Distribution Agreement

Caleb Ziems April 8, 2020

Towards More Robust Methods of Cyberbullying Detection

By

Caleb Ziems

Ymir Vigfusson, Ph.D.
Adviser

Department of Computer Science

Ymir Vigfusson, Ph.D.
Adviser

Eugene Agichtein, Ph.D.
Committee Member

Marjorie Pak, Ph.D.
Committee Member

Phillip Wolff, Ph.D.
Committee Member

2020

Towards More Robust Methods of Cyberbullying Detection

By

Caleb Ziems

Ymir Vigfusson, Ph.D.
Adviser

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements for the degree of
Bachelor of Science with Honors

Department of Computer Science

2020

Abstract

Towards More Robust Methods of Cyberbullying Detection

By Caleb Ziems

Cyberbullying is a pervasive problem in online communities. To identify cyberbullying cases in large-scale social networks, content moderators depend on machine learning classifiers for automatic cyberbullying detection. However, existing models remain unfit for real-world applications, largely due to a shortage of publicly available training data and a lack of standard criteria for assigning ground truth labels. In this study, we address the need for reliable data using an original annotation framework. Inspired by social sciences research into bullying behavior, we characterize the nuanced problem of cyberbullying using five explicit factors to represent its social and linguistic aspects. We model this behavior using social network and language-based features, which improves classifier performance. Lastly, we develop a method for inferring the target of aggression in the message thread, and we evaluate this approach on hand-labeled data. These results demonstrate the importance of representing and modeling cyberbullying as a social phenomenon.

Towards More Robust Methods of Cyberbullying Detection

By

Caleb Ziems

Ymir Vigfusson, Ph.D.
Adviser

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements for the degree of
Bachelor of Science with Honors

Department of Computer Science

2020

Acknowledgments

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Agreement No. HR0011890019. The project was completed in part at the USC Information Sciences Institute's REU site, supported by NSF Grant No. 1659886; and at Emory University, supported by NSF Grant No. 1553579.

I would like to thank Dr. Fred Morstatter, my project mentor, for his incredible guidance over the summer at USC, and for his continued support during the academic year. I would also like to thank Dr. Ymir Vigfusson, my thesis advisor, who has been a mentor as well as a friend to me. Both Fred and Ymir have been instrumental in shaping my thesis and developing me as a researcher. Together, we published a significant portion of this work at the 14th International AAAI Conference on Web and Social Media (ICWSM) 2020. We received helpful comments and suggestions on our paper during the review process, and for this, I would like to thank the four anonymous reviewers as well as the ICWSM committee for their consideration. Copyright for this work belongs to the Association for the Advancement of Artificial Intelligence (AAAI) © 2020, although the AAAI does not endorse the authors nor Emory University.

Lastly, I would like to thank my wonderful thesis committee for their investment in my academic and personal growth. Dr. Eugene Agichtein has been a great resource both in and out of the classroom. Dr. Marjorie Pak has been a mentor to me ever since my freshman year when I enrolled in her LING 201 Foundations of Linguistics course. With four more linguistics courses behind us now, I am absolutely sure she shaped and molded my understanding of natural language, and for this I am incredibly grateful. Lastly, I thank Dr. Phillip Wolff for advising me since my very first undergraduate research project. He has helped guide my pursuit of an interdisciplinary research career.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

Cyberbullying poses a serious threat to the safety of online communities. The Centers for Disease Control and Prevention (CDC) identify cyberbullying as a "growing public health problem in need of additional research and prevention efforts" [7]. Cyberbullying has been linked to negative mental health outcomes, including depression, anxiety, and other forms of self-harm, suicidal ideation, suicide attempts, and difficulties with social and emotional processing [18, 24, 29]. Where traditional bullying was once limited to a specific time and place, cyberbullying can occur at any hour and from any location on earth [3]. Once the first message has been sent, the attack can escalate rapidly as harmful content is spread across shared media, compounding these negative effects [14, 36].

Internet users depend on content moderators to flag abusive text and to ban cyberbullies from participating in online communities. However, due to the overwhelming volume of social media data produced every day, manual human moderation is often unfeasible. For this reason, social media platforms are beginning to rely instead on machine learning classifiers for automatic cyberbullying detection [35].

The research community has developed increasingly competitive classifiers to detect harmful or aggressive content in text. Despite significant progress in recent years, however, existing models remain unfit for real-world applications. This is due, in part, to shortcomings in the training and testing data [12, 27, 28]. Most annotation schemes have ignored the importance of social context, and researchers have neglected to provide annotators with objective criteria for distinguishing cyberbullying from other crude messages.

To address the urgent need for reliable data, we provide an original annotation framework and an annotated Twitter dataset.[1] The key advantages to our labeling approach are:

- **Contextually-informed ground truth.** We provide annotators with the social context surrounding each message, including the contents of the reply thread and the account information of each user involved.

- **Clear labeling criteria.** We ask annotators to provide labels for five clear cyberbullying criteria. These criteria can be combined and adapted for revised definitions of cyberbullying.

Using our new dataset, we experiment with baseline NLP features and compare results with a newly-proposed set of features. We designed these features to encode the dynamic relationship between a potential bully and victim, using comparative measures from their relative linguistic and social network profiles. Additionally, our features have low computational complexity, so they can scale to web-scale datasets, unlike expensive network centrality and clustering measurements.

Results from our experiments suggest that, although baseline models can reliably detect aggressive language in text, they will fall short of the more subtle goal of cyberbullying detection. With n-grams and dictionary-based features, classifiers prove unable to detect harmful intent, visibility among peers, power imbalance, or the repetitive nature of aggression with sufficiently high precision and recall. However, our proposed feature set improves F_1 scores on all four of these social measures.

[1] https://github.com/cjziems/cyberbullying-representations

Real-world detection systems can benefit from our proposed approach, incorporating the social aspects of cyberbullying into existing models and training these models on socially-informed ground truth labels.

Finally, we propose a rule-based heuristic algorithm for inferring the cyberbullying target or, more generally, the user account in primary focus given a message in an existing conversational thread. This form of social role detection is a necessary step in our approach and an important step in the direction of more robust and socially-informed cyberbullying detection solutions.

Chapter 2

Background

Existing approaches to cyberbullying detection generally follow a common workflow. Data is collected from social networks or other online sources, and ground truth is established through manual human annotation. Machine learning algorithms are trained on the labeled data using the message text or hand-selected features. Then results are typically reported using precision, recall, and F_1 scores. Comparison across studies is difficult, however, because the definition of cyberbullying has not been standardized. Therefore, an important first step for the field is to establish an objective definition of cyberbullying.

2.1 Defining Cyberbullying

Some researchers view cyberbullying as an extension of more "traditional" bullying behaviors [10, 22, 25]. In one widely-cited book, the psychologist Dan Olweus defines schoolyard bullying in terms of three criteria: **repetition**, **harmful intent**, and an **imbalance of power** [21]. He then identifies bullies by their intention to "inflict injury or discomfort" upon a weaker victim through repeated acts of aggression.

Social scientists have extensively studied this form of bullying as it occurs among adolescents in school [16, 17]. However, experts disagree whether cyberbullying should

Table 2.1: Definitions of Cyberbullying

Work	AGGR	REP	HARM	PEER	POWER
Al-garadi et al. [1]	✓		✓		
Chatzakou et al. [3]	✓	✓	✓		✓
Hosseinmardi et al. [11]	✓	✓			✓
Huang et al. [13]		✓	✓		
Reynolds et al. [26]		✓	✓		
Rosa et al. [27]	✓	✓	✓	✓	
Sugandhi et al. [34]		✓	✓		
Van Hee et al. [35]	✓		✓		

be studied as a form of traditional bullying or a fundamentally different phenomenon [16, 22]. Some argue that, although cyberbullying might involve repeated acts of aggression, this condition might not necessarily hold in all cases, since a single message can be otherwise forwarded and publicly viewed without any repetitive behaviors from the author [32, 36]. Similarly, the role of power imbalance is uncertain in online scenarios. Power imbalances of physical strength or numbers may be less relevant, whereas bully anonymity and the permanence of online messages may be sufficient to render the victim defenseless [31].

The machine learning community has not reached a unanimous definition of cyberbullying either. They have instead echoed the uncertainty of the social scientists. Moreover, some authors have neglected to publish any objective cyberbullying criteria or even a working definition for their annotators, and among those who do, the formulation varies. This disagreement has slowed progress in the field, since classifiers and datasets cannot be as easily compared. Upon review, however, we found that all available definitions contained a strict subset of the following criteria: aggression (AGGR), repetition (REP), harmful intent (HARM), visibility among peers (PEER), and power imbalance (POWER). The datasets built from these definitions are outlined in Table 2.1.

2.2 Existing Cyberbullying Datasets

According to Van Hee et al. [35], data collection is the most restrictive "bottleneck" in cyberbullying research. Because there are very few publicly available datasets, some researchers have turned to crowdsourcing using Amazon Mechanical Turk or similar platforms.

In most studies to date, annotators labeled individual messages instead of message threads, ignoring social context altogether [1, 13, 20, 26, 30, 34]. Only three of the papers that we reviewed incorporated social context in the annotation process. Chatzakou et al. [3] considered batches of time-sorted tweets called *sessions*, which were grouped by user accounts, but they did not include message threads or any other form of context. Van Hee et al. [35] presented "original conversation[s] when possible," but they did not explain when this information was available. And Hosseinmardi et al. [12] was the only study to label full message reply threads as they appeared in the original online source. This information is summarized in Table 2.2 along with the size and the cyberbullying class balance for each dataset.

Table 2.2: Existing Cyberbullying Datasets

Work	**Source**	**Size**	**Balance**	**Context**
Al-garadi et al. [1]	Twitter	10,007	6.0%	✗
Chatzakou et al. [3]	Twitter	9,484	-	✓
Hosseinmardi et al. [11]	Instagram	1,954	29.0%	✓
Huang et al. [13]	Twitter	4,865	1.9%	✗
Reynolds et al. [26]	Formspring	3,915	14.2%	✗
Rosa et al. [27]	Formspring	13,160	19.4%	✗
Sugandhi et al. [34]	Mixed	3,279	12.0%	✗
Van Hee et al. [35]	AskFM	113,698	4.7%	✓

2.3 Modeling Cyberbullying Behavior

A large body of work has been published on cyberbullying detection and prediction, primarily through the use of natural language processing techniques. Most common approaches have relied on lexical features such as n-grams [12, 35, 37], TF-IDF vectors [9, 19, 34], word embeddings [40], or phonetic representations of messages [39], as well as dictionary-based counts on curse words, hateful or derogatory terms, pronouns, emoticons, and punctuation [1, 6, 26, 30]. Some studies have also used message sentiment [30, 34, 35] or the age, gender, personality, and psychological state of the message author according to text from their timelines [1, 6]. These methods have been reported with appreciable success as shown in Table 2.3.

