You Don't Know How I Feel: Insider-Outsider Perspective Gaps in Cyberbullying Risk Detection

Seunghyun Kim,[1] **Afsaneh Razi,**[2]
Gianluca Stringhini,[3] **Pamela Wisniewski,**[2] **Munmun De Choudhury,**[1]
[1]Georgia Institute of Technology, [2]University of Central Florida, [3]Boston University
seunghyun.kim@gatech.edu, afsaneh.razi@knights.ucf.edu,
gian@bu.edu, pamwis@ucf.edu, munmund@gatech.edu

Abstract

Cyberbullying is a prevalent concern within social computing research that has led to the development of several supervised machine learning (ML) algorithms for automated risk detection. A critical aspect of ML algorithm development is how to establish ground truth that is representative of the phenomenon of interest in the real world. Often, ground truth is determined by third-party annotators (i.e., "outsiders") who are removed from the situational context of the interaction; therefore, they cannot fully understand the perspective of the individuals involved (i.e., "insiders"). To understand the extent of this problem, we compare "outsider" versus "insider" perspectives when annotating 2,000 posts from an online peer-support platform. We interpolate this analysis to a corpus containing over 2.3 million posts on bullying and related topics, and reveal significant gaps in ML models that use third-party annotators to detect bullying incidents. Our results indicate that models based on the insiders' perspectives yield a significantly higher recall in identifying bullying posts and are able to capture a range of explicit and implicit references and linguistic framings, including person-specific impressions of the incidents. Our study highlights the importance of incorporating the victim's point of view in establishing effective tools for cyberbullying risk detection. As such, we advocate for the adoption of human-centered and value-sensitive approaches for algorithm development that bridge insider-outsider perspective gaps in a way that empowers the most vulnerable.

Introduction

Bullying is a prolonged, repetitive, and aggressive behavior from one individual directed to another (Smith et al. 1999). It has been shown to have a long-term negative impact on victims, especially youth. The damage inflicted on young victims of bullying can last throughout their lives, negatively influencing their health, wealth, social, and mental well-being (Zwierzynska, Wolke, and Lereya 2013). What exacerbates these negative impacts is that bullying comes in various forms and happens across multiple environments, ranging from home, school, to one's workplace. In recent years, bullying has transcended offline person-to-person circumstances to include *cyberbullying*, a form of bullying that occurs online, and has affected more than half of youth within the U.S. (Anderson 2018) and ∼65% of all youth within their lifetime (Brochado, Soares, and Fraga 2017).

Given its prevalence, experts agree that cyberbullying is a problem that must be addressed in order to protect the mental health, safety, and well-being of our youth (Thomas, Connor, and Scott 2015) and because the damage inflicted on young victims of bullying can last throughout their lives (Zwierzynska, Wolke, and Lereya 2013). Young people rely on social media as a means to make new friends, develop their social networks, as well as forming bridging and bonding social capital (Ellison, Steinfield, and Lampe 2007); however, ∼70% of youth have experienced "drama" amongst their online friends (Lenhart et al. 2015) and express concerns towards social media websites in tackling cyberbullying (Hamm et al. 2015). Meanwhile, youth and young adults also use social media to make sensitive disclosures and seek support around important personal issues, including bullying victimization and sexual abuse (De Choudhury and De 2014; Andalibi et al. 2016). However, in some cases, youth report being further traumatized due to cyberbullying that result from these sensitive online disclosures, which were meant for seeking support (Razi, Badillo-Urquiola, and Wisniewski 2020) or to gain therapeutic benefits (Ernala et al. 2017). Although cyberbullying is typically against the terms of service of most platforms, the problem still persists; it is nearly impossible for a handful of content moderators engaged in volunteer labor to manually keep up with the increasing volumes of online interactions (Van Royen et al. 2015).

Consequently, there have been many attempts over the years to build and evaluate sophisticated and robust computationally-driven systems to detect bullying, whether offline or online (see (Rosa et al. 2019) for a comprehensive review). An effective automated or a mixed-initiative system that combines machine and human efforts to detect cyberbullying could offer helpful resources to victims of cyberbullying. Such systems can augment the abilities of content moderators of online platforms so that they can intervene and mitigate behaviors that may be deemed inappropriate per the norms of a community (Van Cleemput, Vandebosch, and Pabian 2014). However, recent work (Ziems, Vigfusson, and Morstatter 2020) points out key limitations, such as the lack of publicly available training data and a robust standard for determining ground truth, that have made existing cyberbullying detection algorithms unfit for real-world use. Notably, to

date, most research on automated detection of cyberbullying has leveraged third-party annotators or "**outsiders**" (rather than victims or "**insiders**") to label training datasets for cyberbullying ground truth, e.g., (Singh, Ghosh, and Jose 2017; Kwak, Blackburn, and Han 2015), which may not be sensitive to the victims' narratives regarding their own experiences.

Personal experiences are core to knowing whether a situation warrants consideration as bullying (Barlińska, Szuster, and Winiewski 2013; Van Cleemput, Vandebosch, and Pabian 2014; Gualdo et al. 2015; Kwak, Blackburn, and Han 2015). Given that bullying perspectives differ based on the vantage point of the person assessing the situation, and that detection algorithms typically rely on third-party annotators for ground truth curation, we pose the following research questions:

RQ1: *How do classification algorithms perform while using differing perspectives from the "insiders" (victims) or the "outsiders" (third-party annotators) as ground truth?*

RQ2: *How do these perspectives differ when automatically assessing social media posts for bullying?*

To answer these research questions, we partnered with TalkLife, an online peer-support platform, to obtain posts categorically labeled by post authors based on the following topics: *Bullying, Mental Health, Family, Friends*, and *Relationships*. We scoped a large corpus of 2,362,428 posts to sample 2,000 posts that were then manually coded by five third-party annotators. Next, we trained two sets of models based on "insiders" (i.e., post authors') categorization of their own posts versus the "outsiders" (i.e., third-party annotators') classifications. To this end, we trained multiple classifiers from more interpretable approaches such as logistic regression models as well as state-of-the-art approaches in cyberbullying detection, such as deep neural networks (Founta et al. 2019). We compared the influence of these perceptions on the performance of the cyberbullying risk detection models. Overall, we found statistically significant differences between the performances of the insider and outsider models (RQ1). The models trained based on the insider annotations had higher recall for bullying than the models based on the outsider annotations. The outsider models were likely to flag "obvious" bullying incidents, which resulted in higher precision but with the trade-off of lower recall. A deeper analysis corresponding to RQ2 revealed that these outsider models failed to capture the rich and diverse narrative framings of self-reports of bullying incidents, as well as instances where a reference to bullying was not explicitly made.

Our findings highlight the importance of incorporating the victims' points of view and impressions, and negotiating the trade-off between the differing perspectives, when developing cyberbullying risk detection models on online platforms. Drawing attention to the critical limitation of past literature, we argue the need for human-centeredness and value-sensitive approaches in developing these algorithms.

Content Warning and Privacy Considerations Topics discussed in this paper are sensitive (e.g., bullying, sexual trauma) and may be triggering. Given the sensitive nature of this work, we took several measures to protect the privacy of the users whose data is used here. To reduce traceability and discoverability of a potentially vulnerable population, personally identifiable information (e.g., screen names) was removed prior to it being shared with the researchers, and all quotations included in this paper are paraphrased.

Related Work

An Overview of Cyberbullying Detection Research

The domain of cyberbullying detection research is fairly mature with numerous machine learning and computational social scientists developing automated detection algorithms based on supervised or semi-supervised machine learning methods (Hosseinmardi et al. 2015; Rafiq et al. 2015; Zhao, Zhou, and Mao 2016; Soni and Singh 2018; Cheng et al. 2019). According to two recent systematic reviews of automated cyberbullying detection research (Rosa et al. 2019; Salawu, He, and Lumsden 2020), the key challenges in this research area include: 1) a lack of publicly available datasets that contain labeled data for establishing ground truth, 2) a lack of consensus on how cyberbullying is conceptualized, and 3) a lack of uniformity as to how these detection systems are evaluated. As such, it was concluded that most of these detection systems have little, if any, application to the real-world (Rosa et al. 2019).

In this work, we address some of these limitations by analyzing a real-world dataset that contains categorized (or labeled) data from the original authors of the posts (i.e., insiders). We compare these ground truth labels to those of third-party annotators (i.e., outsiders) to evaluate how these perspectives impact the performance of cyberbullying detection. In the next section, we describe how the use of third-party annotators is common practice when developing cyberbullying detection algorithms and how this practice may impact the overall quality of the detection systems.

