Scalable and Timely Detection of Cyberbullying in Online Social Networks

by

Rahat Ibn Rafiq

B.Sc., Bangladesh University of Engineering and Technology, 2012

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Computer Science

2018

This thesis entitled:
Scalable and Timely Detection of Cyberbullying in Online Social Networks
written by Rahat Ibn Rafiq
has been approved for the Department of Computer Science

_____

Professor Dr. Shivakant Mishra

_____

Professor Dr. Eric Rozner

Date _____

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

Rafiq, Rahat Ibn (Ph.D., Computer Science)

Scalable and Timely Detection of Cyberbullying in Online Social Networks

Thesis directed by Professor Dr. Shivakant Mishra

The exponential growth of popularity of online social networks in the last decade has unfortunately paved the way for the threat of cyberbullying to rise to an unprecedented level. So a research that provides insights into the analysis of cyberbullying incidents and building a system that is highly scalable and responsive is of unparalleled need. This dissertation gathers insights into cyberbullying incidents in video and image-based social networks (Vine and Instagram respectively) and then presents a system solution that makes use of the gained insights to improve efficiency and efficacy of cyberbullying detection. First, it presents detailed analyses of cyberbullying incidents in the Vine social network by collecting data and labeling them by CrowdFlower. Second, it performs a thorough investigation of the differentiating factors of cyberbullying in online social networks. Third, it implements a highly scalable and responsive system solution for cyberbullying detection along with a comprehensive evaluation of its performances in terms of timeliness and scalability against a highly popular online social network. Fourth, it outlines design, implementation and preliminary user experience analysis of an android application, BullyAlert, that was developed to enable guardians to get adaptive notifications for cyberbullying based on their individual subjective tolerance levels. Finally, it shows that using textual and video feature greatly improves cyberbullying detection classifier's performances.

Dedication

To my wife, Romena Yasmin, whose patience has helped me to see the light even in my darkest of hours and my parents whose unconditional love and support have put me where I am now.

Acknowledgements

At first, I would like to express my heartfelt gratitude to my Ph.D. supervisor Professor Shivakant Mishra for always being there for me when I needed any academic and research direction. His unwavering support has enabled me to complete this thesis.

In addition to my supervisor, I would also like to thank Professor Richard Han and Qin Lv, without whose able advice and counsel, this research would not have been possible. Finally, I would like to thank Dr. Homa Hosseinmardi and Dr. Sabrina Mattson who have been a great source of help during my research.

This work was supported by the US National Science Foundation (NSF) through grant CNS 1528138.

Contents

Chapter

Tables

Table

Figures

Figure

Chapter 1

Overview

1.1 Thesis Statement

The first part of this thesis provides valuable insights into the nature of cyberbullying in Vine, a video-based online social network. It investigates profanity word usage and categories of media shared, labels the media sessions (video and its associated comments) and then develops an accurate classifier model that detects cyberbullying in Vine. The second part explores the differentiating factors of cyberbullying by leveraging labeled data-sets from two social networks, namely Vine and Instagram. The third part introduces the practical challenges of scalability and responsiveness when building a cyberbullying detection system. It tackles those challenges by implementing a solution that consists of two novel components, namely dynamic priority scheduler and incremental feature extraction-classification. Then it demonstrates the improvements that the aforementioned components help to materialize over the current state-of-the-art. The fourth part of the dissertation deals with the challenges encountered when facing the subjective tolerance levels of the guardians for cyberbullying. It develops an android application, BullyAlert which allows guardians to get personalized cyberbullying notifications based on their individual tolerance levels. After that, it delineates a preliminary user-experience analysis of the application. Finally, the thesis goes beyond the basic textual features and explores video-content and topical features and incorporates those to get an improved performance for cyberbullying classification.

1.2 Contributions

This dissertation makes the following contributions.

- Collecting data from Vine, an online video-based social network
- Labeling shared-media and associated comments by CrowdFlower and use the labeled data-set as ground-truth to develop cyberbullying detection classifier
- Analyzing differentiating factors of cyberbullying using labeled data-sets from Vine and Instagram
- Presenting design, implementation and performance evaluation of a scalable and highly responsive cyberbullying detection system
- Outlining design and implementation of an Android application, BullyAlert that allows users to get personalized cyberbullying notifications based on their individual tolerance levels
- Performing a preliminary analysis of user-experience of BullyAlert
- Incorporating topical and video features into cyberbullying detection classifier to achieve better performance

Chapter 2

Introduction

Online Social Networks (OSNs) have seen an exponential growth in recent times. With the advent of the advancements and innovations made in this area, the threats of online predators, stalkers, and cyberbullies have also reached an unprecedented extent. The constant threat of cyberbullying in these multitudes of social networks has become so expansive and pervasive that it has been reported that in America alone, more than fifty percent of teenage OSNs users have been affected by the threat of cyberbullying [23]. While real-life bullying may involve verbal and/or physical assault, cyberbullying is different in the sense that it occurs under the umbrella of an electronic context that is available 24/7, thereby rendering the victims vulnerable to its threats on a constant and relentless basis. This unique feature of cyberbullying subjects the victim to devastating psychological effects that later cause nervous breakdown, low self-esteem, self-harm, clinical depression and in some extreme cases, suicides [45],[34]. Recently there have been some disturbing press reports about some teens committing suicides after being victimized by cyberbullies in OSNs like Facebook [36] and Ask.fm [105]. To make matters worse, nine suicide cases have already been attributed to cyberbullying in Ask.fm alone [11]. Although the causes of these suicides cannot be directly or solely attributed to cyberbullying, it has been reported as one of the potential factors [15]. Figure 4.1 shows an example of a cyberbullying instance in Vine.

Mobile social networks like Vine, Instagram and SnapChat have been hugely popular among teenagers, thus representing a potential target for investigating cyberbullying behavior. The importance of a holistic and elaborate research to develop a methodical and complete understanding

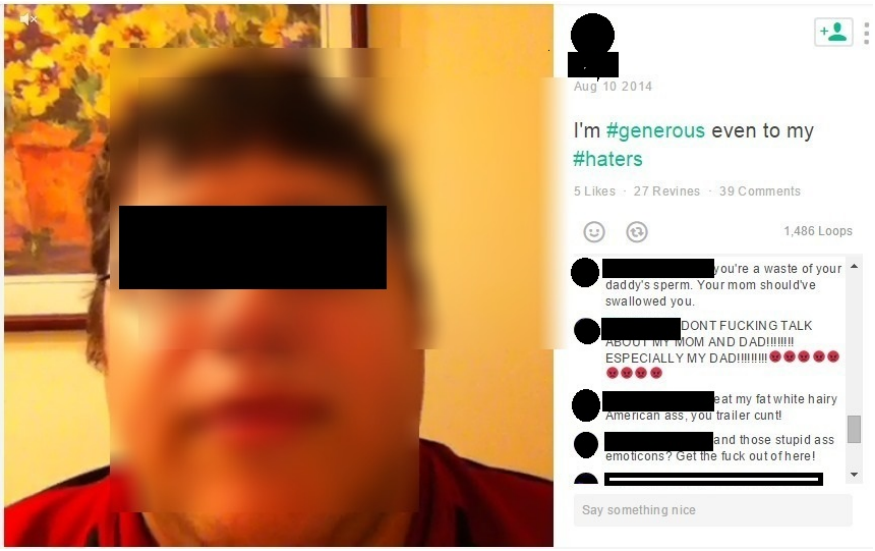


Figure 2.1: An example of cyberbullying on Vine. The image is just a snapshot of the 6-second video.

of cyberbullying behavior in OSNs is significant so as to make sure we can thwart the inadvertent and potentially destructive consequences it may lead the vulnerable victims to. A thorough understanding of cyberbullying behavior can be properly utilized to build an effective and efficient system that can accurately detect potential instances of cyberbullying and take necessary measures to tackle the situation. Vine (purchased by Twitter) in particular is interesting because it offers the opportunity to explore cyberbullying in the context of video-based communication, which has been gaining popularity recently. It is a popular mobile application that enables its registered users to record and edit six-second looping videos, which they can share on their profiles for others to see, like, revine (similar to retweeting) and comment upon. Cyberbullying can happen in Vine in many ways, including posting mean, aggressive and hurtful comments, recording video of others without their knowledge and then sharing the Vines as a way to make fun of or mock them, and playing "the slap game" in which one person records video while another person slaps or hits a person in order to record a reaction. They later share the Vine for the world to see. There are even violent versions called "knock-out" where someone punches an unsuspecting person in an attempt to knock them out [37].

There are four key challenges a potential cyberbullying detection system has to address.

- A clear distinction between cyberaggression and cyberbullying has to be made by performing thorough analyses of the labeled media sessions from a social network. Cyberaggression is defined as a type of behavior in an electronic context that is meant to intentionally harm another person [65]. Cyberbullying is defined in a stronger and more specific way as an aggressive behavior that is *carried out repeatedly* in OSNs against a person who *cannot easily defend himself or herself*, creating a power imbalance [65],[86],[55],[65],[81],[104]. Thus, in order to understand cyberbullying behavior, the factors of repetition of aggression and imbalance of power must be considered.

- Scalability challenges of a potential cyberbullying detection solution. While progress has been made to improve the performance of classifiers for cyberbullying detection [79, 31, 54, 49], scalability (and timeliness, as described in the next point) challenges have largely been ignored. OSNs, of course, involve an enormous amount of data, on the order of several hundred gigabytes per day. For example, it has been reported that for Vine, around 39 million videos have been shared since it was introduced [102] while for Instagram, the amount of shared media is 40 billion[67].

- Timeliness challenge of raising alerts whenever cyberbullying incidents are suspected. Cyberbullying is different from traditional, face-to-face bullying, because it can occur 24/7, and perpetrators can stay anonymous and have easy access to sophisticated tools to launch attacks. The consequences of cyberbullying can be disastrous, which is why it is extremely important to provide the necessary support to the victims as early as possible. So, a timely detection of cyberbullying is a vital necessity.

- Different parents may wish to get different levels of notifications based on their own individual tolerance/preference levels for cyberbullying. So it is imperative to design a system that accommodates the individual tolerance levels of the guardians, thus making the alert

levels personalized and adaptive.

This thesis aims to address these four challenges.

Chapter 3

Related Work

Related works in the area of cyberbullying can be partitioned into four sections. They are: definition of cyberbullying, cyberbullying research in online social networks, cyberbullying detection techniques and systems, applications and tools for detection of cyberbullying in online social networks. Past works in each of these four areas are explored in the following four sections.

3.1 Definition of Cyberbullying

Cyberbullying is defined as an aggressive, intentional act that is carried out by a group or an individual, using electronic/digital/multi-modal forms of contact/messaging/communication, repeatedly against a victim who cannot easily defend him or herself [103, 110]. One huge distinction between traditional bullying and cyberbullying is that the perpetrator of cyberbullying really wants to hurt the feelings of the victim [112]. Intending to hurt the feelings of the victim, imbalance of power and the repetitive nature are the unique traits of cyberbullying. Although cyberbullying is sometimes defined as an electronic form of face-to-face bullying rather than a distinct phenomenon [65], considering cyberbullying as merely the electronic form of face-to-face bullying may overlook intricacies of these behaviors [33], such as repetition of aggression and imbalance of power in an electronic context. Repetition in cyberbullying is problematic to contextualize, as there can be differences between the perpetrator and victim when it comes to the conceptualization of how many incidents occur and their potential consequences. A single aggressive act such as uploading an embarrassing picture to the internet can result in continued and widespread ridicule and humiliation

for the victim. While the aggressive act is not repeated, the damages caused by the act is relived by the victim through an elongated humiliation [33]. Power imbalance in an electronic context can be defined as the perpetrators having superior technological skills [33] or the victim being "shy" or "modest" and the perpetrator knowing the victim in real world[112]. From over-viewing the existing literature, around eight types of cyberbullying behaviors can be recognized [72, 85, 119]. The types are the following:

- **Flooding** involves the bullies sending repeated frequent nonsensical comments/posts in order to not allow the targeted victim to participate in the conversation [72]
- **Masquerade** involves the bullies pretending to mimic or impersonate the target victim [119]
- **Flaming/Bashing** involves an online fight where the bully sends and/or posts insulting, hurtful and vulgar contents to the targeted victim privately or publicly in an online group [119]
- **Trolling** involves purposely publishing comments which disagree with other comments in order to incite arguments or negative emotions although the comments themselves might not be vulgar or hurtful in themselves [75]
- **Harassment** is the kind of conversation where the bullies frequently send insulting and rude messages to the victim [75, 119]
- **Denigration**, also called "dissing", happens when the bullies send or publish gossips or untrue statements about the victims in order to damage the victims' friendships/reputations [75, 119]
- **Outing** occurs when bullies send or publish private or embarrassing information in public chat-rooms or forums. This type of cyberbullying is similar to the denigration. However, in the outing, the relationship between bully and victim is close [75, 119]

- **Exclusion** involves intentionally excluding someone from an online group. This type of cyberbullying happens among youth and teenagers more prominently [75, 85]

3.2 Cyberbullying Research on Online Social Networks

Analysis and detection of cyberbullying/profanity/harassing incidents in several online social networks like Twitter [96], Ask.fm [47], YouTube, FormSpring [30], chat-services [62] have been performed.

Twitter is a text-based social network where a user can update their status by not more than 280 characters [44]. Opinion mining and sentiment analysis techniques have been used to detect cyberbullying in Twitter [96]. A negative word list was leveraged to streamline the tweets that contained those negative words [96]. After that, a sentiment classifier was built with four classes: negative with bullying intentions, negative without bullying intentions, positive or good content and neutral. The labeling of the tweets was performed using the Amazon Mechanical Turk platform which was then employed to build and evaluate the classifier. The reported performance was 67.3%.

The relationship between cyberbullying and anonymity in online social networks have been explored in depth as well [47]. Ask.fm is a semi-anonymous online social network, where the users have the option to hide their identity when posting questions/comments on a profile [6]. In [47], using snowball sampling [10], around $30,000$ profiles were collected. These profiles were then analyzed using interaction graphs, word graphs, frequency distributions and network properties such as reciprocity, clustering coefficient, and the influence of negativity on in-degree and out-degree. It was found that the most vulnerable users were the least active in terms of online social network activity, such as receiving/posting likes.

Researches based on tracking and categorization of internet predators on online chat services have also been performed [62]. 288 chat-logs were collected from PJ [59] website, a project where the volunteers pose as teens and tweens to trap potential sexual predators. Identified categories of the terms and phrases frequently used by the predators were: deceptive trust development,

grooming, isolation, and approach. The idea was to distinguish between predators and victims and to this aim, their developed clustering methods were able to achieve an accuracy of 93%.

Researches have been performed to detect instances of harassment in online social networks and chat services as well. In [126], online social networks were partitioned into two groups: discussion style and chat style. In discussion style environments, there are various threads, usually with multiple posts that populate each of those threads. Users can start a new thread or participate in an existing thread by posting comments. Each thread contains posts that adhere to a predefined topic. On the other hand, in chat style environments, ongoing conversations are more casual and usually, each conversation only consists of a few words with little information. Topical and sentimental features were used to train the supervised classifier to detect harassment after collecting data from Kongregate (chat style) and MySpace(discussion style).

It will be interesting to have further insights into cyberbullying behavior in multi-modal online social networks like Vine and Instagram where users can share videos and images respectively. In comparison to textual cyberbullying, these social networks also provide potential perpetrators to harass the victim though posting harmful images or insulting videos instead of just posting mean comments. Moreover, an in-depth analysis of the correlation between the media contents and cyberbullying behavior can also better our understanding of cyberbullying behavior in online social networks. Finally, delving into the details of cyber-aggression and cyberbullying and investigating the potential distinguishing factors between these two behaviors are also some untapped areas of future research.

3.3 Cyberbullying Detection Techniques

In this section, researches focusing on effective and efficient cyberbullying detection techniques are outlined briefly.

Researches have been proposed based on text mining paradigm for detection issues that are closely related to cyberbullying such as such as online sexual predator recognition [62] and spam detection [108]. Modeling the detection of textual cyberbullying has been a cornerstone of cyber-

bullying research [31] where the problem of cyberbullying detection in Twitter was decomposed into a problem of detecting discussions on sensitive topics, thus rendering the problem into a text classification sub-problem. Three topics were identified as sensitive: sexuality, race/culture, and intelligence. Upon collecting comments pertaining to the aforementioned sensitive topics, the final step was to determine the profanity content of those comments in order to detect cyberbullying. JRip classifier was reported to be the best performing classifier in this technique.

Comparison of different approaches to building effective machine learning classifiers for cyberbullying have also been investigated [27], namely human expert system, supervised machine learning model and a hybrid system combining both the machine learning and the expert system. Labeled data from YouTube was used to evaluate each of these three systems. In the evaluation, it was reported that the expert model outperformed all the machine learning models. The machine learning models' sensitivity to the class skew of the data-set (10% bullying and 90% non-bullying) was attributed to this under-performance. The hybrid approach was reported to have performed better than both the expert model and the machine learning model. Other techniques such as building query terms of phrases and words pertaining to cyberbullying have been developed in the past to detect instances of cyberbullying. In [63], the researchers used labeled data from FormSpring.me and went on to build the most effective query terms for efficient detection of cyberbullying leveraging two models: language and machine learning. It was reported that the terms generated by the machine learning model were the better performing one, yielding both high recall and precision than its language model counterpart.

Initial works in cyberbullying detection techniques have mostly concentrated on the conversations' content though they did not attend to the characteristics of the actors involved in cyberbullying. Social studies demonstrated that men and women bully each other in a different way. For example, women tend to employ aggressive communication styles, such as excluding someone from a group of conspiracy against them whereas men tend to use more words and phrases threatening outrage [21, 75]. In [4, 75], it was shown that pronouns like "I", "you", "she", etc. are used more by females and noun specifiers such as,"a", "the", "that" are used prominently by males.

These findings motivated several cyberbullying researchers to include gender-specific information in cyberbullying detection techniques. Gender-specific information in online social networks has been reported to be useful in improving the performance of a cyberbullying detection system [26].

Graph models in OSNs have also been actively used in cyberbullying research. Researchers in [76] presented a graph model to extract cyberbullying network. This then led to identifying the most active predators and victims through a ranking algorithm. They improved the classification performance by applying a weighted TF-IDF function, in which bullying-like features were scaled by a factor of two. Techniques to detect cyberbullies and cyber-predators have also been proposed in the past[62, 73] . A cyber predator is a person who uses the Internet to hunt for victims to take advantage of them in several ways, including sexually, emotionally, psychologically or financially. Cyber predators know how to manipulate kids, creating trust and friendship where none should exist [51]. Online sexual predator related researches identified communication and text-mining techniques to differentiate predators and victims by analyzing the one-to-one conversations [62, 73]. In [57], the online predator detection problem was partitioned into two sub-problems, namely identifying predators and recognizing predator's conversation techniques/lines for identifying them [75]. Three stages were then proposed: pre-filtering stage, feature extraction stage, and classification stage. For the feature extraction stage, two categories of features were leveraged: lexical and behavioral features [75]. Lexical features were described as those features that could be derived from the raw text of the conversation between the victim and the potential predator, for example, unigrams and bigrams [84, 113], number of emoticons used and the weighted TF-IDF or the cosine similarities. The behavioral features included the number of questions asked, intention(grooming, hooking) to capture the action of the users [75].. For classifying predators, several approaches were investigated by the researchers, namely, decision trees [61], Neural Network [113] and Maximum-Entropy [35].

3.4 Systems, Applications and Tools

Several applications, systems, and tools have been developed in recent years to detect potential cyberbullying instances. Some of these applications were developed for the guardians to

help them protect their youth from the adverse consequences of cyberbullying in online social networks. Other proposed tools work independently of the guardians and try to contain the impacts of cyberbullying by negating the negative/profane comments on the victim's online social network profile.

The youth monitoring applications include Net Nanny, eBlaster and IamBigBrother [120, 121, 52]. These applications make use of packet sniffer, which examines all the outgoing and incoming traffic in a network and then apply a filter to only see the useful blocks of data [109]. There are several problems regarding this approach, for example, they are too intrusive and require the guardians to gloss over many trivial data. Moreover, these tools are based on simple keyword detection and not a sophisticated classification algorithm [109].

Applications have also been proposed to support cyberbullying victims in online social networks. One such tool, iAnon [5] automatically detects Ask.fm users who are at risk and allows third-party "do-gooders" to anonymously send friendly encouraging messages to the victims. The idea is to help mitigate feelings of depression and loneliness that are often felt by victims of cyberbullying by providing an online support system. In order to detect the vulnerable users for depression and anxiety due to cyberbullying, the authors first detect individual posts that have traces of cyberbullying in them. If the ratio of bullied posts over the number of non-bullying posts exceeds a certain threshold, the authors classify the users as being at risk and thus making him/her a candidate for the iAnon.

Because the online social networks have thousands of users with millions of comments, images and videos pouring in every hour, it is imperative for a cyberbullying detection system to take into account the scalability issues. Scalability has been a major aspect of research in building real-time systems for different services, for example, malware detection services [98]. Scalability is a major area of research in flasher detection in online video chat services as well [122] where the researchers developed a real-time misbehavior detection system for Chatroulette, an online video chat service.

Although some applications have been developed for cyberbullying, these are more or less based on word-based detection of profanity. Design and implementation of a scalable and responsive

cyberbullying detection system that alerts the guardians when a potential cyberbullying instance takes place in a victim's profile is still an unexplored research area.

Chapter 4

Detection of Cyberbullying Instances in Vine, a video-based Social Network

Mobile social networks like Instagram, Vine, and SnapChat are booming in popularity, spurred by the revolution in smart-phone usage, and therefore represent a natural target for investigating cyberbullying. Vine (purchased by Twitter) in particular is interesting because it offers the opportunity to explore cyberbullying in the context of video-based communication, which has been gaining popularity recently. Vine is a mobile application that allows users to record and edit six-second looping videos, which they can share on their profiles for others to see, like and comment upon. Cyberbullying can happen in Vine in many ways, including posting mean, aggressive and hurtful comments, recording video of others without their knowledge and then sharing the Vines as a way to make fun of or mock them, and playing "the slap game" in which one person records video while another person slaps or hits a person in order to record a reaction. They later share the Vine for the world to see. There are even violent versions called "knock-out" where someone punches an unsuspecting person in an attempt to knock them out [37]. Figure 4.1 provides an illustration where the profile owner is victimized by hurtful and aggressive comments posted by others.

In the following research, a distinction between cyberaggression and cyberbullying is made. Cyberaggression is defined as a type of behavior in an electronic context that is meant to intentionally harm another person [65]. Cyberbullying is defined, in a stricter and more specific way, as an aggressive behavior that is *carried out repeatedly* in OSNs against a person who *cannot easily defend himself or herself*, creating a power imbalance [65, 86]. Thus in order to understand cyberbullying, the factors of repetition of aggression and imbalance of power must be taken into account.

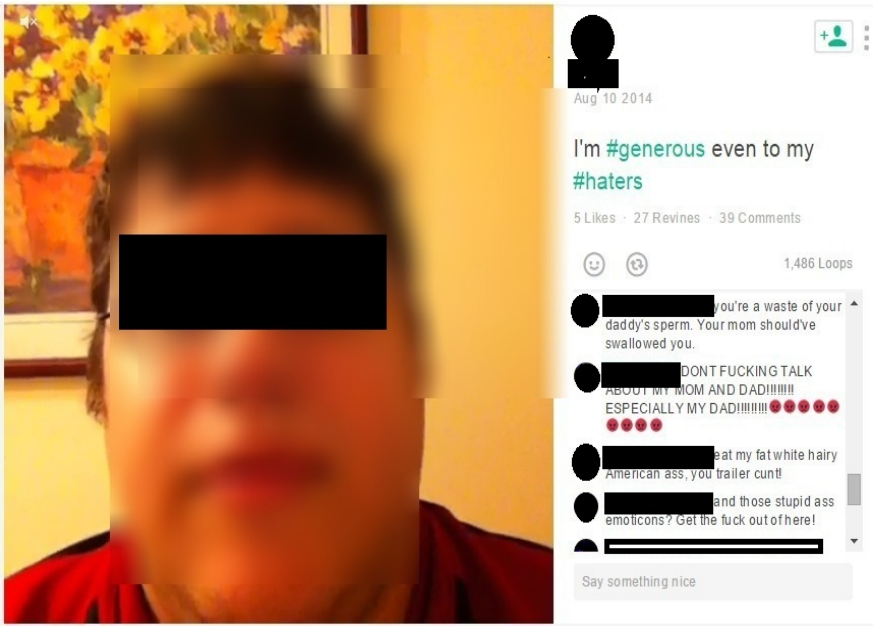


Figure 4.1: An example of cyberbullying on Vine. The image is just a snapshot of the 6-second video.

This chapter makes the following contributions:

- It investigates cyberbullying behavior in Vine, a video-based mobile social network by labeling the videos along with the comments associated with them according to the appropriate definition of cyberaggression and cyberbullying.
- It presents a thorough analysis of the labeled videos, the associated comments, different features, and meta-data of the media-sessions and the relationship between these features and both cyberaggression and cyberbullying.
- It presents an elaborate development and evaluation of classifiers to effectively identify instances of cyberbullying based on the labeled data and all the features associated with the videos and comments.

4.1 Data Collection and Labeling Methodology

The following subsections briefly describe the data collection from Vine and the labeling methodology used to label cyberbullying instances.

4.1.1 Data Collection

To collect data from Vine, we applied the snowball sampling method [10] in which we selected one random user u_s as a seed and then collected all the users that u_s is following. We then repeated this process for each new user u_i, i.e., collecting all users followed by u_i. The reason that we traversed the following link instead of the follower link is that in social networks like Vine, there are some well-known celebrities and popular users who tend to have a lot of followers, whereas it is relatively rare to come across a user who is following a large number of users. Thus, to keep the number of users in the network manageable, we traced the following network. By applying the aforementioned policy, we collected Vine information for $59,560$ users. For each user, we collected the user-id and profile information such as user-name, full name, location (if any), profile description, number of videos posted by that user and the post-ids, the number of followers who follow that user and their user-ids and the number of users that the user is following and their user-ids. After collecting all the videos posted by these users, we collected all the comments associated with the videos, user-ids of the users who commented on that video, total number of likes and user-ids who liked that video, number of times that video has been viewed and the number of times it was re-posted or shared by some other users. We refer to each posted video along with all the likes and comments associated with it a *media session*. In total, about 652K media sessions were collected.

After collecting the media sessions, we selected those media sessions that have at least 15 associated comments. Reasons for this particular filtering are twofold. Firstly, As it can be seen from Figure 4.2, the percentage of posts in Vine having less than 15 comments is quite low. Secondly and most importantly, our ultimate goal was to detect cyberbullying in the media sessions, and in order to identify cyberbullying in a media session, we needed a sufficient number of comments so

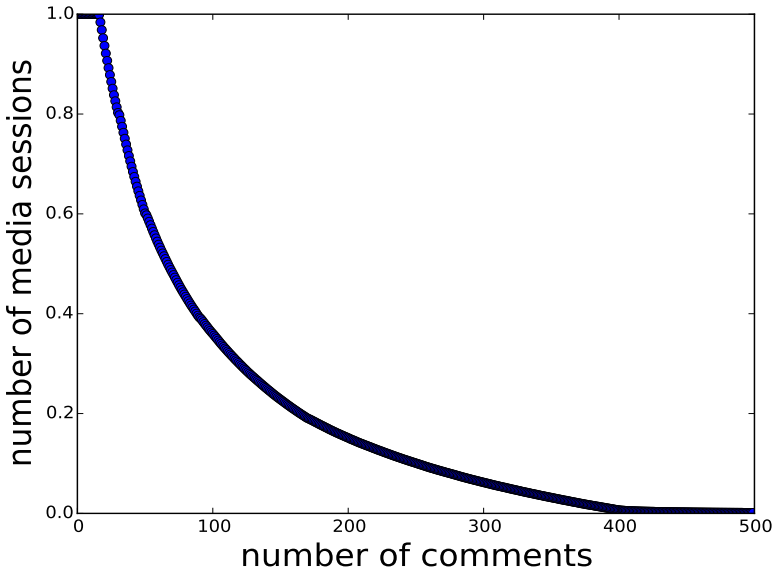


Figure 4.2: CCDF Distribution of media sessions' comments in Vine

that our labelers could make a contextual assessment of the frequency/repetition of aggression that fits the definition of cyberbullying [65, 86].

This filtering gave us 436K media sessions. We computed the profanity of each one of these media sessions. For this purpose, we followed the profanity word dictionary provided in [114]. We considered a comment in a media session profane if that comment had at least one profane word in it. We acknowledge the fact that cyberbullying can also take place where profane words are not used but we felt that detection of profanity word usage would give us good insights into an important form of cyberbullying occurring in media sessions.

Figure 4.3 shows the complementary cumulative distribution function (CCDF) [115] of the percentage of profanity for our media sessions. We called a media session x percent profane if x percent of the comments associated with that media session had at least one profane word in it. The figure shows that most of the media sessions have less than 25 percent profanity. The fraction of media sessions with more than 40 percent profanity was fairly low. **A key finding of this profanity analysis of media sessions is that in Vine, the percentage of high profanity-containing media sessions is quite low**.

