2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

A Socio-linguistic Model for Cyberbullying Detection

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

***Abstract*—Cyberbullying is a serious threat to both the short and long-term well-being of social media users. Addressing this problem in online environments demands the ability to automatically detect cyberbullying and to identify the roles that participants assume in social interactions. As cyberbullying occurs within online communities, it is also vital to understand the group dynamics that support bullying behavior. To this end, we propose a socio-linguistic model which jointly detects cyberbullying content in messages, discovers latent text categories, identifies participant roles and exploits social interactions. While our method makes use of content that is labeled as bullying, it does not require category, role or relationship labels. Furthermore, as bullying labels are often subjective, noisy and inconsistent, an important contribution of our paper is effective methods for leveraging inconsistent labels. Rather than discard inconsistent labels, we evaluate different methods for learning from them, demonstrating that incorporating uncertainty allows for better generalization. Our proposed socio-linguistic model achieves an 18% improvement over state-of-the-art methods.**

Introduction

Bullying has long presented physical, emotional and psychological risks to children, youth and adults. As such, there is an extensive body of knowledge aimed at understanding and preventing bullying. Far less is known about cyberbullying, the newest form of interpersonal aggression. Cyberbullying occurs in an electronic environment [1], from online forums to social media platforms such as Twitter and Facebook. As it can occur at any time or location, cyberbullying poses new psychological risks, while also influencing physical well being [1], [2], [3]. 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.

A necessary first step in understanding and preventing cyberbullying is detecting it, and here our goal is to automatically flag potentially harmful *social media* messages. These messages introduce unique challenges for natural language processing (NLP) techniques. As they are unusually short and rife with misspellings and slang, when treated with traditional text pre-processing, these messages can be stripped to only one or two words. This sparsity makes cyberbullying messages especially ill-suited for methods which depend on sufficiently large training corpora to generalize well. Solutions which can augment poor textual data with domain knowledge or social data might outperform those which rely on text alone.

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 cyberbullying. Thus their labeling is subjective, and even labels with high inter-annotator agreement may be incorrect. Rather than throwing out annotations with low inter-annotator agreement, we propose a series of probabilistic models which can directly incorporate uncertainty. Furthermore, we show that modeling uncertainty in the training data can improve the performance of all models, demonstrating that a probabilistic approach is well suited for this domain.

We develop a series of probabilistic models of increasing sophistication. We build these models with Probabilistic Soft Logic (PSL) [4], a recently introduced highly scalable probabilistic modeling framework. Our first model makes use of text, sentiment and collective reasoning. Next, we incorporate seed-words and latent representations of text categories. Finally, we make use of social information by inferring relational ties and social roles.

Our models are evaluated on a dataset of youth interactions on the social media platform Twitter. Twitter has emerged as a fertile environment for bullying [5]. Twitter's ability to provide a veil of anonymity can facilitate bullying [6], [7]. Whittaker and Kowalski [8] found that though fewer survey participants used Twitter (69.4%) compared to Facebook (86.5%), a higher percentage of participants experienced cyberbullying on Twitter (45.5%) than Facebook (38.6%) (and other platforms). We compare our models to a baseline N-Grams model. This model is comparable to standard bag-of-n-grams approaches. Additionally, we compare to an implementation of a state-of-the-art approach.

Our contributions include strategies for learning from uncertain annotations and linguistic models which demonstrate the utility of domain knowledge and collective reasoning. In addition, we introduce two novel latent-variable models which categorize attacks and identify user roles by exploiting inferred relational ties. These models are advantageous in that they combat the sparsity of textual data and provide interpretation of latent phenomena, such as relational power dynamics, in cyberbullying.

Modeling Preliminaries

In order to model the intricate dependencies between language and participants in social interactions, we propose a collective probabilistic approach. We construct our models with Probabilistic Soft Logic[1] [4], a probabilistic programming

IEEE/ACM ASONAM 2018, August 28-31, 2018, Barcelona, Spain
978-1-5386-6051-5/18/$31.00

[1]Open-source software available at: http://psl.linqs.org

framework which offers several advantages for this domain: annotators' uncertainty can be fully represented with random variables in $[0, 1]$, logical rules provide an intuitive formulation of socio-linguistic structure, and collective inference of both target and latent variables is highly efficient.

In PSL domain knowledge is encoded with weighted logical rules. These rules can describe complex relationships and, crucially, capture dependencies not only from observed features to target variables, but *between* target variables. This expressivity allows us to encode both linguistic and social relationships between social media messages and users. Finally, PSL provides an intuitive framework for representing latent abstractions and an efficient procedure for inferring their values.

To illustrate PSL in the cyberbullying context, consider 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 which expresses their similarity as a value between 0 and 1. Similarity can be described with many measures; for example, as the cosine distance between document embeddings or as the Jaccard distance between one-hot representations of the documents. In our linguistic models, we define similarity as the cosine distance between documents in a learned embedding space, while in our latent models, we introduce a range of similarity-like measures. Additionally, we introduce the predicate BULLYINGCONTENT, which takes a message as an argument and whose truth value indicates bullying. Together a predicate and its arguments form a logical atom; unlike in Boolean logic, PSL atoms can assume soft truth values in $[0, 1]$. With these predicates and a weight w_{sim} which reflects the relative importance of this rule, we define our rule in PSL as follows:

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

Combined with data, a PSL model defines a joint probability distribution over messages. This distribution is expressed with a *hinge-loss* Markov random field (HL-MRF) [4], a general class of conditional, continuous probabilistic graphical models. HL-MRFs provide the advantage of both high efficiency and expressivity. Formally, a HL-MRF describes the following 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.