The Use of Third-Party Annotations for Establishing Ground Truth

One of the key challenges when preparing a new training dataset for a supervised computational model is creating ground truth labels. Given the large-scale datasets required for robust analysis, this task can prove to be labor-intensive and daunting. Consequently, many researchers resort to hiring a group of annotators (Schenk, Guittard et al. 2009), or use crowdsourcing tactics (Hosseinmardi et al. 2015; Rafiq et al. 2015) to manually code the data based on a pre-defined codebook that attempts to describe the phenomena of interest (Dinakar, Reichart, and Lieberman 2011; Dadvar et al. 2013; Singh, Ghosh, and Jose 2017). This general practice may be appropriate where the classification task is straightforward, such as identifying specific objects in images. However, when the task involves perspective-taking and nuance—like the case of cyberbullying—these approaches of ground truth curation may not be as appropriate.

Yet, most published cyberbullying detection systems continue to utilize the practice of using third-party ("**outsider**") annotators in their research. For instance, Dinakar et al. investigated the use of classification to detect cyberbullying comments on YouTube (Dinakar, Reichart, and Lieberman 2011). They had two annotators, one of which was a youth

educator, label data related to sexuality, race, and intelligence. Dadvar et al. followed a similar approach to collect and label 4.6k YouTube comments (Dadvar et al. 2013). Other studies on cyberbullying detection (Dinakar, Reichart, and Lieberman 2011; Huang, Singh, and Atrey 2014; Singh, Ghosh, and Jose 2017) report using students or other individuals as third-party annotators, but often do not report on the past experiences of these individuals, their positionality to the dataset, or how they were trained to accurately complete the annotation task.

In order to scale up the manual labeling process and reduce costs, many researchers (Hosseinmardi et al. 2015; Rafiq et al. 2015; Chatzakou et al. 2017) are now using crowdsourcing platforms like Amazon Mechanical Turk (Crowston 2012) for annotation tasks. For instance, Hosseinmardi et al. (Hosseinmardi et al. 2015) used annotators on CrowdFlower to label cyberbullying incidents on Instagram. Following their lead, other researchers have also used CrowdFlower for gathering annotations toward cyberbullying dataset development (Rafiq et al. 2015; Chatzakou et al. 2017; Founta et al. 2019). Hosseinmardi et al.'s work (Hosseinmardi et al. 2015), however, is notable because they established a systematic way to monitor the quality of the data labeling process, which is something rarely done in most studies; yet, the labels still heavily relied on the personal decision-making of the individual annotators, who were unknown to the researchers. Reflecting more deeply on the robustness of using crowdsourced data labels, Founta et al. (Founta et al. 2018) found that agreement between annotators decreased as the number of possible abuse categories increases. Therefore, they reduced the abusive behavior into four categories (i.e., normal, spam, abusive, and hateful). While these categories may have been more stable from an algorithmic perspective, such monolithic labels may negate the nuance within cyberbullying experiences, which has been a general criticism of cyberbullying detection systems in critical reviews of the literature (Rosa et al. 2019; Salawu, He, and Lumsden 2020).

Acknowledging the Victim's Perspective

In this paper, we refer to the victims' perspectives as the **insider perspectives**, because the individuals have a first-hand experience of the incident. It is understandable as to why the victim's perspective is rarely taken into account when annotating social media data. First, when scraping and analyzing publicly available social media trace data, it is nearly impossible and ethically questionable to contact the original poster about their experience. Second, research has shown that victims rarely report such incidents on online platforms or the authorities (Kwak, Blackburn, and Han 2015). Yet, research has found that the role of the individual (e.g., bully, victim, or bystander) and the context both play a crucial role in how one perceives whether bullying has occurred or not. For instance, Gualdo et al. (Gualdo et al. 2015) found differences in how youth perceived bullying depending on their role (e.g., bullies, victims, bully-victims), as well as the contexts in which bullying occurred (e.g., online versus offline). Victims of cyberbullying often reported being less impacted by cyberbullying incidents than bullies anticipated, unless the victims experienced both offline and online bullying (Gualdo et al. 2015). The way one perceives a cyberbullying event might drastically differ based on the past experiences of that individual, and thus could influence the labeling of the data.

Third-party annotators are by definition outsiders as they are not the direct victims of the specific cyberbullying events they are annotating. Yet, their lived experiences (e.g., their own bullying experiences) may influence their judgment, as well as their differing viewpoints from those of the victims'. Although in many cases there is a clear and distinct sign that an event falls under the category of cyberbullying, in other cases, the situation is less clear. As such, third-party annotators may fail to "read between the lines" regarding the relationship between the victim and bully, the underlying message being sent, or other nuanced details within the interaction. Yet, assessing the robustness of the annotations made by these outsiders is important as it lays the foundation for the models trained in the majority of cyberbullying detection research. To our knowledge, no studies on automated cyberbullying detection have attempted to incorporate the bullying victims' perspectives in ground truth generation, which is a key contribution of our study.

TalkLife Categories			
Relationships	**Family**	Self Harm	**Friends**
Hopes	**Bullying**	Health	Work
Music	Helpful Tips	Parenting	Education
Religion	LGBT	Pregnancy	Positive
Mental Health	My Story	Poetry	Eating Disorders
Addiction	Self-Care	I Need Help	New Parents
Gaming	Grief	Anxiety	Disabilities
Depression	Life Hacks	Politics	Ask TalkLife
Others			

Table 1: TalkLife Categories: the categories are ordered in the way they are introduced when a user makes a post on TalkLife; the categories selected for this study are marked in the table in **bold**

Category	Number of Posts	Number of Users
Bullying	40248	32907
Mental Health	434456	112892
Family	227243	84417
Friends	676729	126799
Relationships	983752	212850

Table 2: Number of posts and users in each Category used for dataset creation

Data

Data Collection

We licensed a de-identified dataset from an online peer-support platform called TalkLife (https://www.talklife.co/), a platform that social computing and computational social science literatures (Pruksachatkun, Pendse, and Sharma 2019; Saha and Sharma 2020) have recently started to explore. The time frame of this dataset spanned 2012-2018. The entire dataset contains about five million posts and 15 million comments made by over 400 thousand users.

Positive-bullying Criteria
1) Aggressive behavior between the post author and others (such as the post mentions or implies a specific individual/group that harassed/insulted/assaulted the post author; the post author feels threatened/insulted/judged by others; the post author is being sexually or otherwise harassed on or outside the platform)
2) Self-narrative, where the post author talks about a past bullying experience, on or outside the platform
Negative-bullying Criteria
1) The post has no clear connection to bullying
2) The post shows signs of depression/loneliness/sadness without clearly articulating harassment to be a cause
3) The post talks about a specific topic that seems more appropriate to other categories than Bullying

Table 3: Rulebook developed for the annotation of posts

We provide a brief overview of the platform and its basic features, as this information is critical to our ensuing approach. This globally used online platform gives youth and young adults the ability to disclose their personal struggles and difficulties for help, advice, and support. The platform operates primarily as a mobile app targeted towards youth to provide peer support. Approximately 70% of the users on this platform are between the ages of 15 and 24. While nationality was not a variable included in the dataset, most of the users of this platform were English speakers, primarily from the United States, United Kingdom, and Europe. Commonly discussed topics range from the day-to-day goings on in their lives, relationships with family or friends, mental health issues, to bullying. Accordingly, when a user creates a post, they are prompted to select one out of over thirty categories that best fit the topic of the post. This feature of enabling users to select a category for their posts is a typical affordance of many social media platforms. For instance, on the social media platform Reddit, post authors self-label their post within a set of categories or "subreddits" (De Choudhury and De 2014). The most popular labeled Talklife categories in the dataset included Relationships, Self-harm, Friend, Mental health, Bullying, etc. Table 1 gives the entire list of categories on the TalkLife platform.

To scope the dataset, we performed an iterative manual inspection of post categories, taking into consideration the ways in which bullying and cyberbullying concepts have been defined in the literature. Removing irrelevant categories (e.g., Life Hacks, Poetry, Religion, Work, Music), we chose five inter-related categories for our analysis: 1) Bullying, 2) Mental Health, 3) Family, 4) Friends, and 5) Relationships. The number of posts and unique users for each category are shown in Table 2.

