Fine-grained Analysis of Cyberbullying using Weakly-Supervised Topic Models

Yue Zhang and Arti Ramesh
SUNY Binghamton
{yzhan202, artir}@binghamton.edu

***Abstract*—The possibility of anonymity and lack of effective ways to identify inappropriate messages have resulted in a significant amount of online interaction data that attempt to harass, bully, or offend the recipient. In this work, we perform a fine-grained quantitative and qualitative linguistic analysis of messages exchanged using one such recent web/smartphone application—*Sarahah*, that allows friends to exchange messages anonymously. We first develop a weakly supervised hierarchical framework using seeded topic models to automatically categorize *Sarahah* messages into different coarse and fine-grained bullying categories. Our linguistic analysis reveals that a significant number of messages exchanged using this platform ($\sim$ 20%) include inappropriate, hurtful, or profane language intended to embarrass, offend, or bully the recipient. We then present a detailed analysis of the messages and corresponding users' responses to these messages in the different bullying categories by comparing them across different linguistic and psychological attributes such as sentiment and psycho-linguistic categories from Linguistic Inquiry Word Count (LIWC). Finally, we perform a comparative analysis of messages exchanged on Sarahah to an existing labeled cyberbullying dataset from the Formspring social network on the severity of bullying, coarse-grained bullying categories, and anonymity. Our analysis sheds light on the different categories of bullying and the effect each category has on the recipient and helps quantify the different types and amounts of negativity existing in online social media.**

Introduction

The recent years have witnessed the rise and prevalence of cyberbullying in online interactions. Statistics from three different studies presented on *Cyberbullying.org* state a steady increase in percentage of cyberbullying incidents between 2000 to 2013 [1]. Severe psychological problems can entail from cyberbullying incidents such as development of low self-esteem, depression, and suicidal tendencies [2], making it an important problem to study. Online cyberbullying incidents tend to be digitally preserved for a considerable length of time, further aggravating the effect on individuals experiencing them.

Anonymity has been shown to be a contributing factor in cyber harassment and bullying. Previous work on *Ask.fm* and *Yik-Yak* have shown that the possibility of anonymity significantly propels the number of cyberbullying messages [3, 4]. In this work, we focus on one such anonymous messaging application that topped the download charts in the period between July—October, 2017 on App Store: *Sarahah*. This application could be added to existing social networking applications such as Facebook, Twitter, and Snapchat, allowing users who are friends on these networks to exchange anonymous messages. The application soon transformed into a breeding ground for hate [5]. Most previous work on cyber bullying has been in settings such as Ask.fm, Youtube and Formspring, where the people exchanging bullying or hateful comments need not necessarily know each other at a personal level. Our analysis especially brings forth the amount of negative content in messages exchanged between people who know each other on social networking sites, making it more personal than other instances of bullying. Since we don't have direct access to messages exchanged through Sarahah, we collect this data via Twitter, where recipients share the messages they receive on Sarahah on their Twitter feed, sometimes along with a response to the message. This unique aspect of this data (cyberbullying messages + corresponding user responses) helps in understanding the impact of bullying on users who receive these messages. While this data is not representative of all cyberbullying content on the web/social media and can potentially be biased due to its collection using the Twitter API [6], it offers a rich and important source of information to study this growing problem.

Stopbullying.gov [7] defines cyberbullying as *sending, posting, or sharing negative, harmful, false, mean, or personal information about someone else that can bring about embarrassment or humiliation to the recipient*. Applying this definition of cyberbullying to our specific problem setting where the sender is anonymous and part of the recipient's social network, and closely examining users' responses to different kinds of messages, we identify different coarse and fine-grained bullying categories. The sensitive nature of cyberbullying content poses challenges in labeling it using crowdsourcing, making unsupervised/weakly-supervised models a lucrative choice for analyzing this data.

Specifically, we make the following contributions:

1) We first leverage Latent Dirichlet Allocation (LDA) [37], to perform a linguistic analysis and identify the different types of messages exchanged using Sarahah. Using our linguistic analysis, we identify the different coarse and fine-grained bullying categories and identify words corresponding to each of these categories.
2) We then leverage a seeded variant of LDA, seeded LDA [8], to develop hierarchical weakly-supervised models for categorizing the messages into coarse and fine-grained categories. We observe that the most prevalent form of bullying in our data is inappropriate flirting, followed by sexually offensive, and hate messages.

3) We then examine differences in psycholinguistic categories from Linguistic Inquiry and Word Count (LIWC) [38] and sentiment in both the bullying messages and responses to these messages across the different bullying categories. We observe that responses to sexually offensive and hate messages have a significant amount of anger and negative emotion when compared to the responses to other messages. Though flirting could be perceived as harmless under normal circumstances, we observe that a significant number of these messages contain words that can potentially make the recipient uncomfortable. We isolate inappropriate flirting messages and observe that users' responses to these messages contain a significant expression of anger and negative emotion making it an important bullying category to study.
4) Finally, we leverage an existing labeled cyberbullying dataset, Formspring [9], to perform a comparative analysis of messages exchanged on Sarahah and Formspring on the severity, coarse-grained bullying categories, and anonymity. Our comparisons show that both Formspring and Sarahah have similar coarse-grained bullying categories indicating the reusability of our weakly supervised models across platforms. From our psycholinguistic comparison using LIWC, we observe that Sarahah messages contain significantly more profane and hurtful language than Formspring, indicating the escalation of bullying in a personal setting, where the messages are targeted at a specific person (Sarahah) rather than random impersonal instances of bullying found on the web (Formspring). This comparative analysis helps understand the similarities and differences between the different cyberbullying categories in social media and the importance of studying our specific problem setting.

Our models and entailing analysis help in understanding and quantifying the alarming amount of negativity existing between people who know each other at a personal level in an online setting. Our analysis sheds light on the different types of bullying, their prevalence, and consequent effect on the recipient through their responses to these bullying messages.

