Analyzing Labeled Cyberbullying Incidents on the Instagram Social Network

Homa Hosseinmardi[1(✉)], Sabrina Arredondo Mattson[2], Rahat Ibn Rafiq[1], Richard Han[1], Qin Lv[1], and Shivakant Mishra[1]

[1] Computer Science Department, University of Colorado Boulder, Boulder, CO, USA
{homa.hosseinmardi,rahat.ibnrafiq,
richard.han,qin.lv,shivakant.mishra}@colorado.edu
[2] Institute of Behavioral Science, University of Colorado Boulder, Boulder, CO, USA
sabrina.mattson@colorado.edu

Abstract. Cyberbullying is a growing problem affecting more than half of all American teens. The main goal of this paper is to study labeled cyberbullying incidents in the Instagram social network. In this work, we have collected a sample data set consisting of Instagram images and their associated comments. We then designed a labeling study and employed human contributors at the crowd-sourced CrowdFlower website to label these media sessions for cyberbullying. A detailed analysis of the labeled data is then presented, including a study of relationships between cyberbullying and a host of features such as cyberaggression, profanity, social graph features, temporal commenting behavior, linguistic content, and image content.

1 Introduction

As online social networks (OSNs) have grown in popularity, instances of cyberbullying in OSNs have become an increasing concern. In fact more than half of American teens have reported being the victims of cyberbullying [1]. Moreover, research has found links between experiences of cyberbullying and negative outcomes such as decreased performance in school, absenteeism, truancy, dropping out, and violent behavior [2], and potentially devastating psychological effects such as depression, low self-esteem, suicide ideation, and even suicide [3–6], that can have long term effects in the future life of victims [7]. Incidents of cyberbullying with extreme consequences such as suicide are now routinely reported in the popular press. For example cyberbullying of Jessica Logan via her image shared in Facebook and MySpace and of Hope Sitwell with her image shared in MySpace is attributed to their suicides [8], [9].

Given the gravity of the consequences cyberbullying has on its victims and its rapid spread among middle and high school students, there is an immediate and pressing need for research to understand how cyberbullying occurs in OSNs today, so that effective techniques can be developed to accurately detect cyberbullying. In [6], it is reported that experts in the field of cyberbullying could favor automatic monitoring of cyberbullying on social networking sites and propose effective follow-up strategies.

T.-Y. Liu et al. (Eds.): SocInfo 2015, LNCS 9471, pp. 49–66, 2015.
DOI: 10.1007/978-3-319-27433-1_4

Our work makes the important distinction between cyberaggression and cyberbullying. Cyberaggression is defined as aggressive online behavior that uses digital media in a way that is intended to cause harm to another person[10]. Examples include negative content and words such as profanity, slang and abbreviations that would be used in negative posts such as hate, fight, wtf. Cyberbullying is one form of cyberaggression that is more restrictively defined as (1) an act of aggression online with (2) an imbalance of power between the individuals involved and (3) repetition of the aggression [2,10–15]. Similar to traditional bullying, it is the combination of the aggressive behavior, repeated acts, and the victim's inability to defend himself or herself that severely impacts many teens [2]. Particularly important in the context of cyberbullying, is the permanent nature of the online posts (until they're removed), the ease and wide distribution in which aggressive posts can be made, the difficulty of identifying the behavior, the ability to be connected and exposed to online interaction 24/7, and the growing number of potential victims and perpetrators [4]. The power imbalance can take on a variety of forms including physical, social, relational or psychological [14,16–18], such as a user being more technologically savvy than another [2], a group of users targeting one user, or a popular user targeting a less popular one [19]. Repetition of cyberbullying can occur over time or by forwarding/sharing a negative comment or photo with multiple individuals [19].

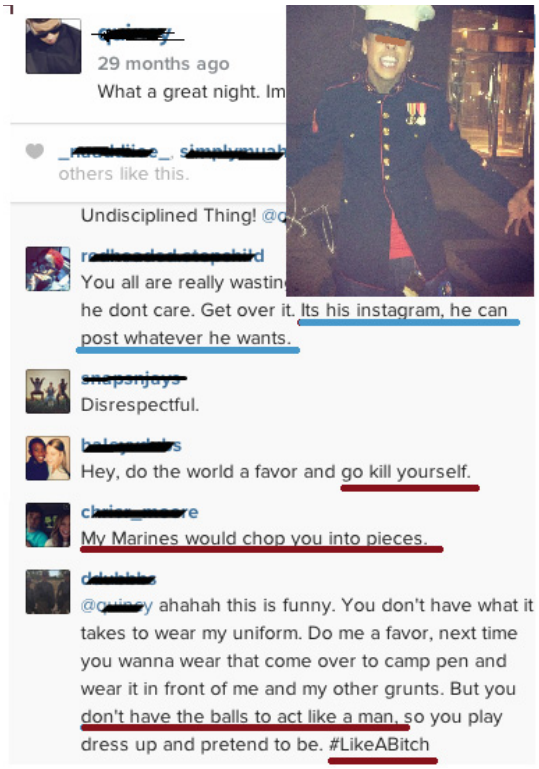

Fig. 1. An example of comments posted on Instagram. To give more room for the text, we have moved the associated image to overlay some of the text.

Facebook, Twitter, YouTube, Ask.fm, and Instagram have been listed as the top five networks with the highest percentage of users reporting experience of cyberbullying [20]. Instagram is of particular interest as it is a media-based social network, which allows users to post and comment on images. An example of an Instagram media session is shown in Figure 1. Cyberbullying in Instagram can happen in different ways, including posting a humiliating image of someone else by perhaps editing the image, posting mean or hateful comments, aggressive captions or hashtags, or creating fake profiles pretending to be someone else [21].