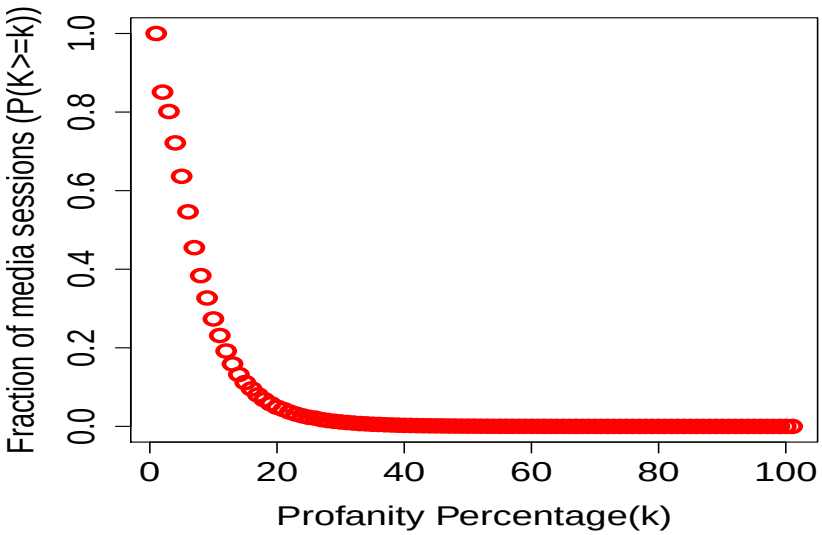


Figure 4.3: CCDF of profanity percentage and fraction of media sessions.

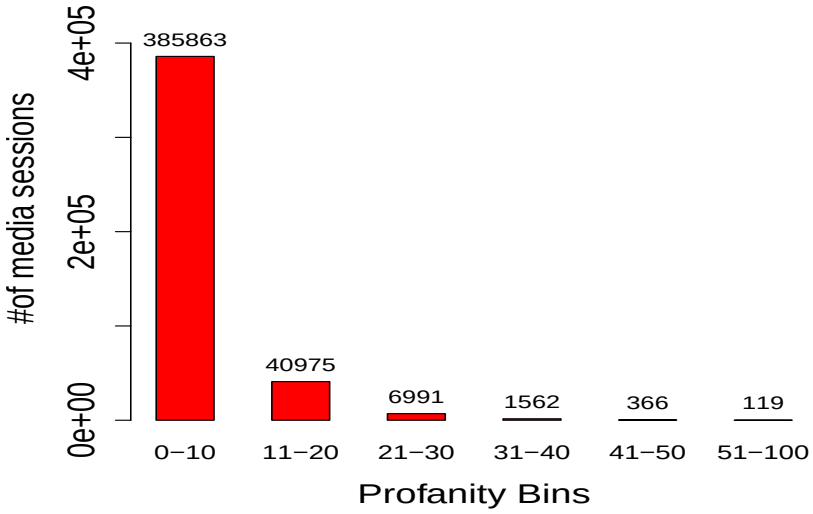


Figure 4.4: Distribution of media sessions with different percent of comments containing profanity.

Our next step was to collect a sub-sample from these media sessions so that we could conduct our labeling survey. For this purpose, we created 6 bins where each bin represented a range of % of comments with profanity. The ranges we selected are $0 \sim 10\%$, $11 \sim 20\%$, $21 \sim 30\%$, $31 \sim 40\%$,

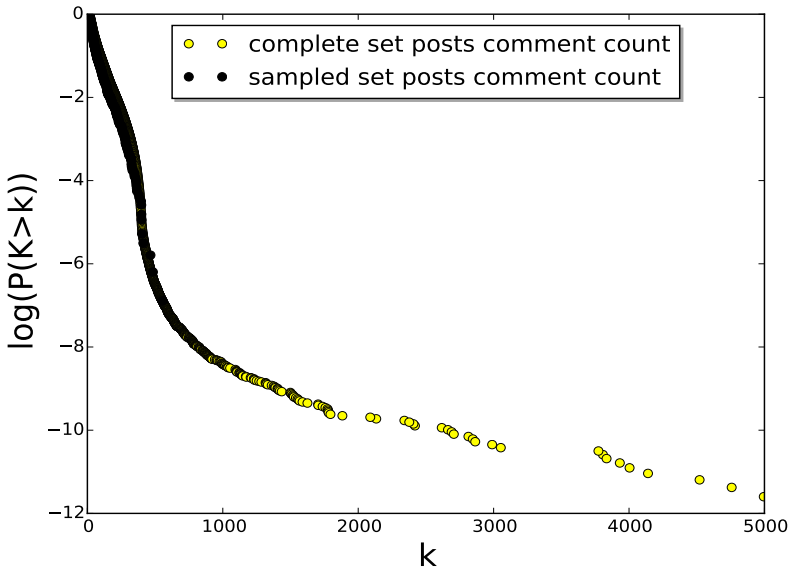


Figure 4.5: Complementary Cumulative Distribution Function (CCDF)of the number comments for both the sampled and complete set of media sessions

$41 \sim 50\%$ and lastly $51 \sim 100\%$.

Figure 4.4 shows the distribution of media sessions associated with each of these bins. After that, we randomly sampled 170 media sessions from each of the first 5 bins and 119 media sessions from the last bin, as it had only that many media sessions. That gave us in total 969 media sessions, each belonging to a distinct user, providing a broad distribution of media sessions with differing profanity for our labelers.

After sampling, we compared the post comments associated with the 969 sampled media sessions with the complete set of 435876 media sessions. Figure 4.5 shows the CCDF of the number of comments received in the complete set of media sessions and the sampled set of media sessions. It can be seen that both the sampled and the complete set follow the same distribution until the point where the number of comments received is around 500. After that point, the complete set of media sessions show a long tail. We hypothesize that the reasons for this phenomenon are twofold. Firstly, there are some popular users in Vine who tend to have a lot of followers and therefore a lot of comments on their media sessions. Secondly, Vine supports Revining, which allows a certain

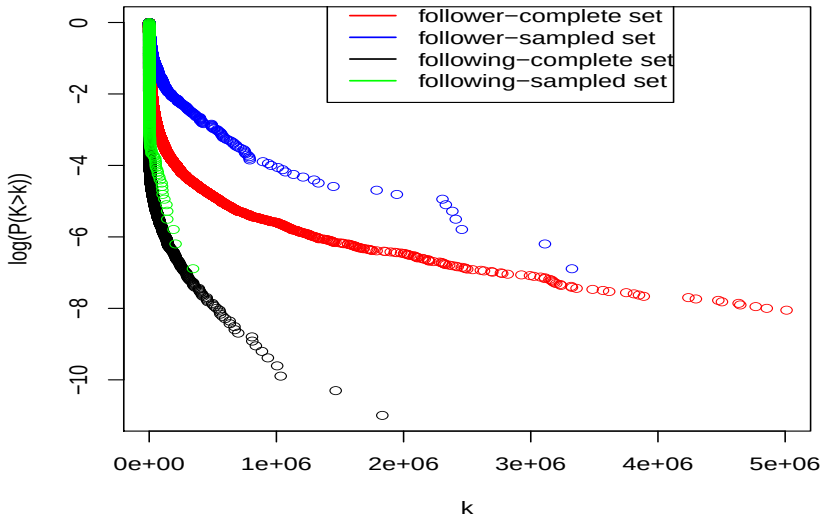


Figure 4.6: Complementary Cumulative Distribution Function (CCDF)of the number of followers and followings for both the sampled and complete set of users

user to repost a video from someone's profile in his/her own profile. In this particular case, all the comments the user receives in his profile are actually associated with the original video that was posted in the original user's profile. So the more a user's video gets revined, the more comments will be associated with that media session. So we think the popular user's media sessions that were revined a lot of times by others might have contributed for the long tail.

In addition to this, we compared the number of followers and followings for the complete set of 59560 users with the distinct 969 users whose media sessions have been sampled. Figure 4.6 shows the CCDF of the number of followers and followings for both the sampled and complete set of users. It can be seen that both the sampled set and the complete set of users show the same distribution for the number of followings. But when it comes to the number of followers, the complete set of users has a longer tail compared to the sampled set. We attribute this happening to the presence of a considerable number of popular users like celebrities, artists, band profiles etc. in the complete set of users. We can also see that the distribution of followers for the sampled users falls slowly compared to the complete set of users. This is because the sampled set of users

was collected after we collected and sampled the media sessions with each media session falling into a different bin. Because we have bins where we have media sessions with 40 or 60 percent negative comments, it is really hard to find a media session from a popular user who has thousands of comments with a considerable amount of negativity. So the sampled users are much less likely to be popular compared to the complete set of users, which is why their distribution of the number of followers falls more slowly.

4.1.2 Labeling Methodology

In this section, we outline the way we designed our labeling survey for the set of media sessions we sampled from the complete set of media sessions as described in Section 4.1.1. While designing the survey, our first goal was to choose the appropriate definitions of cyberbullying and cyberaggression. In order to understand cyberaggression and cyberbullying in Vine, we designed our survey to incorporate both the video shared and its associated comments so that the human labelers could make an informed and contextual decision when participating in the survey. Figure 4.7 depicts an example of an instance of a media session in our survey. The video is on the left while a scrollable interface contains all the comments associated with that shared video along with the user-names who commented to help the participants decide whether the aggressiveness is repetitive. With the help of an expert in Behavioral Science, we decided to ask the labelers two questions, whether the media session is an instance of cyberaggression or not and whether the media session is an instance of cyberbullying or not. Prior to labeling, participants were given the definitions and distinctions between cyberbullying and cyberaggression along with related examples. Each media session was labeled by five contributors.

4.1.3 Quality Control

Because we used CrowdFlower, a crowd-sourcing website, we had to make sure the participants were of the highest quality. First, to make sure that the prospective participants were elaborately trained prior to the participation, they were given clear instructions explaining them

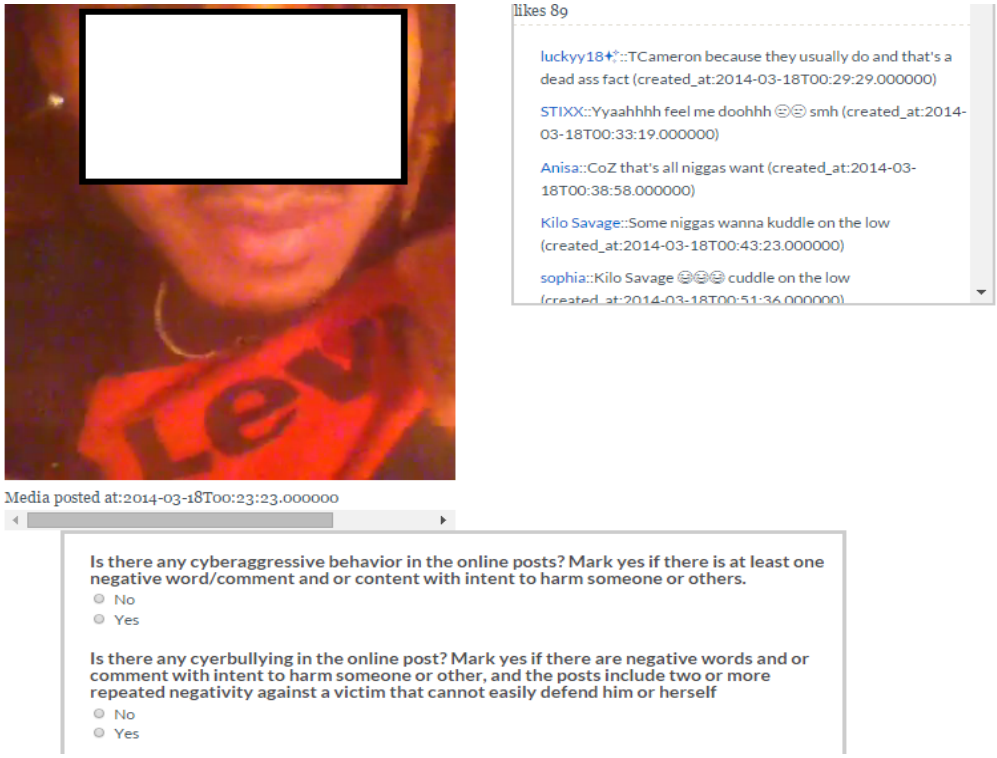


Figure 4.7: An example of cyberbullying labeling. The labeler would be shown the 6-second video, though here we can only show a snapshot of the video. The comments associated with the media are on the right in a scrollable interface.

Table 4.1: Survey Statistics

	Results
Trusted Judgments	4795
Untrusted Judgments	156
Average Test Question Accuracy-trusted	86%
Average Test Question Accuracy-untrusted	44%
Average Test Question Accuracy-all	69%
Total Contributors	106
Agreement on Cyberaggression question	76.6%
Agreement on Cyberbullying question	79.49%

the distinctions between cyberaggression and cyberbullying along with answers to some example set of media sessions. After that, to filter out users with questionable quality, the potential labelers were asked to answer a set of test questions. The labelers needed to answer a minimum number of test questions to be qualified to participate in the survey.

In addition to using the test questions, random test questions were asked in the middle of the actual survey to monitor the quality of the survey. To ensure that the users did not just rush through the job, a minimum threshold amount of time was also set to filter out labelers who hurried through the job because we assumed that at least a minimum amount of time was required to carefully peruse the comments associated with the media session and give contextually knowledgeable answers to the questions asked in the survey. **It is worth mentioning that the demographics of the labelers such as age, gender, were not taken into account during the whole process. We acknowledge that labeling cyberbullying is a subjective matter that might depend on the demographics of the labelers. We leave exploring the relationship between the demographics of the participants and their labeling decisions as a future work.**

4.2 Analysis of Cyberbullying Labeling

Each of the sampled media sessions were submitted to CrowdFlower for labeling of cyberaggression and cyberbullying by five different participants. The incentive for the survey was money. Table 4.1 shows the statistics of the survey. A judgment was considered trusted if the trust score

was at least 0.8, which was computed by CrowdFlower by incorporating the contributor's performance in answering the test questions and his/her overall trust score in CrowdFlower [24], thus giving us in total 4795 trusted judgments for 959 media sessions with 10 test questions. Average test question accuracy percentages for the trusted, untrusted and all contributors were 86%,44% and 69% respectively. The contributors showed 76.6% and 79.49% agreement for the two questions, namely whether the media session constituted cyberaggression or not and whether the media session constituted cyberbullying or not.

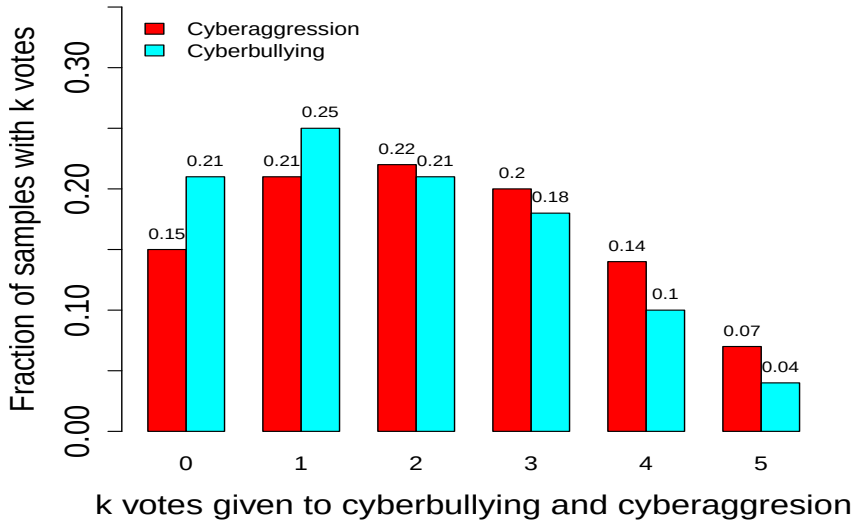


Figure 4.8: Fraction of media sessions that have been voted k times as cyberaggression and cyberbullying.

During the survey, CrowdFlower assigned a degree of trust [24] to each labeler that was computed from the percentage of correctly answered test questions. This was then incorporated with the majority voting method [116] to assign a confidence level to each survey question's answer [25]. We took into account this weighted confidence level given by CrowdFlower to decide whether a label was dependable or not. By taking the answers with a confidence level of 50 percent or more, we show in Figure 4.8 the distribution of the labeled answers for the questions asked about cyberaggression and cyberbullying. A higher number of votes for a particular question for a given

media session means higher trust and confidence level for the given answer. Five votes for a question mean an agreement that is unanimous. Figure 4.8 shows the percentage of media sessions that has been voted k times as cyberaggression and cyberbullying respectively. As it can be seen from the figure, most of the probability mass is around $0, 1, 2$ number of votes for both cyberaggression and cyberbullying. Also, it is seen that only 0.21 and 0.14 fraction of the sampled posts have received 4 or more votes for cyberbullying and cyberaggression respectively, which shows that labeling cyberaggression and cyberbullying is less unanimous than for Instagram [49]. Further investigation is needed to identify whether the motion/looping videos exhibited in Vine media sessions are a contributing factor for this lack of unanimity among labelers.

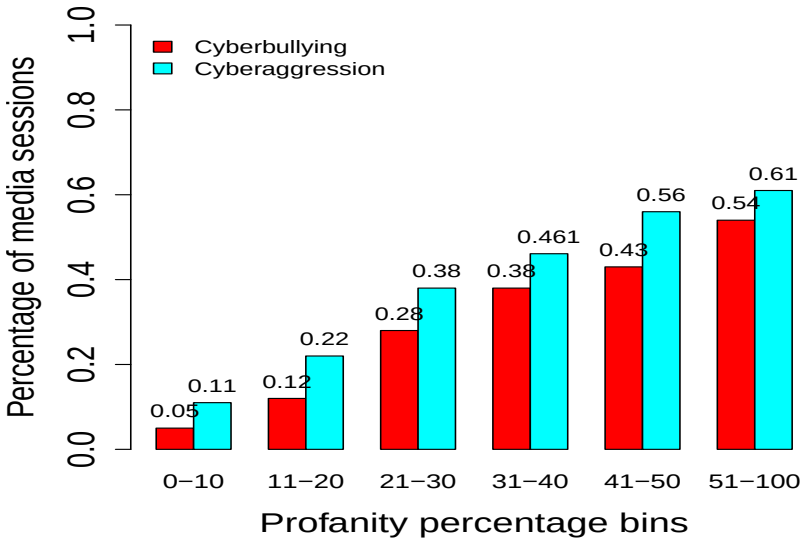


Figure 4.9: Percentage of posts labeled as instances of cyberbullying and cyberaggression for each profanity percentage bins

Next, we show in Figure 4.9 the percentage of the media sessions labeled as cyberbullying and cyberaggression for each profanity bins. The figure clearly shows a pattern of increasing instances of cyberaggression and cyberbullying as the profanity percentage in the media session increases. However, out of media sessions with more than 50 percent profanity, only 54 and 61 percent of media sessions have been labeled as cyberbullying and cyberaggression respectively. This strongly

suggests that we cannot simply employ the percentage of profanity in a media session as the primary indicator of cyberaggression or cyberbullying. Our classifier will need to be more sophisticated. **As a result, we were able to claim that profanity in a Vine media session can be one of the many indicators of cyberbullying but not the only one**.

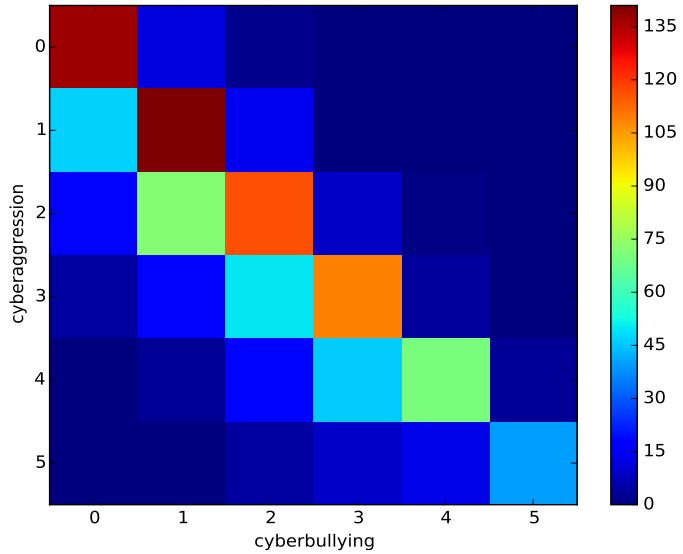


Figure 4.10: Two dimensional distribution of number of media sessions as a function of the number of votes given for cyberaggression versus the number of votes given for cyberbullying, assuming five labelers.

Figure 4.10 shows a two-dimensional heatmap investigating the distribution of media sessions as a function of the number of votes each media session received for cyberaggression and cyberbullying. We plot this heatmap to understand the relationship between labeled cyberaggression and labeled cyberbullying media sessions. From the figure, we see that a significant portion of media sessions lies along the diagonal, which shows strong agreement between cyberaggression and cyberbullying receiving the same number of votes from the labelers. This is expected as we know cyberbullying is one form of cyberaggression so if there is an instance of cyberbullying in a media session, it is also likely that the media session also exhibits cyberaggression. The strength of energy along the diagonal slowly decreases along the diagonal as we move from low (0) to high number

of votes (5) which means strong agreement for the media sessions in terms of receiving as low as 0 or 1 votes but not as much for votes as high as 5 votes. We hypothesize that this is because determining whether a media session has cyberaggression was pretty straightforward. Thus, when a media session had no cyberaggression it was most likely that the media session did not exhibit cyberbullying too, which is why the top left the portion of the diagonal shows such strong energy. On the contrary, determining whether a media session exhibited cyberbullying was not as straightforward as cyberaggression because the labelers had to take into account the imbalance of power and repetitions of aggression. This is why when a media session shows a good amount of cyberaggression and thus receiving a high number $(4, 5)$ of votes for it, there is not as much agreement for cyberbullying.

The area below the diagonal also shows a fair amount of energy, which is for the media sessions that have more cyberaggression votes than cyberbullying votes. This means there are a good portion of media sessions (300 out of 969) that received more votes for cyberaggression than cyberbullying. If we look more closely, we observe that, of the media sessions that received as few as 0 or 1 votes for cyberbullying, a good portion of them (162) received as high as 2,3 or 4 votes for cyberaggression. **This analysis enabled us to claim that in Vine, not all media sessions that exhibit cyberaggression are instances of cyberbullying**.

We also observe a small fraction of media sessions(45) that lies just above the diagonal, which means some labelers have labeled a media session as cyberbullying but not cyberaggression. When we investigated these labeled data, we saw that the confidence scores for the cyberbullying questions of these media sessions were almost close to 50%. The way CrowdFlower assigns this confidence score allows one question to have an answer for a media session as, for example, bullying, to have a confidence score of more than 50% even if only two out of the five labelers tagged that question as bullying. This happens when the trust scores of those two labelers are far greater than the other two labelers. Surely enough, when we took a threshold of 60% confidence level to make sure at least three labelers agree on an answer, only 10 of such media sessions prevailed. Moreover, after performing a careful examination of these 10 media sessions, it was seen that those media

sessions lacked any profanity but seemed more like prolonged arguments that contained lots of barbed sarcastic comments among only two or three people. We suspect this collection of lengthy arguments lacking profanity and containing thinly veiled sarcasm made the contributors label those as cyberbullying but not as acts of aggression.

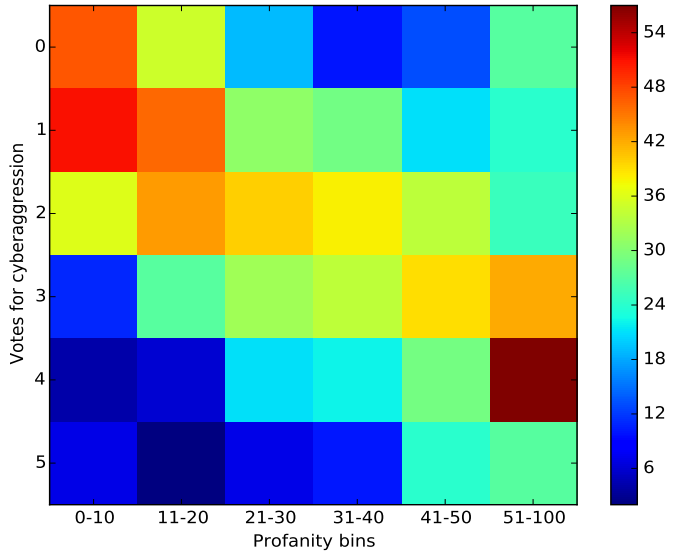


Figure 4.11: Two dimensional distribution of number of media sessions as a function of the number of votes given for cyberaggression for different profanity bins, assuming five labelers.

To further understand the loose relationship between profanity and both cyberaggression and cyberbullying, we plot two heatmaps in Figure 4.11 and 4.12. From the two heatmaps, it can be seen that a significant number of media sessions with a very high percentage of profane comments received as low as 0 or 1 votes for both cyberaggression and cyberbullying. This again clearly shows that just profanity word usage alone in the comments of a media session cannot be the only indicator of whether a media session is an instance of cyberaggression or cyberbullying. For example, we observed many users who employ profanity words as a show of affection. However, there is still a trend in which the main energy/mass for media sessions with low profanity percentages is concentrated among low numbers of votes for cyberaggression and cyberbullying, while media sessions with higher profanity percentages concentrate their mass around higher numbers of votes

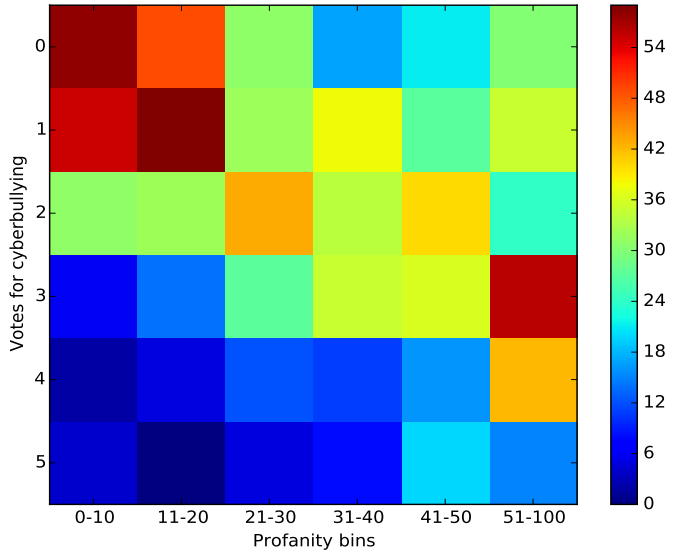


Figure 4.12: Two dimensional distribution of number of media sessions as a function of the number of votes given for cyberbullying for different profanity bins, assuming five labelers.

for cyberaggression and cyberbullying. **This shows that although profanity usage cannot be the only indicator, it has the potential to be one of the indicators to identify instances of media sessions in Vine that exhibit cyberaggression and cyberbullying.**

4.3 Classifier Performance

Based on the labeled data from CrowdFlower, we proceeded to design and evaluate classifiers that could detect cyberbullying behavior in Vine. In this section, we sketch the approaches we undertook in developing the classifier. The following subsections are organized as follows: subsection 4.3.1 describes the features we considered to develop our classifier, subsection 4.3.2 discusses the pre-processing techniques we used prior to designing the classifier and finally subsection 4.3.3 investigates different classifiers' performances with the features considered.

Table 4.2: Features Considered

profile owner features	number of followers, number of followings, user description polarity, user description subjectivity
media session features	number of likes, number of comments, number of revines, media caption polarity, media caption subjectivity
comment features	percentage of negative comments, average number of profane words per comment, average negative comment polarity, average negative comment subjectivity, average profile-owner comment polarity, average profile owner comment subjectivity, average other comment polarity, average other comment subjectivity

4.3.1 Feature Description

To design the classifier, we considered, in total, three categories of features: profile owner features, media-session features and comment features. To extract sentiment information, we use python sentiment library [71]. The library gives as output polarity and subjectivity value of a particular text. Texts have a polarity (negative/positive, -1.0 to +1.0) and a subjectivity (objective/subjective, +0.0 to +1.0) showing how negative and subjective a particular text is. The library is reported to have an accuracy of 75 percent [71] when applied to an English movie review data-set [83]. This convinced us to use this library to extract sentiments from the texts when designing sentiment features. The features considered for the aforementioned three types are shown in Table 4.2. The profile owner features include features belonging to the particular user information, for example, number of followers and followings and so on. The media session features contain features that belong to a particular media session shared by a user, for example, number of likes and comments for that media session. Comment features include textual and sentiment features extracted from the set of comments belonging to a particular media session.

4.3.2 Pre-processing

Before extracting the features from the labeled media session data-set, we employed several pre-processing techniques to the texts, namely removing white spaces, unrecognized characters, punctuation and making the text lower case. We tagged a particular comment as negative when

Table 4.3: Brief Description of Features Used

		Description	Range
Profile owner features	number of followers	total number of users following this user	0-1
	number of followings	total number of users this user follows	0-1
	user description polarity	polarity of the user description text on the profile	0-1
	user description subjectivity	subjectivity of the user description text on the profile	0-1
Media Session Features	number of likes	number of likes for this media	0-1
	number of comments	number of comments for this media	0-1
	number of revines	number of revines for this media	0-1
	media caption polarity	polarity of the media caption	0-1
	media caption subjectivity	subjectivity of the media caption	0-1
Comment Features	Percentage of negative comments	percentage of comments with at least one negative word in them using [114]	0-1
	average number of profane words per negative comment	ratio of total profane words in the comments and total number of negative comments	0-1
	average negative comment polarity	average polarity of the negative comments in the media session	0-1
	average negative comment subjectivity	average subjectivity of the negative comments in the media session	0-1
	average profile owner negative comment polarity	average polarity of the negative comments posted by the profile owner	0-1
	average profile owner negative comment subjectivity	average subjectivity of the negative comments posted by the profile owner	0-1
	average other negative comment polarity	average polarity of the negative comments posted by others	0-1
	average other negative comment subjectivity	average subjectivity of the negative comments posted by others	0-1

there was at least one negative word in it according to the negative word dictionary [114]. To extract sentiments, we used the python library as described in section 4.3.1. We did an average of the sentiment polarity and sentiment subjectivity of all the negative comments, negative comments belonging to the profile-owner and others individually. This was done with the intuition that in a cyberbullying media session, the profile owner is prone to react to the negative comments posted by others by more negativity, for example, anger or sadness. The same approach for extracting sentiments was done for the captions belonging to the media and texts belonging to the user description text.

