Multimodal Labeling and Characterization of Social Network Data for Detection and Prediction of Cyberbullying

by

H. Hosseinmardi

B.S., Khajeh Nasir Toosi University, 2005

M.S., University of Tehran, 2008

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

2016

ProQuest Number: 10108742

All rights reserved

INFORMATION TO ALL USERS
The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript
and there are missing pages, these will be noted. Also, if material had to be removed,
a note will indicate the deletion.

ProQuest 10108742

Published by ProQuest LLC (2016). Copyright of the Dissertation is held by the Author.

All rights reserved.
This work is protected against unauthorized copying under Title 17, United States Code
Microform Edition © ProQuest LLC.

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

This thesis entitled:
Multimodal Labeling and Characterization of Social Network Data for Detection and Prediction of Cyberbullying
written by H. Hosseinmardi
has been approved for the Department of Computer Science

Prof. Richard Han

Prof. Vanja Dukic

Prof. Qin Lv

Prof. Shivakant Mishra

Prof. Tom Yeh

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.

Hosseinmardi, H. (Ph.D., Computer Science)

Multimodal Labeling and Characterization of Social Network Data for Detection and Prediction of Cyberbullying

Thesis directed by Prof. Prof. Richard Han

This dissertation provides insight into the problem cyberbullying in social networks by investigating profanity usage, ground truth labeling of cyberbullying, and characterization of relationships between cyberbullying and a variety of factors, including profane word usage, social graph features, temporal commenting behavior, linguistic content, and multimedia modality. This thesis investigates how labeled cyberbullying compares to labeled cyberaggression, the utility of labeling images involved in cyberaggression and cyberbullying incidents, the propagation of cyberbullying in social networks, and detecting/predicting cyberbullying in social networks.

Dedication

To my husband, Amir Ghasemianlangroodi, for all the support, both intellectually and spiritually and to my parents for being there for me all the time throughout my life.

Acknowledgements

Thanks to all cyberbullying project members. Foremost, I would like to express my sincere gratitude to my advisor Prof. Richard Han for the continuous support of my PhD study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this dissertation report.

Besides my advisor, I would like to thank the rest of my committee: Prof. Qin Lv, Prof. Shivakant Mishra, Prof. Tom Yeh,and Prof. Vanja Dukic. I would also acknowledge Rahat Ibn Rafiqh, for his contributions in data collection, and Dr. Sabrina Arredondo Mattson, for her advice on cyberbullying problem, which were very fruitful for shaping up my ideas and research.

Contents

Chapter

1 Overview 1

1.1 Thesis Statement 1

1.2 Research Findings 1

2 Introduction 3

3 Related Work 6

4 Analysis of Negative and Positive Word Usage in a Semi-anonymous Social Network 12

4.1 Overview 12

4.2 Data Collection 13

4.2.1 Ask.fm 13

4.2.2 Description of Collected Data 14

4.2.3 Modeled Interaction Network 16

4.2.4 Extracting Positive and Negative Words 17

4.3 Network Statistics 18

4.3.1 Building the Interaction Graph 18

4.4 Cyberbullying Types 27

4.5 Users with Cutting Behavior 29

4.6 Summary 30

5 Analysis of Negative and Positive Word Usage Across Social Networks 34

5.1 Overview 34

5.2 Methodology 35

5.2.1 Instagram and Ask.fm Features 36

5.2.2 Data Collection 37

5.3 Analysis of Instagram User Behavior 38

5.4 Analysis of Ask.fm User Behavior 40

5.5 User Behavior Across Instagram and Ask.fm 42

5.6 Impact of Anonymity on User Behavior in Ask.fm 49

5.7 Summary 51

6 Labeling and Characterization of Cyberbullying and Cyberaggression in Social Networks 53

6.1 Overview 53

6.2 Labeling 56

6.2.1 Quality Control 56

6.2.2 Image Labeling 57

6.3 Analysis and Characterization of Ground Truth Instagram Data 58

6.3.1 Labeling and Negativity Analysis 59

6.3.2 Temporal and Graph Properties Analysis 62

6.3.3 Linguistic and Psychological Analysis 64

6.3.4 Image Content Analysis 66

6.4 Summary 68

7 Detection and Prediction of Cyberbullying Incidents 70

7.1 Detection of Cyberbullying Incidents 70

7.2 Prediction of Cyberbullying Incidents 73

7.3 Summary 75

8 Analyzing Factors Impacting Revining and Propagation of Cyberbullying Incidents on the Vine Social Network 77

8.1 Overview 77

8.2 Dataset 80

8.3 User Behavior on Vine 81

8.3.1 Basic Analysis 81

8.3.2 Revining Behavior and Followers and Following 83

8.3.3 Revining and Temporal Behavior 84

8.3.4 Revining and Hash Tags 86

8.3.5 Revining and Followers 87

8.3.6 Revine Tree 88

8.4 Labeled Data Analysis 90

8.4.1 Revining and the Content/Emotion of Vines 90

8.4.2 Revining and Cyberbullying Vines 92

8.5 Summary 93

8.6 Acknowledgments 94

9 Conclusions 95

9.1 Summary 95

9.2 Future Work 98

Bibliography 99

Appendix

Tables

Table

4.1 The first row shows the negative words with the highest frequency and the second row shows the negative words with the highest eigenvector centrality (values decrease from left to right). 18

4.2 Average number of questions and words per user 19

4.3 Average reciprocity and degree for different groups 24

4.4 Local clustering coefficient for some social networks. Clustering coefficient values for Flickr, Orkut and You Tube are from paper [62] and values for Facebook, Twitter, Gplus are from the website [60]. 26

4.5 Local clustering coefficient for defined group 26

4.6 Average number of questions and words per user with "cutting" label 29

4.7 Average reciprocity and degree for users with "cutting" label 30

5.1 Comparison of (positive and negative) posts by profile owners and friends across four different user groups. 44

5.2 The first row shows the negative words with the highest frequency and the second row shows the negative words with the highest eigenvector centrality in Ask.fm. The third row shows the negative words with the highest frequency in Instagram and the fourth row shows the negative words with the highest eigenvector centrality in Instagram. 46

6.1 Labeling process statistics. Trusted judgments are the ones made by trusted contributors. 58

6.2 Mean values of social graph properties for cyberbullying versus non-cyberbullying samples and aggression versus non-cyberaggression. ($*p < 0.05$). 63

7.1 Cyberbullying detection's classifier performance 71

7.2 Labeling process statistics. Trusted judgments are the ones made by trusted contributors. 74

7.3 Mean values of social graph properties for cyberbullying versus non-cyberbullying samples and cyberaggression versus non cyberaggression, ($^{**}p < 0.05, *p < 0.1$). . . 74

7.4 Cyberbullying prediction's classifier performance. User properties are followers, following and total number of shared media objects. 75

8.1 Collected features for users and seed vines. 80

8.2 Statistics on the Vine dataset, for 55,744 users. 82

8.3 Pearson correlation between the collected Vine features. Only for the vines with $*$ is the p-value larger than 0.001, while for the rest the p-value is smaller than 0.001. . 82

8.4 The tags with highest frequency in 4 different bins of revining. 86

8.5 Mean of user features for cyberbullying and non-cyberbullying class. 92

8.6 Mean of post features for cyberbullying and non-cyberbullying class. 92

Figures

Figure

2.1 Typical public profile on Ask.fm with questions/comments (usually anonymous) to the profile owner and answers from the profile owner. 4

2.2 An example of comments posted on Instagram. To give more room for the text, we have moved the associated image to overlay some of the text. 4

3.1 10 mostly used profanity words by females and males in MySpace dataset 10

4.1 Probability that the total number of likes and question+answer pairs for a user is greater than or equal to k 15

4.2 Total number of likes versus total number of answered questions a user receives . . . 16

4.3 Complementary cumulative distribution function for negative and non-negative interaction graphs/matrices. 20

4.4 Percentage of common users among two subsets of x% highest in-degree and x% highest out-degree. Measurements for Flicker and Web have been obtained from [62] 21

4.5 Cumulative distribution function for ratio of out-degree to in-degree. Measurements for Flicker and Web have been obtained from [62] 21

4.6 Reciprocity for non-negative and negative interaction graphs/matrices 22

4.7 An example of a profile with both positive and negative comments. 23

4.8 Mean local clustering coefficient versus degree 27

4.9 An observation on occurrence of religious cyberbullying in Ask.fm, visualized using Gephi [28]. 28
4.10 An observation on occurrence of appearance cyberbullying in Ask.fm, visualized using Gephi [28]. 32
4.11 Frequency of negative words seen with the word “fat” as an example of appearance-based cyberbullying 32
4.12 Frequency of negative words seen with the word “muslim” as an example of religious cyberbullying 32
4.13 Frequency of negative words seen with words “lesbian”, “bisexual”, “gay” as an example of LGB cyberbullying 33
4.14 Word usage with the word “cut” in Ask.fm. The size of the circle indicates Eigenvector Centrality score 33
4.15 Frequency of negative words seen with word ”cut”. 33

5.1 Complementary cumulative distribution functions (CCDFs) of **follows** and **followed by** of normal and common Instagram users. 39
5.2 CCDFs of the number of shared pictures, number of likes received, and number of comments, of normal and common Instagram users. 40
5.3 CCDFs of the percentage of negative and positive posts by normal and common public Instagram users. Red and purple show positivity and blue and green show negativity. 41
5.4 CCDFs for the number of likes and comments on Ask.fm 41
5.5 CCDFs of the percentage of negative and positive posts by normal and common Ask.fm users. Red and purple show positivity and blue and green show negativity. . 42
5.6 Correlations of profile comments between two social networks for different liwc categories, separated by owner and friend comments ($p < 10^{-5}$). 43

5.7 CCDFs of the percentage of positive posts by common Instagram and Ask.fm users. Pink and purple show common Instagram users and light and dark green show common Ask.fm users. 45

5.8 CCDFs of the percentage of negative posts by common Instagram and Ask.fm users. Pink and purple show common Instagram users and light and dark green show common Ask.fm users. 45

5.9 Correlation matrix, where the characters stand for A: Ask.fm, I: Instagram, P: Positivity, N: Negativity, O: Owner's comment, F: Friend's comments, L: Likes, C: Comments, M: Media, Fb: Followed by, Fls: Follows ($p < 10^{-5}$). 48

5.10 An example of a profile in which negative words from a bystander which is not a cyberbullying question against profile owner. 49

5.11 CCDFs of the percentage of positive and negative posts for Ask.fm users. Pink and purple show positiviyu and light and dark green show negativity. 50

5.12 An example of a profile having posts with negative words from an anonymous user which is cyberbullying question against profile owner 51

5.13 An example of a profile having posts with affectionate negative words from a non-anonymous user 51

6.1 An example of comments posted on Instagram. 54

6.2 An example of the labeling survey, showing an image and its corresponding comments, and the survey questions. 57

6.3 Fraction of media sessions that have been voted k times as cyberaggression (left) or cyberbullying (right). 60

6.4 Two-dimensional distribution of media sessions as a function of the number of votes given for cyberaggression versus the number of votes given for cyberbullying, assuming five labelers. 60

6.5 Percentage of media sessions that have been labeled as cyberaggression (left) and cyberbullying (right) versus their negativity percentages. 62
6.6 Pearson's correlation between the number of votes and the number of comments that arrive in $\leq x$ seconds after their previously received comment. 63
6.7 (left) Ratio of LIWC values of cyberbullying/cyberaggression labeled media sessions to non-cyberbullying/non-cyber aggression class in Linguistic categories. (right)Ratio of LIWC values of cyberbullying/cyberaggression labeled media sessions to non-cyberbullying/non-cyberaggression class in Personal Concerns categories. 65
6.8 Ratio of LIWC values of cyberbullying/cyberaggression labeled media sessions to non-cyberbullying/non-cyberaggression class in Psychological categories. 66
6.9 Fraction of image categories for all media sessions, cyberbullying and cyberaggression classes. 67
6.10 Fraction of images which have been labeled as cyberbullying and cyberaggression for each content category. 68

8.1 An example of the Vine social network with revining. 78
8.2 CCDFs of the number of followers and followings in Vine. 81
8.3 CCDFs of the number of revines, loops, comments, and likes in Vine. 81
8.4 CCDF of the number of tags. 83
8.5 Mean and median number of users who follow back a user versus number of followings. 83
8.6 Mean and median number of original vines versus number of followers. 84
8.7 Mean and median number of original vines versus number of followings. 84
8.8 Mean and median number of revines of the original vines versus number of followers. 85
8.9 Mean and median number of revines of the original vines versus number of followings. 85
8.10 CCDF of maximum number of times a seed vine has been revined per day. 85
8.11 The time lag between the first revine and the original vine. 85

8.12 (left) Percentage of vines posted at time x during a day. (right) Normalized number of likes, revines, comments and loops for posts received at time x in a day. Each graph has been normalized to one by dividing to its highest value. 86
8.13 Visualization of the tags associated with the vines 87
8.14 CCDF of the percentage of the reviners who are the followers of the source of the original vine. 88
8.15 Mean and median number of additional recipients of the vine via revining. 88
8.16 CCDF of number of hops a video get propagated through the network. 89
8.17 Revine tree of followers of a seed Vine video who revined the post. 89
8.18 Mean and median of number of times a video has been revined versus the content of the video. 91
8.19 Mean and median number of times a seed video has been revined versus the emotional content of the video. 91

Chapter 1

Overview

1.1 Thesis Statement

This dissertation shows that achieving accurate prediction of cyberbullying in social networks relies upon comprehensive labeling of multimodal data. It provides insight into the problem of cyberbullying in social networks by analyzing profane and positive word usage in online social networks, investigating the difficulty in labeling media sessions for cyberaggression and cyberbullying behavior, evaluating propagation of such behavior in a social network, and predicting victims of such behavior.

1.2 Research Findings

(1) When people have highly negative profiles without any positive support in Ask.fm, a text-based semi-anonymous social network, they also have the lowest activities in terms of both in-degree and out-degree, that is they are the least sociable.

(2) Higher profanity usage in Ask.fm was observed compared to Instagram, an image-based social network. Targeted users are the same across both social networks. Also a profile owner who is negative on one network generally tends to be negative on the other network, and similarly for positive behavior.

(3) As the percentage of negativity in Instagram media sessions increases, the percentage of cyberbullying in media sessions also increases, up to a maximum of 40-50% negativity. After that, as the percentage of negativity in a discussion increases, the likelihood of cyberbullying

behavior decreases, e.g. media sessions with very high percentage of negativity above 60-70% actually correspond to a lower likelihood of cyberbullying.

(4) There exists high correlation between unigram features extracted from Instagram text comments, such that reducing the dimension from more than two hundred thousands to ten, we achieve higher precision and recall in terms of detecting cyberbullying.

(5) Cyberbullying in Instagram has a higher probability of occurring when multimedia sessions contain certain types of image content. Using image and user features, we are able to predict the likelihood of receiving cyberbullying comments for a media session.

(6) Cyberbullying video sessions in Vine have been shared less, and also propagated less through the network compared to non-cyberbullying videos sessions. In fact, in terms of the mean number of hops of video propagation for different emotions, we found the highest number of hops in average is from positive emotions and the lowest is from negative emotions.

Chapter 2

Introduction

One of the most pressing problems in high schools is bullying. However, with today's online and mobile technologies, bullying is moving beyond the schoolyards via cell phones, social networks, online text, video and images, etc. As bad as fighting and bullying were before the advent of the mobile Internet, the recording and posting of hurtful content online has magnified the harmful reach of bullying, enabling 24/7 bullying. In fact more than half of American teens have been the victim of cyberbullying [17].

Although cyberbullying may not cause any physical damage initially, it has potentially devastating psychological effects like depression, low self-esteem, suicide ideation, and even suicide [35, 25]. For example, Phoebe Prince, a 15 year old high school girl, committed suicide after being cyberbullied by negative comments in the Facebook social network [31]. Hannah Smith, a 14 year old, hanged herself after negative comments were posted on her Ask.fm page, a popular social network among teenagers [93]. Cyberbullying is such a serious problem that nine suicides have been linked with cyberbullying on the Ask.fm Web site alone [13]. Cyberbullying was viewed as a contributing factor in the death of these teenagers [3]. Given the gravity of the problem and its rapid spread among middle and high school students, there is an immediate and pressing need for research to understand how cyberbullying occurs today, so that techniques can be rapidly developed to accurately detect, prevent, and mitigate cyberbullying. While most current studies have focused on the prevalence and impact of cyberbullying in education and psychology [78, 90, 105], our interest is in understanding how social networks are being used to enable cyberbullying. This

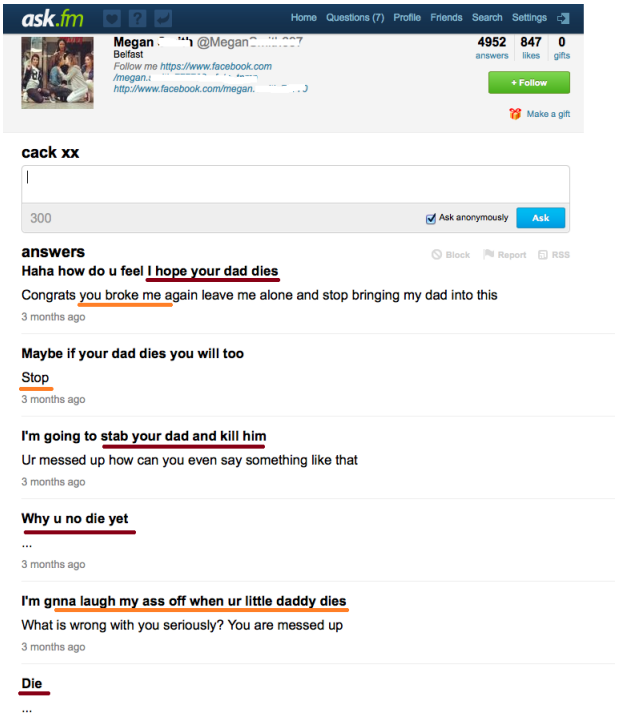

Figure 2.1: Typical public profile on Ask.fm with questions/comments (usually anonymous) to the profile owner and answers from the profile owner.

Figure 2.2: An example of comments posted on Instagram. To give more room for the text, we have moved the associated image to overlay some of the text.

research will identify and characterize cyberaggression (intentional negative electronic behavior) and will explore the extent to which we can identify and characterize cyberbullying. Detection of cyberbullying in OSNs is challenging for a number of reasons. First, the sheer volume of postings in OSNs renders any manual attempts to detect cyberbullying completely infeasible. Second, it is not at all clear what constitutes cyberbullying as it depends on a number contextual factors. Finally, there are a multitude of OSNs that differ from one another in a variety of ways, such as the type of user postings (text, images, videos, etc.), ability to post messages anonymously, methods of responding to user posts (liking individual postings vs. a single response to a group of postings), ability to make friends vs. following a user, whether the primary method of posting is via a mobile phone, etc. Therefore, an effective solution for automatically detecting cyberbullying must account for variations among different OSNs. Figures 2.1 and 2.2 illustrate examples of attacks on the semi-anonymous social network Ask.fm, and the mobile social network Instagram, respectively.

The main goals of this research are to first understand the nature of cyberbullying in online social networks, and then develop methods that can automatically detect and predict incidents of cyberbullying in online and mobile social networks. In particular, this dissertation investigates the definition of cyberbullying, how to conduct ground truth labeling of social network data for cyberbullying, characterization and analysis of the collected data, including a study of how cyberbullying propagates in social networks, and finally automated detection and prediction of cyberbullying.

Chapter 3

Related Work

Most of the previous research on cyberbullying focused on the psychological component of cyberbullying. Only recently have researchers begun to explore computational methods for detecting cyberbullying. Prior computational work on cyberbullying has applied a text analysis approach to online comments. Researchers have collected data from the website Formspring.me, a question-and-answer formatted website that contains a high percentage of bullying content [44, 49, 48], and labeled the data as follows: 1) bullying or normal post? 2) scale of bullying, between 0-10. 3) explicit or implicit bullying? profanity word, if any? 4) optional comment. The labeled data was used, in conjunction with machine learning techniques provided by the Weka tool kit, to train a computer to recognize bullying content. Both a C4.5 decision tree learner and an instance-based learner were able to identify the true positives with 78.5% accuracy.

Researchers sought to detect bullying on Twitter text data [32]. Cyberbullying was detected by looking for inappropriate words using a Näive Bayes classifier. They track potential bullies, their followers and the victims. For this purpose, first all of the sentences that include some term of abuse were found, then it was determined if they had a negative, positive or neutral meaning. Furthermore, they look for information like username, gender, and followers to check the behavior of its connection. If followers also had abusive behavior and any bullying relationship with the same victim, they would be also considered for further monitoring. A small dictionary of four words "Gay", "Homo", "Dike" and "Queer" was considered to classify the tweets. The authors observed that most of the potential bullies have been detected as spammers faster by Twitter's

spam detector and deleted. This fact has made it hard to track their behavior. At the end, an accuracy of 70% in detecting negative tweets was reported, though no mention of the key metric of false positive rate was provided.

Authors in [86, 71] try to detect cyberbullying on an unofficial school website. They gathered and labeled 2,999 entries to train their SVM classifier. Out of these entries, 1495 has been labeled as harmful and 1504 has been labeled as non-harmful. They reported an accuracy of around 88% in their best defined scenario.

A sophisticated exploration of textual cyberbullying has been conducted [23]. Bullies achieve power against victims by exploiting sensitive topics such as: race, culture, sexuality, physical appearance, intelligence, and social rejection. For example the usage of LGBT[1] stereotypes without profanity words highlights the need to move beyond conventional statistical learning methods for effective detection. Instances of cyberbullying where the abuse is more indirect and does not involve the use of profanity or stereotypical words are likely to be misclassified by supervised learning methods. In order to understand the way objects relate to each other, the Open Mind Common Sense (OMCS) project has collected common sense statements from 1999. This knowledge allows understanding hidden meaning implied by comments and detecting normal behaviors. They employ common sense reasoning to detect these more contextual indirect forms of cyberbullying on YouTube and Formspring social networks. They collected a Formspring dataset of both negative and positive samples, with comments from videos on both sensitive and non-sensitive topics. It was observed that supervised machine learning approaches are suitable for sentences that include the negative words, as machine learning works based on generalization, but did not work well and misclassified implicit bullying sentences. They first classified bullying sentences versus non-bullying ones. At the second round, they classified bullying sentences according to the three subjects of sexuality, race & culture and intelligence, since prior research had found that the most common cyberbullying categories among middle and high school students fall into one of the following: sexual, race/culture, intelligence, physical attributes and social rejection [61]. They used Näive

[1] Lasbian,Gay,Bisextual,Transgender

Bayes, a tree based learner and an SVM method. The results show the highest accuracy was obtained with a binary SVM classifier, which could detect Sexuality with accuracy 79%, Race with accuracy 71.8% and intelligence with accuracy 77.2% on the explicit form of verbal bullying of YouTube dataset. Turning to common sense reasoning, they modeled their data mathematically, and considered each concept as a node and the common sense that connects two concepts as the label of the link to construct a graph. For example in the sentence "girls like pink bows", "girls" and "pink bows" are two concepts and "like" is the relation between them. Different combinations of concepts and relationships can produce positive or negative polarity for the sentence. Then, a large matrix with rows showing concepts and columns reveals relationships between concepts, where each element of the matrix indicates whether the corresponding row and column have negative, positive, or zero (when the meaning of the combination is unknown) polarity. Indirect cyberbullying was detected with accuracy 70% and false positive of 30% on the Formspring dataset.

In the phase after detecting explicit or implicit bullying, the researchers propose an idea of how to interact with a bully. Designing a "Reflective User Interface" has been proposed by [89] as a more effective way of interaction, instead of just deleting a bullying post. For example, one approach is to send a notification for the user and propose a better reflection of the situation. It can be like a message that appeared for a user before posting the message online and educating him/her. Other approaches are to notify the victim and give him/her the option of ignoring or deleting the bully, talking with adolescents or reporting abuse to the social network provider. These options can be helpful to make the bully more aware of the consequences of his/her behavior [23].

A limitation of most of this prior work is that it uses a definition of cyberbullying when labeling comments that does not consider either the frequency of negativity or the imbalance of power between the victim and bully. It is more accurate to say that most of the prior work on cyberbullying is more appropriately described as research either on the topic of cyberaggression, which does not consider frequency of negativity or power imbalance, or on negative word analysis. This latter definition suffers further in that not all occurrences of negative words constitute cyberaggression, much less cyberbullying.

All these prior works focused on text-based analysis of cyberaggression, and did not explore social network relationships in their investigation of cyberbullying. Homan **et al.** studied the social structure of LGBT youths with depression in the TrevorSpace social network [36]. As a result, many papers suggest that text analysis alone is not enough for accurately detecting cyberbullying [19, 64, 80, 21], and that there is the potential for substantial improvement by incorporating other features besides text. The research group in Human Media Interaction Group, University of Twente studied textual cyberbullying detection, investigating whether integration of user's profile information improves the accuracy of cyberbullying detection in social networks [19]. Incorporation of the user's information, their characteristics, and post-harassing behavior, for instance, posting a new status in another social network as a reaction to their bullying experience, will improve the accuracy of cyberbullying detection [20]. Figure 3.1 shows males and females have different patterns of profanity word usage. Therefore, incorporation of the user's information, such as age and gender, will improve the accuracy of cyberbullying detection. In their last paper they have collected comments on videos with sensitive topics [21]. For a set of 3825 unique users with in average 15 comments per user they labeled the user as bully or not based on the comments' content. This work is trying to detect users with aggressive behavior, not cyberbullying incidents. They have proposed a system to monitor and raise an alarm whenever someone is detected as bullying, to alleviate the burden of manually looking at all profiles.

Further, some researchers, including us have observed that not all examples of cyberaggression lead to cyberbullying, i.e. some cyberaggressive comments may be directed at bullies instead of victims [39, 112, 111, 64]. Therefore, we need to label roles to correctly detect cyberbullying. Prior work has begun to label roles such as victim, bullier, defender, assistant, accuser, and other [112, 111, 64], but has reported at best detection accuracies less than 50% for the roles of bullies and victim. Preliminary work has also explored anonymity in the context of cyberaggression [64, 116, 76], but it is far too early to establish a strong link between anonymity and cyberbullying.