Table 2.3: State of the Art in Cyberbullying Detection. Here, results are reported on either the Cyberbullying (CB) class exclusively or on the entire (total) dataset.

Work	Model	Precision	Recall	F1	Class
Zhang et al. [39]	CNN	99.1%	97.0%	98.0%	total
Al-garadi et al. [1]	Random Forest	94.1%	93.9%	93.6%	total
Nahar et al. [20]	SVM	87.0%	97.0%	92.0%	CB
Sugandhi et al. [34]	SVM	91.0%	91.0%	91.0%	total
Soni and Singh [33]	Naïve Bayes	80.2%	80.2%	80.2%	total
Zhao et al. [40]	SVM	76.8%	79.4%	78.0%	total
Xu et al. [37]	SVM	76.0%	79.0%	77.0%	total
Hosseinmardi et al. [12]	Logistic Regression	78.0%	72.0%	75.0%	CB
Yao et al. [38]	CONcISE	69.5%	79.4%	74.1%	CB
Van Hee et al. [35]	SVM	73.3%	57.2%	64.3%	total
Singh et al. [30]	Proposed	82.0%	53.0%	64.0%	CB
Rosa et al. [27]	SVM	46.0%	-	45.0%	CB
Dadvar et al. [6]	SVM	31.0%	15.0%	20.0%	CB
Huang et al. [13]	Dagging	76.3%	-	-	CB

Some researchers argue, however, that lexical features alone may not adequately represent the nuances of cyberbullying. Hosseinmardi et al. [11] found that in Instagram media sessions containing profane or vulgar content, only 30% were acts of cyberbullying. They also found that while cyberbullying posts contained a moderate proportion of negative terms, the most negative posts were not considered cases of cyberbullying by the annotators. Instead, these negative posts referred to politics, sports, and other domestic matters between friends [11].

The problem of cyberbullying cuts deeper than merely the exchange of aggressive language. The meaning and intent of an aggressive post is revealed through conversation and interaction between peers. Therefore, to properly distinguish cyberbullying from other uses of aggressive or profane language, future studies should incorporate key indicators from the social context of each message. Specifically, researchers can measure the author's status or social advantage, the author's harmful intent, the presence of repeated aggression in the thread, and the visibility of the thread among peers [11, 27, 28].

Since cyberbullying is an inherently social phenomenon, some studies have naturally considered social network measures for classification tasks. Several features have been derived from the network representations of the message interactions. The degree and eigenvector centralities of nodes, the k-core scores, and clustering of communities, as well as the tie strength and betweenness centralities of mention edges have all been shown to improve text-based models [13, 30]. Additionally, bullies and victims can be more accurately identified by their relative network positions. For example, the Jaccard coefficient between neighborhood sets in bully and victim networks has been found to be statistically significant [5]. The ratio of all messages sent and received by each user was also found significant.

These findings show promising directions for future work. Social network features may provide the information necessary to reliably classify cyberbullying. However, it may be prohibitively expensive to build out social networks for each user due to time constraints and the limitations of API calls [38]. For this reason, alternative measurements of online social relationships should be considered.

In the present study, we leverage prior work by incorporating linguistic signals into our classifiers. We extend prior work by developing a dataset that better reflects the definitions of cyberbullying presented by social scientists, and by proposing and evaluating a feature set that represents information pertaining to the social processes that underlie cyberbullying behavior.

Chapter 3

Data

Here, we provide an original annotation framework and a new dataset for cyberbullying research, built to unify existing methods of ground truth annotation. In this dataset, we decompose the complex issue of cyberbullying into five key criteria, which we drew from the social science and machine learning communities. These criteria can be combined and adapted for revised definitions of cyberbullying.

3.1 Data Collection

We collected a sample of 1.3 million unlabeled tweets from the Twitter Filter API. Since cyberbullying is a social phenomenon, we chose to filter for tweets containing at least one "@" mention. To restrict our investigation to original English content, we removed all non-English posts and retweets (RTs), narrowing the size of our sample to 280,301 tweets.

Since aggressive language is a key component of cyberbullying [11], we ran the pre-trained classifier of Davidson et al. [8] over our dataset to identify hate speech and aggressive language and increase the prevalence of cyberbullying examples [1]. This

[1] Without this step, our positive class balance would be prohibitively small. See Appendix A for details.

gave us a filtered set of 9,803 aggressive tweets.

We scraped both the user and timeline data for each author in the aggressive set, as well as any users who were mentioned in one of the aggressive tweets. In total, we collected data from 21,329 accounts. For each account, we saved the full user object, including profile name, description, location, verified status, and creation date. We also saved a complete list of the user's friends and followers, and a 6-month timeline of all their posts and mentions from January 1$^{\text{st}}$ through June 10$^{\text{th}}$, 2019. For author accounts, we extended our crawl to include up to four years of timeline content. Lastly, we collected metadata for all tweets belonging to the corresponding message thread for each aggressive message.

3.2 Annotation Task

We presented each tweet in the dataset to three separate annotators as a Human Intelligence Task (HIT) on Amazon's Mechanical Turk (MTurk) platform. By the time of recruitment, 6,897 of the 9,803 aggressive tweets were accessible from the Twitter web page. The remainder of the tweets had been removed, or the Twitter account had been locked or suspended.

We asked our annotators to consider the full message thread for each tweet as it was displayed on Twitter's web interface. We also gave them a list of up to 15 recent mentions by the author of the tweet, directed towards any of the other accounts mentioned in the original thread. Then we asked annotators to interpret each tweet in light of this social context, and had them provide us with labels for five key cyberbullying criteria. We defined these criteria in terms of the *author* account ("who posted the given tweet?") and the *target* ("who was the tweet about?" – not necessarily the first mention). If the tweet wasn't "about anyone," we said, "just put who the tweet is directed towards." We also stated that "if the target is not on Twitter or their handle cannot be identified" the annotator should "please write *OTHER*." With this framework established, we gave annotators the definitions for our five cyberbullying criteria as follows.

1. **Aggressive language:** (AGGR) Regardless of the author's intent, the language of the tweet could be seen as aggressive. The user either addresses a group or individual, and the message contains at least one phrase that could be described as *confrontational, derogatory, insulting, threatening, hostile, violent, hateful,* or *sexually abusive.*

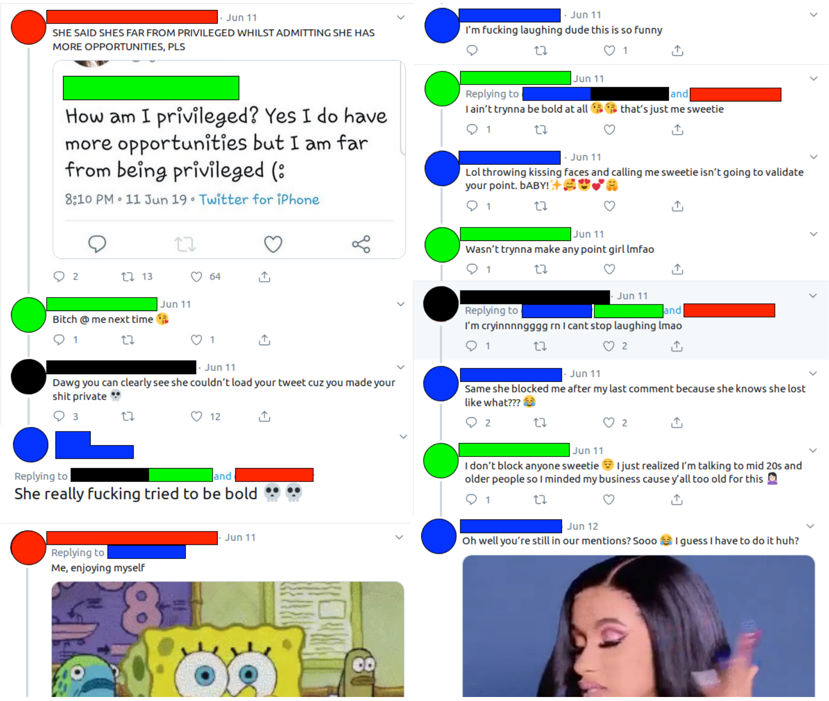
2. **Repetition:** (REP) The target user has received at least two aggressive messages in total (either from the author or from another user in the visible thread).

3. **Harmful intent:** (HARM) The tweet was designed to tear down or disadvantage the target user by causing them distress or by harming their public image. The target does not respond agreeably as to a joke or an otherwise lighthearted comment.

4. **Visibility among peers:** (PEER) At least one other user besides the target has liked, retweeted, or responded to at least one of the author's messages.

5. **Power imbalance:** (POWER) Power is derived from authority and perceived social advantage. Celebrities and public figures are more powerful than common users. Minorities and disadvantaged groups have less power. Bullies can also derive power from peer support.