After extracting the features, min-max normalization was applied to the feature vector to fit the values of the features into a range from 0 to 1. The ranges for the features after the min-max normalization process along with a brief description of the features are shown in Table 4.3. We applied L1 Regularization for feature selection [80]. Others features were considered as well, for instance, media loop count, average polarity/subjectivity of the comments for a media session. We only present the features, techniques, and approaches that gave us the best performing classifier in terms of accuracy, precision, and recall, as described in section 4.3.3.

4.3.3 Classifier Investigation

Based on the labeled data from CrowdFlower, we proceeded to design and evaluate classifiers that could detect cyberbullying behavior in Vine using the features described in section 4.3.1 and 4.3.2. During the survey, CrowdFlower assigned a degree of trust to each labeler that is computed from the percentage of correctly answered test questions. This degree of trust was then incorporated with the majority voting method to assign a confidence level to each survey question's answer. We take into account this weighted confidence level given by CrowdFlower to design our classifier. By taking the labeled media sessions with at least 60% confidence to make sure we had at least 3 out of 5 people agreeing on the labeling, we saw that about 31% of the media sessions were labeled as cyberbullying, which created an unbalanced data set. To make the data-set balanced, we applied Synthetic Minority Over-sampling Technique (SMOTE) [18] and

used 10-fold cross validation [117] to evaluate the performances of the classifiers. Several classifiers were employed namely AdaBoost, DecisionTree, Random Forest, Extra Tree classifier, SVM Linear, SVM Polynomial, SVM RBF (radial basis function), SVM Sigmoid, k-NN, Naive Bayes, Neural network classifiers like Perceptron, Ridge classifier and Logistic Regression. When investigating the classifiers' performances, we used several combinations of the three types of features that gave the best performances in terms of accuracy, precision and recall (by applying L1 Regularization [58]). In addition to the accuracy, we also considered precision and recall to reduce the false positives and negatives. Only those feature combinations were considered that helped the classifiers to attain the maximum accuracy, precision, and recall.

Table 4.4: Different classifier's accuracy percentage performance using media, user and comment features

		Metrics				
		Accuracy	Precision	Recall	bullying Precision	bullying Recall
User	k-NN	56	56	56	56	53
	AdaBoost	56	56	56	55	52
	RandomForest	**70**	**65**	**64**	**67**	**63**
	ExtraTree	67	70	67	70	60
Media	LogisticRegression	60	60	60	60	55
	AdaBoost	64	65	64	65	59
	ExtraTree	**72**	**74**	**72**	**79**	**60**
Comment	LogisticRegression	72	78	76	79	76
	AdaBoost	**76**	**80**	**72**	**81**	**74**
	RandomForest	75	79	76	74	70
	SVM RBF	70	79	70	79	70
	SVMLinear	71	80	71	80	69

Table 4.4 shows the best performing classifiers' performance when using the profile owner, media session and comment features. In addition to the accuracy, precision and recall metrics, we also considered two other metrics namely cyberbullying precision and cyberbullying recall that illustrate the precision and recall performance of the classifiers for the cyberbullying class. The reasons for including these two additional metrics are twofold [58]. First, data-set that we used to train and evaluate our classifier was imbalanced. Sometimes high accuracy in imbalanced data-sets

can be misleading because high performance in the majority class can also lead to overall high accuracy [58]. Second, in this problem setting, we wanted to make sure the penalty for missing the minority class, that is cyberbullying class, is more. Only the results of the classifiers that yielded the best results across these five evaluation metrics are presented in the table. We applied several combinations of the first three types of features namely profile owner features, media session features, and comment features and found that just by using comment features, AdaBoost classifier gave an accuracy, precision, recall, cyberbullying precision, cyberbullying recall of 76,80,72,81 and 74 respectively. , we created two sets, namely sets with media sessions with less and more than 30 percent profanity in the comments respectively to further examine our best performing classifier's performance. We found that the precision and recall scores of AdaBoost were 63 and 91 percent for the media sessions with less than 30 percent profanity. For media sessions with more than 30 percent profanity, the precision and recall scores were 77 and 76 percent respectively.

4.4 Conclusions

This chapter makes the following contributions. To our knowledge, this is the first research paper to conduct a detailed investigation of cyberbullying in the context of a video-based mobile social network, namely Vine. An appropriate definition of cyberbullying was given that differentiated itself from cyberaggression by including repetition of aggression and imbalance of power in an electronic context. Then, that definition was incorporated in labeling the media sessions of Vine. Next, detailed analyses of the labeled media sessions were performed. Finally, using the labeled media sessions and features derived from user, media session, comment meta data, different classifiers' performances are presented across different performance metrics.

The key findings from this research are as follows. First, we found that the percentage of high profanity-containing media sessions in Vine is quite low. Second, we discovered that a significant fraction of the high profanity-containing media sessions was not labeled as cyberbullying, though in general there was a trend towards increasing identifications of cyberbullying as the percentage of profanity increased. This suggested that the percentage of profanity in media sessions should

not be used as the sole indicator of cyberbullying, but should be supplemented by other input features to the classifier. Third, we found that not all media sessions that exhibit cyberaggression are instances of cyberbullying, validating the need to apply a stricter definition of cyberbullying. Fourth, we demonstrated that AdaBoost achieved the best classification performance, using a combination of profile owner, media session, comment features and unigrams.

4.5 Acknowledgments

I would like to thank Homa Hosseinmardi and Sabrina Arredondo Mattson for their contributions towards the completion of this research. This work was supported by the US National Science Foundation (NSF) through grant CNS 1528138.

Chapter 5

Identifying Differentiating Factors for Cyberbullying in Vine and Instagram

The past decade has seen an unprecedented growth of Online Social Networks (OSNs). Unfortunately, this rise has also paved the way for online predators, stalkers and cyberbullying to wreak havoc on the psyches of potential victims. The threats of cyberbullying in these social networks are constant, pervasive and expansive and have led to some very serious consequences. It has been reported that in the United States alone, more than fifty percent of teenage OSN users have been affected by the threat of cyberbullying [23]. Cyberbullying has the potential to be more damaging than real-life bullying since it follows children and teens outside of their schools, e.g., even in their homes where they were safe earlier. The factors of availability and relentlessness make cyberbullying a very serious threat to the potentially vulnerable victims. The constant threat of cyberbullying in online social networks has led to devastating psychological effects in victims such as nervous breakdowns, low self-esteem, self-harm, clinical depression and in some extreme cases, suicides [45, 34]. In recent years, the pervasiveness and availability of social networks for cyberbullying have resulted in the suicide of numerous teens [36, 105, 11]. Therefore, a holistic and elaborate research to identify the differentiating factors for cyberbullying is of paramount importance.

In this chapter, we focus on the analysis of cyberbullying on Instagram and Vine, which are especially popular with the current youth. Cyberbullying in Instagram can happen in different ways, including sharing a humiliating/insulting/edited image of a victim, posting mean and hateful comments on victim's profile, including aggressive captions on shared media or hash-tags, or even creating fake profiles pretending to be someone else [100]. Vine is a popular video-based social

network that allows its users to record and edit six-second looping videos, which they can share on their profiles for others to see, like, revine and comment upon. Cyberbullying can happen in Vine in many ways, including posting mean, aggressive and hurtful comments, recording video of others without their knowledge and then sharing the Vines as a way to make fun of or mock them, and playing "the slap game" in which one person records video while another person slaps or hits a person in order to record a reaction. They later share the Vine for the world to see. There are even violent versions called "knock-out" where someone punches an unsuspecting person in an attempt to knock them out [37]. Figure 5.1 provides an illustration where the profile owner is victimized by hurtful and aggressive comments posted by others in Vine and Instagram respectively.

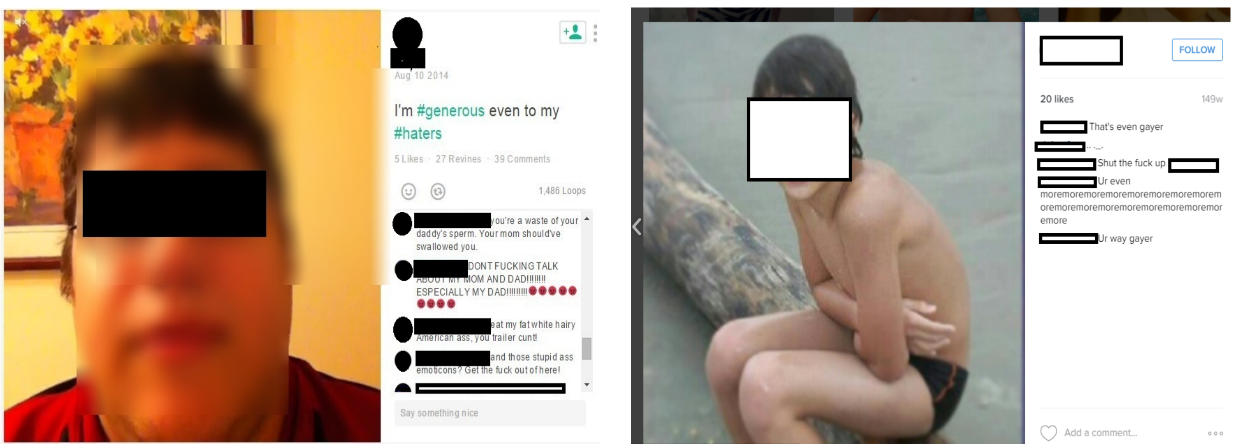


Figure 5.1: Examples of cyberbullying in the (L) Vine and (R) Instagram online social networks.

In online postings, cyberaggression is defined as a type of behavior in an electronic context that is meant to harm another person (e.g., verbal abuse from an anonymous user online). Cyberbullying is cyberaggression that is carried out repeatedly, against a person who cannot easily defend himself or herself, and where the bully has power over the victim [65, 86]. Thus in order to understand cyberbullying behavior, the factors of repetition of aggressive behavior and imbalance of power between victims and perpetrators must be considered. Previous works [49, 48] have reported that not all media sessions (shared media + associated comments) that exhibit cyberaggression are necessarily instances of cyberbullying. In this chapter, we go deeper to identify distinguishing features that differentiate cyberbullying postings from non-cyberbullying postings. In particular,

we make the following contributions:

- We investigate whether the numbers of unique commenters, unique positive sentiment commenters, and unique negative sentiment commenters have any influence in making a media session a cyberbullying one for both Vine and Instagram.
- We conduct a temporal analysis of all comments and comments belonging to the profile owner to investigate any differentiating patterns between cyberbullying and non-cyberbullying media sessions.
- We perform a text-content analysis of the comments associated with the media sessions to check for any distinguishing factors between cyberbullying and non-cyberbullying media sessions.

5.1 Data Set

We use labeled data from Instagram [48] and Vine from chapter 4, which label each media session as an instance of cyberbullying, cyber aggression, both, or neither. The data was originally collected using snowball sampling and labeled using the crowdsourcing work platform CrowdFlower (See [48] and chapter 4 for the detailed methodology for data collection and labeling). To improve the quality of our analysis, we filter the data-set to include only media sessions with a high confidence score of being correct. For each media session, each judgment is given a trust score that incorporates the overall trust score of a labeler with the score that the labeler got while answering the test questions given on the survey (administered during the labeling process). This trust value is, in turn, incorporated with the majority voting method to assign a confidence score to the label given to a particular media session.

For our analysis, we only use media sessions with a confidence score of 90% or higher. For Vine, this filtering reduced 983 media sessions to 42 cyberbullying media sessions and 213 non-cyberbullying media sessions. For Instagram, this filtering reduced 2216 media sessions to 239 cyberbullying media sessions and 769 non-cyberbullying media sessions. The reason for using

a high confidence score is twofold. Firstly, it means that the labelers were unanimous in their labeling of a particular media session. Secondly, it gives us a manageable number of media sessions to perform 3D temporal analysis of comments.

5.2 Analysis of Unique Commenters

We first investigate whether the number of unique commenters has any possible influence when it comes to making a media session an instance of cyberbullying. Here, the number of unique commenters means the number of distinct users who comment of a media session. We consider the total number of unique commenters, the total number of unique positive sentiment commenters and the total number of unique negative sentiment commenters. For this purpose, we take the comments associated with the labeled media sessions for both Vine and Instagram and perform sentiment analysis of all the comments using Python's NLTK library. NLTK computes polarity for each comment that shows how negative or positive a particular comment's sentiment is. After getting all the comments and getting their corresponding sentiments, we generate CCDF (Complementary Cumulative Distribution Function) of the number of unique commenters (Figure 5.2), number of unique positive sentiment commenters (Figure 5.3) and the number of unique negative sentiment commenters (Figure 5.4) vs the percentage of total cyberbullying (non-cyberbullying) media sessions for Vine and Instagram, respectively.

In Figure 5.2, the red and blue plots stand for the cyberbullying and non-cyberbullying media sessions respectively. The X-axis denotes the number of unique commenters and the Y-axis denotes the percentage of cyberbullying (non-cyberbullying) media sessions out of total cyberbullying (non-cyberbullying) media sessions having at least that many numbers of unique commenters. It is evident from the figure that for both Vine and Instagram, the number of unique commenters tends to have the same pattern for cyberbullying and non-cyberbullying. The same indistinguishable trend for both labels is also seen for the total number of unique positive sentiment commenters from Figure 5.3. This means that **for both Vine and Instagram, cyberbullying and non-cyberbullying media sessions tend to have the same trend when it comes to the number**

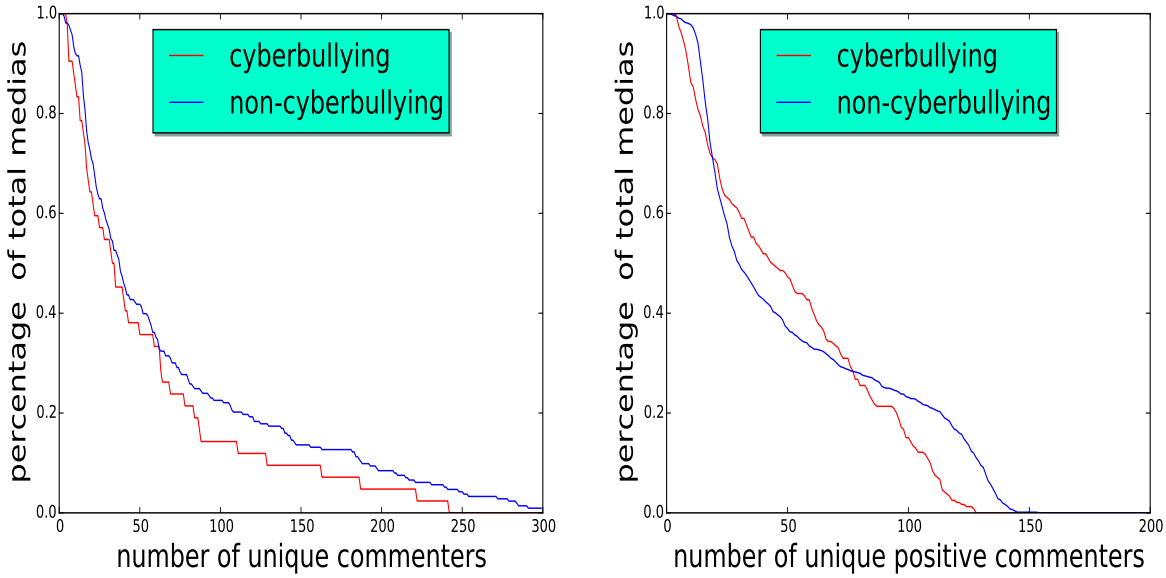


Figure 5.2: CCDF of number of unique commenters vs percentage of total cyberbullying(non-cyberbullying) media sessions for (L) Vine and (R) Instagram.

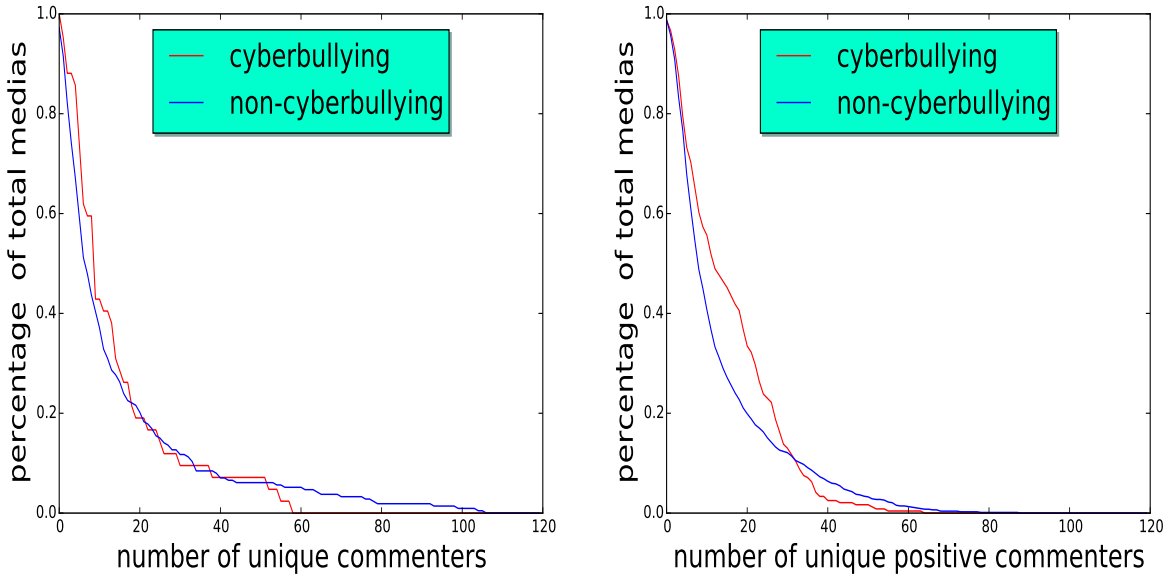


Figure 5.3: CCDF of the number of unique positive sentiment commenters vs percentage of cyberbullying and non-cyberbullying media sessions for (L) Vine and (R) Instagram.

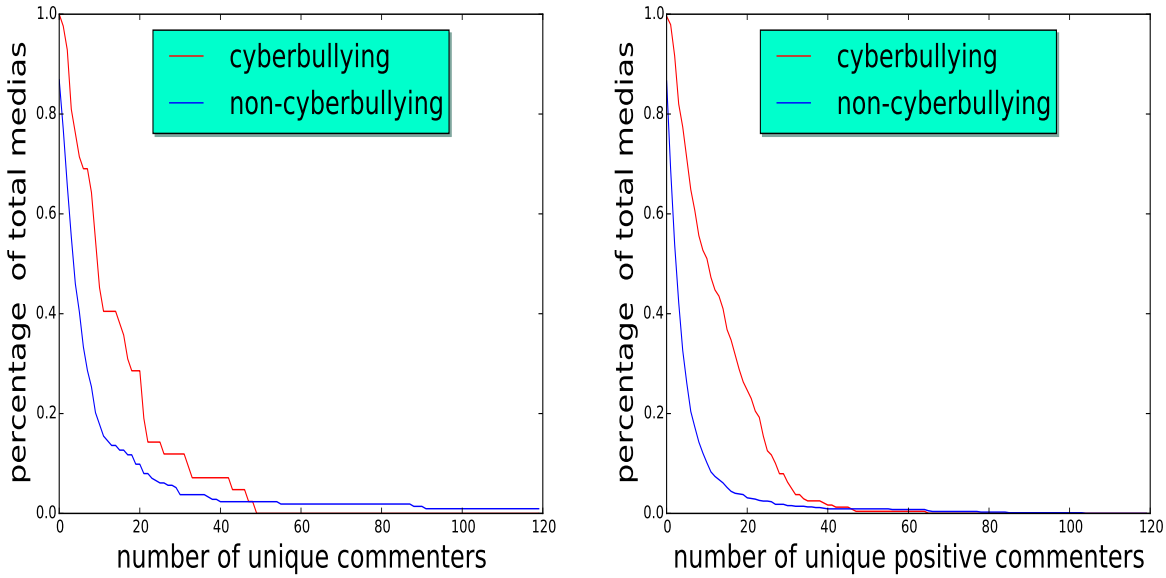


Figure 5.4: CCDF of number of unique negative sentiment commenters vs percentage of total cyberbullying (non-cyberbullying) media sessions for (L) Vine and (R) Instagram.

of unique commenters and the number of unique positive sentiment commenters.

However, for the number of unique negative commenters (Figure 5.4), cyberbullying and non-cyberbullying sessions differ from one another. It is seen that, for both Vine and Instagram, the number of unique negative commenters trend for cyberbullying media sessions falls much more slowly than for non-cyberbullying sessions. The figure shows that the percentage of cyberbullying media sessions having at least a certain number of negative unique commenters is much more than that of non-cyberbullying media sessions. This means that **cyberbullying media sessions are likely to have more unique negative sentiment commenters for Vine and Instagram**. We believe this is because, in a cyberbullying media session, perpetrators often gang up against the victim and thus spiking up the number of unique negative sentiment commenters. It can also be seen that after 40 unique negative sentiment commenters, the non-cyberbullying trend starts to show a long tail, which is not seen for the cyberbullying trend. This is because some non-cyberbullying media sessions belong to celebrities and famous brands that have a large number of comments from a large number of followers, and sometimes the commenters express awe with expletives and/or swear words in those media sessions, thus contributing to the long tail.

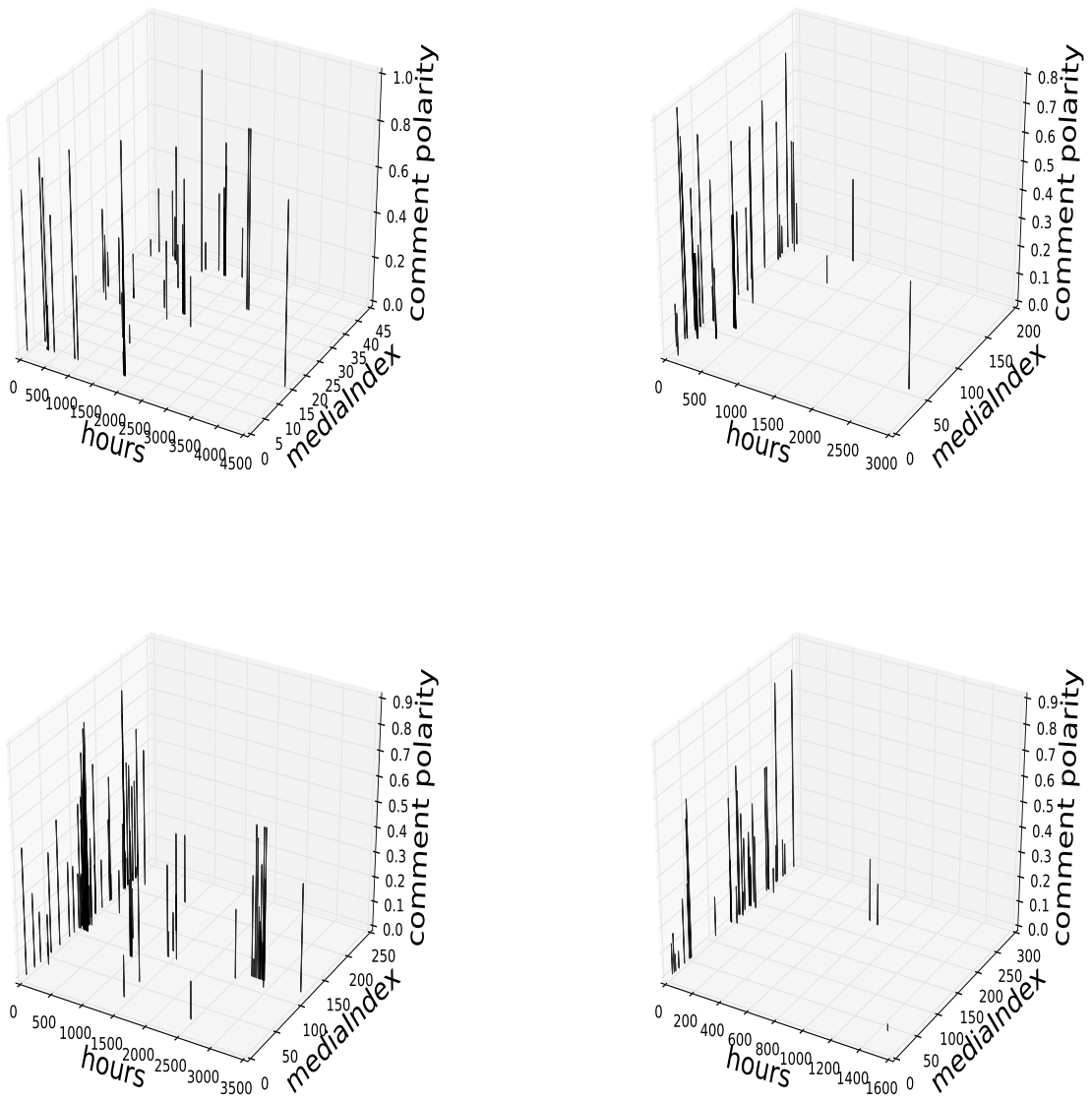


Figure 5.5: Polarity of negative sentiment profile owner comments as hours move on since the media session has been posted for cyberbullying and non-cyberbullying Vine and Instagram media sessions. (Vine cyberbullying upper left, Vine not-cyberbullying upper right, Instagram cyberbullying lower left, Instagram not-cyberbullying lower right)

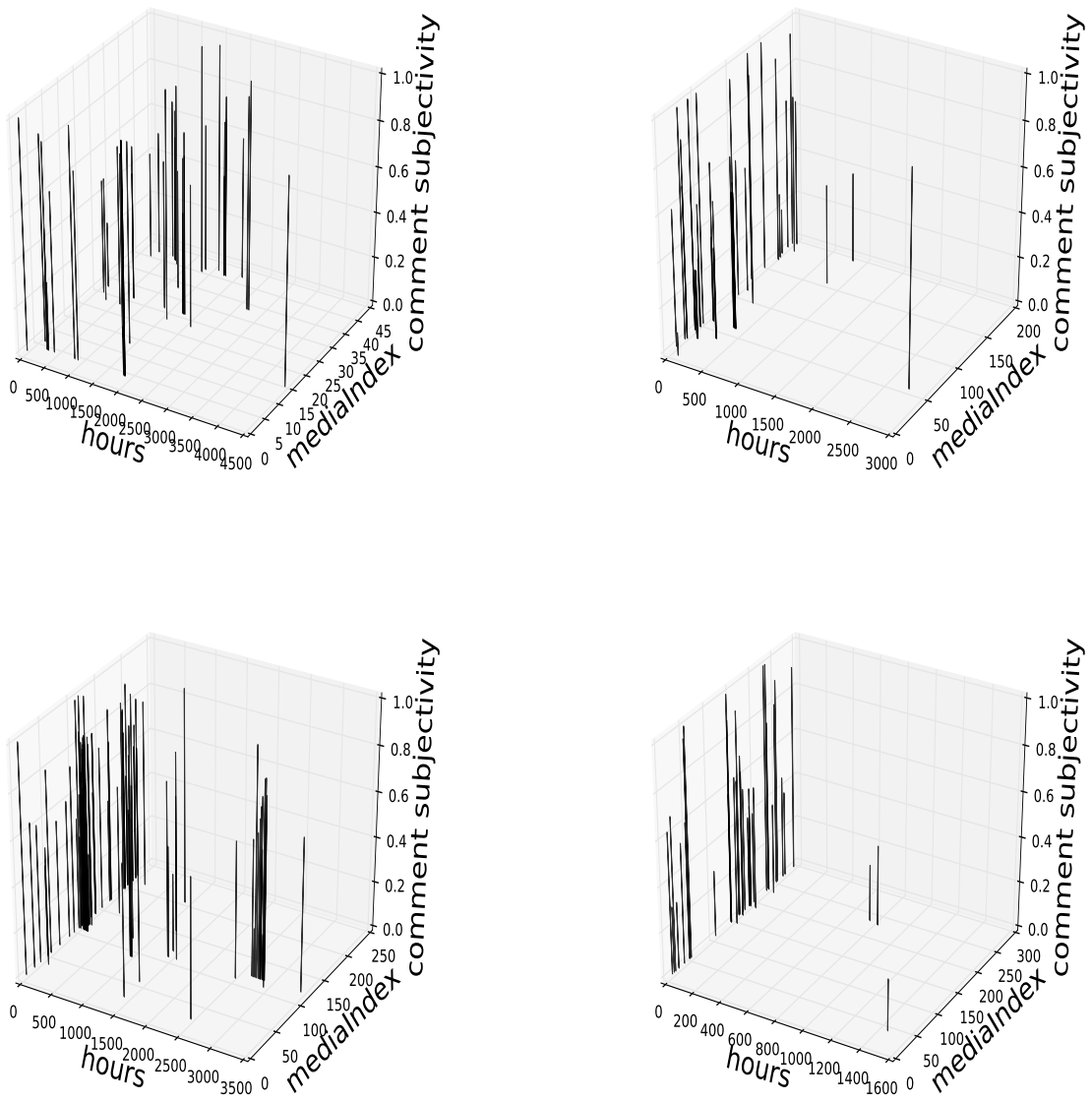


Figure 5.6: Subjectivity of negative sentiment profile owner comments as hours move on since the media session has been posted for cyberbullying and non-cyberbullying Vine and Instagram media sessions.(Vine cyberbullying upper left, Vine not-cyberbullying upper right, Instagram cyberbullying lower left, Instagram not-cyberbullying lower right)

5.3 Temporal Analysis of Profile Owner Comments

We now conduct a temporal analysis of the negative sentiment comments belonging to the profile owner, the person who posted the original media session. The intuition is that since the definition of cyberbullying involves repeated aggression, the profile owners might also use negativity to defend themselves against that repeated aggression. We also intuit that the profile owner's defensive negative sentiment comments are more likely to be spread across the temporal frame due to the repetition of aggression in a cyberbullying instance.