Authors in [66] sought to find bullies and victims by looking at the in-degree and out-degree of aggressive messages of nodes. They proposed using a semantic and weighted TFIDF features to

Figure 3.1: 10 mostly used profanity words by females and males in MySpace dataset

detect cyberaggressive comments and achieved precision 86% and recall 98% on a balanced MySpace dataset. After using text features for detecting cyberbullying comments, they ranked users as bullies and victims based on number of outgoing and incoming cyberbullying comments. This work did not consider the full definition of cyberbullying and did not explore temporal behavior. Also most of the previous work on textual cyberbullying is on offline detection, even though the ultimate goal is detecting cyberbullying in near realtime. [68] consider semisupervised cyberbullying detection of unbalanced datasets with F-measure 47%. Unfortunately they have not provided enough details so that we can verify this approach.

The other work employing social graph properties of the users besides text features for detection of cyberbullying used a Twitter corpus containing about 900,000 posts of 27135 users[41]. The authors chose 4853 messages where the receiver was known based in the character "@" and the history on both ends is available in the dataset. Precision 76% and ROC 75% was obtained by combining text and social features. One of the interesting findings was that the higher the number of friends for a node, the lower the likelihood of the cyberbullying. On the other hand the higher the interaction between two nodes, the higher the chance of a bullying incident.

Potha, et. al [84, 85] labeled question pairs of sexual predators, the same data as 2010 Kontostathis [47], by providing the complete dialogue not only a window. They ask AMT labelers

to label the comments within four categories. In order to classify a set comment as cyberbullying or not, they look at the similarity of labeled sequence and the SVD features extracted to capture the behavior of individual linguistic information. The highest degree of similarity was observed between the second row of the SVD matrix and the class label vector. They also tried to predict the severity of the next coming message from a predator, by using 1-3 previously posted comments. The correctness of the proposed method was evaluated by RMSFE (Root Mean Square Forecasting Error) and MAPE (Mean Absolute Percent Error).

Another work labeled 85,000 Ask.fm posts, [101, 102, 103], and besides labeling a comment (question/answer) as aggressive or not (3-level: zero, moderate and serious), also labeled the roles and categories. The labeled roles can be victim, harasser, bystander-defender and bystander-assistant. Each post is also labeled for one of the categories: threat, sexual, insult, curse, defense, defamation and encouragement. As a proof-of-concept, the highest F-score obtained was 55.39% for classifying cyberbullying posts from non-cyberbullying posts (precision 51% and recall 60%).

These works largely focus on text-based analysis and unlike our work do not examine the multi-modal features associated with media objects such as images or videos belonging to those comments, as in Instagram.

Chapter 4

Analysis of Negative and Positive Word Usage in a Semi-anonymous Social Network

4.1 Overview

In particular, this chapter chooses to focus on analyzing the Ask.fm social network for the following key reasons. First, Ask.fm is a major source of cyberbullying on the Internet. In fact, it ranks as the fourth worst site in terms of percentage of young users bullied according to a recent survey [100], after Facebook, YouTube and Twitter. Second, very little research has been published to date concerning the Ask.fm social network. Third, Ask.fm is a highly popular and rapidly growing social network, with over 130 million registered users as of September 2014 [7]. Finally, Ask.fm provides publicly accessible data.

We analyzed negative user behaviors at Ask.fm, a popular OSN that has led to many cases of cyberbullying, some resulting in suicide. We examined the occurrence of negative words in users' question+answer profiles along with the "likes" of these questions+answers. Our results indicate that negativity exhibited in users' online behaviors, such as the type of words used in postings, type of postings liked, and level of OSN activities, can potentially help identify cyberbullying.

One of our big challenges in analyzing the Ask.fm social network is that it behaves as a semi-anonymous social network. User profiles are public, but postings to each profile by users other than the owner are by default anonymous. In addition, we cannot obtain from the public profiles which users are following which other users. As a result, it is not possible for us to construct a social graph based on friendships. However, we observe that another type of graph called an

interaction graph [109] can be extracted from the "likes" of comments. We use this insight to build and analyze interaction graphs that embed social relationship information, to help identify negative user behavior in this semi-anonymous social network.

This chapter makes the following contributions. It is the first work to provide a detailed characterization of key properties of the important Ask.fm semi-anonymous social network. Second, it builds and analyzes interaction and word graphs and finds that properties of the interaction graph such as in-degree and out-degree are strongly related to the amount of negative user behavior expressed on a profile, i.e. highly positive profiles exhibit the highest degree of sociability in terms of liking others and being liked by others, whereas profiles with a high number of negative questions exhibit the lowest degree of sociability.

4.2 Data Collection

In this section, we discuss key aspects of the Ask.fm social network, what information was collected from Ask.fms publicly accessible profiles, e.g. questions, answers, and likes, and our process of building interaction and word graphs.

4.2.1 Ask.fm

We observe that Ask.fm is generally an example of a semi-anonymous social network, in that the identity of users who post questions/comments to a profile is typically anonymous (though posters may choose to reveal their identity, we have seen this happen only rarely), whereas the identity of the target user is publicly known. People can search Ask.fm users via their name, id or email address. This is unlike non-anonymous social networks such as Facebook and Twitter, where the identity or ID of posters and target users are both publicly visible. Thus, in Ask.fm, any one (even people without an account) may post on another user's profile, and this posting of a question or comment is usually done anonymously (in fact, that is the default).

There are some policies specific to Ask.fm. First, only the target user may post an answer to a question/comment on this site. Only after answering a question will the question and answer appear

on his/her profile. Further, a user may choose to "like" at the granularity of a question+answer pair, but cannot like the question nor the answer individually. Liking is non-anonymous, so that the identity of the likers is publicly known. Another feature of Ask.fm is that users may follow other users. However, this relationship data is not available publicly and only the profile owner knows who he/she is following. Even the user who is being followed can only know how many followers he/she has (which is not publicly available to other users), not who is following him/her.

4.2.2 Description of Collected Data

The data we can extract from a common profile includes the following fields: userID, personal information (if any, as it is optional), total number of answers, total number of likes, content of answered questions posted on a user's page, and the userID of people who liked the questions+answers.

An interaction graph can be constructed from the "likes" of answered questions, i.e. each directed edge in the graph connects user i to neighbor j if user i has liked a question+answer pair in j's profile. Note that the edges are not bidirectional, so that i liking an question+answer on j's profile does not imply j liking one of i's question+answer. In order to extract this interaction graph from Ask.fm, we conducted a breadth-first search starting from a couple of random seed nodes. Seed nodes should have non-zero liked question+answer pairs. For each seed node we found all nodes that liked an answered question on its page (incoming edges are publicly known for each profile). However by only looking at a user's profile, we are not privy to any of the nodes that this profile owner liked (outgoing edges). In the second step, we collect the profile information of all neighbors of the seed node. These steps are then repeated. Ultimately, the outgoing edges from a profile can be reconstructed from the incoming edges of other profiles. Note that because we're crawling profiles using breadth-first search, we can only find that subset of the outgoing edges for each profile that happen to be incoming edges on other crawled profiles. The only way to find all outgoing edges for each profile is to crawl the entire Web site, which is impractical. Since the breadth-first search is terminated before crawling all profiles, then this is called snowball sampling, and results in an interaction graph wherein all the internal nodes have been fully crawled

or sampled, but also a typically small fraction of nodes on the edge that have not yet been crawled. Our analysis below focuses only on fully sampled nodes in the interaction graph. Using snowball sampling, 30K profiles were crawled from October to December 2013.

To provide some general context, we first perform an analysis of the total number of question+answer pairs and likes per Ask.fm user. We observe they both have heavy tail distributions in Figure 4.1. The total number of likes has a heavier tail than the total number of question+answer pairs. Looking at the tail, we observe that as the total number of answered questions in a user profile increases, the total number of likes also grows, Figure 4.2. The red line is the best fit obtained through linear regression in the transformed (log-log) space. It seems when a user is more popular and active, they receive more questions, and more user visits and likes on his/her page. Also pushing a like button is easier than writing a question, which makes the number of likes larger than the number of questions.

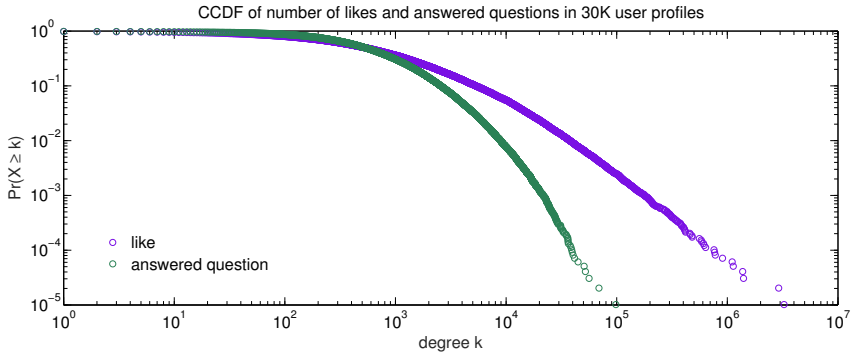

Figure 4.1: Probability that the total number of likes and question+answer pairs for a user is greater than or equal to k

Figure 4.2 describes the correlation between the number of answered question+answer pairs and the number of likes. The correlation value when the number of answered questions is less than 50 is -0.05. This shows that when the number of answered questions in a user page is low, the number of people who like these users does not have any special relation with the number of answered questions. When the number of answered questions is larger than 50, the correlation between the number of likes and the number of answered questions increases to 0.33. This means

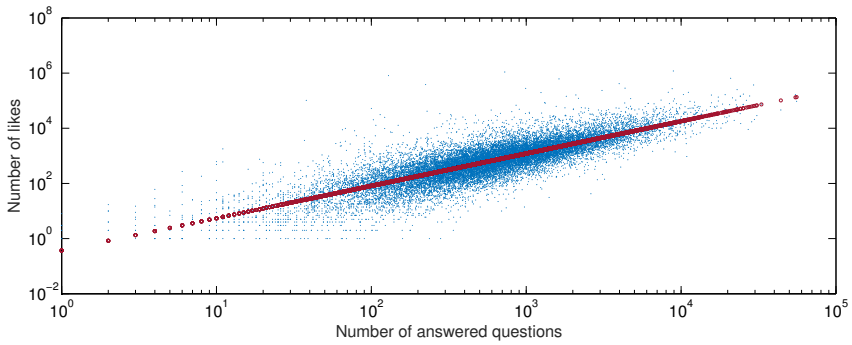

Figure 4.2: Total number of likes versus total number of answered questions a user receives

that as the number of question+answer pairs increases beyond around 50, the number of likes becomes weakly related to the number of answered questions.

4.2.3 Modeled Interaction Network

The preceding analysis provided a general overview of likes and answered questions in Ask.fm, but we are more interested in likes and answered questions in the context of cyberbullying. Our search for cyberbullying on the Ask.fm is based on the supposition that repetitive negative words represent the core of the abusive text posted on profiles. Following the occurrence of negative words led us to many examples of cyberbullying. However, after a preliminary analysis of the answers by the profile owner, we found that many such examples seemed to have no effect on the targeted user, namely based on the answers the target seemed indifferent to the negative comments. In contrast, there were other more serious cases where the target seemed particularly vulnerable, making statements like "you broke me" (See Figure 2.1).

Looking at the profiles, we observed that the users who seemed most vulnerable to negative questions were often those who were most isolated, with few "likes" and also rarely liking others' comments. In contrast, users who were subject to but seemingly indifferent to negative questions appear to have a fair number of likes and also seemed somewhat active in liking others' comments.

Based on these two observations that (i) cyberbullying often involves posting questions with negative words and (ii) vulnerable targets of cyberbullying (based on their answers) seem isolated,

we sought to build and analyze social graphs and word graphs derived from our data that would capture the negativity and isolation of users.

In order to capture the greatest degree of interaction between the users, we collected the top 15 questions for each user that had the highest number of likes. For some users with very low activity, the number of highest liked questions was less than 15. Analyzing the most popular questions of profiles, we have built our network modeled as a directed bipartite network. In order to build our graph, if user i likes a question+answer in the page of user j, then there is a link from i to words on that question and a link from those words to node j.

From this bipartite graph, we seek to derive characteristics that can highlight the negativity and isolation associated with targeted users. In order to project a bipartite network with adjacency matrix B, to the network of words W, we have $W = BB^T$. Then we can similarly build the network of users with adjacency matrix U from our bipartite network. That is, based on this data, our idea is to build and use a like-based interaction graph between users (U) and examine the balance of in-degree vs out-degree of the users with high degree of negativity in their pages and small number of positive posts; such users have both low in-degree and low out-degree in graph terminology. In contrast, users who received on average the same amount of negativity but have positive questions at the same time, show healthier in-degree and out-degree.

4.2.4 Extracting Positive and Negative Words

We next consider which negative words are pertinent to our analysis of negative user behavior. The natural approach is to select those negative words that have the highest frequency in the sampled profiles. However, we observe that it is the collective effect of negative words that is exploited by cyberbulliers. Negative words that may be commonly used but more in isolation rather than in a collective fashion would be less likely to create a strong effect of cyberbullying. Therefore, we sought to create a word graph that measures the relationship between words, i.e. are they being used together on the same profile to bully a victim. By one mode projection from a bipartite graph comprised of 30K users and a dictionary of around 1500 negative words (obtained

from [110, 1]), we constructed a word graph wherein each word signifies a node in the graph and each edge indicates that the two words have been used in the same profile.

We found that there is a cluster of connected negative words in the center of the word graph, but there are also many negative words that were not connected to any word. Out of 1500 words, the eigenvector centrality of 968 words is zero, which means 968 negative words either have not been seen in any profile or have not been seen together in a profile. These words were eliminated from our analysis. The top row of table 4.1 shows the remaining negative words that had the highest frequency of appearance. The second row of Table 4.1 shows the negative words with the highest eigenvector centrality. Though there is a fair degree of overlap with the highest frequency words, we note that the sets do indeed differ. Since eigenvector centrality captures to some extent the collective negativity of cyberbullying, we focused our word graph analysis below on the 80 negative words with eigenvector centrality values larger than 0.5. A similar approach was followed for a collection of 1000 positive words. 80 positive words with highest eigenvector centrality were chosen for the following analysis.

hoe	n**ga	stupid	slut	kill	p**sy	cut	suck	gay	fat	sex	bad	d**k	die
f**k	ass	bitch	s**t	hate	d**k	ugly	sex	bad	suck	p**sy	fat	gay	stupid

Table 4.1: The first row shows the negative words with the highest frequency and the second row shows the negative words with the highest eigenvector centrality (values decrease from left to right).

4.3 Network Statistics

4.3.1 Building the Interaction Graph

We collected and analyzed about 30K profiles that had a complete list of likes, using the snowball sampling method, gathering the top 15 most liked questions. We found that there were on average 14.5 most liked questions per user, that is most users had close to 15 questions with likes. We also focused on a collection of 80 negative words and 80 positive words with the highest eigenvector centrality as explained previously.

We found that 46.0% have at least one question including one of the top negative words. We

Average number of answers per user	14.5
Average number of negative questions per user	0.778
Average number of positive questions per user	4.57
Average number of negative words per user	1.04
Average number of positive words per user	5.42

Table 4.2: Average number of questions and words per user

term such a question a "negative" question. Furthermore, 7.74% of the users have at least three questions with top negative words, and 96.0% users had at least one question with a top positive word in their profiles. We term such a question a "positive" question. Table 4.2 shows the average number of positive and negative questions and words per user.

By a one mode projection of the bipartite graph, we were able to obtain the like-based interaction graph U between these 30K users in the Ask.fm social network. In the adjacency matrix U, created from the users' interaction graph, edges are weighted. In fact the weight of each edge e_{ij} , from node i to node j is a vector $[n_1\ n_2]$, where n_1 is the number of questions in node j including at least one negative word (from our selected dictionary) and n_2 is the number of questions with no negative word liked by node i. Since we are most interested in negative behavior, we group positive and neutral behavior together into a "non-negative" category.

For further analysis, we have divided matrix U into two matrices U_{neg} and $U_{non-neg}$, where U_{neg} (negative adjacency matrix) is the adjacency matrix with weights n_1 and $U_{non-neg}$ (non-negative adjacency matrix) is the adjacency matrix with weights n_2. Figure 4.3 shows the CCDF for the in-degree and out-degree distribution of matrices U_{neg} and $U_{non-neg}$. Authors in [62] show that in-degree and out-degree distributions in social networks are approximately the same. We can see from Figure 4.3 the in-degree and out-degree for non-negative degree distribution is approximately the same for non-negative degree distributions. **However, the negative in-degree and negative out-degree distributions of the interaction graphs clearly differ, indicating that negative behavior in the Ask.fm semi-anonymous social network has different properties than previously reported in other social networks**.

Figure 4.4 shows the percentage of common users among x% highest in-degree and x% highest

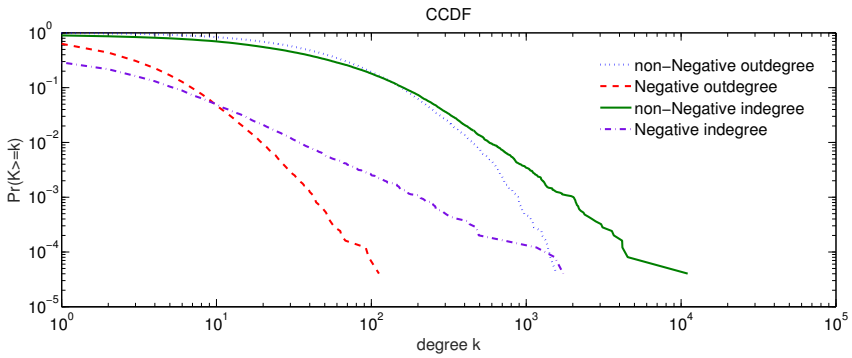

Figure 4.3: Complementary cumulative distribution function for negative and non-negative interaction graphs/matrices.

out-degree. In fact this figure shows whether the users with high out-degree are the same people with high in-degree. Authors in [62] show in social networks for users with 1% highest in-degree and out-degree, more than 60% of users are common. This value is less than 20% for the Web. Here we observe that Ask.fm's semi-anonymous social network exhibits behavior that is between previously analyzed social networks and the Web. Overlap of the users with 1% highest in-degree and out-degree is more than 20% and less than 30%. It seems the correlation between in-degree and out-degree in an interaction graph built from like-based interactions is much less than the correlation in friendship-based social graphs. Friendship is usually a symmetric relation. However, in like-based interactions symmetry is less probable, which causes less correlation between in-degree and out-degree. In addition, there is the ambiguity in Ask.fm as to whether a like is for a question or answer. This decreases the correlation between in-degree and out-degree because we do not know whether the liker is liking the question as a support of the question's content or the answer to defend the profile owner.

Figure 4.5 shows the CDF for the ratio of out-degree to in-degree. It has been shown in [62] that for social networks more than 50% of users have in-degree within 20% of their out-degree. However in the Ask.fm this value is around 16%. Again we observe that the relationship between in-degree and out-degree in Ask.fm is weaker than what was found for prior social networks and is stronger than the Web.

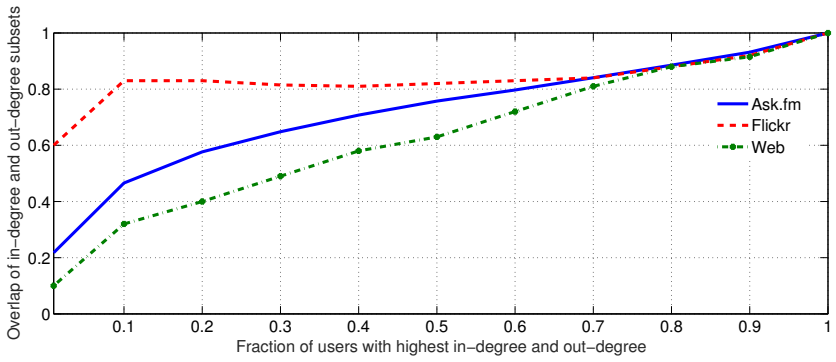

Figure 4.4: Percentage of common users among two subsets of x% highest in-degree and x% highest out-degree. Measurements for Flicker and Web have been obtained from [62]

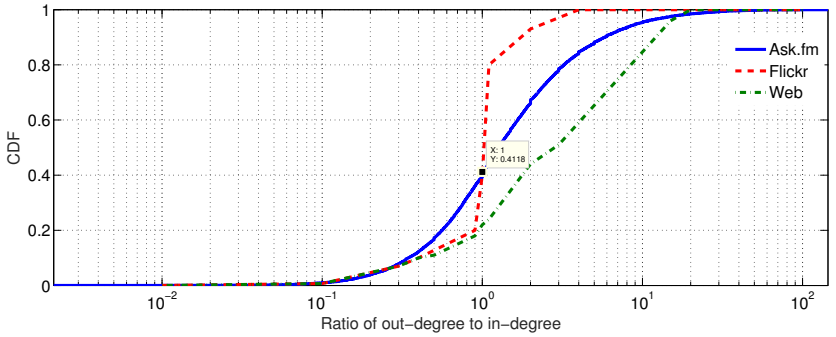

Figure 4.5: Cumulative distribution function for ratio of out-degree to in-degree. Measurements for Flicker and Web have been obtained from [62]

The mean reciprocity of the interaction graph/matrix U is 28.2%, which is a low number compared to other social graphs like Yahoo! 360 with reciprocity 84% and Flickr with reciprocity 68%, [54]. However Twitter has a more similar structure to Ask.fm (in the sense that there is no friendship concept between users) and has an even lower reciprocity equal to 22.1% [54]. Ask.fm's network negative reciprocity is a very low 3.61%, which shows how much users like each others' negative questions. This gives an insight that users who both have negative posts do not tend to like each others' negative questions. Reciprocity between non-negative questions is a far higher 27.9%.

Figure 4.6 shows the mean reciprocity for the two graphs/matrices U_{neg} and $U_{non-neg}$ versus out-degree. It shows that in normal (non-negative) behavior the more active a user is, the more

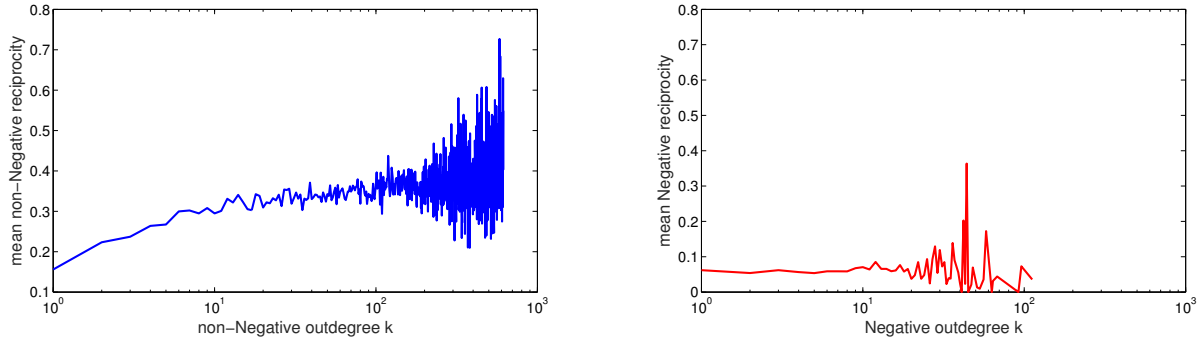

Figure 4.6: Reciprocity for non-negative and negative interaction graphs/matrices

he/she will receive likes from other users or vice versa. On the other hand, the mean reciprocity does not increase with out-degree in the negative matrix. When the negative out-degree is high, this means that a user is active and likes others user's negative questions. However this type of active user receives a low negative in-degree in return. We consider two likely possibilities that explain this result. First, the user is supportive and popular, liking answered questions often to show support for the answerer of a negative question. Due to the lack of liking granularity in Ask.fm, this like is recorded as a like of a negative question, even though the intent of the user was to like the answer. Indeed, we found cases on Ask.fm where a profile owner asked a liker why they liked a negative question, and the liker responded that in fact they were liking the profile owner's response instead. Such a supportive user would be unlikely to receive many negative questions on their profile. Alternatively, the user is a bully and unpopular, frequently liking other's negative questions, and not receiving many likes of negative questions in return. Again, we see that negative behavior follows a different pattern than non-negative behavior.

Next, we explore the relationship between the degree of negativity (and positivity) on a profile and the profile's graph properties. We compute the average number of negative questions in a set of profiles with negative questions that the profile owner answers to show his/her unhappiness about the questions. The average number of negative questions in 150 profiles of this type was 2.87 and therefore we chose a threshold of 3 negative questions to define negative groups. In fact, we

segmented the user base into 4 different groups:

(1) Highly Negative: users with at least 3 negative posts and no positive posts (HN)

(2) Highly Positive: users with more than 10 positive posts (HP)

(3) Positive-Negative: users with at least 3 negative posts and more than 4 positive posts (PN)

(4) Others (OTR)

The definition of these groups helps us identify the properties of the most negative profiles, namely the HN users who have no positive support, while also allowing us to contrast them with the graph properties of other users who have some or a lot of positive words in support. They were chosen based on our observations that targets of negative behavior could be roughly divided into a group that receives support from bystanders and a group that has been left alone and doesn't receive any support with positive posts. Figure 4.7 shows an example of a mixed profile with both positive and negative comments, representative of the PN group.

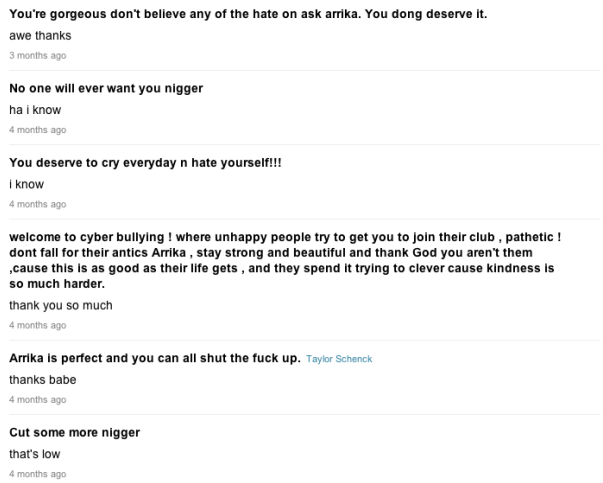

Figure 4.7: An example of a profile with both positive and negative comments.

Table 4.3 summarizes some of the key results of our graph analysis. We measure the average reciprocity of the four different groups, as well as the in-degree and out-degree, based on the interaction graphs U_{neg} and $U_{non-neg}$ calculated earlier. Note the in-degree has two subcategories,

	HN	HP	PN	OTHR
Negative reciprocity	0.131	0	0.106	0.066
Non-Negative reciprocity	0.210	0.379	0.246	0.339
Negative in-degree	6.39	0	14.0	2.16
non-Negative in-degree	21.1	112	49.4	70.1
Negative out-degree	3.48	2.73	3.73	3.59
non-Negative out-degree	32.8	117	46.0	70.1
Total number of likes	1027	5995	1765	2906
Ratio of likes per answer	1.59	2.58	1.94	2.73

Table 4.3: Average reciprocity and degree for different groups

pertaining to negative in-degree and non-negative in-degree, i.e. a user X will have a negative in-degree if another user likes one of the questions posted on X's profile that has a negative word in it, while another user liking a question without a negative word will count towards the non-negative in-degree of X. Similarly, out-degree has negative and non-negative subcategories.

