Weakly Supervised Machine Learning for Cyberbullying Detection

Elaheh Raisi

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Computer Science and Applications

Bert Huang, Chair
Naren Ramakrishnan
Madhav V. Marathe
Gang Wang
Richard Y. Han

March 28, 2019
Blacksburg, Virginia

Keywords: Machine Learning, Weak Supervision, Cyberbullying detection, Social Media,
Co-trained Ensemble

Weakly Supervised Machine Learning for Cyberbullying Detection

Elaheh Raisi

(ABSTRACT)

The advent of social media has revolutionized human communication, significantly improving individuals lives. It makes people closer to each other, provides access to enormous real-time information, and eases marketing and business. Despite its uncountable benefits, however, we must consider some of its negative implications such as online harassment and cyberbullying. Cyberbullying is becoming a serious, large-scale problem damaging people's online lives. This phenomenon is creating a need for automated, data-driven techniques for analyzing and detecting such behaviors. In this research, we aim to address the computational challenges associated with harassment-based cyberbullying detection in social media by developing machine-learning framework that only requires weak supervision. We propose a general framework that trains an ensemble of two learners in which each learner looks at the problem from a different perspective. One learner identifies bullying incidents by examining the language content in the message; another learner considers the social structure to discover bullying.

Each learner is using different body of information, and the individual learner co-train one another to come to an agreement about the bullying concept. The models estimate whether each social interaction is bullying by optimizing an objective function that maximizes the consistency between these detectors. We first developed a model we referred to as participant-vocabulary consistency, which is an ensemble of two linear language-based and user-based models. The model is trained by providing a set of seed key-phrases that are indicative of bullying language. The results were promising, demonstrating its effectiveness and usefulness in recovering known bullying words, recognizing new bullying words, and discovering users involved in cyberbullying. We have extended this co-trained ensemble approach with two complementary goals: (1) using nonlinear embeddings as model families, (2) building a fair language-based detector. For the first goal, we incorporated the efficacy of distributed representations of words and nodes such as deep, nonlinear models. We represent words and users as low-dimensional vectors of real numbers as the input to language-based and user-based classifiers, respectively. The models are trained by optimizing an objective function that balances a co-training loss with a weak-supervision loss. Our experiments on Twitter, Ask.fm, and Instagram data show that deep ensembles outperform non-deep methods for weakly supervised harassment detection. For the second goal, we geared this research toward a very important topic in any online automated harassment detection: fairness against particular targeted groups including race, gender, religion, and sexual orientations. Our goal is to decrease the sensitivity of models to language describing particular social groups. We encourage the learning algorithm to avoid discrimination in the predictions by adding an unfairness penalty term to the objective function. We quantitatively and qualitatively evaluate the effectiveness of our proposed general framework on synthetic data and data from Twitter using post-hoc, crowdsourced annotation. In summary, this dissertation introduces a weakly supervised machine learning framework for harassment-based cyberbullying detection using both messages and user roles in social media.

Weakly Supervised Machine Learning for Cyberbullying Detection

Elaheh Raisi

(GENERAL AUDIENCE ABSTRACT)

Social media has become an inevitable part of individuals social and business lives. Its benefits, however, come with various negative consequences such as online harassment, cyberbullying, hate speech, and online trolling especially among the younger population. According to the American Academy of Child and Adolescent Psychiatry,[1] victims of bullying can suffer interference to social and emotional development and even be drawn to extreme behavior such as attempted suicide. Any widespread bullying enabled by technology represents a serious social health threat.

In this research, we develop automated, data-driven methods for harassment-based cyberbullying detection. The availability of tools such as these can enable technologies that reduce the harm and toxicity created by these detrimental behaviors. Our general framework is based on consistency of two detectors that co-train one another. One learner identifies bullying incidents by examining the language content in the message; another learner considers social structure to discover bullying. When designing the general framework, we address three tasks: First, we use machine learning with weak supervision, which significantly alleviates the need for human experts to perform tedious data annotation. Second, we incorporate the efficacy of distributed representations of words and nodes such as deep, nonlinear models in the framework to improve the predictive power of models. Finally, we decrease the sensitivity of the framework to language describing particular social groups including race, gender, religion, and sexual orientation.

This research represents important steps toward improving technological capability for automatic cyberbullying detection.

[1] http://www.aacap.org/

Dedication

To the most important people in my life: my family

Acknowledgments

I would like to express my deepest gratitude to my advisor, Professor Bert Huang, for his unwavering support and invaluable advice from when I joined his lab, through to completion of my Ph.D. degree. Without his guidance and persistent help this dissertation would not have been possible. I truly feel fortunate to have had the opportunity to work with him during past five years. He had been not only an amazing scientist, but also a wonderful person. He has shown me, by his example, what an outstanding scientist and tremendous mentor should be: patient, smart, persistent, passionate, supportive, kind, positive, humble, and flexible. My journey through graduate school was incredible, and I owe a huge part of it to him.

I very much appreciate the rest of my dissertation committee members, Professor Naren Ramakrishnan, Professor Madhav Marathe, Professor Gang Wang, and Professor Richard Han for accepting to serve on my committee and generously offering their time, support, and guidance. Their helpful comments and suggestions improved the presentation and contents of this dissertation.

Many thanks to all my awesome Hokie friends I have come to know in Virginia Tech for always being there for me. I could not imagine how boring and empty my life would be without them in Blacksburg. Their friendship means a lot to me.

I would also like to thank my lab mates for their company, support, and sharing their learning experiences. We had so many great and brilliant times together.

Last but not least, I would like to thank my parents, two sisters, grandmother, and niece (who joined our family last year) for their endless love, unbelievable support, and encouragement. I feel incredibly lucky to be surrounded by an exceptional family.

Contents

1 Introduction 1

1.1 What is Bullying? 1

1.2 Social Media 3

1.3 What is Cyberbullying 4

1.3.1 Cyberbullying Impacts 7

1.4 Cyberbullying Awareness and Prevention 8

1.5 Automated Methods for Cyberbullying Detection 9

1.6 Bias in Social Media 10

1.6.1 Bias in Machine Learning Models 11

1.7 Fairness in Machine Learning 14

1.8 Bias in Online Harassment Systems 15

1.9 Proposed Framework 16

1.10 Main Contributions 17

1.11 Overview and outline 18

2 Literature Review 20

2.1 Online Harassment and Cyberbullying Detection 20

2.2 Weakly Supervised Machine Learning and Vocabulary Discovery 23

2.3 Fairness in Machine Learning 24

3 Weakly Supervised Cyberbullying Detection with Participant-Vocabulary Consistency 28

3.1 Introduction 28

3.2 General Framework 30

3.2.1 Alternating Least Squares 32

3.3 Experiments . . . 33

3.3.1 Data Processing . . . 33

3.3.2 Baselines . . . 34

3.3.3 Human Annotation Comparisons . . . 35

3.3.4 Qualitative Analysis . . . 41

3.3.5 Bully and Victim Score Analysis . . . 43

3.3.6 Bully and Victim Network Analysis . . . 44

3.3.7 Fairness Analysis . . . 50

3.4 Conclusion . . . 53

4 Co-trained Ensembles of Embedding Models 55

4.1 Introduction . . . 55

4.2 Co-trained Ensembles . . . 56

4.2.1 General Framework . . . 57

4.2.2 Models . . . 58

4.3 Experiments . . . 59

4.3.1 Precision Analysis . . . 59

4.3.2 Qualitative Analysis . . . 64

4.4 Conclusion . . . 67

5 Reduced-Bias Co-Trained Ensembles for Weakly Supervised Cyberbullying Detection 68

5.1 Introduction . . . 68

5.2 Reduced-Bias Co-Trained Ensembles . . . 70

5.3 Experiments . . . 70

5.4 Conclusion . . . 76

6 Complementary Experiments 78

6.1 Inter-Annotator Agreement . . . 78

6.2 Explainable Node Representations . . . 79

6.3 Weak Supervision Against Full Supervision . . . 79

7 Summary and Outlook 81

7.1 Significance and Contribution . . . 81

7.2 Future prospects 82
7.3 What to Do When Cyberbullying is Detected 83
7.4 Practical Advice 84
7.5 Other Applications 84
7.6 Conclusion 85
Bibliography 86

List of Figures

1.1 Rate of students between the age of 12 and 18 were bullied at school in 2015. This figure is taken from the report by U.S. Department of Education and Department of Justice.[2] . . . 2

1.2 Tremendous rise in social media usage. Figure is reproduced from a report by Pew Research Center.[3] 69% of adults in the U.S. adults use at least one social media site in January 2018, while only 43% used at least one social media site in January 2010. . . . 4

1.3 The rates of cyberbullying victimization have varied between 2007 and 2016. On average, about 28% of the students who have been surveyed have been the victim of cyberbullying at least once. This figure is reproduced from [183]. . 6

1.4 According to the Pew Research Center, 67% and 33% of younger adults and adults above 30 faced any form of harassment. This figure is taken from [78]. 7

3.1 Schematic of the weak supervision paradigm for training a cyberbullying detector. 29

3.2 Precision@k for bullying interactions on Ask.fm (top), Instagram (middle), and Twitter (bottom). . . . 37

3.3 Precision@k for bullying phrases on Ask.fm (top), Instagram (middle), and Twitter (bottom). . . . 40

3.4 Two examples of true positives by PVC that other baselines were not be able to detect. These examples are clearly intense and toxic, but their concentration of obvious swear words may not be high enough for the baseline approaches to identify. . . . 41

3.5 Examples of harassment detected by PVC and verified by annotators. These examples do not have very obvious offensive-language usage, so methods beyond simple query-matching may be necessary to find them. . . . 42

3.6 Examples of false positives by PVC: interactions identified by PVC that annotators considered non-harassment and appear correctly labeled. These examples include usage of offensive language but may require sophisticated natural language processing to differentiate from harassing usage. . . . 42

3.7 Two examples of non-bullying conversations mistakenly flagged by PVC containing the keyword "black". . . . 44

3.8 Conversation of users with high bully score (top) and high victim score (bottom). The user with a high bully score is sending toxic messages to other users; the user with a high victim score is receiving intense negative messages while responding to some messages reacting the bullies. 45

3.9 Scatter (top) and heatmap (bottom) plots for Twitter. The plots show the distributions of the bully and victim scores. According to the heatmap, most of the users in our Twitter data are not involved in bullying because most of the users occupy the redder region. However, there are moderate number of bullies, victims, and victimized bullies. 46

3.10 Heatmap plots for Instagram (top) and Ask.fm (bottom). In both datasets, users with low bully and low victim scores fall in the denser region (red color), but there are users with high bully and high victim scores to a lesser extent. 47

3.11 An example of a user with high victim score and low bully score. The user receives messages containing offensive words, but they are not used in an aggressive or harassing manner, and the user does not appear to be a victim. 48

3.12 A sub-graph of bullies and victims in Instagram. *Red*, *green*, and *gray* represent bullies, victims, and bystanders, respectively. Victims have several bully and bystander neighbors. In addition, a bully may interact with different people, but they may not bully all of them. 49

3.13 Two sub-graphs of bullies and victims in Ask.fm. *Red*, *green*, and *gray* represent bullies, victims, and bystanders, respectively. Top: a bully at the center has bully, victim, and neutral neighbors. Bottom: a bully has two victim neighbors; one of them has a large community of friends, while the other is isolated. . . 51

3.14 Two examples of bias toward race and sexual orientation in Instagram. In top example, bully is harassing victim using racist slur ("spic"). In the bottom example, the user is bullying the other one using negative words as well as sensitive keywords about race and sexual orientation ("black" and "gay"). . 52

3.15 Sensitive keywords ranks obtained by PVC for Twitter (top), Instagram (middle), and Ask.fm (bottom). The highest word rank in Twitter, Instagram, and Ask.fm is 870 for the word "boy", 1,973, and 677, respectively for the word "queer". 54

4.1 Diagram of the co-trained ensemble of the RNN message learner and the node2vec user learner. 56

4.2 Diagram of the components in the co-trained ensemble of the doc2vec message learner and the node2vec user learner. 57

4.3 Precision@k for bullying interactions on Ask.fm data using the combination of message and user learners, PVC, seed-based, and naive-participant. 60

4.4 Precision@k for bullying interactions on Twitter data using the combination of message and user learners, PVC, seed-based, and naive-participant. The legend in the top left plot applies for all other plots in this figure. 61

4.5 Precision@k for bullying interactions on Instagram data using the combination of message and user learners, PVC, seed-based, and naive-participant. 62

4.6 The difference between the precision@100 score of the co-trained ensemble and the message learner trained alone. The improvements produced by co-training are more pronounced for the doc2vec language model than others, but the co-training also improves the performance of the other models as well. 65

5.1 ROC curve of the four co-trained ensemble methods stratified by fairness type. The performance of substitutional bow_none, doc2vec_node2vec, and emb_none improves over the vanilla method. Removal emb_none also improves over vanilla, but emb_node2vec has worse performance. 73

5.2 ROC curves for the emb_none model on data containing sensitive keywords in the sexual orientation and religion categories. The gap between AUC of sensitive keywords in these two groups reduces, without degradation in performance. 74

5.3 Three examples of conversations given a high score by the vanilla model and given a low score by a reduced-bias model. These conversations discuss sensitive topics but may not represent one user bullying another. 75

5.4 Bullying scores of sensitive keywords as scored by emb_none and emb_node2vec with and without fairness terms. The bullying score of most sensitive keywords reduce when fairness is imposed on the model, and the scores become more uniform than without any fairness terms (vanilla). 76

6.1 Comparing precision@k doc2vec message learner using different user learners. The red color shows the explainable *degree-based* node representation for user classifier. Degree-based user learner decreases the precision of "doc2vec" message learner. 79

6.2 ROC curve of weakly supervised PVC model and fully supervised model. The area under the ROC curve of PVC is lower than both logistic regression and multinomial naive Bayes. 80

List of Tables

3.1 Color-coded bullying bigrams detected in Ask.fm data by PVC and baselines. Terms are categorized according to the aggregate score of annotations. "Bullying" (2 or greater), "Likely Bullying" (1), "Uncertain" (0), and "Not Bullying" (negative) bigrams are shown in red, orange, gray, and blue, respectively. . . 36

3.2 Color-coded bullying bigrams detected in Instagram data by PVC and baselines 38

3.3 Color-coded bullying bigrams detected in Twitter data by PVC and baselines 39

3.4 The average in-degree and out-degree of top-scoring bullies and victims. On average, bullies are sending more messages than victims. They also send more messages than they receive, unlike victims who are more receivers of the messages than senders. 48

4.1 Precision@100 across all methods. The methods with the highest precision@100 for each dataset are bolded. The methods that are among the top four for all three datasets are highlighted. The doc2vec_node2vec and emb_node2vec combinations are among the top methods for all datasets. 64

5.1 False positive rate of models on non-bullying synthetic benchmark statements (using threshold 0.5). Both *removal* and *substitutional* fair models reduce the false positive rate compare to without bias reduction (vanilla). 71

5.2 The category dispersion of four co-trained ensemble methods for three targeted groups. A bold value means the category dispersion of the reduced-bias model is lower than the default method (vanilla). Substitutional emb_none and doc2vec_node2vec have better category dispersion than the vanilla method in all three groups. 73

6.1 Fleiss kappa inter-annotator agreement measure for each dataset and all of them together. 79

Chapter 1

Introduction

In this chapter, we discuss various definitions of bullying and cyberbullying and their adverse consequences on society; we then explain existing prevention efforts and methods and the importance of automatic methods to detect this phenomenon. In addition, we describe the bias and prejudice in social media data, how this bias is transmitted to machine learning models, and causes discriminative decision against some particular social groups. Finally, we briefly describe our proposed framework for cyberbullying detection and how we reduce the bias in our model's prediction.

1.1 What is Bullying?

Bullying is a form of mean behavior in which someone with high social or physical power repeatedly abuse, intimidate, or harm particular target (usually a less powerful person). Some researchers have characterized bullying by three main features: i. hostile intention (a bully has aggressive behavior toward victim), ii. repetition (it happens repeatedly over time), iii. power imbalance (a bully is physically stronger, could reveal humiliating information, or is known to hurt others) [116, 122, 161, 181, 224, 225].

There are many types of bullying. *Physical bullying* occurs with hitting, fighting, yelling, spitting, tripping, pushing, kicking, pinching, and shoving; *Verbal bullying* includes name calling, threatening, taunting, etc. *Social bullying* are behaviors to embarrass someone in public. For example, rumor spreading, leaving someone out on purpose, laughing or making fun of someone. If bullying is done by a group, it is called mobbing [156, 257]. Many bullying cases take place because of the differences in social class such as race, religion, sexual orientation, appearance, etc.

According to the National Center for Educational Statistics (NCES) in 2015, roughly 20% of teenagers between 12 and 18 in the U.S. are being bullied at school [143] despite anti-bullying legislation in all 50 states.[1] In another line of research jointly done by the Bureau of Justice Statistics and National Center for Education Statistic published in 2018, during the 2015-16

[1] https://www.huffingtonpost.com/deborah-temkin/all-50-states-now-have-a_b_7153114.html

school year, 22% of public middle schools declared student bullying happened at least once a week. This percentage for high schools is 15%.[2] Figure 1.1 shows more specific information regarding the type of bullying categorized by gender.

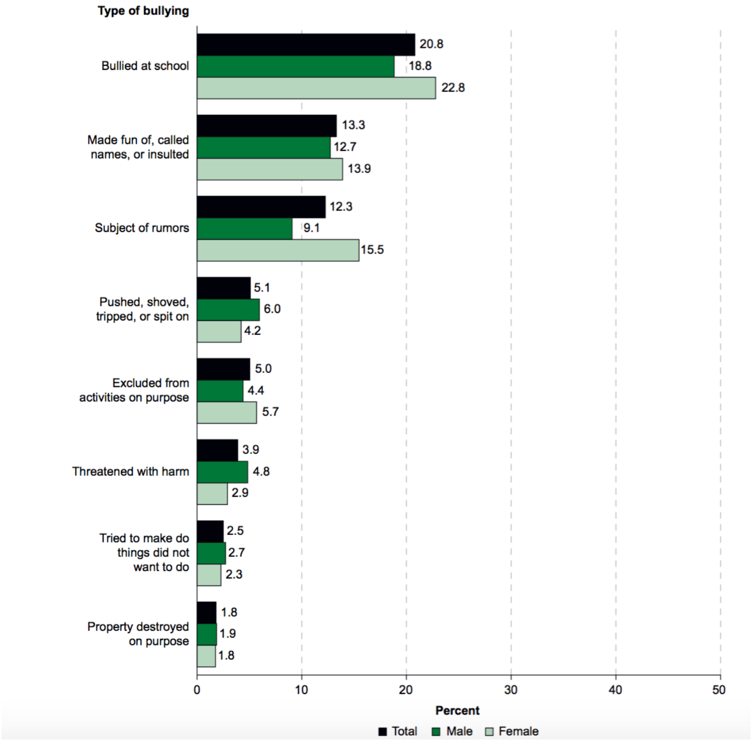


Figure 1.1: Rate of students between the age of 12 and 18 were bullied at school in 2015. This figure is taken from the report by U.S. Department of Education and Department of Justice.[2]

Bullying is linked to serious short and long term health issues including but not limited to physical injury, social and mental health problems, substance use, low self-esteem, and depression [123, 134, 215, 216]. Due to the widespread harm associated with bullying, many local and national policymakers have been taken steps to resolve the problem. In 2011, president Barack Obama, the Department of Health and Human Services, together with the Department of Education called for the unified effort to combat bullying in a Bullying Prevention Conference at the White House [239]. First Lady Melania Trump has also engaged in an anti-bullying effort [188].

The Committee on Injury, Violence, and Poison Prevention released a refined policy statement highlighting bullying and dating violence, and emphasized the importance of this issue in the strategic agenda of the American Academy of Pediatrics [132].

[2] https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2018036

1.2 Social Media

With the rapid growth of information and communication technology, numerous new methods of electronic communication, such as social networking, have appeared. Using modern technologies, time and distance barriers in communication have been eliminated, leading to extensive spread of advanced communication tech tools in everyday life. Social media is one of such tools that have dramatically changed the way groups of people interact and communicate. Kaplan et al. [128] define social media as "a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content." Social media is a great platform for sharing information and ideas, collaboration and extending communication with others. Therefore, one can access an enormous amount of information and real-time news instantly. People are relatively free to post links, sharing pictures and videos, and almost anything in online communities. It makes it possible to communicate with anyone and at any time. Using social media, individuals are able to stay in touch with their friends, meet new people, join groups, and gain exposure to a diverse world.

Social media has become an integral part of individuals' daily lives. Its usage is drastically increasing all around the world. According to the latest research by Brandwatch Analytics, a social media monitoring company [222], as of January 2019, there are 4.2 billion users in the Internet, and 3.397 billion active social media users. Social media users grew by 320 million between September 2017 and October 2018. An example of this tremendous rise is shown in Figure 1.2 based on the statistics collected by the Pew Research Center.[3]

Social media also can be a powerful business tool. Business owners and organizations can connect with their customers, sell their products, marketing a new product or service, make their brand visible, and increase their brand reputation with minimal cost. A study shows that 81% of all small and medium businesses use some kind of social platform [222]. 78% of people who complain to a brand via Twitter expect a response within an hour. There are 60 million active business pages on Facebook.

Along with its tremendous benefits, however, social media could be used as a platform for cybercrimes and inappropriate online activities such as hacking, fraud and scams, spreading false information, trolling, online harassment, and cyberbullying.

In this research, we focus on cyberbullying, which is on the rise worldwide. According to a recent survey by the Pew Research Center report in 2018 [9], 59% of U.S. teens have been bullied or harassed online, which is roughly six-in-ten in the U.S. Their report in 2017 [78] states that around four-in-ten (41%) of Americans have been personally subjected to harassing behavior online, and 66% has witnessed these behaviors directed at others. 88% of 18- to 29-year-olds indicate that they use any form of social media; and 78% among those ages 30 to 49, and 64% among those ages 50 to 64.

These statistics as a whole give an idea about the popularity and growth of social media platforms and their impacts. Because of the accessibility and freedom that these social media sites provide, these platforms are increasingly used for cyberbullying activities [38, 56].

[3] http://www.pewinternet.org/fact-sheet/social-media/

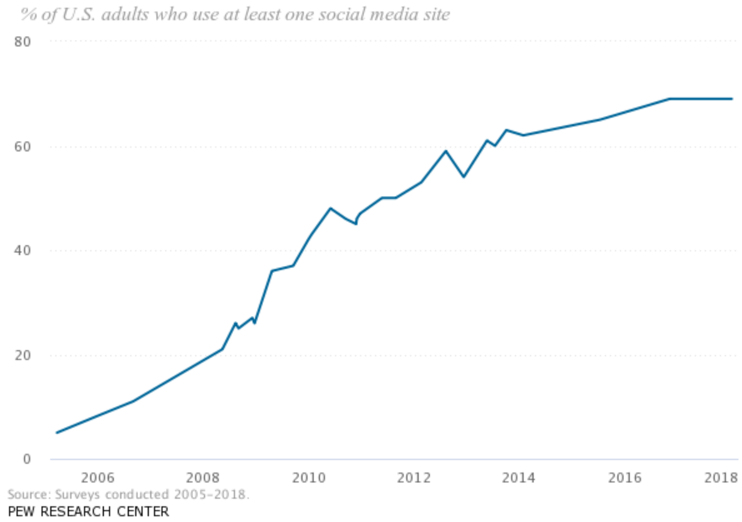


Figure 1.2: Tremendous rise in social media usage. Figure is reproduced from a report by Pew Research Center.[3] 69% of adults in the U.S. adults use at least one social media site in January 2018, while only 43% used at least one social media site in January 2010.

According to a survey by Ditch the Label in 2017, the top social media sites where the highest percentage of users experience cyberbullying are Facebook, Instagram, YouTube, Twitter, and Snapchat [238]. Ask.fm was among the top networking platforms with a very high percentage of cyberbullying incidents until 2014 [237]. In our experiments, we focus on sites with the highest concentration of cyberbullying: Twitter, Instagram, and Ask.fm.

1.3 What is Cyberbullying

As people communicate through social networks more and more, online harassment and cyberbullying are becoming a more serious problem. Dan Olweus, in his book entitled "Bullying at School: What We Know and What We Can Do," defined three criteria for bullying: 1) intention to harm, 2) repetition, and 3) power imbalance between the bully and the victim [181]. With the rise of online harassment and cyberbullying through social media sites, some researchers are seeking formal definitions for cyberbullying; the central question, however, has been whether the same criteria as offline bullying can be used [24, 135, 168, 182, 184, 226, 241, 261]. Applying these three factors for cyberbullying definition is controversial because it does not happen in physical spaces and in particular times. Danah Boyd, in her book "It's Complicated," described four characteristics of the web, which make them different from physical public spaces: persistence, visibility, spreadability, and searchability [33]. Online content is durable; once hurtful information posted online, it remains publicly online forever. The victim and many more people can see it over and over again. So, the victim can experience the bullying repeatedly. When bullying happens online, a wider range of audience can see it, or even spread it.

The other difference in cyberbullying is that bullies do not need to face someone and can hide their true identity. Victims can be easily harassed online because of constantly connected to the Internet. Online bullies do not interact face-to-face with others, and they even could be anonymous. Bullies do not need to be faster, stronger, taller, or older to confront the victim. Therefore physical *power imbalance* in bullying definition cannot hold for cyberbullying. But instead the anonymity, and the fact the bullying can happen anytime and anywhere, they online information are persistent, everyone can see and spread them, make victim powerless [76].

Repetition, the other bullying criterion can be interpreted in other way. Bullying posts could be liked, shared, and distributed by others easily. Therefore, repetition is actually the inherent characteristics of the Web. Additionally, in online conversations, the *intention* of users to bully others is not clear since people do not have face-to-face interaction, and they might get wrong impression [146, 220, 221, 248, 251].

With these considerations, a more realistic definition of cyberbullying is required. The National Crime Prevention Council in America defines cyberbullying "when the Internet, cell phones or other devices are used to send or post text or images intended to hurt or embarrass another person."[4] StopBullying.gov defines cyberbullying as "bullying that takes place using electronic technology[, including] devices and equipment such as cell phones, computers, and tablets as well as communication tools including social media sites, text messages, chat, and websites." Hee et al. [248] define cyberbullying as "content that is published online by an individual and that is aggressive or hurtful against a victim." These cyberbullying definitions are more consistent with the true nature of the Web, and the harms associated with this phenomenon without specifically citing the *power imbalance* and *repetition* as criterion [124, 240]. In this research, we follow these definitions.

In 2014, the National Crime Prevention Council reported approximately 50% of the youth in America are victimized by cyberbullying [20]. In 2014, more than 50% of teenagers were cyberbullying bystanders on social media [37].

Cyberbullying Research Center in 2016 reported around 34% of 12–17 year-old students were cyberbullied in their life; while, approximately 12% of students in this age gap revealed that they have been involved in cyberbullying in their life. Moreover, they show that the percentages of people who have experienced cyberbullying at least once during their life have increased from 18% to 34% from 2007 till 2016 [183] as illustrated in Figure 1.3. Additionally, a report by Pew research center in 2017 [78] stated 41% of Americans have personally experienced different forms of cyberbullying, and one-in-five (18%) victims expressed they have experienced a severe form of cyberbullying. As shown in Figure 1.4, 67% of adults between age 18 to 29, have experiences any form of cyberbullying; from less sever to more severe, and 33% of adults above 20, have encountered any form of harassment.

In a large-scale study on European kids, 20% of 11 to 16 year olds reported having experience online harassment; Additionally, the chance of being cyberbullied increased 12% from 2010 to 2013 [148].

The following are descriptions of various forms of cyberbullying [22, 83, 83, 107, 152, 217, 260]:

- Online harassment: using mean, negative and hurtful languages against someone

[4] http://www.ncpc.org/cyberbullying

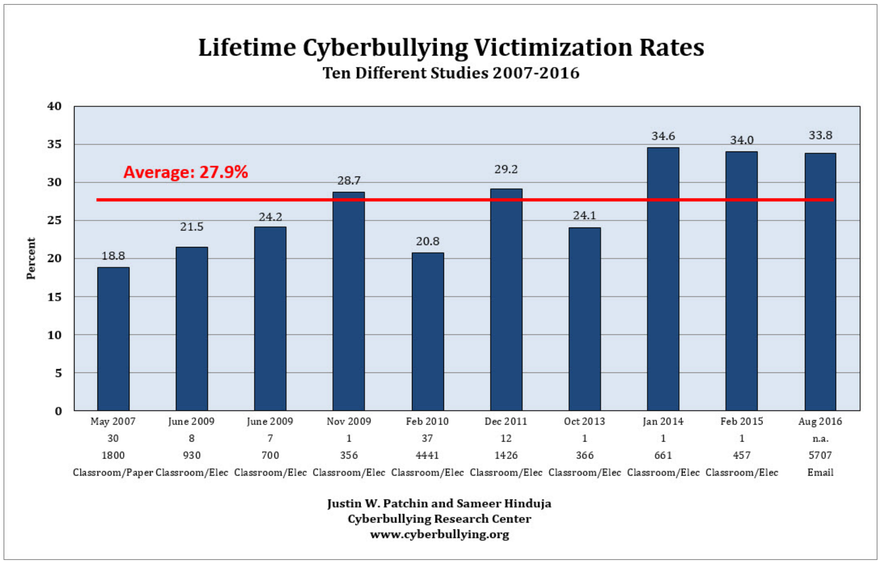


Figure 1.3: The rates of cyberbullying victimization have varied between 2007 and 2016. On average, about 28% of the students who have been surveyed have been the victim of cyberbullying at least once. This figure is reproduced from [183].

- Name-calling
- Rumor spreading: posting untrue or malicious rumors about someone to public
- Uploading obscene or embarrassing photos of someone
- Hacking, identity theft, or impersonation: logging into someone else's account or creating a fake account with another person's information and pictures, and sending out messages pretending to be that user.
- Threats and intimidation
- Outing: posting private, personal or embarrassing information, pictures, or videos in a public
- Trolling: provoking readers into an emotional response through the use of insults or inflammatory language in an online community
- Cyberstalking: stalking or harassing an individual using the Internet or other electronic devices. It includes monitoring, intimidation and threats of harm, and/or offensive comments that can extend to threats of physical harm to the targeted victim.
- Exclusion: leaving someone out of an online group deliberately

In the studies in this dissertation, we focus on harassment, in which harassers (bullies) send toxic and harmful communications to victims.