Related Work

Detecting and understanding bullying on social media has received considerable interest in the recent years. Hosseinmardi et al. [10] identify media sessions on Instagram that have at least one profane word in their comments by users other than the profile owner. Raisi et al. [11] propose a participant-vocabulary consistency model for identifying the instigators and victims of bullying in a social network and simultaneously building a bullying vocabulary by starting with a corpus of social interactions and a seed dictionary of bullying indicators. They evaluate the model on data from Twitter and Ask.fm and show that the proposed method can detect new bullying vocabulary as well as victims and bullies.

Several work consider social interaction features along with textual features to detect cyberbullying [4, 12, 13, 14, 15]. Bigelow et al. [16] use latent semantic indexing to detect cyberbullying. Sanchez et al. [17] leverage sentiment analysis to detect bullying instances in Twitter and visualize the evolution of bullying instances over time. Dani et al. [14] use sentiment features and Zhong et al. [18] investigate the use of content-driven features to detect cyberbullying. Li et al. [19] analyze the negative and positive sense of the words on Instagram and Ask.fm networks. Margono et al. [20] analyze bullying patterns in Indonesia on Twitter. Whittaker et al. [21] examine the prevalence of cyberbullying in college students.

There is also work that draw attention to the broader issue of cyber aggression [10, 22, 23, 24] and trolling [15, 25]. Patton et al. develop automated tools to detect aggressive language on social media and help individuals and groups in performing violence prevention and interruption [26]. Bellmore et al. [27] and Tokunaga et al. [28] study the socio-psychological issues of bullying in social media data. There is also previous work on detecting abusive and hateful speech targeting specific groups including ethnicity, origin, religion, gender, sexual orientation and physical appearance [29, 30]. Dinakar et al. [29] show that individual classifiers perform better than multi-class classifiers for this problem on a Youtube comments dataset.

Perhaps the closest work to our approach is Hee et al.'s work on using supervised classifiers to predict severity and fine-grained bullying events such as insults/threats and sexual talk [31]. Their approach however relies on the presence of labeled data and the performance of the prediction models on finer-grained categories is lower. They attribute this to the presence of fewer training data points in those categories.

In our work, we perform a coarse-to-fine analysis of different forms of cyberbullying, abusive language, and inappropriate messages in data collected from a recent anonymous mobile/web application, Sarahah. Our data and subsequent analysis stands out from existing previous work in the following ways:

1) The messages that we consider in our analysis are exchanged anonymously between people who are friends on the social network, giving us an avenue to study the amount of enmity that anonymity can unleash between friends.
2) Our dataset comprises of messages and corresponding reactions to these messages which helps us in understanding the discomfort caused by messages in the different bullying categories on recipients.
3) The weakly supervised nature of our models helps in understanding how to effectively build models for this problem without the need for expensive training data.
4) While most previous work focus primarily on detecting cyberbullying, our work focuses on performing a detailed analysis of the different factors contributing to/affecting cyberbullying and comparing our setting to bullying on other platforms. Thus, in addition to performing fine-grained categorization of bullying messages, we also present a systematic way to perform a

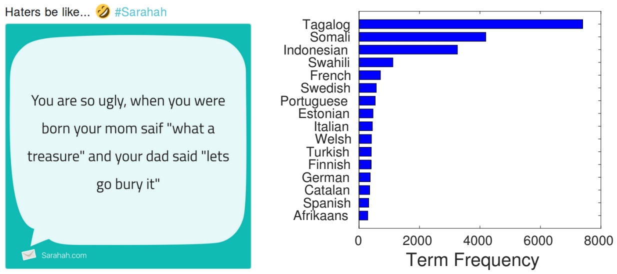
fine-grained analysis, paving the way for designing more informed prediction models in the future.

Data

In this work, we focus on data collected from a recent anonymous mobile application, *Sarahah*, which can be added to many popular social networking platforms such as Facebook and Twitter. It entered the US Apple Store in June 13, 2017 and gradually spread out to Canada, India, and a few other countries. The popularity of the application spiked after a new update was launched by Snapchat on July 5, 2017. Gradually it became the top rated application in *App Store* leaving behind other popular social networking applications such as Twitter, Facebook, and Snapchat [32]. While this application was originally created to exchange anonymous messages, it soon became a breeding ground for hate and a platform for cyberbullying [5]. Owing to the large number of cyberbullying incidents, the Sarahah app was eventually shut down from Apple Store and Android Play Store [33].

We present analysis on data collected from 30^{th} August, 2017 to 15^{th} October, 2017. Since messages exchanged on Sarahah are private, we collect messages that recipients share on their Twitter feed. We collect messages exchanged using Sarahah on Twitter by searching for *#Sarahah* images using the Twitter search API [34]. Since Twitter allows you to only extract tweets posted in the last week, we collect at one-week intervals during the specified period. We also extract Sarahah messages by crawling specific users' Twitter accounts. While the sender remains anonymous in this setting, we know the recipient as he/she shares this message on their Twitter feed. While this subset of messages is only a sample of messages exchanged through Sarahah, we believe that this data is helpful in understanding the key ways the application is used and the distribution of various kinds of messages exchanged through it and can help us understand the types of bullying present in online interaction data. Figure 1(a) shows an example of a message exchanged through *Sarahah*. Since the message exchanged using Sarahah is in the form of images, we use Google's optical character recognition software to extract text from the images [35]. The extracted data has three components: i) the textual message exchanged using Sarahah, ii) user's reaction to the message when the user shares this message on Twitter, and iii) other user-related information extracted from the user's profile.

Since Sarahah messages are generally from friends, they tend to also be in languages other than English. Though English remains the most popular language in our dataset, we found the presence of several languages. We report the distribution of messages across the top 50% languages other than English in Figure 1(b). We use Google's language detection library *langdetect* to detect the language and convert the message and users' response to English for our analysis [36]. In all, we collect 82,193 Sarahah messages and corresponding user responses. Removing duplicates and empty messages, we have 76,278 messages and corresponding user responses. We perform standard NLP preprocessing techniques of stop-word removal, tokenization, and stemming on the messages and responses.



Haters be like... 🤣 #Sarahah

You are so ugly, when you were born your mom saif "what a treasure" and your dad said "lets go bury it"

Sarahah.com

(a) Example Sarahah message and corresponding user reaction

(b) Distribution of different languages in our Sarahah dataset

Fig. 1. An example message from Sarahah (left) and statistics of different languages in our data (right).