From Table 4.3, a key finding is that HN users are distinguished by having the smallest total (negative plus non-negative) in-degree and the smallest total out-degree. It shows they are either not popular in terms of being liked or do not tend to have activity in this social network in terms of liking others. That is, the high degree of negativity that these users are subject to is related to less sociable behavior on this social network. Our results indicate that we could leverage low total in-degree and low total out-degree of Ask.fm's interaction graph to suggest a greater likelihood of determining highly negative profiles of cyberbullying victims on this semi-anonymous social network.

Another major finding is that as the amount of positive support increases, we find a greater in-degree and greater out-degree of the users, that is users become more social and actively like and are liked more often. This is demonstrated first by the PN group, which like the HN group has at least 3 negative posts, but the PN group also has positive support. We observe that the PN group has increased activity in terms of about twice the amount of total in-degree and total out-degree as the HN group. Highly positive HP profiles exhibit the highest sociability in terms of actively liking other profiles and being liked by other profiles. We see that

the total in-degree and the total out-degree are both over 100, which is much higher than the second nearest group OTHR, which has a total in-degree and out-degree around 70 each. This confirms the trend that higher positivity is strongly related to higher sociability on Ask.fm's interaction graph.

An interesting result of this analysis is that we can distinguish the HN group from the PN group not merely by the total in-degree and total out-degree, but also by just the negative in-degree. We see that the negative in-degree of HN is 6.39, while PN's negative in-degree is 14.0, over twice as large. While the mean number of negative questions is the same in both HN and PN, i.e. about 3, why would there be so many more likes for the same number of negative questions for the PN group? The explanation for this lies again we believe with the limited granularity of liking in Ask.fm. In the case of liking a question+answer pair, where the question/comment is negative and the answer is positive, then it would not be clear whether a user is liking a bullier's negative comment or supporting a positive answer from the target of bullying. The higher volume of likes of negative questions we believe is actually users liking the target's positive response. This agrees with an examination of a variety of examples, where we saw that the likes were mostly representative of support and a positive sign in this case.

We also observe that HP users have little interaction with negative posts. That is, their negative out-degree is clearly lower than any of the other three groups. That means that not only do HP users not have negative questions posted on their profiles, but **HP users spend very little effort liking negative questions on other users' profiles, focusing the vast majority of their effort on liking positive questions.** Looking at negative reciprocity, it is clear that HP has negative reciprocity 0 as it does not have any negative in-degree. In the group OTHR it seemed that there existed either a set of users with 1 or 2 negative questions without any limitation on the positive posts, or a set of users with more than 2 negative questions and less than 10 positive questions. However, the average number of negative questions is 0.774, which shows we have mostly users belonging to the first set. We see that the reciprocity on negative questions in this group is lower than the HN and PN groups. The reason is either because they have few negative questions compared to the HN and HP groups (which in average have less), or they have less support (they

have in average 4.76 positive questions compared to 6.53 positive questions for the HP group).

In order to calculate the local clustering coefficient for each node we first turned our network into a simple graph. It means either node i has liked node j's question, or node j has liked node i's question, regardless of the number of likes. We set $U_{ij} = 1$ and $U_{ji} = 1$, otherwise $U_{ij} = U_{ji} = 0$. We make this assumption that either user i receives a like from user j or posts a like on his/her profile, they are in the general category of having some familiarity or "connectedness". The expected clustering coefficient of the network is 0.11 and the averaged local clustering coefficient 0.356, as defined in [69]. Comparing with numbers reported by [60], shown in Table 4.4, the clustering coefficient of Ask.fm is pretty small. In Figure 4.8 the local clustering as a function of degree has been depicted. As we expect, the local clustering coefficient decreases when the degree increases in social networks.

social media	Ask.fm	Facebook	Twitter	Gplus	Flickr	Orkut	You Tube
Clustering Coefficient	0.356	0.606	0.565	0.490	0.313	0.171	0.136

Table 4.4: Local clustering coefficient for some social networks. Clustering coefficient values for Flickr, Orkut and You Tube are from paper [62] and values for Facebook, Twitter, Gplus are from the website [60].

HN	HP	PN	OTHR
0.499	0.237	0.380	0.357

Table 4.5: Local clustering coefficient for defined group

Looking at the clustering coefficient of each group in Table 4.5, we observe that among 4 defined groups, HN has the highest mean local clustering coefficient despite having the lowest degree, while the group HP with highest degree has the lowest mean local clustering. The implication is that users of the HN group probably have a few people that know each other. However, users of the HP group are very social and know many people, and therefore the proportion of their friends that know each other is small. In general, we can see each group that has higher degree has lower mean local clustering.

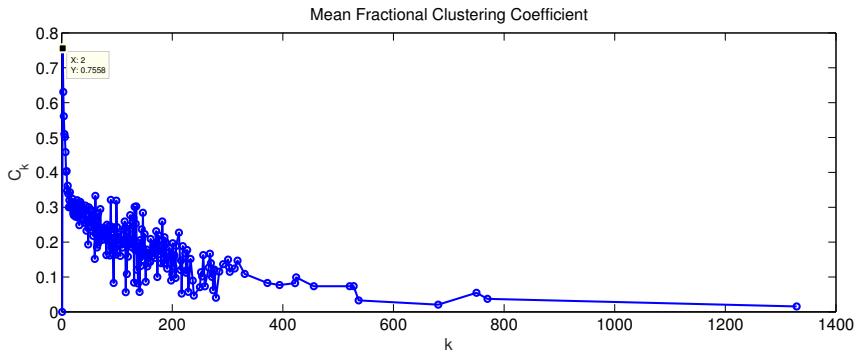

Figure 4.8: Mean local clustering coefficient versus degree

4.4 Cyberbullying Types

We were also interested in understanding whether different types of cyberbullying could be discerned from our data. Cyberbullying manifests itself in a variety of different forms, including racial, religious, sexual and appearance-based cyberbullying. Recall that this words graph has been built with our initial 1500 negative words set. Based on our initial analysis, we believe that groups of words tend to be used together for certain types of cyberbullying, and that these word groups may be sufficiently distinctive to be used as a basis for identifying some kinds of cyberbullying. Our thinking was to develop word graph models that are representative for each kind of cyberbullying.

Towards this end, we constructed word graph models centered around certain types of bullying words. For example, Figure 4.9 illustrates a word graph clustered around the word "muslim", which is associated with religious cyberbullying, while Figure 4.10 illustrate a word graph based around the "fat", which is associated with appearance-based cyberbullying. Words with higher eigenvector centrality are shown with larger diameters. We observe that the two graphs exhibit clearly distinct patterns that differ in the number of words associated with the core word, which words are clustered around the core word, and the strength of association between the words and the core word.

Another way to visualize the differences between religious, appearance-based, sexual (LBGT) and racial cyberbullying is to analyze the distribution of negative words associated with each type of cyberbullying. For this purpose we first find all profiles with a specific word among the 30K

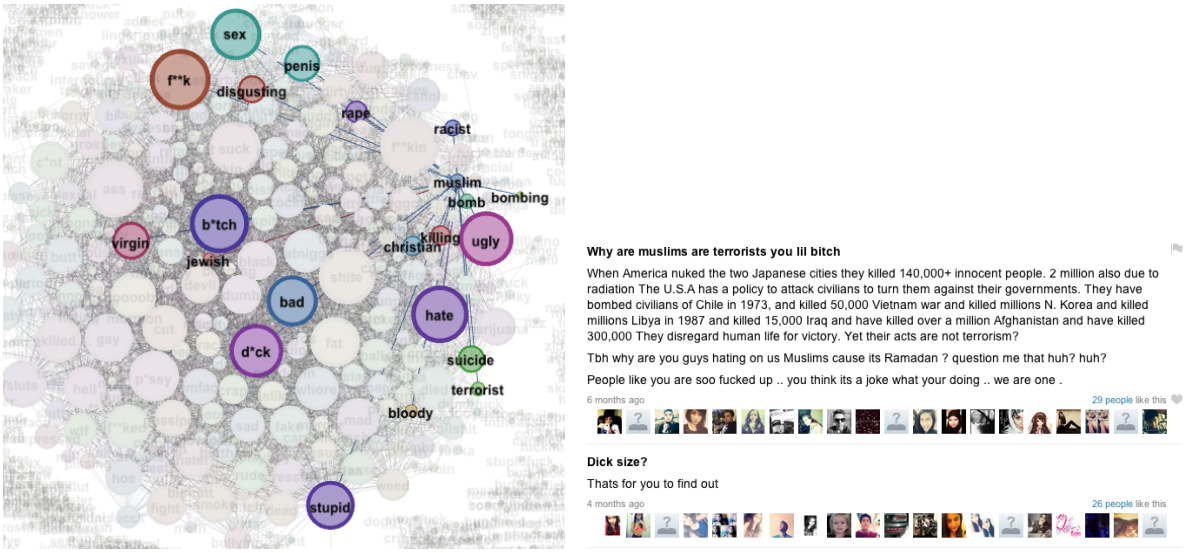

Figure 4.9: An observation on occurrence of religious cyberbullying in Ask.fm, visualized using Gephi [28].

gathered profiles, then for those profiles find the distribution of the 80 negative words of our reduced collection and compute an average over them. Figure 4.11 shows the frequency of negative words seen with the word "fat", while Figure 4.12 plots the frequency of negative words seen with the word "muslim". The top 80 negative words are constant across the x-axis of both graphs. We see that the word "muslim" is used with words in the 0-10 and 40-50 ranges that do not show up for "fat"-based cyberbullying.

Even within a category of cyberbullying, there can be distinctions. For example, in sexually-based LGBT cyberbullying, Figure 4.13 shows the frequency of negative words for three separate words: "lesbian", "bisexual" and "gay". We observe that the distributions are distinguishable both in terms of the presence/absence of peaks at specific word IDs, and also in terms of the height of peaks. In particular, the word "lesbian" has strong peaks of association with the words "ugly" and "f**k" that are absent or only half as strong as for sexual cyberbullying using the word "gay". Also, even though the "fat"-based distribution's shape resembles the "lesbian"-based distribution, the distinction is that the peaks of the former are half the frequency values of the latter.

4.5 Users with Cutting Behavior

One of the most disturbing behaviors that we encountered was the problem of "cutting" (slicing one's wrists). Looking at the profiles of these cases and studying their answers, it seemed that the profile owner exhibited weak confidence and depression problems, sometimes admitting to earlier failed attempts at suicide. In this section we first found 150 profiles for which their owners have explicitly expressed the experience of "cutting" behavior and label them. A graduate student labeled the profile as "cutting" behavior if and only if the profile owner has expressed explicitly in his/her answers that he/she has had such an experience. We observe in Figure 4.14 that among the words that the word "cut" has been connected to are the words "depress", "stressful", "sad", and "suicide" . This association could be used to detect these type of users.

The frequency of negative words used with "cut" has been shown in Figure 4.15. We can see these user profiles have two peaks at words "ugly" and "f**k".

Average number of questions per user	15
Average number of negative questions per user	2.31
Average number of positive questions per user	5.29
Average number of negative words per user	3.38
Average number of positive words per user	6.65

Table 4.6: Average number of questions and words per user with "cutting" label

Table 4.6 illustrates the statistics of the labeled users for "cutting" behavior. We found that such profiles have positive questions close to PN profiles however the average number of negative posts is less by a factor 0.65 compared to the PN and HN groups. This suggests surprisingly but encouragingly that these profiles do receive more support and less negative posts compared to general HN users.

Table 4.7 shows the statistics of the average reciprocity, in-degree and out-degree for cutting victims. **Though we originally expected this group to exhibit behavior similar to the HN group, we found instead that this cutting group appears to exhibit collectively a different behavior than the HN, HP, PN, and OTHR groups measured in Table 4.3**.

The in-degree is 1.5 times more than PN's in-degree and the out-degree is also more by a factor of 1.4. In fact total in-degree and out-degree is more similar to the group OTHR, though there is a marked difference in negative in-degree compared to OTHR. This group has an average number of negative and positive questions most closely related to PN profiles. However, it receives more likes (higher in-degree), and exhibits more activity (higher out-degree) compared to PN groups. This deserves further investigation to explain the reasons behind the differences.

	users with "cutting" label
Negative reciprocity	0.146
Non-Negative reciprocity	0.248
Negative in-degree	17.9
non-Negative in-degree	78.7
Negative out-degree	4.42
non-Negative out-degree	65.0

Table 4.7: Average reciprocity and degree for users with "cutting" label

4.6 Summary

As far as we are aware, this work is the first to present a detailed analysis of user behavior in the Ask.fm social network. We analyzed nearly 30K profiles of Ask.fm users using interaction graphs, word graphs, and frequency distributions, and characterized key properties such as reciprocity, clustering coefficient, and the influence of negativity on in-degree and out-degree. Some of the key findings of the work are that (1) When people have highly negative profiles without any positive support, they also have the lowest activities in terms of both in-degree and out-degree, that is they are the least sociable. (2) As the amount of positive support increases, we find a greater in-degree and greater out-degree of the users, that is users become more social and actively like and are liked more often. For example, users with negative profiles that also receive positive support have higher in-degree and out-degree than owners of highly negative profiles that lack positive support, that is they are more sociable. When people have highly positive profiles, then they also have the highest in-degree and out-degree, which shows that they socialize the most on this social network. This suggests that we may be able to use the interaction graph's in-degree

and out-degree as an indicator of the extent of negativity or positivity on a given profile. (3) The negative in-degree and negative out-degree do not exhibit similar behavior, unlike the similarity of in-degree and out-degree found in other social networks. (4) Finally, our analysis of cutting behavior on Ask.fm reveals that such profiles have surprisingly high positive support, and exhibit a different signature than the other group segments studied.

Acknowledgments

Thank to PhD student, Amir Ghasemianlangroodi for his contribution to this project. I would also like to thank all the cyberbullying project members.

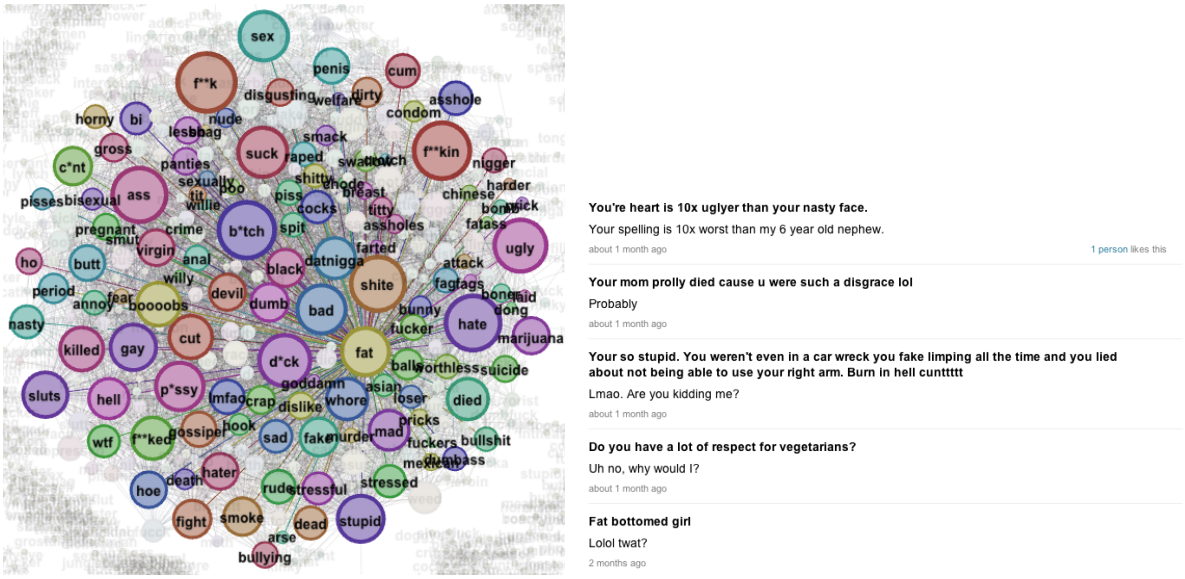

Figure 4.10: An observation on occurrence of appearance cyberbullying in Ask.fm, visualized using Gephi [28].

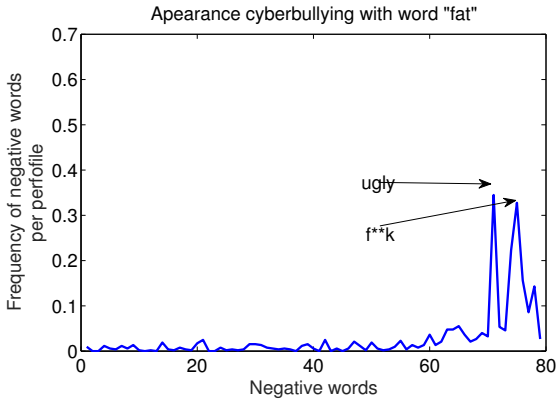

Figure 4.11: Frequency of negative words seen with the word "fat" as an example of appearance-based cyberbullying

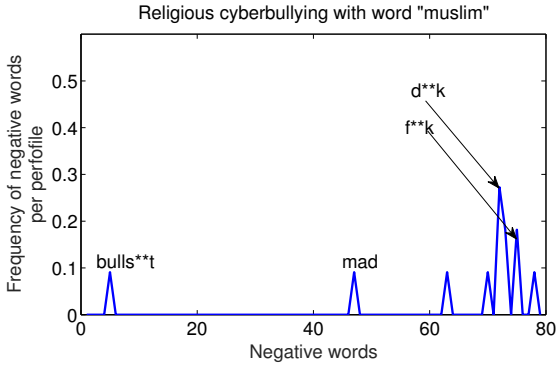

Figure 4.12: Frequency of negative words seen with the word "muslim" as an example of religious cyberbullying

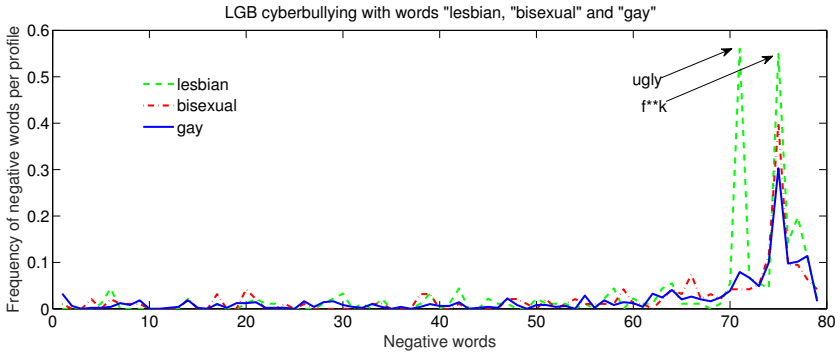

Figure 4.13: Frequency of negative words seen with words “lesbian”, “bisexual”, “gay” as an example of LGB cyberbullying

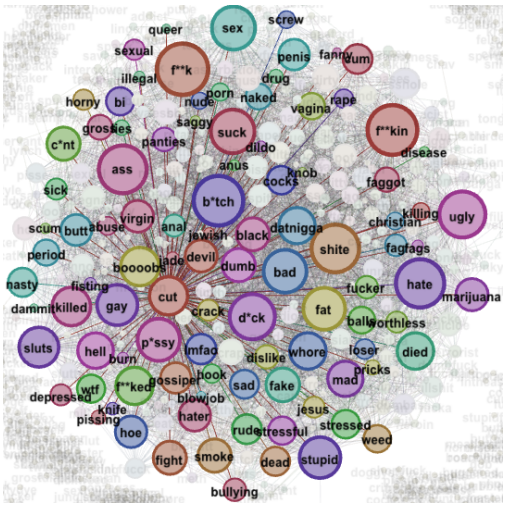

Figure 4.14: Word usage with the word “cut” in Ask.fm. The size of the circle indicates Eigenvector Centrality score

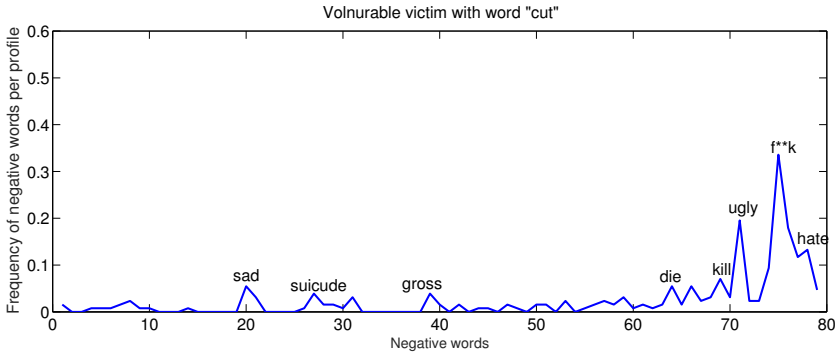

Figure 4.15: Frequency of negative words seen with word ”cut”.

Chapter 5

Analysis of Negative and Positive Word Usage Across Social Networks

5.1 Overview

In the previous chapter we analyzed negative user behaviors at Ask.fm, a popular OSN that has led to many cases of cyberbullying, some resulting in suicide. We examined the occurrence of negative words in users' question+answer profiles along with the likes of these results indicating that negativity exhibited in users' online behaviors, such as the type of words used in postings, type of postings liked, and level of OSN activities, can potentially help identify cyberbullying.

In this chapter we tackle the challenge of understanding how negative user behavior varies across different OSNs, in particular the two OSNs Ask.fm and Instagram. Both OSNs are very popular among teenagers and also rank among the top sites for cyberbullying [100]. These two networks differ from each another in some significant ways. Ask.fm is a semi-anonymous network in which user profiles are public, but postings to these profiles by non-owner users are anonymous by default. On the other hand, Instagram makes the identities of user postings public. Furthermore, users in Ask.fm can post questions and profile owners can provide answers to those questions individually, while users in Instagram can share and comment on media objects such as images. We define four classes of users across these two social networks:

(1) **normal Instagram users**: A sample set of Instagram users with public profiles (2) **normal Ask.fm users**: A sample set of Ask.fm users (3) **common Instagram users**: A sample set of Ask.fm users, for whom we also have collected their Instagram profile data (4) **common Ask.fm users**: A sample set of Instagram users, for whom we also have collected their Ask.fm

profile data

We then compare the posting behaviors of normal and common users in each social network, as well as the posting behaviors of common users across both networks. Together, these comparisons provide a deeper understanding of user behavior both within each OSN as well as across two different OSNs.

Our analysis in this work has revealed the following five important findings. First, we show that more negativity is exhibited at Ask.fm than at Instagram. Second, we show that there is no significant difference in the positivity exhibited by the users in the two social networks. Third, we show that within each social network, common and normal users tend to behave similarly in terms of their positivity and negativity. Fourth, we show that anonymity tends to result in lower negativity. This finding is quite counter-intuitive and we provide some possible reasons for it. Finally, we provide a detailed correlation analysis among different user behaviors and activities in the two networks.

5.2 Methodology

This section explains the design methodology of our study, including the specific questions that we aim to answer, key features of Instagram and Ask.fm, and the data collection process. Detailed analysis results are then presented in the following sections.

To effectively detect cyberbullying in OSNs, we need to understand if and how cyberbullying is affected by OSNs with different features. We chose to study Instagram and Ask.fm in this work, because both networks are very popular among teenagers and have increasingly been used for cyberbullying. More importantly, we have noticed that some Ask.fm users mentioned their Instagram IDs in their profiles, thus allowing for direct comparison of user behaviors across these two social networks.

In this work, we consider four different types of users:

- **Normal Instagram users**: A sample set of Instagram users with public profiles

- **Normal Ask.fm users**: A sample set of Ask.fm users
- **Common Instagram users** who are also Ask.fm users
- **Common Ask.fm users** who are also Instagram users

We are interested in answering the following questions:

(1) Do common and normal Instagram users differ in their posting behaviors?

(2) Do common and normal Ask.fm users differ in their posting behaviors?

(3) Do common Instagram users and common Ask.fm users differ in their posting behaviors?

We consider profanity words usage as a potential indicator of cyberbullying. Therefore we are mostly interested in the negative and positive words usage and the difference across Ask.fm and Instagram among common users. For this purpose we are using a dictionary of 1500 negative words (obtained from [110]) and 1500 positive words (obtained from [2]). Then we calculate the percentage of posts by profile owners and friends/non-owners that include at least one of the words in our dictionaries and consider it as a measure of *negativity* and *positivity* of a user profile.

5.2.1 Instagram and Ask.fm Features

As we explain in the previous section, in Ask.fm, each user maintains his or her profile and other users can ask questions by posting a question on that profile. Ask.fm is a semi-anonymous network in which the identities of the profile owners are always known but by default the identities of the users asking questions are not revealed. However, users asking questions have the option to reveal their identities. In the Ask.fm data that we analyzed, we noticed that on average, about 27% of the posts in users' profiles are non-anonymous. Profile owners may answer some of the questions posted on their profile. Their web pages show only the questions they choose to answer along with their answers.

While Ask.fm is a web and mobile-based social network that focuses primarily on textual content, Instagram is primarily a mobile social network whose main feature is allowing users to

take pictures using their smartphones. These pictures can then be digitally filtered and posted as media on their accounts, and shared through other social networking sites such as Facebook, Flickr, Twitter, and so on. Users can follow other users after their follow requests have been accepted. A user following another user can see the media posts of that user. The default access privilege in Instagram is public, which means everyone can see the media objects posted by a user. However, users have the option to change the access privilege to private, so that only the people who are following them can see their posts. In the Instagram data that we analyzed, we noticed that 39% of users chose to keep their access privilege private. Users in Instagram can like a media post, comment on a media post, and can also tag other users while commenting. Unlike Ask.fm, Instagram is a completely non-anonymous network since the names and profiles of users who like, follow, or comment are visible to everyone.

5.2.2 Data Collection

To compare user posting behaviors across Instagram and Ask.fm, we collected a comprehensive set of user data from each network. For Instagram users, starting from a seed node, 41K user ids were gathered with a snowball sampling method. Among these Instagram ids, there were both public and private profiles, of which only 61% of them have public profiles, or about 25K public profiles. These 25K public user profiles are used as our **normal Instagram users** data. For each public Instagram user, the collected profile data includes the media objects, the comments, the user id of each person followed by the user, the user ids of those who follow the user, and similarly for comments and likes of all shared media objects. For Ask.fm, starting from a seed node and again using snowball sample, 24K complete Ask.fm user profiles were collected and used as the **normal Ask.fm users** data. For each Ask.fm user, we collected the profile information, complete list of questions, their answers and authors (if not anonymized) and user ids of people who have liked each question+answer pair.