Conceptualization of Ground Truth

To build machine learning models that would enable comparing insider and outsider perspectives on the posts, we developed a strategy to gather these perspectives. We used the category label assigned by the author of a post as an indicator of the insider perspectives. We do not have such readily available proxy for the outsiders; therefore, we conducted a series of annotation tasks where third-party annotators were asked to assign exactly one category out of the five categories to each post. While we were specifically interested in gathering the outsiders' perspectives on which posts constituted a bullying narrative versus not, we did not adopt a binary annotation strategy for the posts. First, we did not want to bias the annotators to be more conservative or liberal, one way or the other, in identifying what constitutes bullying in posts. Second, our intention was to be consistent with the mental model of the authors of the posts, when, based on the design of our chosen online platform, they have to pick among several designated categories. Since an annotation task that included all 33 categories would be too cumbersome, we considered the five categories noted above that were the most pertinent to bullying narratives.

Gathering Third-Party Annotations

Background of Annotators Five annotators completed the annotation task. They were all undergraduate students at the institutions of the coauthors of this paper. Three self-identified as male annotators and two identified as female. The annotators were familiar with qualitative content coding methods, were active social media users themselves, and represented diverse cultural and academic backgrounds. In addition, the annotators had previously experienced bullying or seen their narratives online either directly or indirectly – their relevant personal experiences were helpful in unpacking the highly subjective and complex context of the posts to be annotated and the goals of the annotation task.

Annotation Task The annotation process consisted of four tasks. For each task, each annotator was given the same sampled set of posts of which 60% came from Bullying and the remaining 40% was evenly distributed between the other four categories. To avoid posts with obvious references to bullying, posts that contained the word "bully" or its variant were excluded in constructing this sample. Drawing upon criteria and quality control measures used in prior work (Hosseinmardi et al. 2015), the annotators were given detailed instructions on how to label each post along with examples from each category. A total of 2,000 posts (1200 posts from Bullying, 200 posts each from Mental Health, Family, Friends, and Relationships) were used for the annotation process. The size of the dataset, while not huge, is sufficient for and consistent with prior studies in this field of research. For example, in a previous work on examining the role of previous discourses in identifying key elements of public textual cyberbullying, the dataset that was used was a total of 2038 Ask.fm conversation instances (Power et al. 2019).

Once the annotators labeled all of the post batches, we gathered the posts for which two or three of the annotators labeled them as Bullying while the others labeled them as one of the remaining four groups. The annotators returned 267 posts with a clear 'yes' on bullying and 1381 posts with a clear 'no' on bullying. Accordingly, the annotators together re-examined the remaining 352 posts on which there was no agreement, in a two hour workshop to come to a consensus. To minimize the influence of bias from the researchers, the researchers were only present in the workshop discussion

Category	Example Post
Positive-Bullying	My boyfriend slapped me and hasn't apologise and is now telling his family about it. Do I have cause to be upset?
	is it normal to feel like all your friends are againt you like they are talking about you behind your back and you know its happening but you cant do anything about it....if you try to you only get bullied and put down..and they just laugh and hurt you emotionally and physically
Negative-Bullying	You don't have right to punish who apologized you from deep inside. They suffer with guiltiness and also they realized they done something wrong, but not accepting apologize and punishing them is the rude way!
	Why do people hate me?

Table 4: Sample positive-bullying and negative-bullying posts by the annotators

for quality control purposes and facilitation. During the discussion, the annotators established a rulebook which they followed for reaching annotator agreement (see Table 3):

- *Positive-bullying*: A post was classified as bullying by the outsiders if the annotators agreed that the post was clearly talking about bullying experiences such as being harassed at school or being verbally or physically abused by someone that is close to the post author.
- *Negative-bullying*: A post was classified as non-bullying by the outsiders if the annotators identified the post as "not closely related to" bullying or if the annotators were not sure if the post was distinctively related to some form of a harassment experience, such as sharing of personal thoughts on people or open-ended questions that seemed not completely related to these experiences or incidents.

In Table 4, we note some lightly paraphrased examples of posts that belong to the two classes. At the end of this workshop discussion, we gathered posts that received equal to or more than 4 positive votes as Bullying and those that received equal to or less than a single vote as non-Bullying, leading to a **Fleiss' κ score of 0.79**, which was an improvement over the pre-discussion κ score of 0.47. The finalized dataset from the five annotators consisted of 535 bullying posts and 1465 non-bullying posts.

Machine Learning Approach

We adopted a supervised machine learning approach to compare and understand the insider and outsider perspectives in cyberbullying detection. Considering the labels from each group (insiders' self-assignments of posts to "bullying" or another category, and the annotations by the outsiders) as ground truth, our methodological approach below describes how we conceptualized a number of post features using natural language processing techniques, and then constructed a series of machine learning classifiers for the said purpose.

Training and Test Datasets

Due to the intrinsic differences in the labels between the five contributors and the original post authors, as well as the complexity of the annotation task itself, it was not practical to construct a single training dataset which had 50-50 split between posts labeled as positive-bullying and posts labeled as negative-bullying. However, it was crucial that the posts in the training dataset were consistent across the insiders and the outsiders, in order to rule out the impact of unobserved topical confounders in the two perspectives. Therefore, we created two types of training datasets (Figure 1):

(1) The first training dataset ensured we obtain a 50-50 split between the positive and negative bullying classes from the insider perspective. To do so, we started with the 1200 posts described in the previous section that were assigned to the Bullying category by the post authors, and the other 800 that were assigned to one of the four categories: Mental Health, Family, Friends, Relationships. To obtain a 50-50 split, we randomly sampled 800 from the 1200 posts assigned to the Bullying category to obtain the first training dataset, such that 800 posts were positive-bullying and the other 800 negative-bullying according to the post authors' ground truth labels, that is, their categorical assignments of the posts. We refer to this as the **insider-training-1** dataset. We used these same (800 + 800=) 1600 posts to construct an equivalent **outsider-training-1** dataset, however here, the ground truth labels of the posts were derived from the third party annotations.

(2) In a similar manner, to construct a balanced dataset from the outsider perspective, we started with the 535 positive-bullying posts labeled by the annotators in the previous section. To obtain a 50-50 split here, we randomly sampled an equal number of negative-bullying posts (535 out of the 1465 annotated) to create an **outsider-training-2** dataset. Like above, we used these same (535 + 535=) 1070 posts to construct an equivalent **insider-training-2** dataset, where the ground truth labels came from the post authors' assignment of posts to the Bullying or other categories.

Note that we acknowledge that cyberbullying messages are normally a small percentage of all social media messages, hence the natural distribution between bullying and non-bullying posts is likely to be significantly skewed. Although not a true representation of this natural distribution, a balanced dataset with a 50-50 split between the positive and negative classes may be more desirable in our case, as it will allow eliciting differences between the insider and outsider perspectives in a robust and consistent fashion.

Corresponding to the two types of training datasets above, we constructed an equivalent test set for each type, that is, one test set corresponding to the insider/outside-training-1 dataset and another for the insider/outsider-training-2 dataset. Out of the 39,048 unannotated Bullying category posts, and the 2,321,380 unannotated Mental Health, Friends, Family, and Relationship category posts, we randomly sampled respectively 200 posts each corresponding to the first type of training data, and respectively 135 posts each for the second type of training data. They are referred as **testing-1** and

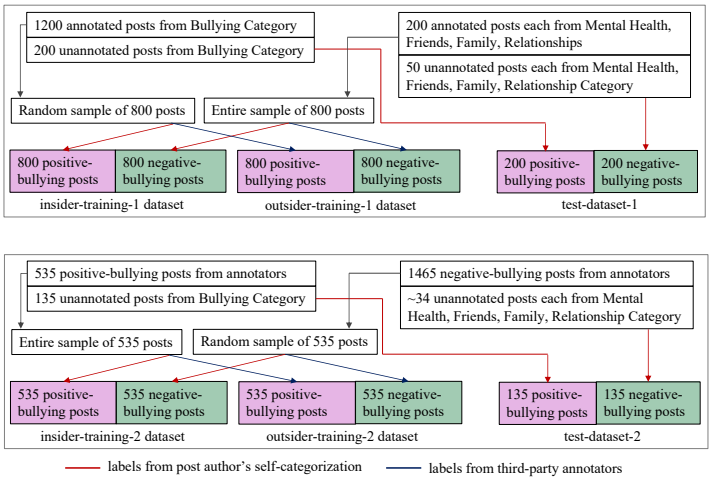

Figure 1: Overview of the training and test datasets

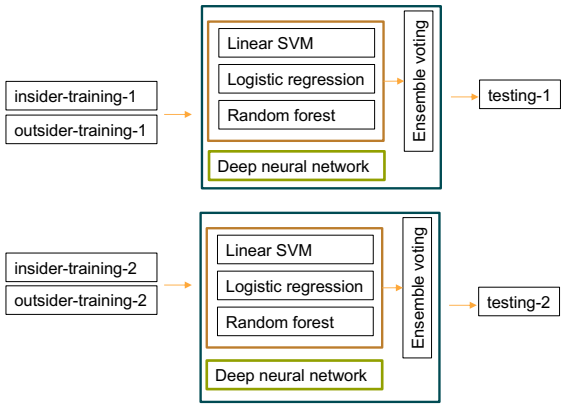

Figure 2: Overview of model training and testing approach

testing-2 in the rest of the paper. With our approach, we thus ensured that we had a 4:1 ratio between the sizes of the training and test datasets of each type, as well as that half of the posts in each test set were positive-bullying and the rest negative-bullying, based on the insider perspective (the post authors' assignment of the posts to the Bullying or other categories). Since our goal here is to contrast the outsider perspective with that of the insiders, for both of these test datasets, we considered the insiders' point of view only.