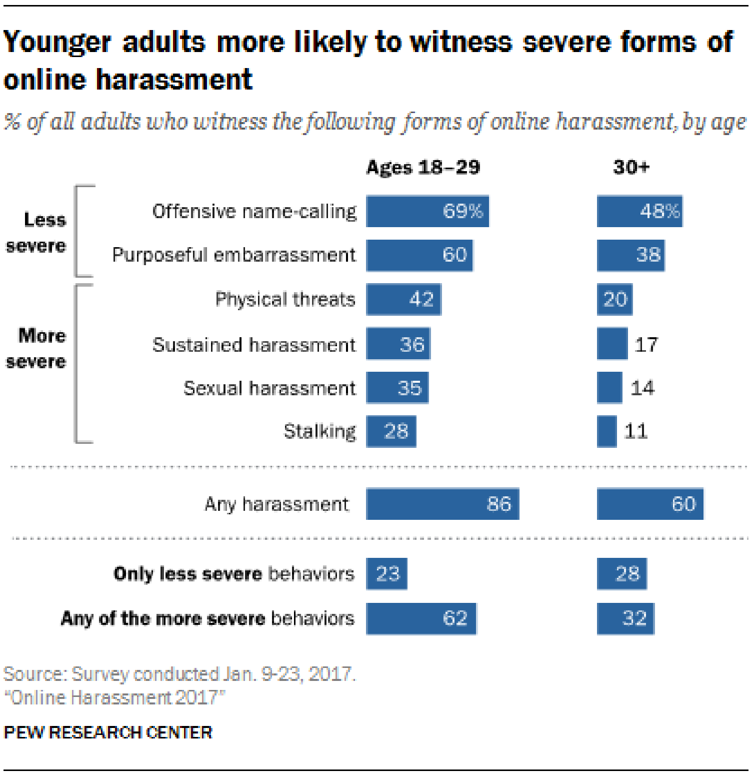


Figure 1.4: According to the Pew Research Center, 67% and 33% of younger adults and adults above 30 faced any form of harassment. This figure is taken from [78].

1.3.1 Cyberbullying Impacts

Cyberbullying has many negative impacts on victims especially on teenagers and young adults [252]. According to the American Academy of Child and Adolescent Psychiatry, targets of cyberbullying often suffer from mental and psychiatric disorders [159]. There is a strong connection between being involved in cyberbullying and experiencing psychiatric and psychosomatic disorders [229].

In online settings, various forms of power, such as anonymity, the constant possibility of threats, and the potential for a large audience, can create power imbalances in cyberbullying [76]. Therefore, the adverse effects of cyberbullying is deeper and longer-lasting compared to physical bullying [38,59]. Indeed, 71% of young generations are concerned about cyberbullying [201]. The consequences can be devastating: depression, low self-esteem, social exclusion and isolation, feelings of sadness, anger, fear, and frequent nightmares. Among teenagers, it could cause learning difficulties and poor academic results [58,60,87,89,106,137,208,212,226]. Other than the psychological risks, it could even influence physical well being; Self-injury, cutting, and even suicide are some harmful consequences of cyberbullying, making cyberbullying a serious social health threat [77,121,145,243]. Research says 24% of victims of cyberbullying had suicidal thoughts [238]. In 2017, the number of children taken to hospitals for attempted suicide or expressing suicidal thoughts became twice from 2008 t0 2015; and this increase linked to the rise in cyberbullying incidents [56].

To highlight the seriousness of the problem, we mention some examples of cyberbullying-related suicide: 15-year old Natasha MacBryde took her life after being cyberbullied on the Formspring social network [149]. In 2013, Hannah Smith, a 14-year old girl, hanged herself after receiving bullying messages on her Ask.fm account [227]. Phoebe Prince, a 15 year old teenager, committed suicide after being bullied through her Facebook social network account [96]. 16-year-old Jade Prest contemplated suicide because of cyberbullying via text messages and chatrooms. Ryan Patrick Halligan, a 7th grade teen, took his life after his peers in school spread a rumor of him being gay, and bullied him online and offline [135]. 19 year-old Abraham Biggs, who was suffering from bipolar disorder, live streamed his suicide. He was a target of online harassment and bullying by a group of people [93].

1.4 Cyberbullying Awareness and Prevention

In a 2016 UNESCO report, the importance of bullying and cyberbullying was discussed by presenting case and policies from all around the world. The report indicates that cyberbullying is not only a public health issue, but it is in fact a violation to human rights [85].

In the United States, all states, except Montana, have passed a law against bullying; and the majority of them extended their anti-bullying statutes to include cyberbullying. Nearly half of the states ban cyberbullying. Some states explicitly define some serious forms of cyberbullying such as threats and cyberstalking as a crime; and it could be reported to the law enforcement. In Louisiana, for instance, if a cyberbully is at least 17 years old, they could be fined up to $500 or confined in a prison or both [140].

Schools often consider some forms of punishments for cyberbullying, including suspension, after school detention, and even expulsion, depending on the severity of the issue. In some US schools it is mandatory to educate students about cyberbullying and its consequences. These cases often take the form of school assemblies [71].

Many researchers in psychology and education disciplines study cyberbullying to understand its characteristics and the psychological motivations behind this issue [100, 226]. They then provide some advice to students, parents and school authorities how to approach cyberbullying incident, and how report them.

Cyberbullying is often considered a violation in the terms of service of many social media sites. They define guidelines information in their safety center. Therefore, people can report bullies to the social media sites by flagging inappropriate content. Social network providers have employed human moderators to monitor the flagged messages and delete harmful content or remove the perpetrators and abusing users in strong cases [3, 4]. However, according to a survey by the National Bullying Center in the UK, 91% of people who reported cyberbullying state that not serious action has been taken by the social media site providers [176], and it might be due the huge volume of complaints received daily.

During the past few years, with the growth of cyberbullying incidents, many national and international institutions have been established aiming individual's cybersafety and security such as the 'Non au harcelement' French campaign, the 'KiVa' Finnish cyberbullying prevention program, or clicksafe.be by Belgian government to improve their citizen's online safety [248]. MTV has launched a campaign called “A Thin Line”[5] to educate people about online abuse and its consequences with the aim of stopping the spread of different forms of online harassment.

October was named National Bullying Prevention Month by PACERs National Bullying Prevention Center in 2006.[6] Every October, schools and organizations join this campaign to raise awareness about bullying and cyberbullying and its negative impacts on students.

Some research has been done to support the victims of cyberbullying. Strategies for cyberbullying prevention such as adult supervision, social and curriculum programs so that bystanders intervene and support the victims have been explored [39]. For example, Zwaan et al. [247] designed an Embodied Conversational to Agent (ECA) to help the victim child dealing with the negative effects of cyberbullying. The agent interacts with the victim, and collects information, and finally gives them the necessary advice.

1.5 Automated Methods for Cyberbullying Detection

Detrimental online behavior such as harassment and cyberbullying is becoming a serious, large-scale problem damaging people’s lives. The anti-harassment policy and community standards provided by social media platforms as well as the ability to flag the harassment and block or report the bully are positive steps toward having a safer online community, but these are not sufficient. Popular social media platforms such as Twitter, Facebook, and Instagram receive a huge amount of such flagged contents daily— therefore, inspecting the massive reported messages and users by human editor is very time consuming and not feasible and efficient; in addition to not scaling well [62].

By looking at the problem from different perspectives, designing automated, data-driven techniques for analyzing and detecting such behaviors will be substantially helpful. Successful detection of cyberbullying would allow large-scale social media sites to identify harmful and threatening situations early and prevent their growth more efficiently. According to the research on the desirability of automated systems to signal potential online harassment, majority of respondents believe such monitoring approach will be useful if follow-up strategies is taken and privacy and autonomy of users are not violated [249, 250].

One technique for online harassment detection could be vocabulary-based in which a list of profane words is provided, then using query to extract cyberbullying content [34,84,211]. These methods give higher scores to the posts with more matching profane words. Consequently, they are very sensitive to the quality of words in the list; a large list would result in high recall but low precision (many false positives). These methods only consider the presence of slurs as a strong indication of bullying; neglecting role of users as well as conversation history. There might be many bullying messages without any profane words in it; and these methods are unable to flag them (false negatives).

Many computational social scientists introduced machine learning approaches for cyberbul-

[5] AThinLine.org

[6] www.pacer.org

lying detection in social media. Most machine learning methods for this problem consider supervised text-based cyberbullying detection, classifying social media posts as "bullying" or "non-bullying." In these approaches, crowdsource workers annotate the data, and then a supervised classifier (SVM, Naive Bayes, Decision Tree, etc.) is applied to classify the posts. This is very similar to methods used for sentiment analysis or topic detection; the key challenge here is extracting *good* features that characterize message and users. Wide range of features have been incorporated by scientists so far for this task; for example features extracted from message content such as swear words, pronouns [34, 72, 84, 172], features expressing the emotive or sentiment of the message such as emoticons [170, 173, 265, 266]; features indicating the user characteristics such as age, gender, sexual orientation, and race [51, 59, 60, 171, 214, 232]; or network-based features like number of followers or friends, in-degree, out-degree, etc. [94, 115, 173, 174, 214, 232].

The complexities underlying these behaviors make automatic detection difficult for existing computational approaches. Fully annotating data requires human intervention, which is costly and time consuming. Additionally, analysis of online harassment requires multifaceted understanding of language and social structures. It is necessary to understand the rapidly changing language of cyberbullying, who is harassing whom, and the social roles of the instigators and the victims. For example, given a corpus of social media communication records, keyword searches or sentiment analyses are insufficient to identify instances of harassment as existing sentiment analysis tools often use fixed keyword lists [186]. Or some bullying sentences do not contain slurs or foul words, and according to sentiment analysis they will be classified as bullying. Using existing methods, effective analysis therefore requires human experts with a working comprehension of societal norms and patterns.

1.6 Bias in Social Media

Dinakar et al. and Sanchez et al. [72, 212] in their cyberbullying detection research observe that "race," "culture," "sexuality," "physical appearance," and "intelligence" are the most common bullying topics used to harass others; as a result, they used the slurs associated with these groups as features for their cyberbullying detection model [211].

Social media reflects societal biases, which may be against individuals based on sensitive characteristics such as gender, race, religion, physical ability, and sexual orientation. Machine learning algorithms trained on such data, in spite of their appreciable performance, may therefore perpetuate or amplify discriminatory attitudes and systematic biases against various demographic groups, causing unfair decision-making. This will significantly harm disadvantaged subpopulations.

The bias might be due to the lack of data for some subpopulations, or simply because of the existing stereotypes or prejudice in the society, which arises due to the lack of consideration in the process of data collecting.

1.6.1 Bias in Machine Learning Models

When the input of an algorithm contains existing social biases, the algorithm then incorporates and perpetuates those biases, which leads to discriminatory decision-making. During the past few years, many scientists in machine learning and social science have performed empirical studies to understand the encoded bias in data driven methods; and they introduced strategies to mitigate such discriminations. We briefly mention some of these case studies.

Creditworthiness Nowadays, many employers, landlords and insurance companies use individuals' credit scores to make decisions about offering jobs, houses, and affordable insurance or loans to their applicants [187]. Using big data and machine learning algorithms to evaluate the credit score of customers simplifies the application process for financial institutions. Human mistakes, labor costs, and processing time can be reduced, while they could identify people with good credit. For instance, FICO scores, which are widely used by the financial services industries in the United States, predict creditworthiness using machine learning algorithms [55]. The problem, however, is that credit scores computed by these algorithms are not race or gender neutral because these prediction models may make use of discriminatory information of data such as type of car driven, first names, or zip codes that are highly correlated with ethnicity [19, 69]. This leads to making biased decision making against protected classes of people.

A study at University of California, Berkeley shows that online mortgage loan lenders charge higher interest rate blacks and Latinos than white borrowers with comparable credit history, which costs the minority population 250000 to 500000 more per year [57]. A report from the Federal Reserve in 2007 showed that African American and Hispanics had lower credit scores than Whites and Asian due to their residency in low-income area [254].

Word Embedding Word embedding is a popular framework that represents text data as d-dimensional vectors of real numbers [162,163]. These vectors are learned from co-occurrence data in a large corpus of text such as news articles, webpages, etc. As a result, words with similar semantic meanings shown to have vectors that are close together. A desirable property of these vectors is that they geometrically capture intrinsic relationships between words, for example analogies such as "man is to king as woman is to queen," are captured by the equality [164]:

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{king}} - \overrightarrow{\text{queen}}$$
$$\overrightarrow{\text{Japan}} - \overrightarrow{\text{France}} \approx \overrightarrow{\text{Tokyo}} - \overrightarrow{\text{Paris}}$$

These embedding used in a variety of downstream machine learning and natural language processing applications such as search or result tracking [175], sentiment analysis [118, 244], and question retrieval [142]. Despite its desirable characteristics, there has been a lot of concerns regarding social stereotypes encoded to this word embedding model. Bolukbasi et al. [31] compared the similarity of the vector of some occupations with the vectors for the words "he" and "she" and found out the most similar jobs to "she" are: homemaker, receptionist, librarian, socialite, hairdresser, and nanny, while the vector of occupations maestro, skipper, protege, philosopher, captain, and architect are most similar to "he,"

Therefore, the solution of x in analogy "man is to computer programmer as woman is to x" will be $x =$ *homemaker*. This shows the unfair behavior of word2vec model trained on data carrying social bias. Some methods have been introduced to rectify this discriminatory behavior against gender from word embedding model [31].

Gender Shades Joy Buolamwini, an MIT researcher, was using a 3D camera with facial recognition software for her studies on social robots. Because the software was not able to detect her darker skin face, she needed to wear a white mask to complete her research [81]. "Gender Shades" is a project initiated by Joy to investigate three commercial face recognition systems. They discovered systematic bias in these automated facial analysis algorithms and datasets with respect to race and gender [36]. They demonstrated that there are significant performance gaps across different populations at classification tasks. The accuracy of all three systems was pretty high in detecting the gender using popular benchmark (around 90%); however, the accuracy of these systems dropped noticeably for females compared to males and for dark-skinned people compared to light-skinned ones. More specifically, the most misclassified group were darker-skinned females with the error rate of roughly 34%.

Wikipedia Wager et al. [255] examined the potential gender inequalities in Wikipedia articles. They found that men and women are covered and featured equally well in many Wikipedia language editions. However, a closer investigation reveals that the way women are portrayed is different from the way men are portrayed. Women on Wikipedia tend to be more linked to men; i.e. the Wikipedia articles about women more often highlight their gender, husbands, romantic relationships, and family-related issues, while this is not true for articles about men. They also discovered that certain words such as "husband" is remarkably more frequent in women-related articles, while "baseball" is more frequent in men-related Wikipedia pages.

Criminal Justice System A recent report by the Electronic Privacy Information Center shows that machine learning algorithms are increasingly used in court to set bail, determine sentences, and even contribute to determinations about guilt or innocence [192, 207]. There are various companies that provide machine learning predictive services such as *criminal risk assessment tools* to many criminal justice stakeholders. These risk assessment systems take in the details of a defendants profile, and then estimate the likelihood of recidivism for criminals to help judges in their decision-making. Once a suspect is arrested, they are pre-trialed using these risk assessment tools. The results will be shown to the judge for the final decision. The judge then decides to release or to incarcerate that person. A low score would lead to a kinder verdict; A high score does precisely the opposite [103].

These tools are in operation in some states at the sentencing and parole stage to predict how likely the person will commit a crime if released from prison [230]. For example, *COMPAS* is an system used in Broward County, Florida in pre-trial hearings. It uses a

137-question questionnaire to give a “risk score” to an individual [10]. However, according to an extensive investigative by *ProPublica*, these algorithms reinforce racial inequalities in law enforcement data. Those algorithms falsely labeled high-risk black defendants as future criminals approximately two times more than white defendants. In other words, populations that have historically been unequally targeted by law enforcement–especially low-income and minority communities are highly likely receive higher scores using those risk assessment tools. Many judges who use these risk-assessments did not know how the scores were determined by the algorithm [10]. The concern is that the system was designed to reduce human bias using machines, but the outcomes continued to increase the same biases.

Microsoft Chat Bot *Tay* was an AI chatbot released by Microsoft via Twitter to mimic and converse with users in real time as an experiment for “conversational understanding.” A few hours after the launch of Tay, some Twitter users (trolls) took advantage of Tay’s machine learning capabilities and started tweeting the bot with racist and sexist conversations. A few hours later, Tay quickly began to repeat these sentiments back to the users and post inflammatory and offensive tweets [193, 256]. Around 16 hours after its release, Microsoft shut down the Twitter account and deleted Tay’s sensitive tweets [18].

Predicting income Based on a research by Chen et al. [47], an income-prediction system falsely classified female employees to be low-income much more than wrongly labeling male employees as high-income. They found that this misclassification issue would decrease 40% by increasing the size of training set size ten times.

Beauty.AI The first international online beauty contest judged by artificial intelligence held in 2016 after the launch of *Beauty.AI*.[7] Roughly 6,000 men and women from more than 100 countries submitted their photos to be judged by artificial intelligence, supported by complex algorithms. Out of 44 winners, the majority of them were White, a handful were Asian, and only one had dark skin; while half of the contestants were from India and Africa [144]. Their algorithm was trained using a large datasets of photos; but the main problem was that the data did not include enough minorities; i.e. there were far more images of white women; and many of the dark-skinned images were rejected for poor lighting. This leads to learning the characteristics of lighter skin to be associated with the concept of beauty [191].

Recommendation Systems Recommender systems have been applied successfully in a wide range of domains, such as entertainment, commerce, and employment [245]. The studies show that in the electronic marketplace, online recommendations can change not only consumers’ preference ratings but also their willingness to pay for products [5, 6].

Recommendation systems are seen as tools to accelerate content discover and lead customer engagement. Examples are Netflix’s personalized movie recommendations, Amazon’s sug-

[7] http://beauty.ai/

gested items to users, Spotify's personalized playlist recommendation [133], generating news feeds, recommending job openings, ads on various websites, etc.

However, there have been raising concerns regarding the fairness of such systems as several cases of biases and discrimination have been observed:

- Lower paying job being recommended to women candidates [231],
- The possibility of expensive flight recommendations were made to MacBook owners [120],
- Showing high paid jobs advertisements to men compare to women jobseeker [65],
- Search with keyword "stress in workplace" result in displaying documents related "women in the work force" [200],
- Discrimination against women in recommending STEM course [269],
- Racial discrimination and gender bias in the advertisements showed to users [234].

Amazon Hiring System In 2014, Amazon created a new recruiting system using machine learning algorithms to review applicants' resumes in order to automatically find the top talent applicants. The system was designed to give job candidates scores ranging from one to five stars [64]. In 2015, they noticed that their system was not rating female applicants similarly to male applicants for technical jobs such as software developer. It was more favorable toward male applicants for technical jobs. The models were trained based on resumes from the past 10 years, which were mostly male applicants. Consequently, Amazon's system prefer men that women for technical jobs. Although they adjusted the system to make it neutral against gender, there was no guarantee that the model wont exhibit discriminatory behaviors in future. Eventually, they stopped the system when they observed such bias in their hiring process using this AI-based system.

1.7 Fairness in Machine Learning

Biases created by machine learning models are increasingly prevalent due to its widespread use in modern era. This might cause expensive loss and harm in individuals' lives. The U.S. Equality Act has defined some protected populations such as age, race, gender, religion, etc. and prohibited discrimination on the basis of membership to these classes when making decisions in broad areas such as housing, employment, public education, public accommodations, credit, and the jury system [48]. For example, most credit decisions are require to comply with U.S. Equality Act such as the Equal Credit Opportunity Act (ECOA) [2] and Fair Credit Reporting Act (FCRA) [1], and some more. These laws prohibit discrimination on the basis sexual orientation, gender, color, disability, religion, national origin, etc. [48]. The problem of using machine learning algorithms while complying with law by U.S. Equality Act (such as ECOA and FCRA) is the lack of transparency and explainability since they need to provide an explanation of results specifically if discrimination observed.

In 2016, the Obama Administration requested investigation of big data and machine learning algorithms to ensure fairness [223]. In addition, the EU passed the General Data Protection Regulation (GDPR) in 2018, in which citizens are allowed to ask for the explanation about the decisions made by machine learning model [194].

These policies enforce the growth of a subfield in machine learning known as "fairness in machine learning" with the goal of creating ethical automatic algorithm. During the past few years, the machine learning community is actively responding to these biases. There are several questions they seek to answer:

What are the reasons for discrimination in a ML model? There are many sources of unfairness in a model. One main reason is the bias in the training data. For example, when there is a lack of samples for the under-represented groups or there is an implicit bias in training data, ML models can perform poorly on these groups.

How to measure fairness in a ML model? Many different notions of fairness have been proposed in literature: demographic parity (requires that a decision be independent of the protected attribute), equality of odds (equal true/false positive rate in all groups), and equality of opportunity (equal true positive rate in all groups).

How to enforce fairness in a ML model? There are several ways to enforce fairness to machine learning models: 1) pre-processing, in which discriminatory characteristics are removed from training data to make it a fair representation of all protected groups; 2) constraint-based, which involves imposing fairness criteria as constraints into the loss function; 3) post-processing in which a trained model is treated as a black box, and using an audit data, the model's output is modified such that desired fairness criteria are met.

1.8 Bias in Online Harassment Systems

A key concern in the development and adoption of machine learning for building harassment detectors is whether the learned detectors are fair. Biases in this context could take the form of bullying detectors that make false detections more frequently on messages by or about certain identity groups because such keywords in identity groups may occur relatively frequently in toxic samples in the training set [74]. For example, the sentences "I am gay" and "I am straight" will be give different toxicity scores. We geared the last part of our research toward fairness against particular targeted groups. Our goal is to decrease the sensitivity of models to language describing particular social groups.

1.9 Proposed Framework

Most machine learning methods for this problem consider supervised text-based cyberbullying detection, classifying social media posts as “bullying” or “non-bullying.” In these approaches, crowdsource workers annotate the data, and then a supervised classifier is applied to classify the posts. There are, however, several challenges related to these approaches. Fully annotating data requires human intervention, which is costly and time consuming. And analysis of online harassment requires multifaceted understanding of language and social structures. Without considering social context, differentiating bullying from less harmful behavior is difficult due to complexities underlying cyberbullying and related behavior. Our approach aims to encode such complexities into an efficiently learnable model. We present an automated, data-driven method for identification of harassment. Our approach uses machine learning with weak supervision, which significantly alleviating the need for human experts to perform tedious data annotation.

Participant-Vocabulary Consistency We use machine learning with weak supervision, which significantly alleviates the need for human experts to perform tedious data annotation. Our weak supervision is in the form of expert-provided key phrases that are highly indicative or counter-indicative of bullying. For example, various swear words and slurs are common indicators of bullying, while positive-sentiment phrases are counter-indicators. The algorithms then extrapolate from these expert annotations to find instances of bullying.

Our proposed general framework for weak supervision is based on consistency of two detectors that co-train one another. In this framework, an ensemble of cyberbullying detectors will seek consensus on whether examples in unlabeled data are cases of cyberbullying, while one of these detectors receives weak supervision. These detectors use two different perspectives of the data: language and social structure. The intuition behind this approach is that the detectors will use different forms of evidence to reach the same conclusion about whether social interactions are bullying. Exploiting different learners aligns with the true nature of cyberbullying that can occur in different directions. To train the models, we construct an optimization problem made up of two sets of loss functions: a co-training loss that penalizes the disagreement between language-based learner and user-based learner, and a weak-supervision loss that is the classification loss on weakly labeled messages.

In the first step, we introduced *participant-vocabulary consistency* (PVC) model, which uses a similar paradigm of viewing the learning tasks as seeking consensus between language-based and user-based perspectives of the problem. The algorithm seeks a consistent parameter setting for all users and key phrases in the data that specifies the tendency of each user to harass or to be harassed and the tendency of a key phrase to be indicative of harassment. The learning algorithm optimizes the parameters to minimize their disagreement with the training data. Thus, it fits the parameters to patterns of language use and social interaction structure. We demonstrate that PVC can discover examples of apparent bullying as well as new bullying indicators.

Co-trained Ensembles of Embedding Models PVC uses simple key-phrase presence and a two-parameter user characterization as its vocabulary and participant detectors, respectively. To improve predictive performance of PVC, we replaced vocabulary and participant detectors with richer language and user classifiers using deep embedding methods. Currently, we are in the progress of applying embedding methods, which represents words and users as vectors of real numbers. When word embeddings are trained using deep learning, the vectors created by word embeddings preserve contextual similarities, so we can extract meaning from text to derive similarities and other relationships between words. Word2vec [162, 163] is a popular word-embedding model that represents words with low-dimensional vectors based on how often words appear near each other. And node2vec [99] is a framework for learning continuous feature representations for nodes in networks. We use word and user vectors as the input to language-based and user-based classifiers, respectively. We examine two strategies when incorporating vector representations of words and users. First, using existing doc2vec [141]—an extension of word embedding— and node2vec models as inputs to the learners. Second, creating new embedding models specifically geared for our specific task of harassment detection, which we train in an end-to-end manner during optimization of the models, by incorporating the unsupervised doc2vec and node2vec loss function into our co-training objective. Our results on Twitter, Instagram, and data show that our weakly supervised deep models improve precision over a non-deep variation of the models.

Reduced-Bias Co-Trained Ensembles Social media reflects societal biases, which may be against individuals based on sensitive characteristics such as gender, race, religion, physical ability, and sexual orientation. Machine learning algorithms trained on such data may therefore perpetuate or amplify discriminatory attitudes against various demographic groups, causing unfair decision-making. A key concern in the development and adoption of machine learning for building harassment detectors is whether the learned detectors are *fair*. Biases in this context could take the form of bullying detectors that make false detections more frequently on messages by or about certain identity groups. In our next step, we extended the co-trained ensemble model to mitigate unfair behavior in the trained model. We add an unfairness penalty to the framework to prevent unintended discrimination against particular social groups. The unfairness term penalizes the model when we observe discrimination in predictions. We explore two unfairness penalty terms, each aiming toward a different notion of fairness. One aims for *removal fairness* and the other for *substitutional fairness*. An ideal, fair language-based detector should treat language describing subpopulations of particular social groups equitably. We quantitatively and qualitatively evaluate the resulting models' fairness on a synthetic benchmark and data from Twitter using post-hoc, crowdsourced annotation.

1.10 Main Contributions

The main contributions of this research are as follows:

Weak Supervision Our framework uses minimal supervision to learn the complex patterns of cyberbullying. Using weak supervision significantly alleviate the need for human experts to perform tedious data annotation. Our weakly supervised approach is built on the idea that it should be inexpensive for human experts to provide weak indicators of some forms of bullying, specifically vocabulary commonly used in bullying messages. The algorithm extrapolates from the weak indicators to find possible instances of bullying in the data.

Co-train Ensemble Our framework consists of an ensemble of two learners that co-train one another. The framework trains an ensemble of two learners in which each learner looks at the problem from a different perspective. One learner identifies bullying incidents by examining the language content in the message; another learner considers the social structure to discover bullying. Individual learners train each other to come to an agreement about the bullying concept.

Non-linear Embedding Our framework incorporates the efficacy of distributed word and graph-node representations. We represent words and users as low-dimensional vectors of real numbers. We use word and user vectors as the input to nonlinear language-based and user-based classifiers, respectively. We examine two strategies when incorporating vector representations of words and users. First, we use existing doc2vec, which is an extension of word embedding, and node2vec models as inputs to the learners. Second, we create new embedding models specifically geared for analysis of cyberbullying, in which word and user vectors are trained during optimization of the model.

Reduce Bias We propose a reduced-biased framework for weakly supervised training of cyberbullying detectors. We penalize the model if we observe discrimination against some particular social groups in the predictions. To penalize the model against discrimination, we add an unfairness penalty term to the objective function.

To the best of our knowledge, this framework is the first specialized algorithm for cyberbullying detection that considers fairness while detecting online harassment with weak supervision.

1.11 Overview and outline

The remaining part of this dissertation is organized as follows:

- Chapter 2 is a thorough literature review of the research has been done on online harassment and cyberbullying detection, machine learning with weak supervision, fairness in machine learning, multi-view learning, and vocabulary discovery.
- Chapter 3 explains our first approach, *Participant-Vocabulary Consistency* or in short *PVC*, which is a machine learning method for simultaneously inferring user roles in harassment-based bullying and new vocabulary indicators of bullying. *PVC* considers social structure and infers which users tend to bully and which tend to be victimized.

To address the elusive nature of cyberbullying, this learning algorithm only requires weak supervision. Experts provide a small seed vocabulary of bullying indicators, and the algorithm uses a large, unlabeled corpus of social media interactions to extract bullying roles of users and additional vocabulary indicators of bullying. The model estimates whether each social interaction is bullying based on who participates and based on what language is used, and it tries to maximize the agreement between these estimates.

- Chapter 4 describes the extension of PVC to benefit from non-linear embedding models. The introduced framework, *Co-trained Ensembles of Embedding Models*, has three distinguishing characteristics. (1) It uses minimal supervision in the form of expert-provided key phrases that are indicative of bullying or non-bullying. (2) It detects harassment with an ensemble of two learners that co-train one another; one learner examines the language content in the message, and the other learner considers the social structure. (3) It incorporates distributed word and graph-node representations by training nonlinear deep models. The model is trained by optimizing an objective function that balances a co-training loss with a weak-supervision loss.

- Chapter 5 presents the *Reduced-Bias Co-Trained Ensembles* for training bullying detectors from weak supervision while reducing how much learned models reflect or amplify discriminatory biases in the data. The goal is to decrease the sensitivity of models to language describing particular social groups. Building off the previously proposed weakly supervised learning algorithm, we modify the objective function by adding an unfairness penalty term to the framework. By penalizing unfairness, we encourage the learning algorithm to avoid discrimination in the predictions and achieve equitable treatment for protected subpopulations. We introduce two unfairness penalty terms: one aimed at removal fairness and and another at substitutional fairness. An ideal, fair language-based detector should treat language describing subpopulations of particular social groups equitably.

- Chapter 6 shows some complementary experiments, which includes comparing our framework against fully supervised model, considering other transparent node representations, and examining inter-annotator agreement.

- Chapter 7 summarizes our research, explains its limitations, and discusses future prospects.

Chapter 2

Literature Review

The three main bodies of research that support our contribution are (1) emerging research investigating online harassment and cyberbullying, (2) research on weakly supervised machine learning and developing automated methods for vocabulary discovery, and (3) research developing fairness in machine learning algorithms.