After collecting Ask.fm user information we observed some users have revealed their Instagram ids in their profile information. Selecting those users that have both their Ask.fm and

Instagram ids enables us to better compare two networks than comparing aggregates/averages of normal but non-common users. However, the incidence of such common users is relatively low, so we needed to collect a large number of Ask.fm profiles to find a sufficient number of common profiles. For this purpose, starting from a seed node, we collected 1M users' information via snowball sampling. Only 4% of the users have mentioned their Instagram id in their profile information, thereby furnishing around 40K users with ids for both social networks. Only about 24K of these profiles were public. From this, we collected complete profile content for an 8K subset of these common users from each of the two social networks, forming our **common Instagram users** data and **common Ask.fm users** data.

5.3 Analysis of Instagram User Behavior

In this section we examine the distributions of different types of activities of normal and common Instagram users. Figure 5.1 shows the complementary cumulative distribution function (CCDF) of the number of users a user follows (*follows*) and the number of users a user is followed by (*followed by*) in the Instagram social network. These CCDFs are shown separately for normal Instagram users and common Instagram users. Note that only those Instagram users who make their profiles public are included in this study. It can be seen that for both common and normal users, there is a large gap between the number of followed bys and follows. The main reason is that Instagram imposes 7,500 as the maximum number of users a user can follow. This was done to limit spam. There is no such limit on the number of users a user can be followed by. Also, we note that the follows behavior is similar for both common and normal users, except that there is a long tail in "normal follows". This is due to the fact that this limit was imposed in January 2013, and the users who were following more than 7,500 users before then continue to do so. Our examination suggests that most of these accounts in this long tail belong to campaigns and advertisements as well as aspiring photographers, models, businesses and people trying to attract attention on the popular page. Comparing the followed by behavior, we observe that there is a large gap between common and normal users' curves, but the two curves eventually converge, and both have long

tails. We believe that these long tails (up to 10^7) are due to the presence of celebrities who tend to be followed by a significantly larger number of users. We also have found a small number of celebrities among our common users, which lead to the convergence of the two curves. In terms of the large gap in followed bys between normal and common Instagram users, it probably comes from the nature of the Instagram social network which is a photo sharing website. Users followed by more than $10^3 - 10^5$ people are less likely to appear in the set of common users. Looking at normal users' profiles, we see users ranging from small music bands, local celebrities or fashions, popular nonprofessional photographers. These users seem to be interested in the photo sharing nature of Instagram and tend not to have profiles in Ask.fm. At the tail we have celebrities, campaigns and advertisements which open accounts at all social networks and that is why the two CCDFs converge.

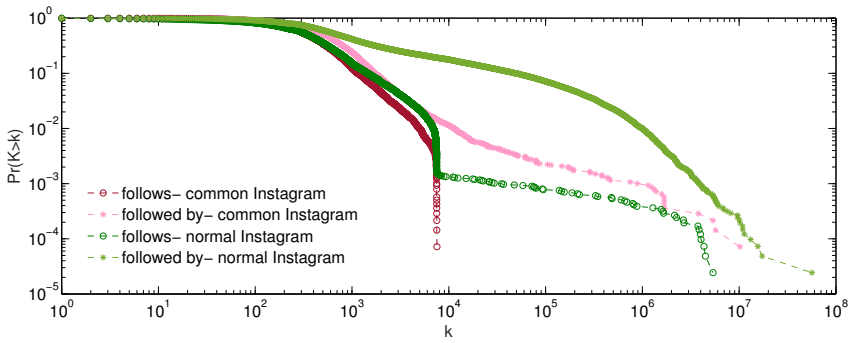

Figure 5.1: Complementary cumulative distribution functions (CCDFs) of **follows** and **followed by** of normal and common Instagram users.

Next, we analyze the following Instagram user activities: the number of media objects shared by a user, the number of likes received on the user's media objects, and the number of comments (by others and the owner) on a user's media objects. Figure 5.2 shows the CCDFs of these three user activities for normal and common Instagram users. We notice that for all these three user activities, normal users are generally more active than common users. We again observe big gaps between common and normal likes and comments. As will be shown later in Section 5.5, there exist fairly high correlations among followed by, likes, and comments. We can see that their CCDFs have

the same pattern.

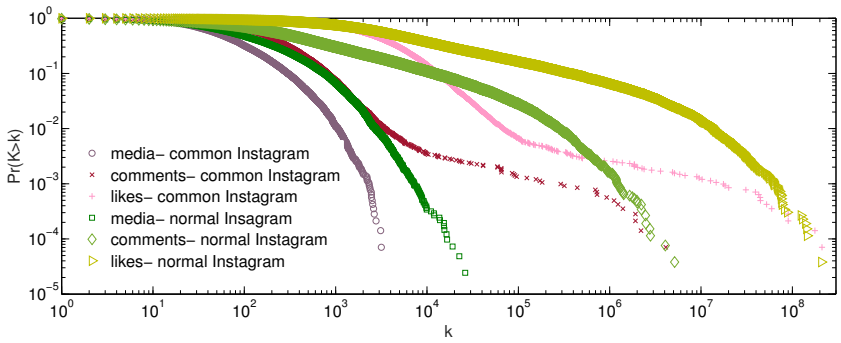

Figure 5.2: CCDFs of the number of shared pictures, number of likes received, and number of comments, of normal and common Instagram users.

Finally, we analyze the user behavior in terms of negativity and positivity in Instagram. To gain deeper insights, we divide all user posts into two groups, posts by profile owners and posts by other non-owner users, whom we will loosely term "friends" in the remainder of this chapter though some of these other users may not be truly friendly. Figure 5.3 shows the CCDFs of negativity and positivity of comments posted by profile owners and their friends, and again separated for normal and common Instagram users. In this figure, dashed lines are the CCDFs for common users and solid lines are for normal users. The first observation we make here is that positivity is always greater than negativity for both profile owners and their friends as well as for normal and common users. **The second important observation we make is that the user behavior is quite similar for normal and common users both for negative and positive user behaviors.**

5.4 Analysis of Ask.fm User Behavior

We first analyze user activity in Ask.fm, which consists of the total number of comments (answer+question pairs) and the total number of likes a user receives. For the rest of the chapter we use the term comments for answer+question pairs which includes both comments by profile owners and comments by other non-owner users (loosely speaking, friends) for Ask.fm too. Note that these two user activities are the only ones that we can gather from the publicly accessible data

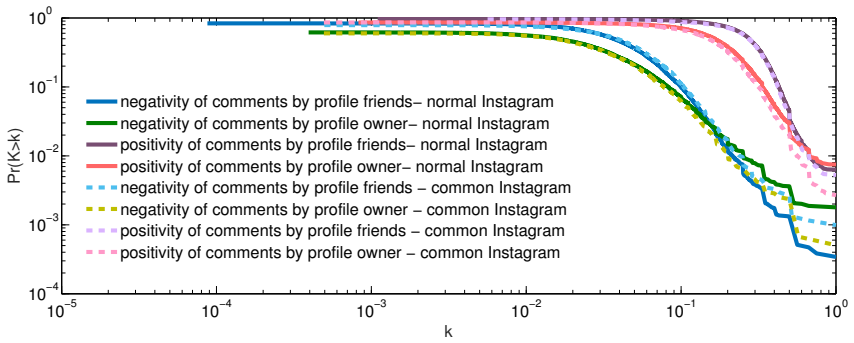

Figure 5.3: CCDFs of the percentage of negative and positive posts by normal and common public Instagram users. Red and purple show positivity and blue and green show negativity.

of Ask.fm. There are some additional activities such as posting a question that is not answered by the profile owner or user following activities that are not publicly available. Figure 5.4 shows the CCDFs of the two publicly available user activities for normal and common users. We observe that common users are slightly more active in terms of received likes than normal users, but that otherwise the activity in terms of number of question+answer pairs is similar. Therefore these common users exhibit opposite behavior in Ask.fm compared to Instagram relative to their normal CCDFs. It shows users who have two social network accounts (Ask.fm and Instagram) are more active/popular than the normal Ask.fm users and less active/popular than normal Instagram users.

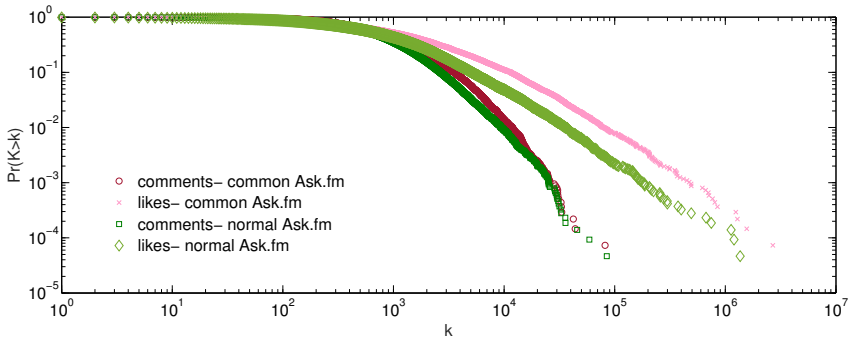

Figure 5.4: CCDFs for the number of likes and comments on Ask.fm

Finally, we analyze the user behavior in terms of negativity and positivity in Ask.fm. As before, we divide all user posts into two groups, those from the profile owner (answers) and those

from friends (questions). Figure 5.5 shows the CCDFs of negativity and positivity of comments for normal and common users. In this figure, dashed lines are the CCDFs for normal users and solid lines are for common users. As for Instagram, we observe on Ask.fm that positivity is always greater than negativity for both questions and answers for normal and common users. **As for Instagram, the second important observation we make is that the user behavior is quite similar for normal and common users on Ask.fm both for negative and positive user behaviors.**

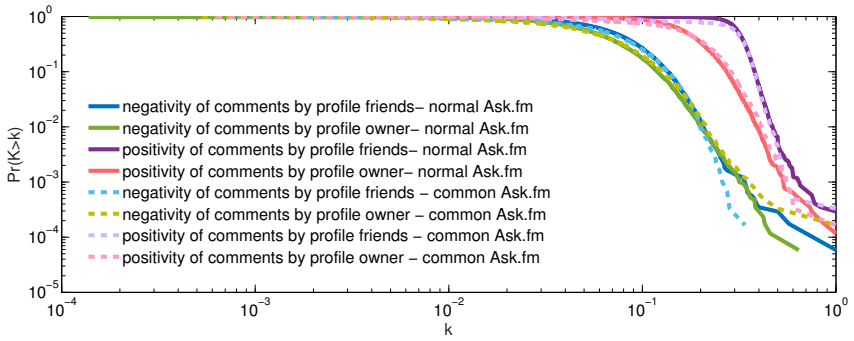

Figure 5.5: CCDFs of the percentage of negative and positive posts by normal and common Ask.fm users. Red and purple show positivity and blue and green show negativity.

5.5 User Behavior Across Instagram and Ask.fm

We now examine the impact of network features on user behavior. To do so, we have identified a set of 8K users who are members of both Ask.fm and Instagram social networks (common users). We will analyze the activities of these users from their postings in the two networks. The key question we seek to answer is "How does their posting behavior differ between the two social networks?"

We use LIWC, which is a text analysis program that counts words and place messages in psychologically meaningful categories [81]. Four different sets of comments are considered for each user; comments by friends and comments by profile owners in Instagram and Ask.fm. We calculate the LIWC values for a subset of categories for these four types of comments for each common

user. In Figure 5.6, we show the correlation for various categories between the two different social networks for the postings by the profile owners and the comments they receive from their friends. Some of these categories have been chosen from [27] that introduces some statistics about categories which teens talk about in their social networks. In this figure, the correlation for the comments posted by a profile owner across two networks is shown with pink bars, and the correlation for the received comments from friends by a profile owner across the two networks is shown in blue bars.

We first note that there are certain categories such as social, cogmachin and time for which the correlation is quite high for both posts by profile owners and by friends. This indicates that the differences in the two social networks have no impact in the discussion of these categories. On the other hand, there are certain categories such as anxiety, inhibition, health, ingestion, work and money for which the correlation is quite low for both posts by profile owners and by friends. This indicates that the differences in the two social networks have a significant impact in the discussion of these categories.

Next we look at the differences in correlations between the comments by profile owners and the comments by friends. For categories like anxiety and anger which are personal characteristics, the correlation between profile owner's comments across networks is higher than correlation of friend's comments. For sad, sexual and death categories the correlation is higher for friend's comments. That is, if an owner has sexual comments directed at them on one network, then there is a likelihood that they will also have sexual comments directed at them on the second network.

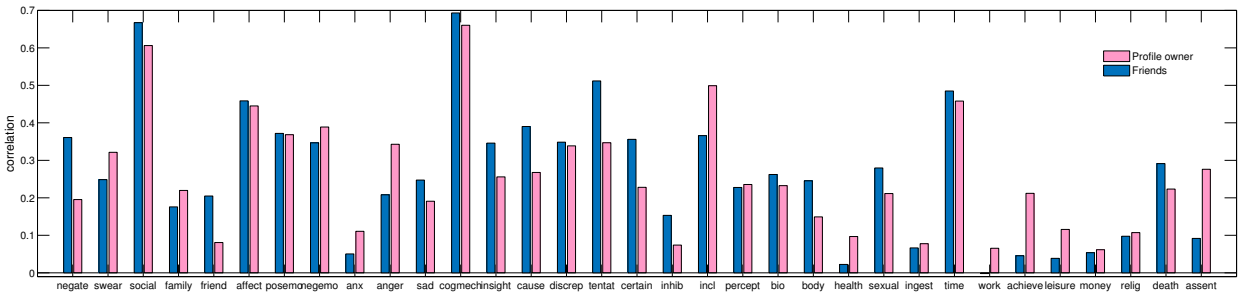

Figure 5.6: Correlations of profile comments between two social networks for different liwc categories, separated by owner and friend comments ($p < 10^{-5}$).

Table 5.1 provides a summary of the statistics of the average number of comments posted by owners or friends in Ask.fm and Instagram among normal and common users. The average number of comments posted by friends is larger than the comments by profile owner by a factor 3.4 for common public Instagram users, and surprisingly by 16.3793 for public normal users. Note that the ratio of owner to friend posts is 1 to 1 as expected for Ask.fm. We further observe that while common users have greater activity than normal users on Ask.fm, the opposite is true on Instagram. This agrees with our earlier analysis of the CCDFs of Figures 2 and 4.

	Normal Instagram	Common Instagram	Normal Ask.fm	Common Ask.fm
Avg # posts by profile owner	195.90	116.06	1216.5	1496.50
Avg # posts by friends	3208.70	394.60	1216.5	1496.50
Avg % negative posts by profile owner	3.72	3.39	6.54	6.47
Avg % negative posts by friends	4.62	4.78	7.79	7.21
Avg % positive posts by profile owner	17.13	16.69	17.97	16.81
Avg % positive posts by friends	27.59	26.4	29.18	27.53

Table 5.1: Comparison of (positive and negative) posts by profile owners and friends across four different user groups.

To understand in more detail whether common users are more or less positive or negative on Instagram vs Ask.fm, we calculated the percentage of positive posts and negative posts for each user profile. We further divided the posts into those written by owners and friends. Figures 5.7 and 5.8 show the CCDFs of positivity and negativity respectively for the common users. **Figure 5.7 shows that the positive behavior of common users across both networks is fairly similar for both owners and friends.** This is confirmed also by the similarity of average percentages of positivity shown in Table 5.1. The table also shows that positivity of friends' posts is consistently higher than owners' posts across all four categories in terms of average percentage of positive posts (by t-test with $p < 0.0001$).

Figure 5.8 compares the negativity of common users across the two networks. **We find among common users that most profile owners are more negative on Ask.fm than**

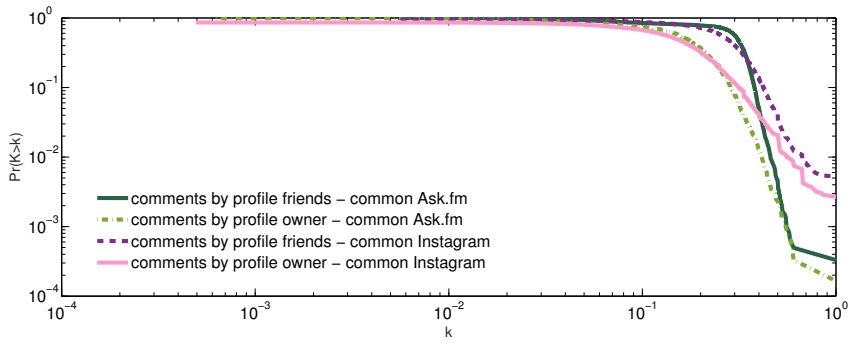

Figure 5.7: CCDFs of the percentage of positive posts by common Instagram and Ask.fm users. Pink and purple show common Instagram users and light and dark green show common Ask.fm users.

Instagram. We also find that most friends are slightly more negative on Ask.fm than Instagram. This is confirmed from Table 5.1, where we observe that the averages of the percentage of negative posts is higher on Ask.fm than Instagram, for both common and normal users (by t-test with $p < 0.0001$). Also, we note from the table that the negative percentages are clearly lower than the positive percentages across all 8 categories. This confirms our earlier analysis for the CCDFs from Figures 3 and 5.

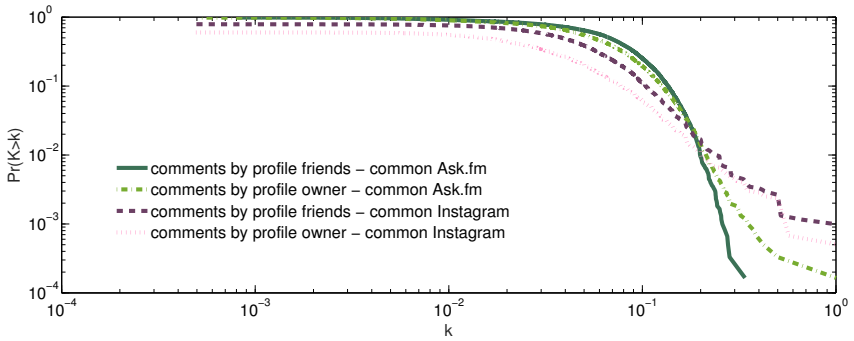

Figure 5.8: CCDFs of the percentage of negative posts by common Instagram and Ask.fm users. Pink and purple show common Instagram users and light and dark green show common Ask.fm users.

In Table 5.2, we show the negative words with the highest frequency of usage in Ask.fm and Instagram. The first row records the negative words with the highest frequency in Ask.fm and the third row records the negative words with the highest frequency in Instagram. The second

and fourth rows show the words with highest eignenvector centrality in Ask.fm and Instagram respectively, that is negative words that are used together. We explained in previous chapter how words with higher eigenvector centrality have been found and how they create a stronger effect of cyberbullying. In order to build the "Word" network, two words are connected if they have been seen together in a profile. After constructing the network, we find the words with high eigenvector centrality which brings up the words with high degree that have been seen together in the same profile. It can be seen that there are some common words among these top sixteen profanity words. However, these words have mostly different orders (highest value is on the left, decreasing in rank to the right). There are some words that can be seen as a potential sign of appearance-based, sexual or LGB cyberbullying.

hoe	n**ga	stupid	slut	kill	p**sy	cut	suck	gay	fat	sex	bad	d**k	die
f**k	ass	bitch	s**t	hate	d**k	ugly	sex	bad	suck	p**sy	fat	gay	stupid
f**k	hate	shit	bad	die	bitch	kill	ass	n**ga	mad	ugly	hell	black	gay
f**k	bitch	shit	hate	ass	kill	n**ga	bad	mad	die	ugly	gay	hell	faggot

Table 5.2: The first row shows the negative words with the highest frequency and the second row shows the negative words with the highest eigenvector centrality in Ask.fm. The third row shows the negative words with the highest frequency in Instagram and the fourth row shows the negative words with the highest eigenvector centrality in Instagram.

Finally, we analyze the correlation among the positivity and negativity of profiles across Ask.fm and Instagram and different social network activities or popularity factors (See Figure 5.9). In the axes legends in this figure, the first (left side) eight x-axis markers or the top eight y-axis markers represent positivity and negativity in various users posts. The right seven markers on x-axis or the lower seven markers on the y-axis represent various user activities. Correlation values only range from -0.2 to 1.0 because no lower values of negative correlation were found.

First, we notice that there is no correlation in the upper right side and lower left side in this figure. This means that there is no correlation between the negativity or positivity percentage of profiles with any user activities such as total comments, likes, followed by or follows in any of the two social networks. Next, we notice that in the lower right side, there is significant correlation with correlation values of 0.89 between Instagram likes and comments (IL and IC), while the correlation

between Ask.fm likes and comments(AL and AC) is lower at 0.46. We also find that the correlation between the number of likes and posted media on Instagram (IL and IM) is only around 0.1. It seems that the users who receive more likes also have more comments (sent+received), however the number of posted media does not seem to be a factor for becoming more popular in terms of receiving more comments or likes or being followed more. Also there is significant correlation, correlation value 0.98, between the number of likes and followed by (IL and IFb). Similarly, there is a high correlation between the number of comments and followed by (IC and IFb). This is not surprising as when you have more fans to follow you, you receive more likes.

Looking at the upper left side, the highest correlation, correlation value 0.7, is between positivity of Ask.fm owner's answers and friend's questions (AOP and AFP). Correlation between negativity of Ask.fm owner's answers and friend's questions is approximately the same with value 0.64 (AON and AFN). **Considering the nature of the Ask.fm social network where there is a specific answer for each question, this means positive questions have been mostly replied with positive answers and the negative questions are replied with negative answers.**

Also, the correlation between answers' positivity and questions' negativity (AOP and AFN) in Ask.fm is around 0.45. Based on our observation of some user profiles at Ask.fm, we noted that a lot of negative questions actually come from user's friends. Some come from upstanders when a cyberbullying case happens for the profile owner and some are just friendly talk between close friends (See Figure 5.10).

On the other hand, we can not see such a correlation between comments by profile owners and comments by friends in Instagram. The reason for this is the fact that in Instagram, one answer can be used in reply to several received comments, e.g. "Thank you all" is just one answer to several greeting messages received. The correlation between Instagram received and sent comments is just around 0.3 for both negativity and positivity, which indicates that there is less correlation between Instagram comments.

Another interesting observation is the correlation of 0.4 between positive behavior of profile

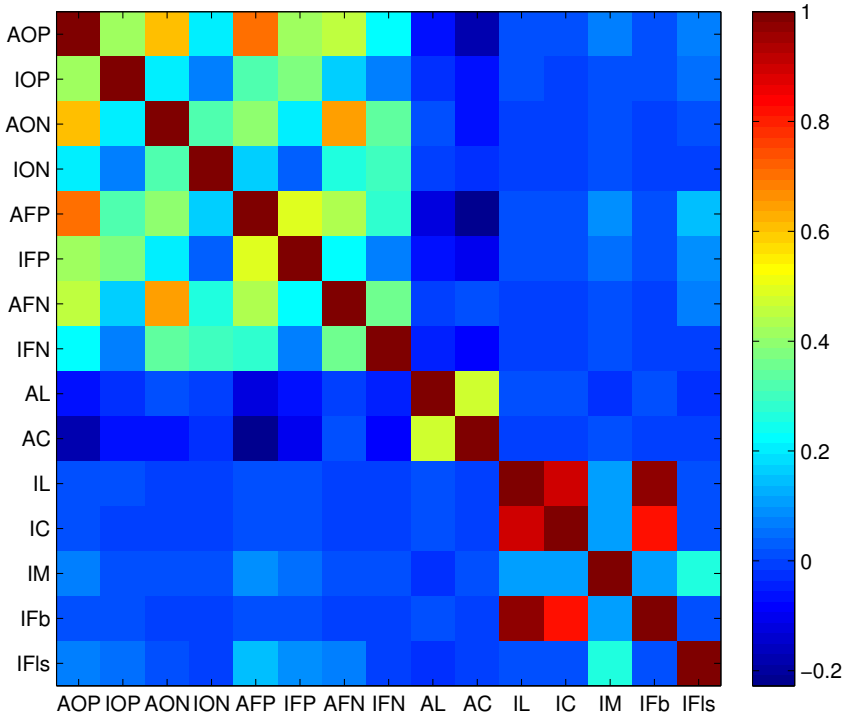

Figure 5.9: Correlation matrix, where the characters stand for A: Ask.fm, I: Instagram, P: Positivity, N: Negativity, O: Owner's comment, F: Friend's comments, L: Likes, C: Comments, M: Media, Fb: Followed by, Fls: Follows ($p < 10^{-5}$).

owners across their Ask.fm and Instagram (AOP and IOP). The correlation of negative behavior of profile owners across the two networks (AON and ION) is a approximatley similar value. However, positive behavior of a profile owner in one network is not correlated with the negative behavior of the profile owner in the other network (AOP and ION or AON and IOP). **These results provide some support for the notion that a profile owner who is positive on one network generally tends to be positive on the other network, and similarly, a profile owner who is negative on one network generally tends to be negative on the other network.**

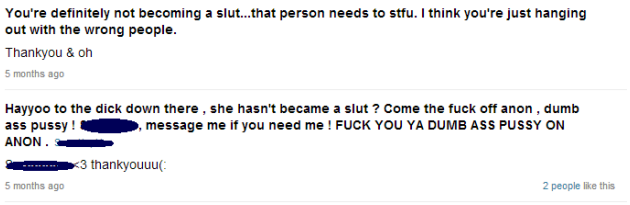

Figure 5.10: An example of a profile in which negative words from a bystander which is not a cyberbullying question against profile owner.

5.6 Impact of Anonymity on User Behavior in Ask.fm

One important feature in a social network that can potentially lead to negative user behavior is anonymity. Intuitively, it seems that if users can post messages anonymously, they will tend to be more negative. Indeed, anecdotal evidence suggests this. For example, according to Natasha MacBryde's father, anonymity had a big role in his daughter's suicide [14]. To understand the impact of anonymity on user behavior, we analyzed the data we collected from Ask.fm. Recall that Ask.fm is a semi-anonymous network in which users asking questions are anonymous by default and some users choose to reveal themselves while asking questions. So, we analyzed negativity in questions that were posted anonymously and compared that with negativity in questions in which users chose to reveal themselves.