Each of these criteria was represented as a binary label, except for **power imbalance**, which was ternary. We asked "Is there strong evidence that the **author** is more powerful than the target? Is the **target** more powerful? Or if there is not any good evidence, just mark **equal**." We recognized that an imbalance of power might arise



[redacted] · Jun 11
SHE SAID SHES FAR FROM PRIVILEGED WHILST ADMITTING SHE HAS MORE OPPORTUNITIES, PLS

[redacted]
How am I privileged? Yes I do have more opportunities but I am far from being privileged (:
8:10 PM • 11 Jun 19 • Twitter for iPhone

[redacted] Jun 11
Bitch @ me next time 😘

[redacted] · Jun 11
Dawg you can clearly see she couldn't load your tweet cuz you made your shit private 💀

[redacted]
Replying to [redacted] [redacted] and [redacted]
She really fucking tried to be bold 💀💀

[redacted] · Jun 11
Replying to [redacted]
Me, enjoying myself

[redacted] · Jun 11
I'm fucking laughing dude this is so funny

[redacted] Jun 11
Replying to [redacted] [redacted] and [redacted]
I ain't trynna be bold at all 😘😘 that's just me sweetie

[redacted] · Jun 11
Lol throwing kissing faces and calling me sweetie isn't going to validate your point. bABY!✨🥰😍💕🤗

[redacted] Jun 11
Wasn't trynna make any point girl lmfao

[redacted] · Jun 11
Replying to [redacted] [redacted] and [redacted]
I'm cryinnnngggg rn I cant stop laughing lmao

[redacted] · Jun 11
Same she blocked me after my last comment because she knows she lost like what??? 😂

[redacted] Jun 11
I don't block anyone sweetie 😌 I just realized I'm talking to mid 20s and older people so I minded my business cause y'all too old for this 🤦🏻‍♀️

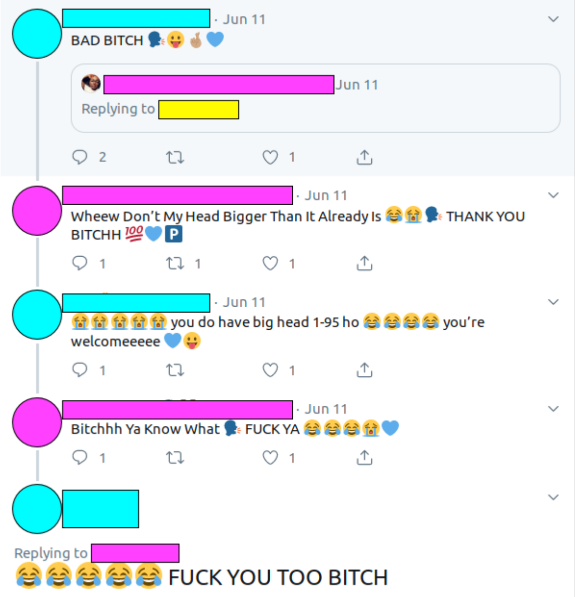
[redacted] Jun 12
Oh well you're still in our mentions? Sooo 😂 I guess I have to do it huh?

Figure 3.1: **Cyberbullying**. This thread demonstrates all five cyberbullying criteria.

in a number of different circumstances. Therefore, we did not restrict our definition to just one form of power, such as follower count or popularity.

For instructional purposes, we provided five sample threads to demonstrate both positive and negative examples for each of the five criteria. Two of these threads are shown here. The thread in Figure 3.1 displays bullying behavior that is targeted against the green user, with all five cyberbullying criteria displayed. The thread includes repeated use of aggressive language such as "she really fucking tried" and "she knows she lost." The bully's harmful intent is evident in the victim's defensive responses. And lastly, the thread is visible among four peers as three gang up against one, creating a power imbalance.

On the other hand, Figure 3.2 shows the importance of context in the annotation process. If we read only the last tweet in the thread, we might decide that the post was cyberbullying, but given the social context here, we can confidently assert that this post is *not* an example of cyberbullying. Although it contains the aggressive



[redacted] · Jun 11
BAD BITCH 🗣😛🤞💙

[redacted] Jun 11
Replying to [redacted]

[redacted] · Jun 11
Wheew Don't My Head Bigger Than It Already Is 😂😭🗣 THANK YOU BITCHH 💯💙🅿

[redacted] · Jun 11
😭😭😭😭😭 you do have big head 1-95 ho 😂😂😂😂 you're welcomeeeee 💙😛

[redacted] · Jun 11
Bitchhh Ya Know What 🗣 FUCK YA 😂😂😂😭💙

[redacted]

Replying to [redacted]
😂😂😂😂😂 FUCK YOU TOO BITCH

Figure 3.2: **Non-cyberbullying**. Although this thread contains repeated use of aggressive language, there is no harmful intent, visibility among peers, or power imbalance.

phrase "F*** YOU TOO B****", the author does not intend harm. The message is part of a joking exchange between two friends or equals, and no other peers have joined in the conversation or interacted with the thread.

After asking workers to review these examples, we gave them a short 7-question quiz to test their knowledge. Workers were given only one quiz attempt, and they were expected to score at least 6 out of 7 questions correctly before they could proceed to the paid HIT. Workers were then paid $0.12 for each thread that they annotated.

We successfully recruited 170 workers to label all 6,897 available threads in our dataset. They labeled an average of 121.7 threads and a median of 7 threads each. They spent an average time of 3 minutes 50 seconds, and a median time of 61 seconds per thread. For each thread, we collected annotations from three different workers, and from this data we computed our reliability metrics using Fleiss's Kappa for inter-annotator agreement as shown in Table 3.1.

We determined ground truth for our data using a 2 out of 3 majority vote as in Hosseinmardi et al. [11]. If the message thread was missing or a *target* user could not

Table 3.1: Analysis of Labeled Twitter Data

Criterion	Positive Balance	Inter-annotator Agreement	Correlation with Bullying
aggression	74.8%	0.23	0.22
repetition	6.6%	0.18	0.27
harmful intent	16.1%	0.42	0.68
visibility among peers	30.1%	0.51	0.07
target power	34.3%	0.37	0.11
author power	3.1%	0.10	-0.02
equal power	59.7%	0.22	-0.09
cyberbullying	0.7%	0.18	–

be identified, we removed the entry from the dataset, since later we would need to draw our features from both the thread and the target profile. After filtering in this way, we were left with 5,537 labeled tweets.

3.3 Cyberbullying Transcends Cyberaggression

As discussed earlier, some experts have argued that cyberbullying is different from online aggression [11, 27, 28]. We asked our annotators to weigh in on this issue by asking them the subjective question for each thread: "Based on your own intuition, is this tweet an example of cyberbullying?" We did not use the cyberbullying label as ground truth for training models; we only used this label to better understand worker perceptions of cyberbullying. Our workers believed cyberbullying depends on a weighted combination of our five criteria, with the strongest correlate being harmful intent as shown in Table 3.1.

Furthermore, the annotators decided that our dataset contained 74.8% aggressive messages as shown in the *Positive Balance* column of Table 3.1. We found that a large majority of these aggressive tweets were not labeled as "cyberbullying." Rather, only 10.5% were labeled by majority vote as cyberbullying, and only 21.5% were considered harmful. From this data, we propose that cyberbullying and cyberaggression are not equivalent classes. Instead, cyberbullying transcends cyberaggression.

Chapter 4

Feature Engineering

We have established that cyberbullying is a complex social phenomenon, different from the simpler notion of cyberaggression. Standard Bag of Words (BoW) features based on single sentences, such as n-grams and word embeddings, may thus lead machine learning algorithms to incorrectly classify friendly or joking behavior as cyberbullying [11, 27, 28]. To more reliably capture the nuances of repetition, harmful intent, visibility among peers, and power imbalance, we designed a new set of features from the social and linguistic traces of Twitter users. These measures were designed to encode the dynamic relationship between the message author and target, using network and timeline similarities, expectations from language models, and other signals taken from the message thread.

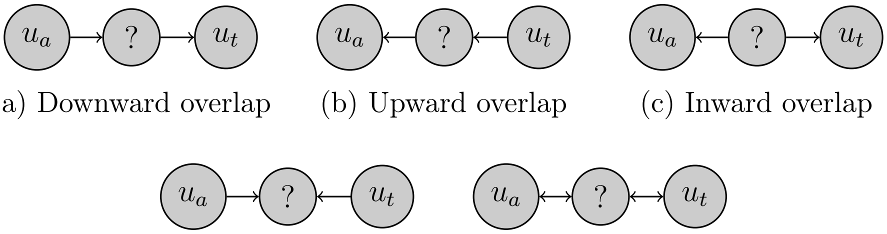
For each feature and each cyberbullying criterion, we compared the cumulative distributions of the positive and negative class using the two-sample Kolmogorov-Smirnov test. We report the Kolmogorov-Smirnov statistic D (a normalized distance between the CDF of the positive and negative class) as well as the p-value with $\alpha = 0.05$ as our level for statistical significance.

4.1 Text-based Features

To construct realistic and competitive baseline models, we considered a set of standard text-based features that have been used widely throughout the literature. Specifically, we used the NLTK library [2] to construct unigrams, bigrams, and trigrams for each labeled message. This parallels the work of Hosseinmardi et al. [12], Van Hee et al. [35], and Xu et al. [37]. Following Zhang et al. [39], we incorporated counts from the Linguistic Inquiry and Word Count (LIWC) dictionary to measure the linguistic and psychological processes that are represented in the text [23]. We also used a modified version of the Flesch-Kincaid Grade Level and Flesch Reading Ease scores as computed in Davidson et al. [8]. Lastly, we encoded the sentiment scores for each message using the Valence Aware Dictionary and sEntiment Reasoner (VADER) of Hutto and Gilbert [15].

4.2 Social Network Features

Network features have been shown to improve text-based models [14, 30], and they can help classifiers distinguish between bullies and victims [5]. These features may also capture some of the more social aspects of cyberbullying, such as power imbalance and visibility among peers. However, many centrality measures and clustering algorithms require detailed network representations. These features may not be scalable for real-world applications. We propose a set of low-complexity measurements that can be used to encode important higher-order relations at scale. Specifically, we measure the relative positions of the author and target accounts in the directed following network by computing modified versions of Jaccard's similarity index as we now explain.