Fine-grained Analysis of Bullying Messages using Weakly Supervised Topic Models

In this section, we develop a hierarchical weakly supervised framework to automatically classify and perform a detailed analysis of bullying messages[1]. We first perform a linguistic analysis of the messages exchanged using Sarahah using LDA. Our analysis paves the way for understanding the nature of these messages and identifying the different coarse and fine-grained bullying categories. We then construct a hierarchical seeded topic model, combining different seeded topic models together to perform a fine-grained classification of Sarahah messages. Our framework uses weak supervision in the form of seed words to identify the different coarse and fine bullying categories.

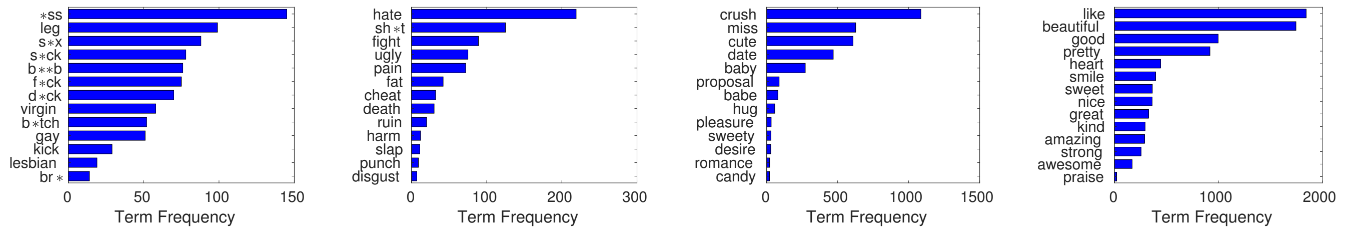
Topic Analysis using Latent Dirichlet Allocation (LDA)

Topic modeling, also known as latent Dirichlet allocation (LDA) is a popular means to analyze document corpora [37]. We first start by using LDA to understand the presence of different bullying related words in our data. We consider each message as a document and run LDA for 10,000 iterations. We use standard values of $\alpha = 0.01$ and $\beta = 0.01$ for the hyperparameters and 30 topics. We consider the top 25 words from topic-word distributions of these 30 topics. Figure 2 gives the most frequent words in the coarse and fine-grained categories identified below. Note that several sexually offensive words and words implying hate/emotional abuse occur $\sim$1600 times. This number is especially alarming considering that the messages are exchanged between people who are "friends" on the social network.

Identifying Coarse and Fine-grained Bullying Categories

We peruse our analysis using LDA, the top words in the different topics in LDA, and bullying categories identified by

[1] The code for our models and analysis is available at https://github.com/yzhan202/zhang-dsaa2018-experiment.



(a) Term frequencies of words in *sexual* topic

(b) Term frequencies of words in *hate* topic

(c) Term frequencies of words in *inappropriate flirting* topic

(d) Term frequencies of words in *praise* topic

Fig. 2. Graphs showing term frequencies of words in *sexual*, *hate*, *flirting*, and *praise* topics.

Hee et al. [31] to identify coarse and fine-grained bullying categories present in Sarahah messages. The coarse-grained bullying categories include:

Sexual: The messages in this category contain explicit sexually offensive words that are intended to harass, intimidate, or make the recipient uncomfortable.

Hate: The messages in this category are intended to convey hatred and emotionally unsettle the recipient. We categorize messages that intend to convey hatred, death threats, and emotional/physical abuse as *hate* messages.

Flirting: While Hee et al.'s classification only contains *sexual* and *hate* categories, we add the *flirting* category as we find that it is a predominant category owing to the personal nature of the communication. The messages in this category are intended to convey a romantic interest toward the recipient. Under an anonymous setting, we notice that a significant percentage of users show negative emotion/sadness in their responses to these messages. Hence, we include this as a bullying category in our analysis.

For each of these coarse-grained categories, we identify fine-grained categories. For the *sexual* category, we identify i) sexually offensive messages *targeting women*, ii) sexually offensive messages *targeting men*, and iii) sexually offensive messages *targeting LGBT community, especially gay/lesbian*. We categorize *hate* into three categories: i) explicit expression of *hatred*, ii) *emotionally abusive* messages, and iii) messages that convey *physical abuse*.

To appropriately understand *flirting* as a bullying category, we further divide *flirting* into i) *inappropriate flirting*, and ii) *flirting through praise/admiration* to isolate the inappropriate flirting messages. They are defined as:

Inappropriate Flirting/Romantic Proposals: The messages in this category do not generally include any offensive words. But, because these messages are anonymous, they can potentially make the recipient uncomfortable. Romantic proposals, asking for personal information, praise/admiration followed by an expression of romantic interest towards the recipient are the common examples of messages in this category. Also, note that our dataset contains messages from many cultures and flirting could be considered inappropriate/offensive in several cultures.

Praise/Admiration: Messages in this category have a touch of flirting through praise/admiration.

Table I gives some example messages in *sexual*, *hate*, *inappropriate flirting*, and *flirting through admiration* bullying categories. Hurtful and offensive words are shown in italics. Figure 2 gives the term frequencies of the top words in *sexual*, *hate*, and *inappropriate flirting* bullying categories.

Cyberbullying category	Example post
Sexual	You need to be *slapped* with my *d∗ck*. What's your *br∗* size? Do you enjoy *an∗l sex*?
Hate	You need to be *slapped*. I want to *punch* you on your face. Here's the *hate* you wanted. *Hate* you *hate* you *hate* you *hate* you.
Flirting	What if I ask you to kiss me? I have a crush on you :(Has anyone ever told you your eyes are so beautiful? I have a huge crush on you.

TABLE I
COARSE-GRAINED BULLYING TOPIC CATEGORIES AND SOME EXAMPLE MESSAGES IN EACH CATEGORY.