Feature Engineering

Each post in the above training and test set was represented as a vector of 1,206 concatenated features described below:

- **Psycholinguistic Attributes (LIWC):** Linguistic Inquiry and Word Count (LIWC) is widely used to extract psycholinguistic features of text and quantify meaning along multiple dimensions (Pennebaker, Francis, and Booth 2001). We used normalized word counts related to 50 categories that span *affect, cognition and perception, interpersonal focus, temporal references, lexical density and awareness, biological concerns*, and *social and personal concerns*, per prior work (De Choudhury and De 2014).
- **Sentiment:** The sentiment of a given post is important as it illustrates the emotion or tone of the post. We used Stanford CoreNLP's deep learning tool to obtain the positive, negative, neutral scores of each post (Manning et al. 2014).
- **Open Vocabulary (Uni/bigrams):** Drawing on prior work where open-vocabulary based approaches have been extensively used to infer psychological attributes of individuals (Schwartz et al. 2013), we also extracted the normalized counts of the top 500 n-grams ($n = 1, 2$).
- **Hate Lexicon:** To capture domain-specific signals as features, we used a lexicon developed in prior work (Saha, Chandrasekharan, and De Choudhury 2019) to quantify words related to offensive and hateful speech. These words relate to gender, sexual orientation, disability, ethnicity, or race—attributes that are common in bullying content. The lexicon was curated through automated classification, crowdsourcing, and moderation by an expert.

Models

Using the features above, we used the two types of training datasets to build binary classifiers for detecting bullying posts: we refer to them as **insider models** (trained on insider-training-1 and insider-training-2 datasets) and **outsider models** (trained on outsider-training-1 and outsider-training-2 datasets) respectively.

We first chose classification approaches that would provide strong performance alongside interpretability: Linear Support Vector Machine, Random Forest, Logistic Regression, and Ensemble Voting. The Ensemble classifier took the predictions from the three aforementioned classifiers and used a majority voting to determine the final prediction of a post.

Next, we implemented the deep neural network model, developed by Founta et al. (2019), to detect abuse in online social media platforms – our purpose here was to compare the performance of the insider and outsider models using state-of-the-art cyberbullying detection approaches. Founta

et al.'s deep-learning approach allows us to capture subtle, hidden commonalities and differences between the various ways abuse might be represented in text, and it provides a global and lightweight solution with the ability to capture latent patterns and structures in the underlying data, without requiring excessive much feature engineering and model tuning. Following text as input, this model first includes an embedding layer, which maps each word to a high-dimensional (typically 25-300 dimensions) vector using pre-trained word embeddings from GloVe (Pennington, Socher, and Manning 2014). The model then includes a recurrent neural network (RNN) layer – specifically a Gated Recurrent Unit or GRU architecture (Chung et al. 2014), which learns sequences of words by updating an internal state. Finally, an attention layer (Bahdanau, Cho, and Bengio 2014) provides a mechanism for the RNN to "focus" on individual parts of the text that contain information related to the task. The last layer is a fully connected output layer with one neuron per class, and a softmax activation function to normalize output between 0 and 1. The output of each neuron represents the probability of the sample belonging to each respective class.

All classifiers except the Ensemble model went through a $k = 10$-fold cross validation for hyperparameter tuning. Each model was further trained on one training set and tested on the corresponding test dataset. Figure 2 gives an overview of this approach after hyperparameter tuning.

Evaluation

We used the average accuracy of the models, the F1-measure, the area under the receiver operating characteristic curve (AUC), and class-specific precision and recall to evaluate our models, combined across the two test sets. While the accuracy and F1 scores return the general performance of the models, precision and recall of each class provide more detailed insights, especially since our training datasets have class imbalances. As we seek to illuminate the influence of insider and outsider perspectives on the performance of the models, class specific metrics are more important, as the cost and importance of the false-positives and false-negatives can be different depending on the application scenario. False positives mean that a model classifies a post that is not really about bullying as bullying, possibly resulting in a user receiving some intervention. In this case, the user could feel startled, but the implications of misclassification are low. The opposite case, however, has a different level of gravity. Classifying a bullying instance as non-bullying (false negative) could mean that the system misses a user that is experiencing bullying, and with the prolonged negative effect of bullying, this could lead to detrimental outcomes.

Results

Comparing Overall Model Performances

To start off, toward answering RQ1, we examine how different features affect the predictions of the models. Accordingly, we performed an ablation study of the feature categories. We trained classifiers across a total of five different feature set combinations: a full feature set (referred to as "Full"), and four sets excluding one of each of the four types of features

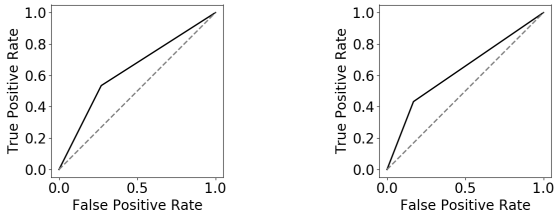

Figure 3: ROC curves of the best performing insider (left) and outsider (right) models. The dotted line depicts the random true positive rate and false positive rate

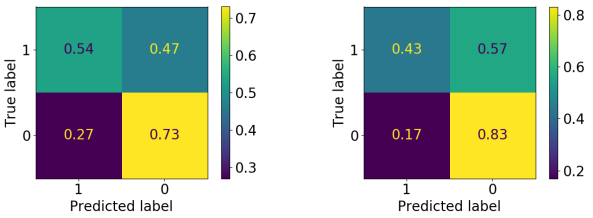

Figure 4: Confusion matrices of the best performing insider (left) and outsider (right) models

(the psycholinguistic attributes referred to as "LIWC", hate lexicon referred to as "Hate", sentiment referred to as "Sent", and uni/bigrams referred to as "*n*-gram").

Tables 5 and 6 illustrate our results; each classification model has five columns that represent the metrics used in this study to evaluate performance. Each metric has two columns which represent the average metric evaluated on the two insider-training and the outsider-training datasets, in that order. The five feature sets each have a row for positive-bullying (Pos-Bully) and negative-bullying (Neg-Bully). The features that were excluded as a part of the ablation study are written at the head of the row; for example, the feature set that excluded hate speech lexicon is shown in the row -Hate. AUC and accuracy (Accr) do not have class-specific numbers unlike the other three metrics as the two numbers represent the performance of the overall model.

For the purposes of our discussion of the overall model performance across the insider and outsider perspectives, we selected the model and the feature set combination that had the highest AUC. Between the two models within each perspective, corresponding to each training dataset, we also chose the model with the best AUC.