The logical rules defined in a PSL model translate to the weighted potentials ϕ, where atoms represent either observed ($\boldsymbol{x}$) or unobserved random variables ($\boldsymbol{y}$). As maximum a posteriori (MAP) inference in a HL-MRF can be formulated as a convex problem, we can tractably and efficiently infer the values to $\boldsymbol{y}$. The optimization technique is based upon the alternating direction method of multipliers ADMM [9]. Next we demonstrate how PSL can be used to template models for cyberbullying detection.

PROBABILISTIC CYBERBULLYING DETECTION MODELS

We propose a series of five probabilistic models. The first three models: N-GRAMS, N-GRAMS++ and SEEDS++, employ linguistic information to detect cyberbullying in text. The next two models: LATENT LINGUISTIC and LATENT SOCIO-LINGUISTIC, utilize latent abstractions to categorize text and label roles. Additionally, LATENT SOCIO-LINGUISTIC infers relational ties.

Linguistic Models

In each model, our goal is to find the MAP assignment to $\boldsymbol{y}$. To do so, we introduce $\textsc{BullyTweet}(T)$ which takes a tweet T, as an argument and whose truth value reflects the extent to which this tweet contains bullying content. By inferring the truth values of $\textsc{BullyTweet}(T)$ for all tweets, we are finding the MAP assignment to $\boldsymbol{y}$.

N-GRAMS: The N-GRAMS model consists of the rules in Table I. To address class imbalance, we introduce a prior in all models, $\neg\textsc{BullyTweet}(T)$ which captures that the majority of messages do not contain bullying content. For each n-gram, we instantiate a weighted rule correlating the presence of this n-gram within a tweet to the extent to which it is a bullying tweet. By training these weights, we learn which n-grams indicate bullying content. Here we consider n-grams to be unigrams and bigrams.

N-GRAMS++: The N-GRAMS++ model contains the rules in Table I and Table II. 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 tweet level and provides some signal even from otherwise uninformative tweets. As bullying messages can be highly charged, we model the valence of a tweet with three predicates: NEGTWEET, POSTWEET and NEUTWEET, respectively expressing the negativity, positivity, and neutrality of a tweet. 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 method, 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 $\textsc{Similar}(T_i, T_j)$ 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.

w_i : $\textsc{HasWord}(T, ng_i) \Rightarrow \textsc{BullyTweet}(T)$

TABLE I: *N-Grams*

w_{neg} : $\textsc{NegTweet}(T) \Rightarrow \textsc{BullyTweet}(T)$
w_{pos} : $\textsc{PosTweet}(T) \Rightarrow \neg\textsc{BullyTweet}(T)$
w_{neu} : $\textsc{NeuTweet}(T) \Rightarrow \neg\textsc{BullyTweet}(T)$
w_c : $\textsc{BullyTweet}(T_i) \wedge \textsc{Similar}(T_i, T_j) \Rightarrow \textsc{BullyTweet}(T)$

TABLE II: *Sentiment and Document Similarity*

The disadvantage of N-GRAMS and N-GRAMS++ is that the weight relating each word to bullying must be tuned

with training data, which is commonly sparse because many words appear in only a few documents. This sparsity can make learning from a small corpus difficult and can hamper generalization to unseen data. While our N-GRAMS++ model mitigates this to some extent, a model which exploits domain knowledge to reduce the number of free parameters may be better suited to this task.

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_{m_2} : HASWORD(T, W) $\wedge$ DIRECTMENTION(W) $\wedge$ INSULT(T)	$\Rightarrow$ BULLYTWEET(T)
w_{m_3} : HASWORD(T, W) $\wedge$ INDIRECTMENTION(W) $\wedge$ INSULT(T)	$\Rightarrow \neg$BULLYTWEET(T)

TABLE III: *Seed Phrases*

SEEDS++: We propose SEEDS++ to address the limitations in correlating n-grams to bullying content with short social media messages. SEEDS++ consists of the rules in Table II and Table III, which relates certain phrases to bullying. Van Hee et al. [10] 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.

Furthermore, we differentiate between attacks directed at individuals and third parties with two rules. These rules specify if a message contains a direct second or indirect third person reference. For example, "you" is a second person reference and "they" is a third person reference. If a word W, (which can also be an n-gram), is a second person reference DIRECTIONMENTION(W) will be 1, and if it is a third person reference INDIRECTMENTION(W) will be 1.

Latent Variable Models

To model unobserved phenomena, we introduce two latent variable models. In the LATENT-LINGUISTIC model, the textual categories of messages and the roles of participants are modeled as latent variables. By describing messages with categories, and users with roles, we can relate roles to types of attacks, rather than expressly connecting roles to low-level linguistic cues.

Next, the SOCIO-LINGUISTIC model expresses unobserved relational ties between participants and connects these relationships to participant roles and textual categories. As the true strength of any relationship between two people is an unobservable quantity, we model relational ties with latent variables. Here, we treat these relational ties as indicators of friendship (although they could also represent other types of social ties) and predict the presence and strength of these ties both with social network and linguistic signals.

Like the linguistic models described earlier, these latent variable PSL models template HL-MRFs. One difference is that in the latent setting the joint probability distribution is defined over $\boldsymbol{x}$, $\boldsymbol{y}$ and $\boldsymbol{z}$, where $\boldsymbol{z}$ is a vector of latent variables. To perform weight-learning in the presence of latent variables, we use the method described by Bach et al. [11].

In the following models we make use of functional constraints in PSL. Consider the predicate CATEGORY(T, C) in $[0, 1]$ which is 1 if tweet T can be described by category C. We allow partial membership to categories while ensuring that a tweet cannot fully belong to more than one category. This is expressed in PSL with the following functional constraint:

$$\infty : \sum_{c \in Categories} \text{CATEGORY}(T, c) = 1.$$

This constraint has infinite weight and ensures that a tweet cannot completely belong to two categories simultaneously. In the following latent models, we use functional constraints in modeling users' roles in tweets, the text categories of tweets and the membership of words in text categories.