Seeded Topic Models

Since we are specifically interested in isolating the messages in different coarse and fine bullying categories rather than general topics identified by a topic model, we leverage a seeded variant of LDA, Seeded LDA [8], to guide the topic model to identify topics of interest. Seeded LDA allows seeding of topics by providing a small set of key words to guide topic discovery influencing both the document-topic and the topic-word distributions [8]. The seed words need not be exhaustive as the model is able to detect other words in the same category via co-occurrence in the dataset. We construct a coarse-to-fine hierarchical seeded LDA model to perform a fine-grained analysis of cyberbullying messages.

Coarse-grained Bullying Categorization

In the next step, we develop a seeded LDA model to categorize messages into coarse-grained categories. For the three coarse-grained bullying categories: i) *sexual*, ii) *hate*, and iii) *flirting* categories, we select the top few words from our words in Figure 2 as seed words for our seeded LDA model. Table II gives the seed words for the coarse-grained bullying categories. We include k un-seeded topics in our model to account for messages that do not fall into these three categories. After experimenting with different values of k and manually evaluating the topics, we find that $k = 2$ gives us the best separation and categorization. We use $\alpha = 0.0001$ and $\beta = 0.0001$ to give us sparse document-topic and topic-word distributions where fewer topics and words with high values emerge, so we can classify the messages to the predominant bullying category. We train the seeded LDA models for 2000 iterations. We first use the document-topic distribution to get the best topic for each Sarahah message. If the best topic of the message is one of the seeded topics which correspond to the coarse-grained bullying categories, then, we classify the message into that category.

Figure 3(a) gives the distribution of messages in the coarse-grained category. Overall, we observe that $21,526$ messages in $76,278$ contain some form of bullying. The most prevalent type of bullying is in the form of inappropriate flirting, contributing to 47% of the bullying messages. This is followed by messages in the *sexual* category that use sexually offensive words, contributing to 7% of the bullying messages. This in turn is followed by messages in the *hate* category (3% of the bullying messages).

Category	Words
Sexual	s∗x, f∗ck, ∗ss, s∗ck, d∗ck, gay
Threats	hate, punch, death, ugly, fat, slap, disgust
Flirting	baby, cute, candy, crush, date, miss, desire

TABLE II

SEED WORDS FOR *sexual*, *hate*, AND *flirting* COARSE-GRAINED BULLYING CATEGORIES

Fine-grained Category	Seed words
Targeting women	virgin, b∗tch, b∗∗b, leg, br∗
Targeting men	d∗ck, s∗ck, b∗ll
Targeting gay/lesbians	gay, lesbian

TABLE III

SEED WORDS FOR FINE-GRAINED CATEGORIES UNDER *sexual* COARSE-GRAINED CATEGORY

Fine-grained Category	Seed words
Hatred	hate, die
Physical Abuse	punch, slap
Emotional Abuse	ugly, fat, disgust

TABLE IV

SEED WORDS FOR FINE-GRAINED CATEGORIES UNDER *hate* COARSE-GRAINED CATEGORY

Fine-grained Category	Seed words
Admiration/praise	pretty, baby, cute, candy, beauty
Romantic proposal	crush, date, hug, desire, love

TABLE V

SEED WORDS FOR FINE-GRAINED CATEGORIES UNDER *flirting* COARSE-GRAINED CATEGORY

Fine-grained Bullying Categorization

To perform fine-grained classification, we first filter the messages in each of the categories and create new datasets that only contain messages in each coarse-grained bullying category. For each of the coarse-grained categories, we now construct seeded topic models to classify the messages into finer-grained categories.

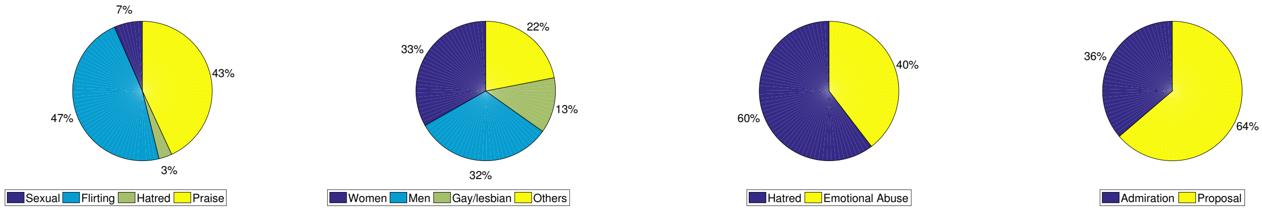
Fine-grained Categories for sexual category: Table III gives the seed words for the three fine-grained bullying categories in *sexual* category. Figure 3(b) gives the distribution of messages across these fine-grained bullying categories. We notice that women and men are targeted almost equally, contributing 33% of messages. We observe 13% of *sexual* messages targeting gay/lesbians and 22% that use sexually offensive words but do not target a particular gender.

Fine-grained Categories for hate category: Table IV gives the seed words for the fine-grained bullying categories in *hate* category. Here, we separate the messages into three fine categories: i) messages that convey hate by explicitly mentioning the words *hate*, *die*, ii) messages that intend to emotionally abuse the recipient by using words such as *ugly* and *disgusting*, iii) messages that mention physical abuse such as *punch* and *slap*. Figure 3(c) gives the distribution of messages across *hate* category. 60% of messages in the *hate* category explicitly use words implying hate. We find that 40% of messages indicate some form of emotional abuse, having mentions such as *you are ugly and disgusting*. We found very few messages (~ 10) in the physical abuse category.

Fine-grained Categories for flirting category: Flirting is a challenging topic to classify further as we find messages both using words indicating praise/admiration and words indicating romantic proposals, asking the recipient out on a date or a kiss, which are more indicative of inappropriate flirting. We identify the words that occur specifically in each of these sub-categories in Table V and use them to classify these messages further. Figure 3(d) gives the fine-grained classification of messages in *flirting* category. We observe that a significant number (64%) of the messages are in proposals category and 36% of the messages are in the flirting through admiration category.

