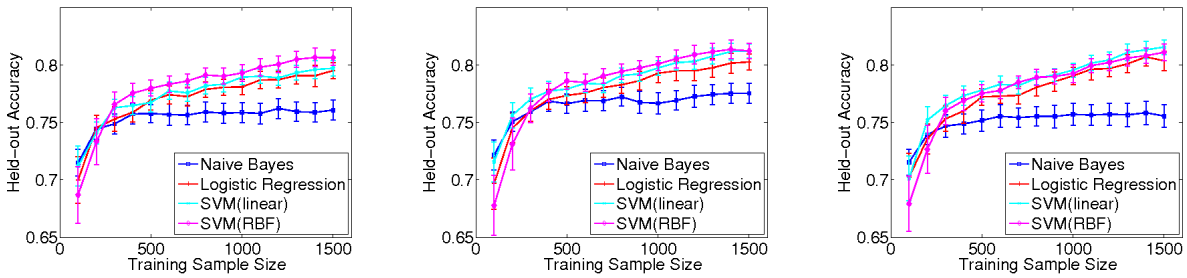
Results. Figure 3 reports the held-out set accuracy as the training set size increases. The error bars are ± 1 standard error. With the largest training set size (1500), the combination of `SVM(linear)` + `1g` achieves an average accuracy 79.7%. `SVM(linear)` + `1g2g` achieves 81.3%, which is significantly better (t-test, $p = 4 \times 10^{-6}$). It shows that including bigrams can significantly improve the classification performance. `SVM(linear)` + `1g2gPOS` achieves 81.6%, though the improvement is not statistically significant ($p = 0.088$), which indicates that POS coloring does not help too 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 also note that these accuracies are much better than the majority class baseline of 61%. On the held-out set, `SVM(linear)` + `1g2g` achieves precision P=0.76, recall R=0.79, and F-measure 0.77.

Discussions. Note that the learning curves are still increasing, suggesting that better accuracy can be obtained if we annotate more training data. As to why the best accuracy is not close to 1, one hypothesis is noisy labels caused by intrinsic disagreement among labelers. Tweets are short and some are ambiguous. Without prior knowledge about the users and their other tweets, labelers interpret 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 and created noisy bullying trace labels.

A future direction is to categorize bullying traces at a finer granularity, e.g., by forms, reasons, etc. This can be solved by multi-class classification methods. Another direction is to extend the classifiers from the "enriched data" to the full range of tweets. Recall that the difference is whether we prefilter the tweets by keywords. Clearly, they have different tweet distributions. Techniques used for *covariate shift* may be adapted to solve this problem (Blitzer, 2008).

4 NLP Task B: Role Labeling

Identifying participants' bullying roles (Figure 1) is another important task, which is also a prerequisite of studying how a person's role evolves over time. For bullying traces in social media, we augment the traditional role system with two new roles: reporter (may not be present during the episode, un-



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

Figure 3: Learning Curves for different feature sets and classification algorithms

like a bystander) and accuser (accusing someone as the bully). 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 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.

4.1 Author's Roles

Methods. Our annotators labeled the author's role for each of the 684 positive bullying traces in Task A (296 R, 162 V, 98 B, 86 A, 42 O). We used the same classifiers and features in Section 3. 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 is `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.

	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 1: Confusion Matrix of Author Role Classification

Table 1 shows the confusion matrix of the best model. Most R and V authors are 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.

Discussions. As tweets are short, our feature representation may not be the best for predicting author's role. Many authors mentioned themselves in the tweets with first-person pronouns, making it advantageous to consider joint classification by merging sections 4.1 and 4.2. Furthermore, assuming roles change infrequently, it may be helpful to jointly classify many tweets authored by the same person.

4.2 Person-Mention's Roles

This sub-task labels each person-mention with a bullying role. It uses Named Entity Recognition (NER) (Finkel, Grenager, and Manning, 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, Roth, and Yih, 2008) but differs critically in that our roles are not tied to specific verb predicates.

Methods. Our annotators labeled each token in the 684 bullying traces with the tags A, B, R, V, O and N for not-a-person. There are 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 which 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 is done individually on each token. We randomly split the 684 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.

	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: Cross Validation Result of Person-Mention Roles

Results. Table 2 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 0.87, 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 outperforms SVM in all measures, showing the value of joint classification.