LATENT-LINGUISTIC: In this model we propose that each tweet can be described with a latent category; these categories can be thought of as speech or dialog acts [12] or as sub-topics. For example, we define three act types: $Attack$, $Teasing$ and $Other$. Furthermore, we define four kinds of attacks: name calling (N), sexual name calling (Sn), silencing (S) and threatening (Th). We use $Attack_i$, where i indexes into $\{N, Sn, S, Th\}$, to indicate the occurrence of each type of attack. We use $NotAttack_j$ to indicate the remaining categories, where j indicates one of teasing (Ts) or other (Ot).

∞ :	$\sum_{c \in Categories}$ TEXTCAT(T, c) $= 1$
∞ :	TEXTCAT(T, Ot)+TEXTCAT(T, Ts) $= 1-$BULLYTWEET(T)
$w_{a_i b}$:	TEXTCAT(T, $Attack_i$) $\Rightarrow$ BULLYTWEET(T)

TABLE IV: *Inferring Text Categories*

∞ :	$\sum_{c \in Categories}$ WORDINCAT(W, c) $= 1$
w_{wtc_i} :	CATEGORY$_i$(W) $\Rightarrow$ WORDINCAT(W, C)
w_{tda} :	WORDINCAT(W, C) $\wedge$ HASWORD(T, W) $\Rightarrow$ TEXTCAT(T, C)

TABLE V: *Words to Categories*

Let TEXTCAT(T, C) be a random variable in $[0, 1]$, such that it is 1 if the text category of T is C. In Table IV we introduce constraints which enforce useful dynamics for this task. The first rule is a functional constraint. This forms a probability vector for each tweet where the entries of the vectors are the text categories. This constraint allows mixed membership while ensuring that a tweet cannot completely belong to more than one category. In the second rule in Table IV, we relate the categories of *teasing* and *other* to non-bullying messages. This rule also sets the truth value of a bullying message to the combined truth values of each of the types of bullying attacks. This is useful as messages can be combinations of categories, for example, a message may include name calling and threats. In the final rule, we relate each type of attack to bullying messages.

As in SEEDS++, the latent models make use of seed words; however, these words now connect specific categories. Additionally, rather than correlating each word to bullying content, these categories are related to bullying. For each word, W, and text category C, we model the extent to which this word belongs in this category, WORDINCAT(W, C). For each word, we learn a probability vector over categories which is encoded with the functional constraint in Table V. The second rule in Table V relates known seed words to certain categories,

where $\text{CATEGORY}_i(W)$ is 1 if a seed word is known to be in CATEGORY_i. Instead of modeling that messages with seed words indicate bullying content, with the final rule in Table V, we model the categories of messages according to their constituent words.

$$w_{wsc} : \text{WORDINCAT}(W_i, C) \wedge \text{WORDSIM}(W_i, W_j) \Rightarrow \text{WORDINCAT}(W_j, C)$$
$$w_{wscc} : \text{WORDINCAT}(W_i, C) \wedge \text{SAMECLUSTER}(W_i, W_j) \Rightarrow \text{WORDINCAT}(W_j, C)$$
$$w_{wcoc} : \text{WORDINCAT}(W_i, C) \wedge \text{COOCCUR}(W_i, W_j) \Rightarrow \text{WORDINCAT}(W_j, C)$$
$$w_{wac} : \text{WORDINCAT}(W_i, C) \wedge \text{ASSOCIATED}(W_i, W_j) \Rightarrow \text{WORDINCAT}(W_j, C)$$

TABLE VI: *Word Associations*

To find potential words which may belong in a given category, we have four collective word rules, shown in Table VI. The first two rules find replacement words for seed words where replacements may have the same semantics as the seeds. For example, we consider the similarity between words in an embedding space with WORDSIM. We also utilize word clusters and consider words in the same cluster as seed words with SAMECLUSTER. We also consider related non-replacement words which may be commonly used *with* seed words. Each pair of words is assigned a co-occurence score, COOCCUR, which is the number of documents this pair occurs in, scaled to be in $[0, 1]$. Word associations can capture both replacements and associations, as words with similar conceptual meaning may be recalled in free association tasks [13]. Thus, we further expand potential category words to include words associated with seed words.

$$w_{gda_i} : \text{NEGTWEET}(T) \Rightarrow \text{TEXTCAT}(T, Attack_i)$$
$$w_{nda_j} : \text{NEUTWEET}(T) \Rightarrow \text{TEXTCAT}(T, NotAttack_j)$$
$$w_{pda_j} : \text{POSTWEET}(T) \Rightarrow \text{TEXTCAT}(T, NotAttack_j)$$

TABLE VII: *Sentiment of Text Categories*

Exactly as we used sentiment to suggest bullying content, we now model the relationship between sentiment and categories. In Table VII the argument $Attack_i$ refers to any of the four attack categories and we express that tweets with negatively charged content can be described as attacks, while those with positive or neutral content are either in the category *Other* or *Teasing*.

$$w_{m2_k} : \text{HASWORD}(T, W_i) \wedge \text{DIRECTMENTION}(W_i) \wedge \text{HASWORD}(T, W_j) \wedge \text{WORDINCAT}(W_j, Attack_k) \Rightarrow \text{DIALOGACT}(T, Attack_k)$$
$$w_{m3} : \text{HASWORD}(T, W_i) \wedge \text{INDIRECTMENTION}(W_i) \wedge \text{HASWORD}(T, W_j) \wedge \text{WORDINCAT}(W_j, Attack_k) \Rightarrow \text{DIALOGACT}(T, Ot)$$

TABLE VIII: *Subjects and Text Categories*

As in SEEDS++, we differentiate between attacks with second and third person mentions. As shown in Table VIII, those tweets with bullying words and second person references may be attacks. Alternatively, messages with third person references are less likely to be attacks.