Differences in LIWC Categories

Linguistic Inquiry Word Count (LIWC) provides a vocabulary for measuring the presence of linguistic signals related to different psychological states [38]. We compare the messages in the different coarse-grained categories across different LIWC categories and report the average values and the corresponding *p*-values using a two-sample *t*-test. Examining these LIWC differences serves: i) as a means to understand the varying levels of different psychological attributes, thus helping us understand the degree and severity of cyberbullying across different groups of messages, and ii) as a means to validate our seeded LDA categorization. The LIWC categories we consider are anger, negemo (negative emotion), sexual (presence of sexually explicit words), body (presence of words related to the human body), sad, and death. Table VI gives the average values and the corresponding *p*-values for the different comparisons. We mainly report the results for the topic combinations where we observe a significant difference in the average values. For most of the comparisons, we observe



(a) Distribution of messages in coarse-grained bullying categories

(b) Distribution of messages in *sexual* fine-grained bullying category

(c) Distribution of messages in *hate* fine-grained bullying category

(d) Distribution of messages in *flirting* fine-grained bullying category

Fig. 3. Graphs showing distribution of messages across coarse and fine-grained bullying categories

LIWC Category	Topic 1	Mean	Topic 2	Mean	p-value $<$
anger	sexual	4.29	hate	18.32	0.00001
anger	hate	18.32	flirting	0.78	0.00001
anger	bullying	8.74	rest	0.66	0.00001
negemo	sexual	6.18	hate	20.94	0.00001
negemo	hate	20.94	flirting	17.62	0.01
negemo	sexual	6.18	rest	4.90	0.00001
negemo	flirting	17.62	rest	4.90	0.00001
negemo	bullying	10.86	rest	4.90	0.00001
death	sexual	0.50	hate	1.98	0.00001
death	sexual	0.50	flirting	0.07	0.00001
death	hate	1.98	flirting	0.07	0.00001
death	bullying	0.97	rest	0.11	0.00001
sexual	sexual	11.77	hate	0.37	0.00001
sexual	sexual	11.77	flirting	1.10	0.00001
sexual	hate	0.37	flirting	1.10	0.05
sexual	flirting	1.10	rest	0.82	0.05
sexual	bullying	8.16	rest	0.82	0.00001
body	sexual	10.11	hate	4.15	0.00001
body	sexual	10.11	flirting	0.56	0.00001
body	bullying	8.22	rest	1.33	0.00001

TABLE VI
DIFFERENCES IN LIWC CATEGORIES FOR SARAHAH MESSAGES ACROSS DIFFERENT COARSE-GRAINED CATEGORIES.

that the p-values are negligible, i.e., p-value $\ll$ 0.00001, indicating that the difference is significant. For the others, we report the nearest number that is greater than the p-value.

From the LIWC differences, we observe that messages in *hate* category contain the most anger words, followed by messages in *sexual* category. Similarly, messages in *hate* category contain more anger than messages in the *flirting* category. For the LIWC comparisons, we isolate messages in the more severe bullying categories of *sexual* and *hate* categories from *flirting* and label them *bullying* to effectively analyze the differences in LIWC scores between bullying messages which contain profane/hurtful words and more subtle bullying in the form of inappropriate flirting. Comparing messages in the *bullying* category with the rest of the messages that are not in the bullying category (*rest*), we observe that bullying messages express significantly more anger than non-bullying messages. Comparing expression of negative emotion, we observe that *hate* messages contain more negative emotion when compared to *sexual* messages. Similarly, we observe that *hate* messages also have more negative emotion than flirting. Comparing messages in *sexual* with *rest*, we find that the former have more negative emotion than the latter. Another interesting comparison is between *flirting* and the *rest*, where we observe that flirting has more incidence of negative emotion when compared to non-bullying messages, in fact second only to the messages in the *hate* category. This indicates the importance of understanding subtle bullying categories such as flirting. Similarly, we find that messages in *hate* category have the most occurrence of death-related words as evident by the high average values for *hate* when compared with all other topic categories and *bullying* with the rest.

The sexual and body LIWC categories get high average values for the *sexual* category, followed by *flirting*, and *hate* categories. Though *sexual* and *flirting* categories have many common words, our model succeeds in achieving a good separation between them with a small set of seed words as evidenced by the high average values in LIWC sexual (11.77 vs. 1.10) and body (10.11 vs. 0.56) categories for *sexual* messages when compared to *flirting* messages. It is also important to note that *flirting* has more sexual words than the rest of the messages second only to the messages in the *sexual* category, though the p-value is significant but lower (0.05) than other comparisons. All these findings indicate that it is an important bullying category to study that can potentially cause significant emotional distress and does not usually contain profane words that can be easily flagged.

ANALYSIS OF USERS' RESPONSES TO SARAHAH MESSAGES

In this section, we present an analysis of users' reactions to Sarahah messages. Our goal here is to answer the question: *What are the different ways in which people react to the bullying messages and how does that vary across the different categories?*

Types of Users' Responses

We primarily observe three types of responses to bullying messages: positive, emotionally affected, and defensive reactions. They are explained as follows:

Positive: Users respond positively to these messages, indicated by words such as *lol*, *funny*.

Emotionally affected: This category captures the responses when users get hurt and emotional.

Defensive: Some users resort to defending themselves by either giving a harsh reply or using words similar to the ones used by the sender.

Table VII, VIII, and IX give some examples of messages and corresponding responses in *sexual*, *hate*, and *flirting* categories.

Response Type	Sarahah Post	User Reaction
Positive	s∗x or chocolates?	Why choose when you can have both.
Emotional	Do you wanna s∗x with me?	Oh dear sarahah! Please dear friends don't play with me
Defensive	Your b∗∗b size is directly proportional to your weight.	Your d∗ck size is also directly proportional to your brain size.

TABLE VII
EXAMPLE SARAHAH MESSAGES AND CORRESPONDING USER RESPONSES IN *sexual* CATEGORY

Response Type	Sarahah Post	User Reaction
Positive	You disgust me.	I will keep disgusting you. Lol.
Emotional	Why do you ruin lives?	I don't think I do, If you feel that way I am sorry.
Defensive	Who would date a ugly person like you?	I am ugly so this hurts me, whoever you are, you are the ugly person not me.
Defensive	You freak me out how ugly u are.	Wait until 24 hours s∗ckers

TABLE VIII
EXAMPLE SARAHAH MESSAGES AND CORRESPONDING USER RESPONSES IN *hate* CATEGORY

Response Type	Sarahah Post	User Reaction
Positive	Are you down to go on a date?	Yes, lets go.
Emotional	I like the fat you.	You're mean. And mean people are ugly.
Defensive	I want to date you, any chances.	No chances.

