Detecting Cyber-bullying from Sparse Data and Inconsistent Labels

Sabina Tomkins Lise Getoor Yunfei Chen Yi Zhang
UC Santa Cruz
satomkin@ucsc.edu {getoor,ychen,yiz}@soe.ucsc.edu

Abstract

Cyber-bullying threatens the emotional, psychological and even physical health of up to 72% [6] of youth. A first step in combating this problem is to automatically detect incidents of cyber-bullying. Detecting incidents of cyber-bullying in text is typically cast as a binary classification problem and requires laborious manual labeling. Unfortunately this hand labeling poses two issues: it is expensive and prone to inconsistency. In this work we address both of these issues. We propose a linguistic model which uses domain knowledge to drastically reduce the number of parameters compared to standard bag-of-words approaches and is thus more appropriate for learning from limited labeled data. Rather than discard inconsistent labels we evaluate several methods for learning from them, demonstrating that incorporating uncertainty allows for better generalization.

1 Introduction

Bullying has long presented physical, emotional, and psychological risks to children, youth, and adults nationwide. As such there is an extensive body of knowledge aimed at understanding and preventing bullying. Far less is known about the newest form of interpersonal aggression: cyber-bullying. Cyber-bullying occurs in an electronic environment [9], from online forums, to social media platforms such as Twitter and Facebook. As it can occur at any time or location, cyber-bullying poses new risks, while also influencing well-being in the classroom [9, 6, 7]. It also introduces new questions of governance and enforcement, as it is less clear in an online environment who can and should police harmful behavior. It is critical to understand cyber-bullying as a new kind of human behavior, and as an extension of bullying which happens in the physical world.

One necessary first step in understanding and preventing cyber-bullying is to detect it, and a particular goal is to be able to automatically flag potentially harmful *social media* messages. However, social media messages introduce unique problems. As they are unusually short, and rife with misspellings and slang, when treated with traditional text cleaning these messages are often stripped to only one or two words. This sparsity, coupled with the cost of obtaining high quality labels, makes them especially ill-suited for bag-of-word approaches which depend on sufficient training data to generalize well. Not only is labeled data costly, but it can be error-prone as annotators are generally third parties who are not directly involved with the incidents of cyber-bullying. Thus their labeling is subjective, and even labels with high inter-annotator agreement may be incorrect.

We propose a novel modeling framework which utilizes collective reasoning and domain knowledge to detect cyber-bullying. Rather than throwing out annotations with low inter-annotator agreement, we propose a set of probabilistic linguistic models which can directly incorporate uncertainty. To overcome the challenges with learning from low-quality tweets, we develop a linguistic model which uses domain knowledge. We demonstrate that models which use domain knowledge are better able to detect cyber-bullying in tweet messages than those that rely on n-grams. Furthermore, we show that

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

modeling uncertainty in the training data can improve the F-Measure of all models, demonstrating that a probabilistic approach is well suited for this domain.

2 Detecting Cyber-Bullying

The problem of determining if a message contains bullying content is typically formulated as a binary classification task, which takes as input a vector of observed variables $\boldsymbol{x}$ and outputs a vector of labels $\boldsymbol{y}$. Typically, $\boldsymbol{y}$ is treated as binary, however in this work we investigate the ability to learn from continuously valued labels representing the degree to which a message contains bullying content. To do so, we relax our labels into the $[0, 1]$ interval for training, and demarcate these training labels $\tilde{\boldsymbol{y}}$.

In this section we introduce three linguistic models with which to detect cyber-bullying. Furthermore we propose three methods for determining the labels used in training $\tilde{\boldsymbol{y}}$. We construct our models with Probabilistic Soft Logic, with which we can encode domain knowledge as well as dependencies between the target variables.

2.1 Probabilistic Soft Logic

We propose a probabilistic approach which can learn from noisy social media interactions, where text is typically short and feature vectors sparse. We define a joint probability distribution over bullying messages using a *hinge-loss* Markov random field (HL-MRF) [1]. HL-MRFs are a general class of conditional, continuous Markov random fields, which provide the advantage of highly-efficient inference while maintaining expressivity.