Figure 5.5 shows the temporal profile owner negative sentiment comment polarity (computed by Python's NLTK library) as the hours passed on since the sharing of the media session for Vine and Instagram. The higher the bars, the more negative a particular comment is. The hours delineate the number of hours passed since the sharing of the media by the profile owner. As seen in the figure, the bars for the cyberbullying media sessions are much more spread across the temporal frame whereas almost all the negative profile owner negative sentiment comments for non-cyberbullying media sessions emerge within the first two weeks. For cyberbullying media sessions, high bars even after a long time since the media has been shared indicate that repeated aggression behaviors have been occurring in those media sessions even after a long time after their emergence, which might have propelled the profile owners to defend themselves with negative sentiment comments. In Instagram, almost all of the profile owner negative sentiment comments for non-cyberbullying media sessions happen within one week of the time when the media session is shared. Similar to Vine, for cyberbullying media sessions, the negative sentiment profile owner comments are also spread across the time-frame. Both these figures confirm our intuition that for **cyberbullying media sessions, the profile owners are much more likely to post highly negative sentiment comments spread across the temporal frame since the media is posted than for a non-cyberbullying media session**.

A similar temporal analysis is performed for the negative comments' subjectivity that determines how severe a negative sentiment comment is. This value is obtained from Python's NLTK

library. Figure 5.6 shows the subjectivity value of each negative sentiment comment belonging to the profile owner as hours pass on since the media has been shared in Vine and Instagram respectively. It is apparent that the subjectivity values of the profile owner's negative sentiment comments are comparatively much higher in the cyberbullying media sessions than those in non-cyberbullying media sessions. This further confirms the observation that is seen from the temporal polarity analysis in the previous paragraph:**the owners react to the repeated aggression by posting negative sentiment comments with high subjectivity across the temporal frame in the cyberbullying media sessions**. For both Vine and Instagram, it is also evident that the highly negative sentiment comments belonging to the profile owners are much denser in the cyberbullying media sessions than the non-cyberbullying media sessions.

5.4 Temporal Analysis of Negative and Positive Sentiment Comments

Now we turn our attention to the temporal analysis of comments posted by other users on a particular media session since the media session is shared. We perform the analysis on all negative and positive sentiment comments where the sentiment was determined by using the Python's NLTK library. We do the temporal analysis for both negative and positive sentiment comments because we think media sessions that are tagged as cyberbullying are more likely to have a higher concentration of negative sentiment comments and lower concentration of positive comments, thus resulting in the imbalance of power as per the definition of cyberbullying.

Figure 5.7 shows the temporal comment polarity for all negative sentiment comments for a particular media session since the sharing of the media session for Vine and Instagram respectively. It is evident from both of these figures that the negative sentiment comments are much more spread up across the temporal frame of each media session in the case of cyberbullying sessions than for the non-cyberbullying sessions. The cyberbullying media sessions have a constant flow of high negative sentiment comments pouring in, even after a considerable amount of time since sharing of the media. On the contrary, the same cannot be said for the non-cyberbullying sessions as the number of negative sentiment comments tend to go down as time moves on. We believe this is a very

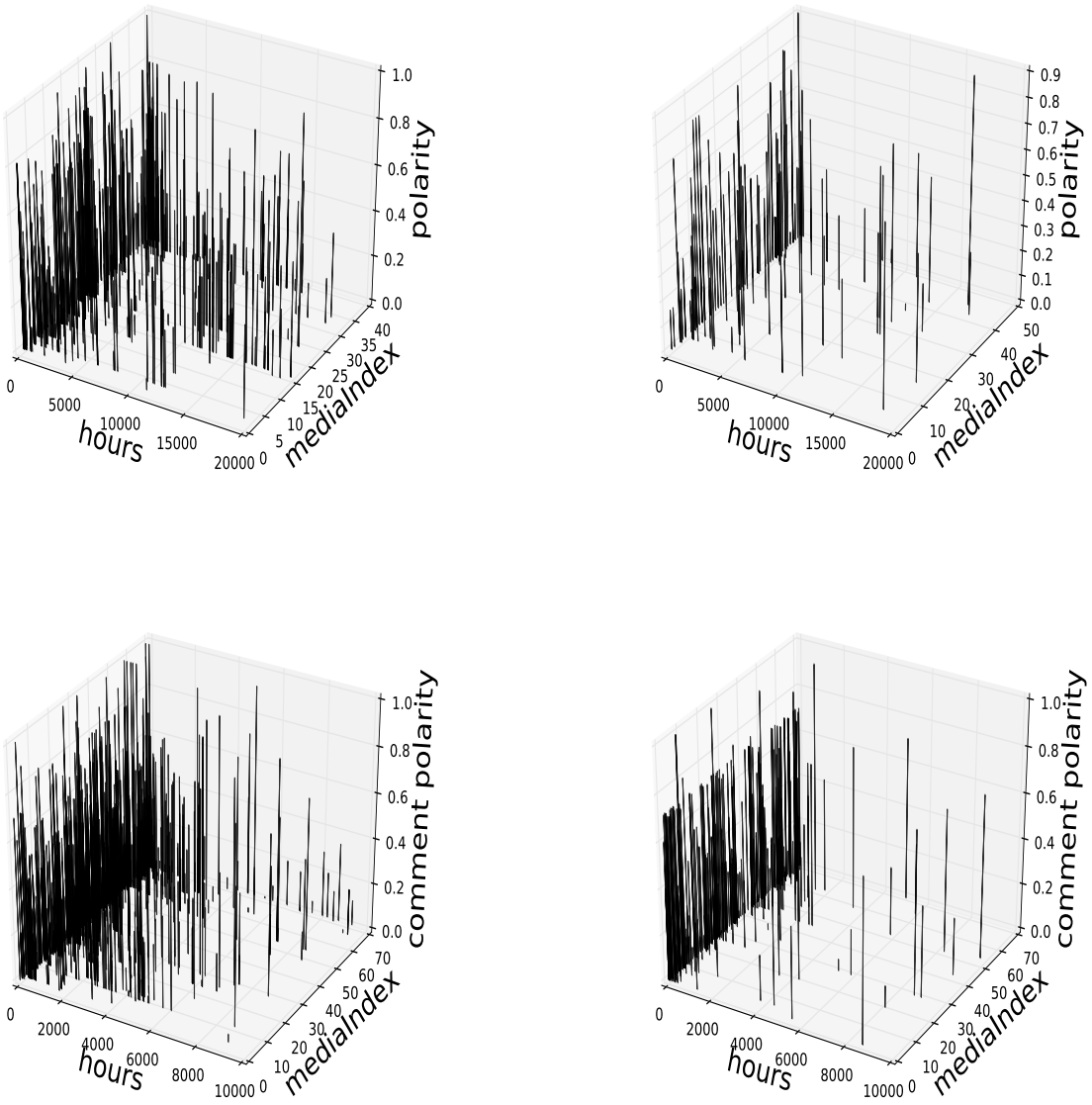


Figure 5.7: Polarity of negative sentiment comments as time moves on since the media session has been posted for cyberbullying and non-cyberbullying media sessions in Vine and Instagram.(Vine cyberbullying upper left, Vine not-cyberbullying upper right, Instagram cyberbullying lower left, Instagram not-cyberbullying lower right)

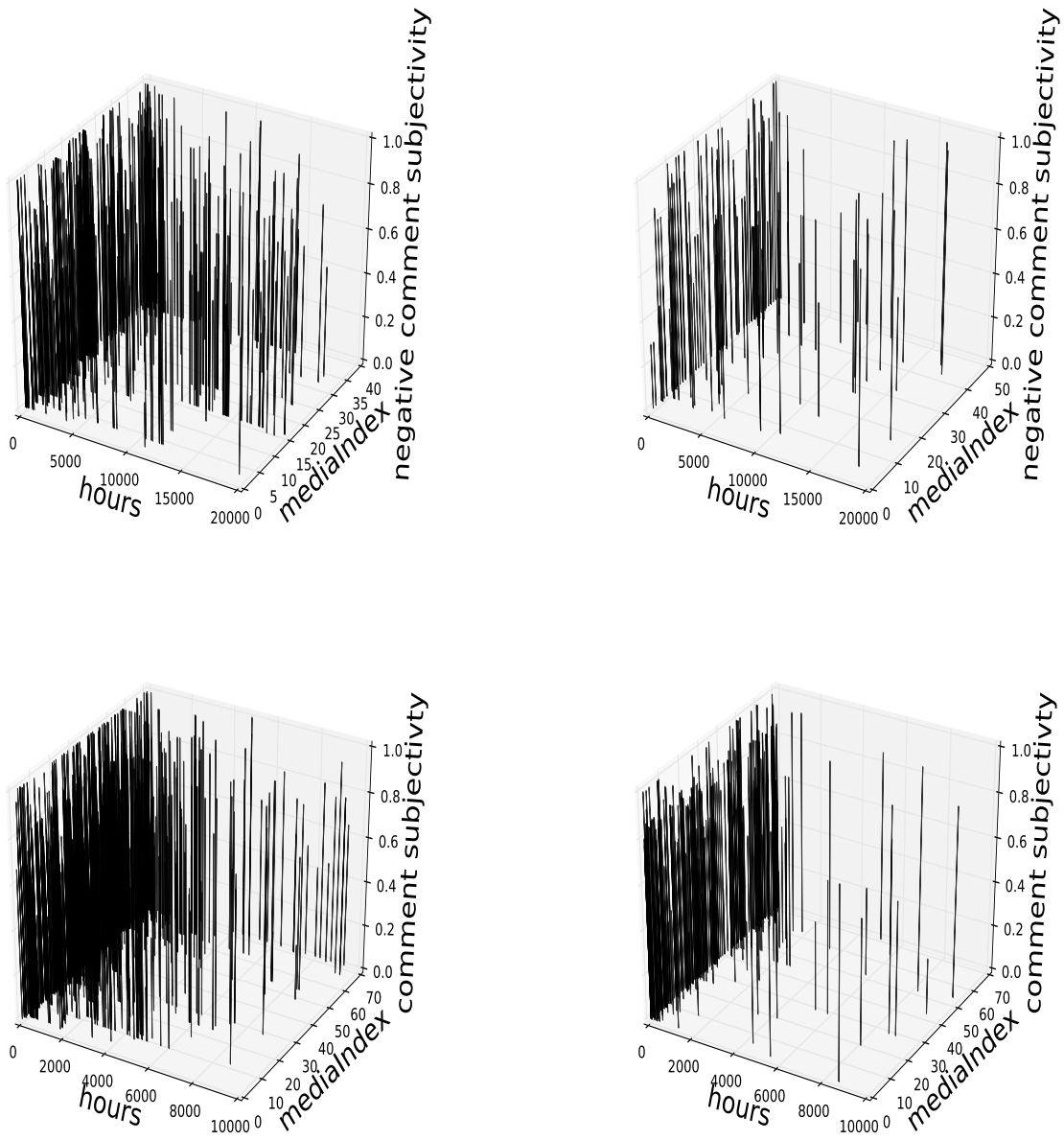


Figure 5.8: Subjectivity of negative sentiment comments as time moves on since the media session has been posted for cyberbullying and non-cyberbullying media sessions in Vine and Instagram.(Vine cyberbullying upper left, Vine not-cyberbullying upper right, Instagram cyberbullying lower left, Instagram not-cyberbullying lower right)

important factor that can differentiate a cyberbullying media session from a non-cyberbullying one. This shows that **in the cyberbullying media sessions, the negative sentiment comments persist even after a long time since the sharing of the media, which confirms the factor of repetition of aggression in the definition of cyberbullying**.

Next, we conduct the same kind of kind of temporal analysis to investigate the subjectivity of the negative sentiment comments for cyberbullying and non-cyberbullying media sessions for both Vine and Instagram. The intuition is that the cyberbullying media sessions should have more negative sentiment comments with comparatively higher subjectivity, thus being more aggressive which in turn results in cyberbullying. Figure 5.8 shows the subjectivity values of all the negative sentiment comments posted for the cyberbullying and non-cyberbullying media sessions since the sharing of the media sessions for both Vine and Instagram. It is apparent from the figures that the cyberbullying media sessions for both Vine and Instagram keep having negative sentiment comments with very high subjectivity spread across the temporal frame since the sharing of the media session. This results in the denser concentration of high bars for the cyberbullying sessions. **So not only the cyberbullying media sessions keep getting more negative sentiment comments even after a long time since the media session is posted, but also the negative sentiment comments tend to have more subjectivity than non-cyberbullying media sessions**.

Now, we conduct a temporal analysis of the polarity of all the positive sentiment comments for cyberbullying and non-cyberbullying media sessions for both Vine and Instagram. The expectation is that the cyberbullying sessions should have less concentrated positive sentiment comments, thus rendering the effect of imbalance of power as delineated in the definition of cyberbullying. Figure 5.9 shows the temporal comment polarity for all positive sentiment comments for a particular media session since the moment the media session has been posted for Vine and Instagram respectively. From both the figures, it is seen that the **density of positive comments coming in for cyberbullying media sessions for both Vine and Instagram is much less than the non-cyberbullying media sessions**. This lesser concentration of positive sentiment comments coupled with the denser concentration of negative sentiment comments with high subjectivity spread

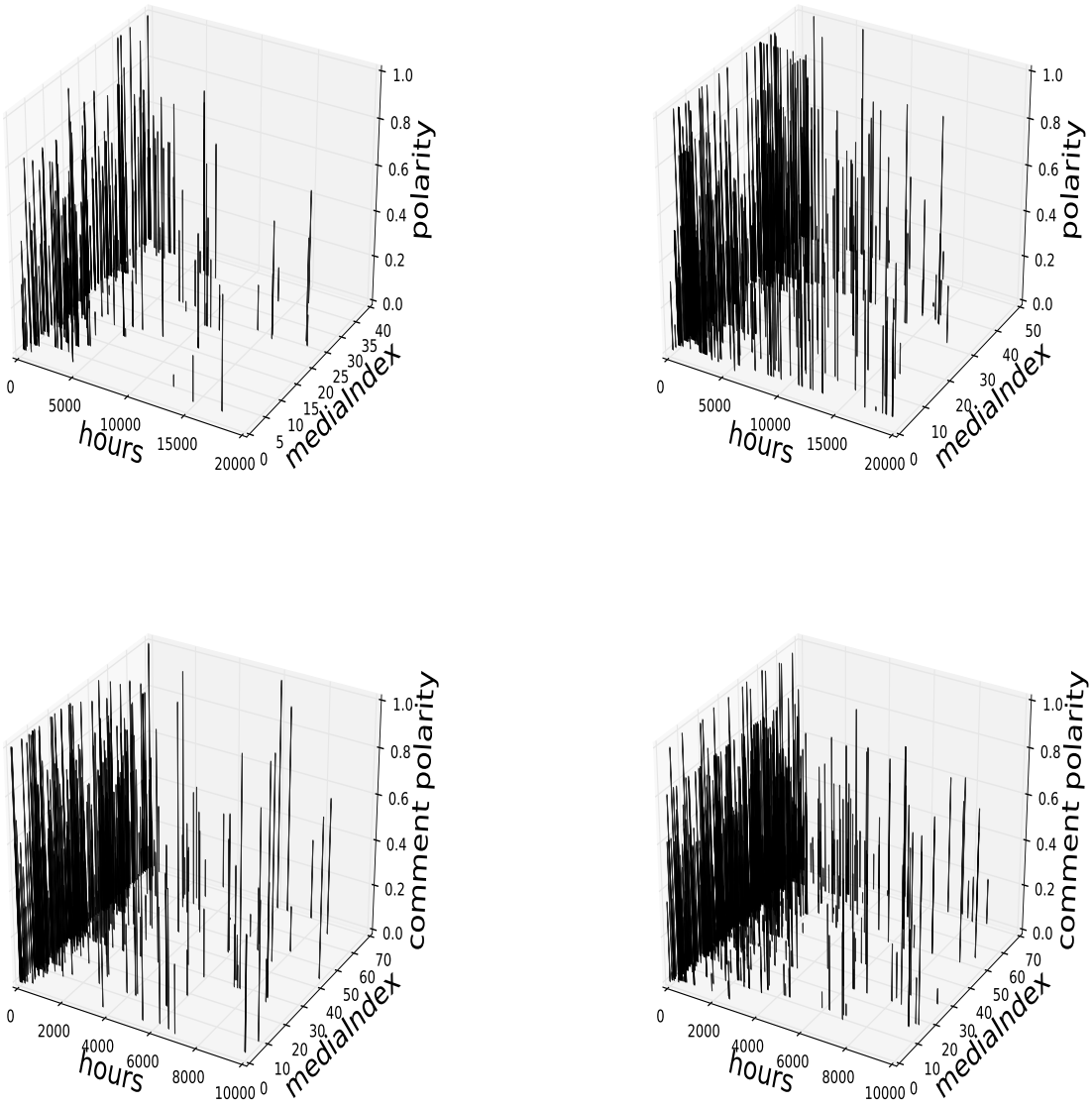


Figure 5.9: Polarity of positive sentiment comments as time moves on since the media session has been posted for cyberbullying and non-cyberbullying media sessions in Vine and Instagram.(Vine cyberbullying upper left, Vine not-cyberbullying upper right, Instagram cyberbullying lower left, Instagram not-cyberbullying lower right)

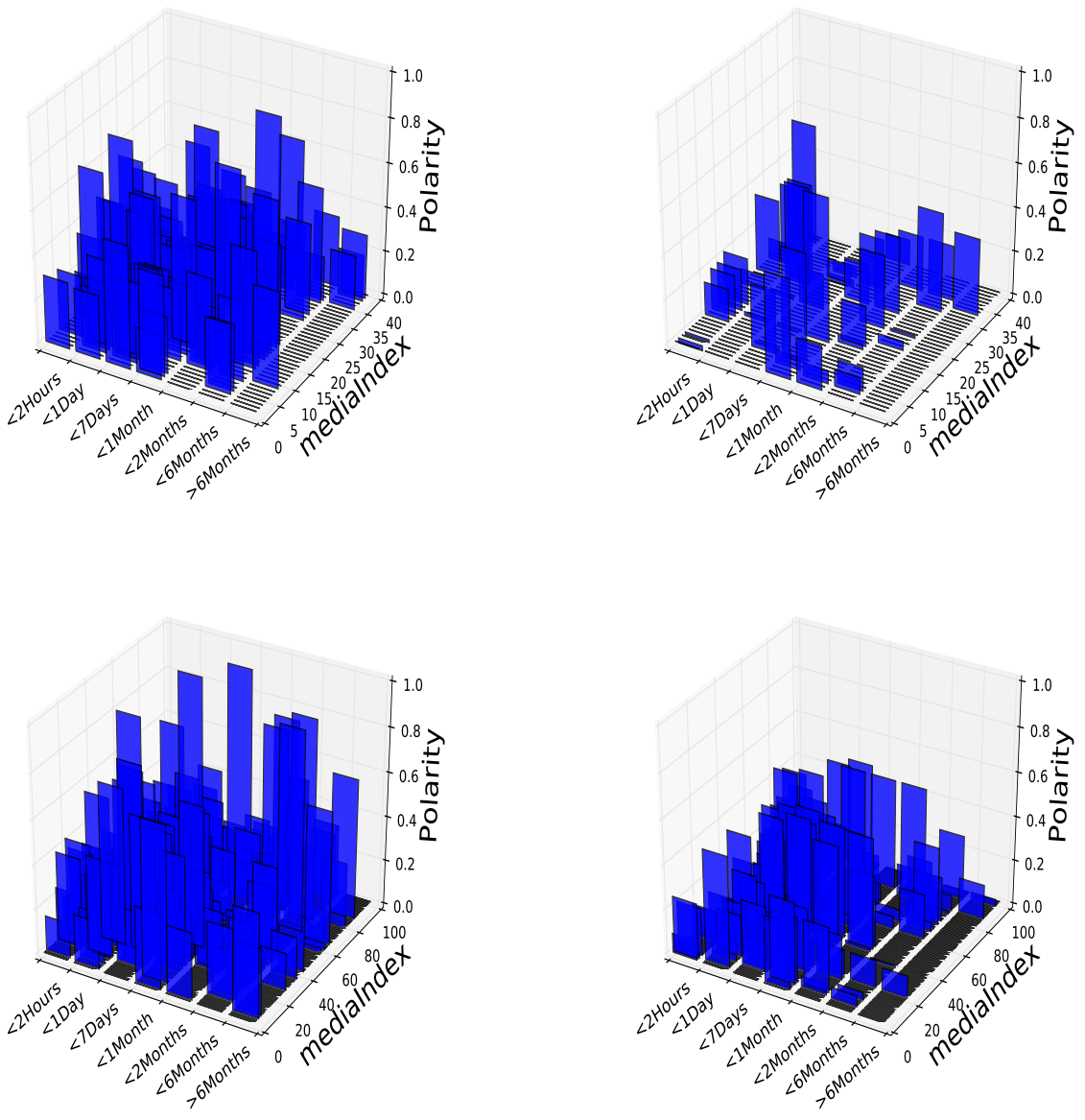


Figure 5.10: Polarity of grouped comments as time moves on since the media session has been posted for cyberbullying and non-cyberbullying media sessions in Vine and Instagram.(Vine cyberbullying upper left, Vine not-cyberbullying upper right, Instagram cyberbullying lower left, Instagram not-cyberbullying lower right)

across the temporal frame instigates the effect of imbalance of power and repeated aggression, thus rendering the media session a cyberbullying one.

Finally, we group the comments belonging to a particular media session based on the time they were posted. We create 7 temporal groups, namely, all comments that came in within the first two hours since the media was shared, within the third hour to one day, within one day to one week, within one week to one month, within one month to two months, within two months to six months and comments that appeared after six months. The intuition of grouping of the comments is two-fold. Firstly, we assume that each comment belongs to a discussion thread on a certain topic that the commenters may have been talking about. Secondly, we try to analyze the overall sentiment of these grouping of comments as temporal discussions to investigate any differentiating pattern for cyberbullying.

Figure 5.10 shows the negative sentiment of these group of comments together as a part of a temporal discussion as time moves on since the sharing of the media. It is seen that for both the social networks, **the negative sentiment polarity of the discussions belonging to cyberbullying media sessions show a higher level**. The height of the bars indicates high negative sentiment discussions. Also, the density of high bars for the cyberbullying sessions in both of these figures indicate that the temporal discussions in the comment section of the cyberbullying media sessions tend to be more negative than the non-cyberbullying discussions. It is also evident that **while for non-cyberbullying media sessions, the negative sentiment temporal discussions tend to fizzle out as time moves on, that is not the case for the cyberbullying media sessions**.

5.5 Analysis of Comments

Finally, we perform a text-content analysis for the comments associated with a media session for both Vine and Instagram. We consider the comments associated with a media session from part of a discussion thread, and our goal is to determine the differences between a discussion thread of a cyberbullying session and a discussion thread of a non-cyberbullying thread.

Figure 5.11: Frequency distribution of words used for the cyberbullying and non-cyberbullying media sessions' comments in Vine and Instagram.(Vine cyberbullying upper left, Vine not-cyberbullying upper right, Instagram cyberbullying lower left, Instagram not-cyberbullying lower right)

Figure 5.12: Top IDF valued distribution of words used for the cyberbullying and non-cyberbullying media sessions' comments in Vine and Instagram.(Vine cyberbullying upper left, Vine not-cyberbullying upper right, Instagram cyberbullying lower left, Instagram not-cyberbullying lower right)

First, we devise a word frequency cloud for the cyberbullying and non-cyberbullying media sessions for both the social networks to get an idea of the words that occur frequently. Figure 5.11 shows the frequency distribution of words of all the media sessions' comments belonging to cyberbullying and non-cyberbullying media sessions for Vine and Instagram respectively. It can be seen from these figures that **negative sentiment words are much more frequent in the discussion comment threads of cyberbullying sessions**.

Next, we do an IDF (Inverse Document Frequency) analysis of the media sessions' comments that measures how common a word is across all media session comment discussions for cyberbullying and non-cyberbullying sessions. The difference between the frequency analysis and IDF analysis is that frequency analysis only take into account the number of times a word appears in a discussion thread whereas IDF analysis gives us words that are common across all cyberbullying and non-cyberbullying comment discussion threads. Thus, a word that appears 10 times in 10 different documents will have lower IDF than a word that appears 10 times in a single document. Figure 5.12 shows the commonly appearing words for cyberbullying and non-cyberbullying media session comment threads for both Vine and Instagram respectively across all the corresponding media session comment threads. The bigger a word is in the word cloud, the more common it is across all the media session comment threads belonging to either cyberbullying or non-cyberbullying label. It is evident that, as it was seen also from the previous paragraph, **a cyberbullying media session comment discussion thread is much more likely to have negative sentiment words**.

To further solidify the aforementioned claim, we use the negative sentiment word list [53] and find out the percentage of cyberbullying (non-cyberbullying) media sessions out of total cyberbullying (non-cyberbullying) media sessions whose comment threads contain those negative sentiment words. We intuit that negative sentiment words appear more in the cyberbullying media sessions than the non-cyberbullying media sessions for both the social networks, thus forming a differentiating factor for cyberbullying. We can see from Figures 5.13 and 5.14, negative sentiment words are much more likely to appear in a cyberbullying media session's associated comments than the non-cyberbullying media sessions, thus further confirming our claim: **a cyberbullying media**

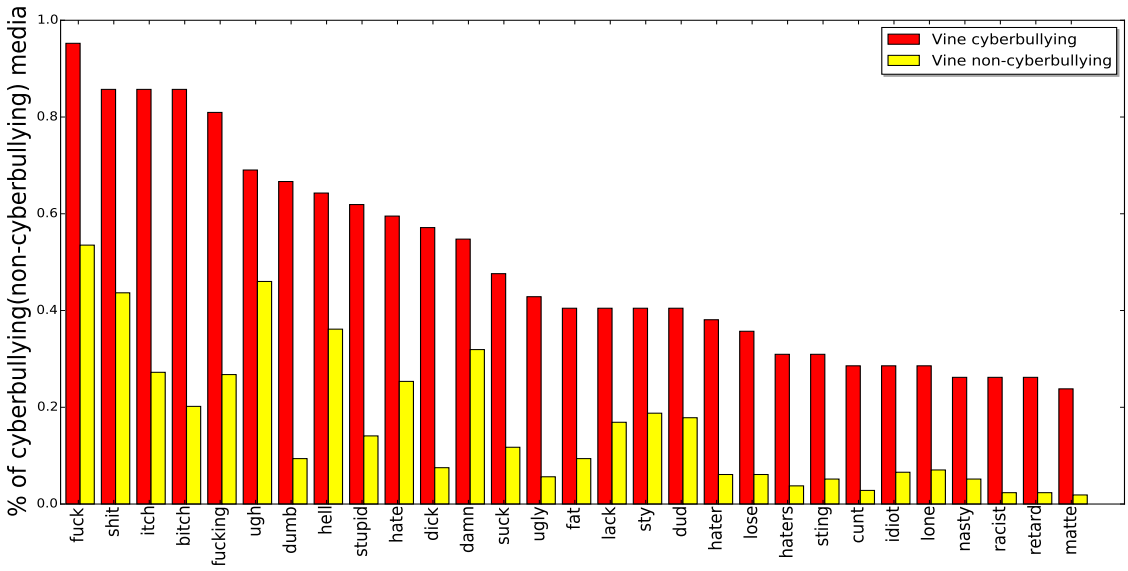


Figure 5.13: Negative sentiment words vs percentage of cyberbullying (non-cyberbullying) media sessions out of total cyberbullying (non-cyberbullying) media sessions' comment threads containing that word in Vine.

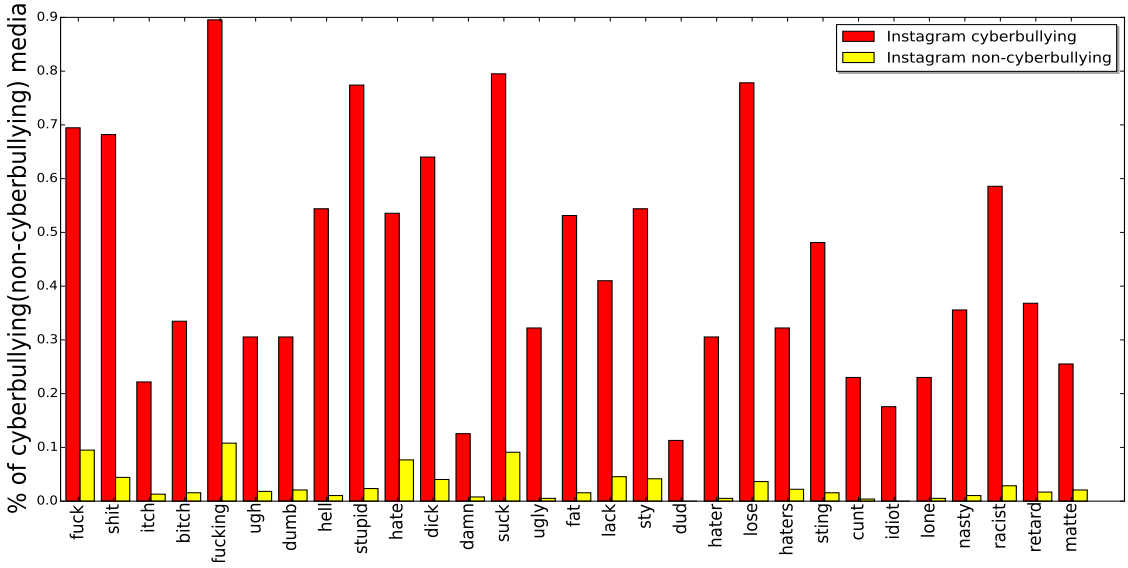


Figure 5.14: Negative sentiment words vs percentage of cyberbullying (non-cyberbullying) media sessions out of total cyberbullying (non-cyberbullying) media sessions' comment threads containing that word in Instagram.

session comment discussion thread is much more likely to have negative sentiment

words.