2.1 Online Harassment and Cyberbullying Detection

A variety of methods have been proposed for cyberbullying detection. These methods mostly approach the problem by treating it as a classification task, where messages are independently classified as bullying or not. Many of the research contributions in this space involve the specialized design of message features for supervised learning. There have been many contributions that design special features.

Language Features Dadvar et al. [60] proposed gender-specific language features to classify users into male and female groups to improve the discrimination capacity of a classifier for cyberbullying detection. Chen et al. [51] study the detection of offensive language in social media, applying the *lexical syntactic feature* (LSF) approach that successfully detects offensive content in social media and users who send offensive messages. Dinakar et al. [72] focus on detecting of textual cyberbullying in YouTube comments. They collected videos involving sensitive topics related to race and culture, sexuality, and intelligence. By manually labeling 4,500 YouTube comments and applying binary and multi-class classifiers, they showed that binary classifiers outperform multi-class classifiers. Reynolds et al. [202] detect cyberbullying based on a language-based method. They used the number, density and the value of offensive words as features. Their research has successfully identified Formspring messages that contain cyberbullying by recording the percentage of curse and insult word in posts. Yin et al. [270] created three types of features: content features (word occurrence weighted by term-frequency inverse-document-frequency, or TF-IDF), sentiment features (offensive words and pronouns), and contextual features (the similarity of a user to her neighbor) to train a model for detecting harassing posts in chat rooms and discussion forums.

Ptaszynski et al. [195] develop a systematic approach to *online patrol* by automatically spotting online slandering and bullying and reporting them to parent-teacher association (PTA) members of schools. At first, they manually gathered a lexicon of curse words indicative of cyberbullying. To recognize informal or jargonized swear words, they calculated word similarity with Levenshtein distance. They then train a support vector machine to classify harmful messages from non-harmful ones. They mentioned that, since new vulgar words appear frequently, finding a way to automatically extract new vulgarities from the Internet is required to keep the lexicon up to date. Researchers have proposed methods that model posts written by bullies, victims, and bystanders using linear support vector machines and designing text-based features on an annotated corpus of English and Dutch [248]. Researchers have also trained a supervised three-class classifiers using language features to separate tweets into three categories: those containing hate speech, only offensive language, and those with neither [67]. Dani et al. [63] use sentiment information to detect cyberbullying behaviors in social media. Nandhini et al. [233] use fuzzy logic and genetic algorithm to find cyberbullying incidents in social network using language features.

Social-Structure Features Some researchers consider social-structure features in cyberbullying analysis. Fore example, Huang et al. [115] investigate whether analyzing social network features can improve the accuracy of cyberbullying detection. They consider the social network structure between users and derived features such as number of friends, network embeddedness, and relationship centrality. Their experimental results showed that detection of cyberbullying can be significantly improved by integrating the textual features with social network features. Tahmasbi et al. [235] investigate the importance of considering user's role and their network structure in detecting cyberbullying. Chatzakou et al. [44, 45] extract features related to language, user, and network; then, they study which features explain the behavior of bullies and aggressors the best. Nahar et al. [172] propose an approach to detect cyberbullying messages from social media data. First, they detect harmful messages using semantic and weighted features. They compute semantic features using *latent Dirichlet allocation* (LDA) and weighted bullying indicators such as a static curse-word vocabulary, weighted by TF-IDF. They identify predators and victims through their user interaction graph. They present a ranking algorithm to find the most active predators and victims. A key point about this work is that their bullying indicator feature sets are limited to a static set of keywords.

Researchers have used probabilistic fusion methods to combine social and text features together as the input of classifier [219]. Tomkins et al. [243] propose a socio-linguistic model that detects cyberbullying content in messages, latent text categories, and participant roles using probabilistic models. They also propose a linguistic model with domain knowledge and reduce the number of parameters appropriate for learning from limited labeled data [242]. Chelmis et al. [46] perform analysis on a large-scale data to identify online social network topology structure features that are the most prominent in enhancing the accuracy in cyberbullying detection methods. Balakrishnan et al. [16] introduce Big Five and Dark Triad models for cyberbullying detection based on seven user personality features (number of hash tags, favorite count, number of mentions, popularity, number of followers and following, and status count)

Linguistic and Statistical Analysis Hosseinmardi et al. [109–112] conducted several studies analyzing cyberbullying on Ask.fm and Instagram, with findings that highlight cultural differences among the platforms. They studied negative user behavior in the Ask.fm social network, finding that properties of the interaction graph—such as in-degree and out-degree—are strongly related to the negative or positive user behaviors [109]. They studied the detection of cyberbullying incidents over images in Instagram, providing a distinction between cyberbullying and cyber-aggression [112]. They also compared users across two popular online social networks, Instagram and Ask.fm, to see how negative user behavior varies across different venues. Based on their experiments, Ask.fm users show more negativity than Instagram users, and anonymity tends to result in more negativity (because on Ask.fm, users can ask questions anonymously) [110]. Rafiq et al. [196] propose a highly scalable multi-stage cyberbullying detection, which is highly responsive in raising alerts.

Rezvan et al. [203, 204] group harassment incidents into five categories: (i) sexual, (ii) racial, (iii) appearance-related, (iv) intellectual, and (v) political, and then perform extensive linguistic and statistical analysis based on the results of type-specific harassment detection method. Bellmore et al. [23] develop machine learning methods to better understand social-psychological issues surrounding the idea of bullying. By extracting tweets containing the word "bully," they collect a dataset of people talking about their experiences with bullying. They also investigate different forms of bullying and why people post about bullying. In another related direction, researchers introduced a user-centric view for hate speech, examining the difference between user activity patterns, the content disseminated between hateful and normal users, and network centrality measurements in the sampled graph [205].

Deep Learning Pitsilis et al. [189] applied recurrent neural networks (RNN) by incorporating features associated with users tendency towards racism or sexism with word frequency features on a labeled Twitter dataset. Al-Ajlan et al. [7] applied convolutional neural network (CNN) and incorporates semantics through the use of word embeddings. Zhao et a. [277] extended stacked denoising autoencoder to use the hidden feature structure of bullying data and produce a rich representation for the text. Kalyuzhnaya et al. [126] classify a tweet as racist, sexist, or neither using deep learning methods by learning semantic word embeddings. Dadvar et al. [61] investigate the performance of several models introduced for cyberbullying detection on Wikipedia, Twitter, and Formspring as well as a new YouTube dataset. They found out that using deep learning methodologies, the performance on YouTube dataset increased.

Others Ashktorab et al. [12] provide a number of potential mitigation solutions for cyberbullying among teenagers and the technologies required to implement these solutions. Yao et al. and Zois et al. [268, 279] formulate cyberbullying detection as a sequential hypothesis testing problem to reduce the number of features used in classification. Li et al. [147] propose a method that take advantage of the parent-child relationship between comments to receive the reaction from a third party to detect cyberbullying. Soni et al. [228] use multiple audio and visual features along with textual features for cyberbullying detection. Cheng et al. [52] introduce a framework that creates a heterogeneous network from social media data, and then a node representations. These learnt numerical vectors are the input of machine learning models. Rosa et al. [208] examined a method Fuzzy Fingerprints for the task of cyberbullying detection, showing its slight improvement over textual cyberbullying detection within the supervised machine learning.

Related research on data-driven methods for analysis and detection of cyberviolence in general includes detection of hate speech [75, 178, 258], online predation [158], and the analysis of gang activity on social media [185], among many other emerging projects.

Non-English Context Much research have been done for detection and analysis of cyberbullying and online harassment in non-English context. Dutch [248], Arabic [101, 169], Malaysia [15], Japanese [195], Indonesian [155], Hindi [67, 136, 153], Spanish [91, 97, 236], German [14, 29, 213, 253] are some examples.

2.2 Weakly Supervised Machine Learning and Vocabulary Discovery

Weakly Supervised Machine Learning To train a supervised machine learning method, input data is required to be labeled. Data annotation is very costly and time-demanding process. Therefore, many researchers have been developing weakly supervised algorithms in which limited number of data is labeled. The intuition behind weak supervision is that the learning algorithm should be able to find patterns in the unlabeled data to integrate with the weak supervision. Weakly supervised approaches have been explored for relation extraction in text [35, 108, 166, 206, 267], and in the domain of knowledge extraction from the Web. It has many applications in computer vision tasks such as semantic segmentation [50, 264] and in automated healthcare tasks such as clinical tagging of medical conditions from electronic health records [102], both cases where human annotation is expensive. Various general approaches for weakly supervised learning have been proposed for classification [177, 199, 278].

Vocabulary Discovery Part of our proposed method simultaneously learns new language indicators of bullying while estimating users' roles in bullying behavior. Learning new language indicators is related to the task of query expansion in information retrieval [154]. Query expansion aims to suggest a set of related keywords for user-provided queries. Massoudi et al. [157] use temporal information as well as co-occurrence to score the related terms to expand the query. Lavrenko et al. [139] introduce a relevance-based approach for query expansion by creating a statistical language model for the query. This commonly-used approach estimates the probabilities of words in the relevant class using the query alone. Mahendiran et al. [151] propose a method based on probabilistic soft logic to grow a vocabulary using multiple indicators (social network, demographics, and time). They apply their method to expand the political vocabulary of presidential elections.

Multi-view Learning Our framework is motivated by ideas studied in *multi-view learning* [167, 190, 262, 263]. Multi-view learning is specifically useful when data is comprised of multiple views of some underlying phenomenon. Algorithms can exploit multi-view information to improve the learning performance. Blum and Mitchell [30] introduced a co-training method for multi-view learning, primarily for semi-supervised problems. These co-training algorithms alternately learn model parameters to maximize the mutual agreement across two distinct views of the unlabeled data. To understand the properties and behaviors of multi-view learning, some researchers have studied its generalization-error via PAC-Bayes and Rademacher complexity theory [276].

2.3 Fairness in Machine Learning

Recent reactions to a Google Jigsaw-released tool for quantifying toxicity of online conversations (see e.g., [218]) have highlighted an important aspect of any automated harassment or bullying detection: fairness, especially in the context of false positives. A serious concern of these detectors is how differently they flag language used by or about particular groups of people. We begin to address this issue with a benchmark analysis in our experiments and propose an algorithm to reduce unfairness.

Discrimination in Machine Learning Models In recent years, machine learning researchers are addressing the need for *fair* machine learning algorithms. Machine learning algorithms can exhibit discriminatory decision making in areas such as recommendation, prediction, and classification [66,82,129,234]. According to Edelman et al., machine learning algorithms trained on biased data would even increase the bias to improve their performance [82]. Islam et al. [119] discuss how language itself contains human-like biases. Bolukbasi et al. [31] demonstrate sexist trends in word embeddings trained on Google News articles.

Young et al. [271] study the dilemmas related to the current data sharing paradigm for (i) privacy, (ii) fairness, and (iii) accountability in the urban transportation data domain. Babaei et al. [13] analyze the users' perceptions of truth in news stories to see how bias in user's perception might affect the methods detecting and labeling fake news. Green et al. [98] study how risk assessments influence human decisions in the context of criminal justice adjudication. They compare the people's prediction about risk with and without the interference of risk assessment. Chen et al. [49] theoretically analyze the bias in outcome disparity assessments using a probabilistic proxy model for the unobserved protected class. Ustun [246] assess a linear classification model in terms of "recourse." They define recourse as changing the decision made by the model using some particular input features (such as income, marital status, or age) by a person.

Chancellor et al. [43] present a taxonomy of issues in predictions made by algorithms for mental health status on social media data. They noted the existing gap among social media researchers. Hu [113] adapt strategic manipulation of models to realize the behaviors causing social inequality. They study the effects of interventions in which a learner reduces the cost of the disadvantaged group to improve their classification performance. Benthall et al. [25]

precede group fairness interventions with unsupervised learning to identify discriminative behaviors. Kannan et al. [127] consider two explicit fairness goals that a college might have for its action policies: granting equal opportunity to students with the same type when graduating from high school, independent of their group membership, and encouraging employers to make hiring decisions that are independent of group membership. Heidari et al. [105] study a fairness measurements in the context of economy; showing that proposed fairness measures, such as equality of odds and predictive value parity, can be translated to Equality of opportunity (EOP) in the economic models.

De-Arteaga et al. [68] study the gender bias in occupation classification. Elzayn et al. [86] explore the general notion of fairness, "equality of opportunity" in several applications and investigate its algorithmic consequences. Obermeyer et al. [180] show that there is a significant bias toward races in a commercial algorithm that is on the operation in most leading Accountable Care Organizations (ACOs) in the U.S. Milli et al. [165] study the trade-off between accuracy to the institution and impact to the individuals being classified, showing any increase in institutional utility causes increasing social burden.

Barocas et al. [17] introduce formal non-discrimination criteria, established their relationships, and illustrate their limitations. Hutchinson et al. [117] explored and compared various notions of fairness and the mathematical method for measuring fairness over past half century. Chouldechova et al. [53] describe how they develop, validate, audit, and deploy a risk prediction model in Allegheny County, PA, USA, in the context of fairness. They studied the results of their analysis, and emphasize main obstacles and biases in data that causes challenges for evaluation and deployment. Kallus et al. [125] introduce the notation of *residual unfairness* and study how imposing fairness to model (under certain conditions) might not be fair toward the target population because of distorting the training data using the biased policies. Friedler et al. [92] compare the behavior of several fairness-enhanced classifiers according to some fairness measures introduced thus far. Their experiments indicate many of fairness measures strongly correlate with one another, and fairness interventions might cause instability to the algorithm.

In order to mitigate discriminatory behavior in machine learning models, three solutions have been introduced: pre-processing, post-processing, and imposing fairness constraints to the model.

Pre-processing Research in this category consists of changing the input data to make it have a true representation of various demographic groups. Feldman et al. [88] proposed the idea of using individual's nonsensitive features to predict their sensitive features. The tight connection between the dataset's sensitive and nonsensitive attributes is the cause of discriminatory algorithms. They proposed changing the responsible nonsensitive elements in the dataset.

Post-processing In some research, they assume re-training the machine learning model is costly and not possible in many cases. Therefore, they change the models' behavior to make it fairer toward particular social group by having black-box access to these models.

Kim et al. [131] developed a framework, multiaccuracy auditing, that ensures accurate predictions across identifiable subgroups using post-processing. Multiaccuracy boost works in any setting where we have black-box access to a predictor and a relatively small set of labeled data for auditing (validation set). During auditing process, if the predictor does not satisfy multiaccuracy, they apply post-processing step to produce a new classifier that is multiaccurate.

Dwork et al. [80] develop a decoupling technique which can be added to any black-box machine learning model to learn different classifiers for different population. InclusiveFaceNet introduced by [210], is a method to improve face attribute detection accuracy across race and gender subpopulation by learning demographic information before training the attribute detection task. Canetti et al. [40] use post-processing a calibrated soft classifier to gain a binary decision, under group fairness constraints, for the case of several protected groups.

Designing Fair Model Various research has been done to design algorithms that make fair predictions toward different demographic groups.

Hardt et al. [104] propose reducing bias after a model is trained on the data by introducing a fairness measure, which they refer to as equal opportunity, for discrimination against a specified sensitive attribute in supervised learning. Zafar et al. [272] introduce another notion of unfairness, *disparate mistreatment*, which is defined in terms of different misclassification rates for people belonging to different social groups.

Dwork et al. [79] considered fairness in the classification problem, where individuals are classified. They defined a metric to determine how much people are similar according to the classification task, and create a fairness constraint such that similar individuals are treated similarly via Lipschitz condition. This work was extended by combining individual fairness with a statistical group-based fairness criteria [273]. Kilbertus et al. [130] look at the discrimination problem from causal reasoning perspective and describe indirect/proxy discrimination in terms of causal graphs; accordingly, they proposed some methodologies to prevent discrimination.

Garg et al. [95] introduce counterfactual fairness in text classification by substituting tokens related to sensitive groups. Then, they introduced three methods for optimizing their counterfactual fairness metric during model training. Their methods are hard ablation, blindness, and counterfactual logit pairing. Zhang et al. [274] examine three fairness measures: demographic parity, equality of odds, and equality of opportunity in the context of adversarial debiasing. They introduced a framework for mitigating undesired biases concerning demographic groups by including a variable for the group of interest and simultaneously learning a predictor and an adversary.

Beutel et al. [26] introduced the multi-head deep neural network, where the model tries to predict the target class with one head while simultaneously preventing the second head from being able to accurately predict the sensitive attribute. Their goal is to exclude any information regarding the sensitive attribute from the latent representation created by a deep model [26]. Menon et al. [160] study the inherent trade-offs in learning classifiers with a fairness constraint by answering two questions: what is the best accuracy we can expect for a given level of fairness? And what is the nature of these optimal fairness-aware classifiers?.

Zhao et al. [275] examine multilabel object classification and visual semantic role labeling task and found out the datasets for these tasks contain significant gender bias. They, then calibrate the models using corpus-level constraints, and applied Lagrangian relaxation for collective inference. Card et al. [41] change the conventional deep architecture using a methodology inspired from non parametric kernel regression to provide transparent explanations for classification predictions. Celis et al. [42] propose a meta-algorithm for classification task in which the input could be a set of fairness constraints with respect to several non-disjoint and multi-valued sensitive attributes.

Madras et al. [150] use a causal modeling approach to learn from biased data. In this method, sensitive feature confounds both the treatment and the outcome. Russell [209] introduces a new search algorithms using a novel set of constraints that could be solved by integer programming techniques to efficiently find coherent counterfactual explanations. Albarghouthi et al. [8] introduce fairness-aware programming where programmers can define their fairness expectations while coding, and have a system that monitors decision-making and report violations of fairness.

Chapter 3

Weakly Supervised Cyberbullying Detection with Participant-Vocabulary Consistency

3.1 Introduction

Analysis of online harassment requires multifaceted understanding of language and social structures. The complexities underlying these behaviors make automatic detection difficult for static computational approaches. For example, keyword searches or sentiment analyses are insufficient to identify instances of harassment, as existing sentiment analysis tools often use fixed keyword lists [186]. In contrast, fully supervised machine learning enables models to be adaptive. However, to train a supervised machine learning method, labeled input data is necessary. Data annotation is a costly and time-demanding process. High-quality hand-labeled data is a key bottleneck in machine learning. Therefore, many researchers have been developing weakly supervised algorithms in which only a limited amount of data is labeled. The intuition behind weak supervision is that the learning algorithm should be able to find patterns in the unlabeled data to integrate with the weak supervision. We use this weak supervision paradigm to significantly alleviate the need for human experts to perform tedious data annotation. Our weak supervision is in the form of expert-provided key phrases that are highly indicative of bullying. For example, various swear words and slurs are common indicators of bullying. The algorithms then infer unknown data values from these expert annotations to find instances of bullying.

The algorithm we present here learns a relational model by using the structure of the communication network. The relational model is trained in a weakly supervised manner, where human experts only need to provide high-fidelity annotations in the form of key phrases that are highly indicative of harassment. The algorithm then extrapolates from these expert annotations—by searching for patterns of victimization in an unlabeled social interaction network—to find other likely key-phrase indicators and specific instances of bullying. Fig. 3.1 illustrates the key components of this machine learning paradigm.

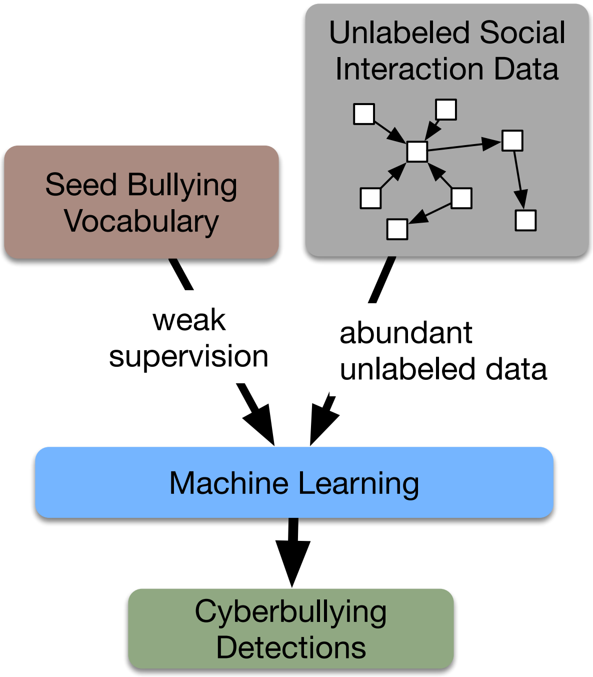


Figure 3.1: Schematic of the weak supervision paradigm for training a cyberbullying detector.

We refer to the proposed method as the *participant-vocabulary consistency* (PVC) model. The algorithm seeks a consistent parameter setting for all users and key phrases in the data that characterizes the tendency of each user to harass or to be harassed and the tendency of a key phrase to be indicative of harassment. The learning algorithm optimizes the parameters to minimize their disagreement with the training data, which takes the form of a directed message network, with each message acting as an edge decorated by its text content. PVC thus fits the parameters to patterns of language use and social interaction structure.

An alarming amount of harassment occurs in public-facing social media, such as public comments on blogs and media-sharing sites. We will use this type of data as a testbed for our algorithms. According to a survey by *ditchthelabel.org* [73], the five sites with the highest concentration of cyberbullying are Facebook, YouTube, Twitter, Ask.fm, and Instagram. We evaluate participant-vocabulary consistency on social media data from three of these sources: Twitter, Ask.fm, and Instagram. We use a human-curated list of key phrases highly indicative of bullying as the weak supervision, and we test how well participant-vocabulary consistency identifies examples of bullying interactions and new bullying indicators.

We conduct wide range of quantitative and qualitative experiments to examine how well PVC identifies examples of bullying interactions and new bullying indicators. In our quantitative evaluation, we use post-hoc human annotation to measure how well PVC fits human opinions about bullying. In our qualitative analysis, we group the identified conversations by PVC into three categories: 1) true positives that other baselines were not be able to detect, 2) true positives not containing very obvious offensive languages, and 3) false positives. We inspect the false positives and notice there are four different types: (1) users talking about other people, not addressing each other in their messages, (2) joking conversations, (3) users talking about some bullying-related topics, (4) conversations with no language indicative of bullying. We also analyze another group of false positives we call *unfair false positives*. Fairness is an important topic when considering any online automated harassment detection. We measure the sensitivity of PVC to language describing particular social groups, such as those defined by race, gender, sexual orientation, and religion. In another set of qualitative evaluations, we show the relationship between the learned user's bully and victim scores in heatmap and scatter plots. We also provide a summary statistics about bullies and victims such as their average in-degree and out-degree. In addition, we showed a few small sub-networks of identified bullies and victims to see how differently they are distributed around each other.

The main contributions of this work are as follows: We present the participant-vocabulary consistency model, a weakly supervised approach for simultaneously learning the roles of social media users in the harassment form of cyberbullying and the tendency of language indicators to be used in such cyberbullying. We demonstrate that PVC can discover examples of apparent bullying as well as new bullying indicators, in part because the learning process of PVC considers the structure of the communication network. We evaluate PVC on a variety of social media data sets with both quantitative and qualitative analyses. This method is the first specialized algorithm for cyberbullying detection that allows weak supervision and uses social structure to simultaneously make dependent, collective estimates of user roles in cyberbullying and new cyberbullying language indicators.

3.2 General Framework

Our weakly supervised approach is built on the idea that it should be inexpensive for human experts to provide weak indicators of some forms of bullying, specifically vocabulary commonly used in bullying messages. The algorithm extrapolates from the weak indicators to find possible instances of bullying in the data. Then, considering the discovered users who tend to be involved in bullying, the algorithm finds new vocabulary that is commonly used by these suspected bullies and victims. This feedback loop iterates until the algorithm converges on a consistent set of scores for how much the model considers each user to be a bully or a victim, and a set of scores for how much each vocabulary key-phrase is an indicator of bullying. The idea is that these vocabulary scores will expand upon the language provided in the weak supervision to related terminology, as well as to language used in different types of bullying behavior. The algorithm considers the entire network of communication, propagating its estimates of bullying roles through the messaging structure and the language used in each message, leading to a joint, collective estimation of bullying roles across the network.

We use a general data representation that is applicable to a wide variety of social media platforms. To formalize the observable data from such platforms, we first consider a set of users U and a set of messages M. Each message $m \in M$ is sent from user $s(m)$ to user $r(m)$. I.e., the lookup functions s and r return the sender and receiver, respectively, of their input message. Each message m is described by a set of feature occurrences $f(m) := \{x_k, \ldots, x_\ell\}$. Each feature represents the existence of some descriptor in the message. In our experiments and in many natural instantiations of this model, these descriptors represent the presence of n-grams in the message text, so we will interchangeably refer to them as vocabulary features.

For example, if m is a Twitter message from user @alice with the text "@bob hello world", then

$$s(m) = \text{@alice}, \quad r(m) = \text{@bob}$$
$$f(m) = \{\text{hello}, \text{world}, \text{hello world}\}.$$

In this representation, a data set can contain multiple messages from or to any user, and multiple messages involving the same pair of users. E.g., @alice may send more messages to @bob, and they may contain completely different features.

To model cyberbullying roles, we attribute each user u_i with a bully score b_i and a victim score v_i. The bully score encodes how much our model believes a user has a tendency to bully others, and the victim score encodes how much our model believes a user has a tendency to be bullied. We attribute to each feature x_k a bullying-vocabulary score w_k, which encodes how much the presence of that feature indicates a bullying interaction.

For each message sent from user u_i to user u_j, we use an additive *participant score* combining the sender's bully score and the receiver's victim score $(b_i + v_j)$. The more the model believes u_i is a bully and u_j is a victim, the more it should believe this message is an instance of bullying. To predict the bullying score for each interaction, we combine the total average word score of the message with the participant score

$$\underbrace{\left(b_{s(m)} + v_{r(m)}\right)}_{\text{participant score}} + \underbrace{\frac{1}{|f(m)|} \sum_{k \in f(m)} w_k}_{\text{vocabulary score}} \ . \tag{3.1}$$

We then define a regularized objective function that penalizes disagreement between the social bullying score and each of the message's bullying-vocabulary scores:

$$\begin{aligned} J(\mathbf{b}, \mathbf{v}, \mathbf{w}) = {} & \tfrac{\lambda}{2} \left(||\mathbf{b}||^2 + ||\mathbf{v}||^2 + ||\mathbf{w}||^2\right) + \\ & \tfrac{1}{2} \sum_{m \in M} \left(\sum_{k \in f(m)} \left(b_{s(m)} + v_{r(m)} - w_k\right)^2 \right) . \end{aligned} \tag{3.2}$$

The learning algorithm seeks settings for the $\mathbf{b}$, $\mathbf{v}$, and $\mathbf{w}$ vectors that are consistent with the observed social data and initial *seed features*. We have used a joint regularization parameter for the word scores, bullying scores, and victim scores, but it is easy to use separate parameters for each parameter vector. We found in our experiments that the learner is not very sensitive to these hyperparameters, so we use a single parameter λ for simplicity. We constrain the seed features to have a high score and minimize Eq. (3.2), i.e.,

$$\min_{\mathbf{b}, \mathbf{v}, \mathbf{w}} \ J(\mathbf{b}, \mathbf{v}, \mathbf{w}; \lambda) \text{ s.t. } w_k = 1.0, \ \forall k : x_k \in S, \tag{3.3}$$

where S is the set of seed words. By solving for these parameters, we optimize the consistency of scores computed based on the participants in each social interaction as well as the vocabulary used in each interaction. Thus, we refer to this model as the participant-vocabulary consistency model.

3.2.1 Alternating Least Squares

The objective function in Eq. (3.2) is not jointly convex, but it is convex when optimizing each parameter vector in isolation. In fact, the form of the objective yields an efficient, closed-form minimization for each vector. The minimum for each parameter vector considering the others constant can be found by solving for their zero-gradient conditions. The solution for optimizing with respect to **b** is

$$\arg\min_{b_i} J = \frac{\sum\limits_{m \in M | s(m) = i} \left(\sum\limits_{k \in f(m)} w_k - |f(m)| v_{r(m)} \right)}{\lambda + \sum\limits_{m \in M | s(m) = i} |f(m)|}, \tag{3.4}$$

where the set $\{m \in M | s(m) = i\}$ is the set of messages that are sent by user i, and $|f(m)|$ is the number of n-grams in the message m. The update for the victim scores **v** is analogously

$$\arg\min_{v_j} J = \frac{\sum\limits_{m \in M | r(m) = j} \left(\sum\limits_{k \in f(m)} w_k - |f(m)| b_i \right)}{\lambda + \sum\limits_{m \in M | r(m) = j} |f(m)|}, \tag{3.5}$$

where the set $\{m \in M | r(m) = j\}$ is the set of messages sent to user j. Finally, the update for the **w** vector is

$$\arg\min_{w_k} J = \frac{\sum\limits_{m \in M | k \in f(m)} \left(b_{r(m)} + v_{s(m)} \right)}{\lambda + |\{m \in M | k \in f(m)\}|}, \tag{3.6}$$

where the set $\{m \in M | k \in f(m)\}$ is the set of messages that contain the kth feature or n-gram.

Each of these minimizations solves a least-squares problem, and when the parameters are updated according to these formulas, the objective is guaranteed to decrease if the current parameters are not a local minimum. Since each formula of the **b**, **v**, and **w** vectors does not depend on other entries within the same vector, each full vector can be updated in parallel. Thus, we use an alternating least-squares optimization procedure, summarized in Algorithm 1, which iteratively updates each of these vectors until convergence.

Algorithm 1 outputs the bully and victim score of all the users and the bullying-vocabulary score of all n-grams. Let $|M|$ be the total number of messages and $|W|$ be the total number of n-grams. The time complexity of each alternating least-squares update for the bully score, victim score, and word score is $O(|M| \cdot |W|)$. No extra space is needed beyond the storage of these vectors and the raw data. Moreover, sparse matrices can be used to perform the indexing necessary to compute these updates efficiently and conveniently, at no extra cost in storage, and the algorithm can be easily implemented using high-level, optimized sparse matrix libraries. E.g., we use `scipy.sparse` for our implementation.