(a) Downward overlap (b) Upward overlap (c) Inward overlap

(d) Outward overlap (e) Bidirectional overlap

Figure 4.1: Graphical representation of the **neighborhood overlap measures** of author u_a and target u_t

4.2.1 Neighborhood Overlap

Let $N^+(u)$ be the set of all accounts followed by user u and let $N^-(u)$ be the set of all accounts that follow user u. Then $N(u) = N^+(u) \cup N^-(u)$ is the neighborhood set of u. We consider five related measurements of neighborhood overlap for a given author u_a and target u_t, listed here.

$$\begin{aligned}
\mathbf{down}(u_a, u_t) &= \tfrac{|N^+(u_a) \cap N^-(u_t)|}{|N^+(u_a) \cup N^-(u_t)|} \\
\mathbf{up}(u_a, u_t) &= \tfrac{|N^-(u_a) \cap N^+(u_t)|}{|N^-(u_a) \cup N^+(u_t)|} \\
\mathbf{in}(u_a, u_t) &= \tfrac{|N^-(u_a) \cap N^-(u_t)|}{|N^-(u_t) \cup N^-(u_t)|} \\
\mathbf{out}(u_a, u_t) &= \tfrac{|N^+(u_a) \cap N^+(u_t)|}{|N^+(u_a) \cup N^+(u_t)|} \\
\mathbf{bi}(u_a, u_t) &= \tfrac{|N(u_a) \cap N(u_t)|}{|N(u_a) \cup N(u_t)|}
\end{aligned}$$

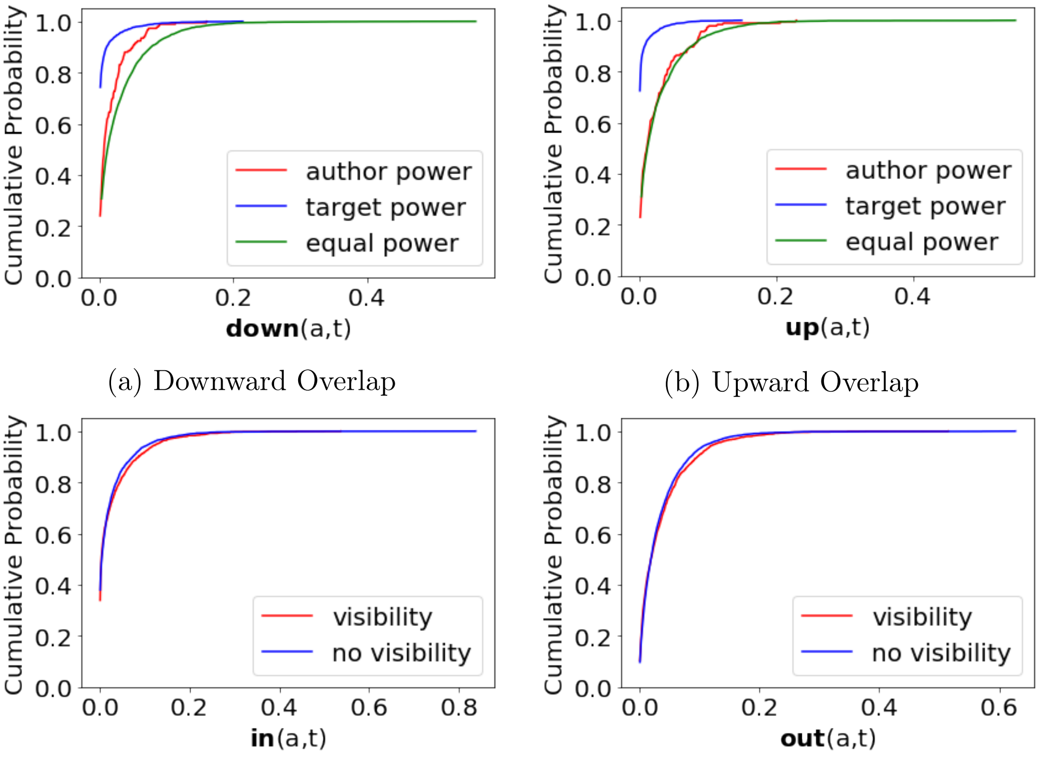
Downward overlap measures the number of two-hop paths from the author to the target along following relationships; upward overlap measures two-hop paths in the opposite direction. Inward overlap measures the similarity between the two users' follower sets, and outward overlap measures the similarity between their sets of friends. Bidirectional overlap then is a more generalized measure of social network similarity. We provide a graphical depiction for each of these features on the right side of Figure 4.1.

High downward overlap likely indicates that the target is socially relevant to the author, as high upward overlap indicates the author is relevant to the target. Therefore, when the author is more powerful, downward overlap is expected to be lower and upward overlap is expected be higher. This trend is slight but visible in the cumulative distribution functions of Figure 4.2 (a): downward overlap is indeed lower when the author is more powerful than when the users are equals ($D = 0.143$). However, there is not a significant difference for upward overlap ($p = 0.85$). We also observe that, when the target is more powerful, downward and upward overlap are both significantly lower ($D = 0.516$ and $D = 0.540$ respectively). It is reasonable to assume that messages can be sent to celebrities and other powerful figures without the need for common social connections.

Next, we consider inward and outward overlap. When the inward overlap is high, the author and target could have more common visibility. Similarly, if the outward overlap is high, then the author and target both follow similar accounts, so they might have similar interests or belong to the same social circles. Both inward and outward overlaps are expected to be higher when a post is visible among peers. This is true of both distributions in Figure 4.2. The difference in outward overlap is significant ($D = 0.04$, $p = 0.03$), and the difference for inward overlap is short of significant ($D = 0.04$, $p = 0.08$).

4.2.2 User-based features

We also use basic user account metrics drawn from the author and target profiles. Specifically, we count the friends and followers of each user, their verified status, and the number of tweets posted within six-month snapshots of their timelines, as in Al-garadi et al. [1], Chatzakou et al. [3], and Hosseinmardi et al. [12].



(a) Downward Overlap

(b) Upward Overlap

(c) Inward Overlap

(d) Outward Overlap

Figure 4.2: Cumulative Distribution Functions for **neighborhood overlap** on relevant features. These measures are shown to be predictive of *power imbalance* and *visibility among peers*.

4.3 Timeline Features

Here, we consider linguistic features, drawn from both the author and target timelines. These are intended to capture the social relationship between each user, their common interests, and the surprise of a given message relative to the author's timeline history.

4.3.1 Message Behavior

To more clearly represent the social relationship between the author and target users, we consider the messages sent between them as follows:

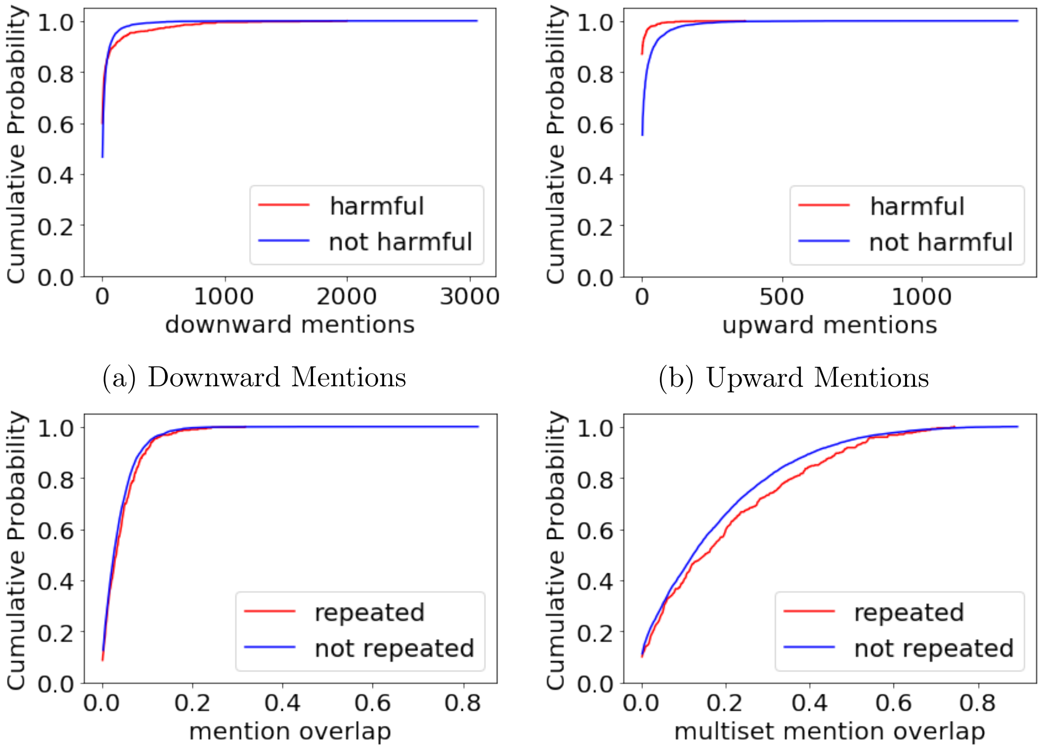
- *Downward mention count:* How many messages has the author sent to the target?

- *Upward mention count:* How many messages has the target sent to the author?

- *Mention overlap:* Let M_a be the set of all accounts mentioned by author a, and let M_t be the set of all accounts mentioned by target t. We compute the ratio $\frac{|M_a \cap M_t|}{|M_a \cup M_t|}$.

- *Multiset mention overlap*: Let $\hat{M}_a$ be the multiset of all accounts mentioned by author a (with repeats for each mention), and let $\hat{M}_t$ be the multiset of all accounts mentioned by target t. We measure $\frac{|\hat{M}_a \cap^* \hat{M}_t|}{|\hat{M}_a \cup \hat{M}_t|}$ where $\cap^*$ takes the multiplicity of each element to be the sum of the multiplicity from $\hat{M}_a$ and the multiplicity from $\hat{M}_b$

The direct mention count measures the history of repeated communication between the author and the target. For harmful messages, downward overlap is higher ($D = 0.178$) and upward overlap is lower ($D = 0.374$) than for harmless messages, as shown in Figure 4.3. This means malicious authors tend to address the target repeatedly while the target responds with relatively few messages.



(a) Downward Mentions

(b) Upward Mentions

(c) Mention Overlap

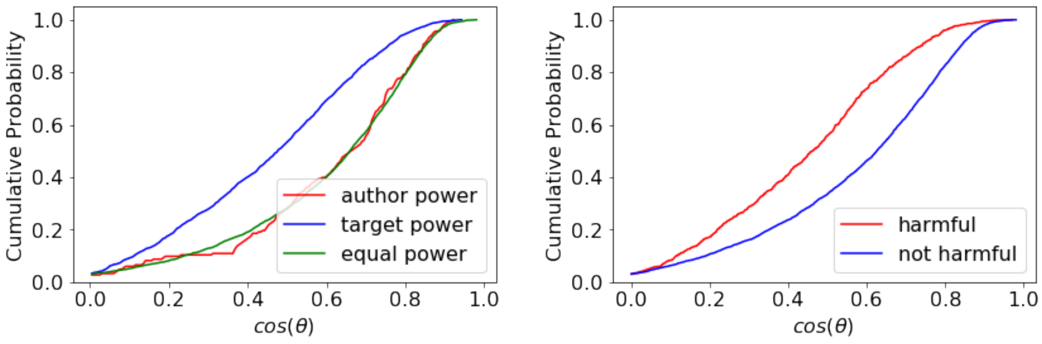
(d) Multiset Mention Overlap

Figure 4.3: Cumulative Distribution Functions for **message behavior** on relevant features. These measures are shown to be indicative of *harmful intent* and *repetition*.