The main goal of this paper is to study cyberbullying in Instagram. To do so, we first collected a large sample of Instagram data comprised of 3,165K media

sessions (images and their associated comments) taken from 25K user profiles. Next, we provided labeling instructions and Instagram media sessions to human labelers at the crowd-sourced CrowdFlower website to identify occurrences of cyberbullying and cyberaggression in Instagram. We then analyzed the labeled data set, reporting the relationship of different features of these media sessions to cyberbullying. This paper make the following important contributions:

- We provide a clear distinction between cyberbullying and general cyberaggression. Cyberbullying is a form of cyberaggression that can have devastating effects on the victims and perpetrators. Most of the prior research in this area is more appropriately described as investigating cyberaggression.
- We obtain ground truth cyberbullying behavior in Instagram by instructing human crowd-sourcers to label Instagram images and their associated comments according to both the more restrictive definition of cyberbullying and the more general definition of cyberaggression. Labelers are provided with the image and its associated comments at the same time to be able to understand the context and label accordingly.
- We present a novel detailed analysis of the labeled media sessions, including the relationships between cyberbullying and a host of factors, such as cyberaggression, profanity, social graph properties (liking, followers/following), the interarrival time of comments, the linguistic content of comments, and labeled image content.

2 Related Works

Prior works that investigated cyberbullying [22–32] are more accurately described as research on cyberaggression, since these works do not take into account both the frequency of aggression and the imbalance of power. These works have largely applied a text analysis approach to online comments, since this approach results in higher precision and lower false positives than simpler list-based matching of profane words [33]. Previous research [27,29,34,35] applied text based cyberbullying on Formspring.me and Myspace dataset. Dinakar *et al.* investigated both explicit and implicit cyberbullying by analyzing negative text comments on YouTube and Formspring profiles [25]. Sanchez and Kumar proposed using a Naive Bayes classifier to find inappropriate words in Twitter text data for bullying detection [26]. They tracked potential bullies, their followers, and the victims. Also some researchers tried to detect bullies and victims by looking at the number of received and sent, beside detecting aggressive comments [36] and [32] . Dadvar *et al.* investigated how combining text analysis with MySpace user profile information such as gender can improve the accuracy of cyberbullying detection in OSNs [23] . Huang *et al.* [37] has consider some graph properties besides text features, however they also worked only over comment-based labeled data. Another work has looked at the time series of posted comments of Formepring dataset, in which each question answer pair was labeled separately as cyberaggression and then their severity predicted [38]. These works largely focus on text-based analysis and unlike our work

do not examine the features associated with the media objects such as images or videos belonging to those comments, as in Instagram. Kansara *et al.* [39] suggest only a framework for using images beside text for detecting cyberbullying.

Other work analyzed profanity usages in Instagram [40] and Ask.fm [41] comments, but did not label the data in terms of cyberbullying. Additional research investigated aspects of the Instagram social network, but not in the context of cyberbullying. For example, [42] explored users' photo sharing experience in a museum. Silva *et al.* [43] considered the temporal photo sharing behavior of Instagram users and Hu *et al.* [44] categorized Instagram images into eight popular image categories and the Instagram users into five types in terms of their posted images. [45] concluded that Instagram users tend to be more active during weekends and at the end of the day, and that Instagram users are more likely to like and comment on the medias that are already popular, thereby inducing the rich get richer phenomenon.

3 Data Collection

Starting from a random seed node, we identified 41K Instagram user ids using a snowball sampling method from the Instagram API. Among these Instagram ids, 25K (61%) users had public profiles while the rest had private profiles. Due to the limitation on the private profiles' lack of shared information, the 25K public user profiles comprise our sample data set. For each public Instagram user, the collected profile data includes the media objects (videos/images) that the user has posted and their associated comments, user id of each user followed by this user, user id of each user who follows this user, and user id of each user who commented on or liked the media objects shared by the user. We consider each media object plus its associated comments as a *media session*.

Labeling data is a costly process and therefore in order to make the labeling of cyberbullying more manageable, we sought to label a smaller subset of these media sessions. To have a higher rate of cyberbullying instances, we considered media sessions with at least one profanity word in their associated comments. We tag a comment as "negative" using an approach similar to [41]. For this set of 25K users, 3,165K unique media sessions were collected, where 697K of these sessions have at least one profane word in their comments by users other than the profile owner, where a profane word is obtained from a dictionary [46], [47].

In addition, we needed media sessions with enough comments so that labelers could adequately assess the frequency or repetition of aggression, which is an important part of the cyberbullying definition. We selected a threshold of 15 as a lower bound on the number of comments in a media session, considering that the average ratio of comments posted by users other than friends to comments posted by the profile owner in an Instagram profile is around 16 [40]. At the end 2,218 media sessions (images and their associated comments) were selected randomly for the task of labeling.

4 Cyberbullying Labeling

In this section, we explain the design and methodology for labeling the selected set of media sessions. In Instagram, each media session consists of a media object posted by the profile owner and the corresponding comments for the media object. For example, Figure 1 illustrated a media session in which hateful comments were posted for an Instagram image on the profile of the owner. Such a media session was used in the labeling process, in which labelers were shown both the image and the associated text comments in order to make determinations for cyberaggression and cyberbullying.

With input from a social science expert, co-author Mattson, we designed simple instructions to help human contributors identify whether the media session constituted an act of cyberaggression or cyberbullying. During the instructional phase prior to labeling, contributors were given the aforementioned definitions of cyberaggression and cyberbullying along with related examples. The example questions provided more details to help contributors accurately label the online behavior and distinguish between cyberaggression and cyberbullying based on the social science definitions they were provided.

In order to provide quality control, we only permitted the highest-rated contributors on CrowdFlower to have access to our job. Next, a mentoring phase was provided for the potential contributors that included instructions and a set of example media sessions with the correct label. Further, to monitor the quality of the contributors and filter out the spammers, potential contributors were asked to answer a set of test questions in two phases: quiz mode and work mode. Potential contributors needed to answer correctly a minimum number of test questions to pass the quiz mode and qualify as a contributor for the job. We also incorporated quality control checks during the labeling process (work mode) by inserting random test questions. A contributor was filtered out if he/she failed this work mode.

Finally, a minimum time threshold was set to filter out contributors who rushed too quickly through the labeling process. The minimum number of test questions to get back high-quality data was recommended by CrowdFlower. More details about the labeling process statistics have been provided in the appendix.