Discussions. Table 3 shows the confusion matrix of person-mention role labeling by 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	9306

Table 3: Confusion Matrix of Person-Mention Roles by CRF

5 NLP Task C: Sentiment Analysis

Sentiment analysis on participants involved in a bullying episode is of significant importance. As Figure 4 suggests, there are a wide range of emotions in bullying traces. For example, victims usually experience negative emotions such as depression, anxiety and loneliness; Some emotions are more violent or even suicidal. Detecting at-risk individuals via sentiment analysis enables potential interventions. In addition, social scientists are interested in sentiment analysis of bullying participants to understand their motivations.

In the present paper we investigate a special form of sentiment in bullying traces, namely teasing. 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.

Methods. One first task is to identify teasing bullying traces. We formulated it as a binary classification problem, similar to classic positive/negative sentiment classification (Pang and Lee, 2004). Our annotators labeled each of the 684 bullying traces in Task A as teasing (99) or not (585). We used the same feature representations, classifiers and parameter tuning as in Section 3 and 10-fold cross validation procedure.

Results. 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 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."

	predicted as	
	Tease	Not
Tease	52	47
Not	26	559

Table 4: Confusion Matrix of Teasing Classification

Discussions. 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.

6 NLP Task D: Latent Topic Modeling

Methods. Given the large volume of bullying traces, methods for automatically analyzing what people are talking about are needed. Latent topic models allow us to extract the main topics in bullying traces to facilitate understanding. We used latent Dirichlet allocation (LDA) (Blei, Ng, and Jordan, 2003) as our exploratory tool. Specifically, we ran a collapsed Gibbs sampling implementation of LDA (Griffiths and Steyvers, 2004).

The corpus consists of 188K enriched tweets from Aug. 21 to Sept. 17, 2011 that are classified as

bullying traces by our classifier in Task A. We performed stopword removal and further removed word types occurring less than 7 times, resulting in a vocabulary of size 12K. We set the number of topics to 50, Dirichlet parameter for word multinomials to $\beta = 0.01$, Dirichlet parameter for document topic multinomial to $\alpha = 1$, and ran Gibbs sampling for 10K iterations.

Results. Space precludes a complete list of topics. Figure 4 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 $p(\text{word} \mid \text{topic})$. The topic names are 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, Ng, and Jordan, 2003). The results were similar, thus we omit discussion of them.

Discussions. Some recovered topics, including the ones shown here, provide valuable insight 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).

Figure 4: Selected topics discovered by latent Dirichlet allocation.

7 Conclusion and Future Work

We introduced social media as a large-scale, near real-time, dynamic data source for the study of bullying. Social media offers a broad range of bullying traces that include but go beyond cyberbullying. In the present paper, we have identified several key problems in using social media to study bullying and formulated them as familiar NLP tasks. Our baseline performance with standard off-the-shelf approaches shows that it is feasible to learn from bullying traces.

Much work remains in this new research direction. In the short term, we need to develop specialized NLP tools for processing bullying traces in social media, similar to (Ritter et al., 2011; Liu et al., 2010), to achieve better performance than models trained on formal English. In the long term, we need to tackle the problem of piecing together the underlying bullying episodes from fragmental bullying traces. Consider two separate bullying episodes with the following participants and roles:

E1: B: Buffy, V: Vivian & Virginia, O: Debra

E2: B: Burton, V: Buffy, O: Irene

The corresponding bullying traces can be three posts in this order:

$\mathbf{w}_1$ Debra: *Virginia, I heard Buffy call you and Vivian fat–ignore her!*

$\mathbf{w}_2$ Buffy to Irene: *Burton picked on me again because I'm only 5'1*

$\mathbf{w}_3$ Vivian: *Buffy I'm not fat! Stop calling me that.*

Reconstructing E1, E2 from $\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3$ is challenging for a number of reasons: (1) There is no explicit episode index in the posts. (2) Posts from a single episode may be dispersed in time (e.g., $\mathbf{w}_1, \mathbf{w}_3$ belong to E1, but not $\mathbf{w}_2$), each containing only part

of an episode. (3) The number of episodes and people can grow indefinitely as more posts arrive. (4) People may switch roles in different episodes (e.g., Buffy was the bully in E1 but the victim in E2). Joint probabilistic modeling over multiple posts using social network structures hold great promise in solving this problem.

To facilitate bullying research in the NLP community, we make our annotations and software publicly available at `http://research.cs.wisc.edu/bullying`.

Acknowledgments

We thank Wei-Ting Chen, Rachael Hansen, Ting-Lan Ma, Ji-in You and Bryan Gibson for their help on the data and the paper.

References

[American Psychological Association2004] American Psychological Association. 2004. APA resolution on bullying among children and youth. `http://www.apa.org/about/governance/council/policy/bullying.pdf`.