Algorithm 1 Participant-Vocabulary Consistency using Alternating Least Squares

```
procedure PVC($b, v, w, \lambda$)
    Initialize $\mathbf{b}$, $\mathbf{v}$, and $\mathbf{w}$ to default values (e.g., 0.1).
    while not converged do
        $\mathbf{b} = \left[\arg\min_{b_i} J\right]_{i=1}^{n}$        ▷ update b using Eq. (3.4)
        $\mathbf{v} = \left[\arg\min_{v_i} J\right]_{i=1}^{n}$        ▷ update v using Eq. (3.5)
        $\mathbf{w} = \left[\arg\min_{w_k} J\right]_{k=1}^{|\mathbf{w}|}$        ▷ update w using Eq. (3.6)
    return $(\mathbf{b}, \mathbf{v}, \mathbf{w})$        ▷ return the final bully, victim score of users and the score of n-grams
```

3.3 Experiments

We apply participant-vocabulary consistency to detect harassment-based bullying in three social media data sets, and we measure the success of weakly supervised methods for detecting examples of cyberbullying and discovering new bullying indicators. We collect a dictionary of offensive language listed on NoSwearing.com [179]. This dictionary contains 3,461 offensive unigrams and bigrams. We then compare human annotations against PVC and baseline methods for detecting cyberbullying using the provided weak supervision. We also compare each method's ability to discover new bullying vocabulary, using human annotation as well as cross-validation tests. Finally, we perform qualitative analysis of the behavior of PVC and the baselines on each data set.

To set the PVC regularization parameter λ, we use three-fold cross-validation; i.e., we randomly partition the set of collected offensive words into three complementary subsets, using each as a seed set in every run. We do not split the user or message data since they are never directly supervised. For each fold, we use one third of these terms to form a seed set for training. We refer to the remaining held-out set of offensive words in the dictionary as *target words*. (The target words include bigrams as well, but for convenience we refer to them as target words throughout.) We measure the average area under the receiver order characteristic curve (AUC) for target-word recovery with different values of λ from 0.001 to 20.0. The best value of λ should yield the largest AUC. The average AUC we obtain using these values of λ for three random splits of our Twitter data (described below) ranged between 0.905 and 0.928, showing minor sensitivity to this parameter. Based on these results, we set $\lambda = 8$ in our experiments, and for consistency with this parameter search, we run our experiments using one of these random splits of the seed set. Thus, we begin with just over a thousand seed phrases, randomly sampled from our full list.

3.3.1 Data Processing

Ask.fm, Instagram, and Twitter are reported to be key social networking venues where users experience cyberbullying [28, 73, 217]. Our experiments use data from these sources.

We collected data from **Twitter**'s public API. Our process for collecting our Twitter data set was as follows: (1) Using our collected offensive-language dictionary, we extracted tweets containing these words posted between November 1, 2015, and December 14, 2015. For every curse word, we extracted 700 tweets. (2) Since the extracted tweets in the previous step were often part of a conversation, we extracted all the conversations and reply chains these tweets were part of. (3) To avoid having a skewed data set, we applied snowball sampling to expand the size of the data set, gathering tweets in a wide range of topics. To do so, we randomly selected 1,000 users; then for 50 of their followers, we extracted their most recent 200 tweets. We continued expanding to followers of followers in a depth-10 breadth-first search. Many users had small follower counts, so we needed a depth of 10 to obtain a reasonable number of these background tweets.

We filtered the data to include only public, directed messages, i.e., @-messages. We then removed all retweets and duplicate tweets. After this preprocessing, our Twitter data contains 180,355 users and 296,308 tweets. Once we obtained the conversation structure, we then further processed the message text, removing emojis, mentions, and all types of URLs, punctuation, and stop words.

We used the **Ask.fm** data set collected by Hosseinmardi et al. [110]. On Ask.fm, users can post questions on public profiles of other users, anonymously or with their identities revealed. The original data collection used snowball sampling, collecting user profile information and a complete list of answered questions. Since our model calculates the bully and victim scores for every user, it does not readily handle anonymous users, so we removed all the question-answer pairs where the identity of the question poster is hidden. Furthermore, we removed question-answer pairs where users only post the word "thanks" and nothing else, because this was extremely common and not informative to our study. Our filtered data set contains 260,800 users and 2,863,801 question-answer pairs. We cleaned the data by performing the same preprocessing steps as with Twitter, as well as some additional data cleaning such as removal of HTML tags.

We used the **Instagram** data set collected by Hosseinmardi et al. [111], who identified Instagram user IDs using snowball sampling starting from a random seed node. For each user, they collected all the media the user shared, users who commented on the media, and the comments posted on the media. Our Instagram data contains 3,829,756 users and 9,828,760 messages.

3.3.2 Baselines

Few alternate approaches have been established to handle weakly supervised learning for cyberbullying detection. The most straightforward baseline is to directly use the weak supervision to detect bullying, by treating the seed key-phrases as a search query.

To measure the benefits of PVC's learning of user roles, we compare against a method that extracts participant and vocabulary scores using only the seed query. For each user, we compute a bullying score as the fraction of outgoing messages that contain at least one seed term over all messages sent by that user and a victim score as the fraction of all incoming messages that contain at least one seed term over all messages received by that user. For each message, the participant score is the summation of the sender's bullying score and the receiver's victim score. We also assign each message a vocabulary score computed as the fraction of seed terms in the message. As in PVC, we sum the participant and vocabulary scores to compute the score of each message. We refer to this method in our results as the *naive participant* method.

We also compare against existing approaches that expand the seed query. This expansion is important for improving the recall of the detections, since the seed set will not include new slang or may exclude indicators for forms of bullying the expert annotators neglected. The key challenge in what is essentially the expansion of a search query is maintaining a high precision as the recall is increased. We compare PVC to two standard heuristic approaches for growing a vocabulary from an initial seed query. We briefly describe each below.

Co-occurrence (CO) returns any word (or n-gram) that occurs in the same message as any of the seed words. It extracts all messages containing any of the seed words and considers any other words in these messages to be relevant key-phrases. All other words receive a score of 0. We should expect co-occurrence to predict a huge number of words, obtaining high recall on the target words but at the cost of collecting large amounts of irrelevant words.

Dynamic query expansion (DQE) is a more robust variation of co-occurrence that iteratively grows a query dictionary by considering both co-occurrence and frequency [198]. We use a variation based on phrase relevance. Starting from the seed query, DQE first extracts the messages containing seed phrases; then for every term in the extracted messages, it computes a relevance score (based on [139]) as the rate of occurrence in relevant messages: $\text{relevance}(w_i, d, D) = |d \in D : w_i \in d| / |D|$, where $|D|$ indicates the number of documents with at least one seed term. Next, DQE picks k of the highest-scoring keywords for the second iteration. It continues this process until the set of keywords and their relevance scores become stable. Because DQE seeks more precise vocabulary expansion by limiting the added words with a parameter k, we expect it to be a more precise baseline, but in the extreme, it will behave similarly to the co-occurrence baseline. In our experiments, we use $k = 4{,}000$, which provides relatively high precision at the cost of relatively low recall.

3.3.3 Human Annotation Comparisons

The first form of evaluation we perform uses post-hoc human annotation to rate how well the outputs of the algorithms agree with annotator opinions about bullying. We enlisted crowdsourcing workers from Amazon Mechanical Turk, restricting the users to Mechanical Turk Masters located in the United States. We asked the annotators to evaluate the outputs of the three approaches from two perspectives: the discovery of cyberbullying relationships and the discovery of additional language indicators. First, we extracted the 100 directed user pairs most indicated to be bullying by each method. For the PVC and naive-participant methods, we averaged the combined participant and vocabulary scores, as in Eq. (3.1), of all messages from one user to the other. For the dictionary-based baselines, we scored each user pair by the concentration of detected bullying words in messages between the pair. Then we collected all interactions between each user pair in our data. We showed the annotators the anonymized conversations and asked them, “Do you think either user 1 or user 2 is harassing the other?” The annotators indicated either “yes,” “no,” or “uncertain.” We collected five

Table 3.1: Color-coded bullying bigrams detected in Ask.fm data by PVC and baselines. Terms are categorized according to the aggregate score of annotations. "Bullying" (2 or greater), "Likely Bullying" (1), "Uncertain" (0), and "Not Bullying" (negative) bigrams are shown in red, orange, gray, and blue, respectively.

	Detected Bullying Words Color-Coded by Annotation: Bullying, Likely Bullying, Uncertain, Not Bullying.
PVC	oreo nice, massive bear, bear c*ck, f*cking anus, ure lucky, f*g f*g, d*ck b*tch, ew creep, f*cking bothering, rupture, f*cking p*ssy, support gay, house f*ggot, family idiot, b*tch b*tch, p*ssy b*tch, loveeeeeee d*ck, f*cking c*nt, penis penis, gross bye, taste nasty, f*cking f*cking, dumb hoe, yellow attractive, b*tch p*ssy, songcried, songcried lika, lika b*tch, b*tch stupid, um b*tch, f*cking obv, nice butt, rate f*g, f*cking stupid, juicy red, soft juicy, f*cking d*ck, cm punk, d*ck p*ssy, stupid f*cking, gay bestfriend, eat d*ck, ihy f*g, gay gay, b*tch f*cking, dumb wh*re, s*ck c*ck, gay bi, fight p*ssy, stupid hoe
DQE	lol, haha, love, tbh, hey, yeah, good, kik, ya, talk, nice, pretty, idk, text, hahaha, rate, omg, xd, follow, xx, ty, funny, cute, people, cool, f*ck, best, likes, ily, sh*t, beautiful, perfect, girl, time, going, hot, truth, friends, lmao, answers, hate, ik, thoughts, friend, day, gonna, ma, gorgeous, anon, school
CO	bby, ana, cutie, ikr, ja, thnx, mee, profile, bs, feature, plz, age, add, pls, wat, ka, favourite, s*cks, si, pap, promise, mooi, hii, noo, nu, blue, ben, ook, mn, merci, meh, men, okk, okayy, hbu, zelf, du, dp rate, mooie, fansign, english, best feature, basketball, meisje, yesss, tyy, shu, een, return, follow follow

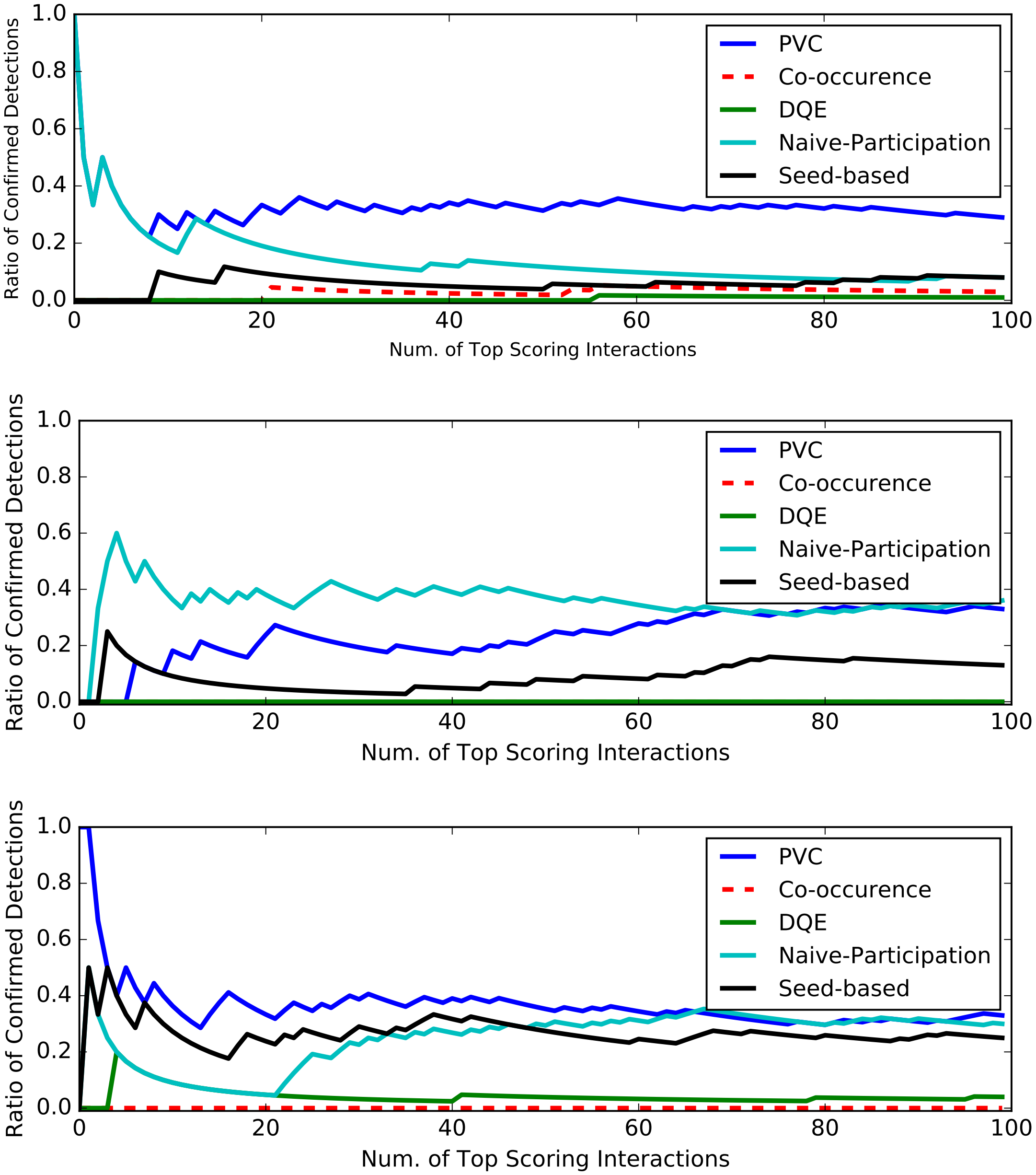


Figure 3.2: Precision@k for bullying interactions on Ask.fm (top), Instagram (middle), and Twitter (bottom).

Table 3.2: Color-coded bullying bigrams detected in Instagram data by PVC and baselines

	Detected Bullying Words Color-Coded by Annotation: Bullying, Likely Bullying, Uncertain, Not Bullying.
PVC	b*tch yas, yas b*tch, b*tch reported, *ss *ss, treated ariana, kitty warm, warm kitty, chicken butt, happy sl*t, jenette s*cking, kitty sleepy, follower thirsty, ariana hope, *ss b*tch, tart deco, sleepy kitty, hatejennette, *ss hoe, b*tch b*tch, sl*t hatejennette, pays leads, deco, happy kitty, fur happy, black yellow, bad *ss, bad b*tch, yellow black, pur pur, kitty pur, black black, d*ck b*tch, boss *ss, b*tch s*ck, soft kitty, nasty *ss, kitty purr, stupid *ss, *sss *ss, stupid b*tch, puff puff, bad bad, b*tch *ss, *ss foo, d*ck *ss, ignorant b*tch, hoe hoe, *ss bio, nasty b*tch, big d*ck
DQE	love, lol, cute, omg, beautiful, haha, good, nice, amazing, pretty, happy, wow, awesome, great, cool, perfect, best, guys, day, time, hahaha, gorgeous, god, pic, girl, people, birthday, tttt, life, man, follow, hair, lmao, hot, yeah, going, happy birthday, wait, better, hope, picture, baby, hey, sexy, ya, damn, sh*t, work, adorable, f*ck
CO	hermoso, sdv, sigo, troco, meu deus, troco likes, lindaaa, eu quero, fofo, perfect body, kinds, music video, girls love, allow, lls, spray, shoulders, wait guys, jet, niners, good sh*t, wie, damnnn, garden, post comments, stalk, rail, captain, belieber, sweety, convo, orders, smash, hahaha true, good girl, spider, au, best night, emotional, afternoon, gallery, degrees, hahahahahahah, oui, big time, por favor, beautiful photo, artwork, sb, drooling

annotations per conversation.

Second, we asked the annotators to rate the 1,000 highest-scoring terms from each method, excluding the seed words. These represent newly discovered vocabulary the methods believe to be indicators of harassment. For co-occurrence, we randomly selected 1,000 co-occurring terms among the total co-occurring phrases. We asked the annotators, "Do you think use of this word or phrase is a potential indicator of harassment?" We collected three annotations per key-phrase.

In Fig. 3.2, we plot the precision@k of the top 100 interactions for each data set and each method. The precision@k is the proportion of the top k interactions returned by each method that the majority of annotators agreed seemed like bullying. For each of the five annotators, we score a positive response as +1, a negative response as -1, and an uncertain response as 0. We sum these annotation scores for each interaction, and we consider the interaction to be harassment if the score is greater than or equal to 3. In the Ask.fm data, PVC significantly dominates the other methods for all thresholds. On the Twitter data, PVC is better than baselines until approximately interaction 70, when it gets close to the performance of the naive-participant baseline. In the Instagram data, PVC is below the precision of the naive-participation score until around interaction 65, but after that it improves to be the same as naive-participant. Co-occurrence, while simple to implement, appears to expand the dictionary too liberally, leading to very poor precision. DQE expands the dictionary more selectively, but still leads to worse precision than using the seed set alone.

In Fig. 3.3, we plot the precision@k for indicators that the majority of annotators agreed

Table 3.3: Color-coded bullying bigrams detected in Twitter data by PVC and baselines

	Detected Bullying Words Color-Coded by Annotation: Bullying, Likely Bullying, Uncertain, Not Bullying.
PVC	singlemost biggest, singlemost, delusional prick, existent *ss, biggest jerk, karma bites, hope karma, jerk milly, rock freestyle, jay jerk, worldpremiere, existent, milly rock, milly, freestyle, *ss b*tch, d*ck *ss, *ss hoe, b*tch *ss, adore black, c*mming f*ck, tgurl, tgurl sl*t, black males, rt super, super annoying, sl*t love, bap babyz, love rt, f*ck follow, babyz, jerk *ss, love s*ck, hoe *ss, c*nt *ss, *ss c*nt, stupid *ss, bap, karma, *ss *ss, f*ggot *ss, weak *ss, bad *ss, nasty *ss, lick *ss, d*ck s*cker, wh*re *ss, ugly *ss, s*ck *ss, f*ck *ss,
DQE	don, lol, good, amp, f*ck, love, sh*t, ll, time, people, yeah, ve, man, going, f*cking, head, didn, day, better, free, ya, face, great, hey, best, follow, haha, big, happy, gt, hope, check, gonna, thing, nice, feel, god, work, game, doesn, thought, lmao, life, c*ck, help, lt, play, hate, real, today,
CO	drink sh*tfaced, juuust, sh*tfaced tm4l, tm4l, tm4l br, br directed, subscribe, follow check, music video, check youtube, checkout, generate, comment subscribe, rt checkout, ada, follback, marketing, featured, unlimited, pls favorite, video rob, beats amp, untagged, instrumentals, spying, download free, free beats, absolutely free, amp free, free untagged, submit music, untagged beats, free instrumentals, unlimited cs, creative gt, free exposure, followers likes, music chance, soundcloud followers, spying tool, chakras, whatsapp spying, gaming channel, telepaths, telepaths people, youtube gaming, dir, nightclub, link amp, mana

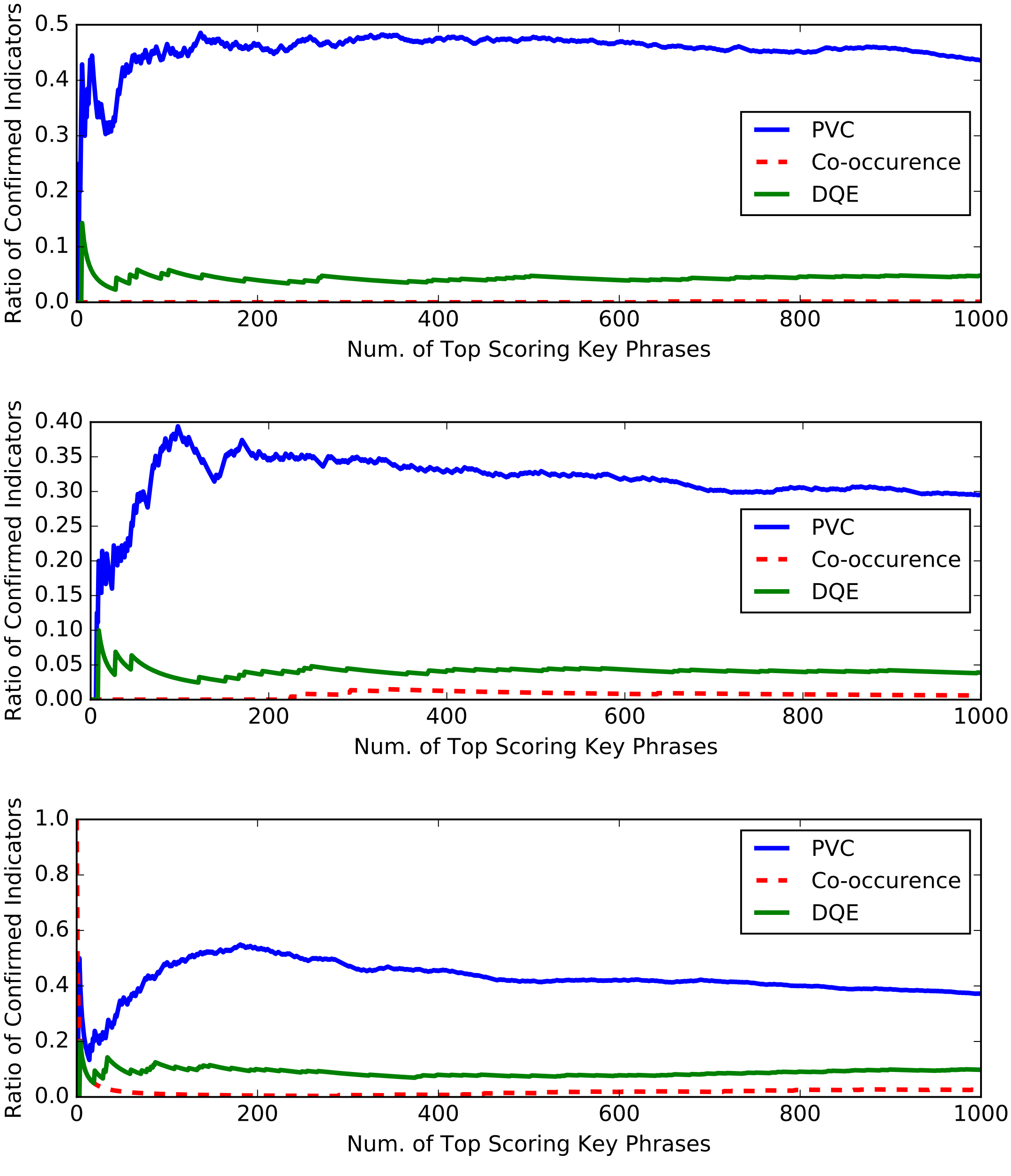


Figure 3.3: Precision@k for bullying phrases on Ask.fm (top), Instagram (middle), and Twitter (bottom).

were indicators of bullying. On all three data sets, PVC detects bullying words significantly more frequently than the two baselines, again demonstrating the importance of the model's simultaneous consideration of the entire communication network.

It is useful to note that the performance of the algorithm is directly affected by the quality and quantity of seed words. A better hand-picked seed set will result in higher precision as our model is founded based on this set. If the number of indicator words in the seed set increases, we expect increased recall but decreased precision. Adding a poor indicator word to the seed set may result in reducing both precision and recall, because the algorithm may identify non-bullying conversations as bullying, and consequently increasing the false positive rate. Moreover, by filtering the seed set, it is possible to focus PVC on particular topics of bullying.

3.3.4 Qualitative Analysis

We analyzed the 1,000 highest-scoring, non-seed terms produced by PVC, DQE, and co-occurrence and categorized them based on the annotations. Table 3.1, Table 3.2, and Table 3.3 list the first 50 words (censored) for Ask.fm, Instagram, and Twitter. The words are color-coded. Using a scoring system of +1 for when an annotator believes the word is a bullying indicator, 0 when an annotator is uncertain, and -1 when an annotator believes the word is not a bullying indicator, we print a word in red if it scored 2 or greater, orange if it scored a 1, gray if it scored a 0, and blue if it scored any negative value. These newly detected indicators suggest that PVC is capable of detecting new offensive words and slang (shown in red).

User1: lmao don't call me a b*tch. I don't know you, the tweet was just funny, b*tch."
User2: then you @ her and not me you little instigating *ss irrelevant hoe. Run along, b*tch.
User1: hy you mad? Lmao you're irrelevant as f*ck, b*tch. You can get out of my mentions you're a piece of sh*t.
User2: When dumb random *ss irrelevant hoes mention me, they get response. Now get your c*nt *ss on somewhere bruh '

User1: IS A FAKE *SS B*TCH WHO DOESNT DESERVE A MAN'
User2: b*tch you gave a BJ to a manager at McDonalds so you could get a free BigMac'
User1: B*TCH YOU SRE CONFUSING ME WITH YOUR MOTHER'
User2: YOUR MOM HAS BEEN IN THE PORN INDUSTRY LONGER THAN I HAVE BEEN ALIVE'
User1: B*TCH TAKE THAT BACK'
User2: TAKE WHAT BACK?
User1: YOUR RUDE DISRESPECTFUL COMMENTS'
User2: I ONLY SEE THE TRUTH YOU HOE'
User1: TBH DONT GET ME STARTED. I CAN BREAK YOU IN TWO MINUTES.'
User2: DO IT B*TCH, YOUR FAT *SS WILL GET TIRED OF TYPING IN 1 MINUTE'

Figure 3.4: Two examples of true positives by PVC that other baselines were not be able to detect. These examples are clearly intense and toxic, but their concentration of obvious swear words may not be high enough for the baseline approaches to identify.

User1: You don't get to call me stupid for missing my point."
User2: I said you're being stupid, because you're being stupid. Who are you to say who gets to mourn whom? Read the link.
User1: You miss my point, again, and I'm the stupid one? Look inwards, f*ckwad.

User1: Stupid she doesnt control the show she cant put it back on you idiot
User1: She isnt going to answer you stupid
User1: Its spelled Carly stupid
User1: She wont answer you stupid

Figure 3.5: Examples of harassment detected by PVC and verified by annotators. These examples do not have very obvious offensive-language usage, so methods beyond simple query-matching may be necessary to find them.

User1: LOL he's so nasty, he has a banana shaped faced
User2: Lmao . No comment
User1: L**** is a whore .
User2: Yeah, I'm over him so he can go whore around on his new girlfriend

User1: your f*cking awesome then (:
User2: damn f*cking right < 333333
User1: pretty good looking and seem cool (:
User2: i seem f*cking awesome

User1: if they act like hoes then they getting called hoes'
User2: that's awful thing to say
User1: what! it's true so if a girls a hoe, acts like a hoe and cheats like a hoe she isn't a hoe?
User2: and what exactly is a hoe? And what do you call men who cheat? Hoes too?'
User1: lets just end this conversation because I feel like you're gonna block me soon and I'd rather not lose another friend
User2: no, I mean, if you can say "if she act like a hoe then she gets called a hoe" then I would like to know what a hoe is'
User1: could mean wh*re, could imply she sleeps around, could mean she's just a evil f*ck face that flirts with you and then goes

User1: are u a vegetarian
User2: my parents are rude and wont let me but i dont like meat rip
User1: same dont like it that much but cant live without chicken :/
User2: i hate chicken what
User1: chicken is lyf wyd :/
User2: the only good thing about chicken are nuggets :/
User1: im not demanding i love all shapes and sizes : /
User2: chicken is gross :/

Figure 3.6: Examples of false positives by PVC: interactions identified by PVC that annotators considered non-harassment and appear correctly labeled. These examples include usage of offensive language but may require sophisticated natural language processing to differentiate from harassing usage.

We inspected the interactions PVC identified in the three datasets and found three categories of note. (1) The first type were represented by bullying conversations containing negative words identified by PVC, but not by other baselines. Two of such cases are shown in Fig. 3.4. (2) Some cases of conversations contained little prototypical bullying language, such as the slurs from the seed query and beyond. We hypothesize that PVC discovered these because of a combination of discovering new language and considering the typical roles of the conversation participants. In Fig. 3.5 we show two of these cases. (3) There were interactions that PVC mistakenly identified as harassment, where both we and the annotators consider the interactions be non-harassment. We grouped these false positives into four classes. First, one where users are talking about other people, not addressing each other in their messages. The false positives of this type reveal the importance of considering some keywords like "you," "you are," "your," etc. in our model. One example of such a case is illustrated at the top of Fig. 3.6. Second, some false positives occur when two users are joking with each other using offensive words, which is common among teenagers, for example. The second conversation in Fig. 3.6 is one such example. Third, false positives occur when two users have conversation about some negative topics, as shown in the third conversation in Fig. 3.6. In this example, the users are discussing sexual promiscuity, and while they are not necessarily being civil to each other, the conversation is not necessarily an example of harassment. Finally, a fourth common form of false positive occurs when no negative words are used in the conversation, but because PVC learned new words it believes to be offensive, it flags these conversations. One example is shown at the bottom of Fig. 3.6, where there is nothing particularly obvious about the conversation that should indicate harassment.

We analyze the sensitivity of PVC toward some often targeted groups such as those defined by race, gender, sexual orientation, and religion. Because of bias in social media across these groups, PVC will identify some messages containing keywords describing sensitive groups as bullying. These can be problematic because these words may often be used in innocuous contexts. We call these mistakes *unfair false positives*, meaning that non-bullying conversations containing sensitive keywords are falsely identified as bullying. Two of such cases are shown in Fig. 3.7, where these messages containing the keyword "black" may have been flagged because of their including the word. There might be two reasons why we observe these *unfair false positives*: i) sensitive key phrases describing target groups are included in the seed set, or ii) in the dataset, sensitive key phrases co-occur with seed words. We could address the first case by carefully choosing the seed set such that no sensitive key phrases are included; because otherwise we train the model to treat sensitive keywords as indicators, increasing the rate of unfair false positives. To address the second case, we should change our model by considering fairness in our objective function, which has been done in our last step in this research.