An example of a media session that each crowd-sourcer was asked to label is shown in Figure 2. Each media session was then labeled by five contributors that asked them to use the instructions and definitions we provided to determine whether the post included cyberaggressive behavior or

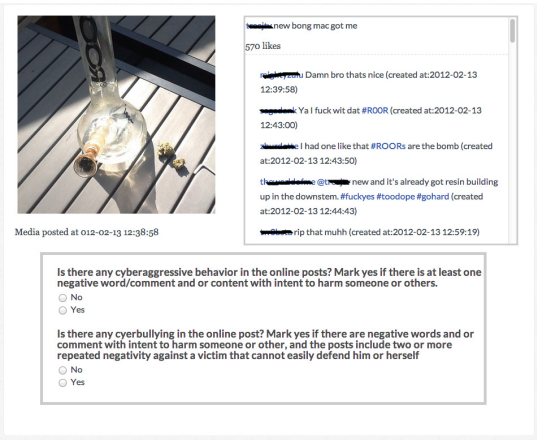

Fig. 2. An example of the labeling study, showing an image and its corresponding comments, and the study questions.

cyberbullying behavior. Specifically we asked 1) Is there any cyberaggressive behavior in the online posts? Mark yes if there is at least one negative word/comment and or content with intent to harm someone or others and 2) Is there any cyberbullying behavior in the online posts? Mark yes if there are negative words and repeated negativity against a victim that cannot easily defend him or herself.

We conducted a separate second phase of labeling focused only on the image contents in order to identify the content and category of the image. We provided separate instructions to label the image contents of media sessions, so that we could investigate the relationship between cyberbullying and cyberaggression and image content. We reasoned that the content or category of an image may help identify incidents of cyberbullying and cyberaggression. More detail regarding image labeling has been explained in the appendix.

5 Analysis and Characterization of Ground Truth Data

We submitted our first study with 2,218 media sessions (images and their associated comments) to CrowdFlower. CrowdFlower assesses a degree of trust for each contributor based on the percentage of correctly answered test questions, as explained in Section 4. This trust value is incorporated by CrowdFlower into a weighted version of the majority voting method called a "confidence level" for each labeled media session. We decided to keep media sessions whose weighted trust-based metric was equal to or greater than 60%. We deemed them to be strong enough support for majority voting from contributors with higher trust. Overall, 1,954 (88%) of the original pure majority-vote based media sessions wound up in this higher-confidence cyberbullying-labeled group. For this higher-confidence data set, 29% of the media sessions belonged to the "bullying" group while the other 71% were deemed to be not bullying.

5.1 Labeling and Negativity Analysis

The distribution of the media sessions based on the number of votes (out of five votes) received for cyberaggression and cyberbullying respectively has been provided in Figure 3. The left chart shows the fraction of samples that have been labeled as cyberaggression k times, and the right chart shows the fraction of samples that have been labeled as cyberbullying k times. The higher the number of votes for a given media session, the more confidence we have that the media session contains an incident of cyberaggression or cyberbullying, with five votes means unanimous agreement. Similarly, the lower the number of votes for a given media session, the more confidence we have that the media session *does not* contain an incident of cyberaggression or cyberbullying, with zero votes means unanimous agreement.The inter-rater agreement Fleiss-Kappa value for cyberbullying is 0.5 and for cyberaggression is 0.52.

We notice that for both cyberaggression and cyberbullying, most of the probability mass is around media sessions labeled by all four or five contributors the

same, i.e. either 0 or 1 votes (about 50% for cyberaggression and about 62% for cyberbullying), or 4 or 5 votes (about 31% for cyberaggression and about 23% for cyberbullying). *Thus, a key finding is that the contributors are mostly in agreement about what behavior constitutes cyberaggression, and what behavior constitutes cyberbullying in Instagram media sessions.* Only about 13−17% of the media sessions have two or three votes, which indicates that there is some disagreement in a small fraction of media sessions about whether the session contains an incident of cyberaggression or cyberbullying. This disagreement can be attributed to the fact that different people have different levels of sensitivity and a conversation may seem normal to one person and hurtful to another.

Next, we observe that about 30% of the media sessions have not been labeled as cyberaggression by any of the five contributors. Since all media sessions contained at least one comment with one or more profane word, this suggests that only employing a profanity usage threshold to detect cyberaggression can produce many false positives. We make a similar observation for cyberbullying. We notice about 40% of the media sessions have not been labeled as cyberbullying by any of the five contributors. Applying a majority voting criterion to a binary label as cyberbullying or not, 30% of the samples have been labeled as cyberbullying. This is despite the fact that all the media sessions contain at least one profane word. This leads us to our second important finding. *A classifier design for cyberbullying detection cannot solely rely on the usage of profanity among the words in image-based discussions, and instead must consider other features to improve accuracy.*

In order to understand the relationship between cyberaggression and cyberbullying, we plotted in Figure 4 a two-dimensional heat map that shows the distribution of media sessions as a function of the number of votes each media session received for cyberaggression and cyberbullying. We observe that a

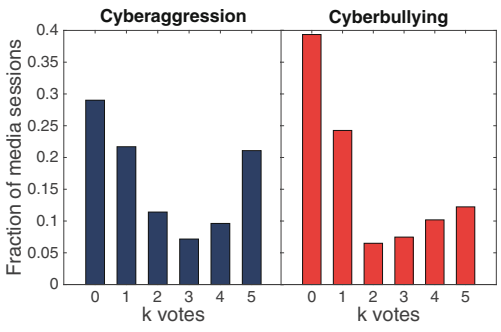

Fig. 3. Fraction of media sessions that have been voted k times as cyberaggression (left) or cyberbullying (right).