In addition to predicting the bullying content of messages, we infer the participants' roles. We refer to users targeted in bullying tweets as victims and authors of those tweets as bullies. Let U be a user who is either mentioned in, or authors, a tweet. We infer each user's role in a tweet, $\text{ROLEINTWEET}(T, U, R)$, where R can be either a victim V, bully B, or other O. We focus on a user's role in a particular tweet as user's roles are often flexible and depend on the context [14]. For the following rules, consider the template:

$$w_{ac} : \text{AUTHOR}(T, U) \wedge \text{BULLYATTRIBUTE}(T) \Rightarrow \text{ROLEINTWEET}(T, U, B).$$

For all rules which follow the template above, the model also includes a rule derived from the corresponding template:

$$w_{mc} : \text{MENTIONS}(T, U) \wedge \text{BULLYATTRIBUTE}(T) \Rightarrow \text{ROLEINTWEET}(T, U, V).$$

$$\infty : \sum_{x \in roles} \text{ROLEINTWEET}(T, U, x) = 1$$
$$\infty : \text{ROLEINTWEET}(T, U_{author}, B) = \text{TEXTCAT}(T, N) + \text{TEXTCAT}(T, Sn) + \text{TEXTCAT}(T, Th)$$
$$\infty : \text{ROLEINTWEET}(T, U_{author}, B) = \text{TEXTCAT}(T, N) + \text{TEXTCAT}(T, Sn) + \text{TEXTCAT}(T, S)$$
$$w_{dab_i} : \text{AUTHOR}(T, U) \wedge \text{TEXTCAT}(T, Attack_i) \Rightarrow \text{ROLEINTWEET}(T, U, B)$$
$$w_{dao_j} : \text{AUTHOR}(T, U) \wedge \text{TEXTCAT}(T, NotAttack_j) \Rightarrow \text{ROLEINTWEET}(T, U, O)$$
$$w_{po} : \text{AUTHOR}(T, U) \wedge \text{POSUSER}(U) \Rightarrow \text{ROLEINTWEET}(T, U, O)$$
$$w_{nb} : \text{AUTHOR}(T, U) \wedge \text{NEGUSER}(U) \Rightarrow \text{ROLEINTWEET}(T, U, B)$$
$$w_s : \text{ROLEINTWEET}(T, U_a, B) \wedge \text{MENTIONS}(T, U_b) \wedge \text{AUTHOR}(T, U_a) \Rightarrow \text{ROLEINTWEET}(T, U_b, V)$$

TABLE IX: *User Roles*

Rather than restrict each user to take exactly one role in each tweet, we allow users to assume roles to varying degrees. We then ensure that the degree to which a user assumes a given role is related to the extent to which they take any other role. For example, if there is strong evidence suggesting that a user is a bully, they cannot also be a victim with high certainty. This hard constraint is expressed in the first line of Table IX. The second and third rules in Table IX are also functional constraints. These rules aggregate the categories of name calling and sexual name calling with threatening and silencing. This ensures that even if the certainty in any one of these categories is weak, if the aggregate certainty is high, we can detect bullying behavior. In the next two rules, we describe the relationship between each text category and role. When a tweet category is a kind of attack, the author is a bully and otherwise the author takes another role.

With the next two rules in Table IX, we correlate a user's propensity to be a bully or a victim with their past history. This decision is inspired by the findings of Hosseinmardi et al. [15]. For each user, we model the sentiment of their aggregate messages such that the truth value of $\text{POSUSER}(U)$ will be close to 1 if U is generally positive in their messages. When a user U is generally negative, the truth value of $\text{NEGUSER}(U)$ will be high. The final rule states that if a bully author mentions a user in a tweet, that user is a victim.

SOCIO-LINGUISTIC: In the final model, we infer relational ties between users and express a number of intuitions about how these ties might influence bullying behavior. We treat ties as positive relationships. For example, a tie between two users might indicate friendship or some other positive association. We model these ties with latent variables as their ground truth strength and nature is unobserved. We express the extent to which U_a is tied to U_b with $\text{TIE}(U_a, U_b)$.

The rules in Table X describe how we infer relational ties. In the first rule, we express that most users in a social network are likely to not have ties. Next, we propose that a user following another indicates a tie. The following rule states that users

may have ties to the ties of their ties. Next, we use linguistic information to predict ties with four rules. When a tweet's author mentions another user in a tweet that is teasing, or a category other than bullying, then those two users may be related. Additionally, if the tweet contains positive or neutral sentiment then the author and mentioned user may be related.

$$
\begin{aligned}
&w_{nf} : \neg\textsc{Tie}(U_a, U_b) \\
&w_{fo} : \textsc{Follows}(U_a, U_b) \Rightarrow \textsc{Tie}(U_a, U_b) \\
&w_{as} : \textsc{Tie}(U_a, U_b) \wedge \textsc{Tie}(U_b, U_c) \Rightarrow \textsc{Tie}(U_a, U_c) \\
&w_{tf} : \textsc{Author}(T, U_a) \wedge \textsc{Mentions}(T, U_b) \wedge \textsc{DialogAct}(T, Ts) \\
&\qquad\qquad \Rightarrow \textsc{Tie}(U_a, U_b) \\
&w_{of} : \textsc{Author}(T, U_a) \wedge \textsc{Mentions}(T, U_b) \wedge \textsc{DialogAct}(T, Ot) \\
&\qquad\qquad \Rightarrow \textsc{Tie}(U_a, U_b) \\
&w_{pf} : \textsc{Author}(T, U_a) \wedge \textsc{Mentions}(T, U_b) \wedge \textsc{PosSent}(T) \\
&\qquad\qquad \Rightarrow \textsc{Tie}(U_a, U_b) \\
&w_{nf} : \textsc{Author}(T, U_a) \wedge \textsc{Mentions}(T, U_b) \wedge \textsc{NeuSent}(T) \\
&\qquad\qquad \Rightarrow \textsc{Tie}(U_a, U_b)
\end{aligned}
$$