Mention overlap is a measure of social similarity that is based on shared conversations between the author and the target. Multiset mention overlap measures the frequency of communication within this shared space. These features may help predict visibility among peers, or repeated aggression due to pile-on bullying situations. We see in Figure 4.3 that repeated aggression is linked to slightly greater mention overlap ($D = 0.07$, $p = 0.07$), but the trend is significant only for multiset mention overlap ($D = 0.08$, $p = 0.03$).

4.3.2 Timeline Similarity

Timeline similarity is used to indicate common interests and shared topics of conversation between the author and target timelines. High similarity scores might reflect users' familiarity with one another, or suggest that they occupy similar social posi-



(a) Timeline Similarity

(b) Timeline Similarity

Figure 4.4: Cumulative Distribution Functions for **timeline similarity** on relevant features. These measures are shown to be predictive of *power imbalance* and *harmful intent*.

tions. This can be used to distinguish cyberbullying from harmless banter between friends and associates. To compute this metric, we represent the author and target timelines as TF-IDF vectors $\vec{A}$ and $\vec{T}$. We then take the **cosine similarity** between the vectors as

$$\cos\theta = \frac{\vec{A} \cdot \vec{T}}{\|\vec{A}\|\|\vec{T}\|}.$$

A cosine similarity of 1 means that users' timelines had identical counts across all weighted terms; a cosine similarity of 0 means that their timelines did not contain any words in common. We expect higher similarity scores between friends and associates.

In Figure 4.4 (a), we see that the timelines were significantly less similar when the target was in a position of greater power ($D = 0.294$). This is not surprising, since power can be derived from such differences between social groups. We do not observe the same dissimilarity when the author was more powerful ($p = 0.58$). What we do observe is likely caused by noise from extreme class imbalance and low inter-annotator agreement on labels for author power.

Turning to Figure 4.4 (b), we see that aggressive messages were less likely to harbor harmful intent if they were sent between users with similar timelines ($D = 0.285$). Aggressive banter between friends is generally harmless, so again, this confirms our intuitions.

4.3.3 Language Models

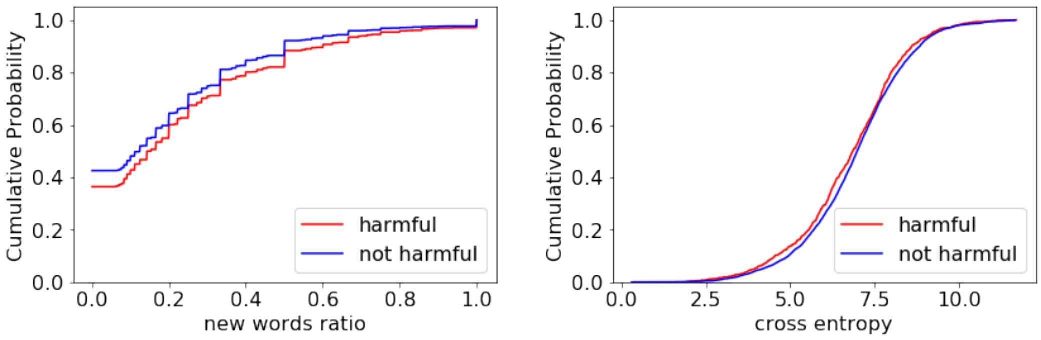
Harmful intent is difficult to measure in isolated messages because social context determines pragmatic meaning. We attempt to approximate the author's harmful intent by measuring the linguistic "surprise" of a given message relative to the author's timeline history. We do this in two ways: through a simple ratio of new words, and through the use of language models.

To estimate historical language behavior, we count unigram and bigram frequencies from a 4-year snapshot of the author's timeline. Then, after removing all URLs, punctuation, stop words, mentions, and hashtags from the original post, we take the cardinality of the set unigrams in the post having zero occurrences in the timeline. Lastly, we divide this count by the length of the processed message to arrive at our **new words ratio**. We can also build a language model from the bigram frequencies, using Kneser-Ney smoothing as implemented in `NLTK` [2]. From the language model, we compute the surprise of the original message m according to its **cross-entropy**, given by

$$\mathrm{H}(m) = -\frac{1}{N}\sum_{i=1}^{N} \log P(b_i)$$

where m is composed of bigrams $b_1, b_2, \ldots, b_N$, and $P(b_i)$ is the probability of the ith bigram from the language model.

We see in Figure 4.5 that harmfully intended messages have a greater density of new words ($D = 0.06$). This is intuitive, since attacks may be staged around new topics of conversation. However, the cross entropy of these harmful messages is slightly lower than for harmless messages ($D = 0.06$). This may be due to harmless jokes, since joking messages might depart more from the standard syntax of the author's timeline.



(a) New Words Ratio

(b) Cross Entropy

Figure 4.5: Cumulative Distribution Functions for **language models** on relevant features. These measures are shown to be predictive of *harmful intent*.

4.4 Thread Features

Finally, we turn to the messages of the thread itself to compute measures of visibility and repeated aggression.

4.4.1 Visibility

To determine the public visibility of the author's post, we collect basic measurements from the interactions of other users in the thread. They are as follows.

- *Message count*: Count the messages posted in the thread
- *Reply message count*: Count the replies posted in the thread after the author's first comment.
- *Reply user count*: Count the users who posted a reply in the thread after the author's first comment.
- *Maximum author favorites*: The largest number of favorites the author received on a message in the thread.
- *Maximum author retweets*: The largest number of retweets the author received on a message in the thread.

4.4.2 Aggression

To detect repeated aggression, we again employ the hate speech and offensive language classifier of Davidson et al. [8]. Each message is given a binary label according to the classifier-assigned class: aggressive (classified as hate speech or offensive language) or non-aggressive (classified as neither hate speech nor offensive language). From these labels, we derive the following features.

- *Aggressive message count*: Count the messages in the thread classified as aggressive

- *Aggressive author message count*: Count the author's messages that were classified as aggressive

- *Aggressive user count*: Of the users who posted a reply in the thread after the author first commented, count how many had a message classified as aggressive

Chapter 5

Model Evaluation

5.1 Experiments

Using our proposed features from Chapter 4 and ground truth labels from our annotation task in Chapter 3, we trained a separate Logistic Regression classifier for each of the five cyberbullying criteria, and we report precision, recall, and F_1 measures over each binary label independently. We averaged results using five-fold cross-validation, with 80% of the data allocated for training and 20% of the data allocated for testing at each iteration. To account for the class imbalance in the *training* data, we used the synthetic minority over-sampling technique (SMOTE) [4]. We did not over-sample testing sets, however, to ensure that our tests better match the class distributions obtained as we did by pre-filtering for aggressive directed Twitter messages.

We compare our results across the five different feature combinations given in Table 5.1. Note that because we do not include thread features in the *User* set, it can be used for cyberbullying prediction and early intervention. The *Proposed* set can be used for detection, sinct it is a collection of all newly proposed features, including thread features. The *Combined* adds these to the baseline text features.

Table 5.1: Feature Combinations

Feature	*BoW*	*Text*	*User*	*Proposed*	*Combined*
n-grams	✓	✓			✓
LIWC, VADER, Flesch-Kincaid		✓			✓
Friend/following counts, tweet count, verified			✓	✓	✓
Neighborhood overlap measures			✓	✓	✓
Mention counts and overlaps			✓	✓	✓
Timeline similarity			✓	✓	✓
New words ratio, cross-entropy			✓	✓	✓
Thread visibility features				✓	✓
Thread aggression features				✓	✓

5.2 Results

The performance of the different classifiers is summarized in Tables 5.2, 5.3, and 5.4. Here, we see that Bag of Words and text-based methods performed well on the aggressive language classification task, with an F_1 score of 83.5%. This was expected and the score aligns well with the success of other published results of Table 2.3.

Cyberbullying detection is more complex than simply identifying aggressive text, however. We find that these same baseline methods fail to reliably detect repetition, harmful intent, visibility among peers, and power imbalance, as shown by the low recall scores in Table 5.3. We conclude that our investigation of socially informed features was justified.

Our proposed set of features beats recall scores for lexically trained baselines in all but the aggression criterion. We also improve precision scores for repetition, visibility among peers, and power imbalance. When we combine all features, we see our F_1 scores beat baselines for each criterion. This demonstrates the effectiveness of our approach, using linguistic similarity and community measurements to encode social characteristics for cyberbullying classification.

Although we achieve moderately competitive scores in most categories, our classifiers are still over-classifying cyberbullying cases. Precision scores are generally much lower than recall scores across all models. To reduce our misclassification of false

Table 5.2: Logistic Regression Precision

Criterion	*BoW*	*Text*	*User*	*Proposed*	*Combined*
aggression	82.5%	82.3%	77.1%	78.7%	**82.6%**
repetition	7.8%	13.4%	7.7%	15.3%	**31.7%**
harmful intent	29.6%	49.4%	35.8%	34.5%	**55.3%**
visibility among peers	30.8%	34.3%	34.0%	42.2%	**46.8%**
author power	1.9%	3.6%	7.6%	9.8%	**17.0%**
target power	43.5%	51.5%	**77.6%**	75.2%	77.0%

Table 5.3: Logistic Regression Recall

Criterion	*BoW*	*Text*	*User*	*Proposed*	*Combined*
aggression	77.0%	84.8%	47.8%	51.6%	**85.6%**
repetition	17.6%	7.3%	49.5%	**64.3%**	26.2%
harmful intent	40.2%	44.4%	63.4%	**67.7%**	52.7%
visibility among peers	34.8%	20.4%	47.1%	**54.2%**	33.7%
author power	6.5%	1.6%	74.1%	**80.0%**	11.9%
target power	49.4%	43.3%	73.3%	**80.8%**	71.1%

Table 5.4: Logistic Regression F_1 Scores

Criterion	*BoW*	*Text*	*User*	*Proposed*	*Combined*
aggression	79.7%	83.5%	59.0%	62.3%	**84.1%**
repetition	10.8%	9.4%	13.3%	24.7%	**28.7%**
harmful intent	34.1%	46.7%	38.7%	45.7%	**53.8%**
visibility among peers	32.7%	25.5%	39.5%	**47.4%**	45.5%
author power	2.9%	2.2%	13.7%	**17.5%**	14.0%
target power	46.2%	47.0%	75.3%	**77.9%**	73.9%

positives and better distinguish between joking or friendly banter and cyberbullying, it may be necessary to mine for additional social features. Overall, we should work to increase all F_1 scores to above 0.8 before we can consider our classifiers ready for real-world applications [27].