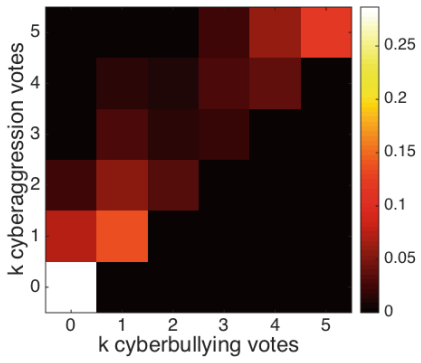

Fig. 4. Two-dimensional distribution of media sessions as a function of the number of votes given for cyberaggression versus the number of votes given for cyberbullying, assuming five labelers.

significant fraction of the sessions exhibit strong agreement in terms of either receiving high numbers of votes for both cyberbullying and cyberaggressions, or receiving low numbers of votes for both cyberbullying and cyberaggression. This can be inferred from the high energy in the upper right and lower left part of the diagonal. In addition, it is promising that the area below the diagonal is essentially zero, meaning no session has received more votes for cyberbullying than for cyberaggression. This conforms with the definition that cyberbullying is a subset of cyberaggression. The Pearson's correlation between number of votes for cyberbullying and number of votes for cyberaggression is 0.9.

We see that the remaining significant energy in the distribution appears in the area above the diagonal. Media sessions in this area exhibit the property that if they receive N_1 cyberbullying votes and N_2 cyberaggression votes, then $N_2 \geq N_1$. The area where $N_1 \leq 2$ and $N_2 \geq 3$ corresponds to cases where there is cyberaggression but not cyberbullying. These observations lead us to our third important finding. *A media session that exhibits cyberaggression does not necessarily exhibit cyberbullying, and a classifier design for cyberbullying detection must go much beyond merely detecting cyberaggression.* This is a very important finding, because as we noted in Section 2, prior work on detecting cyberbullying has mainly focused on detecting cyberaggression as they do not take into account the frequency of aggression or imbalance of power, which are crucial features of cyberbullying.

Finally, we are interested in understanding the relation between cyberbullying/cyberaggression and the percentage of negativity in the comments. We divided all the media sessions into nine different bins based on the percentage of negativity in their comments. Bin $(n_1 - n_2]$ contains all media sessions with bigger than $n_1\%$ and smaller than or equal to $n_2\%$ negativity. None of the media sessions contained more than 90% negative comments. Next, we calculated percentage media sessions for each bin that can be identified as cyberaggression or cyberbullying based on majority of votes, i.e. where the number of votes is 3 or higher.

Figure 5 shows these fractions, left figure for cyberaggression and right figure for cyberbullying. We observe that as the percentage of negativity increases, so does the fraction of media sessions up until 50% negativity for cyberaggression and 60% for cyberbullying. This increase is as expected, since cyberaggression or cyberbullying is typically accompanied with negativity in the postings. However, we notice that the percentage of cyberaggression or cyberbullying starts decreasing after these peaks as the percentage of negativity

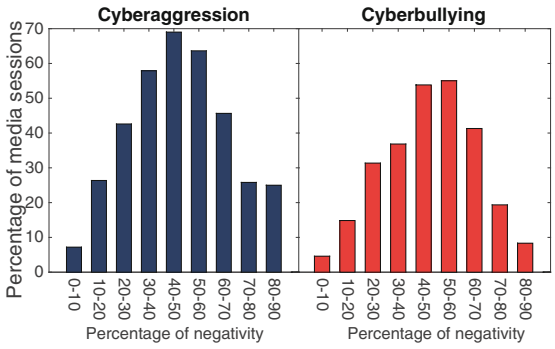

Fig. 5. Percentage of media sessions that have been labeled as cyberaggression (left) and cyberbullying (right) versus their negativity percentages.

increases. This is quite an unexpected result and seems counter-intuitive. To understand this, we examined closely the media sessions that have very high negativity. We noticed that these media sessions typically involved discussions about sports, politics, tattoos, or were just friendly talks. People tend to use lots of profanity words in such discussions, even though they are not insulting any one person in particular. This leads us to our final important finding about negativity analysis. *A media session with a significantly high percentage of negativity (more than 60-70%) typically implies a low probability that the session contains a cyberbullying incident.*

5.2 Temporal and Graph Properties Analysis

Since different comments in a media session are posted at different times, it is important to understand the relationship between the temporal nature of comment postings and cyberbullying/cyberaggression. We define the strength of cyberbullying/cyberaggression as the number of votes received for labeling a media session as cyberbullying/cyber-aggression, and explore the Pearson's correlation of cyberbullying/cyberaggression strength and temporal behavior comment arrivals. We would like to understand how human contributors incorporated the definition of cyberbullying, which includes the temporal notion of repetition of negativity over time, into their labeling. Given the time stamps on the collected comment, we compute the interarrival time between two consequent comments. We then count the number of interarrival times of comments in a media session that have a value less than x = 1min, 5 min, ..., 6 months.

Figure 6 illustrates the correlation between the number votes and the number of comments arrive with $\leq x$ seconds after their previously received comment. We see that there is a correlation of about 0.3 between the strength of support for cyberbullying and media sessions in which there are frequent postings within one hour of previous post. Further, we find that as we expand the allowable interarrival times between comments, the correlation weakens considerably. A similar pattern was observed for cyberaggression. In fact, on average around 40% of the comments arrive in less than 1 hour after previously received comments in cyberbullying media sessions, however only 30% of the comments have been received with the same interarrival time in non-cyberbullying samples ($p < 0.001$, based on t-test). *A key finding here is that media sessions that contain cyberbullying have relatively low comment interarrival times, that is the comments in these media sessions are posted quite frequently.*

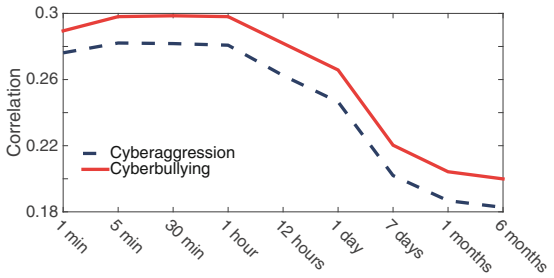

Fig. 6. Pearson's correlation between the number of votes and the number of comments that arrive in $\leq x$ seconds after their previously received comment.

Table 1. Mean values of social graph properties for cyberbullying versus non-cyberbullying samples and aggression versus non-cyberaggression. ($*p < 0.05$).