TABLE X: *Inferring Relational Ties*

We also model social conversational patterns by learning the extent to which users with relational ties use certain text categories. Additionally, we use these ties to discover teasing. Whenever both bullying and teasing terms occur in a message between users with ties, we suggest that that is a teasing, rather than a bullying message.

$$
\begin{aligned}
&w_{fc_i} : \textsc{Tie}(U_a, U_b) \wedge \textsc{Author}(T, U_a) \wedge \textsc{Mentions}(T, U_b) \\
&\qquad\qquad \Rightarrow \textsc{TextCat}(T, C_i) \\
&w_{ft_i} : \textsc{Tie}(U_a, U_b) \wedge \textsc{Author}(T, U_a) \wedge \textsc{Mentions}(T, U_b) \wedge \textsc{HasWord}(T, W_x) \\
&\quad \wedge \textsc{HasWord}(T, W_y) \wedge \textsc{WordInCat}(W_x, Attack_i) \wedge \textsc{WordInCat}(W_y, Ts) \\
&\qquad\qquad \Rightarrow \textsc{TextCat}(T, Ts)
\end{aligned}
$$

TABLE XI: *Ties and Conversation*

$$
\begin{aligned}
&w_{fv} : \textsc{Tie}(U_a, U_b) \wedge \textsc{Mentions}(T_a, U_a) \wedge \textsc{Mentions}(T_b, U_b) \\
&\qquad \wedge \textsc{RoleInTweet}(T_a, U_a, X) \Rightarrow \textsc{RoleInTweet}(T_b, U_b, X) \\
&w_{fb} : \textsc{Tie}(U_a, U_b) \wedge \textsc{Author}(T_a, U_a) \wedge \textsc{Author}(T_b, U_b) \\
&\qquad \wedge \textsc{RoleInTweet}(T_a, U_a, X) \Rightarrow \textsc{RoleInTweet}(T_b, U_b, X) \\
&w_{nb} : \textsc{Author}(T, U) \wedge \textsc{HighPopularity}(U) \Rightarrow \textsc{RoleInTweet}(T, U, B) \\
&w_{po} : \textsc{Mentions}(T, U) \wedge \textsc{HighPopularity}(U) \Rightarrow \neg\textsc{RoleInTweet}(T, U, V) \\
&w_{gu} : \textsc{Tie}(U_a, U_b) \wedge \textsc{Author}(T_a, U_a) \wedge \textsc{Author}(T_b, U_b) \\
&\qquad \wedge \textsc{RoleInTweet}(T_a, U_c, V) \Rightarrow \textsc{RoleInTweet}(T_b, U_c, V)
\end{aligned}
$$

TABLE XII: *Social Behavior*

Finally, peer pressure can have a large influence on youth, who are likely to adopt the behavior of their peers [16], [17]. Here we describe normative behavior with two rules which state that users who share a relational tie are likely to assume similar roles. Moreover, the role a participant takes may depend on their position within an exchange; e.g., whether they are mentioned in or are authoring messages. We restrain the association to friends assuming the same authorship role. To model that users with higher popularity may target users with low popularity, we introduce the predicates HIGHPOPULARITY. Additionally, we capture ganging-up behavior with the last rule in Table XII, where users who share a relational tie may target the same victim.

LEARNING FROM UNCERTAIN ANNOTATIONS

As it is difficult and costly to acquire high-quality annotations in the cyberbullying domain, 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 who label the tweets with a 0 (not bully), 1 (maybe bully) and 2 (bully).

Let a be the average normalized value, e.g. $\frac{1}{6}\sum_{i=1}^{3} a_i$, 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. Finally, we introduce a *Hybrid* method. In this method, when there is high inter-annotator agreement, we treat the label as discrete and set $\tilde{y}$ exactly as in the discrete method. When all annotators are uncertain, or when all three disagree, such that $\frac{2}{3} > a > \frac{1}{3}$, we set $\tilde{y} = a$.

EMPIRICAL EVALUATION

Here we explore cyberbullying on Twitter. We focus on youth and adolescents as they represent a particularly at-risk group [18], and collect tweets by or referencing users under the age of 18. In total, we collect approximately 4.5 million tweets. Of these, a subset were sampled for labeling according to the procedure of Chen et al. [19]. This resulted in a total of 1798 tweets which were labeled by three annotators. The annotators were asked to mark whether each message was a bullying message with 0, 1 and 2 indicating no, maybe and yes. The Fleiss inter-annotator agreement was .14. Approximately 27% of all tweets were labeled as bullying.

All tweets, including those which were not labeled, were used to construct social features, such as if a user mentioned another user, and to build global user features, such as the overall sentiment across a user's tweets. Each user's positive or negative score is assigned by computing the compound score of their entire message history. All users with a compound score less than -0.5 were counted as negative, while all users with a compound score greater than 0.5 were positive. The HIGHPOPULARITY scores are a function of the total number of times a user is mentioned. All users with a high mention count (>5) are considered to have high popularity. Independently from the tweet collection process, we additionally collected user demographics for users in the dataset. For each user, we query using their screen name to collect their followers and followees. By using non-labeled tweets to construct these features, we avoid the potential biases introduced in the selection process for labeling.