5.6 Conclusions

To the best of our knowledge, this is the first research to investigate factors that differentiate a cyberbullying session from a non-cyberbullying one for both Vine and Instagram, two media-based online social networks. Labeled data that used an appropriate definition of cyberbullying has been used to analyze the media sessions. We analyze the number of unique commenters, unique positive sentiment commenters, and unique negative sentiment commenters. We then perform a temporal analysis of the comments belonging to the profile owner and all comments respectively for both social networks. Finally, we conduct a content analysis of the comment threads belonging to the labeled cyberbullying and non-cyberbullying media sessions.

The key findings of this research are as follows. First, for both Vine and Instagram, cyberbullying media sessions are more likely to have more unique negative sentiment commenters. Second, in cyberbullying media sessions, profile owners are much more likely to post highly negative sentiment comments with comparatively higher subjectivity spread across the temporal frame since the sharing of the media than in a non-cyberbullying media session. Third, in the cyberbullying media sessions, negative sentiment comments persist with higher subjectivity even after a long time since the media has been posted, which is not the case for non-cyberbullying media sessions. Fourth, the density of positive comments coming in for cyberbullying media sessions for both Vine and Instagram is much less than that for the non-cyberbullying media sessions across the temporal frame. Fifth, the comment discussion threads across time units belonging to cyberbullying media sessions show a high level of negative sentiment polarity than those belonging to non-cyberbullying sessions. Sixth, while for non-cyberbullying media sessions, negative sentiment discussions tend to fizzle out as time moves on, that is not the case for cyberbullying media sessions in Vine and Instagram. Seventh, a cyberbullying media session comment thread is much more likely to have negative sentiment words than a non-cyberbullying media session.

5.7 Acknowledgments

This work was supported by the US National Science Foundation (NSF) through grant CNS 1528138.

Chapter 6

Scalable and Timely Detection of Cyberbullying in Online Social Networks

Unprecedented growth in the popularity of OSNs, especially among teenagers, has unfortunately resulted in a significant increase in cyberbullying. The main differences between bullying and cyberbullying are the facts that the Internet can help the perpetrators of cyberbullying hide their identities, cyberbullying can be incessant because of the availability and access of the Internet, and the possibility of cyberbullying being viral and exposing the victims to an entire virtual world[106]. Numerous instances [23] and devastating consequences of cyberbullying [36, 105, 11] have led researchers to explore detection of cyberbullying incidents in OSNs like Ask.fm, Instagram, Vine, Twitter etc. [48, 47, 68, 17]. These works have mostly followed the methodology of collecting and labeling data from OSNs and building cyberbullying classifiers. Past works have also investigated the issue of identifying imbalance of power between perpetrators and victims, which is a key feature of cyberbullying, and distinguishing between cyber-aggression and cyberbullying [65], thus paving the way for highly accurate classifiers.

While progress has been made on the accuracy of classifiers for cyberbullying detection, there are two key practical issues that have largely been ignored to date. The first issue concerns the scalability of cyberbullying detection solutions. OSNs, of course, involve an enormous amount of data, on the order of several hundred gigabytes per day. For example, it has been reported that for Vine around 39 million videos have been shared since it was introduced [102] while for Instagram, the amount of shared media is 40 billion[67].

The second issue concerns the timeliness of raising alerts whenever cyberbullying incidents

are suspected. Cyberbullying is different from traditional, face-to-face bullying, because it can occur 24/7, and perpetrators can be anonymous and have easy access to sophisticated tools to launch cyberbullying attacks. Furthermore, the consequences can be disastrous and it is extremely important to provide the necessary support to the victims as early as possible. So, a timely detection of cyberbullying is of paramount importance, so that an alert can be raised as soon as possible.

This chapter proposes a multi-stage cyberbullying detection solution designed to improve the scalability and responsiveness of cyberbullying detection. **To the best of our knowledge, this chapter is the first to propose a scalable and responsive solution to cyberbullying detection in OSNs**. A key property of the solution is that it achieves sufficient classification accuracy while accomplishing these two goals. The solution consists of two key components, namely, a dynamic, multilevel priority scheduler for improved responsiveness, and an incremental feature extraction and classification stage for scaling. Using online social networking data from Vine, we demonstrate the utility of both of these components and show that our complete cyberbullying detection solution is significantly more scalable and responsive than the current state-of-the-art. We the following important contributions:

- We outline an incremental computational design for feature extraction and classification that reuses previous classification results to reduce overhead with minimal impact on accuracy.
- We introduce a dynamic, multi-level priority scheduler that assigns high preference to potential cyberbullying media-sessions, thereby improving the responsiveness of the solution.
- Using real-world data from Vine, we show that our integrated system substantially improves the scalability of cyberbullying, making cyberbullying detection feasible for Vine-scale social networks.
- We further demonstrate how our system scales to monitor much larger, Instagram-scale networks.

6.1 Design Overview

Our goal of this research is to propose a cyberbullying detection system with two key characteristics, namely, scalable enough to handle large OSNs without sacrificing accuracy and timeliness of raising an alert when a cyberbullying instance takes place. To this aim, we face two key challenges. First, how to scale up the system while retaining a reasonable detection performance. Second, how to design a system so as to make sure an alert is raised as soon as a cyberbullying instance takes place while monitoring a large number of media sessions(media and its associated comments in Instagram or Vine). In the following subsections, we describe the two components of our system that address these challenges.

6.1.1 Incremental Classifier

X_n : saved feature vector values for n comments from before for all features;
δn : new comments to be processed;
X_n^i : feature vector value of i-th feature for n comments;
$X_{\delta n}^i$: feature vector value of i-th feature for δn comments;
$|X_n|$: number of total features;
forall i **in** $1, 2, ..., |X_n|$ **do**
  $X_{n+\delta n}^i : X_n^i$+Compute($X_{\delta n}^i$);
end

Algorithm 1: IncrementalFeatureExtraction()

Our first challenge is to build a cyberbullying detection classifier that is scalable when it comes to time and computing resources while also retaining sufficient classification performance. While sophisticated deep learning classifiers have been recently introduced to solve complex problems with high accuracy [56, 66], they come up with considerable computational baggage. For example, in [56], the authors used deep learning in real time to process one 1080p video frame in 644ms using Samsung S7 with leveraging high-performance GPUs(12 GPUs) and 4GB memory. While it is tempting to use deep learning for our system, we want our classifier to be able to leverage lightweight computational resources(Amazon AWS free tier 1GB memory, for example). In addition to being computationally lightweight, we also want our classifier to be faster than the slower current

state-of-the-art AdaBoost developed in chapter 4(as shown later in Table 6.1) to classify when new comments for a media session comes in, without sacrificing accuracy. We see that both these cases (deep learning and AdaBoost), while being highly accurate, does not meet two key challenges, being computational resource-wise lightweight and efficient, respectively when using our lightweight computation resource constraints, mentioned above.

Our approach towards lowering computational resource scalability and improving efficiency while retaining sufficient accuracy is to incorporate **incremental computation** [41, 40] into the design of the potential classifier. Incremental computation reuses data from previous stages within the current stage, thus resulting in less computational complexity. Traditional classifiers need to execute a full run as each new datum arrives, e.g. new comment for a media session. Instead, our approach reuses previous stages' results and combines them with the new comments, thereby reducing computational cost, rendering the solution scalable. We seek ways to apply this incremental approach to both feature extraction and classification stages of the potential classifier. To this aim, we seek to employ a classifier that, during feature extraction stage, uses features which, by nature can be incrementally linear in the sense that once the values corresponding to these features have been computed for the first n comments, then when δn new comments arrive, we only have to compute the individual feature vector values for the new δn comments while reusing the values for the previous n comments to compute the overall feature vector for the $n + \delta n$ comments. This dramatically reduces resource and computation cost because this approach is driven by δn at each run instead of $n + \delta n$. Algorithm 1 provides a pseudo-code of the incremental feature extraction algorithm for our incremental detector. Similarly, the candidate classifier should also be able to leverage this incremental approach during classification stage once the feature vectors are extracted from new data.

We found that Logistic Regression (LR) was the most promising classifier that met all these aforementioned criteria. LR works as follows: if we have n features $a_i, i = 0, 1, 2, 3..., n-1$, after training, LR assigns a weight $w_i, i = 0, 1, 2..., n-1$ to each of those features, and then computes the combined features value $c = \sum_0^{n-1} a_i w_i$. This value is fed to a sigmoid function with output

ranging from 0 to 1[118]. The way we can leverage incremental computation in LR is as follows: LR takes as input a set of features X, and during the training process, the classifier generates a set of weights θ corresponding to those features. When a new media session comes in, the feature extraction step computes a matrix X for that particular media session and computes $C = X\dot{\theta}$, which is then used to make the corresponding prediction. For the incremental feature extraction sub-component, we save the X_{old} value for the previous n comments, compute $X_{\delta n}$ for the new set of δn arrived comments and compute the new X by combining X_{old} and $X_{\delta n}$ instead of computing X all over again for all $n + \delta n$ comments. For the incremental classification part, we only use those components of X that have been changed to compute $C = X\dot{\theta}$ instead of doing the full $X\dot{\theta}$ computation. For this purpose, we save $X_i \times \theta_i, \forall i$, where X_i is the i-th feature at time t. Then we only change the corresponding feature vector value X_i at time $t + \delta t$ if it has been changed by comparing it to the previous saved X_i at time t. If it has been changed, only then we take it into the account to compute $\sum_{\forall i} X_i \times \theta_i$ by simple addition and subtraction instead of full-scale matrix multiplication.

To be able to use incremental computation in the feature extraction stage, LR has to be able to make use of features who are, by nature, amenable to feature extraction. In addition, our incremental logistic regression also has to show sufficient efficiency and scalability over the current state-of-the-art. All these goals have to be met without sacrificing crucial classification qualities like precision and recall. In this section, we provided the design of incremental techniques our LR classifier uses to scale up. From now onward, we will refer to this incremental LR classifier as classifier. In section 6.2.1, we first justify our choice of LR by comparing its execution time, precision and recall with current state-of-the-art in chapter 4 while making use of features that accommodate incremental computation. We then justify using the incremental computation approach in the LR by comparing its scalability performances with standard LR that does not use the incremental approach.

6.1.2 Dynamic Priority Scheduler

While leveraging incremental computation helped our classifier to gain reasonable scalability, we found that it still lacked the other key issue that a potential cyberbullying detection system has to address: responsiveness. As the first step towards tackling this issue, we make use of two key observations. First, **not all media sessions need to be monitored equally**. The theme is to apply limited resources to where they needed the most. As most media sessions are not bullying in nature (shown in chapter 4), we should be able to apply our resources on media sessions that are most likely to result in cyberbullying. This observation makes it natural to build a scheduler that just keeps monitoring media sessions with high priority and discarding all the low priority ones. We call this scheduler Static Priority Scheduler (SPS). Our investigation of the performance of SPS on Vine labeled data from chapter 4 found its precision and recall to be 70 and 58 percent respectively. The low recall value shows that by totally ignoring the media sessions that were given low priority some stage of their lifetimes to achieve responsiveness, we miss a significant portion of potential cyberbullying media sessions. This trade-off between performance and responsiveness led us to make our second observation: **a media session, with its incoming stream of comments, can slowly evolve into a cyberbullying instance even if it started as a normal one at its early stages of lifetime and vice versa**. As new comments arrive for a media session, it may become more or less indicative of cyberbullying, depending on the nature of the newly arrived comments. This means it is important to examine all sessions including the ones with low priority, as some of them may evolve into cyberbullying sessions during later stages.

Investigation results after running SPS and the second observation formed the motivation of the design of our Dynamic Priority Scheduler (DPS). We define two levels of priority, namely high and low, for all media sessions and assign a high priority to all newly created media sessions. Now, our challenge is to accommodate learning based in new comments so as to dynamically vary each media session's priority. After each invocation of our incremental classifier component(Section 6.1.1), a confidence value[12] of how likely a media session contains cyberbullying is generated. We

forall media session m **do**
  $Conf_i^m$: confidence value of the i-th comment session prediction for this media session;
  n: number of total comment session prediction in the confidence history;
  $Avg_{confidence}^m = \frac{\sum_{i=1}^{n} Conf_i^m}{n}$;
  if $Avg_{confidence}^m \geq 0.2$ **and current priority is** LOW **then**
    set current Priority to $HIGH$;
    continue;
  end
  if $Avg_{confidence}^m < 0.2$ **and current priority is** $HIGH$ **then**
    set current Priority to LOW;
    continue;
  end
end

Algorithm 2: SettingPriority()

make use of the history of these confidence values to dynamically change a media session's priority. The reason for using history as opposed to just the most recent confidence value has to do with the definition of cyberbullying. Cyberbullying is defined as an aggressive online behavior that is *carried out repeatedly* against a person who *cannot easily defend himself or herself*, creating a power imbalance [65]. To identify repeated aggressive behavior or whether a victim can defend himself or herself, we need to consider a much longer history than just the most recent confidence value. We calculate the average of all past confidence values for past classifications and current classification of a particular media session and compare that with a threshold value. If the average confidence value is more than a certain threshold value, we assign a high priority to the session and if the average value is lower than the threshold value, we assign a low priority. Algorithm 2 illustrates our priority setting algorithm using an average confidence threshold (0.2 in this example). Upon prioritizing the media sessions, we run our classifier component on high priority media sessions more frequently and postpone the classification and processing of low priority media sessions until a later phase, thus achieving responsiveness without sacrificing recall performances.

In this section, we outlined the motivation and design of introducing our dynamic priority scheduler into the proposed cyberbullying detection system. In Section 6.2.2, we first determine what threshold is appropriate for our DPS by comparing it with our baseline scheduler round-robin

scheduler, where all media sessions always have the same priority and, so, monitored equally. Then we demonstrate through thorough experiments that DPS introduces significant responsiveness gain over round-robin scheduler, thus justifying the utility of this component.

6.1.3 An Example

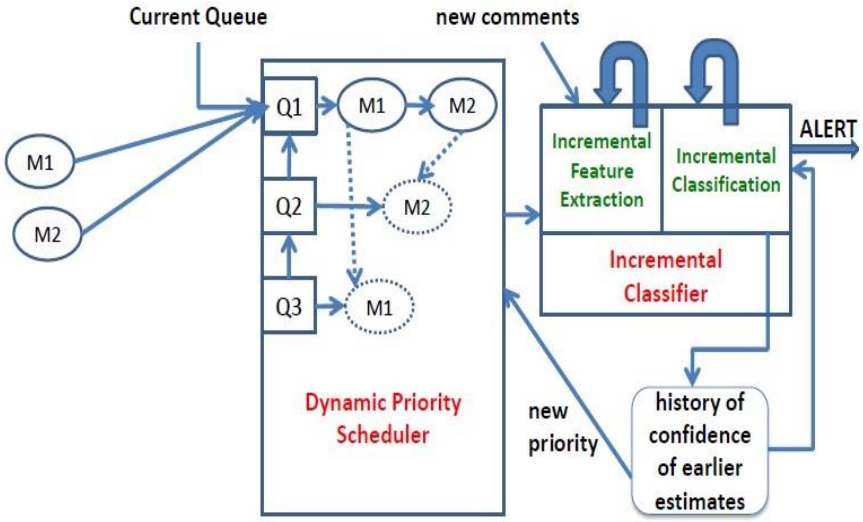


Figure 6.1: Scalable and responsive cyberbullying detection architecture

Consider the example shown in Figure 6.1, where M1 and M2 represent newly created media sessions from Vine in real time. Three separate queues, Q1, Q2, and Q3 are maintained. The scheduler schedules media sessions in queue Q1 (pointed to by current head) for processing one by one in the queue order. After a media session has been processed by the cyberbullying detector component and if no alert is raised, it is placed at the end of either queue Q2 or queue Q3 depending on the new priority assigned to it (discussed in the next subsection). If the session's new priority is high, it is placed in queue Q2, and if the session's new priority is low, it is placed in queue Q3. When all media sessions in queue Q1 have been processed, queue Q2 becomes queue Q1, queue Q3 becomes queue Q2, and queue Q3 becomes empty. In the example shown in Figure 6.1, initially M1 and M2 are assigned high priority and placed in queue Q1. M1 is scheduled first and is processed by

Table 6.1: Comparison of different classifiers using the 983 Labeled Media Sessions

Classifier	Precision	Recall	Time (s)
AdaBoost	0.80	0.72	228
Logistic Regression	0.78	0.76	44.42

the cyberbullying detector component. After this processing, it is assigned a low priority, and so is added at the end of queue Q3. M2 is scheduled next and is processed next. After this processing, it is assigned a high priority, and so is added at the end of queue Q2. At this time queue Q1 is empty, and queue Q2 becomes queue Q1 and queue Q3 becomes queue Q2. This process then continues. Since lower priority processes are eventually elevated into higher level queues, our solution ensures that no media session will starve.

6.2 Performance Evaluation

6.2.1 Incremental Classifier Evaluation

In chapter 4, the AdaBoost classifier was reported to have the best performance based on accuracy, precision and recall values. Table 6.1 compares AdaBoost with logistic regression in terms of precision, recall, and running time. The features AdaBoost classifier used were number of followers and followings, likes and views for media sessions, media caption polarity and subjectivity [71], total number of negative comments, summation of negative comment polarity and subjectivity, total individual comment polarity, total individual comment subjectivity, total negative words, total number of negative comments and unigrams. For Logistic Regression, the features we used were the number of followers, followings, media caption polarity and subjectivity, total individual comment polarity, total individual comment subjectivity, total negative words, and the total number of negative comments. We made sure that the features used by logistic regression were incrementally linear by nature, as noted in Section 6.1.1. The performance values showed in Table 6.1 were obtained using 10-fold cross-validation on the labeled Vine data from chapter 4. We notice that although the AdaBoost classifier achieves a slightly higher precision, logistic regression achieves

higher recall. Furthermore, the running time of logistic regression classifier is significantly less, more than five times faster than that of the AdaBoost classifier. The reason for this is twofold. First, AdaBoost needed unigram features to achieve a high precision and recall, but unigram feature extraction is computationally expensive. In comparison, logistic regression is able to achieve effectively the same precision and better recall while using features that are much more lightweight to compute, thus yielding much lower running time. Second, the AdaBoost classifier is a meta-estimator that begins by fitting a classifier on the original data-set and then fits additional copies of the classifier on the same data-set but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases [97], thus making it computationally expensive compared to much simpler logistic regression. Based on these analyses, we chose the logistic regression classifier for detecting cyberbullying in our solution. **It is worth mentioning that we have employed other classifiers based on different combinations of features too (Decision Tree, Random Forest, Naive Bayes, Perceptron etc)** and only present the classifiers and feature combinations that yielded the best results.

Next, we show that leveraging incremental approach into our logistic regression classifier significantly improves the scalability of the execution stage. To this aim, we defined a baseline solution as consisting of non-incremental feature extraction and a non-incremental logistic regression classifier. As new comments arrive for the baseline solution, it would need to recompute all feature vectors from scratch and recompute the entire logistic regression from scratch. We compared the total running time of the baseline solution with an incremental solution that implemented both incremental feature extraction and incremental logistic regression, as described in Section 6.1.1. We note first that our measurements showed that the fraction of time taken by the logistic regression compared with the feature extraction time was negligible so that the total running time was dominated by feature extraction.

Figure 6.2 shows the average time taken for the standard and the incremental classifiers as the number of comments increases in media sessions. To simplify the plot, we group the comments in sets of 10. The time taken by the standard classification solution goes up almost linearly with the

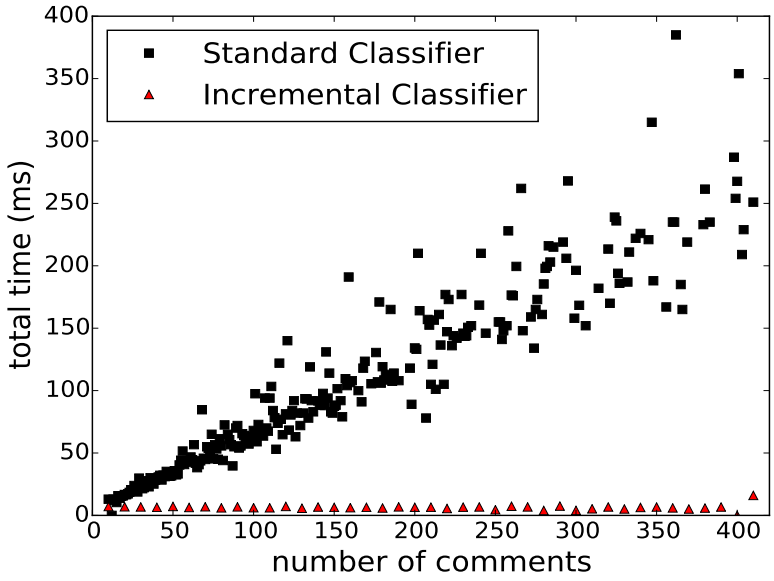


Figure 6.2: Total time taken by standard and incremental classifiers as new comments come in per media session.

number of comments in the media session since the standard solution must recompute all features and regression weights. On the other hand, the time taken for the incremental classifier is basically constant every time a set of 10 additional comments come in because it only has to compute the feature values for the additional 10 new comments. The justification for using 10 comments-set is given in Section 6.2.2.

6.2.2 Dynamic Priority Scheduler Evaluation

In this section, we compare performances of our DPS with round-robin scheduler, a scheduler with no assignment of priority. The aim of performing this comparison is twofold. First, we want to show the gain of responsiveness our scheduler achieve over round-robin scheduler. Second, we also want to investigate several design choices crucial to building our scheduler to decide upon the best choice based on the metrics of performance gain each design choice achieves over round-robin scheduler. We use labeled Vine dataset from chapter 4 to perform the experiments.

First, we need to determine what threshold is appropriate for our solution, as noted in 6.1.2.

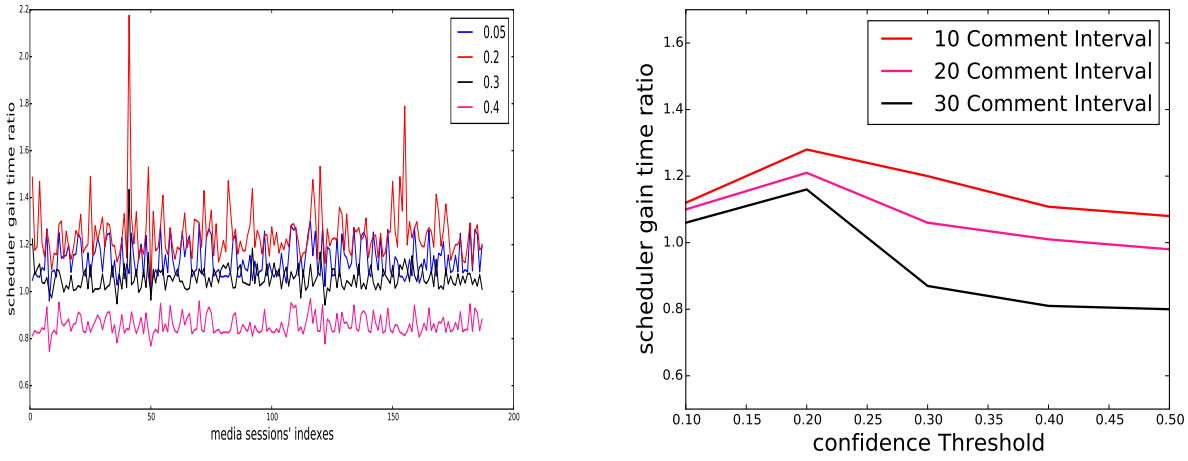


Figure 6.3: Left: Scheduler gain time ratio for different confidence thresholds for labeled cyberbullying media sessions. Right: Scheduler gain time ratio for different confidence thresholds by using different comment increment sizes for different number of media sessions using Vine labeled data from chapter 4.

We also need to determine with what granularity our classifier ingests batches of new comments because this affects the time to first alert. The scheduler will choose a high priority media session to pass to the classifier. In the time between classification attempts, a media session may receive N new comments. If all N comments are input to the classifier at once, and N is quite large, we may delay recognizing cyberbullying, i.e., a burst of negative comments may be swamped by the other positive comments. Therefore, we need to consider comments in small enough batches or intervals so that the classifier can catch cyberbullying with finer granularity and raise the alert early.

The left figure in Figure 6.3 assesses which combination of threshold and interval size produced the best improvement in response time using dynamic prioritization compared with a simple round-robin policy. The round-robin scheduler is defined as one where media sessions are not assigned any priority, and the scheduler simply rotates through all media sessions, with no particular attention being paid to likely cyberbullying sessions.

As can be seen from the figure, **by using a confidence threshold of** 0.2 **and comment increment size of** 10**, we were able to gain the maximum responsiveness over the round-robin scheduler**. We think this is because as the comment increment size goes up to 20 or 30,

the burst of cyberbullying comments can get nullified by the other positive comments, which in turn influences the features(i.e. summation of individual comment sentiments) that are used by our incremental classifier. So 10 comment increment size tends to be the optimal size for having enough context of a comment thread to make a knowledgeable decision about cyberbullying while also not being too big to risk being nullified by other positive comments. The table also confirms that the confidence threshold of 0.2 offers the best speedup for our dynamic priority scheduler. For example, if a media session m has been classified 3 times at t_1, t_2, t_3 with classification decisions not-bullying, not-bullying, not-bullying respectively with confidence values of $0.85, 0, 85, 0.55$, this means even though it has been classified as not cyberbullying, the confidence values of cyberbullying decision is also increasing $(0.15, 0.15, 0.45)$ which makes it a potential candidate for a future cyberbullying session. So we take the average of the previous classification confidence values of cyberbullying class (0.25 in this case) and see that the average confidence value is more than 0.2 and change the priority of this media session as high and insert it in the dynamic priority scheduler. We mention that we tried all possible combinations of comment chunk sizes and confidence thresholds and only present those combinations that yielded the best results. The right figure in Figure 6.3 demonstrates the gain time achieved by our scheduler for each media sessions from the Vine dataset labeled as cyberbullying. This figure further justifies the choice of confidence threshold of 0.2 in our scheduler. These experiments helped us to not only justify our choice of using DPS but also helped us to decide upon the crucial design choices of using confidence threshold of 0.2 and 10 comment increment size.

6.2.3 Alert Performance

Since each media session will be passed sporadically to the classifier by the scheduler, the classifier will generate a sequence of cyberbullying detection decisions over time for each media session. It is therefore worth considering to what extent we should utilize the history of detection decisions in generating the alert. The default is to generate an alert immediately after the classifier decides that the current batch of 10 comments, in combination with earlier content, constitutes

Table 6.2: When to Send Alert

Number of Predicted Cyberbullying Comment Sessions in the History	Precision	Recall
≥ 1	0.68	0.71
≥ 2	**0.71**	**0.71**
≥ 3	0.71	0.71

cyberbullying. However, we wish to be sure and avoid false positives. One option is to decouple the alert from the classification and delay the alert until N positive decisions have been recently seen. This design gives us some flexibility in trading off responsiveness and precision.

For each media session, we maintain an array storing the results of each classification result of that session along with the time of that classification. We use this array to decide when to raise an alert. In particular, we set a threshold value, which is the number of times a media session has been classified as cyberbullying since the last time an alert was raised for that session, or from the beginning if no alert has yet been raised. After experimenting with different number of threshold values(6.2), we find that by raising an alert only when we have at least 2 decisions for cyberbullying since the last time an alert was raised, we achieve the best precision and recall of 0.71 and 0.71 respectively, thus reducing the number of false alarms. This performance is a marked improvement over the Static Priority Scheduler (SPS) described in Section 6.1.2 that had a recall of only 58%. Moreover, the recall is, in fact, an improvement over the standard classifier's 0.66 (See Table 6.1 for comparison). This marked improvement of recall over two baselines (SPS and standard classifier) demonstrates the justification of using incremental classifier component and dynamic priority scheduler along all the associated design choices: that these components are efficient and responsive and also retains sufficient classifier performance when compared to the current state-of-the-art.

6.2.4 Scalability Evaluation

In this section, we first demonstrate the way our proposed system scales when it has to deal with a substantial number of media sessions. For this purpose, we first deploy Amazon AWS free tier 1GB memory virtual machine instances to start monitoring media sessions, implementing both our proposed scheduler and round-robin scheduler. An acceptable responsiveness of the system is our primary goal along with the scalability of the system. In these experiments, we decided that an average alert time under 2 hours is acceptable, which means an alert will be within 2 hours of a cyberbullying instance. We acknowledge that this decision is purely because of the lack of research in this particular area. In the future, we will conduct an elaborate survey to explore the acceptability of a potential cyberbullying system's responsiveness. For the experiments presented in this section, we replicated the 100000 media sessions' traffic from the dataset from chapter 4 up to the scale of 39 million. Those media sessions were gathered by performing snowball sampling after selecting a random seed. We believe the randomness of the seed selection, snowball sampling, and large number (100000) of media sessions in this dataset should enable the scaled up traffic to reasonably approximate the behavior of the overall network.