Label	*Likes	*Media objects	Following	Followers
Non-cyberbullying	9,684.4	1,145.7	668.1	415,676.2
Cyberbullying	7,029.0	1,198.3	626.7	463,073.1
Non-cyberaggression	9,768.6	1,133.7	665.9	421,075.3
Cyberaggression	7,551.3	1,204.3	640.3	440,403.6

Next, we examine the relationship between cyberbullying/cyberaggression and the social network graph features such as the number of likes for a given media object, number of comments posted for a media object, number of users a user is following, and the number of followers of a user. Table 1 shows these numbers, for categories of non-cyberbullying sessions, cyberbullying sessions, non-cyberaggression sessions and cyberaggression sessions. We observe that media sessions that contain cyberbullying/cyberaggression share more media objects than media sessions that do not contain cyberbullying/cyberaggression, but on average receive lower number of likes. Souza *et al.*'s [45] analysis of Instagram users shows there is a positive correlation between number of followers and number of likes for typical Instagram users. Users who receive cyberbullying do not follow the same pattern. In fact, the average number of likes per post for non-cyberbullying sessions is 4 times the average number of likes for cyberbullying sessions, and the average number of likes per post for non-cyberaggression sessions is 4.5 times the average number of likes for cyberaggression sessions. In terms of number of following and followers, the distinction is not as pronounced, although we see that the media sessions with cyberbullying/cyberaggression incidents have more followers and less following compared to the media sessions without cyberaggression/cyberbullying. *The key finding here is that the users of media sessions with cyberbullying/cyberaggression have lower number of likes per post while have more followers.*

5.3 Linguistic and Psychological Analysis

We now focus on the pattern of linguistic and psychological measurements of cyberbullying/cyberaggression media sessions versus non-cyberbullying/non-cyberaggression. For this purpose, we have applied Linguistic Inquiry and Word Count (LIWC), a text analysis program to find which categories of words have been used for cyberbullying/cyberaggression labeled media sessions. LIWC evaluates different aspects of word usages in psychologically meaningful categories, by counting the number of the words across the text for each category [48]. LIWC has often been used for studies on variations in language use across different people. Published papers show that LIWC have been validated to perform well in studies on variations in language use across different peoples [49]. We first analyze the number of words, and usage of pronouns, negations and swear words

(Figure 7). Next, we look at some of the personal concerns such as work, achievements, leisure, etc. (Figure 7). Finally, we investigate some of the psychological measurements such as social, family, friends, etc. (Figure 8). For each of these cases, we first obtain the LIWC values for each media session comment set. We then calculate the average LIWC value for each of the four classes: media sessions with cyberbullying, media sessions with no cyberbullying, media sessions with cyberaggression, and media sessions with no cyberaggression. The bars shown in Figures 7-8 represent the ratio of average LIWC value for cyberbullying class to that of non-cyberbullying, and the ratio of average LIWC value for cyberaggression class to that of non-cyberaggression.

In Figure 7, we first notice that the word count for media sessions with cyberbullying/cyberaggression is significantly higher than for media sessions with no cyberbullying/cyberaggression ($p < 10^{-5}$). Next, as expected, for swear words (e.g., damn, piss) and negations (e.g., never, not), the ratio is higher for cyberbullying/cyberaggression category ($p < 10^{-5}$). It is interesting to note that the ratios for the third person pronouns (she, he, they) are more than 1.3 ($p < 10^{-5}$), the ratio for the first person singular pronoun (i) is 0.85, and the ratios for first person plural and second person pronouns (we, you) is closer to 1. This leads us to our first key finding with respect to the linguistic features. *A user is less likely to directly refer to himself/herself and more likely to refer to other people in third person in postings involving cyberbullying or cyberaggression.*

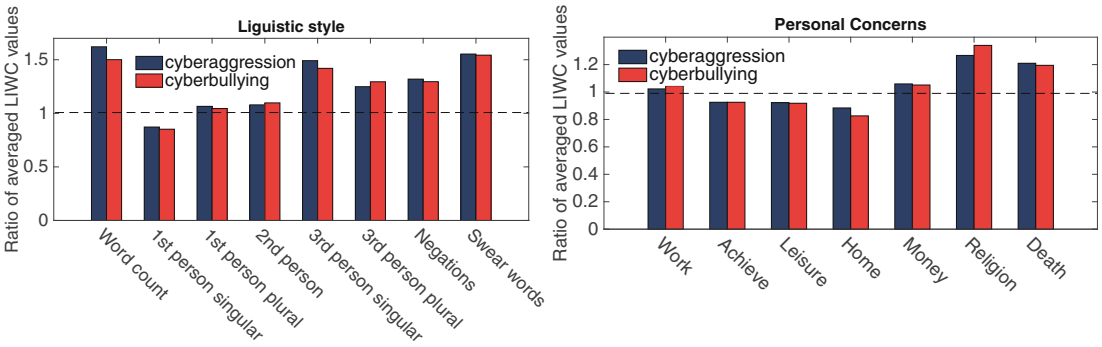

Fig. 7. (left) Ratio of LIWC values of cyberbullying/cyberaggression labeled media sessions to non-cyberbullying/non-cyberaggression class in Linguistic categories. (right)Ratio of LIWC values of cyberbullying/cyberaggression labeled media sessions to non-cyberbullying/non-cyberaggression class in Personal Concerns categories.

For personal concerns (Figure 7), "religion" (e.g., church, mosque) and "death" (e.g., bury, kill) categories have higher ratios (more than 1.2, p < 0.1). This is in line with our findings in our previous work on profanity usage analysis in ask.fm social media, where we observed that there is high profanity usage around words like "muslim" [41]. This suggests that religion-based cyberbullying may be quite prevalent in social media. On the other hand, ratios for personal categories like "work", "money" and 'achieve" are much closer to 1.