[Andrzejewski et al.2011] Andrzejewski, David, Xiaojin Zhu, Mark Craven, and Ben Recht. 2011. A framework for incorporating general domain knowledge into Latent Dirichlet Allocation using First-Order Logic. In *the 22nd IJCAI*, pages 1171–1177.

[Archer and Coyne2005] Archer, John and Sarah M. Coyne. 2005. An integrated review of indirect, relational, and social aggression. *Personality and Social Psychology Review*, 9:212–230.

[Bellmore et al.2004] 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, Nelson, and Sampilo2010] 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, Ng, and Jordan2003] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *JMLR*, 3:993–1022.

[Blitzer2008] Blitzer, John. 2008. *Domain Adaptation of Natural Language Processing Systems*. Ph.D. thesis, University of Pennsylvania.

[Bosse and Stam2011] Bosse, Tibor and Sven Stam. 2011. A normative agent system to prevent cyberbullying. In *WI-IAT 2011*, pages 425–430.

[Brody and Diakopoulos2011] Brody, Samuel and Nicholas Diakopoulos. 2011. Cooooooooooooooollllllllllll!!!!!!!!!!!!!! using word lengthening to detect sentiment in microblogs. In *EMNLP 2011*, pages 562–570.

[Cassidy, Jackson, and Brown2009] 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 Int'l*, 30(4):383–402.

[Chang and Lin2011] Chang, Chih-Chung and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Trans. on Intelligent Systems and Technology*, 2:27:1–27:27.

[Cook et al.2010] 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.

[Dinakar, Reichart, and Lieberman2011] Dinakar, K., R. Reichart, and H. Lieberman. 2011. Modeling the detection of textual cyberbullying. In *International Conference on Weblog and Social Media - Social Mobile Web Workshop*, Barcelona, Spain.

[Fekkes et al.2006] 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.

[Finkel, Grenager, and Manning2005] Finkel, Jenny R., Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *the 43rd ACL*, pages 363–370.

[Fredstrom, Adams, and Gilman2011] Fredstrom, Bridget K., Ryan E. Adams, and Rich Gilman. 2011. Electronic and school-based victimization: Unique contexts for adjustment difficulties during adolescence. *J. Youth and Adolescence*, 40(4):405–415.

[Gildea and Jurafsky2002] Gildea, Daniel and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Comput. Linguist.*, 28(3):245–288.

[Gini and Pozzoli2009] Gini, Gianluca and Tiziana Pozzoli. 2009. Association between bullying and psychosomatic problems: a meta-analysis. *Pediatrics*, 123(3):1059–1065.

[Graham, Bellmore, and Juvonen2007] Graham, Sandra, Amy Bellmore, and Jaana Juvonen. 2007. Peer victimization in middle school: When self- and peer

views diverge. In Joseph E. Zins, Maurice J. Elias, and Charles A. Maher, editors, *Bullying, victimization, and peer harassment: A handbook of prevention and intervention*. Haworth Press, New York, NY, pages 121–141.

[Griffiths and Steyvers2004] Griffiths, Thomas L. and Mark Steyvers. 2004. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235.

[Hall et al.2009] 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.

[Hawker and Boulton2000] 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. *J. of Child Psychology And Psychiatry*, 41(4):441–455.

[Janosz et al.2008] 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? *J. of Adolescent Health*, 43(6):600–608.

[Jimerson, Swearer, and Espelage2010] Jimerson, Shane R., Susan M. Swearer, and Dorothy L. Espelage. 2010. *Handbook of Bullying in Schools: An international perspective*. Routledge/Taylor & Francis Group, New York, NY.

[Juvonen and Graham2001] Juvonen, Jaana and Sandra Graham. 2001. *Peer harassment in school: The plight of the vulnerable and victimized*. Guilford Press, New York, NY.

[Juvonen and Gross2008] Juvonen, Jaana and Elisheva F. Gross. 2008. Extending the school grounds? – Bullying experiences in cyberspace. *J. of School Health*, 78:496–505.

[Kontostathis, Edwards, and Leatherman2010] Kontostathis, April, Lynne Edwards, and Amanda Leatherman. 2010. Text mining and cybercrime. In Michael W. Berry and Jacob Kogan, editors, *Text Mining: Applications and Theory*. John Wiley & Sons, Ltd, Chichester, UK.

[Ladd, Kochenderfer, and Coleman1997] 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.

[Latham, Crockett, and Bandar2010] Latham, Annabel, Keeley Crockett, and Zuhair Bandar. 2010. A conversational expert system supporting bullying and harassment policies. In *the 2nd ICAART*, pages 163–168.