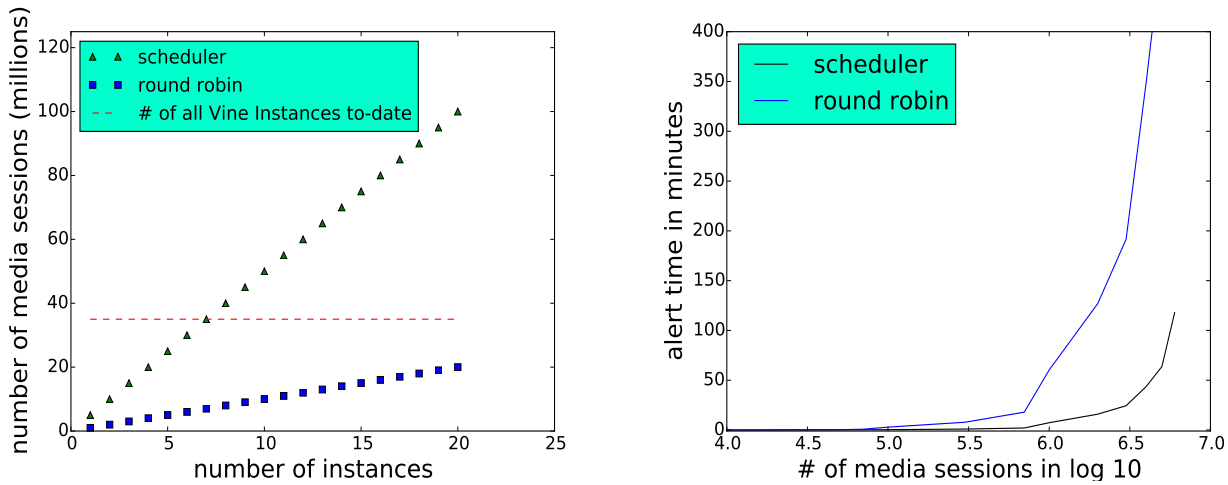


Figure 6.4: Left: Number of media sessions vs number of instances needed to monitor them, keeping average alert time under 2 hours. Right: Average alert time vs number of media sessions for round-robin and dynamic priority scheduler

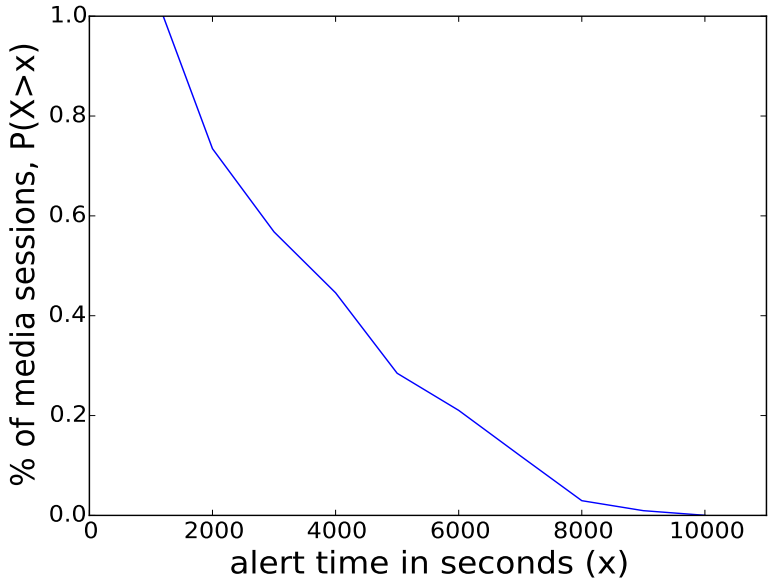


Figure 6.5: CCDF of alert time for 5 million media sessions in 1GB memory amazon AWS instance using dynamic priority scheduler

The left graph in Figure 6.4 shows the number of media sessions that can be processed as the number of AWS instances is increased, keeping the average alert time under 2 hours. This figure shows that the number of media sessions that can be processed increases linearly with the increase in the number of instances, and that our system scales *five times better* than the round robin scheduler. Given that the Vine social network generated about 39 million media sessions since 2012 [102], our system is capable of monitoring Vine-scale social networks with only 8 AWS instances, keeping the average alert time below 2 hours. In contrast, a round-robin scheduler would require upwards of 40 instances.

The right graph in Figure 6.4 shows the average alert time for round robin and dynamic priority scheduler versus the number of media sessions. It can be seen from the figure that we are able to process 5 million media sessions with our proposed system with an average alert time under 2 hours whereas, for round robin, it is 1 million. To show the cost of using our dynamic priority scheduler in terms of worst case scenario, we see in Figure 6.5 that around 10 percent of media sessions get their alerts after 2 hours. This is the cost we pay for postponing the processing of low

priority media sessions in our scheduler.

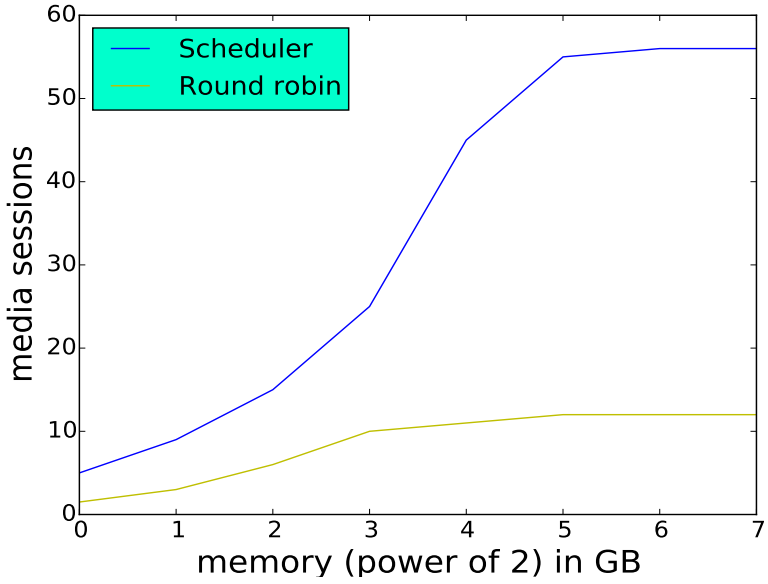


Figure 6.6: Memory vs number of media sessions in millions

Next, we investigate the resources our system will need to monitor a comparatively more popular social network like Instagram. Since its establishment in 2010, Instagram has accumulated over 40 billion media sessions [67], which means almost 6 billion media sessions per year, a number much larger than Vine's 39 million [102]. To accommodate such a high volume social network, based on Figure 6.4, we would need 1,200 AWS instances to keep the average alert time under 2 hours every year. To scale up for such a load, we have two choices. We can either spawn off 1,200 such instances of 1 GB memory or we can increase the memory of our instances to process more media sessions. To further investigate the memory performance, we implemented our system with different memory-sized instances belonging to Amazon AWS services. Figure 6.6 shows the number of media sessions processed by each instance of a particular memory size. The number of media sessions in the Y-axis illustrates the highest number of media sessions that can be processed by that instance without having an average alert time of over two hours. The figure demonstrates that, while increasing memory does help increase the media session monitoring capacity, at a certain point around 32 GB/instance, additional memory no longer enables additional monitoring of media

Table 6.3: Total Time Comparison for Different Approaches and Different Number of Media Sessions (seconds)

Approach	10000	50000	100000
AdaBoost	5674	26784	-
Logistic Regression	1110	5320 (5X)	10438
Incremental Classifier	22	120 (223X)	206

sessions. That is, the graph plateaus around 50 million media sessions so that there is no additional benefit to using 64 GB or 128 GB instances compared to 32 GB instances. We hypothesize that this behavior is due to computation becoming the main bottleneck rather than memory. Therefore, to monitor Instagram-scale social networks, we would need approximately 120 32 GB instances. Note that without our dynamically scheduled incremental classification system, approximately 600 32 GB instances would be required, almost five times as many, as it can be seen from Figure 6.6.

For evaluating the incremental classifier's scaling performance, we compare three types of approaches, namely the best reported Vine cyberbullying classifier from chapter 4 (Standard AdaBoost), Logistic regression without incremental feature extraction or classification (Standard Logistic regression), and Logistic regression with incremental feature extraction and classification (Incremental Classifier). Table 6.3 shows the time needed in seconds for these three approaches to process different numbers of media sessions. The table clearly demonstrates that the choice of using incremental classifier helped us to improve classification time by 223 times faster than AdaBoost for 50,000 media sessions and 5 times faster than the standard logistic regression. For this evaluation purpose, we used the Vine dataset provided in from chapter 4.

Next, we compare the responsiveness of our dynamic priority scheduler against the unprioritized round robin scheduler. The metric we use to compare these two approaches is responsiveness gain, meaning the ratio of time taken by the round-robin scheduler over the time taken by our dynamic priority scheduler to raise an alert. Figure 6.7 shows the responsiveness gain vs. number of media sessions. The gain tends to increase as the number of media sessions goes up, reaching almost 7 times faster responsiveness for $100,000$ media sessions. This improvement

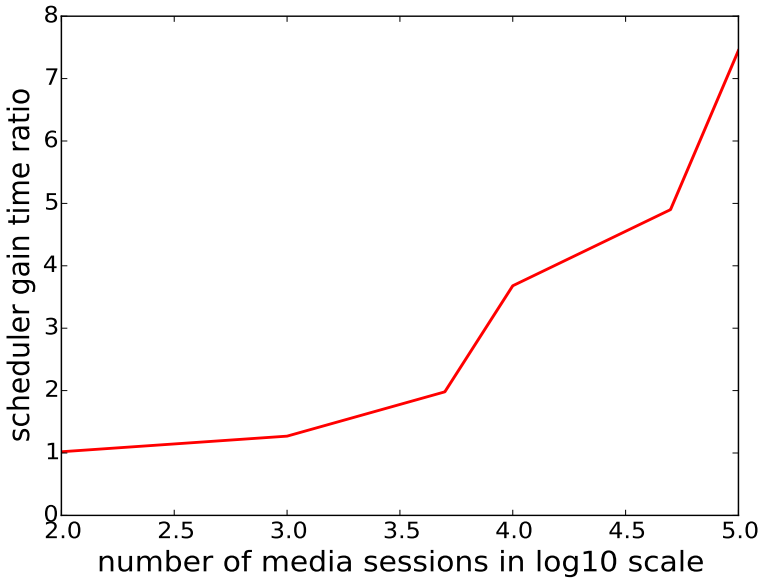


Figure 6.7: Gain time ratio (round-robin scheduler/dynamic priority scheduler)for different number of media sessions.

is due to the fact that our scheduler tends to process cyberbullying media sessions first whereas round robin processes all media sessions at each pass. For this reason, as the number of media sessions grows, so does the improvement of using our dynamic priority scheduler due to its priority processing of cyberbullying media sessions.

For further insights into resource scalability, we present activity graphs of the media sessions from Vine. We investigate the distribution of how long a bullying media session takes to receive its first comment. Figure 6.8 shows that very few bullying media sessions receive their first comment after 500 hours since session creation. Hence, one way to improve scaling is to stop monitoring any session that takes longer than 500 hours for its first comment. Secondly, Figure 6.9 shows the CCDF of activity of media sessions in Vine. A fair percentage of media sessions receive comments even after 10000 hours after initial media posting. In comparison, bullying media sessions received all their comments within 9000 hours, i.e. within a year of their creation. So another way to improve scaling would be to purge out all media sessions that are one year old.

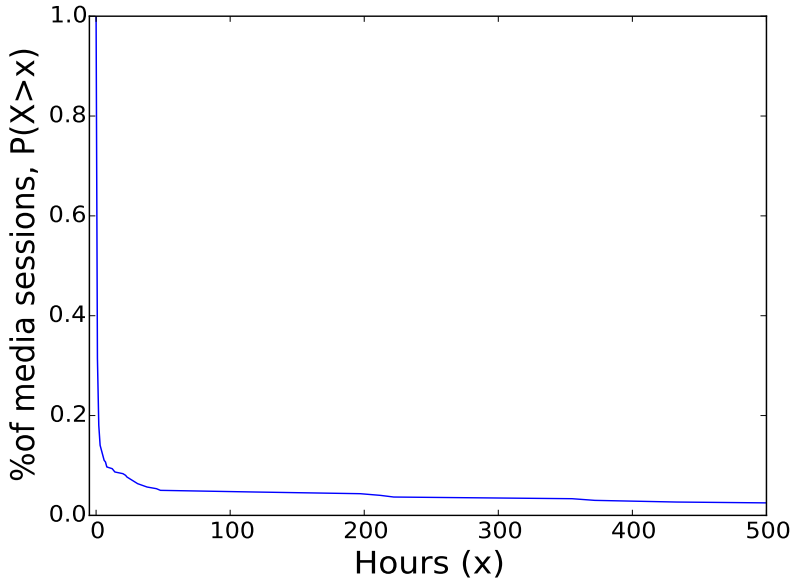


Figure 6.8: CCDF of Time Interval in hours until First Comment For Bullying Media Sessions

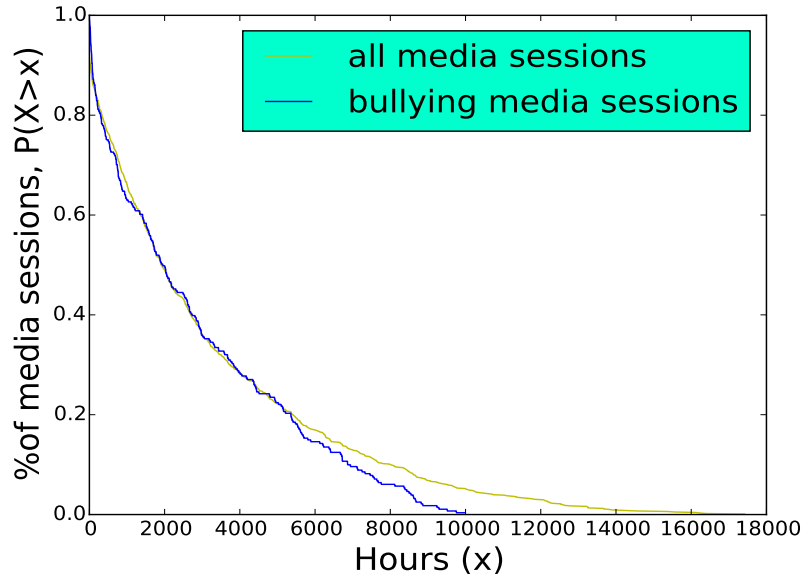


Figure 6.9: CCDF of activity for all and bullying media sessions

6.3 Conclusion

In this work, we have developed a cyberbullying detection system for media-based social networks, consisting of a dynamic priority scheduler and a novel incremental classifier. The evalua-

tion results show that our system substantially improves the scalability of cyberbullying detection, enabling five times more media sessions to be monitored for the same average alert time of 2 hours compared to an un-prioritized system. Moreover, we demonstrate that our system can fully monitor Vine-scale social networks for cyberbullying detection for a year using only eight 1 GB AWS VM instances. We discover the point (32 GB) at which adding memory no longer enables monitoring of more media sessions and project that our system would need 120 32 GB instances to fully monitor Instagram-scale traffic for cyberbullying.

As part of future work, we propose to investigate the plateauing effect that limits the effectiveness of adding more memory, namely that there is likely a computational bottleneck that needs to be further addressed. Portability of our system and the design choices we made (confidence threshold of 0.2 or 10 comment increment sizes) to other OSNs similar in nature such as Instagram and different in nature such as Facebook, SnapChat etc is also a future research topic. Finally, we will delve deeper into the design choices (confidence threshold of 0.2 or 10 comment increment sizes) that we made while implementing our system to have a better understanding of their better performances.

6.4 Acknowledgments

This work was supported by the US National Science Foundation (NSF) through grant CNS 1528138.

Chapter 7

BullyAlert- A Mobile Application for Guardians to Enable Adaptive Cyberbullying Detection

In the previous chapter, we tackled the two challenges, namely scalability, and responsiveness of a potential cyberbullying detection system and evaluated its performance results against two social networks (Vine and Instagram). However, there are two key issues that a potential centralized cyberbullying system presented in chapter 6 faces:

- A centralized cyberbullying detection system still needs to have a lot of computing resources to be able to accommodate all the OSNs in the world. So, even though the system presented in chapter 6 is scalable for two or three social networks, a solution that accommodates all the OSNs currently available will be faced with a daunting challenge of meeting the computational-resource requirements to maintain sufficient responsiveness
- The system presented in chapter 6 makes use of a cyberbullying classifier component that is applied generally to all the guardians. We argue that different guardians will have different tolerance levels, which in turn, might be dependent on their personal preferences, demographic information, location, age, gender and so on. So, a system that allows different levels of cyberbullying alerts to be sent to parents based on their individual tolerance levels for cyberbullying is the most natural solution.

In this section, we present the design and implementation of an Android mobile application for guardians: BullyAlert. This mobile application allows the guardians to monitor the online

social network activities (currently only Instagram) of their kids and get notifications whenever the monitored social networks receive a potential cyberbullying instance. Reasons for developing this mobile application are twofold. First, it allows us to delegate classifier computations of cyberbullying detection to the hand-held devices of the guardians, thereby reducing the computational resources needed for a potential centralized cyberbullying detection system. Second, BullyAlert allows the guardians to give the resident classifier feedback about how right or wrong each notification is. The resident classifier then updates itself accordingly to calibrate its tolerance level with that of the guardian using it. This mechanism allows for personalized cyberbullying notification of an individual guardian.

We make the following contributions in this chapter.

- We propose the design and implementation of an android application, BullyAlert.
- We present a preliminary user-experience analysis of the guardians who downloaded the mobile application by using the current crop of data
- We present a preliminary comparison the behavior of the users who were being monitored by the guardians with the general population of Instagram to derive some initial key insights by leveraging the current collection of data

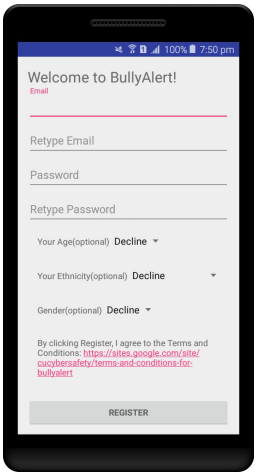
7.1 System Design and Implementation

This section presents the design, implementation, and architecture of BullyAlert. We begin by describing the typical user work-flow through a series of use cases, and then present the architecture and implementation of BullyAlert.

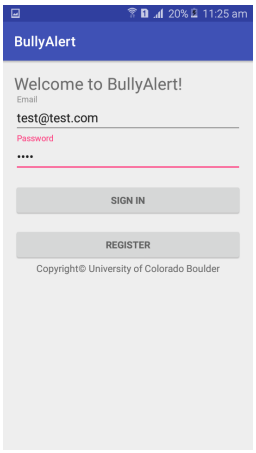
7.1.1 Use Cases

7.1.1.1 Guardian registers

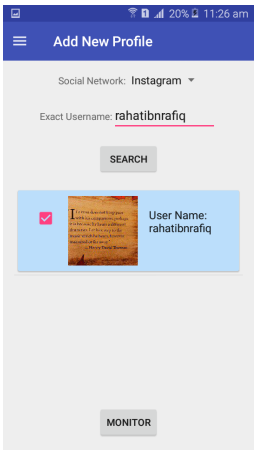
After a guardian downloads the application and opens it, the screen in Table 7.1a is presented. The Guardian has to enter a unique email id and password to be able to register into our system.



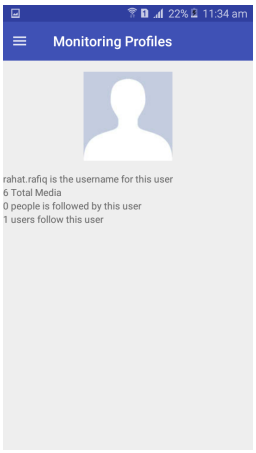
(a) Register



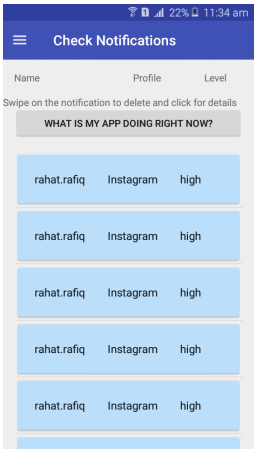
(b) Log In



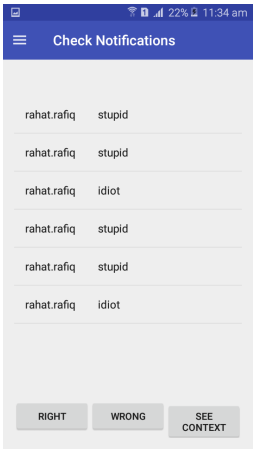
(c) Add Profile



(d) See Monitoring User Details



(e) Notification List



(f) Feedback

Table 7.1: BullyAlert Application

The guardian will also have the option to provide three demographic information: age group, gender, and ethnicity. Each of this information can be selected by a drop-down list. The Guardian has the option to decline giving these information as well.

7.1.1.2 Guardian logins

The login screen shown in Table 7.1b has two input fields. These fields ask for the email and password used to register into the system. After clicking the log-in button, the guardian is directed to the dashboard of the application.

7.1.1.3 Guardian Searches for users to monitor

The Guardian is able to search for public Instagram profiles by going to the user search component shown in Table 7.1c. At first, the guardian selects the social network profile from the drop-down list. *Right now, we are only supporting public Instagram social network profiles.* Then in the text field, the username of the user to be monitored is typed. When the search button is clicked, a list of users matching the username entered is shown along with the associated profile pictures for the guardian to facilitate a better identification. To start monitoring a profile, the guardian has to select the profile and then click the monitor button.

7.1.1.4 Guardian examines user profile information details

The Guardian is able to see the basic profile details of the users being monitored, as shown in Table 7.1d. This page shows guardians the current profile picture, number of total media shared and number of total followers and followings of the user being monitored.

7.1.1.5 Guardian gets a list of notifications

Table 7.1e shows the screen that the guardian sees when a host of notifications are present in the dashboard. The list has three columns, the first column shows the username of the profile where this cyberbullying notification has originated, the second column indicates the social network

and finally, the third column outlines the application's classifier's perceived level of severity for this notification. Currently, the application has two levels of severity, namely low and high. The Guardian is able to see the details of the notification by clicking the individual notification boxes.

7.1.1.6 Guardian examines notification details and give feedback

To enable the guardians to see the full context of a particular notification and give feedback as to whether the application right or wrong in terms of the severity level, table 7.1f is presented. The Guardian has options to click the "see full context" button which will then load not the just the latest comments but also the previous comments of that media session. This enables the guardians to get a full picture of the happenings in the media session. The guardian can give feedback through the two buttons, namely right or wrong. This feedback is then used by the application to calibrate its tolerance level according to that of the guardian.

7.1.2 Architecture and Implementation

This section describes the architecture and implementation of the BullyAlert application's different components. Figure 7.1 presents the architecture diagram of the BullyAlert system. The guardian communicates with the BullyAlert application for registering, logging in and getting notifications for potential cyberbullying instances. The application sends guardian data, notification data, and feedback data to the BullyAlert server. The application also contacts the BullyAlert server for authenticating a user log-in. Moreover, the application implements a polling mechanism by which it periodically collects media session data of the Instagram-users (who are being monitored by the guardians) from the Instagram servers.

7.1.2.1 BullyAlert Server

BullyAlert server is responsible for the following:

- During the registration process, it is responsible for checking that the registration information is verified. It first checks if the email that is being used to register is unique in the

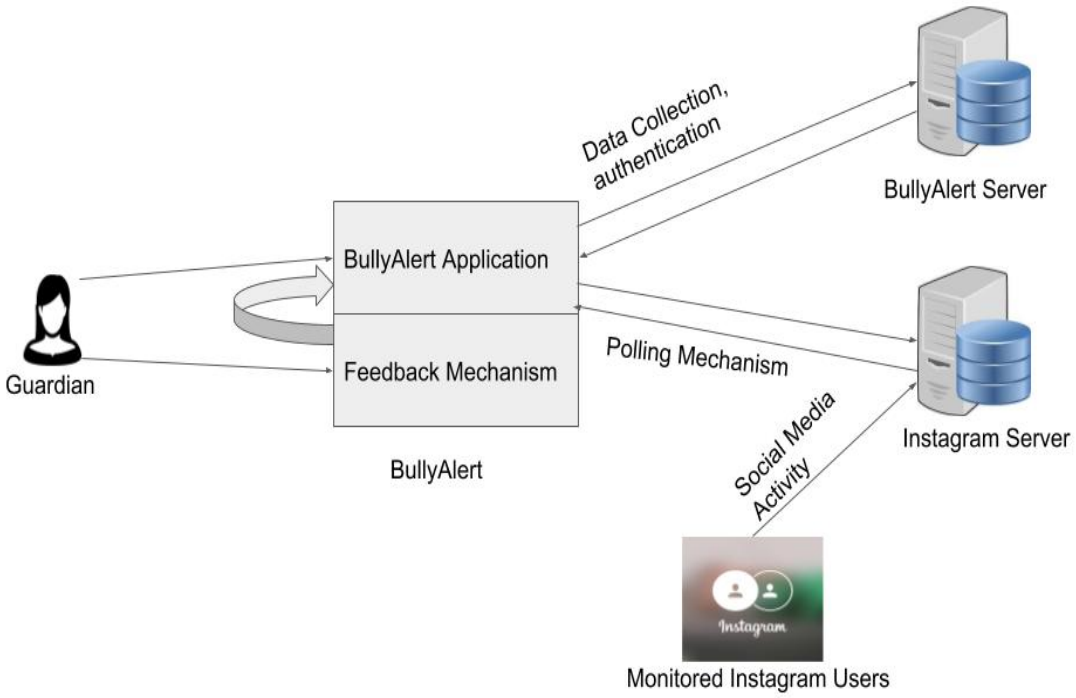


Figure 7.1: BullyAlert Architecture

system and the password is at least of six characters. When the registration is successful, the server stores both the username and the password in encrypted format. In addition to this, the server stores the demographic information provided by the guardian through the registration form, such as age group, gender, and ethnicity.

- During login, the server is responsible to check the login credentials of a guardian with the system's stored credentials. It gives an error if the login credentials are not verified which is then in turn shown to the user.

- Storing all the notifications of the guardians. Every time a guardian receives a notification, the application sends the notification meta-data to the server. The data consists of a list of comments for which the notification was raised, the user-name in whose profile the notification was raised and the severity level of the notification (high or low).

- Storing the feedback that a guardian gives to the application for a particular notification.

Every time the guardian gives the application a feedback (right or wrong), the application sends the relevant data (list of comments, the application's perceived severity level, the guardian's feedback) to the server.

- Storing each guardian's resident classifier information. This is to facilitate continuity of the guardian's classifier so that when the guardian uninstalls the application, the classifier will keep being stored in our server. This means that, if the guardian, at some later time, chooses to re-install the application, the old classifier will become the classifier of the application instead of the general one.

We have used MongoDB, RESTful API and node.js for implementation of this component. The code for this can be found in [91].

7.1.2.2 Adaptive Classifier

An adaptive classifier for each individual guardian is more suitable for our application than a general classifier for every guardian because of the potentially subjective nature of cyberbullying. We hypothesize that each guardian will have their own tolerance level when it comes to cyberbullying, which in turn, can be dependent on several factors, such as gender, age, race etc.

The ways we develop this adaptive classifier are as follows. First, we incorporate a feedback mechanism in our application by which, the guardians, upon receiving a potential cyberbullying notification, will be able to give us a feedback saying how right or wrong the notification is. **We also show the guardians a list of other media sessions which were not deemed as bullying by our application, so as to make sure we also get feedback for media sessions which were not in the potential bullying notifications page. This is to enable the application to keep track of the false negatives in addition to false positives**. Second, we use the logistic regression classifier from chapter 6 for the implementation of the application's resident cyberbullying detection component. Every time an instance of the classifier gets a feedback, the feedback data encapsulate the media session's list of comments for which the alert was raised and

the guardian's label (right or wrong). This datum is considered as a labeled training data for the resident classifier. Feature values described for logistic regression classifier in chapter 6 are extracted for each of these feedback sessions. Upon converting the feedback data into training data, we then perform stochastic gradient descent [89] for the resident classifier. Each parent's individual classifier then reaches a different local optimum, thereby facilitating the adaptive nature of the classifier.

For the guardians whose numbers of feedback are not substantial enough to perform an individual adaptation process, we implement the following. We first collect all the feedback given by all the guardians in our server. Then, based on all these feedback, we update our general classifier that was used by the guardians when they first install our application. We call this **updated general classifier**. This updated general classifier is then propagated to the guardians who don't have enough individual feedback to make sure their classifiers are updated as well. The implementation code can be found in [92].

7.1.2.3 Polling Mechanism

The polling mechanism is responsible for the following:

- When the guardian searches for a particular user by username, this mechanism fetches the user profiles of which the username-string is a match.
- After a monitoring request of a user by a guardian is approved, the polling mechanism starts polling that user profile every hour for any new posts. This is to make sure the app is updated with the latest media postings of the monitored user.
- In addition to polling for newly posted media, this component is also responsible for getting the newest comments for all the media posted by the user. Every time a host of new comments is posted for a media session, this mechanism fetches those new comments and sends this newest media session data to the adaptive classifier component for classification.