The tweets were cleaned according to standard text preprocessing practices. Stop words were removed, as were numbers and non-English words. Words with more than two repeat characters were trimmed, for example "haaappy" became "happy". Only those words which appeared in at least 5 tweets were retained. Sentiment was assigned with the open-source python tool VADER [20].

To calculate the similarity between two tweets, we compute the cosine similarity according to a trained Doc2Vec model

[21]. 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. For this reason, we train a Doc2Vec model on our corpus using Gensim [22] but seed it with pre-trained word vectors[2]. To seed the model, we used the openly available Glove [23] Word2Vecs which were trained on Twitter and are thus appropriate for this domain. The word associations were found using Nelson et al.'s [13] free association norms. Any word which had a positive forward associative strength (FSG) value, was added to the candidate pool of associated words for a given seed word. Clustered words were found using the hierarchical word clusters which were trained on Twitter data, as described in [24] and are published online[3]. To calculate the co-occurrence score, we count the frequency of word-pairs across all documents. For each word we calculate its co-occurrence score for all other words by dividing by the maximum co-occurrence count.

All models are trained using 5-fold cross validation on 80% 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 20% of the data. Here we compare five PSL models: the N-GRAMS, N-GRAMS++, SEEDS++, LATENT-LINGUISTIC and SOCIO-LINGUISTIC.

Additionally, we compare these to an implementation of Van Hee's state-of-the-art approach [10]. Unlike Van Hee, we do not include character level trigrams among the final features, as we did not find their inclusion to be helpful on the validation set. Also, we used VADER to calculate document level sentiment, rather than combining single word scores. Like Van Hee we included: positive, negative, neutral and a combined (compound) score. The classifier is implemented as a support vector machine (SVM) [25] using the Python package scikit-learn [26].

Results

The first evaluation we present is on the ability of the models to detect cyberbullying content in messages, evaluated with F-Measure[4]. Another comparison is between the three labeling strategies. Additionally, we evaluate the ability of the latent variable models to assign participant roles. We also discuss the textual categories discovered by the latent models and the traits of the discovered relational ties.

Detecting Cyberbullying: We compare the five PSL models to Van Hee's approach. To report the detection results, we round $\boldsymbol{y}$ to 0 or 1 values. In Fig. 1, we report the average F-Measure across all labeling strategies for each model (the SVM uses only the discrete strategy). We see that adding collective rules and sentiment, in N-GRAMS++, improves the performance of N-GRAMS, while the seed phrases in SEEDS++ are more powerful than N-GRAMS. The latent models are best at detecting cyberbullying, with SOCIO-LINGUISTIC achieving the highest F-Measure of 63.2.

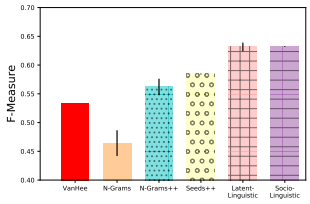

Fig. 1: Collective rules improve the N-GRAMS model, and SOCIO-LINGUISTIC achieves the best performance (bars are standard error).

In Table XIII, we average the results for each labeling strategy across all PSL models. Utilizing uncertainty in any form is beneficial and the Hybrid approach yields the best F-Measure.

	Precision	Recall	F-Measure
Discrete	48.0	70.7	56.0
Soft	47.1	**75.0**	**57.8**
Hybrid	46.9	**79.6**	**58.7**

TABLE XIII: The Soft and Hybrid methods provide statistically significant improvements (shown in bold) in F-Measure and recall over the discrete method.

Role Assignment: Here we consider the effect of social information on detecting roles. To do so, we compare LATENT-LINGUISTIC to SOCIO-LINGUISTIC. For the ground truth, each author of a bullying tweet is labeled as a bully; additionally each user who is mentioned in a bullying tweet is considered to be a victim. In Table XIV, we see that the SOCIO-LINGUISTIC model achieves the best F-Measure with predicting bullies and victims.

	Precision	Recall	F-Measure
	Bullies		
Latent Linguistic	55.4	58.8	57.2
Socio-Linguistic	54.1	**61.5**	**57.7**
	Victims		
Latent Linguistic	81.6	64.4	72.0
Socio-Linguistic	76.7	**70.2**	**73.3**

TABLE XIV: When assigning roles, SOCIO-LINGUISTIC achieves statistically significantly higher F-meaure and recall than LATENT-LINGUISTIC according to a paired t-test.

Roles and Text Categories: Here we inspect the relative frequencies of the bullying categories. To do so, we consider all tweets for which the sum of truth values for the bullying categories exceeds the sum of the truth values of the non-bullying categories. The most common category is name calling, with an average of 70.3% of all tweets predicted to be bullying belonging to this category. Sexual name calling was the second most common category with 24.6%. The remainder of the tweets were predominantly in the threatening category (4.5%), with 0.6% labeled as silencing.

Relational Ties: We analyze ties predicted with SOCIO-LINGUISTIC. We first look at the number of predicted ties. Next, we ask whether bullies have more ties on average and if bullies share ties with attack targets. Ties are discovered by SOCIO-LINGUISTIC using each of the three labeling strategies, and we report the results averaged across the three methods.

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

[3]http://www.cs.cmu.edu/~ark/TweetNLP/cluster_viewer.html

[4]Code online: https://bitbucket.org/linqs/socio-linguistic_cyberbullying

We inferred 131 ties among 421 users in the test set. The average number of ties per user was .40, while it was 1.5 for those users with at least one tie.