The insider model trained on the insider-training-1 dataset (0.63 AUC) was used and the outsider model trained on the outsider-training-2 dataset (0.63 AUC) was used for deeper analysis. Both models were tested on the respective testing set. Setting the deep neural network (NN) model aside for a moment, for the insider model, the best performing model was the ensemble model with the hate lexicon features excluded (-Hate), while it was the logistic regression model with all of the features for the outsider models (Full). Turning attention to the NN model now, although this model improved performance in terms of accuracy, still, the differ-

		Linear SVM										Logistic Regression									
		Prec		Rec		F1		AUC		Accr		Prec		Rec		F1		AUC		Accr	
		In	Out	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out
Full	Pos-Bully	0.61	0.80	0.70	0.25	0.64	0.35	0.62	0.58	0.62	0.58	0.61	0.80	0.72	0.29	0.64	0.40	0.61	0.60	0.61	0.60
	Neg-Bully	0.65	0.55	0.53	0.92	0.57	0.69					0.68	0.57	0.51	0.91	0.54	0.69				
-Hate	Pos-Bully	0.60	0.80	0.72	0.29	0.64	0.39	0.61	0.60	0.61	0.60	0.61	0.80	0.72	0.29	0.64	0.40	0.62	0.60	0.61	0.60
	Neg-Bully	0.66	0.57	0.50	0.91	0.54	0.70					0.68	0.57	0.52	0.91	0.55	0.69				
-Sent	Pos-Bully	0.61	0.83	0.71	0.27	0.64	0.38	0.61	0.60	0.61	0.60	0.61	0.82	0.75	0.29	0.66	0.39	0.62	0.60	0.61	0.60
	Neg-Bully	0.65	0.56	0.51	0.92	0.55	0.70					0.66	0.57	0.48	0.91	0.52	0.69				
-n-gram	Pos-Bully	0.56	0.81	0.76	0.30	0.64	0.39	0.58	0.59	0.58	0.59	0.61	0.79	0.72	0.28	0.64	0.38	0.61	0.59	0.61	0.59
	Neg-Bully	0.64	0.56	0.40	0.88	0.47	0.68					0.68	0.56	0.50	0.90	0.53	0.69				
-LIWC	Pos-Bully	0.52	0.79	0.77	0.28	0.61	0.29	0.52	0.54	0.52	0.54	0.49	0.79	0.73	0.29	0.57	0.30	0.50	0.55	0.50	0.55
	Neg-Bully	0.58	0.54	0.27	0.80	0.32	0.63					0.62	0.54	0.27	0.80	0.27	0.63				

		Random Forest										Ensemble									
		Prec		Rec		F1		AUC		Accr		Prec		Rec		F1		AUC		Accr	
		In	Out	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out
Full	Pos-Bully	0.55	0.30	0.78	0.24	0.62	0.27	0.54	0.54	0.54	0.54	0.60	0.82	0.73	0.28	0.64	0.38	0.61	0.59	0.61	0.59
	Neg-Bully	0.29	0.54	0.31	0.85	0.30	0.65					0.67	0.56	0.49	0.91	0.53	0.69				
-Hate	Pos-Bully	0.59	0.30	0.57	0.24	0.58	0.27	0.58	0.54	0.58	0.54	0.62	0.82	0.72	0.27	0.65	0.37	0.62	0.59	0.62	0.59
	Neg-Bully	0.58	0.54	0.58	0.84	0.58	0.64					0.69	0.56	0.53	0.91	0.57	0.69				
-Sent	Pos-Bully	0.55	0.29	0.79	0.24	0.63	0.26	0.55	0.54	0.55	0.54	0.61	0.84	0.73	0.28	0.65	0.38	0.62	0.59	0.62	0.60
	Neg-Bully	0.30	0.53	0.32	0.84	0.31	0.64					0.66	0.56	0.52	0.91	0.56	0.69				
-n-gram	Pos-Bully	0.56	0.30	0.75	0.24	0.62	0.27	0.55	0.54	0.55	0.54	0.60	0.82	0.73	0.26	0.64	0.36	0.61	0.58	0.61	0.58
	Neg-Bully	0.57	0.54	0.35	0.84	0.35	0.65					0.67	0.56	0.49	0.90	0.53	0.68				
-LIWC	Pos-Bully	0.49	0.32	0.72	0.20	0.56	0.24	0.49	0.54	0.49	0.54	0.50	0.79	0.72	0.28	0.57	0.29	0.50	0.54	0.50	0.54
	Neg-Bully	0.24	0.53	0.27	0.89	0.25	0.66					0.62	0.54	0.28	0.80	0.29	0.63				

Table 5: Model performance of the insider models (In) and outsider models (Out) across different feature sets and training sets, for the SVM, logistic regression, random forest, and ensemble classifiers

		Neural Network Model									
		Prec		Rec		F1		AUC		Accr	
		In	Out	In	Out	In	Out	In	Out	In	Out
Full	PB	0.66	0.80	0.36	0.17	0.48	0.32	0.64	0.64	0.63	0.71
	NB	0.57	0.54	0.80	0.92	0.66	0.68				
-Hate	PB	0.64	0.75	0.39	0.23	0.49	0.33	0.62	0.67	0.65	0.73
	NB	0.51	0.55	0.78	0.91	0.65	0.68				
-Sent	PB	0.71	0.77	0.36	0.21	0.47	0.31	0.65	0.66	0.64	0.71
	NB	0.57	0.54	0.85	0.92	0.68	0.68				
-n-gram	PB	0.69	0.72	0.34	0.21	0.44	0.31	0.66	0.66	0.64	0.71
	NB	0.56	0.54	0.84	0.92	0.67	0.68				
-LIWC	PB	0.67	0.58	0.03	0.03	0.06	0.06	0.57	0.57	0.61	0.67
	NB	0.51	0.51	0.98	0.98	0.67	0.67				

Table 6: Model performance of the insider models (In) and outsider models (Out) across different feature sets and training sets, for the deep neural network based classifier developed by Founta et al. (2019). Here PB and NB are acronyms for positive-bullying and negative-bullying classes

ences between the insider and outsider models across feature sets was comparable to that in the other models in terms of the AUC. Moreover, for this model, the overall performance metrics, particularly recall, across the feature sets, were significantly lower despite the high AUC. This was due to the model predicting the majority as negative-bullying, resulting in a misleadingly high AUC with low recall for the positive-bullying. For this reason, the NN model was not chosen for our ensuing discussion of findings. In addition, we note that due to its opaqueness, this model gives us little information to what contributed to the differences between the insider and the outsider perspectives. We deem this understanding to be critical to the contribution of this paper, for reproducibility purposes and to emphasize when either of two perspectives may be more valued than the other.

In addition, we generated the confusion matrix (Figure 4) and the Receiver Operating Characteristic curves (ROC: Figure 3) of these best performing models for further discussion. Broadly speaking, both the insider and the outside models performed better than random, as shown in Figure 3. However, comparing them mutually, we find from Table 5 that the insider models, across feature and model types, consistently improved over the outsider models both in terms of AUC (up to 3% improvement) and accuracy (up to 4% improvement); a χ^2 test further revealed this difference to be statistically significant ($\chi^2 = 4.88, p < .05$).

We now further examine the differences in class-specific performances, because, ostensibly, depending on the use of these classifiers, better performance in predicting one class might be more desirable over the other (ref. the subsection Evaluation above). While the best insider model had a higher recall for the positive-bullying class (38% higher; $\chi^2 = 72.1, p < .05$), the outsider model had a higher recall (25% higher; $\chi^2 = 58.1, p < .05$) for the negative-bullying class (Table 5). This observation aligns with what we observe in the confusion matrices in Figure 4. We see that the insider model had higher false-positives than false-negatives, while it is the opposite for the outsider model. The higher precision

for the positive-bullying class and the lower precision for the negative-bullying class in the outsider models suggests that the outsider models had a tendency to classify more posts as negative. The insider models, however, labeled more posts as positive. This distinction between the two perspectives was observable across all the models we evaluated in this study.

Understanding Top Predictive Features

Given the differences in performance reported by the classifiers using insider- and outsider-training data, we now turn our attention to understand what elements in posts might guide the labeling decisions by the two perspectives (RQ2). To answer this question, we analyze the most discriminative features for the best performing models; we used the K-best univariate statistical scoring model using mutual information to obtain the relative importance among features, and established their statistical significance using ANOVA.

Looking at the topmost statistically significant features obtained this way, we find that for both the insider and the outsider perspectives, the models were heavily dependent on the LIWC features, compared to the uni/bigram features. Therefore in Figure 5 we report the top 15 LIWC features in terms of their importance given by the respective models. While several of the LIWC categories consistently appear to be highly predictive for both models, their mutual ordering (or rank) is different (Spearman's correlation coefficient $\rho = 0.34, p < .05$). For instance, *humans*, *work*, and *anger* are ranked 1-3 in terms of feature importance in the insider model, however, they are ranked 1, 5, and 2 respectively in the outsider model. *Humans* and *work* both fall under the broader LIWC category of social and personal concerns. Their higher rank in feature importance in the insider models indicate that the narratives or incidents of bullying, as described by the post authors (insiders), are often interlaced with personal stories and experiences. Further, the absolute measure of feature importance for the same LIWC features varies across models. Here, take the example of the *family* feature, which is ranked 1 in terms of feature importance in both models. It is weighted at 0.93 for the insider model, while 0.81 in the outsider. Similarly, the outsider model relies more on the presence of words relating to *anger* in predicting positive-bullying posts (assigning a weight of 0.56), while this feature is weighted at only 0.32 in the insider model, meaning here, its importance is lower. In other words, the outsider models tend to put more emphasis on affective words like *anger* in predicting what post could indicate bullying.