Table 7.2: Data Collection

guardian	username, password, (ethnicity, age group, gender) if provided.
monitored user	username, social network, number of followers, followings, shared medias
notification	username for the profile, number of likes and comments for the media session in question, resident classifier's perceived severity level (high or low), list of comments for which the notification was raised
feedback data	username for the profile, number of likes and comments for the media session in question, resident classifier's perceived severity level (high or low), list of comments for which the notification was raised, guardian data, guardian feedback (right or wrong)
classifier data	application's resident classifier's feature vector's coefficients

7.2 Data Collection

Data collected from the application is stored in our server. Proper encryption methodology is used to make sure sensitive data such as user password are protected. We collect guardian data, monitored user data, notification data, classifier data, and feedback data, as shown in Table 7.2.

7.3 User Data Analysis

(a) age group (b) gender

Table 7.3: BullyAlert Guardian's Demographic Data in Pie Charts

This section presents a preliminary analysis of the data collected until now from BullyAlert. First, it explores the guardian data and then it performs a comparison of social network behav-

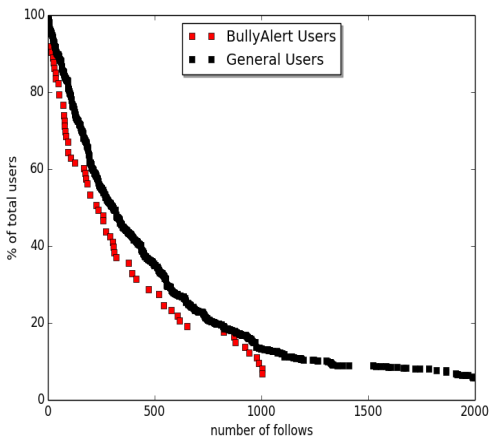
iors between the general Instagram population 4 and the users who were being monitored by the guardians who downloaded our application. **We acknowledge that at the point of writing this thesis, the amount of data we have is not sufficient to derive concrete findings. We are mainly interested in presenting a preliminary analysis of the current collection of data in order to gain some initial insights**. In future, upon collecting a more substantial amount of data, we plan to perform more comprehensive and thorough analyses.

When a guardian registers into our system, in addition to the email and password, we also ask them to provide their gender, ethnicity and age information, if they choose to divulge those. Table 7.3a and 7.3b show the distribution of gender and age-group of the 100 guardians who have downloaded BullyAlert until now. From the distributions, it is fairly clear that a substantial portion of the people chose not to provide the demographic information, 31 and 34 percent for age group and gender respectively. In addition to that, the most prominent age group and gender where 36-45 and female respectively.

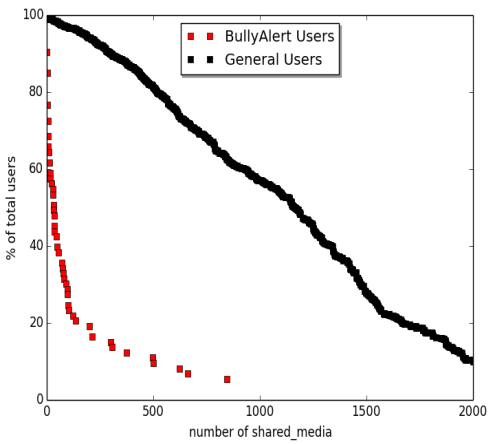
The reasons we collected this demographic information are twofold. First, when we start to get the classifier feedback data for all the guardians for their different tolerance levels, we will want to investigate if there are any correlations between different guardians' tolerance levels and their demographic information. Second, if we do find that people with same demographics tend to have same tolerance levels, we will then want to be able to build different general classifiers for different clusters where each cluster hosts guardians with similar demographics. **While we acknowledge that these 100 guardians' data is an insufficient representation of guardians, we postulate that this preliminary demographics distribution still introduces us to a new systems challenge: what about guardians who do not provide demographic information and thus will not belong to any particular cluster by default?**

Next, we investigate a comparison analysis between the Instagram users who are being monitored by our application and the general Instagram population, a data-set collected from [68]. First, we compare both sets of users' follow and media-sharing activity. Table 7.4a and 7.4b show the CCDF of both set of users' number of people they follow and number of medias they have

shared in their profile. It can be seen that the follows activity of both sets follow the same pattern, which is understandable because a user can only follow so many people. But there is a discernible difference in the media sharing activity. The general population's line tends to fall far slower in the graph than that of the BullyAlert-monitored users, with almost 80 percent of the BullyAlert users having less than 100 shared medias in their profile. This means that the users who are monitored by the guardians tend to be not as active as the general population. This particular observation also poses an interesting system perspective. Because most of the people who are likely to be monitored by our application will not be sharing as much medias, **we can afford to incorporate some sophisticated machine learning classifiers in the application instead of worrying about responsiveness, discussed in chapter 6**. Again, we like to emphasize here that these are preliminary derivations drawn from our current small set of collected data.



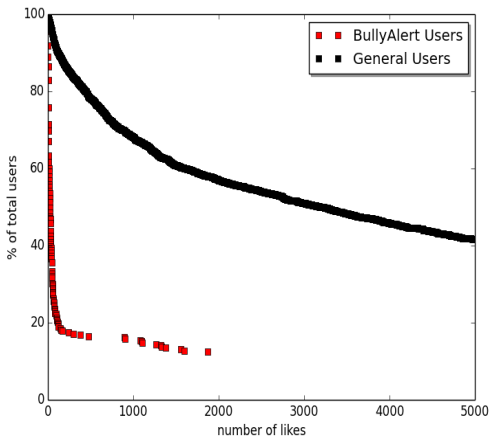
(a) CCDF of follows



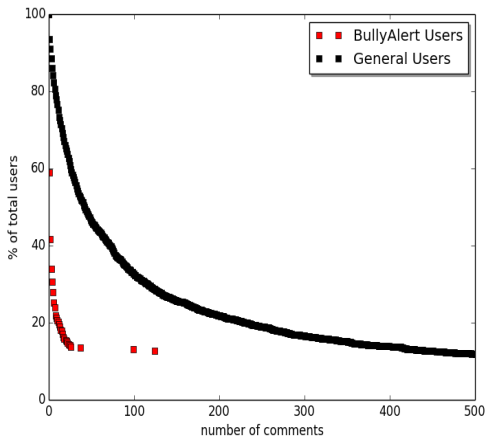
(b) CCDF of shared media

Table 7.4: Comparison between Monitored users of BullyAlert and Instagram population collected in [68]

In continuing the narrative, we also put forth a detailed analysis of activities of other people in the user's profile, for example, the likes and comments received in the shared media sessions. Table 7.5a and 7.5b show the CCDF of the number of likes and comments received for the media sessions for both set of Instagram users. It can be seen that the media sessions shared by the users being



(a) CCDF of likes



(b) CCDF of comments

Table 7.5: Comparison between Monitored users of BullyAlert and Instagram population collected in [68]

monitored are far less active in terms of getting likes and comments than their general counterpart. **This further solidifies our system perspective that our application's classifier will have fewer data to take care for, thus the classifier does not have to be as lightweight as described in chapter 6, based on our current crop of data.** We acknowledge that the current crop of data is not enough to make a decision, so we plan to keep collecting the data to solidify this preliminary insight. Right now, we just present initial analyses with the data we currently have.

7.4 Conclusion

In this chapter, we make the following contributions. First, we outline the motivation and design of a mobile application, BullyAlert, that adapts itself according to individual tolerance level for cyberbullying of the guardian. Second, we present a thorough architecture description of the components implemented to develop BullyAlert. Third, we provide a preliminary user analysis of both the guardians and the users being monitored by the application, and in the process, present several potential system issues/ challenges/perspectives using our current crop of data.

In future, first, we plan to collect notification and feedback data from the guardians as well as

recruit more guardians to have a reasonable data-set. Second, we will investigate our hypothesis that different people tend to have different tolerance levels of cyberbullying by leveraging the feedback data. Third, we will explore, in depth, the feedback behavior to make sure that our classifier is indeed adaptive. This will mean that, for each parent, the classifier will eventually calibrate its tolerance level to that of its host, so the feedback from the guardians will eventually reach "almost all rights". Fourth, we also plan to examine other sophisticated classifiers in the application and compare their performances against the current logistic regression classifier.

7.5 Acknowledgments

This work was supported by the US National Science Foundation (NSF) through grant CNS 1528138.

Chapter 8

Going Beyond Textual Features: Video and Topical Features for Cyberbullying Detection

In chapter 4, we developed a classifier that takes into account user, media and comment features. In chapter 6, we leveraged these lightweight features to develop a highly scalable and timely classifier for cyberbullying detection. We also embedded this fast classifier into the BullyAlert android application, as delineated in chapter 7. In chapter 7.3, we mentioned that (based on the preliminary user analysis), because the application will only be monitoring a few users who are not as active as the general Instagram population, the application's resident classifier can afford to be more sophisticated than the current logistic regression without sacrificing much of the performance metrics. To this aim, in this chapter, we present a short analysis of more complicated feature extractions, such as video (emotions displayed and content shared) and topical features and incorporate these into building a much-improved classifier for cyberbullying detection over the one reported in chapter 4.

In this chapter, we make the following contributions:

- We perform a survey to label the contents of the videos shared and the emotions displayed in the videos in Vine
- We perform an analysis of contents displayed and emotions exhibited in the videos labeled in the survey and investigate the correlation between those and both cyberaggression and cyberbullying. We found that media sessions that exhibited emotions like joy and contents

like people are less likely to be instances of cyberbullying whereas media sessions that exhibited emotions like anger were more likely to be instances of cyberbullying

- We use Latent Dirichlet Allocation (LDA) model to generate latent semantic features of the contents of the labeled media sessions' comments and include it as an additional feature to train. We found that these topical features improve the performance of the classifier reported in chapter 4 significantly.

- We make use of the labeled Vine media sessions' video contents and emotions exhibited and feed these two features to our classifiers. We show that these features yielded a further improvement across the evaluation metrics of the classifier's performance.

8.1 Video Labeling

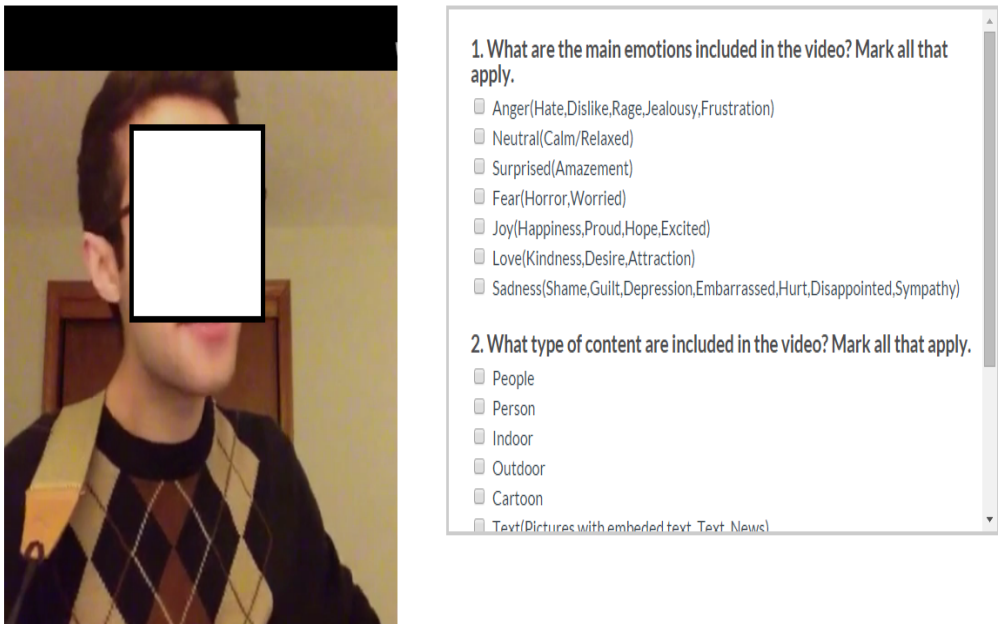


Figure 8.1: An example of video labeling survey on CrowdFlower.

We perform a survey with the sampled data-set from chapter 4 with a view to understanding what kind of videos are being shared in Vine. More importantly, we were interested in what are the contents of the videos that are being shared by the users in Vine and what are the emotions

being displayed in those videos. The goal of this survey was to understand the relation between the video content and cyberbullying in a media session, as to whether some particular categories of videos are more prone to cyberbullying. In this survey, we ask the participants two questions asking them about the content of the video shared and the emotion displayed in that video if the video content contains human presence. Figure 8.1 shows an example of the video labeling survey on CrowdFlower.

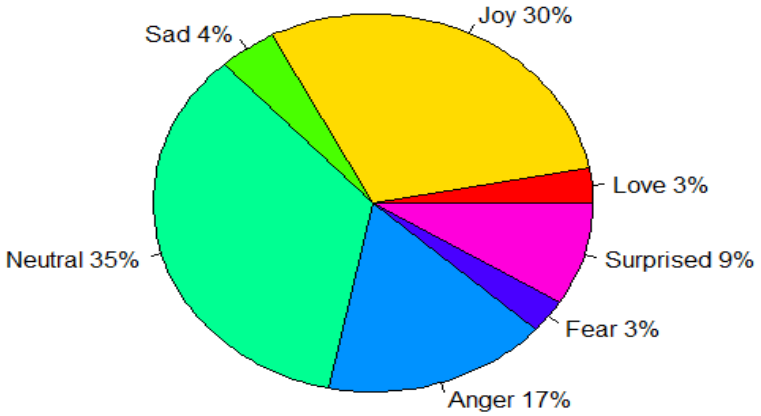


Figure 8.2: Distribution of Emotions exhibited by the media sessions.

Firstly, for the video content, labelers were given the following options to choose from: people, person, indoor, outdoor, cartoon, text, activity, animal and other. Next, the labelers are asked to identify the emotions expressed in the video, and the labelers were given the following options to choose from: neutral, joy, sad, love, surprise, fear and anger. These are the basic human emotions identified in [16].

As a good portion of the videos shared in vine are edited and more like a collage, it is possible to have a video with multiple contents and/or showing multiple emotions. To accommodate this,

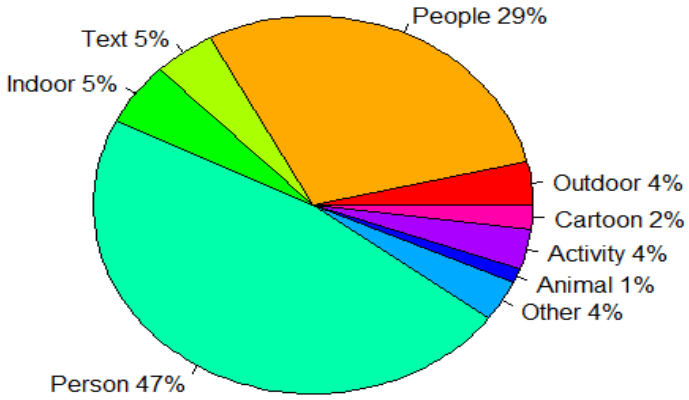


Figure 8.3: Distribution of Contents exhibited by the media sessions.

we allow the labelers to select multiple options while answering the two questions. Each media session is labeled by three labelers for this survey.

8.2 Analysis of Video Labeling

Figure 8.2 and 8.3 provide the distribution of the emotion and content of the videos. Figure 8.2 shows that the most common emotions expressed in the videos are neutral, joy and anger, comprising 82% of the total distribution, whereas the most common content types are person and people, making up 76% of the total distribution as seen from Figure 8.3.

Next, we investigate whether the content and emotion exhibited in the videos have any relation to cyberaggression and cyberbullying. For this, we plotted the distribution of emotion and content categories given that a media session had been voted k times for cyberaggression and cyberbullying from the cyberbullying and cyberaggression labeled dataset in chapter 4. Figures 8.4 and 8.5 show that videos that exhibited anger emotion and were more likely to be labeled

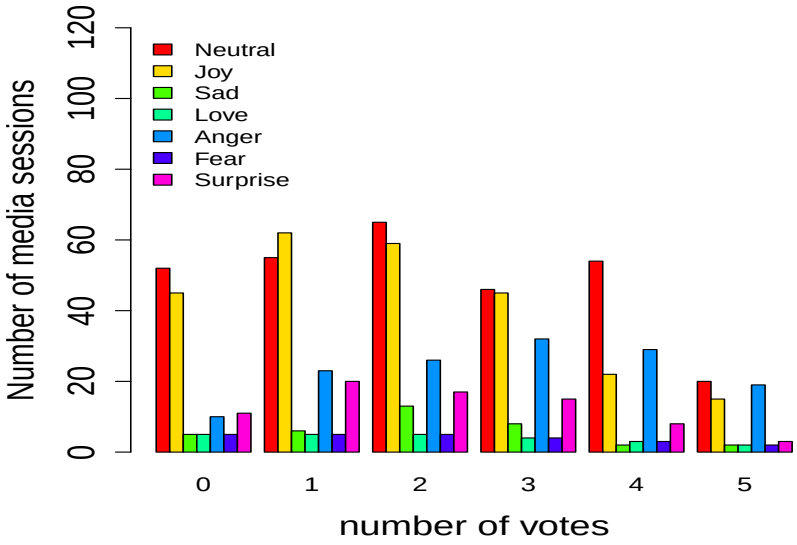


Figure 8.4: Distribution of Emotions in media sessions that were labeled k times as cyberaggression.

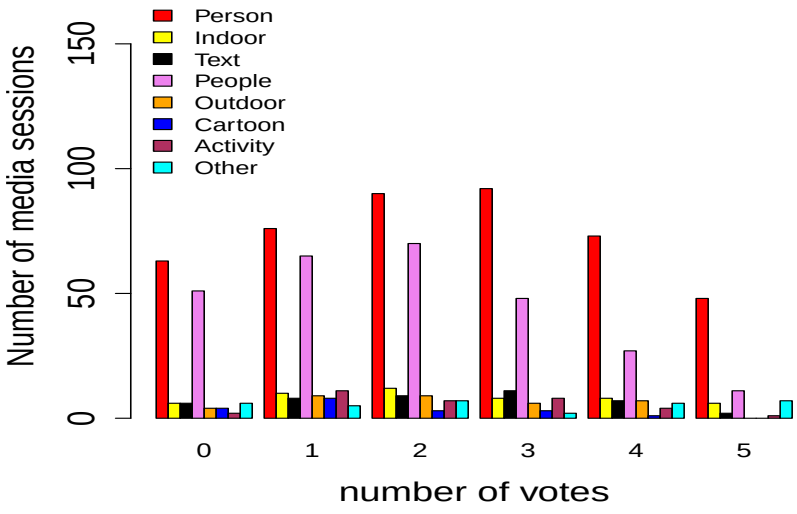


Figure 8.5: Distribution of Contents in media sessions that were labeled k times as cyberaggression.

as cyberaggression whereas for video contents, people and person categories were the primary categories across different number of votes. Similarly, Figures 8.6 and 8.7 show that anger has a positive correlation with cyberbullying whereas emotions like joy and contents like people have a

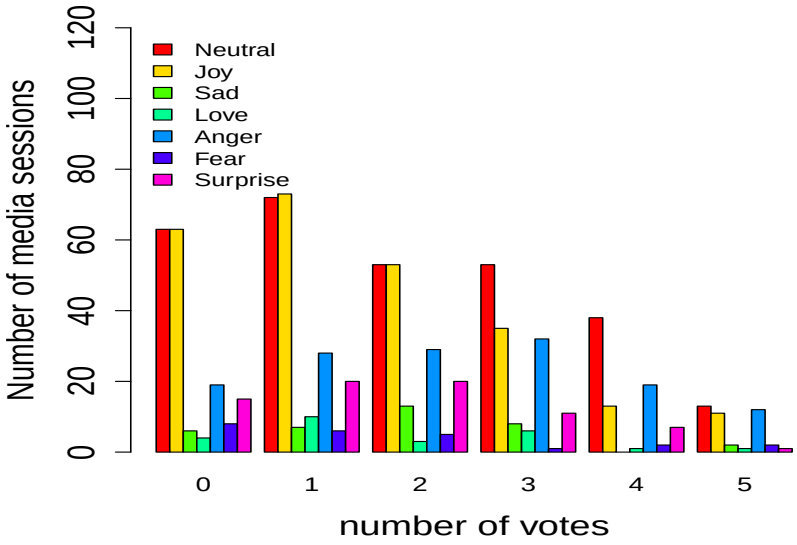


Figure 8.6: Distribution of Emotions in media sessions that were labeled k times as cyberbullying.

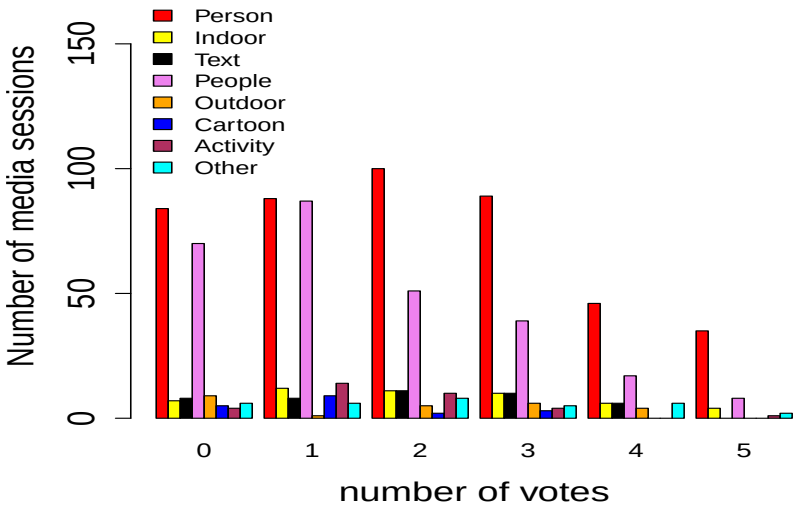


Figure 8.7: Distribution of Contents in media sessions that were labeled k times as cyberbullying.

negative correlation with cyberbullying. These observations may be helpful in classifier design since whenever the content of a video is text or the emotion displayed is joy, there appears to be little support that the media session is an instance of cyberbullying, thus improving our precision and

Table 8.1: Video and Topical Features Considered

video features	emotions exhibited, contents displayed
Latent Semantic Features	top 10 topics based on the comments using LDA

recall and decreasing the chance of mislabeling a media session as cyberbullying. **Therefore, a key finding of our video labeling analysis is that media sessions that exhibited emotions like joy and contents like people were less likely to be instances of cyberbullying whereas media sessions that exhibited emotions like anger were more likely to be instances of cyberbullying**.

8.3 Classifier Performance

Based on the labeled data from chapter 4 and the video content survey data from section 8.1, we proceeded to design and develop cyberbullying classifiers that incorporated both video content and topical features. In this section, we delineate the approaches we undertook in developing the classifier. The following subsections are organized as follows: subsection 8.3.1 describes the features we considered to develop our classifier and subsection 8.3.2 investigates different classifiers' performances with the features considered.

8.3.1 Feature Description

In addition to the profile owner, media-session and comment features described in chapter 4, we considered video features and latent semantic features. Video features include the labeled emotions and contents displayed in the media session by dint of the survey described in section 8.1. To extract topical features, we applied SVD in addition to LDA, investigated different number of topics from 3 to 50 and several normalization techniques other than min-max normalization. We only present the features, techniques, and approaches that gave us the best performing classifier in terms of accuracy, precision, and recall, as described in section 8.3.2.

8.3.2 Classifier Investigation

Based on the labeled data from CrowdFlower, we proceeded to design and evaluate classifiers that could detect cyberbullying behavior in Vine using the features described in section 8.3.1. During the survey, CrowdFlower assigned a degree of trust to each labeler that is computed from the percentage of correctly answered test questions. This degree of trust was then incorporated with the majority voting method to assign a confidence level to each survey question's answer. We take into account this weighted confidence level given by CrowdFlower to design our classifier. By taking the labeled media sessions with at least 60% confidence to make sure we had at least 3 out of 5 people agreeing on the labeling, we saw that about 31% of the media sessions were labeled as cyberbullying, which created an unbalanced data set. To make the data-set balanced, we applied Synthetic Minority Over-sampling Technique (SMOTE) and used 10-fold cross-validation to evaluate the performances of the classifier. Several classifiers were employed namely AdaBoost, DecisionTree, Random Forest, Extra Tree classifier, SVM Linear, SVM Polynomial, SVM RBF (radial basis function), SVM Sigmoid, k-NN, Naive Bayes, Neural network classifiers like Perceptron, Ridge classifier and Logistic Regression. When investigating the classifiers' performances, we used several combinations of the five types of features that gave the best performances in terms of accuracy, precision, and recall. In addition to the accuracy, we also considered precision and recall to reduce the false positives and negatives. Only those feature combinations were considered that helped the classifiers to attain the maximum accuracy, precision, and recall, as shown in Table 8.1.

Table 8.2 shows the best performing classifiers with the best performing combination of features. In addition to the accuracy, precision and recall metrics, we also considered two other metrics namely cyberbullying precision and cyberbullying recall that illustrate the precision and recall performance of the classifiers for the cyberbullying class. The reasons for including these two additional metrics are twofold [58]. First, data-set that we used to train and evaluate our classifier was imbalanced. Sometimes high accuracy in imbalanced data-sets can be misleading because high performance in the majority class can also lead to overall high accuracy. Second,

in this problem setting, we want to make sure the penalty for missing the minority class, that is cyberbullying class, is more. As it can be seen from the table, by adding the LDA features along with profile owner, media session and comment features (denoted as "All features" in the table), a noticeable improvement across all the metrics were attained for almost all the classifiers. Random Forest was the best performing classifiers with accuracy, precision, recall, cyberbullying precision and cyberbullying recall of 86,88,88,90,84 respectively whereas AdaBoost was a close second with 85,86,85,86,85 respectively. **So by adding the topical features to the profile owner, media session and comment features, we got an improvement over the best performing classifier reported in chapter 4**. Then we proceeded to add the video features i.e video contents and emotions displayed and found that this helped the **AdaBoost classifier to give accuracy, precision, recall, cyberbullying precision and cyberbullying recall values of 89,90,88,93,87 respectively, improving the performance further**. This improvement in cyberbullying precision and cyberbullying recall further solidifies our claim in section 8.2 that video features, such as joy and people are less likely to be associated with cyberbullying whereas features such as anger are more likely to be associated with cyberbullying, thus propping up the performances of all the classifiers across the metrics.

In comparison, it was found that for the Instagram social network, SVM linear was the best performing classifier [49] using features such as SVD, unigrams, trigram and image categories. So the justification for using a different classifier for Vine is twofold. Firstly, the SVD, unigram or trigram features did not seem to improve the performances of the classifier across the five metrics in Vine. Secondly, AdaBoost classifier far outperformed linear SVM in terms of performances as can be clearly seen from table 8.2. **These two reasons provide the justification to investigate Vine individually rather than using a generic classifier such as linear SVM for Vine that was reported as the best performing classifier for Instagram**.

Table 8.2: Different classifier's improved percentage performance using LDA and video contents

		Metrics				
		Accuracy	Precision	Recall	bullying Precision	bullying Recall
LDA	**Random Forest**	**79**	**80**	**79**	**84**	**72**
	AdaBoost	73	74	73	76	67
	SVMLinear	65	65	65	65	62
	ExtraTree	80	80	79	85	72
All+LDA	**RandomForest**	**86**	**88**	**88**	**90**	**84**
	AdaBoost	85	86	85	86	85
	SVMLinear	72	75	72	75	73
	ExtraTree	85	89	83	90	81
All+ LDA+ Video	RandomForest	88	90	88	93	84
		89	**90**	**88**	**93**	**87**
	SVMLinear	75	77	74	77	74
	ExtraTree	87	89	87	91	85

8.4 Future works

We plan to consider more sophisticated features like the activities exhibited in the videos shared in Vine, for example, activities related to sports, dancing, walking, etc for our classifiers. We also would like to investigate the cultural differences when it comes to labeling videos as offensive because offensive contents differ from culture to culture. We would like to build automated classifiers so that the video activity category can be automatically inputted to the cyberbullying detection classifier. We also intend to utilize automated emotion detection classifiers as described in [29] and [107]. Finally, investigating social network attributes such as clustering coefficients etc are also planned to be part of our future research works.

Another research direction is to analyze the different types of cyberbullying that take place in OSNs. We plan to label the cyberbullying instances as racial, sexual etc and then design a classifier to detect these different types of cyberbullying. In addition to that, we also plan to explore the different roles played by OSN users like perpetrators, bystanders, and upstanders. Identifying and differentiating these roles may assist us in improving the accuracy of cyberbullying classification. Another future research avenue is to incorporate the improved classifiers developed in this section into the mobile application and compare its performance against the current classifier. In the process, there will be several system challenges such as how to extract topical and video features in a scalable manner, using, for example, deep learning methods such as LSTM [46].

8.5 Conclusion

This chapter presents the following findings. First, we found that videos that showed joy and people were less likely to be labeled as cyberbullying while those exhibiting anger were somewhat more likely to be chosen as cyberbullying. Second, we found that by adding topical features derived from the comments belonging to the media sessions, our best performing classifier improved upon the classifier presented in 4 by almost 10 percent on average across three evaluation metrics. Third, we found that by adding video features, AdaBoost improved to evaluation metrics values of 89,90,88

respectively, increasing the accuracy, precision, recall respectively over the best performing classifier without the video features presented in chapter 4.

8.6 Acknowledgments

This work was supported by the US National Science Foundation (NSF) through grant CNS 1528138.

Chapter 9

Conclusions

This thesis makes the following major contributions:

9.1 Summary

Chapter 4 puts forth a detailed investigation of cyberbullying in Vine, a video-based mobile social network. Upon providing a clear distinction between cyberaggression and cyberbullying, the chapter goes on to provide thorough analyses of labeled media session data collected from Vine. After that, careful application of feature selection techniques and machine learning algorithms were performed to present the best possible classifier performance. The key findings from the chapter are as follows.