TABLE IX
EXAMPLE SARAHAH MESSAGES AND CORRESPONDING USER RESPONSES IN *flirting* CATEGORY

Differences in LIWC Categories and Sentiment

Here, we examine the differences in the LIWC categories and sentiment in users' responses to Sarahah messages. Examining these differences helps us understand the different ways in which users react to these messages and hence enables us to study the different extents the messages affect the recipients.

LIWC Category	Topic 1	Mean	Topic 2	Mean	p-value <
anger	sexual	2.28	hate	3.88	0.005
anger	sexual	2.28	flirting	0.76	0.00001
anger	bullying	2.79	rest	0.91	0.00001
negemo	sexual	4.78	hate	6.86	0.005
negemo	sexual	4.78	flirting	9.56	0.00001
negemo	hate	6.86	flirting	9.56	0.002
negemo	sexual	4.78	rest	2.96	0.002
negemo	flirting	9.56	rest	2.96	0.00001
sexual	sexual	1.80	hate	0.52	0.00001
sexual	sexual	1.80	flirting	0.28	0.00001
sexual	bullying	1.40	rest	0.33	0.00001
sad	bullying	1.41	rest	0.97	0.005
death	hate	0.39	flirting	0.15	0.05
death	sexual	0.31	flirting	0.15	0.06
death	bullying	0.34	rest	0.12	0.00001

TABLE X
DIFFERENCES IN LIWC CATEGORIES FOR USER RESPONSES TO SARAHAH MESSAGES ACROSS DIFFERENT COARSE-GRAINED CATEGORIES.

Sentiment	Topic 1	Mean	Topic 2	Mean	p-value <
Negative	sexual	0.220	hate	0.248	0.003
Negative	hate	0.248	flirting	0.228	0.00001
Negative	flirting	0.228	praise	0.190	0.00001
Negative	bullying	0.410	rest	0.216	0.00001
Negative	flirting	0.228	rest	0.216	0.002

TABLE XI
DIFFERENCES IN NEGATIVE SENTIMENT IN USER RESPONSES TO SARAHAH MESSAGES ACROSS DIFFERENT COARSE-GRAINED CATEGORIES.

Table X gives the differences in LIWC categories across the different coarse-grained categories. We observe that responses to *hate* messages have more anger when compared to *sexual* messages. Similarly, responses to *sexual* messages have more anger than responses to *flirting* messages. And, responses to *bullying* messages (messages in the *sexual* and *hate* category) contain more expression of anger than responses to *rest* (messages that are not in the bullying category). Another difference that is worth noting is that responses to *bullying* messages express more sadness when compared to the rest of the messages. Similarly, analyzing differences in expression of negative emotion in responses, we observe that responses to *flirting* has the most amount of negative emotion, followed by *hate* and *sexual* in that order. This again reaffirms that inappropriate *flirting* is an important bullying category as evident by the nature of responses to these messages.

Further, analyzing the differences in sexual LIWC category, we observe that responses to *sexual* messages have a greater incidence of sexual words than other categories. This helps us understand that the messages can provoke the usage of similar words in response, which in turn also contributes to offensive/profane language on the web. Similarly, we find that responses to *hate* messages have a significant amount of death words, again suggesting the possibility of usage of words similar to bullying messages in defense.

We also compute negative sentiment in responses across the different categories using *SentiWordNet* [39]. The negative sentiment scores are normalized to a value between 0 and 1, with 1 denoting maximum negative sentiment. Now, turning to the differences in negative sentiment across the different coarse-grained topics given in Table XI, we observe that the responses to *hate* messages have the highest negative sentiment. This is followed by responses to *sexual* topic, which in turn is followed by the responses to *flirting* topic. Comparing responses to *bullying* messages with responses to *rest*, we find that the former has higher negative sentiment than the latter. It is also interesting to note that responses to *flirting* category contain a higher average value of negative sentiment, ascertaining that it is an important category to study.

Comparison with Formspring

Formspring is a social network that allows you to post questions either anonymously or non-anonymously and has been previously known to have many instances of cyberbullying [40]. We use the labeled Formspring dataset collected by [9] in our analysis. The data has the following labels: i) cyberbullying (yes/no)—whether the question-answer pair has any cyberbullying words, and ii) severity (1–10)—severity of bullying words in the messages. We use these labels to guide our analysis of Sarahah messages and perform a comparative analysis of Sarahah and Formspring. The messages exchanged via Formspring are likely less personal as it does not necessarily require people to be friends to send messages. Comparing Sarahah data with Formspring brings out the similarities and differences in bullying messages between friends (Sarahah) and other more general cyberbullying instances and also helps in understanding the utility of our models in understanding other cyberbullying data.

Differences in Severity

First, we train a logistic regression classifier using the term frequency of words in Formspring messages as features to predict severity in Sarahah messages. We group severity levels from 1–10 into three levels: low (1–3), medium (4–6), and high (7–10). We use these severity predictions to classify Sarahah messages into three categories: low, medium, and high severity. We find that there are $19,798$ messages with low severity, $1,414$ messages with medium severity, and 308 messages with high severity. Due to the absence of labeled Sarahah data, we resort to evaluating the effectiveness of our predictions by comparing differences in LIWC categories across messages in different severity levels.

In Table XII, we show the LIWC comparisons between low severity messages vs. high severity messages in Sarahah. We note that messages categorized as high severity have a significantly higher mean than messages categorized as low across the LIWC categories of *anger*, *negemo*, and *sexual*. We also perform a similar analysis of user responses to Sarahah messages in the different severity levels. Table XIII gives the differences in LIWC categories across user responses to messages in different severity levels. Again, we observe that across the LIWC categories of *anger*, *negemo*, and *sexual*, we observe a statistically significant difference in the means, with responses to high severity having higher means. This again confirms that high severity bullying messages incite harsher responses, contributing further to the cyberbullying content on the web. We observe from Table XIII that reactions to messages in the low severity category have comparable amount of “sadness” when compared to responses to messages in the high severity category (difference between the means is not statistically significant), making it important to not ignore the low severity bullying messages.