In particular, we collected the profile contents of about 15K Ask.fm users (10.6K normal and 4.4K common users) Figure 5.11 shows the CCDFs of the percentage of positive and negative questions for Ask.fm users. **We notice that the the non-anonymous comments tend to be more negative and more positive than the anonymous comments.** This is a surprising result, since we expected that non-anonymous users will be less negative. To gain a better insight into these behaviors, we delved into some of the users' profile contents and found that the reasons of non-anonymous comments having more negativity is twofold. First, some non-anonymous comments are from users who try to defend a profile owner from abusive comments that were posted anonymously on his/her profile. In turn, these supporting comments from non-anonymous users end up containing some negative words. For example, a non-anonymous comment in one of the

user's profile whose profile content has several negative anonymous comments says *"Hayyoo to the d**k down there , she hasn't became a s**t ? Come the f**k off anon , dumb a** p**sy ! Schuyler , message me if you need me ! F**K YOU YA DUMB A** P*SSY ON ANON"* as it is shown in Figure 5.10. In this comment, the non-anonymous user attacks anon (anonymous user) for abusing the profile owner. Another example of such a post goes like this, *"All you people calling her a s**t, GET OVER YOUR SELF! Your probably the ones who are the s**ts and are fake! Just leave her alone b*tchez! (*****, they are all just jealous of you!)"*.

The second reason for increased negativity in questions posted by non-anonymous users is that some non-anonymous comments are from close acquaintances who use negative words as a mean for showing affection. For example, a non-anonymous comment goes like this/ *"you're my b**ch and I'm always here for you! And you always have to have those stupid memories that do not need to come up but I still love you! See you tomorrow bi***hh!(: love youuuu"* (See Figure 5.13). On the other hand, there exist negative anonymous comments that tend to be actually abusive. For example, *"Wh**e! S**t! Fag! Ugly! A*s! U look like my d**k!"* or *"yhu trynna s**k for a buck b**ch"* as shown in Figure 5.12. So, it is evident from the above discussion and examples that negativity in a comment is not always an indicator of cyberbullying, and that more research, such as labeling comments as cyberbullying or not, will be needed to firmly establish the relationship between anonymity and cyberbullying.

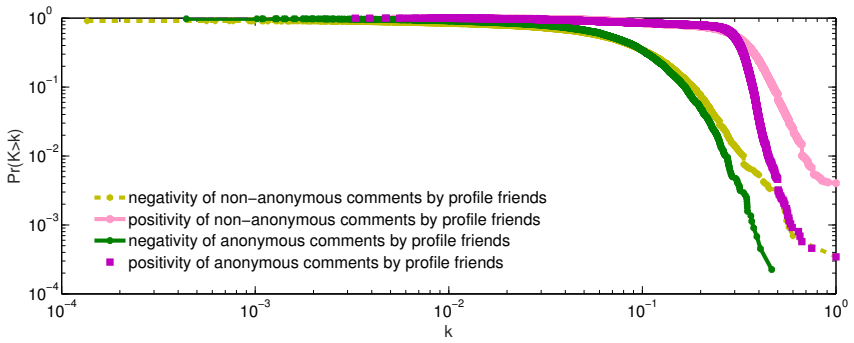

Figure 5.11: CCDFs of the percentage of positive and negative posts for Ask.fm users. Pink and purple show positiviyu and light and dark green show negativity.

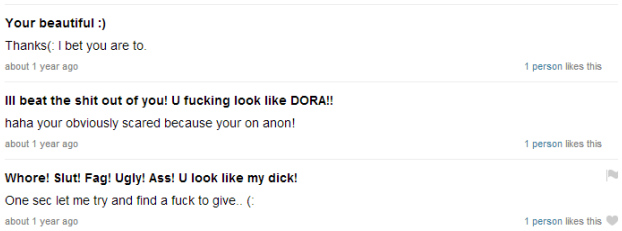

Figure 5.12: An example of a profile having posts with negative words from an anonymous user which is cyberbullying question against profile owner

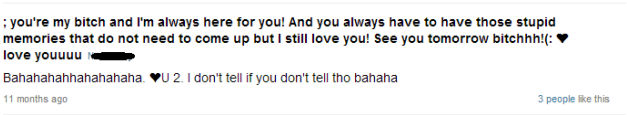

Figure 5.13: An example of a profile having posts with affectionate negative words from a non-anonymous user

5.7 Summary

We have studied user behavior in two social networks, Instagram and Ask.fm, that are among the most popular used for cyberbullying. We have analyzed negative and positive words usage from comments by users common to both the Instagram mobile social network and the Ask.fm online social network. The following comprise our key findings. First, we found that both owners and "friends" who post on common users' profiles have increased negativity on Ask.fm compared to Instagram. Second, there is no difference in the positivity of both owners and friends on each network. Third, common users are similar in their positivity or negativity when compared to normal non-common users on each of the two networks. Fourth, anonymity actually results in less negativity on Ask.fm compared to non-anonymous comments. We have described where more research is needed to label which negative comments constitute cyberaggression. Finally, we found that there is a strong correlation on Ask.fm between owners and friends' negative comments, and similarly for positive comments. There is also some support for the notion that a profile owner who is positive on one network generally tends to be positive on the other network, and similarly for negative behavior.

We found that the users are more negative in Ask.fm than Instagram. It is tempting to attribute the increased negativity on Ask.fm to the anonymity of Ask.fm. However, there are other differences between the two networks that could influence the increased negativity on Ask.fm. For example, Ask.fm users primarily post in text, whereas Instagram is highly focused on posting image media, followed by the commenting. Also, the relationship between Ask.fm owners comments and non-owner "friends" comments is one-to-one, whereas on Instagram, a single owner comment may be directed at multiple friends' comments, such as a "Thank you all". This requires more research into the relationship between anonymity and negativity.

While we have conducted a negative (and positive) word analysis on posting, we would like to extend this research. Not all occurrences of negative words correspond to cyberbullying, e.g. "that's f**king amazing". Towards this end, we would need to specifically label comments as being examples of cyberaggression.

Acknowledgments

Thank to PhD students Rahat Ibn Rafigh and Shaosong Li for their collaboration in this work. I would also like to thank all the cyberbullying group members.

Chapter 6

Labeling and Characterization of Cyberbullying and Cyberaggression in Social Networks

6.1 Overview

Given the gravity of the consequences cyberbullying has on its victims and its rapid spread among middle and high school students, there is an immediate and pressing need for research to understand how cyberbullying occurs in OSNs today, so that effective techniques can be developed to accurately detect cyberbullying. In [104], it is reported that experts in the field of cyberbullying could favor automatic monitoring of cyberbullying on social networking sites and propose effective follow-up strategies.

This thesis makes the important distinction between cyberaggression and cyberbullying. Cyberaggression is broadly defined as any occurrence of using digital media to intentionally harm another person [51]. Examples include negative content and words such as profanity, slang and abbreviations that would be used in negative posts such as hate, fight, wtf. Cyberbullying is one form of cyberaggression that is more restrictively defined as intentional aggression that is **repeatedly carried out** in an electronic context against a person **who cannot easily defend him or herself**[51, 79, 50]. Thus, cyberbullying consists of three main features : (1) an act of aggression online; (2) an imbalance of power between the individuals involved; and (3) it is repeated over time [42, 51, 74, 75, 92]. Similar to traditional bullying, it is the combination of the aggressive behavior, repeated acts, and the victim's inability to defend him or herself that severely impacts many teens [50]. Particularly important in the context of cyberbullying, is the permanent nature of the

online posts (until it's removed), the ease and wide distribution in which aggressive posts can be made, the difficulty of identifying the behavior, the ability to be connected and exposed to online interaction 24/7, and the growing number of potential victims and perpetrators [35]. Repetition of cyberbullying can occur over time or by forwarding/sharing a negative comment or photo with multiple individuals [58]. The power imbalance can take on a variety of forms including physical, social, relational or psychological [24, 63, 75, 87], such as a user being more technologically savvy than another [50], a group of users targeting one user, or a popular user targeting a less popular one [58].

Facebook, Twitter, YouTube, Ask.fm, and Instagram have been listed as the top five networks with the highest percentage of users reporting experience of cyberbullying [100]. Instagram is of particular interest as it is a media-based social network, which allows users to post and comment on images. An example of an Instagram media session is shown in Figure 6.1. Cyberbullying in Instagram can happen in different ways, including posting a humiliating image of someone else by perhaps editing the image, posting mean or hateful comments, aggressive captions or hashtags, or creating fake profiles pretending to be someone else [33].

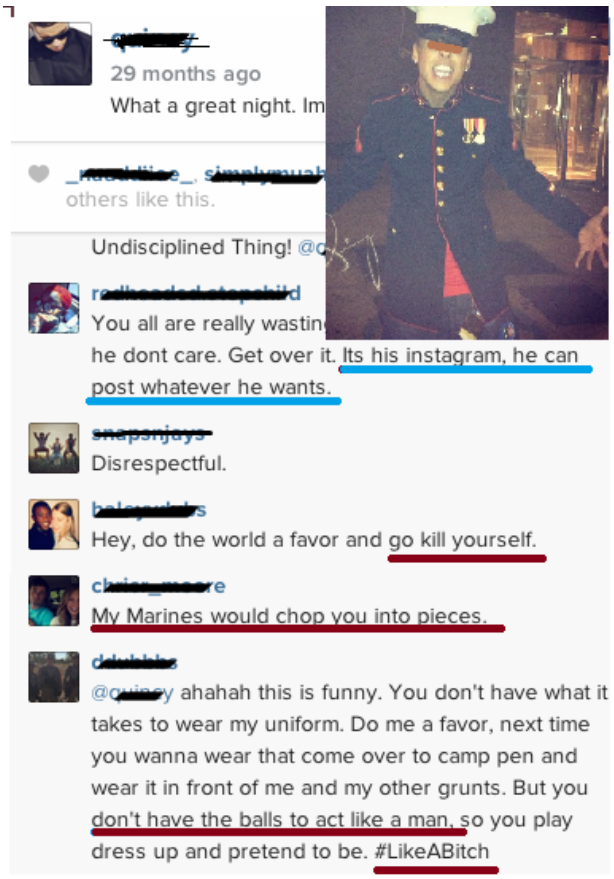

Figure 6.1: An example of comments posted on Instagram.

The main goal of this chapter is to study cyberbullying in Instagram. To do so, we first collected a large sample of Instagram data comprised of 3,165K media sessions (images and their associated comments) taken from 25K user profiles. Next, we provided labeling instructions and Instagram media sessions to human labelers at

the crowd-sourced CrowdFlower website to identify occurrences of cyberbullying and cyberaggression in Instagram. Later in section 8 we apply the same method of labeling to Vine collected data. Then we analyzed the labeled Instagram data set, reporting the relationship of different features of these media sessions to cyberbullying.

As explained in chapter 5, 25K public user profiles were collected. For each public Instagram user, the collected profile data includes the media objects (videos/images) that the user has posted and their associated comments, user id of each user followed by this user, user id of each user who follows this user, and user id of each user who commented on or liked the media objects shared by the user. We consider each media object plus its associated comments as a **media session**. Labeling data is a costly process and therefore in order to make the labeling of cyberbullying more manageable, we sought to label a smaller subset of these media sessions. To have a higher rate of cyberbullying instances, we considered media sessions with at least one profanity word in their associated comments. We tag a comment as "negative" using an approach similar to chapter 4. For this set of 25K users, 3,165K unique media sessions were collected, where 697K of these sessions have at least one profane word in their comments by users other than the profile owner, where a profane word is obtained from a dictionary [1], [110].

In addition, we needed media sessions with enough comments so that contributors could adequately assess the frequency or repetition of aggression, which is an important part of the cyberbullying definition. We selected a threshold of 15 as a lower bound on the number of comments in a media session, considering that the average ratio of comments posted by users other than friends to comments posted by the profile owner in an Instagram profile is around 16 5. All the media session with more that 40% negativity were selected, which is a total of 922 sessions. Also For media sessions with non-zero and less than 40% negativity, 1296 media session we selected randomly for labeling. At the end 2,288 media sessions (images and their associated comments) were ready for the task of labeling.

6.2 Labeling

In this section, we explain the design and methodology for our survey for labeling the selected set of media sessions. Our first challenge is choosing appropriate definitions of terms, which will then be used in ground truth labeling. Based on the literature, a major early choice that we have made is to distinguish between cyberaggression and cyberbullying in the labeling.

In Instagram, each media session consists of a media object posted by the profile owner and the corresponding comments for the media object. The goal in this chapter is to investigate cyberaggression and cyberbullying in this *multi-modal context* (textual comments and media objects). Our labeling process consisted of two separate surveys. The first survey incorporates both the image and the associated text comments when asking the human contributors whether the media session was an instance of cyberbullying or cyberaggression. The second survey is focused on the image contents alone and is used to identify the content and category of the image.

Figure 6.2 illustrates an example of our design for the first labeling survey. On the left is the image, and on the right is a scrollable interface to help the contributor see all of the comments associated with this image. With input from a social science expert, we designed a simple survey that asked the contributors two questions, namely whether the media session constituted cyberaggression or not, and whether the media session constituted cyberbullying or not. During the instructional phase prior to labeling, contributors were given the aforementioned definitions of cyberaggression and cyberbullying along with related examples. Each media session was labeled by five contributors.

6.2.1 Quality Control

In order to provide quality control, we only permitted the highest-rated contributors on CrowdFlower to have access to our job. Next, a mentoring phase was provided for the potential contributors that included instructions and a set of example media sessions with the correct label. Further, to monitor the quality of the contributors and filter out the spammers, potential con-

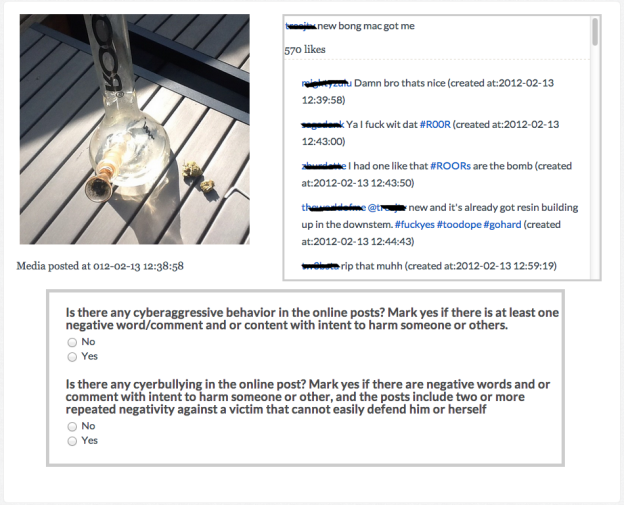

Figure 6.2: An example of the labeling survey, showing an image and its corresponding comments, and the survey questions.

tributors were asked to answer a set of test questions in two phases: quiz mode and work mode. Potential contributors needed to answer correctly a minimum number of test questions to pass the quiz mode and qualify as a contributor for the survey. We also incorporated quality control checks during the labeling process (work mode) by inserting random test questions. A contributor was filtered out if he/she failed this work mode. Finally, a minimum time threshold was set to filter out contributors who rushed too quickly through the labeling process. The minimum number of test questions to get back high-quality data was recommended by CrowdFlower.

Overall, 176 potential contributors worked on the quiz questions, 144 passed the quiz mode, while 31 contributors failed and 1 gave up. The labeled data that we finally obtained were from 139 *trusted* contributors, while the the rest were filtered out during the work mode. Table 6.1 provides the number of trusted judgments and the contributors' accuracy for 11,090 total judgments.

6.2.2 Image Labeling

Next, we conducted a second separate survey just to label the image contents of media sessions, so that we could investigate the relationship between cyberbullying and image content. Human contributors were given detailed instructions for identifying image's content. We first

Table 6.1: Labeling process statistics. Trusted judgments are the ones made by trusted contributors.

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

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

6.3 Analysis and Characterization of Ground Truth Instagram Data

We submitted our first study with 2,218 media sessions (images and their associated comments) to CrowdFlower, a crowd-sourced website. Each media session was labeled by five different contributors. CrowdFlower provides a degree of trust for each contributor based on the percentage of correctly answered test questions explain in section 6.2.1. This trust value is incorporated by CrowdFlower into a weighted version of the majority voting method called a "confidence level" for each labeled media session. We decided to keep media sessions whose weighted trust-based metric was equal to or greater than 60%. We deemed them to be strong enough support for majority voting from contributors with higher trust. Overall, 1,954 (88%) of the original pure majority-vote based media sessions wound up in this higher-confidence cyberbullying-labeled group. For this higher-confidence data set, 29% of the media sessions belonged to the "bullying" group while the

other 71% were deemed to be not bullying.

6.3.1 Labeling and Negativity Analysis

Distribution of the media sessions based on the number of votes (out of five votes) received for cyberaggression and cyberbullying respectively has been provided in Figure 6.3. The left chart shows the fraction of samples that have been labeled as cyberaggression k times, and the right chart shows the fraction of samples that have been labeled as cyberbullying k times. The higher the number of votes for a given media session, the more confidence we have that the media session contains an incident of cyberaggression or cyberbullying, with five votes means unanimous agreement. Similarly, the lower the number of votes for a given media session, the more confidence we have that the media session *does not* contain an incident of cyberaggression or cyberbullying, with zero votes means unanimous agreement.The inter-rater agreement Fleiss-Kappa value for cyberbullying is 0.5 and for cyberaggression is 0.52.

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

Next, we observe that about 30% of the media sessions have not been labeled as cyberaggression by any of the five contributors. Since all media sessions contained at least one comment with one or more profane word, this suggests that only employing a profanity usage threshold to detect

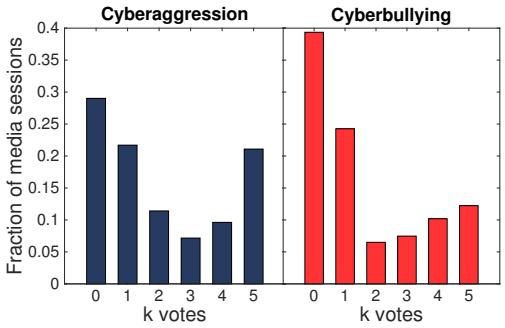

Figure 6.3: Fraction of media sessions that have been voted k times as cyberaggression (left) or cyberbullying (right).

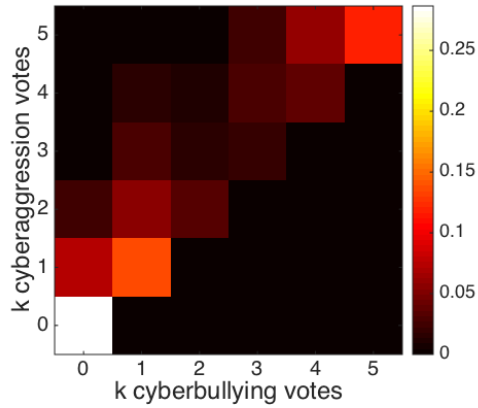

Figure 6.4: Two-dimensional distribution of media sessions as a function of the number of votes given for cyberaggression versus the number of votes given for cyberbullying, assuming five labelers.

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

In order to understand the relationship between cyberaggression and cyberbullying, we plotted in Figure 6.4 a two-dimensional heat map that shows the distribution of media sessions as a function of the number of votes each media session received for cyberaggression and cyberbullying. We observe that a significant fraction of the sessions exhibit strong agreement in terms of either receiving high numbers of votes for both cyberbullying and cyberaggressions, or receiving low numbers of votes for both cyberbullying and cyberaggression. This can be inferred from the high energy in the upper right and lower left part of the diagonal. In addition, it is promising that the area below the diagonal is essentially zero, meaning no session has received more votes

for cyberbullying than for cyberaggression. This conforms with the definition that cyberbullying is a subset of cyberaggression. The Pearson's correlation between number of votes for cyberbullying and number of votes for cyberaggression is 0.9.

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

Finally, we are interested in understanding the relation between cyberbullying/cyberaggression and the percentage of negativity in the comments. We divided all the media sessions into nine different bins based on the percentage of negativity in their comments. Bin $(n_1 - n_2]$ contains all media sessions with bigger than $n_1\%$ and smaller than or equal to $n_2\%$ negativity. None of the media sessions contained more than 90% negative comments. Next, we calculated percentage media sessions for each bin that can be identified as cyberaggression or cyberbullying based on majority of votes, i.e. where the number of votes is 3 or higher. Figure 6.5 shows these fractions, left figure for cyberaggression and right figure for cyberbullying. We observe that as the percentage of negativity increases, so does the fraction of media sessions up until 50% negativity for cyberaggression and 60% for cyberbullying. This increase is as expected, since cyberaggression or cyberbullying is typically accompanied with negativity in the postings. However, we notice that the percentage of cyberaggression or cyberbullying starts decreasing after these peaks as the percentage of negativity increases. This is quite an unexpected result and seems counter-intuitive. To understand this, we examined closely the media sessions that have very high negativity. We noticed that these media

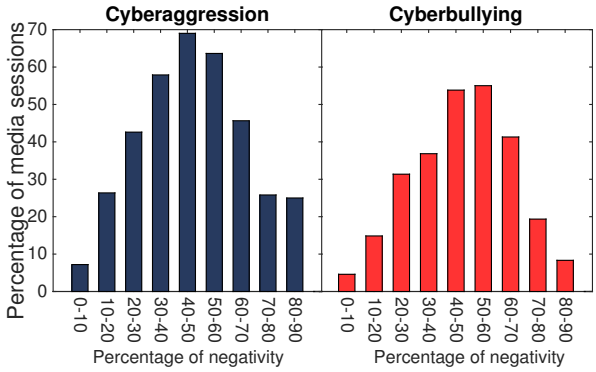

Figure 6.5: Percentage of media sessions that have been labeled as cyberaggression (left) and cyberbullying (right) versus their negativity percentages.

sessions typically involved discussions about sports, politics, tattoos, or were just friendly talks. People tend to use lots of profanity words in such discussions, even though they are not insulting any one person in particular. This leads us to our final important finding about negativity analysis. **A media session with a significantly high percentage of negativity (more than 60-70%) typically implies a low probability that the session contains a cyberbullying incident.**

6.3.2 Temporal and Graph Properties Analysis

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

Figure 6.6 illustrates the correlation between the number votes and the number of comments

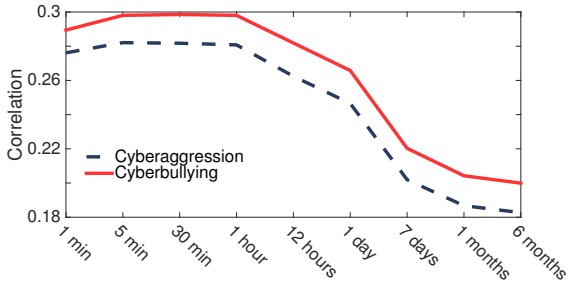

Figure 6.6: Pearson's correlation between the number of votes and the number of comments that arrive in $\leq x$ seconds after their previously received comment.

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

Table 6.2: Mean values of social graph properties for cyberbullying versus non-cyberbullying samples and aggression versus non-cyberaggression. ($*p < 0.05$).

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

Next, we examine the relationship between cyberbullying/cyberaggression and the social network graph features such as the number of likes for a given media object, number of comments posted for a media object, number of users a user is following, and the number of followers of a user. Table 6.2 shows these numbers, for categories of non-cyberbullying sessions, cyberbullying sessions,

non-cyberaggression sessions and cyberaggression sessions. We observe that media sessions that contain cyberbullying/cyberaggression share more media objects than media sessions that do not contain cyberbullying/cyberaggression, but on average receive lower number of likes. Souza **et al.**'s [95] analysis of Instagram users shows there is a positive correlation between number of followers and number of likes for typical Instagram users. Users who receive cyberbullying do not follow the same pattern. In fact, the average number of likes per post for non-cyberbullying sessions is 4 times the average number of likes for cyberbullying sessions, and the the average number of likes per post for non-cyberaggression sessions is 4.5 times the average number of likes for cyberaggression sessions. In terms of number of following and followers, the distinction is not as pronounced, although we see that the media sessions with cyberbullying/cyberaggression incidents have more followers and less following compared to the media sessions without cyberaggression/cyberbullying. *The key finding here is that the users of media sessions with cyberbullying/cyberaggression have lower number of likes per post while have more followers.*

6.3.3 Linguistic and Psychological Analysis

We now focus on the pattern of linguistic and psychological measurements of cyberbullying/cyberaggression media sessions versus non-cyberbullying/non-cyberaggression. For this purpose, we have applied Linguistic Inquiry and Word Count (LIWC), a text analysis program to find which categories of words have been used for cyberbullying/cyberaggression labeled media sessions. LIWC evaluates different aspects of word usages in psychologically meaningful categories, by counting the number of the words across the text for each category [81]. LIWC has often been used for studies on variations in language use across different people. Published papers show that LIWC have been validated to perform well in studies on variations in language use across different peoples [22]. We first analyze the number of words, and usage of pronouns, negations and swear words (Figure 6.7). Next, we look at some of the personal concerns such as work, achievements, leisure, etc. (Figure 6.7). Finally, we investigate some of the psychological measurements such as social, family, friends, etc. (Figure 6.8). For each of these cases, we first obtain the LIWC values

for each media session comment set. We then calculate the average LIWC value for each of the four classes: media sessions with cyberbullying, media sessions with no cyberbullying, media sessions with cyberaggression, and media sessions with no cyberaggression. The bars shown in Figures 6.7-6.8 represent the ratio of average LIWC value for cyberbullying class to that of non-cyberbullying, and the ratio of average LIWC value for cyberaggression class to that of non-cyberaggression.

In Figure 6.7, we first notice that the word count for media sessions with cyberbullying/cyberaggression is significantly higher than for media sessions with no cyberbullying/cyberaggression ($p < 10^{-5}$). Next, as expected, for swear words (e.g., damn, piss) and negations (e.g., never, not), the ratio is higher for cyberbullying/cyberaggression category ($p < 10^{-5}$). It is interesting to note that the ratios for the third person pronouns (she, he, they) are more than 1.3 ($p < 10^{-5}$), ratio for the first person singular pronoun (i) is 0.85, and the ratios for first person plural and second person pronouns (we, you) is closer to 1. This leads us to our first key finding with respect to the linguistic features. *A user is less likely to directly refer to himself/herself and more likely to refer to other people in third person in postings involving cyberbullying or cyberaggression.*

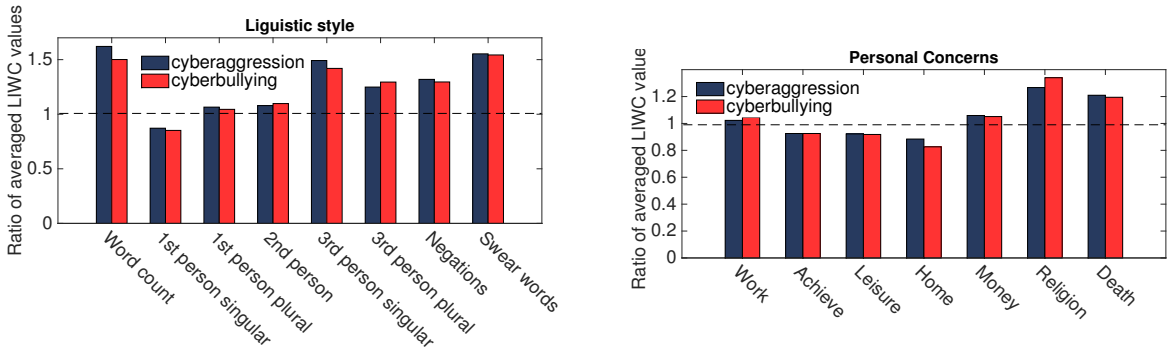

Figure 6.7: (left) Ratio of LIWC values of cyberbullying/cyberaggression labeled media sessions to non-cyberbullying/non-cyber aggression class in Linguistic categories. (right)Ratio of LIWC values of cyberbullying/cyberaggression labeled media sessions to non-cyberbullying/non-cyberaggression class in Personal Concerns categories.