- We found that the percentage of high profanity-containing media sessions in Vine is quite low
- We discovered that a significant fraction of the high profanity-containing media sessions was not labeled as cyberbullying, though in general there was a trend towards increasing identifications of cyberbullying as the percentage of profanity increased. This suggested that the percentage of profanity in media sessions should not be used as the sole indicator of cyberbullying, but should be supplemented by other input features to the classifier
- We found that not all media sessions that exhibit cyberaggression are instances of cyberbullying, thus validating the need to apply a stricter definition of cyberbullying. Fourth,

we demonstrated that AdaBoost achieved the highest accuracy, precision, recall, cyberbullying precision and cyberbullying recall of 76,80,72,81,74 percent respectively using a combination of profile owner, media session, comment features and unigrams.

An investigation of differentiating factors between non-cyberbullying and cyberbullying was outlined in chapter 5. To the best of our knowledge, this is the first research to investigate factors that differentiate a cyberbullying session from a non-cyberbullying one for both Vine and Instagram, two media-based online social networks. Labeled data that used an appropriate definition of cyberbullying from chapter 4 was leveraged to present the following key findings.

- For both Vine and Instagram, cyberbullying media sessions are more likely to have more unique negative sentiment commenters
- In cyberbullying media sessions, profile owners are much more likely to post highly negative sentiment comments with comparatively higher subjectivity spread across the temporal frame since the sharing of the media than in a non-cyberbullying media session
- In the cyberbullying media sessions, negative sentiment comments persist with higher subjectivity even after a long time since the media is posted, which is not the case for non-cyberbullying media sessions
- Density of positive comments coming in for cyberbullying media sessions for both Vine and Instagram is much less than that for the non-cyberbullying media sessions across the temporal frame of the lifetime of the media session

Although a substantial amount of research has been proposed to develop accurate cyberbullying classifiers, two key system challenges, namely scalability, and responsiveness of a potential cyberbullying detection system have largely been ignored. Chapter 6 addresses these two challenges by making the following contributions.

- Design and implementation of a cyberbullying detection system that consists of two novel components: dynamic priority scheduler and incremental classifier computation phase

- Thorough evaluation of the implemented system showing that our system substantially improves the scalability of cyberbullying detection, enabling five times more media sessions to be monitored for the same average alert time of 2 hours compared to an un-prioritized system

- We demonstrate that our system can fully monitor Vine-scale social networks for cyberbullying detection for a year using only eight 1 GB AWS VM instances

- We discover the point (32 GB) at which adding memory no longer enables monitoring of more media sessions and project that our system would need 120 32 GB instances to fully monitor Instagram-scale traffic for cyberbullying.

Our ultimate goal is to provide an efficient, effective and feasible platform for guardians so that they are able to monitor their kids' activities in online social networks with utmost ease and get notifications when a potential cyberbullying takes place. To this aim, chapter 7 makes the following contributions.

- We present the design and implementation of an Android mobile application, BullyAlert. The reason for developing this mobile application is twofold. First, we wanted to move on from a centralized cyberbullying system described in chapter 6 as the sheer number of different social networks as well as their daily active users will still put substantial resource constraint. Second, in chapter 6, a single general cyberbullying detection classifier was developed to be applied to all guardians. We argue that different parents have different tolerance levels for cyberbullying and thus, will want to get different levels of alerts from the system

- We outline the design and implementation of the feedback mechanism and the adaptive classifier mechanism, by dint of which the application will be able to calibrate itself according to the varying tolerance levels of guardians

- Using the preliminary data collected from the application, we present an initial user experience analysis of the guardians who downloaded the application and the social network users who were being monitored by the guardians.

In chapter 8, we present a short analysis of more complicated feature extractions, such as video (emotions displayed and content shared) and topical features and incorporate these features into building a much-improved classifier for cyberbullying detection over the one reported in chapter 4. The following are the major contributions from the chapter.

- We perform a survey to label the contents of the videos shared and the emotions displayed in the videos in Vine
- We present an analysis of contents displayed and emotions exhibited in the videos labeled in the survey and investigate the correlation between those and both cyberaggression and cyberbullying and find that media sessions that exhibited emotions like joy and contents like people are less likely to be instances of cyberbullying whereas media sessions that exhibited emotions like anger were more like to be instances of cyberbullying
- We use Latent Dirichlet Allocation (LDA) model to generate latent semantic features of the contents of the labeled media sessions' comments and include it as an additional feature to train and test the performance of our classifiers and improve the performance of the classifier presented in chapter 4 significantly
- We make use of the labeled Vine media sessions' video contents and emotions exhibited and include these two features as features to our classifiers which yielded a further improvement across the evaluation metrics of the classifier. It is found that, after including the additional video features significantly improve the classifiers' performance

9.2 What Next?

There are several future research avenues that can be traversed. While we have delved deep into the social network Vine, other social networks like YouTube, SnapChat or Facebook are yet to be considered. These social networks are also immensely popular[7, 8, 101, 9], thus making these very interesting candidates to further the research on cyberbullying behavior. In terms of investigating cyberaggression behavior, an empirical differentiation analysis between cyberaggression and cyberbullying can be a future research interest too. Furthermore, the roles of different participants can also be examined, such as upstanders, victims and perpetrators. Last but not the least, by exploring the social connectivity graphs of the social networks, research can be done to analyze the cliques and communities to see which are most vulnerable to cyberbullying activities and why. Finally, investigation of different types of cyberbullying, such as racial, sexual and so on is also one of future research directions.

In chapter 6, we implemented a scalable and responsive cyberbullying detection system. In future, we will extend this research to investigate portability of our system and the design choices we made to other OSNs similar in nature such as Instagram and different in nature such as Facebook, SnapChat etc. We will also examine in depth the plateauing effect that limits the effectiveness of adding more memory, namely that there is likely a computational bottleneck that needs to be further addressed.

In chapter 6 and 7, we employed logistic regression and lightweight features for the development of the systems. Future research directions includes exploring topical and video/image contents into the classifier and be still scalable and responsive. We have shown in chapter 8 that topical and video features improve the performances of cyberbullying detection classifier to a considerable extent. So incorporating these features into both the system and the mobile application is an essential future research direction. Last but not the least, we plan to collect more data from our BullyAlert application to solidify our hypothesis that different people tend to have different tolerance levels of cyberbullying by leveraging the feedback data. Finally, we will investigate, in depth, the feedback

behavior from BullyAlert to make sure that our classifier is indeed adaptive. This will mean that, for each parent, the classifier will eventually calibrate its tolerance level to that of its user, so the feedback from the guardians will eventually reach “almost all rights”.

Bibliography

[1] Fabian Abel, Nicola Henze, Eelco Herder, and Daniel Krause. Linkage, aggregation, alignment and enrichment of public user profiles with mypes. In Proceedings of the 6th International Conference on Semantic Systems, I-SEMANTICS '10, pages 11:1–11:8, New York, NY, USA, 2010. ACM.

[2] Deepak Agarwal, Bo Long, Jonathan Traupman, Doris Xin, and Liang Zhang. Laser: A scalable response prediction platform for online advertising. In WSDM, pages 173–182, New York, NY, USA, 2014. ACM.

[3] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci. Wireless sensor networks: A survey. Comm. ACM, 38(4):393–422, 2002.

[4] Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. Gender, genre, and writing style in formal written texts. TEXT, 23:321–346, 2003.

[5] Zahra Ashktorab, Srijan Kumar, Soham De, and Jennifer Golbeck. ianon: Leveraging social network big data to mitigate behavioral symptoms of cyberbullying. iConference 2014 (Social Media Expo), 2014.

[6] Ask.Fm. `https://support.ask.fm/hc/en-us/articles/115008817448-Anonymity-explained`. [Online; accessed November 8, 2018.].

[7] Salman Aslam. `https://www.omnicoreagency.com/snapchat-statistics/`. [Online; accessed October 27, 2018.].

[8] Salman Aslam. `https://www.omnicoreagency.com/snapchat-statistics/`. [Online; accessed October 27, 2018.].

[9] Salman Aslam. `https://www.omnicoreagency.com/facebook-statistics/`. [Online; accessed October 27, 2018.].

[10] Fabiola Baltar and Ignasi Brunet. Social research 2.0: virtual snowball sampling method using facebook. Internet Research, 22(1):57–74, 2012.

[11] Ryan Broderick. 9 teenage suicides in the last year were linked to cyber-bullying on social network ask.fm. `http://www.buzzfeed.com/ryanhatesthis/a-ninth-teenager-since-last-september-has- committed-suicide`, 2013. [Online;accessed 14-January-2014].

[12] Probability Calibration. `http://scikit-learn.org/stable/modules/calibration.html`. [Online; accessed September, 2017].

[13] Magnus Carlsson. Monads for incremental computing. In Proceedings of the Seventh ACM SIGPLAN International Conference on Functional Programming, ICFP '02, pages 26–35, New York, NY, USA, 2002. ACM.

[14] Francesca Carmagnola and Federica Cena. User identification for cross-system personalisation. Inf. Sci., 179(1-2):16–32, January 2009.

[15] Cyberbullying Research Center. Cyberbullying research center. `http://cyberbullying.us`, 2013. [Online; accessed September, 2013].

[16] changingmings.org. Basic emotions. `http://changingminds.org/explanations/emotions/emotions.htm`, 2015. [Online; accessed April 24, 2015.].

[17] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. Detecting aggressors and bullies on twitter. In WWW '17 Companion, pages 767–768, 2017.

[18] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16:321–357, 2002.

[19] Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. Detecting offensive language in social media to protect adolescent online safety. In Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust, SOCIALCOM-PASSAT '12, pages 71–80, Washington, DC, USA, 2012. IEEE Computer Society.

[20] Hanqiang Cheng, Yu-Li Liang, Xinyu Xing, Xue Liu, Richard Han, Qin Lv, and Shivakant Mishra. Efficient misbehaving user detection in online video chat services. In WSDM '12, pages 23–32, New York, NY, USA, 2012. ACM.

[21] J. F. Chisholm. Cyberspace violence against girls and adolescent females. In Annals of the New York Academy of Sciences, pages 74–89, 2006.

[22] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In Proceedings of the 25th International Conference on Machine Learning, ICML '08, pages 160–167, New York, NY, USA, 2008. ACM.

[23] National Crime Prevention Council. Teens and cyberbullying, 2007. Executive summary of a report on research conducted for National Crime Prevention Council.

[24] CrowdFlower. `https://success.figure-eight.com/hc/en-us/articles/201855679-Guide-to-Contributors-Page`. [Online; accessed October 27, 2018.].

[25] CrowdFlower. `https://success.crowdflower.com/hc/en-us/articles/201855939-How-to-Calculate-a-Confidence-Score`. [Online; accessed March 10, 2016.].

[26] M. Dadvar, F. M. G. de Jong, R. J. F. Ordelman, and R. B. Trieschnigg. Improved cyberbullying detection using gender information. In Twelfth Dutch-Belgian Information Retrieval Workshop, DIR, pages 23–25. University of Ghent, 2012.

[27] Maral Dadvar, Dolf Trieschnigg, and Franciska de Jong. Experts and machines against bullies: A hybrid approach to detect cyberbullies. In Advances in Artificial Intelligence, volume 8436 of Lecture Notes in Computer Science, pages 275–281, Berlin, Germany, May 2014. Springer Verlag.

[28] Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. Improving cyberbullying detection with user context. In Proceedings of the 35th European Conference on Advances in Information Retrieval, ECIR'13, pages 693–696, Berlin, Heidelberg, 2013. Springer-Verlag.

[29] Liyanage De Silva, T Miyasato, and Ryohei Nakatsu. Facial emotion recognition using multi-modal information. In Proceedings of the IEEE Intelligent Conf. Information, Comm. And Signal Processing, volume 1, pages 397 – 401 vol.1, 10 1997.

[30] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. ACM Transactions on Interactive Intelligent Systems (TiiS) - Special Issue on Common Sense for Interactive Systems, 2(3):18:1–18:30, September 2012.

[31] Karthik Dinakar, Roi Reichart, and Henry Lieberman. Modeling the detection of textual cyberbullying., 2011.

[32] Linda DiProperzio. Cyberbullying applications. `http://www.parents.com/kids/safety/internet/best-apps-prevent-cyberbullying/`, 2015. [Online; accessed February 6, 2015.].

[33] Julian J Dooley, Jacek Pyżalski, and Donna Cross. Cyberbullying versus face-to-face bullying. Zeitschrift für Psychologie/Journal of Psychology, 217(4):182–188, 2009.

[34] A. Nocentini E. Menesini. Cyberbullying definition and measurement. some critical considerations. Journal of Psychology, 217(4):320–323, 2009.

[35] Gunnar Eriksson and Jussi Karlgren. Features for modelling characteristics of conversations. In Pamela Forner, Jussi Karlgren, and Christa Womser-Hacker, editors, CLEF (Online Working Notes/Labs/Workshop), 2012.

[36] Russel Goldman. Teens indicted after allegedly taunting girl who hanged herself, bbc news. `http://abcnews.go.com/Technology/TheLaw/teens-charged-bullying-mass-girl-kill/story?id=10231357`, 2010. [Online; accessed 14-January-2014].

[37] Sherri Gordon. 4 apps used for sexting and cyberbullying parents should know about, 2014. [Online; accessed June 11, 2014.].

[38] Jennifer Van Grove. Ask.fm statistics. `http://www.cnet.com/news/ask-fm-the-troubling-secret-playground-of-tweens-and-teens/`, 2013. [Online; accessed February 6, 2015.].

[39] Matthew A. Hammer, Joshua Dunfield, Kyle Headley, Nicholas Labich, Jeffrey S. Foster, Michael Hicks, and David Van Horn. Incremental computation with names. CoRR, abs/1503.07792, 2015.

[40] Matthew A Hammer, Joshua Dunfield, Kyle Headley, Nicholas Labich, Jeffrey S Foster, Michael Hicks, and David Van Horn. Incremental computation with names. In ACM SIGPLAN Notices, volume 50, pages 748–766. ACM, 2015.

[41] Matthew A Hammer, Khoo Yit Phang, Michael Hicks, and Jeffrey S Foster. Adapton: Composable, demand-driven incremental computation. In ACM SIGPLAN Notices, volume 49, pages 156–166. ACM, 2014.

[42] Matthew A. Hammer, Khoo Yit Phang, Michael Hicks, and Jeffrey S. Foster. Adapton: Composable, demand-driven incremental computation. In Proceedings of the 35th ACM SIGPLAN Conference on Programming Language Design and Implementation, PLDI '14, pages 156–166, New York, NY, USA, 2014. ACM.

[43] Keith Harrison. Scalable Detection of Community Cyber Incidents Utilizing Distributed and Anonymous Security Information Sharing. PhD thesis, The University of Texas at San Antonio, 2012. AAI3548643.

[44] Yoni Heisler. `https://www.omnicoreagency.com/snapchat-statistics/`. [Online; accessed October 27, 2018.].

[45] S. Hinduja and J. W. Patchin. Cyberbullying research summary, cyberbullying and suicide, 2010.

[46] Sepp Hochreiter and Jrgen Schmidhuber. Long short-term memory. 9:1735–80, 12 1997.

[47] Homa Hosseinmardi, Amir Ghasemianlangroodi, Richard Han, Qin Lv, and Shivakant Mishra. Towards understanding cyberbullying behavior in a semi-anonymous social network. In IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pages 244 – 252, Beijing,China, 2014. IEEE.

[48] Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. Analyzing Labeled Cyberbullying Incidents on the Instagram Social Network, pages 49–66. Springer International Publishing, Cham, 2015.

[49] Homa Hosseinmardi, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. Prediction of cyberbullying incidents in a media-based social network. In Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, San Francisco,CA,USA, 2016. IEEE.

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

[51] http://cybersafety.wikispaces.com. `http://cybersafety.wikispaces.com`. [Online; accessed March 10, 2016.].

[52] http://www.eblaster.com. `http://www.eblaster.com/`. [Online; accessed March 10, 2016.].

[53] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 168–177. ACM, 2004.

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

[55] Simon C. Hunter, James ME Boyle, and David Warden. Perceptions and correlates of peer-victimization and bullying. British Journal of Educational Psychology, 77(4):797–810, 2007.

[56] Loc N Huynh, Rajesh Krishna Balan, and Youngki Lee. Deepmon: Building mobile gpu deep learning models for continuous vision applications. In Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services, pages 186–186. ACM, 2017.

[57] Giacomo Inches and Fabio Crestani. Overview of the international sexual predator identification competition at pan-2012. In CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012, 2012.

[58] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. An Introduction to Statistical Learning: With Applications in R. Springer Publishing Company, Incorporated, 2014.

[59] PErverted Justice. `http://www.perverted-justice.com/`. [Online; accessed November 8, 2018.].

[60] A. Kontostathis K. Reynolds and L. Edwards. Using machine learning to detect cyberbullying. Proceedings of the 2011 10th International Conference on Machine Learning and Applications and Workshops, 2:241–244, 2011.

[61] A Kontostathis, W. West, A Garron, K. Reynolds, , and L. Edwards. Identify predators using chatcoder 2.0. In CLEF (Online Working Notes/Labs/Workshop), 2012.

[62] April Kontostathis. Chatcoder: Toward the tracking and categorization of internet predators. In PROC. TEXT MINING WORKSHOP 2009 HELD IN CONJUNCTION WITH THE NINTH SIAM INTERNATIONAL CONFERENCE ON DATA MINING (SDM 2009). SPARKS, NV. MAY 2009., 2009.

[63] April Kontostathis, Kelly Reynolds, Andy Garron, and Lynne Edwards. Detecting cyberbullying: query terms and techniques. In Proceedings of the 5th Annual ACM Web Science Conference, pages 195–204. ACM, 2013.

[64] Robin Kowalski, Gary W Giumetti, Amber Schroeder, and Micah R Lattanner. Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth. Psychological bulletin, 140, 02 2014.

[65] Robin M. Kowalski, Sue Limber, Susan P. Limber, and Patricia W. Agatston. Cyberbullying: Bullying in the digital age. John Wiley & Sons, Reading, MA., 2012.

[66] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems 25, pages 1097–1105. Curran Associates, Inc., 2012.

[67] Evan Lepage. Instagram statistics. http://blog.hootsuite.com/instagram-statistics-for-business/, 2015. [Online; accessed February 6, 2015.].

[68] 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 2014 IEEE Fourth International Conference on Big Data and Cloud Computing, pages 355–362, Sydney, Australia, Dec 2014. IEEE.

[69] Rui Li, Kin Hou Lei, Ravi Khadiwala, and Kevin Chen-Chuan Chang. Tedas: A twitter-based event detection and analysis system. In Proceedings of the 2012 IEEE 28th International Conference on Data Engineering, ICDE '12, pages 1273–1276, Washington, DC, USA, 2012. IEEE Computer Society.

[70] S. P. Limber, R. M. Kowalski, and P. A. Agatston. Cyber bullying: A curriculum for grades 6-12. Center City, MN: Hazelden., 2008.

[71] Steven Loria. Python sentiment library. https://github.com/sloria/textblob, 2016. [Online; accessed May 30, 2016].

[72] D. Maher. Cyberbullying: an ethnographic case study of one australian upper primary school class. Youth studies Australia, 27(4),5057, 2008.

[73] India Mcghee, Jennifer Bayzick, April Kontostathis, Lynne Edwards, Alexandra Mcbride, and Emma Jakubowski. Learning to identify internet sexual predation. Int. J. Electron. Commerce, 15(3):103–122, April 2011.

[74] Claire P Monks and Peter K Smith. Definitions of bullying: Age differences in understanding of the term, and the role of experience. British Journal of Developmental Psychology, 24(4):801–821, 2006.

[75] S. Nadali, M. A. A. Murad, N. M. Sharef, A. Mustapha, and S. Shojaee. A review of cyberbullying detection: An overview. In 2013 13th International Conference on Intelligent Systems Design and Applications (ISDA), pages 325 – 330, Bangi, 2013. IEEE.

[76] Vinita Nahar, Xue Li, and Chaoyi Pang. An effective approach for cyberbullying detection, 2013.

[77] Vinita Nahar, Sayan Unankard, Xue Li, and Chaoyi Pang. Sentiment analysis for effective detection of cyber bullying. In Web Technologies and Applications, pages 767–774. Springer, 2012.

[78] Vinita Nahar, Sayan Unankard, Xue Li, and Chaoyi Pang. Semi-supervised learning for cyberbullying detection in social networks. In Databases Theory and Applications, LNCS'12, pages 160–171, 2014.

[79] K. Nalini and L. J. Sheela. Classification of tweets using text classifier to detect cyber bullying. In Emerging ICT for Bridging the Future-Proceedings of the 49th Annual Convention of the Computer Society of India CSI Volume 2, pages 637–645. Springer, 2015.

[80] Andrew Y. Ng. Feature selection, l1 vs. l2 regularization, and rotational invariance. In Proceedings of the Twenty-first International Conference on Machine Learning, ICML '04, pages 78–, New York, NY, USA, 2004. ACM.

[81] Dan Olweus. Bullying at school: What we know and what we can do, 1993.

[82] T O'Neil and D. Zinga. Childrens rights: multidisciplinary approaches to participation and protection. Univ of Toronto Pr,2008, 2008.

[83] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts, 2004.

[84] Javier Parapar, David E. Losada, and Alvaro Barreiro. A learning-based approach for the identification of sexual predators in chat logs. In Pamela Forner, Jussi Karlgren, and Christa Womser-Hacker, editors, CLEF (Online Working Notes/Labs/Workshop), volume 1178 of CEUR Workshop Proceedings. CEUR-WS.org, 2012.

[85] Justin W. Patchin and Sameer Hinduja. Bullies move beyond the schoolyard; a preliminary look at cyberbullying. Youth violence and juvenile justice 4:2, pages 148–169, 2006.

[86] Justin W. Patchin and Sameer Hinduja. An update and synthesis of the research. Cyberbullying Prevention and Response: Expert Perspectives, page 13, 2012.

[87] N. Potha and M. Maragoudakis. Cyberbullying detection using time series modeling. 2014 IEEE International Conference on Data Mining Workshop (ICDMW), pages 373–382, 2014.

[88] M. Ptaszynski, P. Dybala, T. Matsuba, F. Masui, R. Rzepka, K. Araki, and Y. Momouchi. In the service of online order tackling cyberbullying with machine learning and affect analysis. International Journal of Computational Linguistics Research, 1(3):135–154, 2010.

[89] Scikit Library Python. `http://scikit-learn.org/stable/modules/sgd.html`. [Online; accessed October 22, 2018.].

[90] Jacek Pyżalski. Electronic aggression among adolescents: An old house with. Youth culture and net culture: Online social practices, page 278, 2010.

[91] Rahat Ibn Rafiq. `https://github.com/RahatIbnRafiq/cybersafetyapp_servercodes`. [Online; accessed October 22, 2018.].

[92] Rahat Ibn Rafiq. `https://github.com/RahatIbnRafiq/AndroidCodesForCyberbullying`. [Online; accessed October 22, 2018.].

[93] Rahat Ibn Rafiq, Homa Hosseinmardi, Richard Han, Qin Lv, Shivakant Mishra, and Sabrina Arredondo Mattson. Careful what you share in six seconds: detecting cyberbullying instances in vine. In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, pages 617–622, Paris,France, 2015. ACM.

[94] Ashwin Rajadesingan, Reza Zafarani, and Huan Liu. Sarcasm detection on twitter: A behavioral modeling approach. In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, WSDM '15, pages 97–106, New York, NY, USA, 2015. ACM.

[95] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In Proceedings of the 19th International Conference on World Wide Web, WWW '10, pages 851–860, New York, NY, USA, 2010. ACM.

[96] H. Sanchez and S. Kumar. Twitter bullying detection. In NSDI, pages 15–15, Berkeley, CA, USA, 2012. USENIX Association.

[97] Python scikit learn. Adaboost classifier. http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html, 2016.

[98] F. Shaorong and H. Zhixue. An incremental associative classification algorithm used for malware detection. In Future Computer and Communication (ICFCC), 2010 2nd International Conference on, volume 1, pages V1–757–V1–760, May 2010.

[99] Feng Shaorong and Han Zhixue. An incremental associative classification algorithm used for malware detection. In 2nd International Conference on Future Computer and Communication (ICFCC), 2010, pages V1–757–V1–760, May 2010.

[100] T .H. Silva, P .O. S. V. de Melo, J. M. Almeida, J. Salles, and A. A. F. Loureiro. A picture of Instagram is worth more than a thousand words: Workload characterization and application. In 2013 IEEE International Conference on Distributed Computing in Sensor Systems (DCOSS), pages 123–132. IEEE, 2013.

[101] Cooper Smith. Facebook statistics. http://www.businessinsider.com/facebook-350-million-photos-each-day-2013-9, 2013. [Online; accessed February 6, 2015.].

[102] Craig Smith. Vine statistics. http://expandedramblings.com/index.php/vine-statistics/, 2015. [Online; accessed February 6, 2015.].

[103] P. K. Smith, J. Mahdavi, M. Carvalho, S. Fisher, S. Russel, and N. Tippet. Cyberbullying: its nature and impact in secondary school pupils. Journal of Child Psychology and Psychiatry, 49(4):376–385, 2008.

[104] Peter K. Smith, Cristina del Barrio, and R. Tokunaga. Principles of Cyberbullying Research. Definitions, measures and methodology, Chapter: Definitions of Bullying and Cyberbullying: How Useful Are the Terms? Routledge, 2012.

[105] Laura Smith-Spark. Hanna smith suicide fuels calls for action on ask.fm cyberbullying, cnn. http://www.cnn.com/2013/08/07/world/europe/uk-social-media-bullying/, 2013. [Online;accessed 14-January-2014].

[106] StopCyberBullying.org. 5 differences between cyber bullying and traditional bullying. http://onlinesense.org/5-differences-cyber-bullying-traditional-bullying/, 2017. [Online; accessed May 22, 2017.].

[107] Yafei Sun, Nicu Sebe, Michael S. Lew, and Theo Gevers. Authentic emotion detection in real-time video. In Computer Vision in Human-Computer Interaction, pages 94–104. IEEE, 2004.

[108] P.N. Tan, F. Chen, and A. Jain. Information assurance: Detection of web spam attacks in social media. In Proceedings of Army Science Conference,,Orland, Florida. 2010, 2010.

[109] Brett Thom, Brett Thom, Akshaye Dhawan, Lynne Edwards, John P. Dougherty, and Roger Coleman. Safechat: Using open source software to protect minors from internet.

[110] R. S. Tokunaga. Following you home from school: A critical review and synthesis of research on cyberbullying victimization. Computers in Human Behavior, 26:277–287, 2010.

[111] A. Tzamaloukas and J. J. Garcia-Luna-Aceves. Channel-hopping multiple access. Technical Report I-CA2301, Department of Computer Science, University of California, Berkeley, CA, 2000.

[112] Heidi Vandebosch and Katrin Van Cleemput. Defining cyberbullying: A qualitative research into the perceptions of youngsters. CyberPsychology and Behavior, 11(4), 2008.

[113] Esa Villatoro-Tello, Antonio Jurez-Gonzlez, Hugo Jair Escalante, Manuel Montes y Gmez, and Luis Villaseor Pineda. A two-step approach for effective detection of misbehaving users in chats. In Pamela Forner, Jussi Karlgren, and Christa Womser-Hacker, editors, CLEF (Online Working Notes/Labs/Workshop), volume 1178 of CEUR Workshop Proceedings. CEUR-WS.org, 2012.

[114] Luis von Ahn's Research Group. Negative words list form, luis von ahn's research group. http://www.cs.cmu.edu/~biglou/resources/, 2014.

[115] Wikipedia. https://en.wikipedia.org/wiki/Cumulative_distribution_function#Complementary_cumulative_distribution_function_(tail_distribution). [Online; accessed October 27, 2018.].

[116] Wikipedia. https://en.wikipedia.org/wiki/Majority_rule. [Online; accessed October 27, 2018.].

[117] Wikipedia. https://en.wikipedia.org/wiki/Cross-validation_(statistics)#k-fold_cross-validation. [Online; accessed October 27, 2018.].

[118] Wikipedia. logistic regression. https://en.wikipedia.org/wiki/Logistic_regression, 2016. [Online; accessed May 30, 2016].

[119] N.E. Willard. Cyberbullying and cyberthreats: Responding to the challenge of online social aggression, threats, and distress. Champaign, IL: Research, 2007.

[120] www.netnanny.com. https://www.netnanny.com/. [Online; accessed March 10, 2016.].

[121] www.parentalsoftware.org. http://www.parentalsoftware.org/bigbrother.html. [Online; accessed March 10, 2016.].

[122] Xinyu Xing, Yu-li Liang, Sui Huang, Hanqiang Cheng, Richard Han, Qin Lv, Xue Liu, Shivakant Mishra, and Yi Zhu. Scalable misbehavior detection in online video chat services. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, pages 552–560, New York, NY, USA, 2012. ACM.

[123] Xinyu Xing, Yu-li Liang, Sui Huang, Hanqiang Cheng, Richard Han, Qin Lv, Xue Liu, Shivakant Mishra, and Yi Zhu. Scalable misbehavior detection in online video chat services. In KDD, pages 552–560, New York, NY, USA, 2012. ACM.

[124] Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. Learning from bullying traces in social media. In NAACL HLT, pages 656–666. Association for Computational Linguistics, 2012.

[125] Zhi Xu and Sencun Zhu. Filtering offensive language in online communities using grammatical relations. Proceedings of Collaboration, Electronic messaging, Anti-Abuse and Spam Conference 2010, 2010.

[126] Dawei Yin, Zhenzhen Xue, Liangjie Hong, Brian D Davison, April Kontostathis, and Lynne Edwards. Detection of harassment on web 2.0. Proceedings of the Content Analysis in the WEB, 2:1–7, 2009.