Here we consider authors of bullying tweets as bullies and users mentioned in those tweets as victims. Though this approach has shortcomings, it allows us to inspect characteristics of ties between users who have bullied and been victimized. In this approach we do not find that bullies have more ties. The average number of ties for bullies is .19, while it is .20 for victims. Of the ties, 9 were between bullies and victims, where victims had ties to bullies. In contrast, only 4.33 bullies (averaged across the three labeling strategies) had ties to victims, potentially reflecting that power dynamics influence relationships.

DISCUSSION

There is a clear benefit to leveraging document representations and classifying bullying content collectively, as seen in the improvement of N-GRAMS++ over N-GRAMS. Furthermore, when the feature vectors representing short messages are sparse, we see an advantage to using seed phrases (with SEEDS++) rather than n-grams.

In LATENT-LINGUISTIC we describe messages with fine-grained categories. These categories are seeded as in SEEDS++ and through a number of collective word rules, these categories can adaptively expand. This abstraction of messages to textual categories is useful in predicting cyberbullying messages and this is shown by the clear improvement from SEEDS++ to LATENT-LINGUISTIC. It is also useful for being able to interpret cyberbullying in more detail. For example, we see that name-calling is the primary form of cyberbullying in this dataset, followed by sexual name-calling.

Cyberbullying is a social activity, thus in SOCIO-LINGUISTIC we predict relational ties as well as participant roles. A primary question of this model is if we can infer these ties with limited social media data which does not contain explicit relationship links. We find that leveraging social information provides an improvement in performance over LATENT-LINGUISTIC. This suggests that the inferred ties might be meaningful. In denser social networks this improvement might be more pronounced.

By inferring relational ties we are able to better interpret group dynamics. For example, there is support for the theory that bullies are popular within peer-networks, although we did not see this theory reflected in the predicted number of bully ties. We did, however, see that victims were more likely to have ties to bullies than bullies to victims. This supports the idea that it may be socially advantageous to act positively towards bullies, who can hold positions of high social-status, while it is less socially advantageous to exhibit positive behavior towards members with low social-status, such as targets of bullying attacks. One persisting question is how different network structures might impact bully-victim relationships. We have found one example of the influence of social-status within a small discovered social network. Larger datasets might afford more discoveries into the nature of these complex relationships.

One limitation of our work is that we only investigated these dynamics on one social media platform. Yet, online interactions can be influenced by the platform which hosts them [27]. In future work we will broaden this investigation to additional social media platforms. Additionally, while the size of our dataset was comparable to those in published literature [5], we would surely learn more from a larger dataset. Finally, here we only consider the roles of participants in individual tweets. An important next step is to contextualize how users' roles vary and persist across situations.

RELATED WORK

Hodas et. al. analyzed network dynamics of Twitter users to address questions of friendship [28] and found that Twitter dynamics reaffirm well-studied principles of friendship. Kim et al. [29] studied 2193 pairs of users and found that their online interactions were impacted by offline friendships.

Current methods for detecting cyberbullying have largely relied on linguistic classifiers [30], [31], [32], [10], [33]. Using TF-IDF features and contextual features, Yin et al. [31] predict bullying with high accuracy. Chen et al. [30] develop a novel framework which incorporates unique style features and structure. Zhao et al. [34] also make use of word embeddings.

Similar to our work, Raisi and Huang [35] 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. Their work is also similar to ours in that they jointly infer roles and message labels. Critically, their work differs in that they learn from *unlabeled* training data. Also in a similar vein to our work, Reynolds et al. [32] compare a rule-based approach to a SVM and find that the rule-based model obtains higher accuracy. Like Huang et al. [36], we use social network information. The authors demonstrate that social features, such as network edge centrality, can improve classification. Our work differs from theirs in that we jointly infer relationship status while detecting messages, in addition to constructing social features as a pre-processing step. Finally, our work has been highly motivated by Van Hee et al. [10] who go beyond binary classification to label events as belonging to textual categories. We compare our models to an SVM implementation of Van Hee's approach.

Similar to existing linguistic models, we build rich textual features to detect cyberbullying. Additionally, we exploit dependencies and similarities between words, documents and categories using a collective approach to better classify messages and discover latent categories. A critical difference from other fine-grained approaches is that we do not need labels to detect fine-grained bullying categories. Like existing work, we infer participant roles and describe messages with textual categories. Our work differs from previous state-of-the-art in that we address four distinct challenges in one model which jointly infers: participant roles, latent textual categories and relational ties, while detecting cyberbullying messages.

Conclusion

Machine learning models face inherent challenges from social media data which can consist of short messages with misspellings and slang. Cyberbullying detection is made all the more difficult by a reliance on third-party annotators to acquire sufficient data for training. We address these concerns with two categories of models: domain-inspired linguistic models and a socio-linguistic model. The domain-inspired models combat sparsity by reducing the number of parameters which must be learned and by exploiting relations between words and documents collectively. Our socio-linguistic model is capable of inferring relationship ties from limited social media data while detecting cyberbullying. To the best of our knowledge, it is the first model in this domain which jointly infers bullying content, textual categories, participant roles and relationship links. By formulating these tasks jointly, we can learn from social dynamics to provide a statistically significant improvement in both cyberbullying detection and role assignment.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant Numbers CCF-1740850 and IIS-1703331. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] Q. Li, "New bottle but old wine: A research of cyberbullying in schools," *Computers in Human Behavior*, vol. 23, no. 4, 2007.

[2] J. Juvonen and E. F. Gross, "Extending the school grounds? – Bullying experiences in cyberspace," *Journal of School Health*, vol. 78, no. 9, 2008.

[3] C. Katzer, D. Fetchenhauer, and F. Belschak, "Cyberbullying: Who Are the Victims?" *Journal of Media Psychology*, vol. 21, no. 1, 2009.