For personal concerns (Figure 6.7), "religion" (e.g., church, mosque) and "death" (e.g., bury, kill) categories have higher ratios (more than 1.2, p $<$ 0.1). This is in line with our findings in our previous work on profanity usage analysis in ask.fm social media, where we observed that

there is high profanity usage around words like "muslim" [37]. This suggests that religion-based cyberbullying may be quite prevalent in social media. On the other hand, ratios for personal categories like "work", "money" and 'achieve" are much closer to 1.

For psychological measurements (Figure 6.8), we notice that the ratios for "negative emotion", "anger", "body", and "sexual" categories are significantly higher than 1 (more than 1.4, $p < 10^{-5}$), and the ratio for "positive emotion" category is significantly lower than 1 (0.76, $p < 10^{-5}$). Higher ratios for "body" (e.g. face, wear) and "sexual" (e.g. slut, rapist) categories provide evidence for appearance-based and sexual-based cyberbullying in social media. For other psychological measurement categories, such as "social", "friend", etc., the ratios are closer to 1. Based on our observations from Figures 6.7 and 6.8, our final important finding with respect to linguistic features is as follows: *There is a higher probability of cyberbullying in postings involving religion, death, appearance and sexual hints, and cyberbullying posts typically have higher occurrences of negative emotions and lower occurrences of positive emotions.*

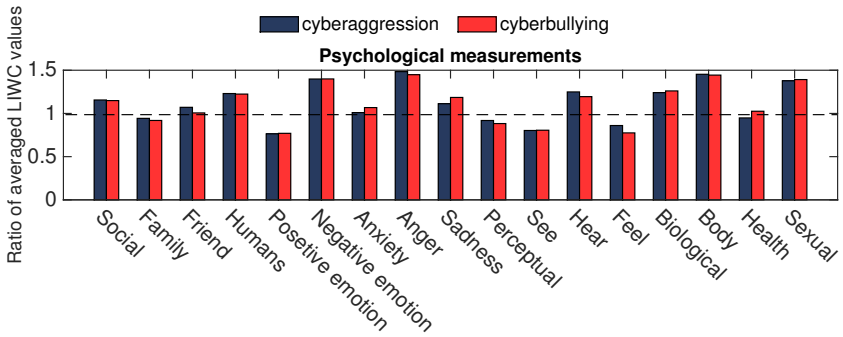

Figure 6.8: Ratio of LIWC values of cyberbullying/cyberaggression labeled media sessions to non-cyberbullying/non-cyberaggression class in Psychological categories.

6.3.4 Image Content Analysis

We now explore the relationship between image content and cyberbullying/cyberaggression in a media session. If the majority of labelers chose a given content category for an image, then that image was counted as belonging to that category. Note that it was possible for contributors

to place an image in more than one category. More than 70% of the images were labeled with only one category, and around 20% were labeled with two categories. However, there were a few images that were labeled with up to eight unique categories.

Figure 6.9 shows the fraction of the contents for all labeled data in the green bar. The "dont know" choice was given as we realized that for some images it is hard to figure out what is in the image. As some images belong to more than one category, the bars will sum up to more than one. First, we observe that the most common labels for image content are Person/People, Text and Sports.

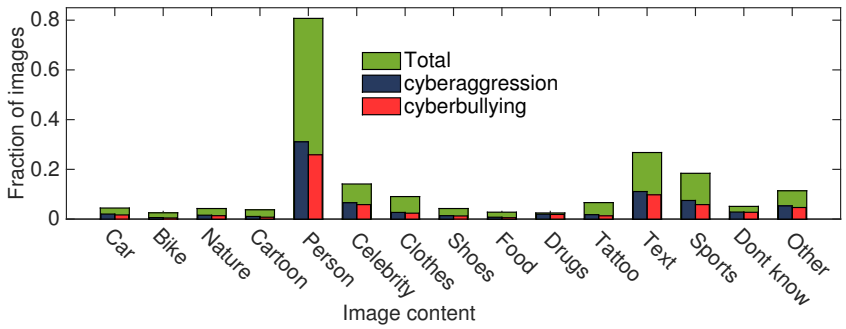

Figure 6.9: Fraction of image categories for all media sessions, cyberbullying and cyberaggression classes.

Next, the heights of the blue and red bars embedded inside each green bar relative to the height of the green bar indicate the fraction of images belonging to the media sessions that contained cyberaggression and cyberbullying respectively. For example, for the "Text" category, about 1/3 of the images with "Text" were associated with media sessions containing cyberaggression and cyberbullying. We observe that for some content categories such as "Drugs", the overall fraction is quite small (green bar height is low), but most of the images in those categories do belong to media sessions with cyberaggression/cyberbullying in them. To see this more clearly, Figure 6.10 plots the fraction of images labeled as cyberaggression/cyberbullying for each content category. We notice that for content category "Drugs", 75% of the images belong to media sessions containing cyberbullying, while for content categories like "Car", "Nature", "Person", "Celebrity", "Text" and

"Sport", 30%-40% of the images belong to media sessions containing cyberbullying/cyberaggression. Also, whenever images contain bike, food, tattoo, etc., there is little cyberbullying occurring. *The key finding here is that certain image contents such as Drug are strongly related with cyberbullying, while some other image contents such as bike, food, etc. have a very low relationship with cyberbullying.*

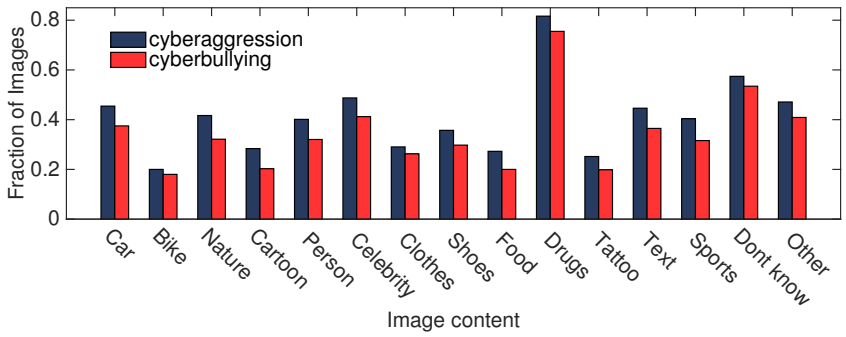

Figure 6.10: Fraction of images which have been labeled as cyberbullying and cyberaggression for each content category.

6.4 Summary

This chapter makes the following major contributions. First, an appropriate definition of cyberbullying that incorporates both frequency of negativity and imbalance of power is applied in large-scale labeling, and is differentiated from cyberaggression. Second, cyberbullying is studied in the context of a media-based social network, incorporating both images and comments in the labeling. We found that labelers are mostly in agreement about what constitutes cyberbullying and cyberaggression in Instagram media sessions. Third, a detailed analysis of the distribution results of labeling of cyberbullying incidents is presented, including a correlation analysis of cyberbullying with other factors derived from images, text comments, and social network meta data. We found a significant number of media sessions containing profanity and cyberaggression were not labeled as cyberbullying, suggesting that detection of cyberbullying must be more sophisticated than merely looking for profanity. We observed that media sessions with very high percentage of negativity

above 60-70% actually correspond to a lower likelihood of cyberbullying. Also, media sessions with cyberbullying exhibit more frequent commenting. We found that users of media sessions containing cyberbullying demonstrate a lower number of likes per post. Finally, cyberbullying has a higher probability of occurring when media sessions contain certain linguistic categories such as death, appearance, religion and sexuality content. Similarly, certain image contents such as "drug" are highly related to cyberbullying while other image categories such as "tattoo" or "food" are not.

In the future, we plan to release the labeled dataset, to support building of classifiers. More features should be examined for their correlation with cyberbullying, such as new image features, etc. Such features can be auto-generated by software rather than requiring human labeling. We also wish to obtain greater detail from the labeling process. Streamlining down to two labeling questions improved the response rate, quality and speed, but limited our ability to ask more detailed questions about other aspects of cyberbullying, such as different types and roles.

Chapter 7

Detection and Prediction of Cyberbullying Incidents

In this section we design and evaluate multimodal classifiers that extend beyond merely the text dimension for detecting and predicting cyberbullying and further incorporate image-based features and user properties. To the best of our knowledge, no other research has addressed the problem of predicting cyberbullying in social networks. In selecting data for labeling, we chose all 922 media sessions with more than 40% negativity (Set40+). For comment sets with less than 40% negativity, we randomly selected a set of around 1296 media sessions (Set0+). Therefore in evaluating the performance of the classifiers, we will look at the result of these two sets separately. For the purpose of prediction of cyberbullying incidents, we added new labeled media sessions, selected randomly from a set of media sessions with zero percent negativity.

7.1 Detection of Cyberbullying Incidents

Based on the labeled ground truth data, we next proceed to detect occurrences of cyberbullying in Instagram media sessions. A majority vote criterion was employed to determine whether a media session was cyberbullying/cyberaggression or not. To design and train the classifier, five-fold cross-validation was applied to the data such that 80% of the data was used for training in each run and 20% was used for testing, where we report the average. To achieve a balanced training data set, we over-sampled the minority class of data labeled as cyberbullying. For the dataset Set40+, we found that 49% of media sessions were labeled as non-cyberbullying and 51% as cyberbullying. For the dataset Set0+, we have 15% non-cyberbullying and 85% cyberbullying examples.

Based on our analysis in Section 6.3, we evaluated three types of features, namely those features obtained from the content of comments, those derived from shared media objects, and those obtained from user graph properties of the profile owner, such as the number of followers or followings. For the text features, we removed characters such as "!", ">", etc, as a preprocessing step. We first focused on unigrams and bigrams. LIWC categories are derived from unigrams and hence implicitly included as part of this text analysis. Next, removing stop words such as "and","or","for", etc. was employed, which proved beneficial. Finally, each feature vector was normalized by removing the mean zero and scaling to unit variance. In this section we focus on using a logistic regression classifier to be able to interpret the capability of the features in detection of cyberbullying and compare it with results obtained from the analysis section.

Table 7.1: Cyberbullying detection's classifier performance

Features	Dataset	F1-measure	Precision	Recall
unigram	Set0+	0.45	0.54	0.40
unigram	Set40+	0.81	0.87	0.77
unigram, bigram	Set0+	0.5	0.46	0.54
unigram, bigram	Set40+	**0.84**	0.85	0.83
Reduced feature set (83 words)	Set0+	0.53	0.47	0.60
Reduced feature set (83 words)	Set40+	0.83	0.88	0.77
Reduced feature set (10 words)	Set0+	**0.58**	0.50	0.68
Reduced feature set (10 words)	Set40+	**0.85**	0.89	0.81

Table 7.1 illustrates the improvement of the results for different features. **As we see, the logistic-regression classifier based on text-based unigrams and bigrams achieves the highest recall (54%) for the Set0+ set and recall 83% on Set40+ with dimension more than 200,000. However by reducing the dimension to 10, we obtained better results in both sets**. The reason is due to the high correlation of unigram features, which causes the collinearity problem.

First, we remove the words with coefficient close to zero. Next, we keep removing the variables with high correlation, high p-value and small coefficient, and check the F1-measure after each

removal to make sure it will not degrade the performance of the model.

In addition, a variety of non-text features were evaluated, including those features extracted from user behavior (number of shared media objects, following, followers), media properties (likes, post time, caption) and image content.

First we add number of words. It was observed that number of words is considerably higher for examples of cyberbullying. Adding this feature does not provide any value to the classifier performance. Looking at the correlation table, we realize there is high correlation between number of words and a set of variables with positive coefficient, namely "bitch", "fuck", "gay", "hate", "shut", "suck", "ugly". Next we add "time interval" variable. This variable also has high correlation with cyberbullying indicator words. Both of these support our correlation analysis for "time interval" and "word count". Another feature related to the media session is number of likes the image has received, however it does not provide any improvement with a very small logistic coefficient in the model.

We then added the image content features. We first remove the features with small coefficient, then examine the correlations. "sick" has positive coefficient and has been seen a lot with Tattoo images. We observe positive correlation between Tattoo and sick. Drugs, Bike and Tattoo are the features with highest percentage for cyberbullying/non-cyberbullying sessions. However, these features are also correlated highly with text features and were unable to improve the classifier's performance.

After media session related features, we considered user related features, meaning total shared media objects, followers and followings. Total number of media is negatively correlated with words with positive coefficients, meaning it has a higher percentage for cyberbullying incidents as we saw in the analysis section. Due to the high correlation and collinearity problem, it causes the performance of the classifier to decrease. Considering follows and followings does not provide any improvement either. There was no significant correlation between these two features and other text features. We should recall the t-test over the mean value of these two features had high p-value, suggesting there is no significant difference between the values of these features for cyberbullying

and non-cyberbullying classes.

Observing the correlation between text features and non-text features brings the insight for prediction of cyberbullying incidents using the non-text features. Next we design a logistic model for prediction of cyberbullying incidents.

7.2 Prediction of Cyberbullying Incidents

The idea here is to predict whether cyberbullying will occur when an image is posted on Instagram, without benefit of using any subsequent comments in the media session, as we had in the cyberbullying detection. That means our prediction is based only on the initial posting of the media object, any image features derived from the object, and any properties observed at the time of posting, such as graph features of the profile owner.

To perform the prediction accurately, we need to augment our previous set of labeled media sessions, which were pre-filtered based on text analysis to contain only those with profanity, to also include media sessions lacking in profanity. This will create a labeled data set that is independent of the content of text comments, namely profanity, and hence usable for prediction, which will not have access to comment contents. Therefore, we randomly selected 1164 media sessions from the ones with no profanity usage with the criteria of having more than 15 comments and labeled the new set by the same methodology as before. In total 80 contributors worked on the quiz mode, 68 passed, 11 failed and 1 gave up. Labeled data was obtained from 59 trusted contributors as 9 contributors were further filtered out during the labeling process, because they either failed the work mode or rushed through their labeling process. We also labeled the image contents of the new set of media sessions. Table 7.2 shows the statistics related to this labeling process. Furthermore, pursuing the same methodology that we used of cyberbullying detection, we only consider the media sessions whose labeling had a confidence level more than 60%. Overall, it gave us 1,142 labeled media sessions.

Table 7.3 shows the user social graph measurements for the labeled media sessions. We observe that p-values are all less than 0.1 for followers and following, which was not the case for

Table 7.2: Labeling process statistics. Trusted judgments are the ones made by trusted contributors.

Trusted Judgments	5,638
Untrusted Judgments	72
Average Test Question Accuracy of Trusted Contributors	82%
Labeled Media Sessions per Hour	8

the same features in Table 6.2. This suggests that these features will be more helpful in prediction of cyberbullying than in detection.

Table 7.3: Mean values of social graph properties for cyberbullying versus non-cyberbullying samples and cyberaggression versus non cyberaggression, ($^{**}p < 0.05, *p < 0.1$).

Label	*Media objects	*Following	Followers
Non-cyberbullying	1,157.8	721.7	**398,283.7
Cyberbullying	1,198.3	626.5	**465,376.1
Non-cyberaggression	1,152.6	724.4	*393,901.6
Cyberaggression	1,204.3	640.3	*440,403.6

The examined features include image contents, media session properties, such as post time and caption, and profile owner social graph properties, such as the number of followers and following and total number of shared media objects. A similar methodology as in the detection section was applied for training and testing of a predictor for a total 3,096 media sessions. Here we have three different sets, with newly added media sessions with no example of cyberbullying (Set0-). Table 7.4 provides the results for each of the three sets. It shows that **cyberbullying incidents can be predicted with recall 99% for Set0+ and recall 61% for Set40+. While the best false positive rate over Set0- is 3% , using only the image contents, media and user meta data based on a logistic regression classifier.** In the previous section and also from the new analysis in this section we observed there is a correlation between non-text features and text features that was the main motivation to use non-text features for prediction of the ensuing comments in a session.

In addition, we were interested to explore if prediction could be improved using only a limited

set of early comments, not the complete set of comments for a media session. Hence we also give in Table 7.4 the results of prediction using the first 15 comments. We see that our predictor is able to improve its recall for Set0+ and Set40+ to 72% and 100% respectively using a logistic regression classifier. The false positive rate over Set0- has reduced to 1% with the help of text features. In future work we will add features extracted from the previously shared media sessions to improve predictor performance.

Table 7.4: Cyberbullying prediction's classifier performance. User properties are followers, following and total number of shared media objects.

Features	Set	F1-measure	Precision	Recall	False Positive
Image content	Set0+	27	15	98	83
Image content	Set40+	56	62	51	37
Image content	Set0-	-	-	-	24
Following, Image content	Set0+	37	23	91	48
Following, Image content	Set40+	62	68	51	18
Following, Image content	Set0-	-	-	-	3
Followers, Following , Image content	Set0+	42	28	88	34
Followers, Following, Image content	Set40+	68	75	60	22
Followers, Following, Image content	Set0-	-	-	-	5
Media objects ,Followers, Following, Image content	Set0+	45	31	87	30
Media objects ,Followers, Following, Image content	Set40+	69	77	62	21
Media objects ,Followers, Following, Image content	Set0-	-	-	-	4
Post time ,User properties, Image content	Set0+	52	38	88	23
Post time ,User properties, Image content	Set40+	67	76	61	22
Post time ,User properties, Image content	Set0-	-	-	-	4
Caption,Post time ,User properties, Image content	Set0+	57	40	99	23
Caption ,Post time ,User properties, Image content	Set40+	67	76	61	22
Caption ,Post time ,User properties, Image content	Set0-	-	-	-	3
Comments, Caption,Post time ,User properties, Image content	Set0+	66	50	100	14
Comments, Caption,Post time ,User properties, Image content	Set40+	75	78	72	22
Comments, Caption,Post time ,User properties, Image content	Set0-	-	-	-	1

7.3 Summary

Based on Instagram labeled data, a logistic regression classifier was used to examine the prediction power of features analyzed in previous section. The designed model can detect cyberbullying incidents with high negativity with 81% recall and 89% precision using only a 10-dimension feature set extracted from comments. Using the same model, recall and precision for cyberbullying incidents with less negativity are 68% and 50% respectively. Where non-text features were not helpful

in detection, some of them had high correlation with text features, give us the insight for using non-text features in prediction of cyberbullying incidents. The non-text features such as image and user meta data were central to cyberbullying prediction, where a logistic regression classifier achieved 72% recall and 78% precision for Set40+ and 100% recall and 50% precision for Set0+. The false positive rate over Set0- is as low as 1%.

Chapter 8

Analyzing Factors Impacting Revining and Propagation of Cyberbullying Incidents on the Vine Social Network

Diffusion of information in the Vine video social network happens via a **revining** mechanism that enables accelerated propagation of news, rumors, and different types of videos. In this chapter we aim to understand the revining behavior in Vine and how it may be impacted by different factors. We first look at general properties of information dissemination via the revining feature in Vine. Then, we examine the impact of video content on revining behavior. Finally, we examine how cyberbullying may impact the revining behavior. The insights from this analysis help motivate the design of more effective information dissemination and automatic classification of cyberbullying incidents in online social networks.

8.1 Overview

Online social networks such as Facebook, Twitter, YouTube, and Vine are attracting more users every day, and they have become an important source of information sharing and propagation [4]. Vine is a video-based online social network and has become increasingly popular recently in the Internet community. There are almost 40 million registered Vine users as of April 2, 2015 while the total number of Vines played everyday is 15 billion [18]. Using a mobile application, Vine users can record and edit six-second looping videos, which they can share on their profiles for others to see, like, and comment upon. An example of the Vine social network is shown in Figure 8.1. All user profiles in Vine are public by default unless users change their privacy policies. In the

public setting, posts are accessible by all Vine users, not only followers. Using the privacy setting, Vine users can limit the access to their posts to their followers only. Users can also limit who can find them or message them. In the home feed page, featured Vine videos in different categories are provided to a user when logged in.

Due to the wide popularity of online social networks, they have been used for sharing news, art, politics [12, 43, 46] and have been researched from different areas, such as marketing and sociology [46]. Vine videos are getting more popular, as they can be easily embedded in Twitter, and the auto-loop feature makes it funny and interesting [70]. One of the most interesting characteristics of Vine is the revining behavior. That is, spreading information in Vine happens via revining a shared 6-second video referred to as a "vine". Users can easily share a video by pressing the revining button provided below the video. Figure 8.1 also illustrates the revining behavior.

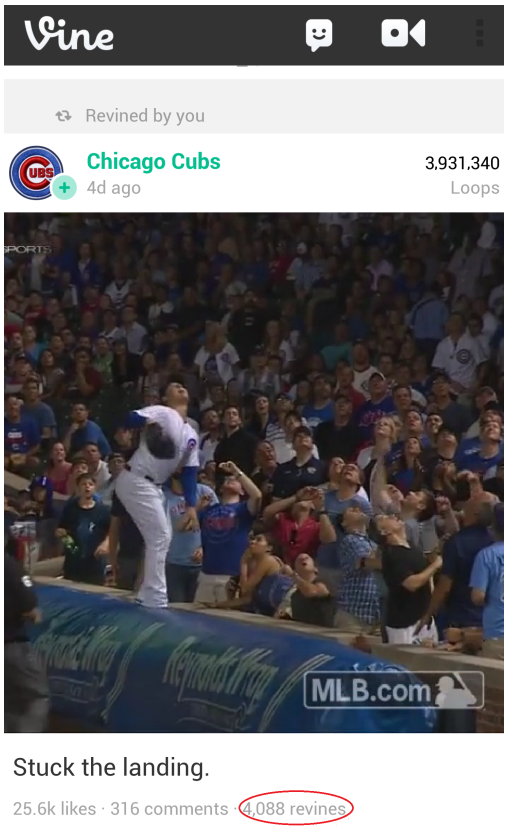

Figure 8.1: An example of the Vine social network with revining.

Previous work [52, 55] provided detailed characterization of Twitter, looking at the interaction among the users and the temporal behavior of users in Twitter. Other works have looked at the retweeting behavior in Twitter [96, 12, 99, 57, 83]. Pezzoni et al. measured the influence of a user based on the average number of times that his/her originated tweets have been retweeted [83]. Beside user properties, they also considered the position of the tweet in the feed as a factor impacting retweeting behavior. The authors of [59, 9] have considered the number of times a recent seed tweet has been retweeted as the influence of the post, and used this to estimate the total influence

and influence score of the user, and then examined a network of influencers. Further, [9] observed that URLs that were rated as having more positive feelings have been spread more. The authors evaluated the impact of Twitter users in different topics, instead of labeling them as influencer or not influencer. While they built the propagation tree by looking at the followers retweeting, [8] looked at the retweeting methods other than formal retweeting mechanism of Twitter. Looking at different domains, they observed the percentage of retweets coming from non-followers is higher than that from followers. Zhao et al. in [117] considered two types of retweets, coming from direct and indirect followers of the user who originated a tweet. They proposed that looking at the influence of the followers is also important for predicting the number of retweets. There are also considerable amount of works trying to predict the number of retweets [82, 114, 40, 117, 15]. In [15], the authors used image features to predict the number of retweets by looking at the image link tweets. Another work [65] analyzed the behavior of the tweets that have been retweeted many times and tried to detect two classes of tweets, first the group of tweets that have been retweeted less than 30 times, and second the group of tweets that have been retweetd more than 100 times. Our work differs from previous works in terms of examining the revining behavior of video-based vines, but leverages prior works in that determining how far a vine has been propagated can be useful as a sense of how influential the post has been.

To date, there has been little prior work examining video-based social networks like Vine. One paper [88] has labeled a small set of about a thousand Vine videos as cyberbullying or not, for the purposes of detecting cyberbullying incidents in Vine. That work does not investigate how vines propagate via revining in the social network, which is the focus of this chapter.

It is the first work to provide a detailed characterization of key properties of Vine, a video-based social network. Second, we labeled the content and emotion of a small set of videos and explore their relationship to revining behavior. Third we reveal the difference in revining behavior between vine videos labeled as cyberbullying, and those that are not.

In the following, we first describe our data collection efforts, basic analysis and then provide more characteristics regarding revining. We then analyze the revining behavior of videos both in

terms of their labeled video content and cyberbullying content.

8.2 Dataset

Vine is a 6-second short video sharing platform, launched in 2012. Users can create videos recorded by their mobile phone camera, and apply edits provided by the mobile app. Using snowball sampling, we collected the complete profile information of 55,744 users in Vine. Specifically, starting from 5 random seed users, we collected data for all users within two tops of the seed users, i.e., followers of the seed users, and users who follow those followers. This gives us in total 390,463 unique seed videos generated by these users (approximately 7 vines for each user). For each user we collected their total number of followers, followings, total number of videos created by the user (seed videos) and total number of videos revined by the user. We also collected the total number of likes, revines, and loops (# of times video has been played) for each user.

Table 8.1: Collected features for users and seed vines.

Follower	list of followers of user
Following	list of followings of user
uSeedPost	number of posted videos originated by a user
uPost	number of total posted videos by a user (originated plus revined)
uLoop	number of times all posted videos by a user have been played
uLike	number of times all originated videos by a user have been liked
PostDate	the exact date and time the seed video was originated
Description	the attached caption and tags to the posted video
Revinee	list of all the users who shared the seed video
pLoop	number of times the seed video has been played
pLike	number of times the seed video has been liked
pComment	number of comments the seed video has been liked
pRevine	number of times the seed video has been revined

For each of the seed videos, we collected the total number of likes and comments associated with the posed video. Also we collected the creation day, how many times the video has been played (i.e., loop counts) in July 2015. Also we looked at how many times each video has been revined by collecting the user id of all the users who have revined it, along with the time stamp when it has been revined. The complete list of users who have revined a video is accessible directly

by collecting information associated to the seed video itself. Table 8.1 summarizes the features that were collected for users and their seed videos.