Further, while both models relied heavily and consistently on LIWC categories like *family* and *humans*, the insider model showed some categories that did not appear in the other. Consider the following post excerpt with a *death* word: "I hate school so much it makes me want to die" (the authors refer to a past harassment experience as the reason), and this excerpt with a *sexual* word: "I'm bisexual and my family wants to kill me because they're muslims and it's a sin and everyone are supporting them. I'm so scared please help this is not a joke"; both of these posts were labeled as positive-bullying in the insider-based model.

The outsider model similarly relied on the categories *negative affect* and *conjunction* for predicting positive-bullying

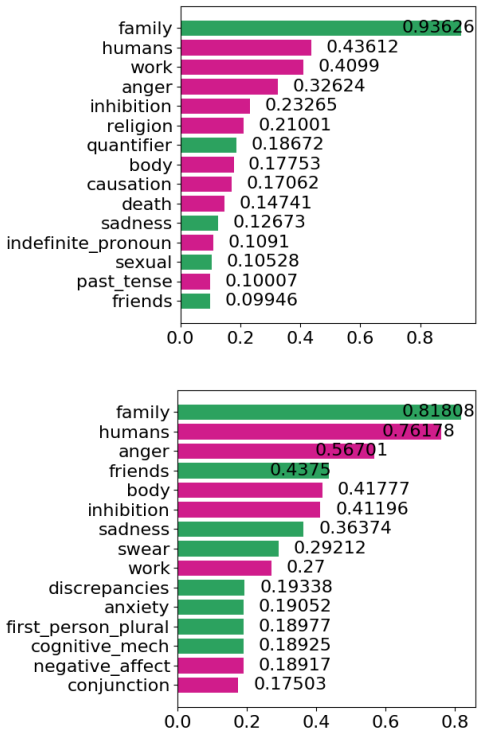

Figure 5: Top LIWC features for the best performing insider (top) and outsider (bottom) models. Features for positive-bullying class are shown in pink while the features for negative-bullying in green, each with its absolute coefficient

posts—these categories do not appear in the top LIWC feature list for the insider models. See this excerpt that consisted of a *negative affect* word: "I'm glad I'm dead, a worthless fuckin' buddah head" – here the outsider model interpreted the expressed negative emotion and low self-esteem to be as a consequence of past harassment. In the top LIWC features for the outsider model, we see more higher ranking features predictive of negative-bullying compared to positive-bullying, indicating the model's propensity to harness more signals to rule out which posts could be about bullying. Notable is the category *swear*, which is weighed at 0.29 in predicting negative-bullying posts, but which is a category that does not appear in the top LIWC features for the insider model.

Additionally, in Table 7 we report the top 30 uni- and bigrams that had high feature importance in the best insider and outsider models. As can be observed clearly, qualitatively speaking, there is little overlap between the types of words and phrases that are harnessed in the prediction task by the models. This indicates that the models are picking up different linguistic cues from the posts in their assessment of what constitutes bullying or non-bullying. We elaborate on these differences further through error analysis, presented next.

Error Analysis on Misclassified Posts

Finally, where are the two models making errors; in particular, what is it that the outsider models are not able to pick up in

Insider model	Outsider model
Unigrams	
know, **afraid**, *parents*, *although*, *guilty*, **bout**, **rude**, *drugs* **kids**, *asked*, **bullshit**, *anymore*, *embarassing*, **guess**, *grown*	*parents*, **rude**, *get*, *know* **point**, **whole**, *guess*, **bullshit**, *life*, *stuff*, *kik*, **assume**, *anymore*, **admit**, **inside**
Bigrams	
much hate, *always stuff*, **people calling**, *around online school boy*, **people dont**, *want live*, **im ugly**, *say things*, **seem like**, **need stop**, **fat disgusting**, **can't deal**, *sometimes feel*, **try make**	**anyone ever**, **always get**, **can't deal**, **fat disgusting**, *always online*, *school boy*, **people dont**, *want live*, *anyone want*, *sometimes feel*, **im ugly**, **real life**, **fuck fuck**, **white people**, **need stop**

Table 7: Top 30 uni- and bigrams in predicting positive- and negative-bullying posts using the best performing insider and outsider models. Each n-gram is color-coded to indicate if it highly predictive of the positive-bullying (bold) or the negative-bullying label (italic)

the classification of the positive-bullying posts (low recall) that insider models can? To answer this, we sampled several posts from the test sets, alongside their predicted labels by the insider and outsider models. We discuss them below:

Capturing implicit references of bullying Here is a paraphrased post that the outsider model classified as negative-bullying, but the insider model correctly classified as positive:

> "I just wanna say that THIS IS NOT OKAY. WEARING A SHORT SKIRT ISN'T A OPEN INVITATION, YOU CANNOT TOUCH SOMEBODY LIKE THAT THINKING IT'S OKAY JUST CAUSE OF WHAT THEY'RE WEARING!!!"

This post, that was shared under the Bullying category, reads almost like an advisory, expressing the author's view on unjustifiably victimizing people, who have experienced a harassment episode, because of the clothes they were wearing. Because the post author associated this post with the Bullying category, what seems like is that this view may be rooted in the author's past experience of a bullying episode. Since there are hardly any direct references to this past experience, the outsider model failed to capture this latent context in the same way that the insider model can. A similar argument can be made for the following post that mentions obliquely about an "incident" and that "[past] pain teaches," however without any explicit reference to a bullying episode:

> "My incidents caused me to learn from it and so I have enacted special procedures in handling, screening and negotiating with new contacts. This favors the fact that I am unable to make new commitments. This is what I learned the hardway. Then I can be ignorant, and sway from people's insanity becoming my insanity. You learn entire life. Every pain teaches"

Framing of experiences Consider this false negative of the outsider model, which is a true positive of the insider:

> "I'm so tired of being picked on by others. I'm so tired of being alone and broken, over and over again, then I'm told that I'm nothing and will never be anything. No one cares about me and no one ever has. literally no one gives a damn about me. That's why I cut myself over and over again. I just want to die so I can stop the pain"

The post author uses words from the LIWC categories *discrepancies*, *death*, *sadness*, and *anger*. Both models treat words from *sadness* as indicators of negative-bullying, as often times sorrow and dejection may be associated with mental health discourse, one type of posts used as negative-bullying examples in our training data. Similarly, both models use words in *anger* as those indicative of positive-bullying, as self-disclosures or personal reports of bullying experiences can often be laden with rage and fury. However, multiple usages of words from the *death* category emphasizes the difference between the learned features of the two models and why the eventual predictions of the two models are different. Words in *death* are more highly weighted in the insider models for positive-bullying prediction, compared to the outsider models – as bullying narratives tend to express hopelessness, a lack of yearning for life, and occasional suicidal thoughts. Further, although the author explicitly cites the cause of their negative feelings about having been bullied by others ("tired of being picked on by others"), this atypical framing was missed by the outsider model. The next post excerpt further hones this point (note "people will beat me down"):

> "The fact is clear, I don't want to get up, I know people will beat me down again. So this time already, I know how to avoid getting beaten is not standing up to them"

In addition, consider this post which was misclassified by the outsider model as positive-bullying when the post was not shared in the Bullying category by the post author:

> "I highly recommend that you all spend some sort of quality time with your loved ones tonight. A movie, a dinner, a chat, anything. [...] And if not, for whatever the reason, then just pm me and I'll keep you occupied [...] These moments are not to be taken for granted. Now go out and have fun people! [...]"

This further supports the points of how the insider model could identify implicit references of bullying as well as the framings of experiences. In summary, these examples illuminate how the different perspectives in establishing the ground truth labeling for the insider and the outsider models influence the performance of their respective cyberbullying detection.

Discussion

An Insider-Outsider Gap in Bullying Perspectives

A key contribution of this research is that we identified a significant gap between insiders and outsiders, when it came to their perspective on identifying bullying related social media posts. In terms of accuracy and AUC, the models' performances were not largely different between the insider models and the outsider models. Yet, the insider model showed that the original authors of the posts were more likely than the third-party annotators to categorize a post as bullying, resulting in higher false-positives but higher recall. In contrast, the outsider models had a higher precision in identifying positive-bullying posts, meaning that when a post was labelled as bullying, it was very likely that it was correct. However, in doing so, the outsider models missed out on a number of posts that

were categorized as positive-bullying from the point of view of the original author of the post.