A HL-MRF describes the following conditional probability density function over vectors of observed, $\boldsymbol{x}$, and unobserved, $\boldsymbol{y}$, continuous random variables:

$$P(\boldsymbol{y}|\boldsymbol{x}) \propto exp\left(-\sum_{j=1}^{m} w_j \phi_j(\boldsymbol{y}, \boldsymbol{x})\right)$$

where ϕ_j is a *hinge-loss* potential, $\phi_j = \max\{l_j(\boldsymbol{x}, \boldsymbol{y}), 0\}^p, p \in \{1, 2\}$, l_j is a linear function of $\boldsymbol{x}$ and $\boldsymbol{y}$ and w_j is the positive weight associated with ϕ_j.

To specify a HL-MRF, we use the templating language Probabilistic Soft Logic (PSL). In PSL, domain knowledge is encoded as weighted rules that capture dependences between both the input and output variables. These rules translate into the weighted potential functions ϕ. For example in our model, we have a rule which says that if two messages are similar, and one contains bullying content, then the other one may be likely to as well. To express this rule we introduce the predicate Similar, which takes two messages as arguments and whose truth value is the similarity between those messages. Additionally, we introduce the predicate BullyingContent, which takes a message as an argument and whose truth value indicates bullying. Using these predicates, and a weight w_{sim}, we define our rule in PSL as follows:

$$w_{sim} : \text{Similar}(T_a, T_b) \wedge \text{BullyingContent}(T_a) \Rightarrow \text{BullyingContent}(T_b)$$

Together a predicate and its arguments form a logical atom; unlike in Boolean logic PSL atoms can assume soft truth values in $[0, 1]$, allowing us to express the uncertainty inherent in annotations. When supplied with data a PSL model defines a unique HL-MRF where atoms represent either observed ($\boldsymbol{x}$) or unobserved random variables ($\boldsymbol{y}$) in the probabilistic model. As MAP inference in a HL-MRF can be formulated as a convex problem, we can tractably and efficiently infer the values to $\boldsymbol{y}$ or $\tilde{\boldsymbol{y}}$ with the PSL software[1], using the alternating direction method of multipliers ADMM [2]. Next we demonstrate how PSL can be used to template linguistic models for cyber-bullying detection.

2.2 Linguistic Patterns of Cyber-Bullying

A question in developing linguistic models of cyber-bullying detection is how exactly to model the words which occur in text. Here we propose two approaches. The first is an *N-Grams* model, which learns the correlation between each n-gram and the bullying content of a message. We compare

[1] `http://psl.linqs.org`

this to the *Seed Phrases* model, which correlates phrases to bullying, and differentiates between subjects of attacks to detect harmful messages. As each of these models has potential weaknesses, we combine them into the *Seeds++* model, which contains all rules from both models to benefit from their respective strengths.

Learning from social media data poses its own unique challenges. Messages are typically shorter than many other natural language documents, and tweets are severely limited to be within 140 characters. Furthermore, the difficulty of extracting linguistic signal from short tweets is compounded by their proclivity to contain misspellings and slang.

Model 1: N-Grams The full *N-Grams* model consists of the rules in Table 1 and Table 2. To address class imbalance we introduce a prior in all models, $\neg$BullyTweet(T), which captures that the majority of messages do not contain bullying content. For each word we instantiate a weighted rule correlating the presence of this word within a tweet to whether or not it is a bullying tweet. By training these weights we learn which words indicate bullying content. Here we consider words to be unigrams and bigrams.

Model 2: N-Grams++ The *N-Grams++* model contains the rules in Table 1 through Table 3. To overcome the sparsity of the feature vectors in the N-Grams model we propose two advanced features: sentiment and a document embedding similarity. Sentiment is assessed at the document level and provides some signal even from otherwise information deficient tweets. As bullying messages can be highly charged, we model the valence of a tweet with SentiTweet(T), where "Senti" is either Negative, Positive, or Neutral. Additionally, we learn distributed representations of each tweet, allowing us to abate some issues of sparsity. By mapping tweets to an embedding space, using any common text embedding methods, we can encode the relationship that documents which are close to each other in that space should have similar labels. To do so we introduce the predicate Similar(T_i, T_j) whose truth value is the cosine similarity between the embeddings of T_i and T_j respectively, scaled to be in [0,1]. By modeling similarity we can explicitly express dependencies between our target variables, thus benefiting from collective inference.