We obtained similar results by replacing our logistic regression model with any of a k-nearest neighbors (KNN) model, random forest, support vector machine (SVM), AdaBoost, or Multilayer Perceptron (MLP). We report all precision, recall, and F_1 scores in Appendix 2, Tables B.1-B.15. We chose to highlight logistic regression here because it can be more easily interpreted. As a result, we can identify the relative importance of our proposed features. The feature weights are also given Tables 5.5-5.9. Here we can observe a trend. The *aggressive language* and *repetition* criteria are dominated by lexical features; the *harmful intent* is split between lexical and historical communication features; and the *visibility among peers* and *target power* criteria are dominated by our proposed social features.

Table 5.5: Top Absolute Weights for Aggressive Language

Rank	*Feature*	*Weight*
1	affect (LIWC)	-1.34
2	sexual (LIWC)	1.07
3	negemo (LIWC)	0.90
4	maximum author retweets	0.86
5	relativ (LIWC)	-0.75
6	bio (LIWC)	-0.69
7	posemo (LIWC)	0.66
8	num chars	-0.64
9	space (LIWC)	0.52
10	upward overlap	0.51

Table 5.6: Top Absolute Weights for Repetition Features

Rank	*Feature*	*Weight*
1	negemo (LIWC)	1.40
2	author verified status	-1.32
3	affect (LIWC)	-1.24
4	cogmech (LIWC)	-0.96
5	relativ (LIWC)	-0.89
6	posemo (LIWC)	0.80
7	social (LIWC)	0.77
8	aggressive user count	0.63
9	upward overlap	0.62
10	number of unique terms	0.61

Table 5.7: Top Absolute Weights for Harmful Intent

Rank	*Feature*	*Weight*
1	number of words	-1.70
2	number of unique terms	1.41
3	bio (LIWC)	-1.05
4	funct (LIWC)	0.95
5	author follower count	-0.90
6	present (LIWC)	0.83
7	you (LIWC)	0.83
8	message count	0.79
9	upward mention count	-0.71
10	verb (LIWC)	-0.67

Table 5.8: Top Absolute Weights for Visibility Among Peers

Rank	*Feature*	*Weight*
1	author follower count	6.29
2	maximum author retweets	-1.63
3	maximum author favorites	1.46
4	aggressive user count	-1.36
5	number of words	-1.16
6	reply user count	1.03
7	number of unique terms	1.02
8	reply message count	-0.91
9	message count	0.77
10	affect (LIWC)	-0.67

Table 5.9: Top Absolute Weights for Target Power

Rank	*Feature*	*Weight*
1	target follower count	2.28
2	author follower count	-1.67
3	bidirectional overlap	-1.22
4	target verified status	1.20
5	upward overlap	-1.11
6	downward overlap	1.04
7	relativ (LIWC)	0.76
8	reply user count	-0.69
9	space (LIWC)	-0.68
10	message count	-0.63

Chapter 6

Inferring The Target User

In Chapter 4, we derived our feature set from the social and linguistic profiles of both the author and target accounts. Implicitly, we assumed we could reliably identify the target user's account. To do so, we used the annotations that were provided to us by Mechanical Turk workers in Chapter 3. For real-world implementation, however, our classification system would need to infer the target user automatically before the drawing the relevant social features from the target profile.

Automatically inferring the target user profile is a challenging NLP task. Here we formally outline the problem, propose an original strategy, and evaluate the results of our preliminary approach. In summary, we take a linguistically informed, rule-based heuristic approach that accounts for pronoun use as well as the reply structure of the conversational thread.

6.1 Problem Formulation

Formally, we can define a message thread as a 6-tuple

$$\mathcal{M} := (U, T, P, \Sigma, \sigma, \tau)$$

where U is the set of user account names, T is the set of post times, $P \subseteq U \times \mathcal{P}(U) \times T$ is the set of directed post interactions, and Σ is the set of message strings, while $\sigma : P \to \Sigma$ maps posts to message strings, and $\tau : P \to T$ maps posts to post times.

Here, each post $p_k \in P$ is a 3-tuple given by (u_k, R_k, t_k) where $u_k \in U$ represents the *author* user account name, $R_k := [u_r]$ gives the ordered list of all recipient user accounts u_r, and t_k gives the time that post p_k was sent. Therefore $\tau(p_k) = t_k$.

Now, given a focus post $f \in P$, we consider the ordered sequence of posts $S := p_1, p_2, \ldots, p_n$ such that $p_1 = f$ and $\tau(p_i) < \tau(p_{i+1})\ \forall i \in \{1, 2, \ldots, n-1\}$. From S, our objective is to identify a user $u_t \in U$ about whom post f was created primarily. We call u_t the *target* user.

6.2 Proposed Algorithm

Building on the formulation outlined in the previous subsection, we will now detail our target identification algorithm Targ($\mathcal{M}$, f) here as Algorithm 1. We will now give a basic overview of our heuristic approach along with the motivation for the decisions we made.

We initialize a variable `user_pool` to store the list of usernames that are candidates for the potential target. At the start of our algorithm, all users are considered as candidates except for u_f, the author of the message in focus.

Next, we iterate forward in time through the thread, considering each post $p_i \in S$. We use the `stanfordnlp` library to run a dependency parse of the string $\sigma(p_i)$. Then we consider each word in the dependency parse in reverse sentence order, scanning for subject dependencies. We scan the string in reverse order because we intend to select as subject the more deeply nested phrases, and these phrases are more likely to appear towards the right of the string. If we find a subject, and the subject is a noun (NNP, NNPS), we try matching it to one of the known usernames in U and return.

Algorithm 1: Targ($\mathcal{M}$, f)

```
Data: Message thread $\mathcal{M} := (U, T, P, \Sigma, \sigma, \tau)$ with focus message
      $f := (u_f, R_f, t_f) \in \mathcal{M}$
Result: Target user $u_t$
initialize $S := p_1, p_2, \ldots, p_n$ such that $p_1 = f$ and $\tau(p_i) < \tau(p_{i+1})$;
user_pool $\leftarrow U \setminus \{u_f\}$;
for $p_i := (u_{r_i}, R_i, t_i) \in S$ do
    Initialize i_subj, you_subj, has_subj, has_verb to False;
    for $w_j \in reverse(dep\_parse(\sigma(p_i)))$ do
        if $pos(w_j) \in [VB, VBZ, VBP, VBD]$ then
            has_verb $\leftarrow$ True;
        end
        else if $dep(w_j) \in [nsubj, nsubjpass, vocative]$ then
            has_subj $\leftarrow$ True; if $pos(w_j) \in [NNP, NNPS]$ then
                return $match\_string(w_j, U)$;
            end
            else if $pos(w_j) = PRP$ then
                if $pers(w_j) = 1$ then
                    i_subj $\leftarrow$ True;
                end
                else if $pers(w_j) = 2$ then
                    you_subj $\leftarrow$ True;
                end
                else if $pers(w_j) = 3$ then
                    user_pool $\leftarrow$ user_pool $\setminus \{R_i[0]\}$;
                end
            end
        end
    end
    if you_subj or (has_verb and not has_subj) then
        return $R_i[0]$;
    end
    if not you_subj then
        user_pool $\leftarrow$ user_pool $\setminus \{R_i[0]\}$;
    end
end
return user_pool[0];
```

If, instead, the given word is personal pronoun (PRP), we mark either first person (`i_subj`) or second person (`you_subj`) accordingly. If the pronoun is third person, we remove the first direct recipient user from the `user_pool` because the author was not likely talking about them.

If we have considered each word in the given post p_i, we have reached the end of the inner loop. If the post was found to have a `you_subj` or an implied second person subject from the combination of `has_verb` and NOT `has_subj`, we just return the first direct recipient user, since the post was explicitly directed towards them. Instead, if there was no `you_subj`, we remove the first direct recipient user from the pool of target candidates since the thread likely did not, at this time step, directly pertain to them.

6.3 Results

We evaluate our proposed model for accuracy using two different label sets. The **MTurk Dataset** we draw directly from Chapter 3 using labels from the original annotation process. The **Hand-Annotated Subset** represents a subset of the original MTurk dataset where we hand-annotated 95 randomly sampled threads. We compare our *Proposed Algorithm* from above with a baseline *First Mention* classifier that simply reports the first recipient – that is, the first direct mention – found in message.

Model	MTurk Dataset	Hand-Annotated Subset
Proposed Algorithm	30.0	**68.4**
First Mention	**91.7**	6.3

Table 6.1: Model Accuracy for Target Inference

We see that for the MTurk dataset, the baseline algorithm surprisingly achieves very high accuracy. This is likely because our annotators defaulted to the first mention when in cases of uncertainty. However, we also see that our approach far surpasses first mention guessing on our hand-annotated subset.

Chapter 7

Conclusion

In this study, we produced an original dataset for cyberbullying detection research. Our labeling scheme was designed to flexibly accommodate the collection of varied but related cyberbullying definitions that have been proposed throughout the literature. In order to accurately represent the nature of cyberbullying, we decomposed this complex issue into five representative characteristics. Our classes distinguish cyberbullying from other related behaviors, such as isolated aggression or crude joking. To help annotators infer these distinctions, we provided them with the full context of each message's reply thread, along with a list of the author's most recent mentions. In this way, we secured a set of reliable labels for more unambiguous cyberbullying representation.

From these ground truth labels, we designed a new set of features to account for each of the five cyberbullying criteria. Unlike previous text-based or user-based features, our feature set encodes the relationship between a message author and target. We show that these features significantly improve the performance of standard text-based models. These results demonstrate the relevance of social-network and language-based measurements to account for the nuanced social characteristics of cyberbullying.

Despite improvements over baseline, our classifiers have not yet attained the high levels of precision and recall that should be expected of real-world detection systems. For this reason, we argue that the challenging task of cyberbullying detection remains an open research problem. We make our dataset publicly available so that it may be used to train more reliable cyberbullying detection models.

Lastly, we proposed a new algorithm for inferring the cyberbullying target from noun and pronoun referents and the structure of the message thread. This step was necessary for a model like ours, which draws relevant features from both the author and target profiles as well as the relationship between them.

7.1 Limitations

Our study was focused on the Twitter ecosystem and a small part of its network. The initial sampling of tweets was based on a machine learning classifier of aggressive English language with a reported F1 score of 0.90 [8]. Even with this filter, only 0.7% of tweets were deemed by a majority of MTurk workers as cyberbullying (Table 3.1). This extreme class imbalance can disadvantage a wide range of machine learning models. Moreover, the MTurk workers exhibited only moderate inter-annotator agreement (Table 3.1).