3.3.5 Bully and Victim Score Analysis

While our proposed model learns parameters that represent the tendencies of users to bully or to be victimized, it does not explicitly model the relationship between these tendencies. We can use the learned parameters to analyze this relationship. We plot users based on

User1: Owwww sexy
User1: Lets do that
User1: Black and yellow hard

User1: Beef Or Chicken ? Coke Or Pepsi ? White Or Black ? Mercedes Or BMW ? Friendship Or Love ? Yummy! Or Tasty ? Traditional Or Love Marriage ? Sister Or Brother ? Action Or Comedy ? Sweet Or Sour ? Chocolate Or Vanilla ? Strawberry Or Raspberry ? Lemon Or Orange ? Money Or Health?
User2: chicken’, ’ coke’, ’ grey;)’, ’ Mercedes’, ’ love’, ’ tasty’, ’ traditional’, ’ neither’, ’ comedy’, ’ sour’, ’ chocolate’, ’ raspberry’, ’ lemon’, ’ both probably:)

Figure 3.7: Two examples of non-bullying conversations mistakenly flagged by PVC containing the keyword “black”.

their bully and victim scores to observe the distributions of the bully and victim scores. We standardize the scores to be between 0 and 1, and in Fig. 3.9, we show the scatter plot of Twitter users according to their learned bully and victim scores as well as the heatmap plot (two-dimensional histogram) to see how dense the populations are in different regions of bully-victim space. The redder the region, the more users have bully and victim scores in the region. In the heatmap, we can see four hotspots: (1) pure bullies, seen as the horizontal cloud, (2) pure victims, seen as the vertical cloud, (3) victimized bullies, seen as the diagonal cloud, and finally, (4) the more dense hotspot is the region with low bully and victim scores. The existence of these hotspots suggests that most of the users in our Twitter data are not involved in bullying, but those that do have a fairly even mix of being bullies, victims, and both. The heatmap plot for Instagram and Ask.fm are also shown in Fig. 3.10. In Fig. 3.8 we show a sample of conversations involving a user with a high bully score (top) and a user with high victim score (bottom). The bully is sending toxic messages to multiple users, and they receive negative response messages as well. The victim, on the other hand, is targeted by three different apparent bullies.

There are also some examples of false positive cases where users are learned to have high bully and victim scores. In one case from Ask.fm shown in Fig. 3.11, a user receives many messages or a few long messages with many offensive words, but the message is not bullying. Instead, users appear to be using strong language in a positive manner.

3.3.6 Bully and Victim Network Analysis

We computed the average in-degree and out-degree of the top 3,000 bullies and victims as shown in Table 3.4. According to the statistics: 1) the in-degree of these top bullies is less than the in-degree of top victims; 2) the in-degree of top bullies is less or equal than their out-degree; and 3) the in-degree of top victims is greater than or equal to their out-degree. These trends suggest that, on average, high-scoring victims receive more messages than high-scoring bullies; they also receive more messages than they send. In contrast, bullies are more senders of messages than receivers.

We also compute the number of top bully and top victim neighbors for each top bully and top victim. In our Twitter data, around 54.63% of top victims have one bully neighbor,

bully to User1: Bernie sanders sucks c*ck f*ggot
bully to User2: ice soccer sucks f*cking c*ck f*ggot
bully to User3: all your merchandise sucks f*cking c*ck f*ggot I will take the money
bully to User4: I'm pretty sure there was one where he sniff his finger after he burrows his finger up his cornhole
bully to User5: stupid *ss race baiting wigger you're not fooling anybody. Get cancer
User6 to bully: All them hood n*ggas busted all over your mom like Moses, letting all their people go wigger.
User2 to bully: chill beaner
User1 to bully: But I'm righ
bully to User1: get mad f*ggot
bully to User5: you're not black stop acting. You are race baiting f*ggot and your mother is a mudshark wh*re who let the whole squad hit it raw
User7 to bully I wish I looked like that still
bully to User7: you look like a f*cking beast
User7 to bully: u look like a f*ggot
User8 to bully: idk it's hard to believe he's actually that much of a loser lol
User8 to bully: lol all the sh*t she talked on smokey was probably her just going off what the trolls said
User8 to bully: well I got that n*ggas info, his moms u tryna have some fun? Lol
User8 to bully: aye u know about
User8 to bully: all the sh*t the trolls said ended up being true smh...

User1 to victim: s*ck a d*ck f*ggot
User2 to victim: If you don't show up in my mentions crying like a huge f*ggot then I won't call you the f*ggot word. F*ggot
User2 to victim: Why don't you cry about it, you huge f*ggot. Cry because somebody disagrees with you.
User3 to victim: African diseases would sicken you even more.
victim: to User2 WOOOO man you morons out are coming out of the woodworks today. You sicken me.

Figure 3.8: Conversation of users with high bully score (top) and high victim score (bottom). The user with a high bully score is sending toxic messages to other users; the user with a high victim score is receiving intense negative messages while responding to some messages reacting the bullies.

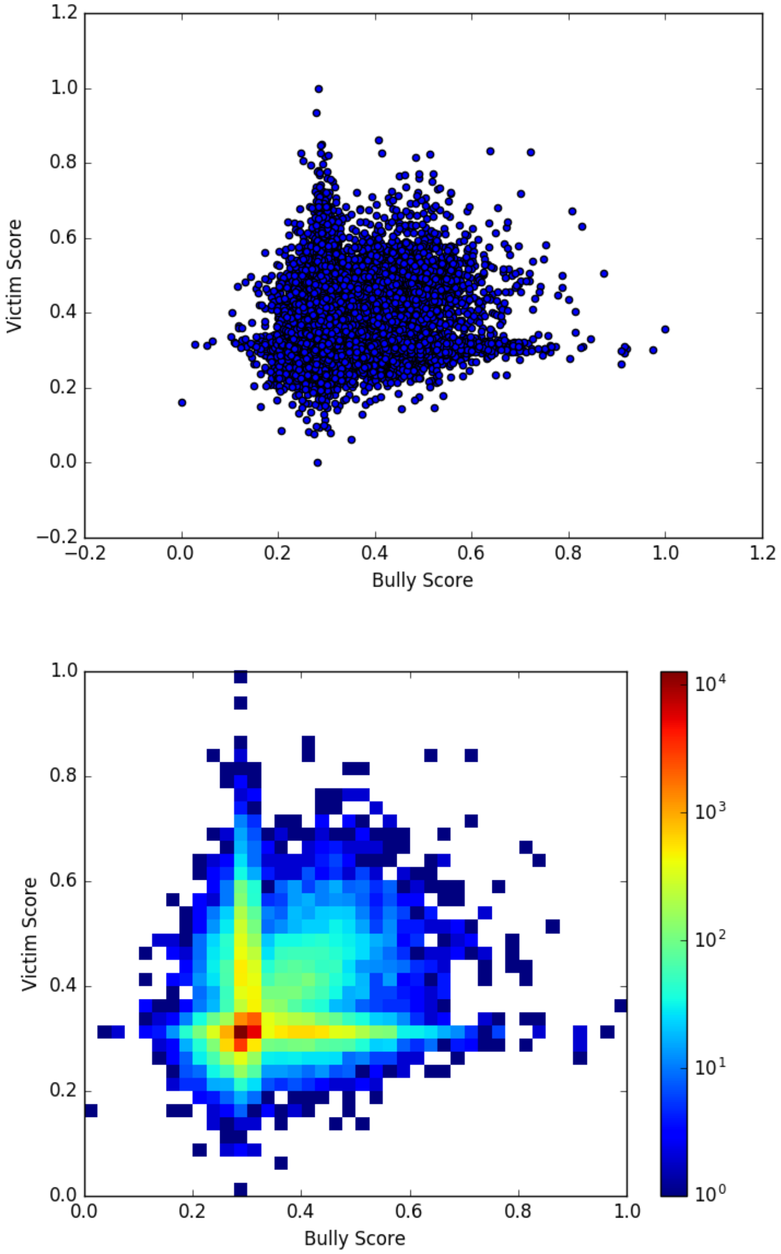


Figure 3.9: Scatter (top) and heatmap (bottom) plots for Twitter. The plots show the distributions of the bully and victim scores. According to the heatmap, most of the users in our Twitter data are not involved in bullying because most of the users occupy the redder region. However, there are moderate number of bullies, victims, and victimized bullies.

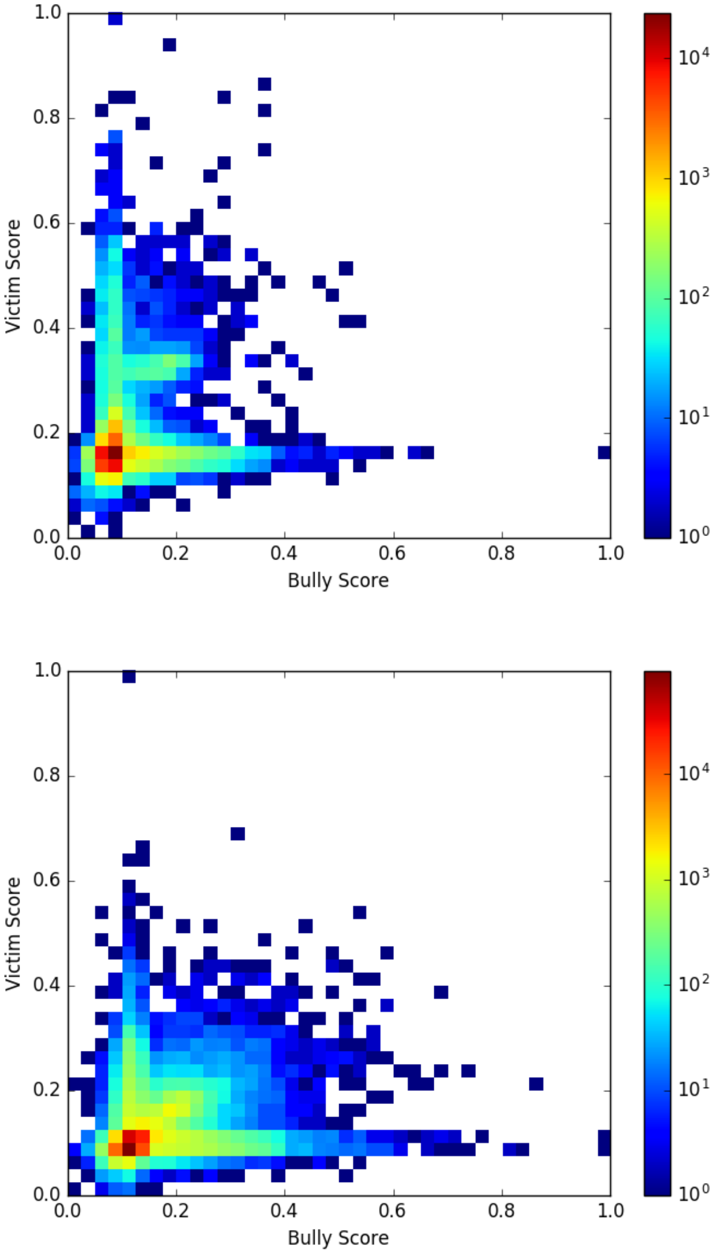


Figure 3.10: Heatmap plots for Instagram (top) and Ask.fm (bottom). In both datasets, users with low bully and low victim scores fall in the denser region (red color), but there are users with high bully and high victim scores to a lesser extent.

User1: I love you so much. I really do. I'm. Just. So. Lesbian. For. You. And were going to go ham when your outta summer school babe :*.
User2: Okay personal. I think you and I would be a boom ass couple.. But ya know
User1: I thought we were going to take it slow between us baby?
User2: Thoughts love?:)
User1: I f*cking love you

Figure 3.11: An example of a user with high victim score and low bully score. The user receives messages containing offensive words, but they are not used in an aggressive or harassing manner, and the user does not appear to be a victim.

while less than 1% of them have two bully neighbors. Around 4% of bullies have one victim neighbor, and less than 0.2% of them have two or three victim neighbors. On Ask.fm, around 4% of victims have one bully neighbor, while 0.5% of them have two or three bully neighbors. Around 2.5% of bullies are surrounded by one victim, while less than 0.8% of them are surrounded by two and three victims. On Instagram, however, the scale is much smaller. Only 14 victims (out of 3,000) are neighboring by at least one top bully. Less than 4% of bullies have one victim neighbor, and less than 1% of them have between two to five victim neighbors. In general, in most detected bullying cases, there is only one bully and one victim in the immediate social graph.

Table 3.4: The average in-degree and out-degree of top-scoring bullies and victims. On average, bullies are sending more messages than victims. They also send more messages than they receive, unlike victims who are more receivers of the messages than senders.

Average Score	Twitter	Ask.fm	Instagram
Average in-degree of bullies	1.081	6.578	0.011
Average out-degree of bullies	2.286	6.578	2.154
Average in-degree of victims	2.181	7.385	100.99
Average out-degree of victims	1.329	7.385	10.29

To gain a better intuition about the network structure among bullies and victims, we illustrate the communication graph among some detected bullies and victims. To extract the bullying subgraphs, we use a depth-2 snowball sampling starting from detected bully and victim users. Since each node might have hundreds of neighbors, we randomly select at most 10 neighbors and collect the subgraph of the second-hop neighbors subject to this random downsampling. Bullies, victims, and bystanders are shown in *red*, *green*, and *gray*, respectively. We illustrate different communities of such users in Instagram in Figure 3.12. In the figure, victims are surrounded by several bullies as well as bystanders. This pattern aligns with the idea that a victim could be targeted by a group of bullies. We also observe that, in most cases, bullies are indirectly connected to each other through bystanders or victims. Another interesting point is that not all of a bully's neighbors are victims. In other words, a bully could interact with different people, but they may not bully all of them.

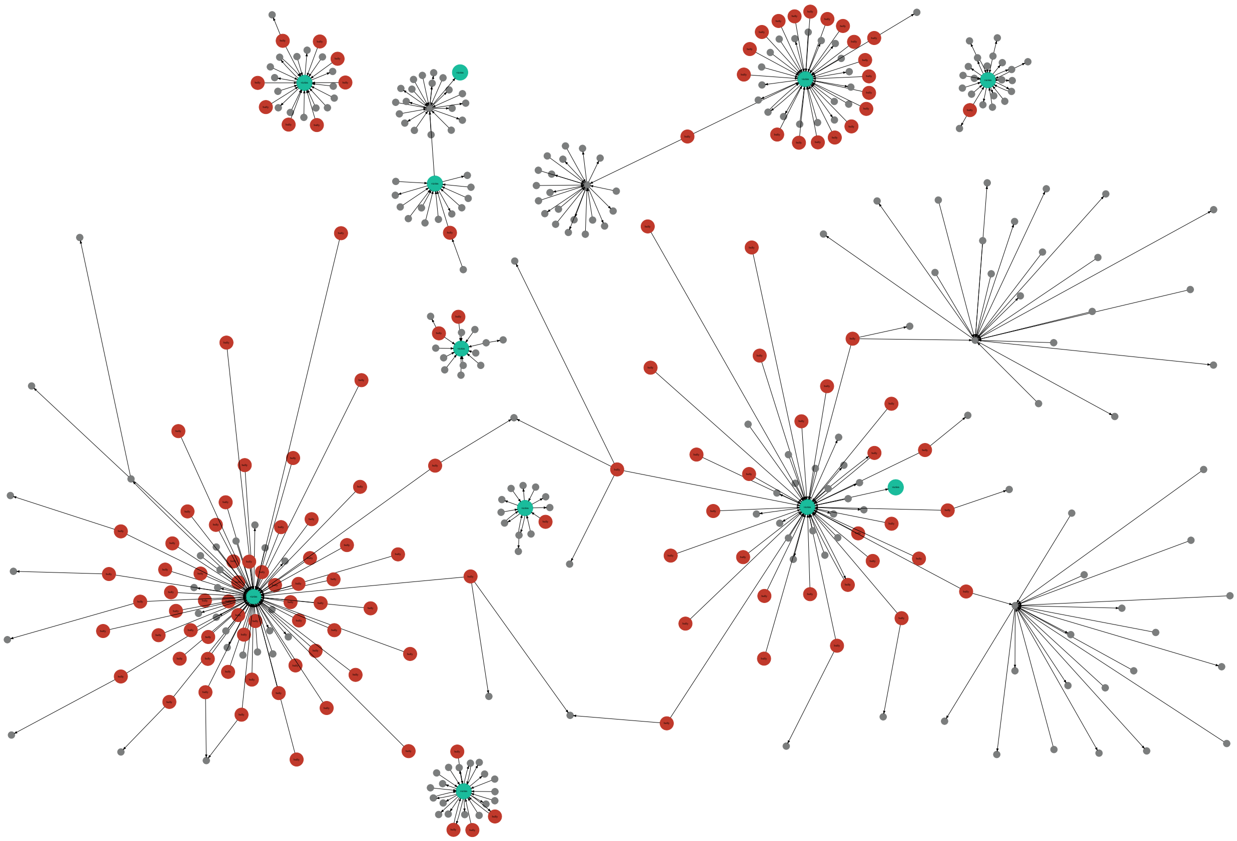


Figure 3.12: A sub-graph of bullies and victims in Instagram. *Red*, *green*, and *gray* represent bullies, victims, and bystanders, respectively. Victims have several bully and bystander neighbors. In addition, a bully may interact with different people, but they may not bully all of them.

In Figure 3.13, we show two sub-graphs from Ask.fm. In the top network, there are two bullies and three victims. The bully at the center has three victim neighbors, and one bully neighbor, showing they might have harsh conversations containing indicator key phrases. In the bottom network, the bully user has two victim neighbors. One victim is interacting with their neighbors, while the other one only communicates with the bully user. In general, there are varieties of structures among users: Victims who are bullied by multiple users, bullies who targeting some of their neighbors but not all of them; bullies and victims with many neighbors; or bullies and victims with one neighbor. Examples of these patterns exist within the subgraphs in Figs. 3.12 and 3.13.

3.3.7 Fairness Analysis

Social media data carries social bias against various demographic groups. Machine learning algorithms trained on this data, therefore, perpetuates this discrimination causing unfair decision-making. Our algorithms are trained on social data; therefore, they might encode society's bias across various demographic groups.

In Fig. 3.14, a conversation from Instagram is shown in which a user is harassing the other using sensitive keywords describing race and gender, in addition to offensive words. In this section, we examine how fair PVC is toward particular groups. First, we created a list of sensitive keywords describing social groups of four types: race, gender, sexual orientation, and religion. We removed all sensitive keywords from our seed words. Then, we train the PVC based on fair seed words, then we sort the words according to their learned word scores. We then plot the computed word rank of the sensitive keywords. Figure 3.15 plots the rank of sensitive keywords for Twitter, Instagram, and Ask.fm. Out of 1,425,111 unigrams and bigrams on Twitter, "boy" has the top rank (870) among the sensitive keywords, while "latina" has the lowest rank (184,094). On Instagram and Ask.fm, the rank of the word "queer" is the highest, while the rank of "sikh" and "heterosexual" are the lowest (On Instagram, out of 3,569,295 words, rank of "queer" is 1,973, while rank of "sikh" is 209,189. On Ask.fm, among 1,960,977 words, ranks of "queer" and "heterosexual" words are 677 and 158,747, respectively). These numbers in the plot indicate that the sensitive keywords are spread, with a few appearing among the most indicative bullying words and others appearing much lower in the ranking. It is worth pointing out that only "boy" on Twitter and "queer" on Ask.fm have listed among the top 1,000 bullying phrases (refer to Fig. 3.3).

Among the selected sensitive keywords, the highest ranked words are *boy* when PVC is trained on Twitter and *queer* when trained on Instagram and Ask.fm. The second highest ranked in all of the datasets is *gay*. The third highest ranked word for Twitter is *black* and in Instagram and Ask.fm is *lesbian*. Comparing gendered words, the rank of *girl* for Twitter is lower than the rank of *boy*; while for both Instagram and Ask.fm, the rank of *girl* is higher than the rank of *boy*. We hypothesize that this happens because in our Twitter data, the word *boy* almost always co-occurs with many other offensive words. Instagram and Twitter are more biased toward *girl*. This trend is consistent with the relative rankings of *woman* versus *man*. Examining the top ten keywords, running PVC on the Twitter data results in four of them belonging to the gender category, two in each of the sexual orientation and race categories, and one in the religion category. Using Instagram data, six out of the ten top keywords describe sexual orientation, three describe gender, and one describes religion. Using Ask.fm, five of the highest ranked keywords are about sexual orientation, three are about race, and two are about gender. Overall, these results may indicate that our Twitter data is more biased about gender, while Instagram and Ask.fm are more biased about sexual orientation. The fact that *queer* appears among the highest ranked sensitive keywords may be a result of its history of being used as a slur that has been reclaimed by the LGBT community. While it is now generally accepted to be simply a descriptive word for a group of people, it is also

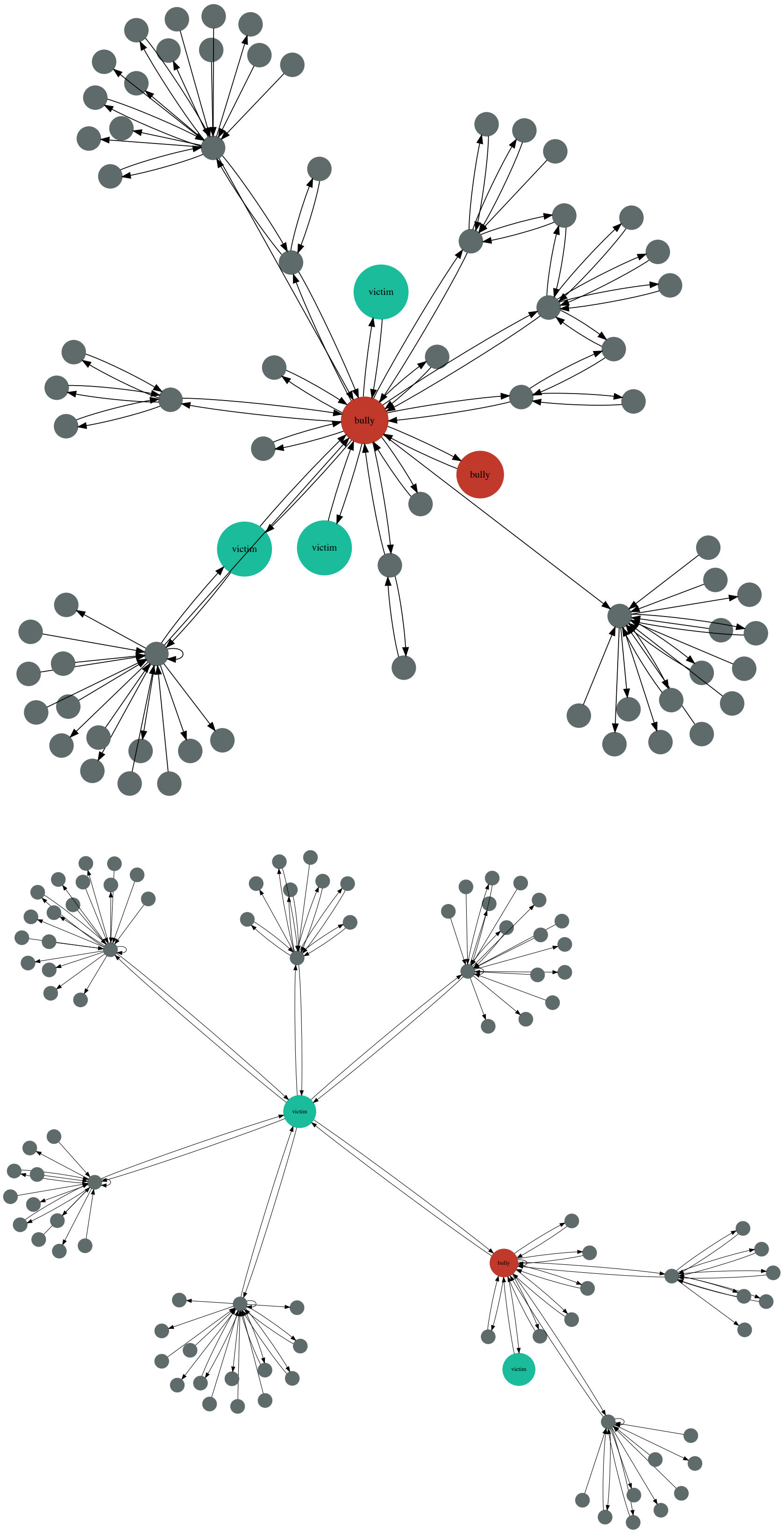


Figure 3.13: Two sub-graphs of bullies and victims in Ask.fm. *Red*, *green*, and *gray* represent bullies, victims, and bystanders, respectively. Top: a bully at the center has bully, victim, and neutral neighbors. Bottom: a bully has two victim neighbors; one of them has a large community of friends, while the other is isolated.

User1: U gay
User1: f*ck you spic
User1: Fagget ass typical spic bitch

User1: U gay
User1: - - is the gayest peoples I have met In my life
User1: - - - u r a f*ggot u can go suck ur dads cock u littles p*ssy f*ggot ur probably on 9 years old bitch IAM 14 dumb hoe f*ggot black bitch suck my cock itll go down ur throat and out of ur mouth u f*ggot black p*ssy
User1: YO BLACK BITCH SHUT YOUR LITTLE BLACK MOUTH ur black cousin can suck my cock too that little bitch probably couldnt fight for sh*t u little black MOTHER F*CKER WHY DONT U GO F*CK UR COUSIN U LITTLES BLACK P*SSYLET U CAN SUCK UR COUSINS DICK TOO BUT THAT SHIT WONT FIT IN YOUR BLACK LITTLE MOUTH I WILL F*CKING HACK UR LITTLE BLACK ASS AN MAKE U SUCK UR DADS DICK SO I SUGGEST U SHUT THE F*CK UP U LITTLE BLACK P*SSY FACE COCK SUCKIN BLACK BITCH SUCK MY DICK AND HAVE A NICE DAY and yo - - - u r unpredictibably retarded and black and suck ur dads an cousins cock u little black bitch
User1: gymnastics 18 lol
User1: - - - shut the f*ck up I will f*cking slap and beat the shit out of u dumbass black little hoe why dont u go f*ck ur cousin he will f*ck ur black butt crack u littles f*ggot and sorry to all the other black people out there these two r just really big d*ck faces

Figure 3.14: Two examples of bias toward race and sexual orientation in Instagram. In top example, bully is harassing victim using racist slur ("spic"). In the bottom example, the user is bullying the other one using negative words as well as sensitive keywords about race and sexual orientation ("black" and "gay").

still often used as a slur. Overall, these analyses provide some assurance that the learned PVC models are not overly reliant on these sensitive keywords, but more study is necessary, and we are planning future work with explicit fairness-encouraging learning objectives.

3.4 Conclusion

In this section, we proposed a weakly supervised method for detecting cyberbullying. Starting with a seed set of offensive vocabulary, participant-vocabulary consistency (PVC) simultaneously discovers which users are instigators and victims of bullying, and additional vocabulary that suggests bullying. These quantities are learned by optimizing an objective function that penalizes inconsistency of language-based and network-based estimates of how bullying-like each social interaction is across the social communication network. We ran experiments on data from online services that rank among the most frequent venues for cyberbullying, demonstrating that PVC can discover instances of bullying and new bullying language. In our quantitative analysis, we compute the precision using post-hoc human annotation to evaluate the detected conversations and key phrases by PVC. In our qualitative analysis, we examined discovered conversations, and classified them into some categories of note. Furthermore, we showed some statistics about bullies and victims as well as the distributions of bully and victim scores. We also showed the network structure between some bullies and victims to visualize the social relation between the two. Motivating our future work of developing methods to train fair language-based detectors, we tested the sensitivity and bias of PVC toward particular social groups.

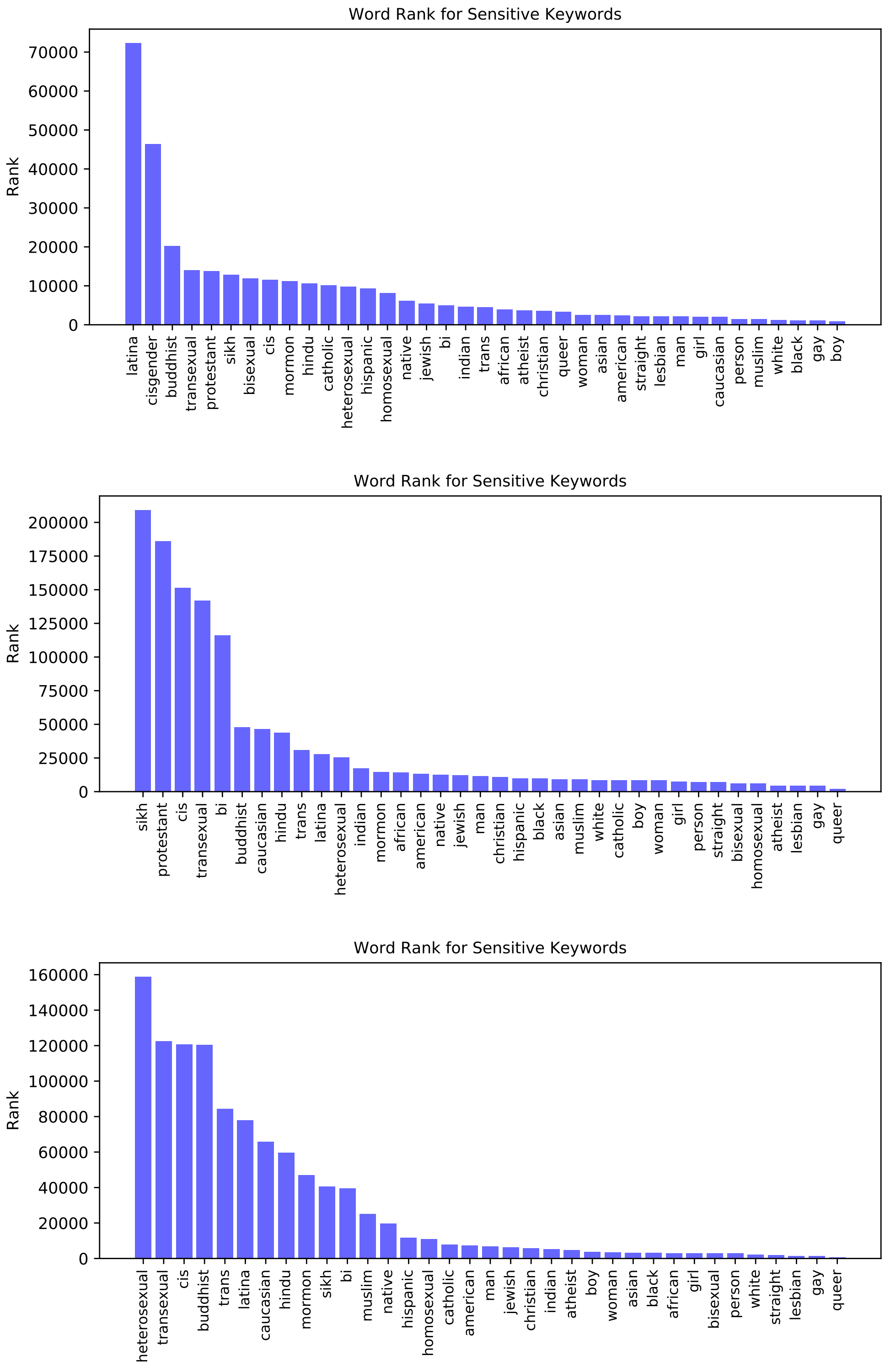


Figure 3.15: Sensitive keywords ranks obtained by PVC for Twitter (top), Instagram (middle), and Ask.fm (bottom). The highest word rank in Twitter, Instagram, and Ask.fm is 870 for the word "boy", 1,973, and 677, respectively for the word "queer".