[Lieberman, Dinakar, and Jones2011] Lieberman, Henry, Karthik Dinakar, and Birago Jones. 2011. Let's gang up on cyberbullying. *Computer*, 44:93–96.

[Little et al.2003] Little, Todd D., Christopher C. Henrich, Stephanie M. Jones, and Patricia H. Hawley. 2003. Disentangling the "whys" from the "whats" of aggressive behavior. *Int'l J. of Behavioral Development*, 27:122–133.

[Liu et al.2010] Liu, Xiaohua, Kuan Li, Bo Han, Ming Zhou, Long Jiang, Zhongyang Xiong, and Changning Huang. 2010. Semantic role labeling for news tweets. In *the 23rd COLING*, pages 698–706.

[Màrquez et al.2005] Màrquez, Lluís, Pere Comas, Jesús Giménez, and Neus Català. 2005. Semantic role labeling as sequential tagging. In *the 9th CoNLL*, pages 193–196.

[Mihalcea and Strapparava2005] Mihalcea, Rada and Carlo Strapparava. 2005. Making computers laugh: Investigations in automatic humor recognition. In *EMNLP 2005*, pages 531–538.

[Moore et al.2003] Moore, Mark H., Carol V. Petrie, Anthony A. Braga, and Brenda L. McLaughlin. 2003. *Deadly lessons: Understanding lethal school violence*. The National Academies Press, Washington, DC.

[Nansel et al.2001] 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. *J. Amer. Medical Assoc.*, 285(16):2094–2100.

[Nishina and Bellmore2010] Nishina, Adrienne and Amy D. Bellmore. 2010. When might aggression, victimization, and conflict matter most?: Contextual considerations. *J. of Early Adolescence*, pages 5–26.

[Nishina and Juvonen2005] Nishina, Adrienne and Jaana Juvonen. 2005. Daily reports of witnessing and experiencing peer harassment in middle school. *Child Development*, 76:435–450.

[Nylund et al.2007] 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.

[Olweus1993] Olweus, Dan. 1993. *Bullying at school: What we know and what we can do*. Blackwell, Oxford, UK.

[Pang and Lee2004] Pang, Bo and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *the 42nd ACL*, pages 271–278.

[Ptaszynski et al.2010] Ptaszynski, Michal, Pawel Dybala, Tatsuaki Matsuba, Fumito Masui, Rafal Rzepka,

and Kenji Araki. 2010. Machine learning and affect analysis against cyber-bullying. In *the 36th AISB*, pages 7–16.

[Punyakanok, Roth, and Yih2008] Punyakanok, Vasin, Dan Roth, and Wen-tau Yih. 2008. The importance of syntactic parsing and inference in semantic role labeling. *Comput. Linguist.*, 34(2):257–287.

[Ratinov and Roth2009] Ratinov, Lev and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *the 13th CoNLL*, pages 147–155.

[Ritter et al.2011] Ritter, Alan, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *EMNLP 2011*, pages 1524–1534.

[Rivers et al.2009] 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.

[Salmivalli1999] Salmivalli, Christina. 1999. Participant role approach to school bullying: Implications for intervention. *J. of Adolescence*, 22(4):453–459.

[Schmidt, Fung, and Rosales2007] Schmidt, Mark W., Glenn Fung, and Rómer Rosales. 2007. Fast optimization methods for l1 regularization: A comparative study and two new approaches. In *the 18th ECML*, pages 286–297.

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

[Settles2011] Settles, Burr. 2011. Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. In *the EMNLP 2011*, pages 1467–1478.

[Smith, Madsen, and Moody1999] 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.

[The American Academy of Pediatrics2009] The American Academy of Pediatrics. 2009. Policy statement–role of the pediatrician in youth violence prevention. *Pediatrics*, 124(1):393–402.

[The White House2011] 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.

[Toutanova et al.2003] Toutanova, Kristina, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *NAACL-HLT 2003*, pages 173–180.

[Vaillancourt et al.2010] 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. *J. of School Violence*, 9:233–250.

[Vandebosch and Cleemput2009] Vandebosch, Heidi and Katrien Van Cleemput. 2009. Cyberbullying among youngsters: profiles of bullies and victims. *New media & society*, 11(8):1349–1371.

[Wang, Iannotti, and Nansel2009] Wang, Jing, Ronald J. Iannotti, and Tonja R. Nansel. 2009. School bullying among adolescents in the united states: Physical, verbal, relational, and cyber. *J. Adolescent Health*, 45(4):368–375.