For psychological measurements (Figure 8), we notice that the ratios for "negative emotion", "anger", "body", and "sexual" categories are significantly higher than 1 (more than 1.4, $p < 10^{-5}$), and the ratio for "positive emotion" category is significantly lower than 1 (0.76, $p < 10^{-5}$). Higher ratios for "body" (e.g. face, wear) and "sexual" (e.g. slut, rapist) categories provide evidence for appearance-based and sexual-based cyberbullying in social media. For other psychological measurement categories, such as "social", "friend", etc., the ratios are closer to 1. Based on our observations from Figures 7 and 8, our final important finding with respect to linguistic features is as follows: *There is a higher probability of cyberbullying in postings involving religion, death, appearance and sexual hints, and cyberbullying posts typically have higher occurrences of negative emotions and lower occurrences of positive emotions.*

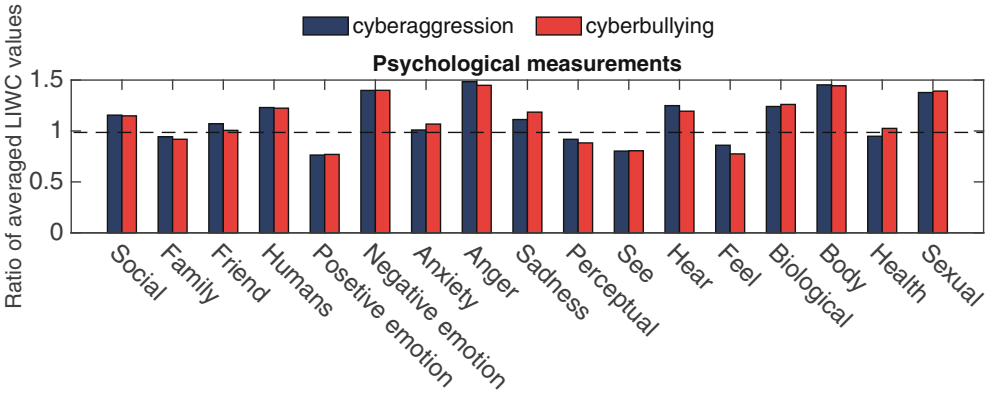

Fig. 8. Ratio of LIWC values of cyberbullying/cyberaggression labeled media sessions to non-cyberbullying/non-cyberaggression class in Psychological categories.

5.4 Image Content Analysis

We now explore the relationship between image content and cyberbullying/cyberaggression in a media session. If the majority of labelers chose a given content category for an image, then that image was counted as belonging to that category. Note that it was possible for contributors to place an image in more than one category. More than 70% of the images were labeled with only one category, and around 20% were labeled with two categories. However, there were a few images that were labeled with up to eight unique categories.

Figure 9 shows the fraction of the contents for all labeled data in the green bar. The "dont know" choice was given as we realized that for some images it is hard to figure out what is in the image. As some images belong to more than one category, the bars will sum up to more than one. First, we observe that the most common labels for image content are Person/People, Text and Sports.

Next, the heights of the blue and red bars embedded inside each green bar relative to the height of the green bar indicate the fraction of images belonging to the media sessions that contained cyberaggression and cyberbullying respectively. For example, for the "Text" category, about 1/3 of the images with "Text"

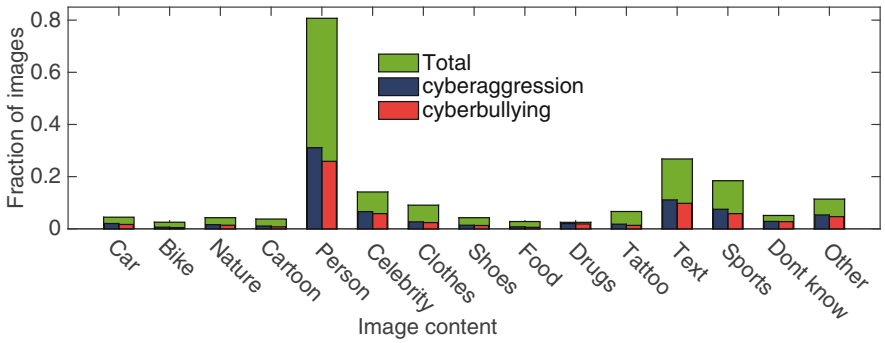

Fig. 9. Fraction of image categories for all media sessions, cyberbullying and cyberaggression classes.

were associated with media sessions containing cyberaggression and cyberbullying. We observe that for some content categories such as "Drugs", the overall fraction is quite small (green bar height is low), but most of the images in those categories do belong to media sessions with cyberaggression/cyberbullying in them. To see this more clearly, Figure 10 plots the fraction of images labeled as cyberaggression/cyberbullying for each content category. We notice that for content category "Drugs", 75% of the images belong to media sessions containing cyberbullying, while for content categories like "Car", "Nature", "Person", "Celebrity", "Text" and "Sport", 30%-40% of the images belong to media sessions containing cyberbullying/cyberaggression. Also, whenever images contain bike, food, tattoo, etc., there is little cyberbullying occurring. *The key finding here is that certain image contents such as Drug are strongly related with cyberbulllying, while some other image contents such as bike, food, etc. have a very low relationship with cyberbullying.*

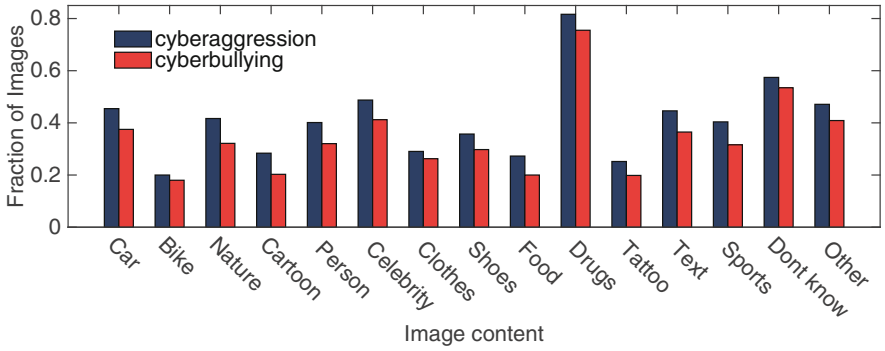

Fig. 10. Fraction of images which have been labeled as cyberbullying and cyberaggression for each content category.