The disadvantage of the *N-Grams* and *N-Grams++* models is that the weight relating each word to bullying must be tuned with training data, which is commonly sparse with some words appearing in only a few documents. This sparsity can make learning from a small corpus difficult, and can hamper generalization to unseen data. A model which exploits domain knowledge to reduce the number of free parameters may be better suited to this task.

w_{nb} :	$\neg$BullyTweet(T)

Table 1: *Class Imbalance Prior*

w_i :	HasWord(T, w_i) $\Rightarrow$ BullyTweet(T)

Table 2: *N-Grams*

w_p :	SentiTweet(T) $\Rightarrow$ BullyTweet(T)
w_c :	BullyTweet(T_i) $\wedge$ Similar(T_i, T_j) $\Rightarrow$ BullyTweet(T)

Table 3: *Sentiment and Document Similarity*

Model 3: Seeds++ To this end we propose *Seeds++*, consisting of the rules in Table 1, Table 3, and Table 4, which relates certain phrases to bullying. Van Hee et al. [15], found seven different categories of bullying messages. In our data we found evidence of three: name calling, threatening, and sexual remarks, as well as a fourth category we refer to as silencing. Each category contains a small set of phrases, for example the insult category contains the phrase, "fat", an example threat phrase is "hurt you", a sexual attack might contain the word "slut" and the silencing category contains the phrase "shut up".

Furthermore, we differentiate between attacks made at individuals and third parties with two rules which express if a message contains a direct second or indirect third person reference. For example, "you" is a second person reference, while "they" is a third person reference. The rule set for *Seeds* is specified in Table 4, though this model also contains the rules from Table 1.

w_n :	Insult(T) $\Rightarrow$ BullyTweet(T)
w_t :	Threat(T) $\Rightarrow$ BullyTweet(T)
w_x :	Sexual(T) $\Rightarrow$ BullyTweet(T)
w_s :	Silencing(T) $\Rightarrow$ BullyTweet(T)
w_{m2} :	Direct(T) $\wedge$ Insult(T) $\Rightarrow$ BullyTweet(T)
w_{m3} :	InDirect(T) $\wedge$ Insult(T) $\Rightarrow$ $\neg$BullyTweet(T)

Table 4: *Seed Phrases*

Model 4: Combined Finally, we combine these rules into a complete model *Combined*, which consists of all of the rules in Table 1 through

Table 4. This model allows us to benefit from the strengths of all rules. All models are evaluated empirically in Section 3.

2.3 Uncertain Annotations

As it is difficult and costly to acquire high-quality annotations of textual data, we explore the possible benefits of learning from low certainty annotations. There are two possible forms of uncertainty in this setting, disagreement between annotators and the uncertainty of individual annotators. Here we have three annotators where each can label a tweet with a 0 (not bully), 1 (maybe bully) and 2 (bully).

Let a be the normalized value assigned by the three annotators, that is $\frac{1}{6}\sum_{i=1}^{3} a_i$, where a_i is the label of the i-th annotator, and $\tilde{y}$ be the final label used in training. We explore three methods for determining $\tilde{y}$. In the *Discrete* method we discard all labels without high inter-annotator agreement. That is, we round all labels such that if $a \geq \frac{2}{3}$, $\tilde{y} = 1$ and $a \leq \frac{1}{3}$, $\tilde{y} = 0$ and all $\frac{2}{3} > a > \frac{1}{3}$ are discarded. We also introduce an alternate method, *Soft*, where $\tilde{y} = a$, which allows us to use all of the information provided by the annotation. In the *Hybrid* method, where there is high inter-annotator agreement we treat the label as discrete, and set $\tilde{y}$ exactly as in the discrete method. Yet when all annotators are uncertain, or when all three disagree, such that $\frac{2}{3} > a > \frac{1}{3}$, we set $\tilde{y} = a$.

