Aggressive, Repetitive, Intentional, Visible, and Imbalanced: Refining Representations for Cyberbullying Classification

Abstract

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 improve classifier performance. These results demonstrate the importance of representing and modeling cyberbullying as a social phenomenon.

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” BIBREF0. 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 BIBREF1, BIBREF2, BIBREF3. Where traditional bullying was once limited to a specific time and place, cyberbullying can occur at any hour and from any location on earth BIBREF4. Once the first message has been sent, the attack can escalate rapidly as harmful content is spread across shared media, compounding these negative effects BIBREF5, BIBREF6.

Internet users depend on content moderators to flag abusive text and 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 BIBREF7.

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 BIBREF8, BIBREF9, BIBREF10. 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. The key advantages to our labeling approach are:

[leftmargin=.2in]

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 existing 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 internet-scale datasets, unlike expensive network centrality and clustering measurements.

Results from our experiments suggest that, although existing NLP models can reliably detect aggressive language in text, these lexically-trained classifiers 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. 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.

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.

Background ::: Defining Cyberbullying

Some researchers view cyberbullying as an extension of more “traditional” bullying behaviors BIBREF16, BIBREF17, BIBREF18. 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 BIBREF19. 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 BIBREF20, BIBREF21. However, experts disagree whether cyberbullying should be studied as a form of traditional bullying or a fundamentally different phenomenon BIBREF20, BIBREF17. 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 repeated actions from the author BIBREF22, BIBREF5. 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 BIBREF23.

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 TABREF1.

Background ::: Existing Sources of Cyberbullying Data

According to BIBREF7, 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 BIBREF11, BIBREF13, BIBREF24, BIBREF14, BIBREF25, BIBREF15. Only three of the papers that we reviewed incorporated social context in the annotation process. BIBREF4 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. BIBREF7 presented “original conversation[s] when possible,” but they did not explain when this information was available. BIBREF8 was the only study to label full message reply threads as they appeared in the original online source.

Background ::: 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 BIBREF8, BIBREF7, BIBREF26, TF-IDF vectors BIBREF27, BIBREF28, BIBREF15, word embeddings BIBREF29, or phonetic representations of messages BIBREF30, as well as dictionary-based counts on curse words, hateful or derogatory terms, pronouns, emoticons, and punctuation BIBREF11, BIBREF31, BIBREF14, BIBREF25. Some studies have also used message sentiment BIBREF25, BIBREF15, BIBREF7 or the age, gender, personality, and psychological state of the message author according to text from their timelines BIBREF11, BIBREF31. These methods have been reported with appreciable success as shown in Table TABREF8.

Some researchers argue, however, that lexical features alone may not adequately represent the nuances of cyberbullying. BIBREF12 found that among 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 BIBREF12.

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 BIBREF12, BIBREF10, BIBREF9.

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 BIBREF13, BIBREF25. 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 BIBREF32. The ratio of all messages sent and received by each user was also 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 BIBREF33. 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.

Curating a Comprehensive Cyberbullying Dataset

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 were drawn from the social science and machine learning communities. These criteria can be combined and adapted for revised definitions of cyberbullying.

Curating a Comprehensive Cyberbullying Dataset ::: 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 BIBREF12, we ran the pre-trained classifier of BIBREF35 over our dataset to identify hate speech and aggressive language and increase the prevalence of cyberbullying examples . This 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^st through June 10^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.

Curating a Comprehensive Cyberbullying Dataset ::: 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 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). 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 the definitions for our five cyberbullying criteria as follows.

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.

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).

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.

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.

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 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 FIGREF18 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.

The final tweet in Figure FIGREF18 shows the importance of context in the annotation process. If we read only this individual message, we might decide that the post is cyberbullying, but given the social context here, we can confidently assert that this post is not cyberbullying. Although it contains the aggressive phrase “FUCK YOU TOO BITCH”, 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 TABREF17.

We determined ground truth for our data using a 2 out of 3 majority vote as in BIBREF12. If the message thread was missing or a target user could not 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.

Curating a Comprehensive Cyberbullying Dataset ::: Cyberbullying Transcends Cyberaggression

As discussed earlier, some experts have argued that cyberbullying is different from online aggression BIBREF12, BIBREF10, BIBREF9. 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 used this label to better understand worker perceptions of cyberbullying. We found that our workers believed cyberbullying will depend on a weighted combination of the five criteria presented in this paper, with the strongest correlate being harmful intent as shown in Table TABREF17.