Chapter 4

Co-trained Ensembles of Embedding Models

4.1 Introduction

In this chapter, we describe how *participant-vocabulary consistency (PVC)* is extended to benefit from deep embedding models. We propose *co-trained ensemble* framework, which consists of two learning algorithms that co-train one another, seeking consensus on whether examples in unlabeled data are cases of cyberbullying or not. One detector identifies bullying by examining the language content of messages; another detector considers the social structure to detect bullying. Training different learners on different perspectives of the problem aligns with the true multi-faceted nature of cyberbullying. Moreover, since the true underlying cyberbullying phenomenon is both linguistic and social, we should expect good models using each of these views to agree with each other, motivating our search for a consistency across the two perspectives.

Similar to *PVC*, *co-trained ensemble* framework is a weakly supervised algorithm. The weak supervision is in the form of expert-provided key phrases that are indicative of bullying or non-bullying. In *PVC* we used a small seed vocabulary of bullying indicators. In *co-trained ensemble*, other than bullying indicators, we use counter-indicators of bullying, which consist of positive-sentiment phrases such as “nice”, “thanks”, “beautiful”, etc.

We represent the language and users as vectors of real numbers with embedding models. For example, doc2vec [141] is a popular word-embedding model that represents documents with low-dimensional vectors (based on ideas from the word2vec per-word embedding [162, 163]). And node2vec [99] is a framework for building continuous feature representations for nodes in networks. We use language and user vectors as the input to language-based and user-based classifiers, respectively. We examine two strategies when incorporating vector representations of language. First, we use existing doc2vec [141] models as inputs to the learners. Second, we create new embedding models specifically geared for our specific task of harassment detection, which we train in an end-to-end manner during optimization of the model, incorporating the unsupervised doc2vec loss function into our co-training objective.

To train the model, we construct an optimization problem made up of a regularizer and two loss functions: a co-training loss that penalizes the disagreement between the deep language-based model and the deep user-based model, and a weak-supervision loss that is the classification loss on weakly labeled messages.

We evaluate our approach on data from Ask.fm, Twitter, and Instagram, which are three of the public-facing social media platforms with a high frequency of cyberbullying. We use two human-curated lists of key phrases indicative and counter-indicative of bullying as the weak supervision, and we assess the precision of detections by variations of the framework. We evaluate the effectiveness of our approach using post-hoc, crowdsourced annotation of detected conversations from the social media data. We quantitatively demonstrate that our weakly supervised deep models improve precision over a previously explored, non-deep variation of the approach.

4.2 Co-trained Ensembles

Our learning framework uses co-trained ensembles of weakly supervised detectors. In this section, we first describe them generally. Then we describe the specific instantiations we use in our experiments. A fundamental principle for our co-trained ensemble framework is the diversity of learners that look at the problem from different perspectives. Our framework trains two detectors; one detector identifies bullying incidents by examining the language content of messages; another detector considers social structure to discover bullying. To formally describe social media data, we consider a set of users U and a set of messages M. Each message $m \in M$ is sent from user $s(m)$ to user $r(m)$. In other words, the lookup functions s and r return the sender and receiver, respectively, of their input message. The input data takes on this form, with some of the messages annotated with weak supervision.

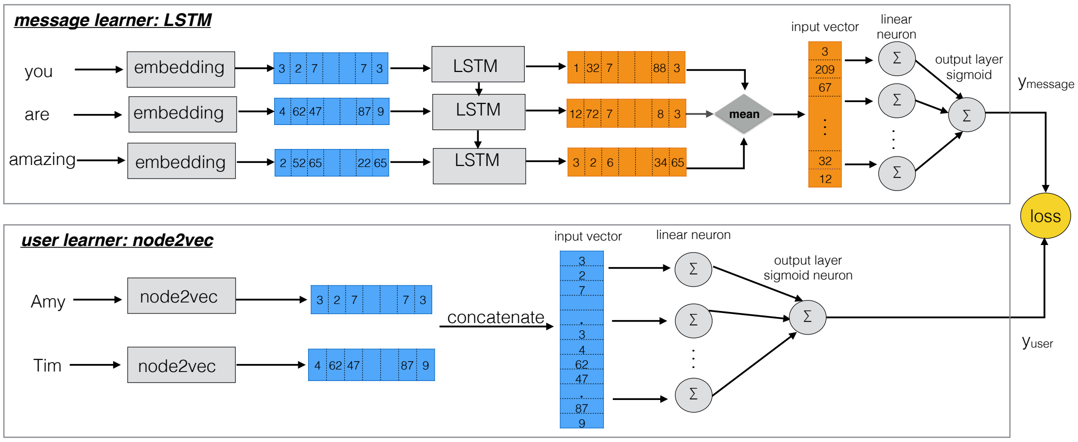


Figure 4.1: Diagram of the co-trained ensemble of the RNN message learner and the node2vec user learner.

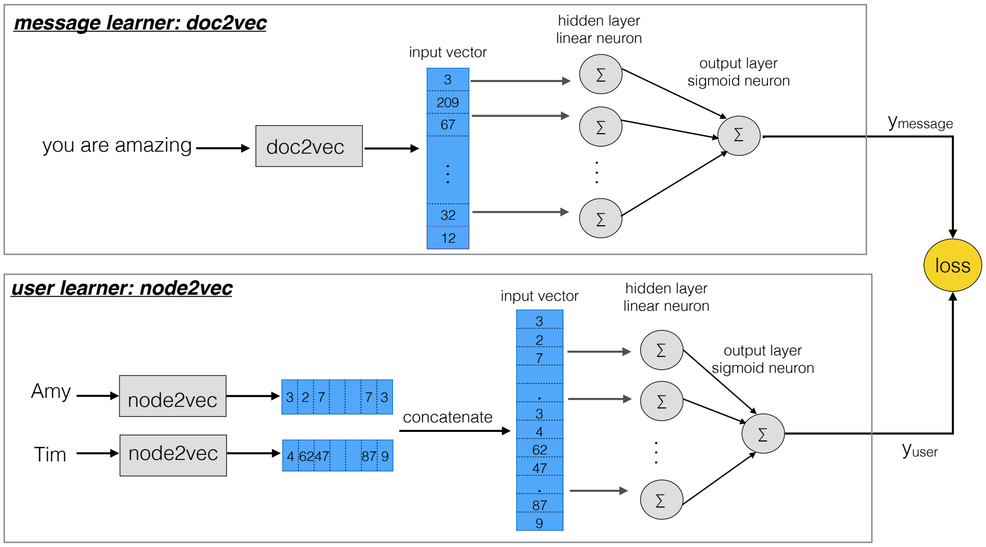


Figure 4.2: Diagram of the components in the co-trained ensemble of the doc2vec message learner and the node2vec user learner.

4.2.1 General Framework

We define two types of classifiers for harassment detection: message classifiers and user-relationship classifiers (or *user classifiers* for short). Message classifiers take a single message as input and output a classification score for whether the message is an example of harassment, i.e., $f : M \mapsto \mathbb{R}$. User classifiers take an ordered pair of users as input and output a score indicating whether one user is harassing the other user, i.e., $g : U^2 \mapsto \mathbb{R}$. For message classifiers, our framework accommodates a generalized form of a weakly supervised loss function ℓ (which could be straightforwardly extended to also allow full or partial supervision). Let Θ be the model parameters for the combined ensemble of both classifiers. The training objective is

$$\min_{\Theta} \underbrace{\frac{1}{2|M|}\sum_{m \in M}\left(f(m;\Theta) - g\left(s(m), t(m);\Theta\right)\right)^2}_{\text{consistency loss}} + \underbrace{\frac{1}{|M|}\sum_{m \in M}\ell\left(f(m;\Theta)\right)}_{\text{weak supervision loss}},$$

where the first loss function is a consistency loss, and the second loss function is the weak supervision loss.

The consistency loss penalizes the disagreement between the scores output by the message classifier for each message and the user classifier for the sender and receiver of the message.

The weak supervision loss relies on annotated lists of key-phrases that are indicative or counter-indicative of harassment. For example, various swear words and slurs are common indicators of bullying, while positive-sentiment phrases such as “thanks” are counter-indicators. It should be noted that counter-indicators were not included in our first approach, *PVC*, explained in previous chapter. Let there be a set of indicator phrases and a set of counter-indicator phrases for harassment. The weak supervision loss ℓ is based on the fraction of indicators and counter-indicators in each message, so for a message containing $n(m)$ total key-phrases, let $n^+(m)$ denote the number of indicator phrases in message m and $n^-(m)$ denote the number of counter-indicator phrases in the message. We bound the message learner by the fraction of indicator and counter-indicator key-phrases in the message:

$$\underbrace{\frac{n^+(m)}{n(m)}}_{\text{Lower Bound}} < y_m < \underbrace{1 - \frac{n^-(m)}{n(m)}}_{\text{Upper Bound}},$$

If this bound is violated, we penalize our objective function using weak supervision loss. The weak supervision loss is then

$$\ell(y_m) = -\log\left(\min\left\{1, 1 + (1 - \tfrac{n^-(m)}{n(m)}) - y_m\right\}\right) - \log\left(\min\left\{1, 1 + y_m - \tfrac{n^+(m)}{n(m)}\right\}\right).$$

This form of penalized bound is a generalization of the log-loss, or cross-entropy; in the rare cases that the weak supervision is completely confident in its labeling of a message being bullying or not, it reduces to exactly the log-loss.

4.2.2 Models

For the message classifiers, we use four learners:

(i) BoW: a randomly hashed bag-of-n-grams model with 1,000 hash functions [259],

(ii) doc2vec: a linear classifier based on the pre-trained doc2vec vector of messages trained on our dataset [141],

(iii) embedding: a custom-trained embedding model with each word represented with 100 dimensions,

(iv) RNN: a recurrent neural network—specifically a long-short term memory (LSTM) network—with two hidden layers of dimensionality 100.

The embedding and RNN models are trained end-to-end to optimize our overall loss function, and the vector-based models (BoW, doc2vec) are trained to only adjust the linear classifier weights given the fixed vector representations for each message.

For the user classifiers, we use a linear classifier on concatenated vector representations of the sender and receiver user nodes. To compute the user vector representations, we pre-train a node2vec [99] representation of the communication graph. Node2vec finds vector representations that organize nodes based on their network roles and communities they belong to. The pre-trained user vectors are then the input to a linear classifier that is trained during the learning process

There are eight combinations of message and user learners (including the option to use no user learner, in which case we are simply using the weak supervision loss to train the message classifiers). Figure 4.1 illustrates the model architecture for a co-trained ensemble of an RNN message learner and a node2vec user learner. Similarly, Figure 4.2 shows the model architecture for a co-trained ensemble of an doc2vec message learner and a node2vec user learner. The other possible combinations of message and user learners are analogously structured.

4.3 Experiments

Mirroring the setup initially used to evaluate PVC in Chapter 3, we construct our weak supervision signal by collecting a dictionary of 3,461 offensive key-phrases (unigrams and bigrams) [179]. We augment this with a list of positive opinion words collected by Hu & Liu [114]. The offensive phrases are our weak indicators and the positive words are our counter-indicators. We used three datasets in our experiments.

We use the same **Twitter** data explained in Chapter 3. The data was collected from Twitter's public API, by extracting tweets containing offensive-language words posted between November 1, 2015 and December 14, 2015. Then we extracted conversations and reply chains that included in these tweets. We then used snowball sampling to gather tweets in a wide range of topics. After some preprocessing, the Twitter data contains 180,355 users and 296,308 tweets.

Our **Ask.fm** dataset is the same as Ask.fm in previous chapter. we use a subsample of the Ask.fm dataset collected by Hosseinmardi et al. [110]. On Ask.fm, users can post questions on public profiles of other users, anonymously or with their identities revealed. The original data collection used snowball sampling, collecting user profile information and a complete list of answered questions. Since our model calculates the bully and victim scores for every user, it does not readily handle anonymous users, so we removed all the question-answer pairs where the identity of the question poster is hidden. Furthermore, we removed question-answer pairs where users only post the word "thanks" and nothing else, because this was extremely common and not informative to our study. Our filtered dataset contains 260,800 users and 2,863,801 question-answer pairs.

Our **Instagram** dataset in this chapter is a lighter version of Instagram in Chapter 3. We also use a subsample of the Instagram dataset collected by Hosseinmardi et al. [111] via snowball sampling. For each user, they collected all the media the user shared, users who commented on the media, and the comments posted on the media. We filter the data to remove celebrities and public-figure accounts. Our filtered Instagram data contains 656,376 users and 1,656,236 messages.

4.3.1 Precision Analysis

As in the evaluation from the previous chapter, we use post-hoc human annotation to measure how well the outputs of the algorithms agree with annotator opinions about bullying. We asked crowdsourcing workers from Amazon Mechanical Turk to evaluate the cyberbullying

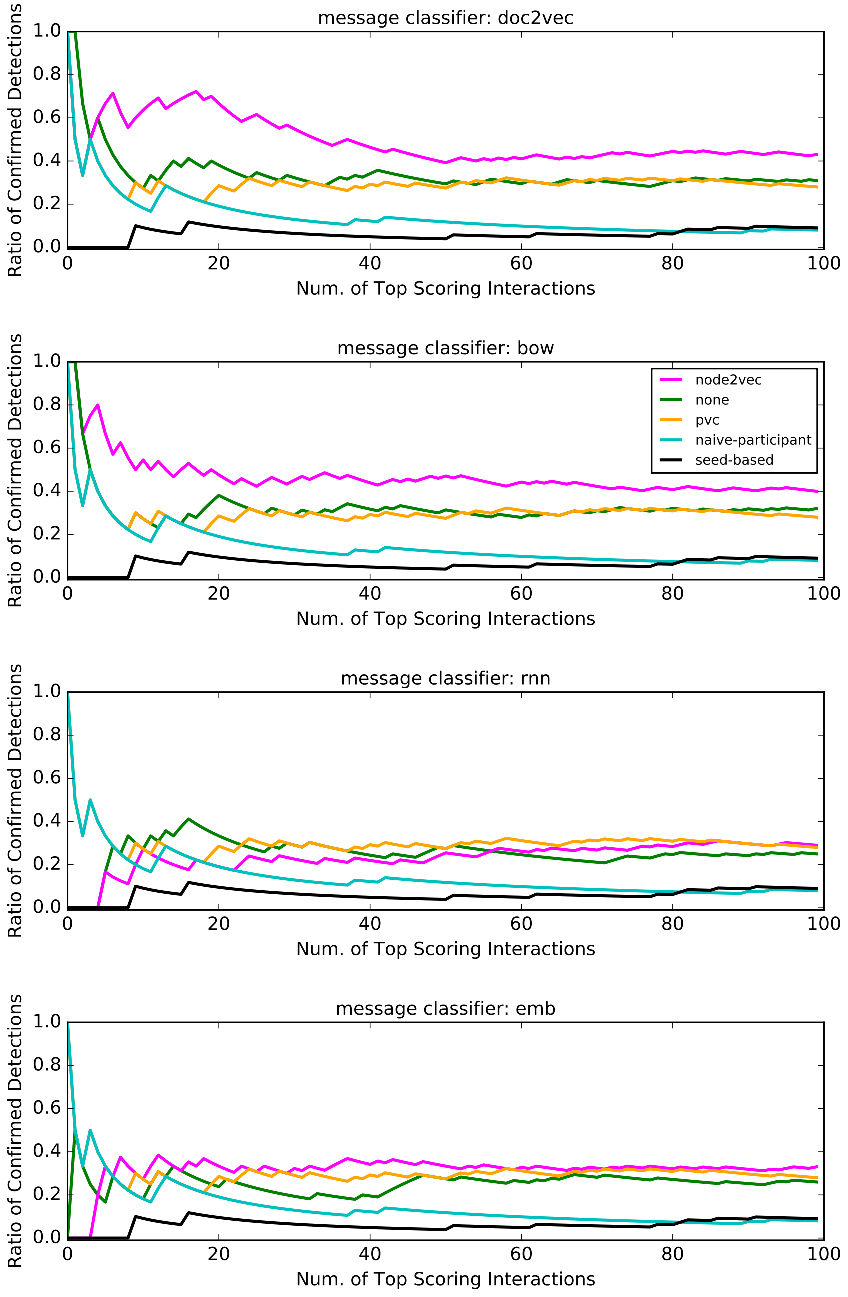


Figure 4.3: Precision@k for bullying interactions on Ask.fm data using the combination of message and user learners, PVC, seed-based, and naive-participant.

interactions discovered by all the methods. First, we averaged the user and message classification scores of each message. Then, we extracted the 100 messages most indicated to be bullying by each method. Finally, we collected the full set of messages sent between the sender and receiver of these messages. We showed the annotators the anonymized conversations and asked them, “Do you think either user 1 or user 2 is harassing the other?” The annotators indicated either “yes,” “no,” or “uncertain.” We collected five annotations per conversation.

In Fig. 4.4, we plot the precision@k of the top 100 interactions for all the combinations

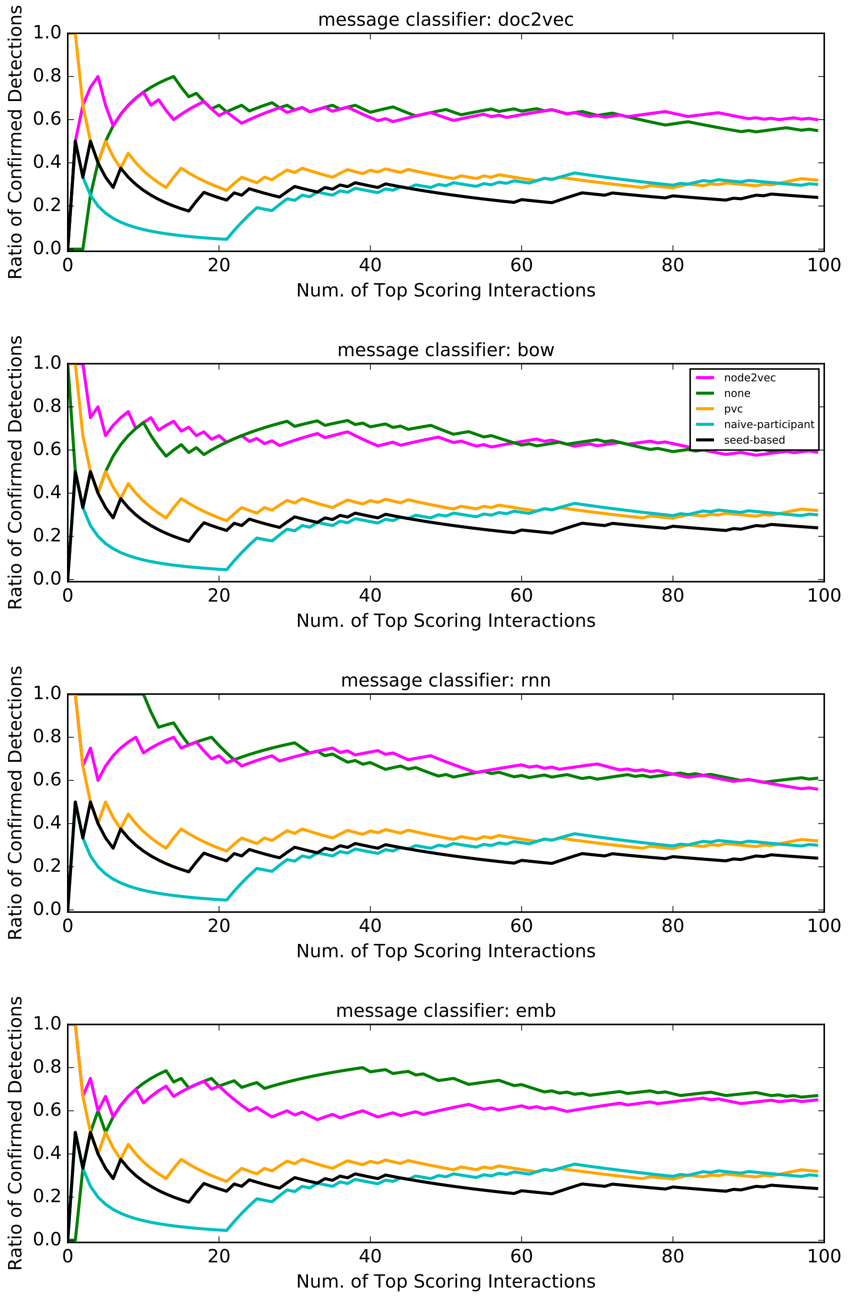


Figure 4.4: Precision@k for bullying interactions on Twitter data using the combination of message and user learners, PVC, seed-based, and naive-participant. The legend in the top left plot applies for all other plots in this figure.

of message and user detectors. We compare these methods with each other and against PVC [197], and two other baselines: *seed-based* and *naive-participant*. The *seed-based* method computes an detection score as the concentration of seed words in the message. The *naive-participant* method computes bully and victim scores for users as the frequency of messages with seed words sent and received, respectively, by the user. The bully and victim scores are then added to a vocabulary score, which is again the concentration of seed words in the message. The precision@k is the proportion of the top k interactions returned by each

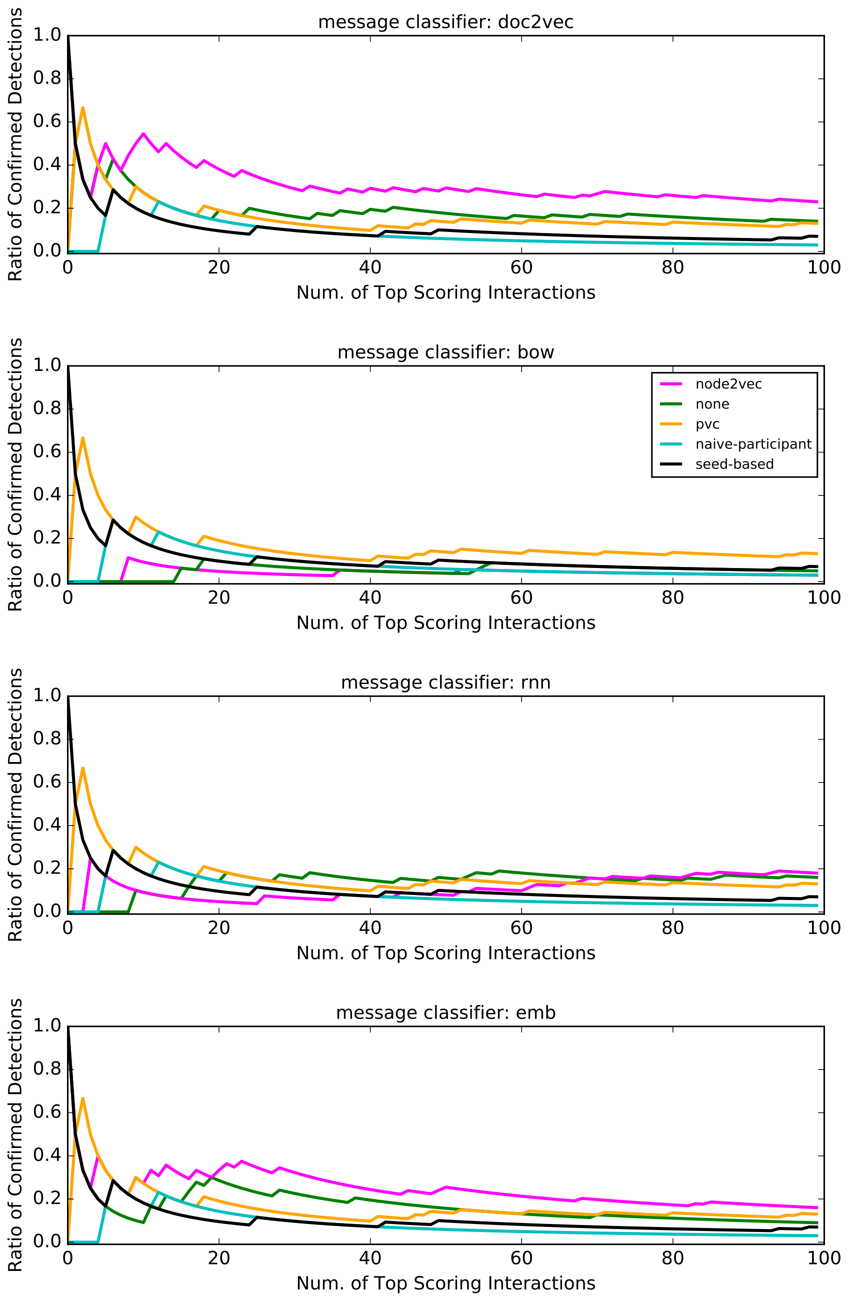


Figure 4.5: Precision@k for bullying interactions on Instagram data using the combination of message and user learners, PVC, seed-based, and naive-participant.

method that the majority of annotators agreed seemed like bullying. For each of the five annotators, we score a positive response as +1, a negative response as -1, and an uncertain response as 0. We sum these annotation scores for each interaction, and we consider the interaction to be harassment if the score is greater than or equal to 3.

We consider the results by grouping them into different message-classifier types. We first examine the doc2vec message detector as shown in top left of Figs. 4.3 to 4.5. The combination of doc2vec message detectors with the node2vec user detector produces the best precision in all three datasets.

We next examine the BoW message detector in the top right of Figs. 4.3 to 4.5. On Ask.fm data, we observe the best precision is achieved by the BoW message detector combined with the node2vec user detector. Interestingly, BoW on Twitter by itself does quite well; its precision is very similar to BoW with the node2vec user learner. The trend, however, is different on Instagram; PVC's performance is better than all variations of BoW. One possible reason why BoW does not do well on the Instagram data may be because we have short and sparse messages in our Instagram data. We also found some conversations were not fully in English, but another language using English characters. This sparsity may cause some tokens to occur only once in the data, causing the bag-of-words representation to be unable to generalize.

The third message detector is the RNN model shown on the bottom left of Figs. 4.3 to 4.5. On Twitter, the RNN by itself and combined with the node2vec user learner achieve higher precision than others. On Ask.fm data, all of the RNN message learner models behave similarly and slightly worse than PVC. On Instagram data, the precision of all RNN methods are similar and low; they have approximately the same performance as PVC.

The last message detector type uses the embedding representation shown in bottom right of Figs. 4.3 to 4.5. On Twitter, the embedding message learner by itself has the best precision. The embedding message learner when combined with node2vec user learner has the second-best precision on Twitter. On Ask.fm and Instagram, however, the combination of the embedding message learner and the node2vec user learner has the best precision.

Comparing across all models we considered, for all three datasets, the BoW message detectors, when combined with the node2vec user detector, had better precision. A similar trend occurs when using the doc2vec message learner. We believe the deep models, because they attempt to extract more semantic meaning of the words, are able to overcome the sparsity of our Instagram data. While the RNN message detector does better than PVC and the other baselines on Twitter, its performance is poor compared to PVC and baselines on Ask.fm and Instagram.

We summarize the precision analysis by answering three major questions:

(i) **Is there any combination that performs the best?** We list in Table 4.1 the precision@100 of all methods. We bold the top four methods with the highest precision@100 for each dataset. Then, we highlight the methods that are among the top four for all three datasets. The statistics indicate that the combinations of the node2vec user learner and the doc2vec or embedding message learners produce the most precise detections.

(ii) **Are deep models better than non-deep models?** We compute the average precision@100 score of deep models vs. non-deep models (PVC, seed-based, and naive-participant). For Twitter, the average precision@100 score of deep models and non-deep models are 0.541 and 0.286 respectively. For Ask.fm, the average score of deep models and non-deep models are 0.295 and 0.15 respectively. For Instagram, the average score of deep models is 0.1216, while the average score of non-deep models is 0.0766. In all three datasets, the average score of precision@100 for deep models is higher than the value in non-deep models. This trend suggests that on average, deep models outperform the non-deep models at detecting cyberbullying instances.

(iii) **Does co-training help?** Figure 4.6 plots the difference between the precision@100 score of the message learner co-trained with a user learner and the message learner by itself for each dataset. If this difference is positive, it indicates that the co-training helps improve precision. For Ask.fm, the co-training improves the precision for all message learners. For Instagram, the co-training improves the precision for all language models except the BoW learner. For Twitter, however, the co-training only helps the doc2vec message learner; the precision for the other two language models reduces slightly with co-training. These summary results suggest that co-training often provides significant improvements, but it can also somewhat reduce precision in some cases.

Table 4.1: Precision@100 across all methods. The methods with the highest precision@100 for each dataset are bolded. The methods that are among the top four for all three datasets are highlighted. The doc2vec_node2vec and emb_node2vec combinations are among the top methods for all datasets.

model	Ask.fm	Twitter	Instagram
doc2vec_node2vec	**0.43**	**0.6**	**0.23**
bow_node2vec	**0.4**	0.59	0.03
emb_node2vec	**0.33**	**0.65**	**0.16**
bow_none	**0.32**	**0.61**	0.05
doc2vec_none	0.31	0.55	0.14
rnn_node2vec	0.29	0.56	**0.18**
emb_none	0.26	**0.67**	0.09
rnn_none	0.25	**0.61**	**0.16**
pvc	0.28	0.32	0.13
seed-based	0.09	0.24	0.07
naive-participant	0.08	0.3	0.03

4.3.2 Qualitative Analysis

We inspected the interactions the eight models identified in the three datasets and found three categories of note. The first category contains bullying conversations detected by most of the models and verified by the annotators. Examples in this category are straightforward true positives because most of these conversations have a high concentration of swear words. Two examples of such conversations follow.