3 Empirical Evaluation

We investigate two questions in this section: which linguistic model can best learn from limited labeled data and which labeling method is most useful in model training. We analyze the performance of five models: *N-Grams*, and *N-Grams++* augmented with collective rules and sentiment features, a domain inspired model *Seeds++*, and *Combined* which integrates strengths from all models. Additionally we compare to a Support Vector Machine (SVM) [4] [2].To inspect the effect of uncertainty, we compare three methods of handling inconsistent labels: *Discrete*, *Soft* and *Hybrid*.

3.1 Data

Here we explore cyber-bullying on Twitter, a popular social media platform. We focus on young children and adolescents as they represent a particularly at-risk group [14] and collect tweets by or referencing users under the age of 18. A total of 1798 tweets were annotated by three students, who were asked to mark each message according to how they felt reading it, with 0, 1, and 2 indicating no, maybe and definitely bullying respectively. The Fliess inter-annotator agreement was .066.

The tweets were cleaned according to standard natural language processing practices. Stop words were removed, as were numbers, and non-English words. Words with repeat characters were trimmed, for example “haaappy” became “happy”. Only those words which appeared in at least 30 documents were retained. Sentiment was assigned with the open-source python tool VADER [5].

To calculate the similarity between two tweets we compute the cosine similarity according to a trained Doc2Vec model [8]. There is a trade-off in document embedding models, where a small domain-specific corpus may not have enough content to properly learn the embedding space yet a publicly available corpus may not fully capture the nuances of a specific domain, such as exchanges among adolescents. For this reason we train a Doc2Vec model on our corpus using Gensim [12], but seed it with pre-trained word vectors[3]. To seed the model we used the openly available Glove [10] Word2Vecs, which were trained on Twitter, and are thus appropriate for this domain.

All models are trained using 5-fold cross validation on 90% of the data. In all folds we maintain a distribution of 30% bully tweets. The reported results are on the final held-out test set of 10% of the data. Here we compare five models, an SVM, and four PSL models: the *N-Grams*, *N-Grams++*, *Seeds++* and *Combined*. When detecting harmful incidents high recall is desirable, especially in situations where predictions are passed to humans for final inspection. However, we also value precision as it could be costly and dangerous to mislabel incidents of aggression. Hence, the F-Measure gives a balanced perspective of the trade-off between precision and recall and is used often in this task, and is more informative than accuracy when there is class imbalance in the data. To evaluate we round all soft predictions to 0 or 1.

[2] We used the implementation in the python package scikit-learn

[3] To initialize the Doc2Vec model with Word2Vecs we use: `https://github.com/jhlau/doc2vec`

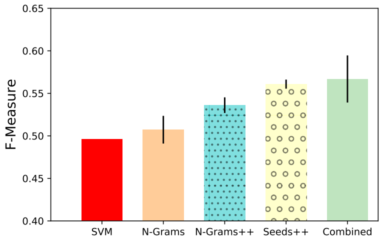

Figure 1: Adding collective rules improves the N-Grams model, while seed phrases are useful overall.

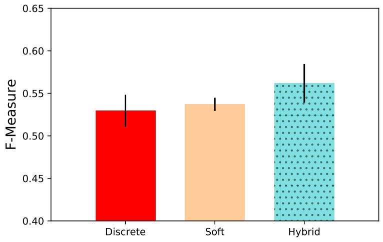

Figure 2: Uncertainty is helpful, with the hybrid labeling method being the most effective

3.2 Results

The PSL models all achieved a higher F-Measure than the SVM baseline. We see that the *Seeds++* model obtained better F-Measure than any of the *N-Grams* models. This is striking, as only a few seed words and phrases (43 total across all categories) were sufficient to obtain these results. Combining all models obtained the best overall F-Measure, with a statistically significantly higher precision recall and accuracy as well, suggesting that the models can adapt to different aspects of the data.