Furthermore, the annotators decided our dataset contained 74.8% aggressive messages as shown in the Positive Balance column of Table TABREF17. 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.

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 BIBREF12, BIBREF10, BIBREF9. 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 allow our classifiers 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 compare 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.

Feature Engineering ::: Text-based Features

To construct realistic and competitive baseline models, we consider a set of standard text-based features that have been used widely throughout the literature. Specifically, we use the NLTK library BIBREF36 to construct unigrams, bigrams, and trigrams for each labeled message. This parallels the work of BIBREF8, BIBREF7, and BIBREF26. Following BIBREF30, we incorporate counts from the Linguistic Inquiry and Word Count (LIWC) dictionary to measure the linguistic and psychological processes that are represented in the text BIBREF37. We also use a modified version of the Flesch-Kincaid Grade Level and Flesch Reading Ease scores as computed in BIBREF35. Lastly, we encode the sentiment scores for each message using the Valence Aware Dictionary and sEntiment Reasoner (VADER) of BIBREF38.

Feature Engineering ::: Social Network Features

Network features have been shown to improve text-based models BIBREF6, BIBREF25, and they can help classifiers distinguish between bullies and victims BIBREF32. 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.

Feature Engineering ::: Social Network Features ::: 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 a and target t, listed here.

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 FIGREF18.

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 FIGREF26 (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 FIGREF26. The difference in outward overlap is significant ($D=0.04$, $p=0.03$), and the difference for inward overlap is short of significant ($D=0.04$, $p=0.08$).

Feature Engineering ::: Social Network Features ::: 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 BIBREF11, BIBREF4, and BIBREF8.

Feature Engineering ::: 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.

Feature Engineering ::: Timeline Features ::: 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 FIGREF38. This means malicious authors tend to address the target repeatedly while the target responds with relatively few messages.

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 FIGREF38 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 ($D=0.08$, $p=0.03$).

Feature Engineering ::: Timeline Features ::: 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 positions. 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

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 FIGREF44 (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 FIGREF44 (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.

Feature Engineering ::: Timeline Features ::: 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 BIBREF36. From the language model, we compute the surprise of the original message m according to its cross-entropy, given by

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

We see in Figure FIGREF47 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.

Feature Engineering ::: Thread Features

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

Feature Engineering ::: Thread Features ::: 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.

Feature Engineering ::: Thread Features ::: Aggression

To detect repeated aggression, we again employ the hate speech and offensive language classifier of BIBREF35. 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

Experimental Evaluation

Using our proposed features from the previous section and ground truth labels from our annotation task, 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) BIBREF39. 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 TABREF58. 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.

The performance of the different classifiers is summarized in Tables TABREF59, TABREF64, and TABREF65. 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 TABREF8. 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 TABREF64. 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.

Similar results were obtained by replacing our logistic regression model with any of a random forest model, support vector machine (SVM), AdaBoost, or Multilayer Perceptron (MLP). We report all precision, recall, and F_1 scores in Appendix 2, Tables TABREF69-TABREF77. We chose to highlight logistic regression 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 in Appendix 2, Tables TABREF78-TABREF78. There we 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.

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 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 BIBREF10.

Discussion ::: Limitations

Our study focuses 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. This classifier has an F1 score of 0.90 BIBREF35. Even with this filter, only 0.7% of tweets were deemed by a majority of MTurk workers as cyberbullying (Table TABREF17). This extreme class imbalance can disadvantage a wide range of machine learning models. Moreover, the MTurk workers exhibited only moderate inter-annotator

agreement (Table TABREF17). We also 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, their 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.

Discussion ::: Future Directions

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.

Conclusion

In this study, we produced an original dataset for cyberbullying detection research and an approach that leverages this dataset to more accurately detect cyberbullying. Our labeling scheme was designed to accommodate the cyberbullying definitions that have been proposed throughout the literature. In order to more 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 new set of labels for more reliable cyberbullying representations.

From these ground truth labels, we designed a new set of features to quantify each of the five cyberbullying criteria. Unlike previous text-based or user-based features, our features measure the relationship between a message author and target. We show that these features 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 methods, our classifiers have not 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.

Acknowledgements

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Agreement No. HR0011890019, and by the National Science Foundation (NSF) under Grant No. 1659886 and Grant No. 1553579.

Appendix 1: Analysis of the Real-World Class Distribution for Cyberbullying Criteria

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%.

Appendix 2: 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.