8.3 User Behavior on Vine

8.3.1 Basic Analysis

We first examine the general behavior of users in Vine, including the number of followers and following, which are user based features, and the number of comments, likes, revines and loops of a video, which are post based features. Figure 8.2 shows the distributions of the number of followers and following of 55,744 users as complementary cumulative distribution functions (CCDFs). Compared with previous work that reported the following and follower distributions in Twitter [55], there is a much larger gap between the distribution of the number of followers and the number of followings of Vine users. Only 1.3% of the Vine users follow more than 10,000 users, however 19.27% have more than 10,000 followers.

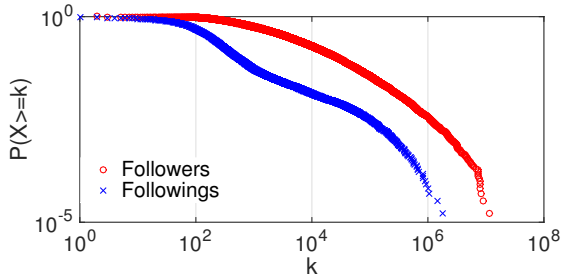

Figure 8.2: CCDFs of the number of followers and followings in Vine.

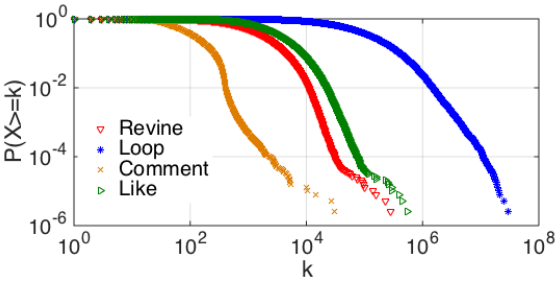

Figure 8.3: CCDFs of the number of revines, loops, comments, and likes in Vine.

Figure 8.3 provides the CCDFs for the number of times a seed video has been played as a loop, has been shared as a revine, has been liked and has been commented upon. The curve for looping is much higher than the other three (i.e., there is a long tail), possibly due to the auto-looping feature. Revining and liking have lower and very similar distributions. Commenting has the distinctly lowest CCDF curve among the features considered for a post.

Table 8.2 provides the mean, median and maximum values for all user features, including the

Table 8.2: Statistics on the Vine dataset, for 55,744 users.

	Follower	Following	uLoop	uSeedPost	uLike	uPost
Mean	23,571.14	1,459.0	5.67×10^5	147.2	3,630.0	705.6
Median	1,434.0	99.0	2.97×10^5	52.0	826.0	160.0
Maximum	1.16×10^7	1.83×10^6	2.10×10^9	1.9×10^4	2.10×10^6	7.10×10^4

total number of followers and followings, total number of posts originated by a user (uSeedPost), total number of videos posted by a user (originated+shared), total number of likes a user has received (uLike) and total number of times all his/her posts have been looped (uLoop). We observed that there are outliers that are a large distance from the mean. Namely, the substantial deviation of the mean from the median for nearly every category shows there are a set of Vine users whose feature values are much higher than the behaviors of most of the population. For example, we noticed celebrities provide such a distortive effect, e.g., one celebrity had 12 million followers.

Table 8.3: Pearson correlation between the collected Vine features. Only for the vines with $*$ is the p-value larger than 0.001, while for the rest the p-value is smaller than 0.001.

	pLike	pComment	pRevine	pLoop	Follower	Following	uLoop	uSeedPost	uLike	uSharedPost
pLike	1.000	**0.672**	**0.840**	**0.596**	**0.473**	-0.009	0.246	-0.018	-0.034	-0.060
pComment	0.672	1.000	**0.692**	0.408	0.209*	0.000	0.131	-0.010	-0.007	-0.03
pRevine	0.840	0.692	1.000	**0.421**	0.295	-0.005	0.144	-0.017	-0.051	-0.046
pLoop	0.596	0.408	0.421	1.000	0.25	-0.006	0.164	-0.033	-0.038	-0.042
Follower	0.473	0.209	0.295	0.256	1.000	0.011	**0.378**	0.141	0.022	0.043
Following	-0.009	0.000*	-0.005	-0.006	0.011	1.000	-0.006	-0.013	0.002*	-0.004*
uLoop	0.246	0.131	0.144	0.164	0.378	-0.006	1.000	0.156	0.030	0.069
uSeedPost	-0.018	-0.010	-0.017	-0.033	0.141	-0.013	0.156	1.000	0.199	**0.512**
uLike	-0.034	-0.007	-0.051	-0.038	0.022	0.002*	0.030	0.199	1.000	0.203
uSharedPost	-0.060	-0.032	-0.046	-0.042	0.043	-0.004*	0.069	0.512	0.203	1.00

Table 8.3 displays the correlations among different user and post based features. pLike, pComment, pRevine and pLoop are post features and uLoop, uLike, uSeedPost (# posts originated by user) and uSharedPost (# posts revined by user) are user features. The highest correlation is among revines and likes for posts, as was seen in Figure 8.3 where their CCDFs were also very similar. Revining is also correlated with the number of comments. The correlation of revines for a post with the loop feature is high but is the lowest among post based features. In terms of correlation of the revining with user based features, there is a positive correlation between revining

of a post and number of followers that the poster of the seed vine has. There is a very small negative correlation between the number of revines and followings.

There is no correlation between the number of followings a user has and the other user features. The number of followers has considerable correlation with the total number of loops for all seed videos posted by the user (uLoops). Also the correlation between user's seed post and user's shared post (uSharedPost) is about 0.5, showing users who tend to share more videos also generate more seed videos.

In total, from all posts, there were 366,517 tags in total and 83,147 unique hashtags. As Figure 8.4 illustrates, only 8.25% of the posts have more than 3 tags and the highest number of tags for a video is 20.

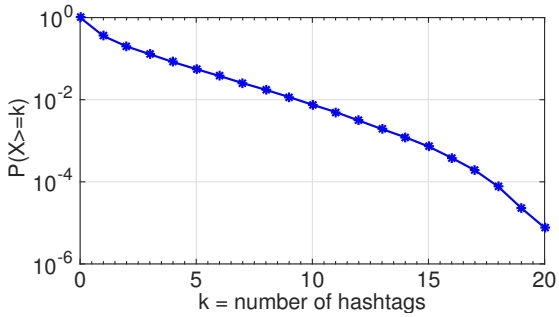

Figure 8.4: CCDF of the number of tags.

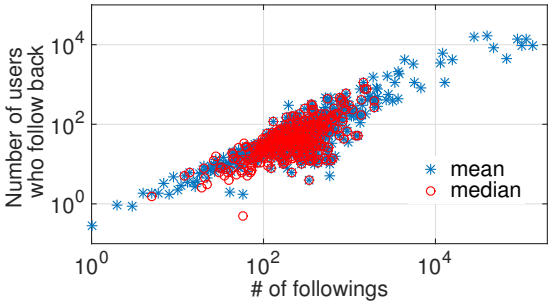

Figure 8.5: Mean and median number of users who follow back a user versus number of followings.

As another property of the Vine social network, we investigated the number of users who follow back their following users. Figure 8.5 shows a positive correlation between the number of followings and the mean number of users who follow back their following users.

8.3.2 Revining Behavior and Followers and Following

In this subsection, we explore in more detail the relationship between revining behavior and the numbers of followers and following. First, we establish a baseline for originated videos. Figure 8.6 plots the correlation between the number of originated vine videos and the number of

followers. For up to about 1000 followers, there is a linear relation between the log number of followers and log number of mean videos and after that there is no positive trend. We also provide the mean and median for the log scale bins in solid and dash lines, [55]. The mean is clearly above the median, indicating again there are outliers who send original vines much more than other users. In Figure 8.7, we plot the correlation between the number of originated vines and the number of followings. Up to about a value of 665, the line is linear and after that there is a negative trend between mean value of vine and followings.

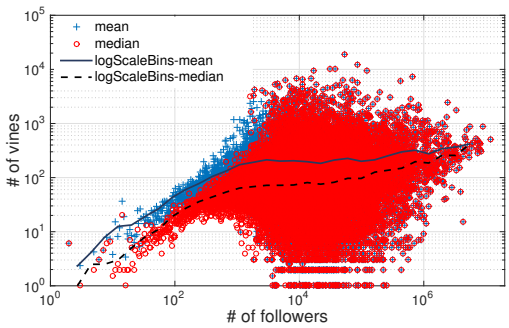

Figure 8.6: Mean and median number of original vines versus number of followers.

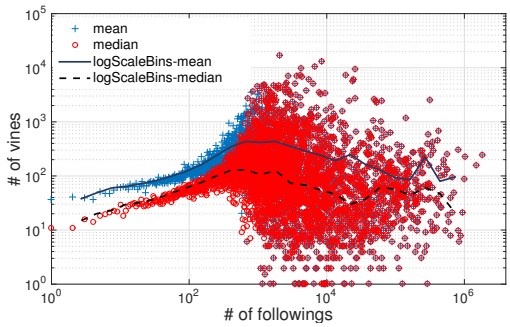

Figure 8.7: Mean and median number of original vines versus number of followings.

In terms of revines, Figure 8.8 shows the relation between the number of followers and the mean number of revines. Figure 8.8 demonstrates that as the number of followers increases, the mean number of revinings increases. For up to 245 followers there is positive correlation, and as followers increase the variance of the mean number of revines increase. On Figure 8.9 it can be seen that as the number of following increase the mean number of revines is approximately constant for values close to 245 and after that the variance starts to increase. But there is no positive correlation between number of followings and the number of revines for a post.

8.3.3 Revining and Temporal Behavior

Next, we explore temporal properties of revining. Figure 8.10 provides the CCDF of the maximum number of times a seed vine has been revined per day. 19% of the vine videos have reached the maximum revining of more than 1000 times a day. A few seed videos have been

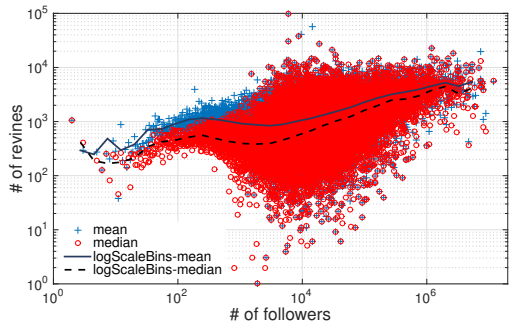

Figure 8.8: Mean and median number of revines of the original vines versus number of followers.

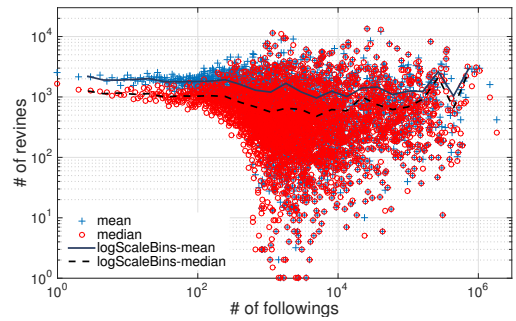

Figure 8.9: Mean and median number of revines of the original vines versus number of followings.

revined more than a maximum of 10,000 times in a day.

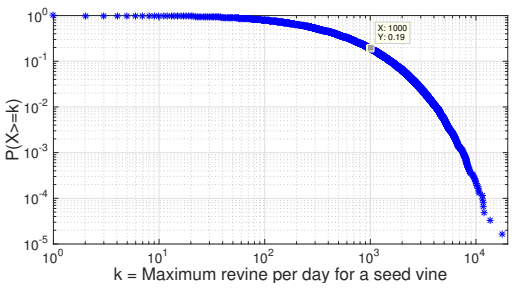

Figure 8.10: CCDF of maximum number of times a seed vine has been revined per day.

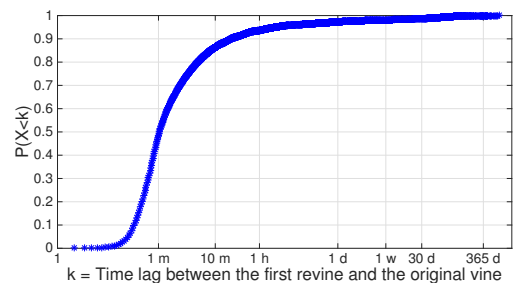

Figure 8.11: The time lag between the first revine and the original vine.

Figure 8.11 shows the time difference between the first revine occurring after an original vine has been posted. 48% of the first revines have occurred within one minute and 93% have happened within one hour.

Figure 8.12 shows the activity of users in terms of posting seed vine videos during the seven days of the week, and across 24 hours. The minimum fraction of the seed videos were posted during the morning, 6am-12pm. The activity increases after that until reaching its peak around 10pm-11pm. The trend of the posting behavior differs somewhat for the weekend, shifting to a higher fraction of posts earlier in the evening, with high proportion of user activity from about 6pm to midnight. On the right side figure, we provide the mean number of likes, comments, revines and loops for the videos posted at time x during a day. It is interesting to see a pattern, showing

the videos posted in the evening received on average more likes, comments and are revined more. However, the number of loops shows a completely different behavior and the highest average number of loops belongs to the videos posted at 11am.

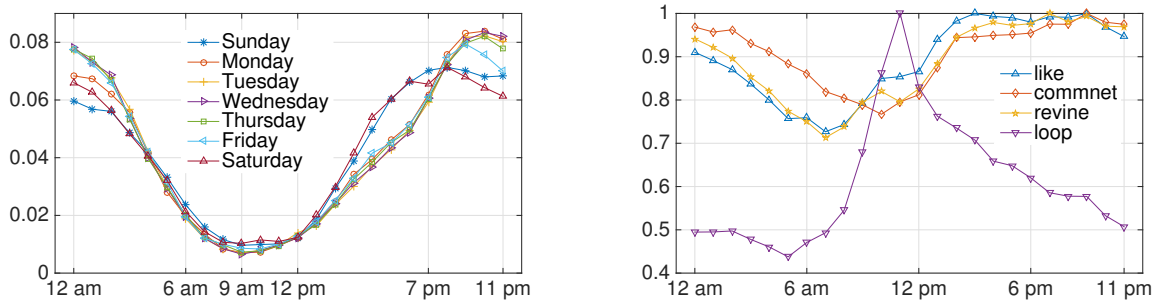

Figure 8.12: (left) Percentage of vines posted at time x during a day. (right) Normalized number of likes, revines, comments and loops for posts received at time x in a day. Each graph has been normalized to one by dividing to its highest value.

8.3.4 Revining and Hash Tags

Another factor we are interested to explore is the correlation between the post's hash tags and the number of times it has been revined. Table 8.4 provides the tags with the highest frequency in four different bins of revining. The hash tags are ordered so that the tags on the left have the highest frequency. Each bin contains a quarter of the seed Vine videos. For example 1-bin contains one fourth of the media sessions with the lowest amount of revining. Comedy is the top tag in all four bins. Football is among the top 10 most frequent tags only in the highest quadrant of the number of revines. Also music and singing only occur in the lowest quadrant of revining. The top ten tags for 2-bin and 3-bin are the same, except reordered.

Table 8.4: The tags with highest frequency in 4 different bins of revining.

1-bin	comedy	funny	remake	6secondcover	lol	loop	revine	music	singing	videoshop
2-bin	comedy	remake	funny	lol	loop	revine	6secondcover	onedirection	teamsour	howto
3-bin	comedy	remake	funny	onedirection	loop	lol	revine	howto	teamsour	6secondcover
4-bin	remake	comedy	funny	onedirection	loop	lol	revine	howto	blackranked	football

A visualization of the tags depicted in Figure 8.13 shows that comedy, funny and remake

have the highest frequency and have been seen among the top three tags for all 4 bins.

Figure 8.13: Visualization of the tags associated with the vines

8.3.5 Revining and Followers

Figure 8.14 provides the CCDF for the percentage of revines that have been made by the followers of the user who posted the original Vine video. Since all users in our dataset have public profiles and their profiles and seed vines are accessible by all Vine users, revines are not just limited to their followers. In fact, for 1.67% of the seed vines, none of the revinings have been by the followers of the poster of the original vine. For 58% of the users, more than 90% of the propagation takes place by users who are not direct followers of the users. These include celebrities, sports pages, local singer pages, fun-related pages, etc. Previous work [8] has reported non-follower retweets range from 78.7% to 98.5% for four different domains "Fundraiser", "News", "Petitions" and "YouTube".

Figure 8.15 displays the relation between the number of additional recipients of the Vine video (users other than the original poster's followers) and the number of followers. There is a positive correlation which shows the higher the number of followers a user has, the greater are the number of users other than followers who access the video post.

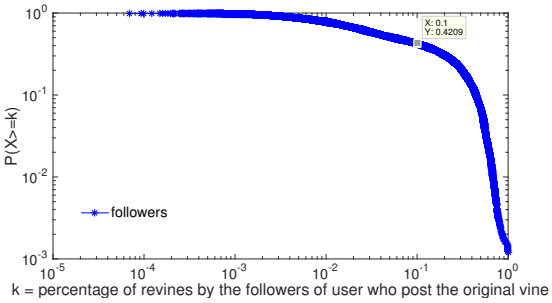

Figure 8.14: CCDF of the percentage of the reviners who are the followers of the source of the original vine.

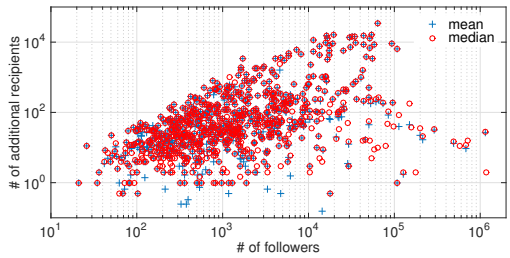

Figure 8.15: Mean and median number of additional recipients of the vine via revining.

8.3.6 Revine Tree

Figure 8.16 shows the number of hops a video gets propagated in Vine. For this purpose we collected the user name of all the users who revined a seed video. We then find the ones who are followers of the poster of the seed video and named them as first hop users. At the next step we collected the followers of the first hop users and compared them with the remaining users in the revining list. The common users, if they exist, are named as second hop users. We continued the process until there are no more users in the x-hop user list. 6.7% of the videos have been propagated 0 hops, meaning none of the revining users have been among their followers. The most common number of hops is 3, i.e., 24.96% of the videos have been propagated 3 hops by their followers. In comparison, for Twitter, the most common number of hops reported is 1 for 85% of the tweets [55]. Also they show the highest number of hops is 10 in tweet propagation, while for Vine the largest number of hops we have discovered is 16. Hence, Vine videos are propagated more on average, and also more in the extreme.

Figure 8.17 shows an example of a revine tree. The seed video has been revined 8401 times. Looking at the links among these 8401 users, 4501 users have connection with at least one of the 8401 users, with 4595 links. This means that around half of the users that were not among the followers of the original poster have revined the video. As the poster of this Vine video has a public profile, other users who do not follow this user also can see his/her profile content and share the

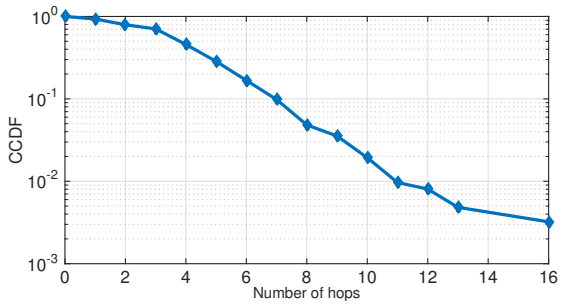

Figure 8.16: CCDF of number of hops a video get propagated through the network.

seed video. The dark blue nodes in the right side shows a propagation of video in the network in a community that is not at all connected to the main network component. Most of the nodes who are not followers of the original poster are single node components that have not lead to more revining.

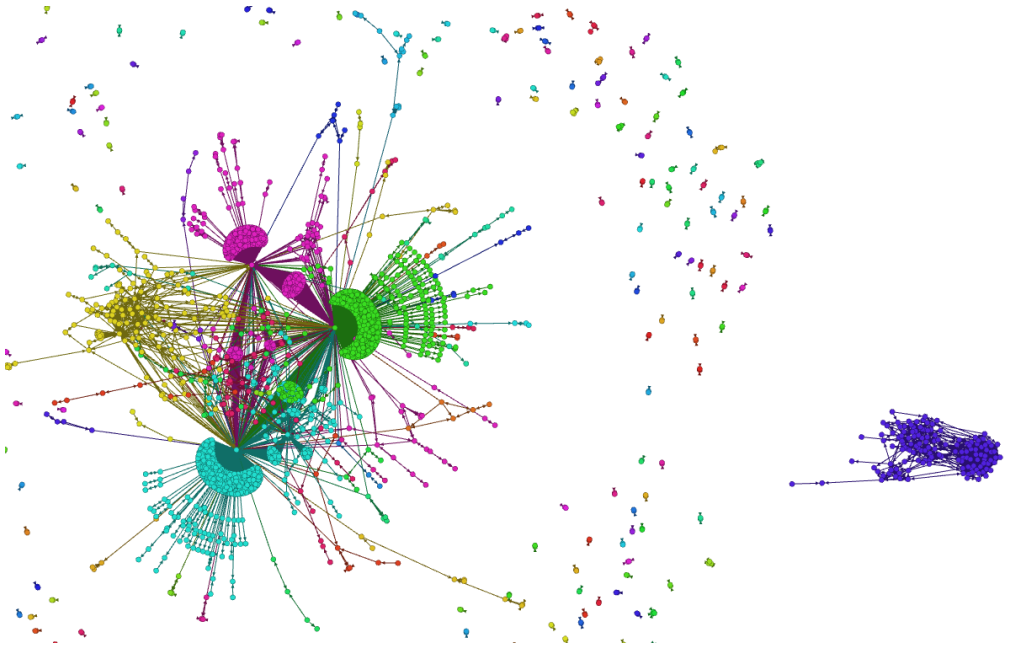

Figure 8.17: Revine tree of followers of a seed Vine video who revined the post.

The node with highest degree in the pink group is the original poster. Looking at the revining time of the other two nodes with highest in-degree (light blue and green groups), both revined the video in the first 24 hours, one being the first user who revined and the other being the 5th user who revined the video. These later reviners have been more influential than the original poster in

the propagation of this vine. Previous work [117] has also observed that real propagation in tweeter happen with a pattern similar to Figure 8.17.

8.4 Labeled Data Analysis

In this work, we are specifically interested in how revining behavior is affected by the content of the videos as well as the presence of cyberbullying in the vines. This requires that we label various Vine videos. We selected a set of 983 vines, and collected the appropriate user features of the seed video posters, and the feature set of the posts. Using CrowdFlower, a crowd-sourced website, we labeled the video content and emotions of these vines according to the methodology from [38]. For the video content, People, Person, Indoor, Outdoor, Cartoon, Text, Activity, Animal and Other were provided as categories in multiple choice format to CrowdFlower contributors. Next, the contributors were asked to identify the emotions expressed in the video, given the following options to choose from: Neutral, Joy, Sad, Love, Surprise, Fear and Anger. The same set of videos were labeled for cyberbullying according to the methodology from [88]. Each media session (including a video and its associated comments) was labeled by five different contributors.

8.4.1 Revining and the Content/Emotion of Vines

We wanted to determine the relationship between the content/emotion of the video and the number of times it gets revined. Figure 8.18 illustrates the mean and median number of times a seed vine is revined for each of the categories we labeled for the vines. We observed “car”, “activity” and “outdoor” have the highest mean. However there is big gap between their mean and median, showing there are outliers with a large amount of revining for these categories. For categories like “food”, “fashion” or “cartoon” the mean is close to the median, revealing the lack of significant outliers. For these categories, our analysis found that in terms of the mean number of hops that the “car” category had been propagated on average 5 hops, which was the highest mean among other categories. Next highest was the“text” category, which was propagated with an average 4.4 hops. The category “activity” and “other” had the minimum average number of hops around 2.7.

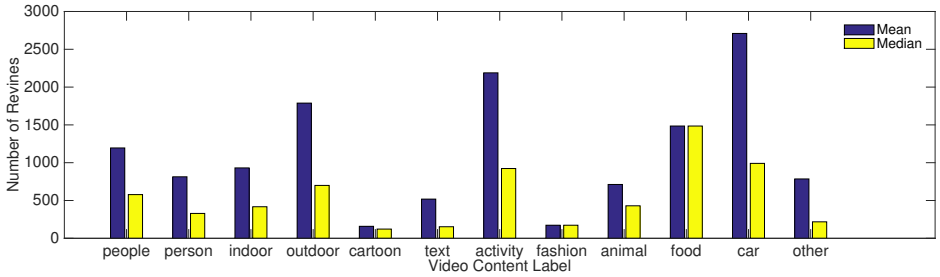

Figure 8.18: Mean and median of number of times a video has been revined versus the content of the video.

Figure 8.19 illustrates the mean and median number of times a seed video has been revined versus the emotional content of the video. We find that "anger" and "sad" have the highest difference between the median and the mean. Positive emotions such as "love", "joy" and "surprised" have higher mean number of revines compared to negative emotion categories such as "sad", "fear" and "anger". "Neutral" behavior seems to behave closer to negative feelings than positive emotions. **In terms of the mean number of hops of revining propagation for different emotions, we found that "love" had the highest mean value 4.4 hops on average and "sad" had the lowest mean value of 2.8 hops.** Previous work [9] also observed that URLs that were rated as having more positive feelings have been spread more. For the rest of the emotion categories the mean value of the category is close to the mean number of hops 3.5.

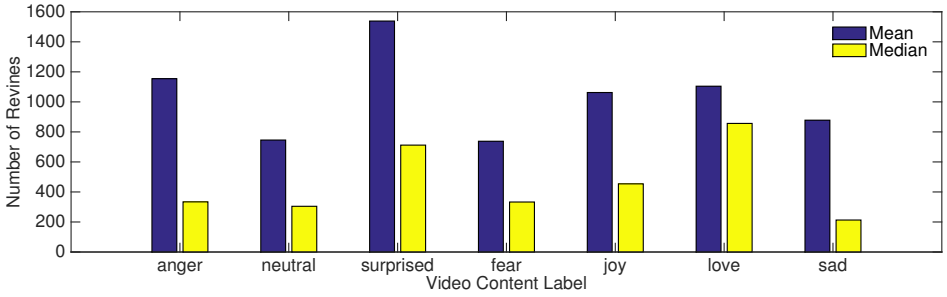

Figure 8.19: Mean and median number of times a seed video has been revined versus the emotional content of the video.