User1:	youre not even sick you dumb b*tch
User2:	I cant f*cking breathe you ugly c*nt
User1:	then how is you alive dumb hoe stupid b*tch *ss c*nt. Kta b*tch

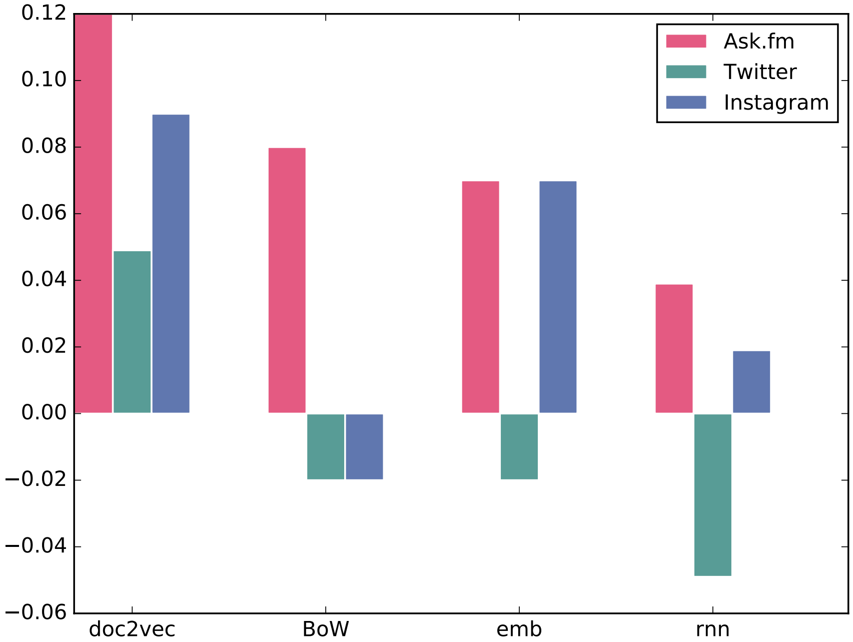


Figure 4.6: The difference between the precision@100 score of the co-trained ensemble and the message learner trained alone. The improvements produced by co-training are more pronounced for the doc2vec language model than others, but the co-training also improves the performance of the other models as well.

User1:	never seen another man on someones d*ck like you. Why you worried about him being ugly. Your prob a dogs *ss yourself.
User2:	just saw your avy lmaooooooooo youre just as ugly, you can take the cape off now freak
User1:	lol ok bud. Keep rtn shoes for sale to your 9 followers f*ckin loser.
User2:	look in the mirror before you can anyone a loser d*ck wad.

The second category of interactions contains conversations with little prototypical bullying language, which are detected by models with the user learner but not by models without user classifiers (i.e., the BoW, doc2vec, RNN, and embedding message learners alone). Because the language-only detectors do not discover these types of conversations, these examples are evidence that considering social structure helps find complicated harassment incidents. Two examples of these challenging true positives follow.

User1: Truth is. You hate me. Rate- my mom said if I have nothing nice to say, I shouldn't say anything at all.
User2: Let me explain why I hate you. Okay so I only hate three people so obviously you have pissed me off enough to get on that list. So for starters, you obviously said you think that T*** and J*** will go to hell. Don't say two of best friends will go to hell because who else would T and J be? Second, you called R*** gay. That's not acceptable either. He even had a girlfriend at the time. You blamed it on your friend P**** or whatever her name is. So you didn't accept what you did and tried to hide it and that didn't work because we ALL know you called him gay multiple times. Another thing is, you are honestly so ignorant and arrogant. You think you are the best of the best and think you have the right to do whatever you want, whenever you want but you cant. I hate to break it to you, but you aren't the little princess you think you are. and you are basically calling me ugly in that rate. But you know what? i know im not the prettiest but at least im not the two-faced, conceited, b*tch who thinks that they can go around saying whatever they want. because saying people will go to hell can hurt more than you think. calling someone gay is really hurtful. youve called me ugly plenty of times, too. so congratulations you have made it on the list of people i hate. and i could go on and on but i think ill stop here. btw; your mom obviously didnt teach you that rule well enough. "buh-bye

User1: listen ** you need to stop , leave me alone , stop harassing me . leave me alone your a creeper ....
User2: Im harassing you ? baha . thats funny bc your the one that started with me , you were the one that said you were gonna fight me . and ** is the one that has the videos . so get your facts right . and Im not gonna waste my time on you . why the hell would I do that ? baha .

The third category of interactions contain non-bullying conversations detected by most models. These false positives are considered by the annotators and us to be non-harassment. In many of these false-positive interactions, the users are joking with each other using offensive words, which is common among younger social media users. These examples include usage of offensive language but may require sophisticated natural language processing to differentiate from harassing usage. Two examples of these false positives follow.

User1: Why you such a b*tch?
User2: i have to stay sassy sl*t xx
User1: Thanks.
User2: youre f*cking welcome.

User1: link plz you b*tch :P I Need A Better f*cking score :P :P
User2: who you callin b*tch, b*tch :P
User1: Motherf*cker :P
User2: f*ckface. :P
User1: Dipsh*t B*tch *sshole *ss :P

Many of the detections by our machine learning models appeared to be correct. Since most of the false positives that we observed were conversations with a high concentration of offensive language, we expect a more refined form of weak supervision than key phrases may help the co-training approach make more nuanced judgments about these cases. Nevertheless, our examination of the detected conversations provided evidence of how effective weakly supervised learning can be at training these deep learning models.

4.4 Conclusion

In this chapter, we present a method for detecting harassment-based cyberbullying using weak supervision. Harassment detection requires managing the time-varying nature of language, the difficulty of labeling the data, and the complexity of understanding the social structure behind these behaviors. We developed a weakly supervised framework in which two learners train each other to form a consensus on whether the social interaction is bullying by incorporating nonlinear embedding models.

The models are trained with an objective function consisting of two losses: a weak-supervision loss and a co-training loss that penalizes the inconsistency between the deep language-based learner and the deep user-based learner. We perform quantitative and qualitative evaluations on three social media datasets with a high concentration of harassment. Our experiments demonstrate that co-training of nonlinear models improves precision in most of the cases.

In the next step, we develop methods to train cyberbullying detection models to avoid learning discriminatory bias from the training data. A serious concern of any automated harassment or bullying detection is how differently they flag language used by or about particular groups of people. Our goal is to design fair models for cyberbullying analysis to prevent unintended discrimination against individuals based on sensitive characteristics including race, gender, religion, and sexual orientations. To tackle this phenomenon mathematically, we will add an unfairness penalty term to the co-trained ensemble framework. The basic idea is to penalize the model when we observe discrimination in the predictions.

Chapter 5

Reduced-Bias Co-Trained Ensembles for Weakly Supervised Cyberbullying Detection

5.1 Introduction

In chapter 4, we introduced a framework called *co-trained ensembles*, which uses weak supervision to significantly alleviate the need for tedious data annotation. This weak supervision is in the form of expert-provided key phrases that are highly indicative or counter-indicative of bullying. In addition, the framework is based on consistency of two detectors that co-train one another. These detectors use two different perspectives of the data: (1) language and (2) social structure. By using different forms of evidence, the detectors train to reach the same conclusion about whether social interactions are bullying. Furthermore, we incorporated distributed word and graph-node representations by training nonlinear deep models. With the advantages of weakly supervised training, there is also a concern that the self-training mechanism used to amplify the weak supervision may also amplify patterns of societal bias. Therefore, in this chapter, we extend the co-trained ensemble model to mitigate unfair behavior in the trained model.

A key concern in the development and adoption of machine learning for building harassment detectors is whether the learned detectors are *fair*. Most machine learning models trained on social media data can inherit or amplify biases present in training data, or they can fabricate biases that are not present in the data. Biases in harassment detectors could be characterized as when harassment detectors are more sensitive to harassment committed by or against particular groups of individuals, such as members of ethnic, gender, sexual orientation, or age groups, which result in more false detections on protected target groups. Recent reactions to a Google Jigsaw tool for quantifying toxicity of online conversations (see e.g., [218]) have highlighted such concerns. A flaw of these detectors is how differently they flag language used by or about particular groups of people. Sinders illustrated that the tool flagged the single-word message “arabs” as “64% similar to toxic language.” Examples such as these illustrate the need for methods that can avoid treating particular identity groups unfairly.

Our goal in the next step is to address discrimination issue against particular groups of people in the context of cyberbullying detection. We add an unfairness penalty to the framework. The unfairness term penalizes the model when we observe discrimination in predictions. We explore two unfairness penalty terms, each aiming toward a different notion of fairness. One aims for *removal fairness* and the other for *substitutional fairness*. For removal fairness, we penalize the model if the score of a message containing sensitive keywords is higher than if those keywords were removed. For substitutional fairness, for each protected group, we provide a list of sensitive keywords and appropriate substitutions. For example, for the keyword "black" describes an ethnicity, substitutions are "asian," "american," "middle-eastern," etc. A fair model would score a message containing any sensitive keyword the same if we replace that sensitive keyword with another; otherwise, we penalize the objective function.

We measure the learned model's fairness on a synthetic data and a dataset of Twitter messages. Our synthetic data is a corpus of sentences generated using the combination of some sensitive keywords describing different attributes: sexual orientation, race, gender, and religion. Mirroring the benchmark established in Chapter 4, we generated statements of identity (e.g., "black woman", "muslim middle-eastern man") that are not harassment. To assess models fairness, we compute the *false-positive rate* on these identity statements. An ideal fair language-based detector should yield a lower false-positive rate on these non-bullying statements. On our Twitter data evaluation, we measure model fairness using the equality of odds gaps [104]. Specifically, we use a criterion we call *category dispersion*, which is the standard deviation of area under the curve (AUC) of receiver operating characteristic (ROC) across multiple keywords in a category of substitutable keywords. A low *category dispersion* is more desirable since it indicates that the model treats the subpopulations equitably. A high *category dispersion* indicates that the model behavior is more favorable toward some subgroups; hence, it discriminates against some other subpopulations. We also test the model's fairness qualitatively by showing conversations with sensitive keywords where their score using the reduced-bias model is much lower than the default model. In another qualitative analysis, we examine the change in the bullying score of sensitive keywords when fairness imposed to the harassment detector.

The main contributions of this work are as follows: We propose a reduced-biased framework for weakly supervised training of cyberbullying detectors. This method is the first specialized algorithm for cyberbullying detection that considers fairness while detecting online harassment with weak supervision. To penalize the model against discrimination, we add an *unfairness penalty term* to the objective function. We introduce two fairness forms: *removal* and *substitutional* fairness. We evaluate the model's fairness on identity statement benchmark by comparing the false-positive rate of the models with and without fairness constraints. Additionally, we measure the model's fairness on Twitter data by introducing *category dispersion* criterion, which represents how well a model behaves equitably across various subpopulations in a category, without accuracy degradation.

5.2 Reduced-Bias Co-Trained Ensembles

In this section, we introduce our approach to reduce bias in co-trained ensemble models. Our goal is to disregard discriminatory biases in the training data and create fair models for cyberbullying detection. We add a term to the loss function to penalize the model when we observe discrimination in the predictions. We investigate the effectiveness of two such *unfairness penalty terms.*

Removal Fairness The motivation for *removal fairness* is that, in a fair model, the score of a message containing sensitive keywords should not be higher than the same sentence without these keywords. Therefore, we penalize the model with

$$\ell(y_m) = \alpha \times -\log\left(\min\left\{1, 1 - (y_m - y_{m^-})\right\}\right). \tag{5.1}$$

where y_m is the score of message containing sensitive keywords, and y_{m^-} is the score of the same message when sensitive keywords are dropped. The parameter α represents to what extent we enforce fairness to the model. In our experiments, we examined three values $\alpha = 1$, 10, 100. The best value resulting better generalization and lower validation error was $\alpha = 10$.

Substitutional Fairness In *substitutional fairness*, for each category of sensitive keyword, we provide a list of sensitive keywords and appropriate substitutions. For example, the keyword “black” describes a type of ethnicity, so substitutions are “asian,” “native-american,” “middle-eastern,” etc. Or the keyword “gay” describes a sexual orientation, so it can be substituted with “straight,” “bisexual,” “bi,” etc. In a fair model, the score of a message containing a sensitive keyword should not change if we replace that sensitive keyword with its substitutions. We penalize the objective function with

$$\ell(y_m) = \frac{\alpha}{|S_c| - 1} \times \sum_{i \in S_c, i \neq k} \left(y_{m(k)} - y_{m(i)}\right)^2. \tag{5.2}$$

where S_c is the set of all sensitive keywords in category c, $|S_c|$ is the cardinality of set S_c, $y_{m(k)}$ is the score of original sentence containing sensitive keyword k in category c, and $y_{m(i)}$ is the score of the substitution sentence with sensitive keyword i in category c.

As in removal fairness, α represents to what extent we want to impose fairness to the group. We tried three values for $\alpha = 1$, 10, 100; and value 10 again led to the best validation error.

5.3 Experiments

Mirroring the setup initially used to evaluate the co-trained ensemble framework, we construct a weak supervision signal by collecting a dictionary of offensive key-phrases (unigrams and bigrams) [179]. We then manually removed all keywords directly related to any particular race, gender, sexual orientation, and religion from the dictionary. Finally, we had a collection of 516 curse words for our weak supervision. We augment this with a list of positive opinion words in [114]. The offensive phrases are our weak indicators and the positive words are our counter-indicators. Out of eight combinations of message and user learners introduced in previous section, we selected the top four models with the highest precision@100 on Twitter for our evaluation: “emb_none,” “emb_node2vec,” “bow_none,” and “doc2vec_node2vec.”

We consider four categories of sensitive keywords: race, gender, religion, and sexual orientation. For each category, we provide a list of keywords. Some example keywords in the race category are “white,” “black,” “caucasian,” “asian,” “indian,” and “latina”; and some example keywords in religion category are “christian,” “jewish,” “muslim,” “mormon,” and “hindu.” In each message, there might be many sensitive keywords. Hence, for computational purposes, we limit the number of substitutions to 20. If the number of substitutions for a message is more than 20, we randomly select 20 substitution sentences.

Evaluation on Synthetic Data We analyze the sensitivity of models toward some often targeted groups on a synthetic benchmark. This benchmark is a corpus of sentences using the combination of sensitive keywords describing different attributes. These statements of identity (e.g., “I am a Black woman,” “I am a Muslim middle-eastern man,” etc.) are not bullying since they are simply stating one’s identity. To assess each models fairness, we compute the *false-positive rate* on these synthetic benchmark statements. An ideal fair language-based detector should yield a lower false-positive rate on these non-toxic statements. Table 5.1 shows the *false-positive rates* (at threshold 0.5) of four aforementioned co-trained ensembles with and without penalizing unfairness. According to the results in Table 5.1, the false-positive rate of these four methods reduces when either removal or substitutional fairness constraints applied.

Table 5.1: False positive rate of models on non-bullying synthetic benchmark statements (using threshold 0.5). Both *removal* and *substitutional* fair models reduce the false positive rate compare to without bias reduction (vanilla).

Method	emb_none	emb_node2vec	bow_none	doc2vec_node2vec
Vanilla	0.8416	0.7055	0.2663	0.0000
Substitutional	0.7685	0.1439	0.0305	0.0000
Removal	0.0000	0.0418	0.0000	0.0000

Evaluation on Twitter We use the data collected for our PVC and co-trained ensembles. We collected data from Twitter’s public API, extracting tweets containing offensive-language words posted between November 1, 2015, and December 14, 2015. They then extracted conversations and reply chains that included these tweets. They then used snowball sampling to gather tweets in a wide range of topics. After some preprocessing, the Twitter data contains 180,355 users and 296,308 tweets.

We evaluate the effectiveness of our approach using post-hoc crowdsourced annotation to analyze the score of fair-imposed model for conversations with sensitive keywords. We extract all of the conversations in Twitter data containing at least one sensitive keyword. Then, we asked crowdsource workers from Amazon Mechanical Turk to evaluate the interactions. We showed the annotators the anonymized conversations and asked them, "Do you think either user 1 or user 2 is harassing the other?" The annotators indicated either "yes," "no," or "uncertain." We collected three annotations per conversation. For each of the three annotators, we score a positive response as +1, a negative response as -1, and an uncertain response as 0. We sum these annotation scores for each interaction, and we consider the interaction to be harassment if the score is greater than or equal to 2.

To measure a model's fairness, we use an "equality of odds" criterion [104], which states all subpopulations in a group experience the same true- or false-positive rate. We generalize a notation "accuracy equity" [70]. We compute the standard deviation of the area under the curve (AUC) of the receiver order characteristic (ROC) for a set of keywords in a category of sensitive keywords. We refer to this measure as *category dispersion*. An ideal, fair language-based detector should treat these keywords equitably that is equivalent to lower category dispersion.

Table 5.2 shows the category dispersion values of methods on each targeted group. We **bold** values when the category dispersion criterion of the reduced-bias method is lower than the vanilla learner (without fairness), which indicates that the reduced-biased model is treating the keywords in the protected category more equitably. According to Table 5.2, the substitutional version of emb_none and doc2vec_node2vec is fairer across all three categories. Substitutional bow_none is fairer in two out of three categories; and substitutional emb_node2vec is fairer for only the religion category. The removal version of emb_none, doc2vec_node2vec, and bow_none has fairer behavior in two out of three categories, while emb_node2vec only treats keywords in the religion category more equitably. To sum up, emb_none and doc2vec_node2vec, when trained with substitutional fairness terms produce lower standard deviation of AUC across keywords for all three tested categories.

An important question, however, is whether there is accuracy degradation as our approach encourages more equitable errors within categories. In Figure 5.1, we plot the ROC curve of four co-trained ensemble methods stratified by fairness type. Surprisingly, the AUC of substitutional bow_none, emb_none, and doc2vec_node2vec actually improve over the default approach. The performance of removal emb_none improves over vanilla, but other methods' performance reduce when adding removal fairness. In Figure 5.2 we compare the ROC curve of emb_none method for some sensitive keywords for the categories of sexual orientation and religion. The ROC curves of sensitive keywords in each group are closer to each other in both removal and substitutional fair methods, while the AUC of most keywords in the substitutional and removal versions are higher than the vanilla approach. This trend indicates that the bias-reduction successfully equalizes the behavior across language describing different subpopulations of people.

Qualitative Analysis We qualitatively test the model's fairness by analyzing the highest scoring conversations identified by the models. An ideal fair model should give a lower score to non-bullying interactions containing sensitive keywords. Figure 5.3 displays three non-bullying conversations highly ranked by the vanilla model, but given a low score by the reduced-bias model.

Table 5.2: The category dispersion of four co-trained ensemble methods for three targeted groups. A bold value means the category dispersion of the reduced-bias model is lower than the default method (vanilla). Substitutional emb_none and doc2vec_node2vec have better category dispersion than the vanilla method in all three groups.

Category	Fairness Type	emb_none	emb_node2vec	bow_none	doc2vec_node2vec
Race	Vanilla	0.0937	0.0525	0.0231	0.0546
	Substitutional	**0.0531**	0.0528	0.0514	**0.0460**
	Removal	**0.0640**	0.0594	0.0665	0.0587
Religion	Vanilla	0.0858	0.0862	0.0716	0.0748
	Substitutional	**0.0657**	**0.0376**	**0.0660**	**0.0494**
	Removal	0.0902	**0.0424**	**0.0190**	**0.0661**
Sexual Orientation	Vanilla	0.1096	0.0791	0.0915	0.0702
	Substitutional	**0.0629**	0.0821	**0.0666**	**0.0623**
	Removal	**0.0894**	0.0914	**0.0467**	0.0784

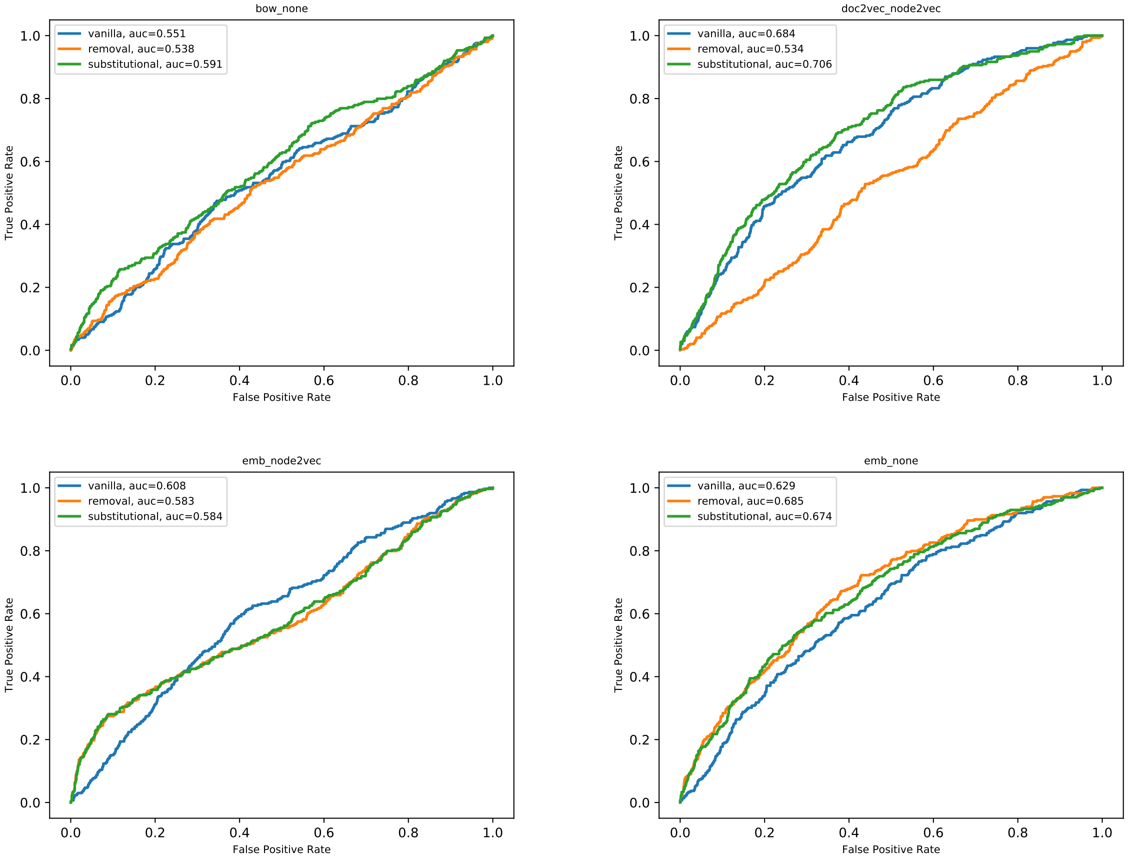
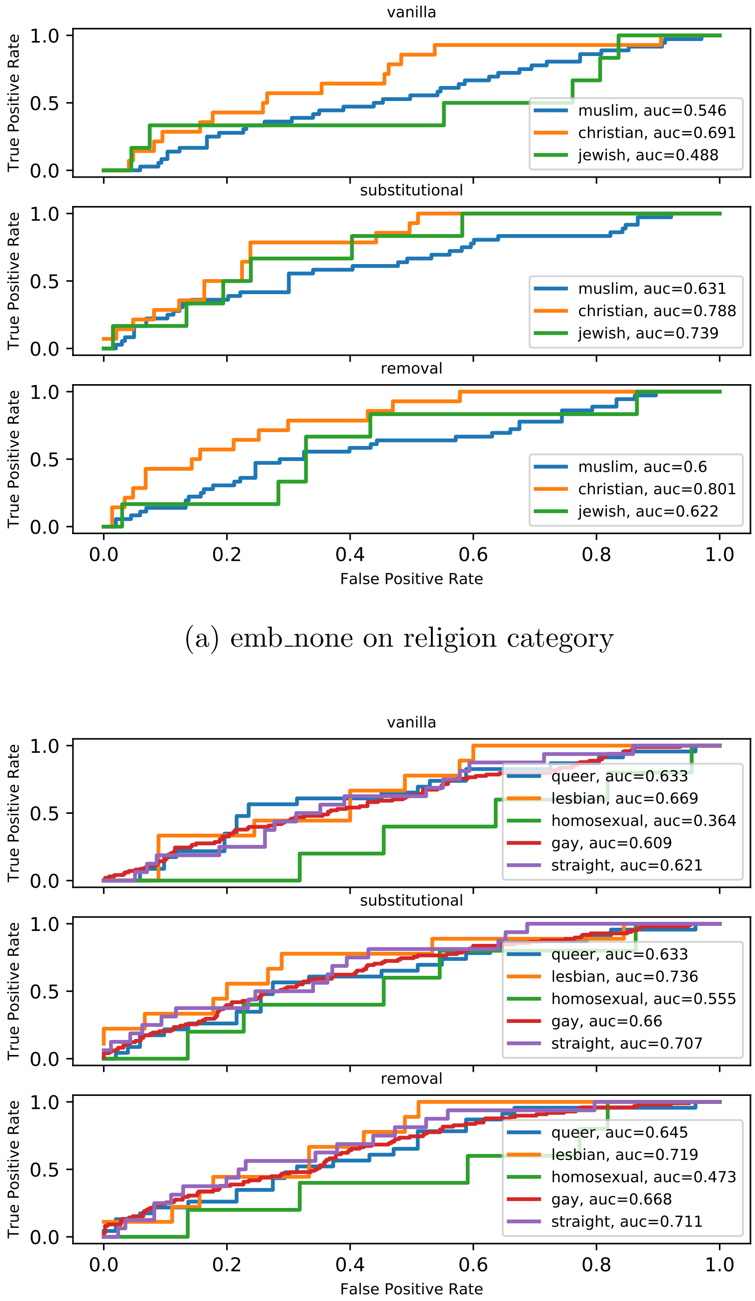


Figure 5.1: ROC curve of the four co-trained ensemble methods stratified by fairness type. The performance of substitutional bow_none, doc2vec_node2vec, and emb_none improves over the vanilla method. Removal emb_none also improves over vanilla, but emb_node2vec has worse performance.



(a) emb_none on religion category

(b) emb_none on sexual orientation category

Figure 5.2: ROC curves for the emb_none model on data containing sensitive keywords in the sexual orientation and religion categories. The gap between AUC of sensitive keywords in these two groups reduces, without degradation in performance.

We also compare the bullying score of messages containing sensitive keywords with and without bias reduction. We consider the scores of embedding message classifiers since they learned a vector representation for all words in corpus. We plot the scores in Figure 5.4. Our findings are as follows: The score of most sensitive keywords in the ethnic category,

User1: all figure that you are, in fact, ***gay*** and theyre freaking out about it? F*CK THOSE PEOPLE.
User2: I think my Uncle has figured it out, him & my cuz are always trying to out me, I feel nervous/sick every time theyre around.

User1: Alienate minorities? Why would anyone, of any color, stand w/ppl that dont believe al lives matter?
User2: All Lives matter Is a gasp of a little racist mind, when Cops shot white boys at racist of ***Blacks*** then OK

User1: you cant really be racist against a religion...
User2: i know my friend but you can be racist against followers of this religion.iam an exMuslim but refuse to be a racist against them
User1: Racist is hating someone based in their skin color, not their faith. Hating someone just for their faith is just bigotry.
User2: remamber the ***Jewish*** and the holocaust or you think it wasnt racist!!!!!!!!!!!! #ExMuslimBecause
User1: ***Jewish*** can be a race or a religion. There is no Muslim race. Theres an Arab race, but you cant assume Arabs are Muslim.

Figure 5.3: Three examples of conversations given a high score by the vanilla model and given a low score by a reduced-bias model. These conversations discuss sensitive topics but may not represent one user bullying another.

such as "black," "indian," "african," "latina," "asian," "hispanic," and "white" reduces when fairness is imposed on the model, but the score of "american" increases. The reason the keyword "white" is scored higher using the vanilla model could be the occurrence of this word in bullying interactions in our dataset. In the gender category, the score of "boy" increases when imposing fairness, but the score of "girl" either reduces or does not change. The bullying score of "woman," however, increases using fairer methods, while the score of "man" increases using substitutional fairness and reduces slightly using removal fairness. In the religion category, the bullying score of "muslim," "christian," and "jewish" reduce when fairness is imposed, but the score of "protestant" score increases using fair models. In the sexual orientation category, the score of "gay," "transexual," "queer," "lesbian," "bi," "heterosexual," and "straight" reduce using biased-reduced models. It is worth noting that "gay" and "transexual" are the highest bullying score without fairness enforcement.

By scrutinize the behavior of the emb_none model, we observe a remarkable change in the score of sensitive keywords when fairness is imposed on the model. Its vanilla version gives the highest bullying score to most sensitive keywords, but the scores reduce noticeably when fairness is imposed. On the other hand, the score of some keywords using the vanilla method is significantly low, but when fairness terms are applied, their scores increase. An open question is whether this significant variation is desirable especially with considering the performance improvement of emb_none in our quantitative analysis.

Discussion One question that might arise is which fairness objective best provides the desired fairness? Considering our quantitative analysis in Table 5.2, which follows the equality of odds criterion, substitutional constraints improve the AUC gap of more groups

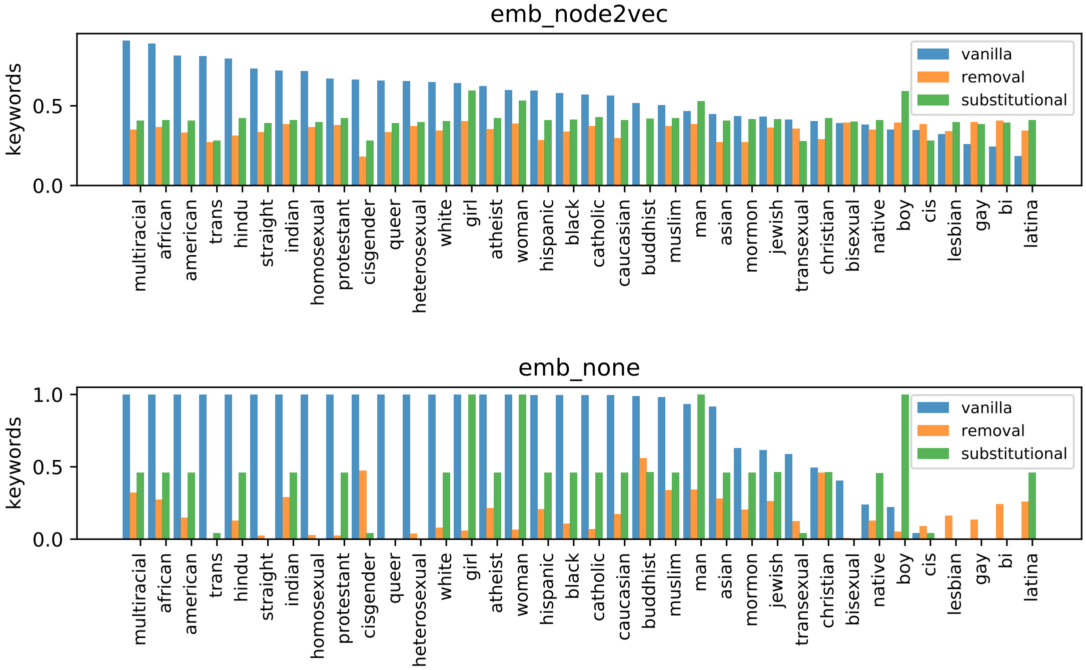


Figure 5.4: Bullying scores of sensitive keywords as scored by emb_none and emb_node2vec with and without fairness terms. The bullying score of most sensitive keywords reduce when fairness is imposed on the model, and the scores become more uniform than without any fairness terms (vanilla).

for each method. However, the difference is not significant. Both constraints reduce the false-positive rate on the synthetic benchmark. Another question is which combination of model architecture and fairness objective has better behavior? Table 5.2 suggests emb_none and doc2vec_node2vec with substitutional fairness produce lower AUC gap between keywords for all three groups. One might ask about the trade-off between accuracy and fairness. As shown in Figure 5.2, adding a fairness term reduces the accuracy of emb_node2vec, but emb_node2vec is also unable to produce fair predictions either. This pattern suggests that emb_node2vec is not compatible with the introduced fairness constraints. The emb_none and doc2vec_node2vec models have better fairness behavior with the substitutional objective, and their accuracy also improves. The removal objective, however, reduces the accuracy when added to most models.