6 Conclusions and Future Work

This paper makes the following major contributions. First, an appropriate definition of cyberbullying that incorporates both frequency of negativity and imbalance of power is applied in large-scale labeling, and is differentiated from cyberaggression. Second, cyberbullying is studied in the context of a media-based social network, incorporating both images and comments in the labeling. We found that labelers are mostly in agreement about what constitutes cyberbullying and cyberaggression in Instagram media sessions. Third, a detailed analysis of the distribution results of labeling of cyberbullying incidents is presented, including a correlation analysis of cyberbullying with other factors derived from images, text comments, and social network meta data. We found a significant number of media sessions containing profanity and cyberaggression were not labeled as cyberbullying, suggesting that detection of cyberbullying must be more sophisticated than merely looking for profanity. We observed that media sessions with very high percentage of negativity above 60-70% actually correspond to a lower likelihood of cyberbullying. Also, media sessions with cyberbullying exhibit more frequent commenting. We found that users of media sessions containing cyberbullying demonstrate a lower number of likes per post. Finally, cyberbullying has a higher probability of occurring when media sessions contain certain linguistic categories such as death, appearance, religion and sexuality content. Similarly, certain image contents such as "drug" are highly related to cyberbullying while other image categories such as "tattoo" or "food" are not.

In the future, we hope to build upon this analysis. We hope to examine more features for their correlation with cyberbullying, such as new image features, mobile sensor data, etc. Such features should be auto-generated by software rather than requiring human labeling. We also wish to obtain greater detail from the labeling process. Streamlining down to two labeling questions improved the response rate, quality and speed, but limited our ability to ask more detailed questions about other aspects of cyberbullying, such as different types and roles.

Acknowledgments. This work was supported in part by the National Science Foundation under awards CNS-1162614 and CNS-1528138.

References

1. National Crime Prevention Council: National Crime Prevention Council (2011). http://en.wikipedia.org/wiki/Cyberbullying (accessed July 6, 2011)
2. Kowalski, R.M., Giumetti, G.W., Schroeder, A.N., Lattanner, M.R.: Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth (2014)
3. McNamee, D.: Cyberbullying 'causes suicidal thoughts in kids more than traditional bullying (2014). http://www.medicalnewstoday.com/articles/273788.php (accessed May 31, 2015)

4. Hinduja, S., Patchin, J.W.: Cyberbullying research summary, cyberbullying and suicide (2010)
5. Menesini, E., Nocentini, A.: Cyberbullying definition and measurement. some critical considerations. Journal of Psychology **217**, 320–323 (2009)
6. Van Royen, K., Poels, K., Daelemans, W., Vandebosch, H.: Automatic monitoring of cyberbullying on social networking sites: From technological feasibility to desirability. Telematics and Informatics **32**, 89–97 (2015)
7. Strickland, A.: Bullying by peers has effects later in life (2015). http://www.cnn.com/2015/05/08/health/bullying-mental-health-effects/index.html (accessed May 2015)
8. cbcNews: Jessica Logan - Victims of bullying (2008). http://www.cbsnews.com/pictures/victims-of-bullying/11/ (accessed May 31, 2015)
9. NoBullying.com: The top six cyberbullying case ever (2015). http://nobullying.com/six-unforgettable-cyber-bullying-cases/ (accessed May 31, 2015)
10. Kowalski, R.M., Limber, S., Limber, S.P., Agatston, P.W.: Cyberbullying: Bullying in the digital age. John Wiley & Sons (2012)
11. Patchin, J.W., Hinduja, S.: An update and synthesis of the research. Cyberbullying prevention and response: Expert perspectives, p. 13 (2012)
12. Hunter, S.C., Boyle, J.M., Warden, D.: Perceptions and correlates of peer-victimization and bullying. British Journal of Educational Psychology **77**, 797–810 (2007)
13. Olweus, D.: Bullying at school: What we know and what we can do. Blackwell (1993)
14. Olweus, D.: School bullying: Development and some important challenges. Annual Review of Clinical Psychology **9**, 751–780 (2013)
15. Smith, P.K., del Barrio, C., Tokunaga, R.: Definitions of bullying and cyberbullying: how useful are the terms? routledge. In: Principles of Cyberbullying Research. Definitions, Measures and Methodology (2012)
16. Dooley, J.J., Pyżalski, J., Cross, D.: Cyberbullying versus face-to-face bullying. Zeitschrift für Psychologie/Journal of Psychology **217**, 182–188 (2009)
17. Monks, C.P., Smith, P.K.: Definitions of bullying: Age differences in understanding of the term, and the role of experience. British Journal of Developmental Psychology **24**, 801–821 (2006)
18. Pyżalski, J.: Electronic aggression among adolescents: An old house with. Youth culture and net culture: Online social practices, p. 278 (2010)
19. Limber, S.P., Kowalski, R.M., Agatston, P.A.: Cyber bullying: A curriculum for grades 6–12. Hazelden, Center City (2008)
20. Ditch the Label Anti Bullying Charity: The annual cyberbullying survey 2013 (2013). http://www.ditchthelabel.org/annual-cyber-bullying-survey-cyber-bullying-statistics/
21. Hinduja, S.: Cyberbullying on Instagram (2013)
22. Ptaszynski, M., Dybala, P., Matsuba, T., Masui, F., Rzepka, R., Araki, K., Momouchi, Y.: In the service of online order tackling cyberbullying with machine learning and affect analysis (2010)
23. Dadvar, M., de Jong, F.M.G., Ordelman, R.J.F., Trieschnigg, R.B.: Improved cyberbullying detection using gender information. In: Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012), Ghent, Belgium, pp. 23–25. University of Ghent, Ghent (2012)
24. Reynolds, K., Kontostathis, A., Edwards, L.: Using machine learning to detect cyberbullying. In: Fourth International Conference on Machine Learning and Applications, vol. 2, pp. 241–244 (2011)