Overall, we found that outsiders were more conservative than insiders when categorizing a post as bullying. The error analysis gave us further insights into the nuance between the insider and outsider models, suggesting that third-party annotators were less likely to infer bullying from implicit references to having been bullied or when individuals recounted their pain resulting from bullying, rather than recounting the incident. While this gap may seem obvious, explicitly examining how insider perspectives differ from outsider perspectives demonstrates why incorporating this perspective is valuable. This work is a first attempt to empirically establish this gap in perspectives and provides valuable guidance on what may contribute to this gap, as well as what the gap implies for real-world use of cyberbullying detection algorithms.

One might argue that a possible solution to closing the insider-outsider gap could be to recruit annotators that have previously experienced bullying. However, this would also not be entirely the same as using the labeling from the victims themselves—the past experiences of the annotators could be very different from those of the victims. Further, narrative inquiry approaches in cyberbullying research (Bowler, Mattern, and Knobel 2014) argue that the perception of whether a particular experience should be considered as bullying or not should be regardless of the perpetrator's intent or bystander's impression. Bullying victimization should instead be dependent on how the victim perceives the event and how the incident impacts the victim based on their lived experience (Dredge, Gleeson, and De la Piedad Garcia 2014). For instance, Dregde et al.'s interviewed youth who had previously experienced negative interactions online and found that research-based definitions of cyberbullying lacked the key criterion for which youth classified their own cyberbullying experiences, which was the emotional, social, and behavior impact of the interaction on the victim. For these reasons, because our insider models incorporated the viewpoints of the victims to whom the specific instance of cyberbullying in a post was directed at, they were able to gather a more context-specific perspective.

Implications for Cyberbullying Detection

Our work bears several implications for the rich legacy of research building automated cyberbullying detection systems targeted for real-world use. While this work focuses on a single platform, TalkLife with unique affordances, our approach of incorporating insider perspectives in cyberbullying detection algorithms is applicable beyond this particular platform. Many social media platforms have unique affordances that enable users to create and associate ecologically valid insider labels for their shared content. For example, Reddit's own structure of associating a subreddit for each post already provides a categorical label to each user's post. In addition, Facebook pages also provide us with tags which give us the insiders' perspective of their posts, while on Twitter users can assign hashtags to their posts. Consequently, with these type of self-reported and self-initiated insider annotations, it may be possible to not just gather data on insider perspectives unobtrusively in other platforms with comparable affordances as TalkLife, but also do so in a scalable fashion spanning thousands of users and posts, that does not require explicit involvement of the insiders.

Ultimately, an ideal model, to be ready for real-world use, would achieve both high precision (i.e., with the outsider model) and high recall (i.e., with the insider model). However, there is usually a direct trade-off between the two when designing machine learning models in a practical scenario. Below we use a value-based approach to consider when one model may be more appropriate than the other.

When the insider perspective is valued more: The gravity of false negatives can far outweigh that of false positives for some applications. In these types of applications, it is often important that every instance of bullying is detected, meaning recall should be sought for over precision. For instance, in an online bullying peer support platform, it may be important to have a system that can detect almost all bullying instances, which will be possible by adopting a very generous threshold when implementing automated cyberbullying detection algorithms. By ensuring the recall is higher, it will be possible to reach out to more users and offer help and interventions, rather than miss a significant number of vulnerable individuals. So for designing such a system, researchers should give more importance to using victim's perspectives.

When the outsider perspective is valued more: On the other hand, there could be other applications where it is not necessary to detect all instances of bullying. An example includes online games for teens where most of the users use vulgar language more than usual. Here, it may be more important to only detect serious cases to ensure community norms, users' expectations, and the collective "health" of the community ecosystem are maintained while removing deviant users engaging in bullying activities. Therefore, the outsider's view may be more appropriate for labeling the datasets in this instance, leading to improved precision of the detection models, and making sure that no user is removed by mistake in trying to uphold the community norms.

As individuals' perceptions of the world are subjective, subjective input may not always indicate the outcomes in the real world. Moreover, machine learning systems, even if they are "human-aware," often fail to recognize the evolving realities, constraints, and needs of different stakeholders, without whose persistent feedback, the systems may create disconnects that undermine practical initiatives and even alienate key users of the system. Combining both victims' perspectives and outsiders' can help to reduce such biases and help researchers to craft more robust solutions. We note two approaches to achieve this. 1) We suggest the iterative and critical involvement of the victims throughout the design of the cyberbullying detection algorithms, as advocated in the emergent literature in **human-centered machine learning** (Baumer 2017). 2) Given the unique perspectives and values held by the insiders and the outsiders and to trade-off their differences, we emphasize the adoption of **value-design algorithm design** (Zhu et al. 2018).

Limitations and Future Research

We note some limitations of our work. First, we chose de-

mographically diverse college students for annotators as they were within the age range of the primary demographic who post on TalkLife. Yet, the fact that they were college students presents some biases. Future research could further quantify differences in perspectives through a wider variety of factors and in a systematic fashion, investigating the connection between the annotation of bullying narratives and the annotator's personal attributes, such as personality, prior experience with bullying, age, gender, education, cultural values, and views on regulation of online content.

Second, while the feature sets were chosen carefully, there are other features that could be considered to further evaluate the perspective gaps, such as post author metadata or their historical posts. Another limitation would be that although TalkLife is a widely used social media platform, it has a specific purpose of usage (peer support) unlike more general purpose platforms like Facebook or Reddit; this could have influenced the way the users wrote their posts. Third, we utilized the target online platform's affordance of associating a post with a category as a proxy for the victim's perspective. While self-disclosure of sensitive experiences on social media has been used in prior research to build machine learning models (e.g., to detect mental health challenges (Coppersmith, Harman, and Dredze 2014)), it does not allow fully capturing the victim's own interpretation. Future work should also consider the human-centered and value-sensitive design approaches recommended above, thereby enriching our understanding of insiders' perspectives through surveys, interviews, and more involved annotations of posts provided by the direct victims of cyberbullying incidents.

Finally, we recognize the sensitivity of the data and task at hand; while our research shows that gleaning insights from insiders is valuable, in doing so, algorithms may inadvertently jeopardize the victims, since the insider perspective data risks traceability to sensitive, personal information about the victims. Therefore, future research should explore privacy-preserving ways to guide algorithm development that harnesses this perspective, as well as to support replicable and reproducible research through sharing of datasets labeled with insider (and outsider) perspectives.

Conclusion

We examined how one's personal experience and perception of a cyberbullying incident can influence the performance of risk detection algorithms. Models using training sets labeled by the post authors of the target online platform were more effective in achieving a high recall in identifying bullying posts. Essentially, the insider model enabled capturing implicit references of bullying, as well as its different perceptions; aspects difficult to account for from an outsider's perspective. Our research highlights implications for improving cyberbullying detection systems by incorporating the victim's perspective.

Acknowledgements

This study is supported by the United States National Science Foundation under grant IIP-1827700. 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

Andalibi, N.; Haimson, O. L.; De Choudhury, M.; and Forte, A. 2016. Understanding social media disclosures of sexual abuse through the lenses of support seeking and anonymity. In *Proc. CHI*, 3906–3918.

Anderson, M. 2018. A majority of teens have experienced some form of cyberbullying. https://www.pewresearch.org/internet/2018/09/27/a-majority-of-teens-have-experienced-some-form-of-cyberbullying/. Accessed: 2021-04-14.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* .

Barlińska, J.; Szuster, A.; and Winiewski, M. 2013. Cyberbullying among adolescent bystanders: Role of the communication medium, form of violence, and empathy. *Journal of Community & Applied Social Psychology* 23(1): 37–51.

Baumer, E. P. 2017. Toward human-centered algorithm design. *Big Data & Society* 4(2): 2053951717718854.

Bowler, L.; Mattern, E.; and Knobel, C. 2014. Developing design interventions for cyberbullying: A narrative-based participatory approach. *iConference 2014 Proceedings* .

Brochado, S.; Soares, S.; and Fraga, S. 2017. A Scoping Review on Studies of Cyberbullying Prevalence Among Adolescents. *Trauma, Violence, & Abuse* 18(5): 523–531. doi:10.1177/1524838016641668. URL https://doi.org/10.1177/1524838016641668.

Chatzakou, D.; Kourtellis, N.; Blackburn, J.; De Cristofaro, E.; Stringhini, G.; and Vakali, A. 2017. Mean birds: Detecting aggression and bullying on twitter. In *Proc. WebSci*, 13–22.