We acknowledge that notions of *harmful intent* and *power imbalance* can be subjective, since they may depend on the particular conventions or social structure of a given community. For these reasons, we recognize that cyberbullying still has not been unambiguously defined. Moreover, the underlying constructs are difficult to identify. In this study, we did not train workers to recognize subtle cues for interpersonal popularity, nor the role of anonymity in creating a power imbalance.

Furthermore, because we lack the authority to define cyberbullying, we cannot assert a two-way implication between cyberbullying and the five criteria outlined here.

It may be possible for cyberbullying to exist with only one criterion present, such as harmful intent. Our five criteria also might not span all of the dimensions of cyberbullying. However, they are representative of the literature in both the social science and machine learning communities, and they can be used in weighted combinations to accommodate new definitions.

The main contribution of our paper is not that we solved the problem of cyberbullying detection. Instead, we have exposed the challenge of defining and measuring cyberbullying activity, which has been historically overlooked in the research community. Furthermore, we have proposed a target identification algorithm and a new set of features that can be used together to augment future efforts towards more robust and socially-informed methods for cyberbullying detection.

7.2 Future Work

Cyberbullying detection is an increasingly important and yet challenging problem to tackle. A lack of detailed and appropriate real-world datasets stymies progress towards more reliable detection methods. With cyberbullying being a systemic issue across social media platforms, we urge the development of a methodology for data sharing with researchers that provides adequate access to rich data to improve on the early detection of cyberbullying while also addressing the sensitive privacy issues that accompany such instances.

Once the data becomes available, the first major objective for future work is to increase performance through the use of new features or improved models. New features could include approximations of higher-order social network measures like community clustering, edge centrality, or node similarity scores [5]. Other features might be drawn from related image content [12], including links and profile images. New models might take a number of different approaches. Previously, Singh et al. [30]

used a fusion approach to account for inter-dependencies between features, and Soni and Singh [33] used point processes to represent the temporal dynamics of message threads. Neural networks might also prove successful, following the results of Zhang et al. [39].

The second major objective is to achieve and demonstrate computational efficiency. Detection algorithms must be fast in order to scale to real-world systems. Towards this end, some researchers have considered online algorithms [38] as well as semi-supervised approaches [20]. Future studies should include detailed estimates on the run times of their proposed methods, including the time needed to make API calls.

Appendix A

Real-World Class Distribution

To understand the real-world class distribution for the cyberbullying criteria, we randomly selected 222 directed English tweets from an unbiased sample of drawn from the Twitter Decahose stream across the entire month of October 2016. Using the same methodology given in the paper, we had these tweets labeled three times each on Amazon Mechanical Turk. Again, ground truth was determined using 2 out of 3 majority vote. Upon analysis, we found that the positive class balance was prohibitively small, especially for *repetition*, *harmful intent*, *visibility among peers*, and *author power*, which were all under 5%.

Table A.1: Analysis of Unfiltered Decahose Data

Criterion	**Positive Balance**	**Inter-annotator Agreement**	**Correlation with Bullying**
aggression	6.3%	0.23	0.68
repetition	0.9%	0.04	0.46
harmful intent	1.4%	0.31	0.75
visibility among peers	0.17%	0.51	0.11
target power	22.5%	0.23	0.11
author power	3.6%	0.04	0.06
equal power	64.7%	0.15	-0.14
cyberbullying	2.7%	0.25	-

Appendix B

Model Evaluation

For the sake of comparison, we provide precision, recall, and F_1 scores for five different machine learning models: k-nearest neighbors (KNN), random forest, support vector machine (SVM), AdaBoost, and Multilayer Perceptron (MLP). Then we provide feature weights for our logistic regression model trained on each of the five cyberbullying criteria.

Table B.1: KNN Precision

Criterion	*BoW*	*Text*	*User*	*Proposed*	*Combined*
aggression	**86.0%**	82.7%	74.6%	76.0%	82.6%
repetition	6.9%	7.4%	7.0%	**13.3%**	8.9%
harmful intent	19.8%	21.2%	**29.7%**	29.2%	23.4%
visibility among peers	30.8%	31.4%	30.1%	**34.7%**	32.4%
target power	37.0%	38.2%	64.0%	**64.1%**	49.1%

Table B.2: Random Forest Precision

Criterion	*BoW*	*Text*	*User*	*Proposed*	*Combined*
aggression	77.6%	**80.1%**	78.3%	78.7%	79.7%
repetition	6.5%	6.8%	7.7%	**16.1%**	10.8%
harmful intent	18.4%	28.1%	33.2%	33.4%	**43.1%**
visibility among peers	28.7%	32.7%	34.8%	**42.8%**	35.1%
target power	39.3%	43.3%	**77.9%**	74.5%	69.6%

Table B.3: SVM Precision

Criterion	*BoW*	*Text*	*User*	*Proposed*	*Combined*
aggression	84.1%	**88.1%**	77.4%	79.2%	86.6%
repetition	6.7%	7.0%	6.9%	16.7%	**20.1%**
harmful intent	17.9%	21.7%	33.7%	**34.4%**	30.5%
visibility among peers	29.8%	30.6%	33.9%	40.2%	**40.9%**
target power	36.2%	39.8%	**75.4%**	71.3%	47.8%

Table B.4: AdaBoost Precision

Criterion	*BoW*	*Text*	*User*	*Proposed*	*Combined*
aggression	**82.6%**	81.6%	77.0%	77.5%	81.6%
repetition	7.8%	9.0%	7.3%	16.6%	**25.8%**
harmful intent	29.1%	46.4%	34.3%	39.9%	**60.0%**
visibility among peers	30.5%	32.9%	35.9%	45.8%	**46.1%**
target power	42.5%	46.5%	78.0%	**78.2%**	77.9%

Table B.5: MLP Precision

Criterion	*BoW*	*Text*	*User*	*Proposed*	*Combined*
aggression	**82.8%**	78.8%	76.7%	77.4%	78.3%
repetition	7.7%	8.7%	8.6%	16.9%	**19.6%**
harmful intent	27.4%	42.8%	37.3%	38.4%	**46.8%**
visibility among peers	30.1%	34.0%	34.3%	**41.6%**	38.5%
target power	39.6%	45.2%	**74.3%**	72.0%	68.6%

Table B.6: KNN Recall

Criterion	*BoW*	*Text*	*User*	*Proposed*	*Combined*
aggression	34.8%	37.8%	**58.3%**	56.6%	33.5%
repetition	65.5%	63.3%	29.0%	45.2%	**70.1%**
harmful intent	75.7%	77.2%	56.5%	56.0%	**82.5%**
visibility among peers	70.6%	74.0%	43.7%	48.4%	**78.1%**
target power	71.3%	73.7%	72.4%	75.0%	**85.0%**

Table B.7: Random Forest Recall

Criterion	*BoW*	*Text*	*User*	*Proposed*	*Combined*
aggression	56.4%	**78.5%**	43.7%	45.3%	76.2%
repetition	36.2%	24.9%	46.3%	**64.7%**	29.9%
harmful intent	42.4%	35.1%	**78.4%**	78.2%	53.5%
visibility among peers	48.1%	30.6%	**50.5%**	49.9%	32.5%
target power	60.1%	38.0%	79.0%	**81.9%**	76.7%

Table B.8: SVM Recall

Criterion	*BoW*	*Text*	*User*	*Proposed*	*Combined*
aggression	9.6%	26.0%	50.4%	**55.7%**	31.2%
repetition	**94.0%**	83.0%	38.9%	48.5%	52.1%
harmful intent	67.6%	**76.7%**	70.3%	68.5%	74.3%
visibility among peers	86.8%	**94.0%**	53.3%	58.1%	33.2%
target power	92.6%	46.0%	72.8%	80.1%	**92.7%**

Table B.9: AdaBoost Recall

Criterion	*BoW*	*Text*	*User*	*Proposed*	*Combined*
aggression	75.0%	**86.4%**	65.9%	77.4%	86.3%
repetition	23.8%	4.1%	26.8%	**31.2%**	17.8%
harmful intent	44.4%	37.8%	**57.0%**	52.8%	50.8%
visibility among peers	41.0%	15.4%	42.8%	**43.1%**	32.0%
target power	56.0%	39.4%	**81.8%**	81.0%	75.6%

Table B.10: MLP Recall

Criterion	*BoW*	*Text*	*User*	*Proposed*	*Combined*
aggression	64.1%	**86.5%**	65.5%	68.0%	85.6%
repetition	26.8%	6.8%	22.5%	**27.1%**	12.6%
harmful intent	51.0%	33.3%	**57.0%**	**57.0%**	37.2%
visibility among peers	51.6%	23.5%	45.6%	**50.2%**	26.5%
target power	61.6%	37.5%	**76.5%**	76.2%	65.6%

Table B.11: KNN F_1

Criterion	*BoW*	*Text*	*User*	*Proposed*	*Combined*
aggression	49.5%	51.9%	**65.5%**	64.9%	47.6%
repetition	12.5%	13.2%	11.3%	**20.5%**	15.7%
harmful intent	31.4%	33.3%	**38.9%**	38.3%	36.5%
visibility among peers	42.8%	44.1%	35.6%	40.4%	**45.8%**
target power	48.7%	50.3%	67.9%	**69.1%**	62.2%

Table B.12: Random Forest F_1

Criterion	*BoW*	*Text*	*User*	*Proposed*	*Combined*
aggression	65.2%	**79.3%**	56.0%	57.5%	77.9%
repetition	11.0%	10.6%	13.2%	**25.8%**	15.8%
harmful intent	25.6%	31.1%	46.6%	46.8%	**47.7%**
visibility among peers	35.7%	30.8%	41.2%	**46.1%**	33.6%
target power	47.4%	39.9%	**78.4%**	78.0%	72.8%

Table B.13: SVM F_1

Criterion	*BoW*	*Text*	*User*	*Proposed*	*Combined*
aggression	16.9%	37.7%	60.9%	**65.4%**	42.1%
repetition	12.6%	13.0%	11.8%	24.8%	**28.9%**
harmful intent	28.1%	33.8%	45.6%	**45.8%**	43.3%
visibility among peers	44.3%	46.1%	41.4%	**47.4%**	28.6%
target power	52.0%	35.8%	74.1%	**75.4%**	63.1%