25. Dinakar, K., Jones, B., Havasi, C., Lieberman, H., Picard, R.: Common sense reasoning for detection, prevention, and mitigation of cyberbullying. ACM Trans. Interact. Intell. Syst. **2**, 18:1–18:30 (2012)
26. Sanchez, H., Kumar, S.: Twitter bullying detection. In: NSDI 2012, Berkeley, CA, USA, p. 15. USENIX Association (2012)
27. Kontostathis, A., Reynolds, K., Garron, A., Edwards, L.: Detecting cyberbullying: query terms and techniques. In: Proceedings of the 5th Annual ACM Web Science Conference, pp. 195–204. ACM (2013)
28. Xu, J.M., Jun, K.S., Zhu, X., Bellmore, A.: Learning from bullying traces in social media. In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 656–666. Association for Computational Linguistics (2012)
29. Nahar, V., Al-Maskari, S., Li, X., Pang, C.: Semi-supervised learning for cyberbullying detection in social networks. In: Wang, H., Sharaf, M.A. (eds.) ADC 2014. LNCS, vol. 8506, pp. 160–171. Springer, Heidelberg (2014)
30. Nahar, V., Li, X., Pang, C.: An effective approach for cyberbullying detection. In: Communications in Information Science and Management Engineering, CISME 2013 (2013)
31. Dinakar, K., Reichart, R., Lieberman, H.: Modeling the detection of textual cyberbullying. In: The Social Mobile Web (2011)
32. Nahar, V., Unankard, S., Li, X., Pang, C.: Sentiment analysis for effective detection of cyber bullying. In: Sheng, Q.Z., Wang, G., Jensen, C.S., Xu, G. (eds.) APWeb 2012. LNCS, vol. 7235, pp. 767–774. Springer, Heidelberg (2012)
33. Sood, S., Antin, J., Churchill, E.: Profanity use in online communities. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1481–1490. ACM (2012)
34. Nandhini, B., Sheeba, J.: Cyberbullying detection and classification using information retrieval algorithm. In: Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015), p. 20. ACM (2015)
35. Nandhini, B.S., Sheeba, J.: Online social network bullying detection using intelligence techniques. Procedia Computer Science **45**, 485–492 (2015)
36. Nalini, K., Sheela, L.J.: Classification of tweets using text classifier to detect cyber bullying. In: Satapathy, S.C., Govardhan, A., Raju, K.S., Mandal, J.K. (eds.) Emerging ICT for Bridging the Future - Volume 2. AISC, vol. 338, pp. 637–645. Springer, Heidelberg (2014)
37. Huang, Q., Singh, V.K., Atrey, P.K.: Cyber bullying detection using social and textual analysis. In: Proceedings of the 3rd International Workshop on Socially-Aware Multimedia, pp. 3–6. ACM (2014)
38. Potha, N., Maragoudakis, M.: Cyberbullying detection using time series modeling. In: 2014 IEEE International Conference on Data Mining Workshop (ICDMW), pp. 373–382. IEEE (2014)
39. Kansara, K.B., Shekokar, N.M.: A framework for cyberbullying detection in social network (2015)
40. Hosseinmardi, H., Rafiq, R.I., Li, S., Yang, Z., Han, R., Mishra, S., Lv, Q.: Comparison of common users across Instagram and Ask.fm to better understand cyberbullying. In: The 7th IEEE International Conference on Social Computing and Networking (SocialCom) (2014)

41. Hosseinmardi, H., Ghasemianlangroodi, A., Han, R., Lv, Q., Mishra, S.: Towards understanding cyberbullying behavior in a semi-anonymous social network. In: Advances in Social Networks Analysis and Mining (ASONAM 2014), pp. 244–252 (2014)
42. Weilenmann, A., Hillman, T., Jungselius, B.: Instagram at the museum: communicating the museum experience through social photo sharing. In: Proc. of the SIGCHI Conf. on Human Factors in Computing Systems, CHI 2013, pp. 1843–1852 (2013)
43. Silva, T.H., de Melo, P.O.S.V., Almeida, J.M., Salles, J., Loureiro, A.A.F.: A picture of Instagram is worth more than a thousand words: workload characterization and application. In: DCOSS, pp. 123–132. IEEE (2013)
44. Hu, Y., Manikonda, L., Kambhampati, S.: What we instagram: a first analysis of instagram photo content and user types. In: Proc. of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM 2014) (2014)
45. Araujo, C.S., Correa, L.P.D., da Silva, A.P.C., Prates, R.O., Meira Jr., W.: It is not just a picture: revealing some user practices in instagram. In: 2014 9th Latin American Web Congress (LA-WEB), pp. 19–23. IEEE (2014)
46. NoSwearing.com: (Bad word list and swear filter) (accessed November 10, 2014)
47. von Ahn's Research Group, L.: Negative words list form. Luis von Ahn's Research Group (2014)
48. Pennebaker, J.M., Francis, M.E., Booth, R.J.: Linguistic Inquiry and Word Count. Lawerence Erlbaum Associates, Mahwah (2001)
49. De Choudhury, M., Gamon, M., Counts, S., Horvitz, E.: Predicting depression via social media. In: International AAAI Conference on Weblogs and Social Media, Boston, MA (2013)

Appendix

Labeling Statistics

Overall, 176 potential contributors worked on the quiz questions, 144 passed the quiz mode, while 31 contributors failed and 1 gave up. The labeled data that we finally obtained were from 139 *trusted* contributors, while the the rest were filtered out during the work mode. Table 2 provides the number of trusted judgments and the contributors' accuracy for 11,090 total judgments.

Table 2. Labeling process statistics. Trusted judgments are the ones made by trusted contributors.

Trusted Judgments	10987
Untrusted Judgments	103
Average Test Question Accuracy of Trusted Contributors	89%
Labeled Media Sessions per Hour	6

Image Labeling

Human contributors were given detailed instructions for identifying image's content. We first sampled 1,200 images from the selected subset of media sessions to determine a suitable set of representative categories to be used in the labeling. A graduate student examined all the images and classified them to different possible categories. Then, a social science expert checked the categories again and revised them. Some of the dominant categories identified were the presence of a human in the image, as well as text, clothes, tattoos, sports and celebrities. We then asked contributors to identify which of the aforementioned categories were present in the image. Multiple categories could be selected for a given image. Each media session was labeled by three different contributors. At the end, our social science expert checked a set of random media sessions and images to confirm the quality of the labeled data for both studies.