LIWC	Severity	Mean	Severity	Mean	p-value $<$
anger	low	0.798	high	5.7634	0.00001
negemo	low	2.9397	high	7.077	0.00001
sad	low	1.588	high	0.3982	0.00001
sexual	low	0.6721	high	1.8776	0.00001

TABLE XII
Differences in LIWC categories between Sarahah messages falling under low/high severity.

LIWC	Severity	Mean	Severity	Mean	p-value $<$
anger	low	0.2156	high	1.8106	0.00001
negemo	low	0.9537	high	2.8785	0.00001
sad	low	0.4615	high	0.4226	0.7112
sexual	low	0.0866	high	0.3353	0.00001

TABLE XIII
Differences in LIWC categories between user responses to Sarahah messages falling under low/high severity.

LIWC	Data 1	Mean	Data 2	Mean	p-value $<$
anger	Sarahah	4.2877	Formspring	1.8744	0.00001
sexual	Sarahah	11.7727	Formspring	0.8861	0.00001
negemo	Sarahah	6.1787	Formspring	2.9947	0.00001

TABLE XIV
Differences in LIWC categories between Sarahah and Formspring messages in *sexual* category.

LIWC	Data 1	Mean	Data 2	Mean	p-value $<$
anger	Sarahah	18.3236	Formspring	1.0075	0.00001
negemo	Sarahah	20.9391	Formspring	2.1218	0.00001

TABLE XV
Differences in LIWC categories between Sarahah and Formspring messages in *hate* category.

LIWC	Data 1	Mean	Data 2	Mean	p-value $<$
sexual	Sarahah	1.1546	Formspring	0.4112	0.00001
negemo	Sarahah	2.0444	Formspring	1.9672	0.5225

TABLE XVI
Differences in LIWC categories between Sarahah and Formspring messages in *flirting* category.

LIWC	Data 1	Mean	Data 2	Mean	p-value $<$
anger	Sarahah	3.382	F-Anon	1.5093	0.00001
negemo	Sarahah	8.465	F-Anon	2.24	0.0001
sexual	Sarahah	3.314	F-Anon	0.9309	0.00001
anger	Sarahah	3.382	F-NA	0.8495	0.00001
negemo	Sarahah	8.465	F-NA	2.1022	0.008
sexual	Sarahah	3.314	F-NA	0.3638	0.003

TABLE XVII
Differences in LIWC categories between Sarahah messages and anonymous (F-Anon) and non-anonymous (F-NA) Formspring messages.

Differences in Coarse-grained Bullying Categories

We apply our coarse-grained seeded LDA model on Formspring data to understand the distribution of messages across the different coarse-grained bullying categories. We find that there are $3,692$ messages in sexual category, $2,118$ messages in hate category, and $7,346$ messages in flirting category.

The percentage of flirting messages in Formspring is lower than Sarahah data, while the percentage of sexual messages is higher. Comparing LIWC differences in coarse-grained bullying categories between Sarahah and Formspring, we find that Formspring has overall less negative emotion, anger, sexual words than Sarahah as given in Tables XIV, XV, and XVI. These differences further illustrate the alarming amount of negativity and bullying that can exist in anonymous messages exchanged between friends in a social network (Sarahah) as opposed to an impersonal setting where the concerned users may not know each other personally (Formspring), further emphasizing the importance of studying various forms and settings of cyberbullying in detail. These similarities and differences help us understand more about bullying in different settings.

Differences in Anonymity

Since Formspring provides the user with an option to post anonymously and non-anonymously, we examine the differences in LIWC categories across anonymous Formspring messages (F-Anon), non-anonymous Formspring messages (F-NA), and Sarahah in Table XVII. It is interesting to note that the negativity is less in non-anonymous Formspring messages when compared to anonymous Formspring messages, indicated by the lower means for the *anger*, *sexual*, and *negemo* categories. This indicates that users overall tend to be harsher in anonymous circumstances. We find that Formspring has lower mean values overall across the LIWC categories, again signaling an enhanced possibility of cyberbullying in an anonymous setting where users have personal connections with each other.

Conclusion

In this work, we performed an extensive quantitative and qualitative analysis of the presence of bullying, abusive, and profane language online by studying the content of messages exchanged using a recently released anonymous web/mobile messenger application, *Sarahah*. Our analysis brings forth the different types of bullying present in a unique setting where anonymous exchanges are exchanged between people who are friends on a social network. We performed fine-grained analysis using seeded LDA, categorizing the messages into different coarse and fine-grained bullying categories. Our analysis revealed that responses to messages in *sexual*, *hate*, and *flirting* categories garner high average values for negative emotion, anger, and negative sentiment, making it important to study bullying at a fine-grained level. We used this categorization to perform a comparative analysis across messages in the different bullying categories within Sarahah and with another labeled cyberbullying dataset, Formspring, studying the differences in psychological LIWC categories and sentiment. Our comparative analysis with Formspring data shows the potential applicability of our models and analysis to other cyberbullying content on the web. There are several exciting directions to go from here. We plan to extend our analysis and models to fine-grained bullying categories to accurately understand the purpose, nature and effect of these messages. Extending our models to include subtle signals of irony and sarcasm in both the messages and the responses that may not necessarily use the standard keywords can help in thoroughly identifying bullying messages. Since each person's reaction to bullying messages is different, understanding finer grained signals in users' responses can help us perform personalized bullying detection and prevention.

References

[1] "Cyberbullying Facts," https://cyberbullying.org/facts, Online; Accessed: 2018-05-24.

[2] S. K. Schneider, L. O'donnell, A. Stueve, and R. W. Coulter, "Cyberbullying, school bullying, and psychological distress: A regional census of high school students," *American Journal of Public Health*, vol. 102, no. 1, pp. 171–177, 2012.

[3] E. Aboujaoude, M. W. Savage, V. Starcevic, and W. O. Salame, "Cyberbullying: Review of an old problem gone viral," *Journal of Adolescent Health*, vol. 57, no. 1, pp. 10–18, 2015.