We also see that utilizing uncertainty in any form is beneficial. However a hybrid approach yields the best results. As annotations are often inconsistent, and annotators may not have high confidence in situations in which they have no personal knowledge, it is interesting to see that even uncertain information can be useful.

4 Related Work

There is a body of work on developing textual features for detecting cyber-bullying with linguistic classifiers [3, 17, 13, 15, 16]. Using TF-IDF features and contextual features Yin et al. [17] predict bullying with high accuracy. Also using context, Chen et al. [3] develop a novel framework which incorporates unique style features and structure.

Similar to our work Raisi and Huang [11] use a small seed vocabulary to indicate bullying events. They leverage participant roles to expand the set of candidate bullying terms, and better predict bullying. Their approach corroborates the idea that seed indicators can be successful in this task. Also in a similar vein to our work, Reynolds et al. [13] compare a rule based approach to a Support Vector Machine, and find that the rule-based model obtains higher accuracy. Finally, our work has been highly motivated by Van Hee et al. [15] who go beyond binary classification to label events as belonging to one of several textual categories pertaining to bullying conversations.

We build on existing work by building rich textual features to detect cyber-bullying. Unlike existing approaches we exploit a collective setting to leverage document similarity. Like [11] we employ a small set of seed words, which we develop from the categories of [15] for a domain inspired model.

5 Conclusion

Detecting cyber-bullying is a critical task in today's inter-connected world, where the virtual manifestations of relationships have strong bearings on the emotional health of all parties. Traditional supervised machine learning approaches, reliant on large amounts of labeled data may suffer in this domain, where high quality annotations are expensive, and feature vectors are particularly sparse. To address these issues we propose a linguistic model which uses domain knowledge to detect harmful messages, obtaining better F-Measure than N-Grams variants and a SVM baseline. We investigate the utility of uncertain annotations, discovering that incorporating soft labels can improve performance.

Acknowledgements

This work is supported by the National Science Foundation under Grant No. 1703281.

References

[1] Stephen H. Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. Hinge-loss markov random fields and probabilistic soft logic. *Journal of Machine Learning Research (JMLR)*, 2017. To appear.

[2] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 2011.

[3] Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. Detecting offensive language in social media to protect adolescent online safety. SOCIALCOM-PASSAT, 2012.

[4] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Mach. Learn.*, 20, 1995.

[5] Clayton J. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *ICWSM*, 2014.

[6] Jaana Juvonen and Elisheva F. Gross. Extending the school grounds? – Bullying experiences in cyberspace. *Journal of School Health*, 2008.

[7] Catarina Katzer, Detlef Fetchenhauer, and Frank Belschak. Cyberbullying: Who Are the Victims? *Journal of Media Psychology*, 2009.

[8] Quoc V Le and Tomas Mikolov. Distributed representations of sentences and documents. In *ICML*, 2014.

[9] Qing Li. New bottle but old wine: A research of cyberbullying in schools. *Computers in Human Behavior*, 2007.

[10] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.

[11] Elaheh Raisi and Bert Huang. Cyberbullying detection with weakly supervised machine learning. *ASONAM*, 2017.

[12] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 2010.

[13] Kelly Reynolds, April Kontostathis, and Lynne Edwards. Using machine learning to detect cyberbullying. ICMLA, 2011.

[14] Robert Slonje, Peter K Smith, and Ann FriséN. The nature of cyberbullying, and strategies for prevention. *Computers in human behavior*, 2013.

[15] Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Veronique Hoste. Detection and fine-grained classification of cyberbullying events. In *Proceedings of Recent Advances in Natural Language Processing, Proceedings*, 2015.

[16] Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. Learning from bullying traces in social media. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT, 2012.

[17] Dawei Yin, Brian D. Davison, Zhenzhen Xue, Liangjie Hong, April Kontostathis, and Lynne Edwards. Detection of Harassment on Web 2.0. In *Content Analysis in the Web 2.0 (CAW2.0)*, 2009.