[4] S. H. Bach, M. Broecheler, B. Huang, and L. Getoor, "Hinge-loss markov random fields and probabilistic soft logic," *Journal of Machine Learning Research (JMLR)*, vol. 18, no. 109, 2017.

[5] A. Bellmore, A. J. Calvin, J.-M. Xu, and X. Zhu, "The five Ws of bullying on Twitter: Who, What, Why, Where, and When," *Computers in Human Behavior*, vol. 44, no. C, 2015.

[6] P. Galán-García, J. G. d. l. Puerta, C. L. Gómez, I. Santos, and P. G. Bringas, "Supervised machine learning for the detection of troll profiles in twitter social network: application to a real case of cyberbullying," 2014.

[7] R. S. Tokunaga, "Following you home from school: A critical review and synthesis of research on cyberbullying victimization," *Computers in Human Behavior*, vol. 26, no. 3, 2010.

[8] E. Whittaker and R. M. Kowalski, "Cyberbullying via social media," *Journal of School Violence*, vol. 14, no. 1, 2015.

[9] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, 2011.

[10] C. Van Hee, E. Lefever, B. Verhoeven, J. Mennes, B. Desmet, G. De Pauw, W. Daelemans, and V. Hoste, "Detection and fine-grained classification of cyberbullying events," in *Recent Advances in Natural Language Processing (RANLP)*, 2015.

[11] S. H. Bach, B. Huang, J. Boyd-Graber, and L. Getoor, "Paired-dual learning for fast training of latent variable hinge-loss mrfs," in *International Conference on Machine Learning (ICML)*, 2015.

[12] J. R. Searle, *Expression and meaning: Studies in the theory of speech acts*. Cambridge University Press, 1985.

[13] D. L. Nelson, C. L. McEvoy, and T. A. Schreiber, "The university of south florida free association, rhyme, and word fragment norms," *Behavior Research Methods, Instruments, & Computers*, vol. 36, no. 3, 2004.

[14] C. Salmivalli, "Participant role approach to school bullying: implications for interventions," *Journal of Adolescence*, vol. 22, no. 4, 1999.

[15] H. Hosseinmardi, A. Ghasemianlangroodi, R. Han, Q. Lv, and S. Mishra, "Towards understanding cyberbullying behavior in a semi-anonymous social network," in *International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2014.

[16] K. E. Bauman and E. S. T., "On the importance of peer influence for adolescent drug use: commonly neglected considerations," *Addiction*, vol. 91, no. 2, 1996.

[17] C. J. Ferguson, C. S. Miguel, and R. D. Hartley, "A multivariate analysis of youth violence and aggression: The influence of family, peers, depression, and media violence," *The Journal of Pediatrics*, vol. 155, no. 6, 2009.

[18] R. Slonje, P. K. Smith, and A. Frisén, "The nature of cyberbullying, and strategies for prevention," *Computers in Human Behavior*, vol. 29, no. 1, 2013.

[19] Y. Chen, L. Zhang, A. Michelony, and Y. Zhang, "4is of social bully filtering: identity, inference, influence, and intervention," in *CIKM*, 2012.

[20] C. J. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text." in *International Conference On Web And Social Media (ICWSM)*, 2014.

[21] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents." in *International Conference on Machine Learning (ICML)*, 2014.

[22] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *LREC Workshop on New Challenges for NLP Frameworks*, 2010.

[23] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Conference on Empirical Methods on Natural Language Processing (EMNLP)*, 2014.

[24] O. Owoputi, B. O'Connor, C. Dyer, K. Gimpel, and N. Schneider, "Part-of-speech tagging for twitter: Word clusters and other advances," *Technical Report, Machine Learning Department. CMU-ML-12-107.*, 2012.

[25] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, 1995.

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, 2011.

[27] F. Giglietto, L. Rossi, and D. Bennato, "The open laboratory: Limits and possibilities of using facebook, twitter, and youtube as a research data source," *Journal of Technology in Human Services*, vol. 30, no. 3-4, 2012.

[28] N. O. Hodas, F. Kooti, and K. Lerman, "Friendship paradox redux: Your friends are more interesting than you." *International Conference On Web And Social Media (ICWSM)*, 2013.

[29] Y. Kim, F. Natali, F. Zhu, and E. Lim, "Investigating the influence of offline friendship on twitter networking behaviors," in *Hawaii International Conference on System Sciences (HICSS)*, 2016.

[30] Y. Chen, Y. Zhou, S. Zhu, and H. Xu, "Detecting offensive language in social media to protect adolescent online safety," in *IEEE International Conferences on Privacy, Security, Risk and Trust and on Social Computing (SOCIALCOM-PASSAT)*, 2012.

[31] D. Yin, B. D. Davison, Z. Xue, L. Hong, A. Kontostathis, and L. Edwards, "Detection of Harassment on Web 2.0," in *Content Analysis in the Web 2.0 (CAW2.0)*, 2009.

[32] K. Reynolds, A. Kontostathis, and L. Edwards, "Using machine learning to detect cyberbullying," in *IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2011.

[33] J.-M. Xu, K.-S. Jun, X. Zhu, and A. Bellmore, "Learning from bullying traces in social media," in *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*, 2012.

[34] R. Zhao, A. Zhou, and K. Mao, "Automatic detection of cyberbullying on social networks based on bullying features," in *International Conference on Distributed Computing and Networking (ICDCN)*, 2016.

[35] E. Raisi and B. Huang, "Cyberbullying detection with weakly supervised machine learning," in *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2017.

[36] Q. Huang, V. K. Singh, and P. K. Atrey, "Cyber bullying detection using social and textual analysis," in *International Workshop on Socially-Aware Multimedia (SAM)*, 2014.