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

[5] "The Skinny on Sarahah," https://cyberbullying.org/sarahah, Online; Accessed: 2018-05-24.

[6] F. Morstatter, J. Pfeffer, H. Liu, and K. M. Carley, "Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose." in *Proceedings of the Conference on Social and Information Networks*, 2013.

[7] "StopBullying.gov," https://www.stopbullying.gov/, Online; Accessed: 2017-09-12.

[8] J. Jagarlamudi, H. Daumé, III, and R. Udupa, "Incorporating lexical priors into topic models," in *Proceedings of the Conference of the European Chapter of the ACL (EACL)*, 2012.

[9] K. Reynolds, A. Kontostathis, and L. Edwards, "Using machine learning to detect cyberbullying," in *Proceedings of the International Conference on Machine learning and applications and workshops (ICMLA)*, 2011.

[10] H. Hosseinmardi, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra, "Prediction of cyberbullying incidents in a media-based social network," in *Proceedings of the Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2016.

[11] E. Raisi and B. Huang, "Cyberbullying detection with weakly supervised machine learning," in *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2017.

[12] Q. Huang, V. K. Singh, and P. K. Atrey, "Cyber bullying detection using social and textual analysis," in *Proceedings of the Workshop on Socially-Aware Multimedia*, 2014.

[13] V. Nahar, X. Li, and C. Pang, "An effective approach for cyberbullying detection," *Communications in Information Science and Management Engineering*, vol. 3, no. 5, p. 238, 2013.

[14] H. Dani, J. Li, and H. Liu, "Sentiment informed cyberbullying detection in social media," in *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)*, 2017.

[15] S. Kumar, J. Cheng, and J. Leskovec, "Antisocial behavior on the web: Characterization and detection," in *Proceedings of the 26th International Conference on World Wide Web Companion*, ser. WWW '17 Companion, 2017.

[16] J. L. Bigelow, A. Edwards (Kontostathis), and L. Edwards, "Detecting cyberbullying using latent semantic indexing," in *Proceedings of the Workshop on Computational Methods for CyberSafety*, 2016.

[17] H. Sanchez and S. Kumar, "Twitter bullying detection," *ser. NSDI*, vol. 12, pp. 15–15, 2011.

[18] H. Zhong, H. Li, A. Squicciarini, S. Rajtmajer, C. Griffin, D. Miller, and C. Caragea, "Content-driven detection of cyberbullying on the instagram social network," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2016.

[19] H. H. S. Li, Z. Yang, Q. Lv, R. I. R. R. Han, and S. Mishra, "A comparison of common users across instagram and ask.fm to better understand cyberbullying," in *Proceedings of the Conference on Big Data and Cloud Computing (BdCloud)*, 2014.

[20] H. Margono, X. Yi, and G. K. Raikundalia, "Mining indonesian cyber bullying patterns in social networks," in *Proceedings of the Australasian Computer Science Conference*, 2014.

[21] E. Whittaker and R. M. Kowalski, "Cyberbullying via social media," *Journal of School Violence*, vol. 14, no. 1, pp. 11–29, 2015.

[22] L. Corcoran, C. M. Guckin, and G. Prentice, "Cyberbullying or cyber aggression?: A review of existing definitions of cyber-based peer-to-peer aggression," *Societies*, vol. 5, no. 2, pp. 245–255, 2015.

[23] C. Van Hee, E. Lefever, B. Verhoeven, J. Mennes, B. Desmet, G. De Pauw, W. Daelemans, and V. Hoste, "Automatic detection and prevention of cyberbullying," in *Proceedings of the International Conference on Human and Social Analytics*, P. Lorenz and C. Bourret, Eds., 2015.

[24] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali, "Mean birds: Detecting aggression and bullying on twitter," in *Proceedings of the ACM Web Science Conference (WebSci)*, 2017.

[25] J. Cheng, M. Bernstein, C. Danescu-Niculescu-Mizil, and J. Leskovec, "Anyone can become a troll: Causes of trolling behavior in online discussions," in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW)*, 2017.

[26] D. U. Patton, K. McKeown, O. Rambow, and J. Macbeth, "Using natural language processing and qualitative analysis to intervene in gang violence: A collaboration between social work researchers and data scientists," 2016.

[27] A. Bellmore, A. J. Calvin, J.-M. Xu, and X. Zhu, "The five ways of bullying on twitter: who, what, why, where, and when," *Computers in Human Behavior*, vol. 44, pp. 305–314, 2015.

[28] R. S. Tokunaga, "Following you home from school: A critical review and synthesis of research on cyberbullying victimization," *Computers in Human Behavior*, vol. 26, no. 3, pp. 277–287, 2010.

[29] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the detection of textual cyberbullying." *The Social Mobile Web*, vol. 11, no. 02, 2011.

[30] W. Warner and J. Hirschberg, "Detecting hate speech on the world wide web," in *Proceedings of the Workshop on Language in Social Media*, 2012.

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

[32] "Sarahah Trends by Google Trends," https://trends.google.com/trends/explore?q=sarahah, Online; Accessed: 2018-05-24.

[33] "Sarahah application blocked by iphone and android," https://www.independent.co.uk/life-style/gadgets-and-tech/news/sarahah-banned-iphone-android-download-app-store-down-not-work html, Online; Accessed: 2018-05-24.

[34] Twitter api Python Library. [Online]. Available: https://pypi.python.org/pypi/twitter

[35] R. Smith, "An overview of the tesseract ocr engine," in *Proceedings of the Conference on Document Analysis and Recognition (ICDAR)*, 2007.

[36] T. Piehn and A. Piehn, "Voice enabled digital camera and language translator," Feb. 20 2001, uS Patent App. 09/789,220.

[37] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, Mar. 2003.

[38] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: Liwc 2001," *Mahway: Lawrence Erlbaum Associates*, vol. 71, no. 2001, p. 2001, 2001.

[39] "Natural Language Tool Kit (NLTK) in Python." [Online]. Available: {http://www.nltk.org}

[40] "Formspring, Cyberbullying, and Alexis Pilkington," https://cyberbullying.org/formspring, Online; Accessed: 2018-05-24.