Cheng, L.; Li, J.; Silva, Y. N.; Hall, D. L.; and Liu, H. 2019. Xbully: Cyberbullying detection within a multi-modal context. In *Proc. WSDM*, 339–347.

Chung, J.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* .

Coppersmith, G.; Harman, C.; and Dredze, M. 2014. Measuring post traumatic stress disorder in Twitter. In *Proc. ICWSM*.

Crowston, K. 2012. Amazon mechanical turk: A research tool for organizations and information systems scholars. In *Shaping the Future of ICT Research*, 210–221. Springer.

Dadvar, M.; Trieschnigg, D.; Ordelman, R.; and de Jong, F. 2013. Improving cyberbullying detection with user context. In *Proc. ECIR*, 693–696. Springer.

De Choudhury, M.; and De, S. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Proc. ICWSM*.

Dinakar, K.; Reichart, R.; and Lieberman, H. 2011. Modeling the detection of textual cyberbullying. In *Proc. ICWSM*.

Dredge, R.; Gleeson, J.; and De la Piedad Garcia, X. 2014. Cyberbullying in social networking sites: An adolescent victim's perspective. *Computers in human behavior* 36: 13–20.

Ellison, N. B.; Steinfield, C.; and Lampe, C. 2007. The benefits of Facebook "friends:" Social capital and college students' use of online social network sites. *Journal of computer-mediated communication* 12(4): 1143–1168.

Ernala, S. K.; Rizvi, A. F.; Birnbaum, M. L.; Kane, J. M.; and De Choudhury, M. 2017. Linguistic markers indicating therapeutic outcomes of social media disclosures of schizophrenia. *Proc. ACM-HCI* 1(CSCW): 1–27.

Founta, A. M.; Chatzakou, D.; Kourtellis, N.; Blackburn, J.; Vakali, A.; and Leontiadis, I. 2019. A unified deep learning architecture for abuse detection. In *Proc. WebSci*, 105–114.

Founta, A. M.; Djouvas, C.; Chatzakou, D.; Leontiadis, I.; Blackburn, J.; Stringhini, G.; Vakali, A.; Sirivianos, M.; and Kourtellis, N. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proc. ICWSM*.

Gualdo, A. M. G.; Hunter, S. C.; Durkin, K.; Arnaiz, P.; and Maquilón, J. J. 2015. The emotional impact of cyberbullying: Differences in perceptions and experiences as a function of role. *Computers & Education* 82: 228–235.

Hamm, M. P.; Newton, A. S.; Chisholm, A.; Shulhan, J.; Milne, A.; Sundar, P.; Ennis, H.; Scott, S. D.; and Hartling, L. 2015. Prevalence and effect of cyberbullying on children and young people: A scoping review of social media studies. *JAMA pediatrics* 169(8): 770–777.

Hosseinmardi, H.; Mattson, S. A.; Rafiq, R. I.; Han, R.; Lv, Q.; and Mishra, S. 2015. Detection of cyberbullying incidents on the instagram social network. *arXiv:1503.03909* .

Huang, Q.; Singh, V. K.; and Atrey, P. K. 2014. Cyber Bullying Detection Using Social and Textual Analysis. In *Proceedings of the 3rd International Workshop on Socially-Aware Multimedia*, SAM '14, 3–6. New York, NY, USA: ACM. ISBN 9781450331241. doi:10.1145/2661126.2661133. URL https://doi-org.ezproxy.net.ucf.edu/10.1145/2661126.2661133.

Kwak, H.; Blackburn, J.; and Han, S. 2015. Exploring cyberbullying and other toxic behavior in team competition online games. In *Proc. CHI*, 3739–3748.

Lenhart, A.; Smith, A.; Anderson, M.; Duggan, M.; and Perrin, A. 2015. Teens, technology and friendships .

Manning, C. D.; Surdeanu, M.; Bauer, J.; Finkel, J. R.; Bethard, S.; and McClosky, D. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proc. ACL*, 55–60.

Pennebaker, J. W.; Francis, M. E.; and Booth, R. J. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71(2001): 2001.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Power, A.; Keane, A.; Nolan, B.; and O'Neill, B. 2019. The Role of Previous Discourse in Identifying Public Textual Cyberbullying. *Journal of Computer-Assisted Linguistic Research* 3(1): 1–20.

Pruksachatkun, Y.; Pendse, S. R.; and Sharma, A. 2019. Moments of change: Analyzing peer-based cognitive support in online mental health forums. In *Proc. CHI*, 1–13.

Rafiq, R.; Hosseinmardi, H.; Han, R.; Lv, Q.; Mishra, S.; and Mattson, S. 2015. Careful what you share in six sec: Detecting cyberbullying instances in Vine. In *Proc. ASONAM*, 617–622.

Razi, A.; Badillo-Urquiola, K.; and Wisniewski, P. J. 2020. Let's Talk about Sext: How Adolescents Seek Support and Advice about Their Online Sexual Experiences, In *Proc. CHI*. In *Proc. CHI*, 1–13. ISBN 9781450367080. doi:10.1145/3313831.3376400. URL https://doi-org.ezproxy.net.ucf.edu/10.1145/3313831.3376400.

Rosa, H.; Pereira, N.; Ribeiro, R.; Ferreira, P. C.; Carvalho, J. P.; Oliveira, S.; Coheur, L.; Paulino, P.; Simão, A. V.; and Trancoso, I. 2019. Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior* 93: 333–345.

Saha, K.; Chandrasekharan, E.; and De Choudhury, M. 2019. Prevalence and psychological effects of hateful speech in online college communities. In *Proc. WebSci*, 255–264.

Saha, K.; and Sharma, A. 2020. Causal Factors of Effective Psychosocial Outcomes in Online Mental Health Communities. In *Proc. ICWSM*.

Salawu; He; and Lumsden. 2020. Approaches to Automated Detection of Cyberbullying: A Survey. *IEEE Transactions on Affective Computing* 11(1): 3–24.

Schenk, E.; Guittard, C.; et al. 2009. Crowdsourcing: What can be Outsourced to the Crowd, and Why. In *Workshop on open source innovation, Strasbourg, France*, volume 72, 3. Citeseer.

Schwartz, H. A.; Eichstaedt, J. C.; Kern, M. L.; Dziurzynski, L.; Ramones, S. M.; Agrawal, M.; Shah, A.; Kosinski, M.; Stillwell, D.; Seligman, M. E.; et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one* 8(9): e73791.

Singh, V. K.; Ghosh, S.; and Jose, C. 2017. Toward multimodal cyberbullying detection. In *Proc. CHI-EA*, 2090–2099.

Smith, P. K.; Catalano, R.; Slee, P.; Morita, Y.; Junger-Tas, J.; and Olweus, D. 1999. *The nature of school bullying: A cross-national perspective*. Psychology Press.

Soni, D.; and Singh, V. 2018. Time reveals all wounds: Modeling temporal characteristics of cyberbullying. In *Proc. ICWSM*.

Thomas, H. J.; Connor, J. P.; and Scott, J. G. 2015. Integrating traditional bullying and cyberbullying: challenges of definition and measurement in adolescents–a review. *Educational psychology review* 27(1): 135–152.

Van Cleemput, K.; Vandebosch, H.; and Pabian, S. 2014. Personal characteristics and contextual factors that determine "helping,""joining in," and "doing nothing" when witnessing cyberbullying. *Aggressive behavior* 40(5): 383–396.

Van Royen, K.; Poels, K.; Daelemans, W.; and Vandebosch, H. 2015. Automatic monitoring of cyberbullying on social networking sites: From technological feasibility to desirability. *Telematics and Informatics* 32(1): 89–97.

Zhao, R.; Zhou, A.; and Mao, K. 2016. Automatic detection of cyberbullying on social networks based on bullying features. In *Proc. ICDC&N*, 1–6.

Zhu, H.; Yu, B.; Halfaker, A.; and Terveen, L. 2018. Value-sensitive algorithm design: Method, case study, and lessons. *PACM-HCI* 2(CSCW): 1–23.

Ziems, C.; Vigfusson, Y.; and Morstatter, F. 2020. Aggressive, Repetitive, Intentional, Visible, and Imbalanced: Refining Representations for Cyberbullying Classification. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, 808–819.

Zwierzynska, K.; Wolke, D.; and Lereya, T. S. 2013. Peer victimization in childhood and internalizing problems in adolescence: a prospective longitudinal study. *Journal of abnormal child psychology* 41(2): 309–323.