Table B.14: AdaBoost F_1

Criterion	*BoW*	*Text*	*User*	*Proposed*	*Combined*
aggression	78.6%	**83.9%**	71.0%	77.5%	**83.9%**
repetition	11.7%	5.6%	11.5%	**21.6%**	20.9%
harmful intent	35.1%	41.6%	42.8%	45.4%	**55.0%**
visibility among peers	34.9%	21.0%	39.1%	**44.3%**	37.8%
target power	48.3%	42.7%	**79.8%**	79.6%	76.7%

Table B.15: MLP F_1

Criterion	*BoW*	*Text*	*User*	*Proposed*	*Combined*
aggression	72.2%	**82.5%**	70.7%	72.4%	81.8%
repetition	12.0%	7.6%	12.4%	**20.7%**	15.2%
harmful intent	35.7%	37.3%	45.0%	**45.8%**	41.3%
visibility among peers	38.0%	27.7%	39.2%	**45.5%**	31.4%
target power	48.2%	41.0%	**75.4%**	74.0%	67.0%

Appendix C

Combined Cyberbullying Classifier

Table C.1: F_1 Scores for Combinations of Cyberbullying Criteria

Cyberbullying Criteria	*BoW*	*Text*	*User*	*Proposed*	*Combined*
AGGR, REP	10.3%	7.8%	13.8%	26.6%	26.5%
AGGR, HARM	34.5%	47.3%	43.4%	44.4%	54.3%
AGGR, PEER	25.0%	21.7%	34.0%	38.3%	30.0%
AGGR, POWER	38.3%	39.1%	67.5%	67.8%	65.4%
REP, HARM	5.8%	5.2%	7.7%	15.0%	13.8%
REP, PEER	1.9%	2.9%	5.2%	10.8%	4.7%
REP, POWER	2.4%	4.2%	10.3%	9.9%	12.1%
HARM, PEER	10.5%	13.8%	17.5%	17.9%	20.5%
HARM, POWER	20.6%	37.0%	49.8%	49.4%	55.8%
PEER, POWER	15.2%	10.4%	34.4%	33.2%	23.3%
AGGR, REP, HARM	5.8%	5.2%	7.7%	15.0%	13.8%
AGGR, REP, PEER	3.7%	0.9%	5.0%	10.8%	3.5%
AGGR, REP, POWER	5.3%	4.4%	9.6%	9.7%	9.8%
AGGR, HARM, PEER	9.3%	18.3%	18.3%	19.5%	25.5%
AGGR, HARM, POWER	23.6%	34.9%	49.8%	49.2%	56.4%
AGGR, PEER, POWER	11.1%	11.5%	31.9%	29.7%	19.1%
REP, HARM, PEER	1.9%	4.8%	3.0%	6.6%	10.0%
REP, HARM, POWER	2.4%	4.0%	10.2%	9.9%	6.8%
REP, PEER, POWER	0.9%	0.0%	4.5%	4.1%	0.0%
HARM, PEER, POWER	7.5%	16.8%	16.8%	16.3%	22.4%
AGGR, REP, HARM, PEER	1.9%	4.8%	3.0%	6.6%	10.0%
AGGR, REP, HARM, POWER	2.4%	4.0%	10.2%	9.9%	6.8%
AGGR, REP, PEER, POWER	0.9%	0.0%	4.5%	4.1%	0.0%
AGGR, HARM, PEER, POWER	8.2%	15.4%	16.0%	15.7%	20.6%
REP, HARM, PEER, POWER	0.0%	0.0%	3.9%	4.7%	0.0%
AGGR, REP, HARM, PEER, POWER	0.0%	0.0%	3.9%	4.7%	0.0%

Bibliography

[1] Mohammed Ali Al-garadi, Kasturi Dewi Varathan, and Sri Devi Ravana. Cybercrime detection in online communications: The experimental case of cyberbullying detection in the twitter network. *Computers in Human Behavior*, 63: 433–443, 2016.

[2] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

[3] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. Mean birds: Detecting aggression and bullying on twitter. In *Proceedings of the 2017 ACM on web science conference*, pages 13–22. ACM, 2017.

[4] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *JAIR*, 16:321–357, 2002.

[5] Charalampos Chelmis, Daphney-Stavroula Zois, and Mengfan Yao. Mining patterns of cyberbullying on twitter. In *ICDMW*, pages 126–133. IEEE, 2017.

[6] Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. Improving cyberbullying detection with user context. In *European Conference on Information Retrieval*, pages 693–696. Springer, 2013.

[7] Corinne David-Ferdon and Marci F Hertz. Electronic media and youth violence; a CDC issue brief for researchers. 2009.

[8] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Eleventh International AAAI Conference on Web and Social Media*, 2017.

[9] Karthik Dinakar, Roi Reichart, and Henry Lieberman. Modeling the detection of textual cyberbullying. In *fifth international AAAI conference on weblogs and social media*, 2011.

[10] Sameer Hinduja and Justin W Patchin. Cyberbullying: An exploratory analysis of factors related to offending and victimization. *Deviant behavior*, 29(2):129–156, 2008.

[11] Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. Analyzing labeled cyberbullying incidents on the instagram social network. In *International conference on social informatics*, pages 49–66. Springer, 2015.

[12] Homa Hosseinmardi, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. Prediction of cyberbullying incidents in a media-based social network. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 186–192. IEEE, 2016.

[13] Qianjia Huang, Vivek Kumar Singh, and Pradeep Kumar Atrey. Cyber bullying detection using social and textual analysis. In *Proceedings of the 3rd International Workshop on Socially-Aware Multimedia*, pages 3–6. ACM, 2014.

[14] Yun-yin Huang and Chien Chou. An analysis of multiple factors of cyberbullying among junior high school students in taiwan. *Computers in Human Behavior*, 26(6):1581–1590, 2010.

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

[16] Robin M Kowalski and Susan P Limber. Psychological, physical, and academic correlates of cyberbullying and traditional bullying. *Journal of Adolescent Health*, 53(1):S13–S20, 2013.

[17] Qing Li. Cyberbullying in schools: A research of gender differences. *School psychology international*, 27(2):157–170, 2006.

[18] Kimberly Miller. Cyberbullying and its consequences: How cyberbullying is contorting the minds of victims and bullies alike, and the law's limited available redress. *S. Cal. Interdisc. LJ*, 26:379, 2016.

[19] Vinita Nahar, Xue Li, Chaoyi Pang, and Yang Zhang. Cyberbullying detection based on text-stream classification. In *The 11th Australasian Data Mining Conference (AusDM 2013)*, 2013.

[20] Vinita Nahar, Sanad Al-Maskari, Xue Li, and Chaoyi Pang. Semi-supervised learning for cyberbullying detection in social networks. In *Australasian Database Conference*, pages 160–171. Springer, 2014.

[21] Dan Olweus. Bullying at school. In *Aggressive behavior*, pages 97–130. Springer, 1994.

[22] Dan Olweus. Cyberbullying: An overrated phenomenon? *European Journal of Developmental Psychology*, 9(5):520–538, 2012.

[23] James W Pennebaker, Roger J Booth, and Martha E Francis. Liwc2007: Linguistic inquiry and word count. *Austin, Texas: liwc. net*, 2007.

[24] Megan Price, John Dalgleish, et al. Cyberbullying: Experiences, impacts and

coping strategies as described by australian young people. *Youth Studies Australia*, 29(2):51, 2010.

[25] Juliana Raskauskas and Ann D Stoltz. Involvement in traditional and electronic bullying among adolescents. *Developmental psychology*, 43(3):564, 2007.

[26] Kelly Reynolds, April Kontostathis, and Lynne Edwards. Using machine learning to detect cyberbullying. In *2011 10th International Conference on Machine learning and applications and workshops*, volume 2, pages 241–244. IEEE, 2011.

[27] H Rosa, N Pereira, R Ribeiro, PC Ferreira, JP Carvalho, S Oliveira, L Coheur, P Paulino, AM Veiga Simão, and I Trancoso. Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior*, 93:333–345, 2019.

[28] Semiu Salawu, Yulan He, and Joanna Lumsden. Approaches to automated detection of cyberbullying: A survey. *IEEE Transactions on Affective Computing*, 2017.

[29] Hugues Sampasa-Kanyinga, Paul Roumeliotis, and Hao Xu. Associations between cyberbullying and school bullying victimization and suicidal ideation, plans and attempts among canadian schoolchildren. *PloS one*, 9(7):e102145, 2014.

[30] Vivek K Singh, Qianjia Huang, and Pradeep K Atrey. Cyberbullying detection using probabilistic socio-textual information fusion. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 884–887. IEEE Press, 2016.

[31] Robert Slonje and Peter K Smith. Cyberbullying: Another main type of bullying? *Scandinavian journal of psychology*, 49(2):147–154, 2008.

[32] Robert Slonje, Peter K Smith, and Ann Frisén. The nature of cyberbullying, and strategies for prevention. *Computers in human behavior*, 29(1):26–32, 2013.

[33] Devin Soni and Vivek Singh. Time reveals all wounds: Modeling temporal characteristics of cyberbullying. In *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[34] Rekha Sugandhi, Anurag Pande, Abhishek Agrawal, and Husen Bhagat. Automatic monitoring and prevention of cyberbullying. *International Journal of Computer Applications*, 8:17–19, 2016.

[35] Cynthia Van Hee, Gilles Jacobs, Chris Emmery, Bart Desmet, Els Lefever, Ben Verhoeven, Guy De Pauw, Walter Daelemans, and Véronique Hoste. Automatic detection of cyberbullying in social media text. *PloS one*, 13(10):e0203794, 2018.

[36] Tracy E Waasdorp and Catherine P Bradshaw. The overlap between cyberbullying and traditional bullying. *Journal of Adolescent Health*, 56(5):483–488, 2015.

[37] Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. Learning from bullying traces in social media. In *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 656–666. Association for Computational Linguistics, 2012.

[38] Mengfan Yao, Charalampos Chelmis, Daphney Zois, et al. Cyberbullying ends here: Towards robust detection of cyberbullying in social media. In *The World Wide Web Conference*, pages 3427–3433. ACM, 2019.

[39] Xiang Zhang, Jonathan Tong, Nishant Vishwamitra, Elizabeth Whittaker, Joseph P Mazer, Robin Kowalski, Hongxin Hu, Feng Luo, Jamie Macbeth, and

Edward Dillon. Cyberbullying detection with a pronunciation based convolutional neural network. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 740–745. IEEE, 2016.

[40] Rui Zhao, Anna Zhou, and Kezhi Mao. Automatic detection of cyberbullying on social networks based on bullying features. In *Proceedings of the 17th international conference on distributed computing and networking*, page 43. ACM, 2016.