8.4.2 Revining and Cyberbullying Vines

Looking at the first revining of the videos for each group, the lag between an original seed vine and its first revine is less than 1 minute for 40% of cyberbullying vine videos and 50% for non-cyberbullying vine videos. The average time that a cyberbullying vine was revined is 4.01 days versus a mean of 1.27 days for non-cyberbullying vines (p-value 0.05 for t-test). This means the first revines happen faster for the non-cyberbullying group. We also looked at the longest time between two consecutive revining of a seed post for cyberbullying and non-cyberbullying Vine sessions. The average value over the longest inactivity period for cyberbullying is 94.91 days versus a mean of 111.15 days for non-cyberbullying (p-value 0.01 for t-test).

We found that the mean number of revining hops for cyberbullying videos was 3.21 while the mean number of hops for non-cyberbullying videos is 3.78 (p-value 0.003 for t-test). *Not only are non-cyberbullying videos shared more, but they also propagate through the network more deeply compared to cyberbullying videos.*

Tables 8.5 and 8.6 provide the mean values for the two different classes of cyberbullying and non-cyberbullying versus user and post features. The owner of cyberbullying video sessions have fewer followers and they also follow fewer users. They have lower number of total likes for their posted videos and the total number of loops is also smaller. However they have shown more activity in terms of posting seed Vine videos compare to non-cyberbullying class users.

Table 8.5: Mean of user features for cyberbullying and non-cyberbullying class.

	Follower	Following	uLike	uLoop	uSeedPost	uPost
cyberbullying	9.10×10^4	2.27×10^3	5.97×10^3	1.5×10^3	5.12×10^2	1.53×10^3
non-cyberbullying	1.07×10^5	3.03×10^3	7.35×10^3	4.15×10^7	4.44×10^2	1.38×10^3

Table 8.6: Mean of post features for cyberbullying and non-cyberbullying class.

	pRevine	pLoop	pComment	pLike
cyberbullying	719.97	139,534.5	88.5	5,973.3
non-cyberbullying	1,095.6	118,326.2	76.3	7,355.8

Table 8.6 shows that cyberbullying labeled vines have been revined less, though they have been looped more. Also they receive more comments though fewer likes compared to the mean value of the non-cyberbullying labeled vines.

8.5 Summary

As far as we are aware, this work is the first to present a detailed analysis of user behavior in the Vine social network. We analyzed 55,744 profiles of Vine users, along with 390,463 seed Vine videos generated by these users. There are six key findings. First, the number of followers has a positive correlation with the number of seed vines for up to 1000 followers, then the variance of the number of seed vines grows too large. There is also a positive correlation between the number of followers and those revining. Second, for 42% of the users, only 10% of the revines come from the followers, meaning they are responsible for only a small percentage of the propagation of the posts. Third, there is a positive correlation between the number of additional recipients and the followers, supporting the first two statements. Fourth, Vine users are most active in the late evening hours; they are most likely to post original videos after 10pm and receive the greatest number of likes, comments and revines. However the peak of the loops belong to the videos posted at 11am. Fifth, using a smaller set of labeled videos based on their content and emotions, we observed that the videos containing the emotion of "love" have been propagated more deeply, and videos containing the emotion of "sadness" have traveled the lowest number of hops in the network compared with the other emotions. Finally, when examining the differences between videos identified as cyberbullying and those that were not, we found that cyberbullying video sessions were revined less, traveled fewer hops in the network, but looped more and received more comments. Additionally, users who posted cyberbullying videos have fewer followers and following, but they are more active in posting seed videos. We plan to incorporate these findings into the design of more effective information dissemination and automatic classification of cyberbullying incidences in online social networks.

8.6 Acknowledgments

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

Chapter 9

Conclusions

This thesis makes the following major contributions:

9.1 Summary

Chapter 4 present a detailed analysis of user behavior in the Ask.fm social network. We analyzed nearly 30K profiles of Ask.fm users using interaction graphs, word graphs, and frequency distributions, and characterized key properties such as reciprocity, clustering coefficient, and the influence of negativity on in-degree and out-degree. Some of the key findings of the work are that (1) When people have highly negative profiles without any positive support, they also have the lowest activities in terms of both in-degree and out-degree, that is they are the least sociable. (2) As the amount of positive support increases, we find a greater in-degree and greater out-degree of the users, that is users become more social and actively like and are liked more often. For example, users with negative profiles that also receive positive support have higher in-degree and out-degree than owners of highly negative profiles that lack positive support, that is they are more sociable. When people have highly positive profiles, then they also have the highest in-degree and out-degree, which shows that they socialize the most on this social network. This suggests that we may be able to use the interaction graph's in-degree and out-degree as an indicator of the extent of negativity or positivity on a given profile. (3) The negative in-degree and negative out-degree do not exhibit similar behavior, unlike the similarity of in-degree and out-degree found in other social networks. (4) Finally, our analysis of cutting behavior on Ask.fm reveals that such profiles have surprisingly

high positive support, and exhibit a different signature than the other group segments studied.

A user behavior analysis across Ask.fm and Instagram social networks is provided in chapter 5. We found that the users are more negative in Ask.fm than Instagram. It is tempting to attribute the increased negativity on Ask.fm to the anonymity of Ask.fm. However, there are other differences between the two networks that could influence the increased negativity on Ask.fm. For example, Ask.fm users primarily post in text, whereas Instagram is highly focused on posting image media, followed by the commenting. Also, the relationship between Ask.fm owners comments and non-owner "friends" comments is one-to-one, whereas on Instagram, a single owner comment may be directed at multiple friends' comments, such as a "Thank you all". This requires more research into the relationship between anonymity and negativity.

The following major contributions were found in chapters 6 and 7. First, an appropriate definition of cyberbullying that incorporates both frequency of negativity and imbalance of power is applied in large-scale labeling, and is differentiated from cyberaggression. Second, cyberbullying is studied in the context of a media-based social network, incorporating both images and comments in the labeling. We found that labelers are mostly in agreement about what constitutes cyberbullying and cyberaggression in Instagram media sessions. Third, a detailed analysis of the distribution results of labeling of cyberbullying incidents is presented, including a correlation analysis of cyberbullying with other factors derived from images, text comments, and social network meta data.

We found a significant number of media sessions containing profanity and cyberaggression were not labeled as cyberbullying, suggesting that detection of cyberbullying must be more sophisticated than merely looking for profanity. We observed that media sessions with very high percentage of negativity above 60-70% actually correspond to a lower likelihood of cyberbullying. Also, media sessions with cyberbullying exhibit more frequent commenting. We found that users of media sessions containing cyberbullying demonstrate a lower number of likes per post. Finally, cyberbullying has a higher probability of occurring when media sessions contain certain linguistic categories such as death, appearance, religion and sexuality content. Similarly, certain image

contents such as “drug” are highly related to cyberbullying while other image categories such as “tattoo” or “food” are not.

Using Instagram labeled data, logistic regression can detect cyberbullying with 81% recall and 89% precision using only text features of media sessions with more than 40% negativity. Non-text features such as image and user meta data were central to cyberbullying prediction, where a logistic regression classifier achieved 75% recall and 88% precision for the same set of media sessions. In cyberbullying sessions with low negativity, unigrams are not as powerful features as for the cases with high negativity.

Chapter 8, as far as we are aware, is the first to present a detailed analysis of user behavior in the Vine social network, introducing six key findings. First, the number of followers has a positive correlation with the number of seed vines for up to 1000 followers, then the variance of the number of seed vines grows too large. There is also a positive correlation between the number of followers and those revining. Second, for 42% of the users, only 10% of the revines come from the followers, meaning they are responsible for only a small percentage of the propagation of the posts. Third, there is a positive correlation between the number of additional recipients and the followers, supporting the first two statements. Fourth, Vine users are most active in the late evening hours; they are most likely to post original videos after 10pm and receive the greatest number of likes, comments and revines. However the peak of the loops belong to the videos posted at 11am. Fifth, using a smaller set of labeled videos based on their content and emotions, we observed that the videos containing the emotion of “love” have been propagated more deeply, and videos containing the emotion of “sadness” have traveled the lowest number of hops in the network compared with the other emotions. Finally, when examining the differences between videos identified as cyberbullying and those that were not, we found that cyberbullying video sessions were revined less, traveled fewer hops in the network, but looped more and receive more comments. Additionally, users who posted cyberbullying videos have fewer followers and following, but they are more active in posting seed videos.

9.2 Future Work

While we have conducted a negative (and positive) word analysis on posting in Ask.fm and Instagram, we would like to extend this research. Not all occurrences of negative words correspond to cyberbullying, e.g. "that's f**king amazing". We would need to specifically label comments as being examples of cyberaggression.

We plan to incorporate the findings from Vine data analysis into the design of more effective information dissemination and automatic classification of cyberbullying incidences in online social networks.

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

Designing sophisticated classifiers for combining features is required, for example designing different models for different negativity bins. We aim to build end to end services for real users along with collaboration with parents of high school students. Designing incremental classifiers to apply to large scale real time datasets is needed in this regard.

Labeling the comments individually, without providing the context for the contributors. Compare the result with session-based labeling. We can verify whether having the context and image, can help the contributors to provide higher quality labeling. Also we can examine the accuracy of the classifiers in detecting cyberbullying incident with the provided labels in any of the two mentioned approaches.

We have collected the history of the users for prediction of cyberbullying.

Bibliography

[1] Bad word list and swear filter. [accessed 10-November-2014].

[2] Positive word research, 2013. [accessed 25-July-2014].

[3] Cyberbullying research center: "http://cyberbullying.us". Retrieved September, 2013.

[4] Samir Akrouf, Laifa Meriem, Belayadi Yahia, and Mouhoub Nasser Eddine. Social network analysis and information propagation: A case study using flickr and youtube networks. International Journal of Future Computer and Communication, 2(3):246, 2013.

[5] Aris Anagnostopoulos, Ravi Kumar, and Mohammad Mahdian. Influence and correlation in social networks. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 7–15. ACM, 2008.

[6] 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).

[7] Ask.fm wikipedia page, 2014. [accessed 14-January-2014].

[8] Norhidayah Azman, David Millard, and Mark Weal. Patterns of implicit and non-follower retweet propagation: Investigating the role of applications and hashtags. 2011.

[9] Eytan Bakshy, Jake M Hofman, Winter A Mason, and Duncan J Watts. Everyone's an influencer: quantifying influence on Twitter. In Proceedings of the fourth ACM international conference on Web search and data mining, pages 65–74. ACM, 2011.

[10] Michael S Bernstein, Andrés Monroy-Hernández, Drew Harry, Paul André, Katrina Panovich, and Gregory G Vargas. 4chan and/b: An analysis of anonymity and ephemerality in a large online community. 2011.

[11] A. Binns. Facebooks ugly sisters: Anonymity and abuse on formspring and ask.fm. In Media Education Research Journal, 2013.

[12] Danah Boyd, Scott Golder, and Gilad Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter. In System Sciences (HICSS), 2010 43rd Hawaii International Conference on, pages 1–10. IEEE, 2010.

[13] R. Broderick. 9 teenage suicides in the last year were linked to cyber-bullying on social network ask.fm, 2013. [accessed 14-January-2014].

[14] Hunting the Internet Bullies. http://www.bbc.co.uk/programmes/b01c00y3.

[15] Ethem F Can, Hüseyin Oktay, and R Manmatha. Predicting retweet count using visual cues. In Proceedings of the 22nd ACM international conference on Conference on information & knowledge management, pages 1481–1484. ACM, 2013.

[16] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and P Krishna Gummadi. Measuring user influence in twitter: The million follower fallacy. ICWSM, 10(10-17):30, 2010.

[17] National Crime Prevention Council. http://en.wikipedia.org/wiki/Cyberbullying, Retrieved on July 6, 2011.

[18] Craig Smith. By the numbers: 25 amazing Vine statistics. http://expandedramblings.com/index.php/vine-statistics/, 12 July 2012. [Online; accessed 6-Aug-2015].

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

[20] Maral Dadvar, Roeland Ordelman, Franciska de Jong, and Dolf Trieschnigg. Towards user modelling in the combat against cyberbullying. In Natural Language Processing and Information Systems, pages 277–283. Springer, 2012.

[21] Maral Dadvar, Dolf Trieschnigg, and Franciska de Jong. Experts and machines against bullies: A hybrid approach to detect cyberbullies. In Advances in Artificial Intelligence, pages 275–281. Springer, 2014.

[22] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. In ICWSM, 2013.

[23] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. ACM Trans. Interact. Intell. Syst., 2(3):18:1–18:30, September 2012.

[24] 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.

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

[26] Quan Fang, Jitao Sang, Changsheng Xu, and Yong Rui. Topic-sensitive influencer mining in interest-based social media networks via hypergraph learning. Multimedia, IEEE Transactions on, 16(3):796–812, 2014.

[27] A. Forte, M. Dickard, R. Magee, and D. E. Agosto. What do teens ask their online social networks?: Social search practices among high school students. In Proc. of the 17th ACM Conf. on Computer Supported Cooperative Work & Social Computing (CSCW'14), pages 28–37, 2014.

[28] Gefi. The open graph viz platform.

[29] Malcolm Gladwell. The tipping point: How little things can make a big difference. Little, Brown, 2006.

[30] Jacob Goldenberg, Sangman Han, Donald R Lehmann, and Jae Weon Hong. The role of hubs in the adoption process. Journal of Marketing, 73(2):1–13, 2009.

[31] R. Goldman. Teens indicted after allegedly taunting girl who hanged herself, bbc news, 2010. [accessed 14-January-2014].

[32] S. Kumar H. Sanchez. Twitter bullying detection. NSDI'12, pages 15–15, Berkeley, CA, USA, 2012. USENIX Association.

[33] S. Hinduja. Cyberbullying on Instagram. 2013.

[34] S. Hinduja and J. W. Patchin. Bullying Beyond the Schoolyard: Preventing and Responding to Cyberbullying. Thousand Oaks, CA: Corwin Press, 2009.

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

[36] C. M. Homan, N. Lu, X. Tu, M. C. Lytle, and V.M.B. Silenzio. Social structure and depression in TrevorSpace. In Proc. of the 17th ACM Conf. on Computer Supported Cooperative Work; Social Computing, CSCW '14, pages 615–625, 2014.

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

[38] Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. Detection of cyberbullying incidents on the instagram social network. CoRR, abs/1503.03909, 2015.

[39] Homa Hosseinmardi, Rahat Ibn Rafiq, Shaosong Li, Zhili Yang, Richard Han, Shivakant Mishra, and Qin Lv. Comparison of common users across Instagram and Ask.fm to better understand cyberbullying. In The 7th IEEE international Conference on Social Computing and Networking (SocialCom), 2014.

[40] Dongxu Huang, Jing Zhou, Dejun Mu, and Feisheng Yang. Retweet behavior prediction in twitter. In Computational Intelligence and Design (ISCID), 2014 Seventh International Symposium on, volume 2, pages 30–33. IEEE, 2014.

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

[42] 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.

[43] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we twitter: Understanding microblogging usage and communities. In Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, pages 56–65. ACM, 2007.

[44] A. Kontostathis K. Reynolds and L. Edwards. Using machine learning to detect cyberbullying. Machine Learning and Applications, Fourth International Conference on, 2:241–244, 2011.

[45] Edward Keller and Jon Berry. One american in ten tells the other nine how to vote, where to eat and, what to buy. they are the influentials, 2003.

[46] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 137–146. ACM, 2003.

[47] 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.

[48] April Kontostathis, Andy Garron, Kelly Reynolds, Will West, and Lynne Edwards. Identifying predators using chatcoder 2.0. In CLEF (Online Working Notes/Labs/Workshop), 2012.

[49] 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.

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

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

[52] Balachander Krishnamurthy, Phillipa Gill, and Martin Arlitt. A few chirps about Twitter. In Proceedings of the first workshop on Online social networks, pages 19–24. ACM, 2008.

[53] Andrey Kupavskii, Alexey Umnov, Gleb Gusev, and Pavel Serdyukov. Predicting the audience size of a tweet. In ICWSM, 2013.

[54] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In Proceedings of the 19th International Conference on World Wide Web, WWW '10, pages 591–600, New York, NY, USA, 2010. ACM.

[55] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In Proceedings of the 19th international conference on World wide web, pages 591–600. ACM, 2010.

[56] Kyumin Lee, Jalal Mahmud, Jilin Chen, Michelle Zhou, and Jeffrey Nichols. Who will retweet this?: Automatically identifying and engaging strangers on twitter to spread information. In Proceedings of the 19th international conference on Intelligent User Interfaces, pages 247–256. ACM, 2014.

[57] Jure Leskovec, Mary McGlohon, Christos Faloutsos, Natalie S Glance, and Matthew Hurst. Patterns of cascading behavior in large blog graphs. In SDM, volume 7, pages 551–556. SIAM, 2007.

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

[59] Jalal Mahmud. Why do you write this? prediction of influencers from word use. In Eighth International AAAI Conference on Weblogs and Social Media, 2014.

[60] J. McAuley and J. Leskovec. Stanford network analysis project. http://snap.stanford.edu/data/egonets-Facebook.html, 2012. [accessed 14-January-2014].

[61] F. Mishna, C. Cook, J. Daciuk T. Gadalla, and S. Solomon. Cyberbullying behaviors among middle and high school students. Jul2010, Vol. 80 Issue 3, p362-374.

[62] Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and analysis of online social networks. In Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement, IMC '07, pages 29–42, New York, NY, USA, 2007. ACM.

[63] 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.

[64] Michael J Moore, Tadashi Nakano, Akihiro Enomoto, and Tatsuya Suda. Anonymity and roles associated with aggressive posts in an online forum. Computers in Human Behavior, 28(3):861–867, 2012.

[65] Mohamed Morchid, Richard Dufour, Pierre-Michel Bousquet, Georges Linarès, and Juan-Manuel Torres-Moreno. Feature selection using principal component analysis for massive retweet detection. Pattern Recognition Letters, 49:33–39, 2014.

[66] Vinita Nahar, Xue Li, and Chaoyi Pang. An effective approach for cyberbullying detection. In Communications in Information Science and Management Engineering, CISME'13, 2013.

[67] 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.

[68] 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.

[69] M. E. J. Newman. Networks, an introduction. published by Oxford.

[70] Nikon Candelle. Why Vine videos are becomming more popular than Instagram videos? `http://www.rantic.com/articles/vine-videos-becoming-popular-instagram-videos/`, 31 Jan 2015. [Online; accessed 16-Aug-2015].

[71] Taisei Nitta, Fumito Masui, Michal Ptaszynski, Yasutomo Kimura, Rafal Rzepka, and Kenji Araki. Cyberbullying detection based on category relevance maximization. 2013.

[72] D. Olweus. Peer harassment: A critical analysis and some important issues. In J. Juvonen and s. Graham (Eds.), Peer harassment in school. New York: Guilford., 2001a.

[73] D. Olweus. Olweus core program against bullying and antisocial behavior: A teacher handbook. Bergen, Norway: Research Center for Health Promotion, University of Bergen. 2001b.

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

[75] Dan Olweus. School bullying: Development and some important challenges. Annual review of clinical psychology, 9:751–780, 2013.

[76] Eli Omernick and Sara Owsley Sood. The impact of anonymity in online communities. In Social Computing (SocialCom), 2013 International Conference on, pages 526–535. IEEE, 2013.

[77] R. Ottoni, D. Las Casas, J. P. Pesce, W. Meira Jr., C. Wilson, A. Mislove, and V. Almeida. Of pins and tweets investigating how users behave across image- and text-based social networks. In Proc. of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM'14), 2014.

[78] R. Kowalski P. W. Agatston and S. Limber. Students' perspectives on cyber bullying. Journal of Adolescent Health, 2007.

[79] Justin W Patchin and Sameer Hinduja. An update and synthesis of the research. Cyberbullying prevention and response: Expert perspectives, page 13, 2012.

[80] Galán-García Patxi, José Gaviria de la Puerta, Carlos Laorden Gómez, Igor Santos, and Pablo García Bringas. Supervised machine learning for the detection of troll profiles in Twitter social network: Application to a real case of cyberbullying. In International Joint Conference SOCO13-CISIS13-ICEUTE13, pages 419–428. Springer, 2014.

[81] J. M. Pennebaker, M. E. Francis, and R. J. Booth. Linguistic Inquiry and Word Count. Lawerence Erlbaum Associates, Mahwah, NJ, 2001.

[82] Sasa Petrovic, Miles Osborne, and Victor Lavrenko. Rt to win! predicting message propagation in twitter. In ICWSM, 2011.

[83] Fabio Pezzoni, Jisun An, Andrea Passarella, Jon Crowcroft, and Marco Conti. Why do i retweet it? an information propagation model for microblogs. In Social Informatics, pages 360–369. Springer, 2013.

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

[85] Nektaria Potha and Manolis Maragoudakis. Time series forecasting in cyberbullying data. In Engineering Applications of Neural Networks, pages 289–303. Springer, 2015.

[86] Michal Ptaszynski, Pawel Dybala, Tatsuaki Matsuba, Fumito Masui, Rafal Rzepka, Kenji Araki, and Yoshio Momouchi. In the service of online order tackling cyberbullying with machine learning and affect analysis. 2010.

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

[88] Rahat Ibn Rafiq, Homa Hosseinmardi, Sabrina Mattson, Richard Han, Qin Lv, and Shivakant Mishra. 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 (ASONAM 2015), 2015.

[89] D. Schon. The reflective practitionor. Basic Book:New York, January 1, 1959.

[90] A. Schrock and D. Boyd. Problematic youth interaction online: Solicitation, harassment, and cyberbullying. Computer-Mediated Communication in Personal Relationships, 2011.

[91] 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 DCOSS, pages 123–132. IEEE, 2013.

[92] P. K. Smith, C. del Barrio, and R. Tokunaga. In book: Principles of Cyberbullying Research. Definitions, measures and methodology, Chapter: Definitions of Bullying and Cyberbullying: How Useful Are the Terms? Routledge, 2012.

[93] L. Smith-Spark. Hanna smith suicide fuels calls for action on ask.fm cyberbullying, cnn, 2013. [accessed 14-January-2014].

[94] Mona E Solberg, Dan Olweus, and Inger M Endresen. Bullies and victims at school: Are they the same pupils? British Journal of Educational Psychology, 77(2):441–464, 2007.

[95] Camila Souza Araujo, Damilton Correa, Luiz Paulo, Ana Paula Couto da Silva, Raquel Oliveira Prates, and Wagner Meira. It is not just a picture: Revealing some user practices in instagram. In Web Congress (LA-WEB), 2014 9th Latin American, pages 19–23. IEEE, 2014.

[96] Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H Chi. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In Social computing (socialcom), 2010 IEEE second international conference on, pages 177–184. IEEE, 2010.

[97] Jie Tang, Jimeng Sun, Chi Wang, and Zi Yang. Social influence analysis in large-scale networks. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 807–816. ACM, 2009.

[98] Jie Tang, Sen Wu, and Jimeng Sun. Confluence: Conformity influence in large social networks. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 347–355. ACM, 2013.

[99] Io Taxidou and Peter M Fischer. Online analysis of information diffusion in Twitter. In Proceedings of the companion publication of the 23rd international conference on World wide web companion, pages 1313–1318. International World Wide Web Conferences Steering Committee, 2014.

[100] Ditch the Label Anti Bullying Charity. The annual cyberbullying survey 2013. 2013.

[101] Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Véronique Hoste. Automatic detection and prevention of cyberbullying. In Proceedings of the First International Conference on Human and Social Analytics (HUSO 2015), St. Julians, Malta, 2015/10/11 2015.

[102] Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Veronique Hoste. Detection and fine-grained classification of cyberbullying events. In Galia Angelova, Kalina Bontcheva, and Ruslan Mitkov, editors, Proceedings of Recent Advances in Natural Language Processing, Proceedings, pages 672–680, 2015.

[103] Cynthia Van Hee, Ben Verhoeven, Els Lefever, Guy De Pauw, Véronique Hoste, and Walter Daelemans. Guidelines for the fine-grained analysis of cyberbullying. 05/2015 2015.

[104] Kathleen Van Royen, Karolien Poels, Walter Daelemans, and Heidi Vandebosch. Automatic monitoring of cyberbullying on social networking sites: From technological feasibility to desirability. Telematics and Informatics, 32(1):89–97, 2015.

[105] H. Vandebosch and K. V. Cleemput. Defining cyberbullying: A qualitative research into the perceptions of youngsters. CyberPsychology and Behavior, 2008.

[106] Heidi Vandebosch and Katrien Van Cleemput. Defining cyberbullying: A qualitative research into the perceptions of youngsters. CyberPsychology & Behavior, 11(4):499–503, 2008.

[107] A. Weilenmann, T. Hillman, and B. Jungselius. Instagram at the museum: Communicating the museum experience through social photo sharing. In Proc. of the SIGCHI Conf. on Human Factors in Computing Systems, CHI '13, pages 1843–1852, 2013.

[108] Gabriel Weimann. The influentials: People who influence people. SUNY Press, 1994.

[109] Christo Wilson, Alessandra Sala, Krishna P. N. Puttaswamy, and Ben Y. Zhao. Beyond social graphs: User interactions in online social networks and their implications. ACM Trans. Web, 6(4):17:1–17:31, November 2012.

[110] Negative words list. Negative words list form, luis von ahn's research group, 2014.

[111] Jun-Ming Xu, Hsun-Chih Huang, Amy Bellmore, and Xiaojin Zhu. School bullying in twitter and weibo: a comparative study. Reporter, 7(16):10–14, 2014.

[112] Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. Learning from bullying traces in social media. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 656–666. Association for Computational Linguistics, 2012.

[113] S. Kambhampati. Y. Hu, L. Manikonda. What we instagram: A first analysis of instagram photo content and user types. In Proc. of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM'14), 2014.

[114] Haihao Yu, Xu Feng Bai, ChengZhe Huang, and Haoliang Qi. Prediction of users retweet times in social network. International Journal of Multimedia and Ubiquitous Engineering, 10(5):315–322, 2015.

[115] Jing Zhang, Biao Liu, Jie Tang, Ting Chen, and Juanzi Li. Social influence locality for modeling retweeting behaviors. In Proceedings of the Twenty-Third international joint conference on Artificial Intelligence, pages 2761–2767. AAAI Press, 2013.

[116] Kaiping Zhang and René F Kizilcec. Anonymity in social media: Effects of content controversiality and social endorsement on sharing behavior. 2014.

[117] Huidong Zhao, Gang Liu, Chuan Shi, and Bin Wu. A retweet number prediction model based on followers' retweet intention and influence. In Data Mining Workshop (ICDMW), 2014 IEEE International Conference on, pages 952–959. IEEE, 2014.

[118] Adam G Zimmerman. Online aggression: The influences of anonymity and social modeling. 2012.