5.4 Conclusion

Fairness is one of the most important challenges for automated cyberbullying detection. As researchers develop machine learning approaches to detect cyberbullying, it is critical to ensure these methods are not reflecting or amplifying discriminatory biases. In this chapter, we introduce a method for training a less biased machine learning model for cyberbullying analysis. We add unfairness penalties to the learning objective function to penalize the model when we observe discrimination in the model's predictions. We introduce two fairness penalty terms based on removal and substitutional fairness. We use these fairness terms to augment co-trained ensembles, a weakly supervised learning framework introduced in Chapter 4. We evaluate our approach on a synthetic benchmark and real data from Twitter. Our experiments on the synthetic benchmark show lower *false-positive rates* when fairness is imposed on the model. To quantitatively evaluate model's fairness on Twitter, we use an equality of odds measure by computing the standard deviation of AUC of for messages containing sensitive keywords in a category. A fair model should treat all keywords in each category equitably, meaning lower standard deviation. We observe that two ensemble learners, when augmented with substitutional fairness, reduce the gap between keywords in three groups, while their detection performance actually improves. We did not always observe such behavior when models were augmented with removal fairness. In addition, we qualitatively evaluate the framework, extracting conversations highly scored by the vanilla model, but not so by the bias-reduce models. These conversations tended to be sensitive, but not examples of bullying. We therefore demonstrate the capability to reduce unfairness in cyberbullying detectors trained with weak supervision.

Chapter 6

Complementary Experiments

In this chapter, we address some questions might arise regarding how well annotators make the same labeling decision for bullying conversations, considering a straightforward node representation for users in the network, and comparison of our model against a fully supervised model.

6.1 Inter-Annotator Agreement

To gauge our model's performance, we asked crowdsourcing workers from Amazon Mechanical Turk to label the cyberbullying interactions discovered by all the methods. We showed the annotators the anonymized conversations and asked them, "Do you think either user 1 or user 2 is harassing the other?" The annotators indicated either "yes," "no," or "uncertain." We collected five annotations per conversation. In order to ensure annotation process is consistent and reliable, we measure of the degree to which crowdsource workers agree on judgments. This is known as "inter-annotator agreement" [11].

The most straightforward strategy is to compute the percent agreement among annotators. This is called "raw agreement" or "observed agreement" [21]. The problem with this measure is that it does not consider the possibility of the agreement occurring by chance. Cohen's kappa coefficient [54] is more robust inter-annotator agreement measure in which the probability that annotators label the data by chance is taken into consideration. It computes the proportion of actual agreement over the agreement by chance and the maximum agreement achievable over chance agreement taking into account pairwise agreement [27]. Cohen kappa is used when there are only two annotators. For more than two annotators, the *Fleiss kappa* measure [90] is introduced, which was an extension of Cohen's kappa measure of degree of consistency for two or more annotators. We use Fleiss kappa to evaluate inter-annotator's agreement. In Table 6.1, we list the Fleiss kappa measure for annotation on each dataset separately and altogether. The Fleiss kappa measure between 0.41 and 0.60 indicates the moderate agreement among annotators [138]. Therefore, based on the figures in Table 6.1, all labelers moderately agree about bullying conversation in each dataset; annotators agree with each other on Instagram slightly more, and on Twitter slightly less.

Table 6.1: Fleiss kappa inter-annotator agreement measure for each dataset and all of them together.

Dataset	Twitter	Ask.fm	Instagram	Total
Fleiss kappa measure	0.442	0.457	0.455	0.49

6.2 Explainable Node Representations

For node embedding, we used node2vec to map the users to a vector of real number. Node2vec is inspired by SkipGram word embedding in which a network can be represented as an ordered sample sequences of nodes. The problem, however, is that this representation seems like a black box. Our goal in this experiment is to see if using explainable node representation performs similar to more complicated non-linear node2vec.

For each node in the graph, we computed three interpretable representation such that each user is represented by a vector of three numbers: [in-degree, out-degree, and PageRank]. We performed the experiments on Twitter data using the doc2vec message learner. As shown in Figure 6.1, degree-based node representation does not improve the precision; it even reduces the message classifier's performance. These results are consistent with the experiments we showed in Chapter 3, Section 3.3.6, stating that there are varieties of structures among users.

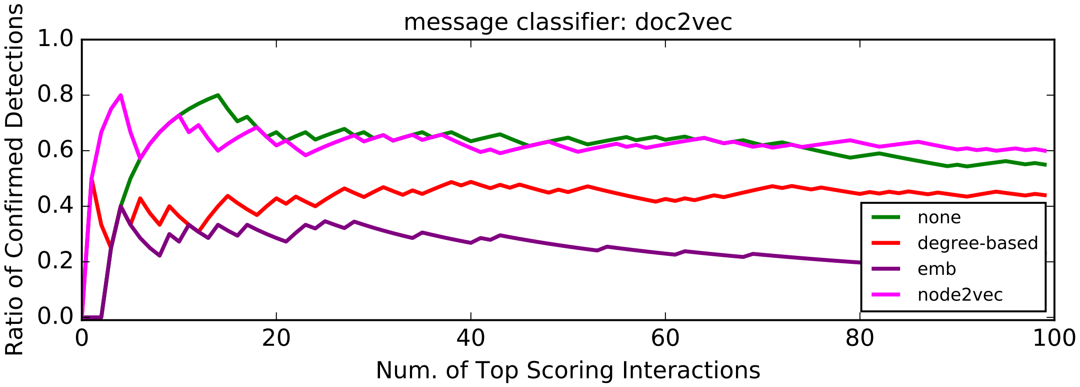


Figure 6.1: Comparing precision@k doc2vec message learner using different user learners. The red color shows the explainable *degree-based* node representation for user classifier. Degree-based user learner decreases the precision of "doc2vec" message learner.

6.3 Weak Supervision Against Full Supervision

One major contribution of this research is using machine learning with weak supervision, which significantly alleviates the need for human experts to perform tedious data annotation. It is important to see how well our framework performs compared to a fully supervised model, or what is the trade-off between fully supervised and weakly supervised method.

We generated a synthetic training data consisting of 500 messages, and test data with 300 messages. To generate a synthetic data, we randomly assign two scores to each user: bully and victim scores. Then, we decide whether the conversation is bullying or not based on bullying score 0.5 x (bully + victim). For example if the bullying score is higher than 0.7, the conversation is bullying. Next, we define 50 neutral words and 5 bad words; and assign the probability to words according to the conversation type. For bullying conversations, the probability of bad words is 0.05, and the probability of neutral words is 0.01 (the summation of the probability of all words should be 1). For non-bullying conversations, the probability of non-bullying words is 0.018 and the probability of bullying words is 0.01. Then, we randomly assign words to the conversation according to the bullying score.

Then, we compared PVC with two fully supervised models: multinomial naive Bayes, and logistic regression. According to the Figure 6.2 the AUC of weakly supervised model (PVC) is smaller than that of fully supervised models. It should be noted that the performance reduction of the weakly supervised model is not significant considering the lower cost of human annotations.

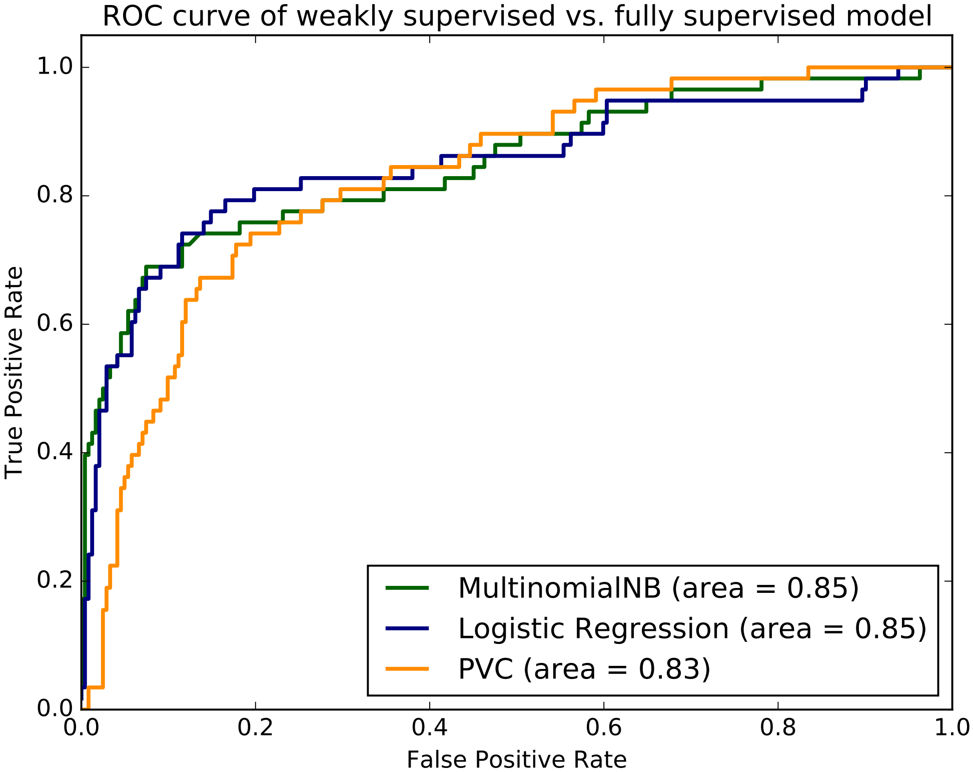


Figure 6.2: ROC curve of weakly supervised PVC model and fully supervised model. The area under the ROC curve of PVC is lower than both logistic regression and multinomial naive Bayes.

Chapter 7

Summary and Outlook

7.1 Significance and Contribution

Early detection of harmful social media behaviors such as cyberbullying is necessary for identifying threatening online abnormalities and preventing them from increasing. Finding a technical solution to handle this problem is not easy because of the time-varying nature of language, the difficulty of labeling the data, and complexity of understanding the social structure behind these behaviors. In this research, we aim to improve automated detection of cyberbullying, which is key step toward automated systems for analyzing modern social environments that can adversely influence mental health. We develop a general framework with weak supervision in which two learners co-train each other to form a consensus whether the social interaction is bullying. In the first step, we introduced participant-vocabulary consistency (PVC), which simultaneously discovers which users are instigators and victims of bullying, and additional vocabulary that suggests bullying. These quantities are learned by optimizing an objective function that penalizes inconsistency of language-based and network-based estimates of how bullying-like each social interaction is across the social communication network. Experiments on three social media platforms with high frequency of cyberbullying, Twitter, Ask.fm, and Instagram, demonstrate that PVC can discover instances of bullying and new bullying language.

In the next step, we incorporated nonlinear embeddings models in which the language and users are represented as vectors of real numbers. The models are trained by constructing an optimization problem consisting of two losses: weak-supervision loss that is the classification loss on weakly labeled messages, and co-training loss that penalizes the inconsistency between the deep language-based learner and the deep user-based learner. Our quantitative and qualitative evaluations on Twitter, Instagram, and Ask.fm show that co-training ensemble of deep models can improve precision in most of the cases.

Fairness is one of the most important challenges for automated cyberbullying detection. As researchers develop machine learning approaches to detect cyberbullying, it is critical to ensure these methods are not reflecting or amplifying discriminatory biases. In the next step, we introduced a less biased machine learning model for cyberbullying analysis. We added unfairness penalties to the learning objective function to penalize the model when we observe discrimination in the model's predictions. We introduced two fairness penalty terms based on removal and substitutional fairness. We use these fairness terms to augment co-trained ensembles. We evaluate our approach on a synthetic benchmark and real data from Twitter. Our experiments on the synthetic benchmark show lower *false-positive rates* when fairness is imposed on the model. To quantitatively evaluate model's fairness on Twitter, we use an equality of odds measure by computing the standard deviation of AUC of for messages containing sensitive keywords in a category. A fair model should treat all keywords in each category equitably, meaning lower standard deviation. We observe that two ensemble learners, when augmented with substitutional fairness, reduce the gap between keywords in three groups, while their detection performance actually improves. We did not always observe such behavior when models were augmented with removal fairness. In addition, we qualitatively evaluate the framework, extracting conversations highly scored by the vanilla model, but not so by the bias-reduce models. These conversations tended to be sensitive, but not examples of bullying. We therefore demonstrate the capability to reduce unfairness in cyberbullying detectors trained with weak supervision.

7.2 Future prospects

This is ongoing research, and our proposed framework can be improved in several ways. One idea is to explore the usage of different user embedding models beyond node2vec. Another suggestion is to extend the framework by distinguish user roles as bullies or victims such that there would be three learners co-training each other: a message learner, a sender learner, and a receiver learner. Separating the user learner into sender and receiver learners would help identify the directional structure of bullying. Our weak supervision was only on language of social media; but it could be strengthened by weak supervision over users.

Our proposed framework used two learners; one for the language, and the other for user. One suggestion to make the model stronger is using more than two classifiers. For example one can take the temporal aspect of social data into consideration, such that the model can decide bullying occurrences based on the time. One approach would be investigating the reactions of message receivers as an indicator of bullying. These models will classify messages as bullying interactions by tracking the follow-up messages to see if they are responding to a bully or not. It can be done by easily spotting the indicators used in reaction to bullying. Another strategy is using a hidden Markov model (HMM) for which every users tendency to bully or be bullied considered latent variables. By training an HMM, we can infer the probability that each user is a bully or is victimized by bullying, given the history of his or her messages each day. It also could be useful to examine a theory of social behavior stating that deviant behavior can be learned as a result of exposure to such behaviors. This theory can be modeled by estimating the probability that users contribute to cyberbullying given their neighbors engagement in bullying behaviors. This can make a connection to the idea of information diffusion, which has been broadly studied in graph mining literature.

The other strategy to develop methods for ensemble co-training is learning in an online, streaming setting, where the new data can be added to the model with minimal computational cost. This way, the input data is presented as a sequence, and models are updated by arrival of new data. The initial approach is to view each new batch of observations as a data sample for stochastic gradient descent. Therefore, the new data will be regarded as the entire dataset, and then one can take one step of stochastic gradient descent. Some modern stochastic gradient methods such as Adam and RMSprop, which have been introduced to rectify the problem of noise, can be used. They are capable of continually optimizing using unlimited data streams. Different strategies for splitting the batches of data could be examined, from small batches to large batches, to find the trade-off between their advantages. One important step to enhance the performance of the current framework is to use more advanced NLP (e.g., sentiment analysis like humor/sarcasm detection) to reduce false positive rates (e.g., joking).

In one direction of our research, we examined the fairness in our model; and refined our framework to reduce discrimination against some protected groups in model's prediction. Then, we evaluated the model using state-of-the-art fairness criterion called *equality of odds*, which states all subpopulations in a group experience the same true- or false-positive rate. One could consider another fairness criterion: *calibration*, which analyzes the importance of selecting a good threshold that determines the bullying messages. A fixed cutoff would result in different false/true positive rates on ROC curves for different keywords. Even if their ROC curves lie close to one another, the false/true positive rate imbalance could cause labeling of messages with one keyword as bullying more frequently than other related keywords.

We evaluated our framework on our Twitter, Ask.fm and Instagram data. The model could be tested on larger datasets and other social media data. Last but not least, we applied these fairness objectives to a framework introduced for cyberbullying detection, but it can be used for any online behavior detection method. It can therefore be generalized to other context such as detecting trolling, vandalism, hate speech, etc.

7.3 What to Do When Cyberbullying is Detected

Automated detection is only one problem among many that must be solved to adequately address the cyberbullying phenomenon. What to do when cyberbullying is detected is an important open problem. Providing detected bullies and victims advice, filtering content, or initiating human intervention are possible actions an automated system could take upon detection, but how to do any of these tasks in a manner that truly helps is a key open question. The the possible actions could be human interference, deletion of offensive terms, blacklist highly scored users, and many more.

Users of many social media platforms such as Twitter and Facebook are able to report harmful behaviors that violate their terms of service agreement. All of those reported users and messages are analyzed by human workers for the proper action, which will be deletion of inappropriate content or suspending the reported user. The goal of automatic online harassment detection is to reduce the workload of human employees.

Dinakar et al. [71] designed reflective user interfaces to prevent bullying behaviors by making some pauses before sending the negative message such that the user could re-think before action. One strategy is disabling the "send" button for a few second; and another one is highlighting the offensive parts of the message. *ReThink*[1] is a mobile app created for such purposes. When inappropriate words are used in a message, *ReThink* shows a message to the sender and asks them re-thinking before sending the message.

Bosse and Stam [32] introduced an agent-based system to motivate users change their behaviors if they are potentials of cyberbullying. Normative agents use some methodologies to detect different types of norm violations such as insulting. Then using rewards and punishments, they try to enforce the change of the user's behavior toward acceptable social norms. Their system has been deployed in some virtual environments, and tested on children between 6 and 12 years old. The results show that these normative agents help to decrease amount of norm violations on the long term.

7.4 Practical Advice

We trained and evaluated our framework on our collected Twitter dataset as well as Ask.fm and Instagram. To make the model operate in real worlds, the first step is to train on a large dataset consisting of millions of messages and users. The other important factor is selecting a precise seed set with the extremely offensive words. In order to reduce unfair false positives, it is necessary that group-specific slurs should **not** be in the seed set. We represented users in network as vectors of real numbers using node2vec embedding model. It is important to consider other advanced network embedding models in which the goal is to assign similar vectors to the users with similar behaviors. One unavoidable issue of our framework (and many other harassment detection models) is false-positives. We suggest reducing the false-positive rates by applying some advanced natural language processing techniques using sentiment analysis for joke detection. The proposed model will help to identify potential harmful behaviors in social media; however because of the possibility of false positive rate, help of human is needed to make a final decision about the message being bullying or not. The data annotated by humans in the loop will be used for improving the model's performance.

7.5 Other Applications

We introduce a framework for cyberbullying detection in social media using weak supervision. We provide a small seed of offensive words that their presence in the message shows there is a high tendency that bullying happening. Our framework consists of two different classifiers that look at the problem from different perspectives. One looks at the language in a social interaction; and another looks at social users. They co-train one another to come to an agreement about cyberbullying incidents. By some adjustments in the seed set, this framework can be generalized to other applications.

[1] https://itunes.apple.com/us/app/rethink-stop-cyberbully-%20ing/id1035161775?mt=8

- Bot detection
- Detecting abnormal behavior in social media (e.g., trolling, vandals)
- Detecting influencers in social media (with the purpose of marketing, election)
- Detecting illegal activities in social media (online threats, hate crime)
- In general identifying abnormal behavior when some entities communicating with each other using some context such as data or text

7.6 Conclusion

In this research, we aim to improve automated detection of cyberbullying, which is key step toward automated systems for analyzing modern social environments that can adversely influence mental health. We developed a framework for detecting harassment-based cyberbullying using weak supervision, which significantly alleviates the need for human experts to perform tedious data annotation.

First, we proposed participant-vocabulary consistency (PVC) in which starting with a seed set of offensive vocabulary, it discovers which users are instigators and victims of bullying, and additional vocabulary that suggests bullying. Second, we introduced weakly supervised co-trained ensemble framework in which two learners train each other to form a consensus on whether the social interaction is bullying by incorporating nonlinear embedding models. One detector identifies bullying by examining the language content of messages; another detector considers the social structure to detect bullying. Third, we presented a less biased machine learning model for cyberbullying analysis using weak supervision with the aim of reducing the reflection and amplification of discriminatory biases in the data while learning the model. Building off our weakly supervised learning algorithm, we added unfairness penalties to the learning objective function. By penalizing unfairness, we encourage the learning algorithm to avoid discrimination in the predictions and achieve equitable treatment for protected subpopulations. An ideal, fair language-based detector should treat language describing subpopulations of particular social groups equitably.

Early detection of harmful social media behaviors such as cyberbullying is necessary to identify threatening online abnormalities and prevent them from increasing. However, automated detection is only one problem among many that needs to be solved to adequately address the cyberbullying phenomenon. This research represents important steps toward improving technological capability for automatic cyberbullying detection.

Bibliography

[1] Us congress. 1970-10. 15 u.s.c. 1681 ff.: Fair credit reporting act. [Online]. Available from: https://www.govinfo.gov/content/pkg/STATUTE-84/pdf/STATUTE-84-Pg1114-2.pdf.

[2] Us congress. 1974-10. 15 u.s.c. 1691 ff.: Equal credit opportunity act. [Online]. Available from: https://www.ecfr.gov/cgi-bin/text-idx?tpl=/ecfrbrowse/Title12/12cfr202_main_02.tpl.

[3] Report abusive behavior on Twitter. [Online]. Available from: https://help.twitter.com/en/safety-and-security/report-abusive-behavior, 2019.

[4] Report Something on Facebook. [Online]. Available from: https://www.facebook.com/help/263149623790594/, 2019.

[5] G. Adomavicius, J. Bockstedt, S. Curley, and J. Zhang. De-biasing user preference ratings in recommender systems completed research paper. *24th Workshop on Information Technology and Systems*, 01 2014.

[6] G. Adomavicius, J. C. Bockstedt, S. P. Curley, and J. Zhang. Do recommender systems manipulate consumer preferences? a study of anchoring effects. *SSRN Electronic Journal*, 24, 12 2013.

[7] M. A. Al-Ajlan and M. Ykhlef. Deep learning algorithm for cyberbullying detection. *International Journal of Advanced Computer Science and Applications*, 9(9), 2018.

[8] A. Albarghouthi and S. Vinitsky. Fairness-aware programming. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 211–219, New York, NY, USA, 2019. ACM.

[9] M. Anderson. A majority of teens have experienced some form of cyberbullying. [Online]. Available from: http://www.pewinternet.org/2018/09/27/a-majority-of-teens-have-experienced-some-form-of-cyberbullying/, 2018.

[10] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias. [Online]. Available from: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing/, 2016.

[11] R. Artstein. *Inter-annotator Agreement*, pages 297–313. Springer Netherlands, Dordrecht, 2017.

[12] Z. Ashktorab and J. Vitak. Designing cyberbullying mitigation and prevention solutions through participatory design with teenagers. In *Proc. of the CHI Conf. on Human Factors in Computing Systems*, pages 3895–3905, 2016.

[13] M. Babaei, A. Chakraborty, E. M. Redmiles, and M. Cha. Analyzing biases in perception of truth in news stories and their implications for fact checking. In *FAT* '19*, 2019.

[14] X. Bai, F. Merenda, C. Zaghi, T. Caselli, and M. Nissim. RuG at GermEval: Detecting Offensive Speech in German Social Media. In J. Ruppenhofer, M. Siegel, and M. Wiegand, editors, *Proceedings of the GermEval 2018 Workshop*, Wien, Austria, 2018. AW Austrian Academy of Sciences.

[15] V. Balakrishnan. Cyberbullying among young adults in malaysia. *Comput. Hum. Behav.*, 46(C):149–157, May 2015.

[16] V. Balakrishnan, S. Khan, T. Fernandez, and H. Arabnia. Cyberbullying detection on twitter using big five and dark triad features. *Personality and Individual Differences*, 141:252–257, 04 2019.

[17] S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning.* fairmlbook.org, 2018.

[18] E. Baron. The rise and fall of microsofts hitler-loving sex robot. [Online]. Available from: `http://www.siliconbeat.com/2016/03/25/the-rise-and-fall-of-microsofts-hitler-loving-sex-robot/`, 2016.

[19] E. BARY. How artificial intelligence could replace credit scores and reshape how we get loans. [Online]. Available from: `https://www.marketwatch.com/story/ai-based-credit-scores-will-soon-give-one-billion-people-access-to-banking-services-2018-10-09`.

[20] S. Bastiaensens, H. Vandebosch, K. Poels, K. V. Cleemput, A. DeSmet, and I. D. Bourdeaudhuij. Cyberbullying on social network sites. an experimental study into bystanders' behavioural intentions to help the victim or reinforce the bully. *Computers in Human Behavior*, 31:259–271, 2014.

[21] P. S. Bayerl and K. I. Paul. What determines inter-coder agreement in manual annotations? a meta-analytic investigation. *Comput. Linguist.*, 37(4):699–725, Dec. 2011.

[22] J. Bayzick, A. Edwards, and L. Edwards. Detecting the presence of cyberbullying using computer software. 02 2019.

[23] A. Bellmore, A. J. Calvin, J.-M. Xu, and X. Zhu. The five W's of bullying on Twitter: Who, what, why, where, and when. *Computers in Human Behavior*, 44:305–314, 2015.

[24] B. Belsey. Cyberbullying: An emerging threat for the Always On generation. [Online]. Available from: `https://www.ucalgary.ca/esa/node/189`, 2005.

[25] S. Benthall and B. D. Haynes. Racial categories in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 289–298, New York, NY, USA, 2019. ACM.

[26] A. Beutel, J. Chen, Z. Zhao, and E. H. Chi. Data decisions and theoretical implications when adversarially learning fair representations. *CoRR*, abs/1707.00075, 2017.

[27] P. K. Bhowmick, P. Mitra, and A. Basu. An agreement measure for determining inter-annotator reliability of human judgements on affective text. In *Proceedings of the Workshop on Human Judgements in Computational Linguistics*, HumanJudge '08, pages 58–65, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[28] A. Bifet and E. Frank. Sentiment knowledge discovery in Twitter streaming data. *Intl. Conf. on Discovery Science*, pages 1–15, 2010.

[29] B. Birkeneder, J. Mitrović, J. Niemeier, T. Leon, and H. Siegfried. upInf - Offensive Language Detection in German Tweets. In J. Ruppenhofer, M. Siegel, and M. Wiegand, editors, *Proceedings of the GermEval 2018 Workshop*, pages 71 – 78, sep 2018.

[30] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, COLT' 98, pages 92–100. ACM, 1998.

[31] T. Bolukbasi, K. Chang, J. Y. Zou, V. Saligrama, and A. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *CoRR*, abs/1607.06520, 2016.

[32] T. Bosse and S. Stam. A normative agent system to prevent cyberbullying. volume 2, pages 425 – 430, 09 2011.

[33] d. boyd. *It's Complicated*. Yale University Press, 2014.

[34] U. Bretschneider, T. Wöhner, and R. Peters. Detecting online harassment in social networks. In *ICIS*, 2014.

[35] R. C. Bunescu and R. J. Mooney. Learning to extract relations from the web using minimal supervision. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, Prague, Czech Republic, June 2007.

[36] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In S. A. Friedler and C. Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91, New York, NY, USA, 23–24 Feb 2018. PMLR.

[37] C. by The Futures Company. 2014 teen internet safety survey. [Online]. Available from: `https://www.cox.com/content/dam/cox/aboutus/documents/tween-internet-safety-survey.pdf`, 2014.

[38] M. A. Campbell. Cyber bullying: An old problem in a new guise? *Australian Journal of Guidance and Counselling*, 15:68–76, 2005.

[39] M. A. Campbell. Cyber bullying: An old problem in a new guise? *Australian Journal of Guidance and Counselling*, 15(1):6876, 2005.

[40] R. Canetti, A. Cohen, N. Dikkala, G. Ramnarayan, S. Scheffler, and A. Smith. From soft classifiers to hard decisions: How fair can we be? In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 309–318, New York, NY, USA, 2019. ACM.

[41] D. Card, M. Zhang, and N. A. Smith. Deep weighted averaging classifiers. *CoRR*, abs/1811.02579, 2018.

[42] L. E. Celis, L. Huang, V. Keswani, and N. K. Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 319–328, New York, NY, USA, 2019. ACM.

[43] S. Chancellor, M. L. Birnbaum, E. D. Caine, V. M. B. Silenzio, and M. De Choudhury. A taxonomy of ethical tensions in inferring mental health states from social media. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 79–88, New York, NY, USA, 2019. ACM.

[44] D. Chatzakou, N. Kourtellis, J. Blackburn, E. D. Cristofaro, G. Stringhini, and A. Vakali. Mean birds: Detecting aggression and bullying on twitter. *CoRR*, abs/1702.06877, 2017.

[45] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali. Detecting aggressors and bullies on Twitter. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, pages 767–768, 2017.

[46] C. Chelmis, D.-S. Zois, and M. Yao. Mining patterns of cyberbullying on twitter. *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 126–133, 2017.

[47] I. Chen, F. D. Johansson, and D. A. Sontag. Why is my classifier discriminatory? In *NeurIPS*, 2018.

[48] J. Chen. Fair lending needs explainable models for responsible recommendation. *CoRR*, abs/1809.04684, 2018.

[49] J. Chen, N. Kallus, X. Mao, G. Svacha, and M. Udell. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 339–348, New York, NY, USA, 2019. ACM.

[50] L.-C. Chen, S. Fidler, A. L. Yuille, and R. Urtasun. Beat the mturkers: Automatic image labeling from weak 3d supervision. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[51] Y. Chen, Y. Zhou, S. Zhu, and H. Xu. Detecting offensive language in social media to protect adolescent online safety. *Intl. Conf. on Social Computing*, pages 71–80, 2012.

[52] L. Cheng, J. Li, Y. N. Silva, D. L. Hall, and H. Liu. Xbully: Cyberbullying detection within a multi-modal context. In *Proceedings of the Twelfth ACM International Conference on Web Searchand Data Mining, WSDM 2019, Melbourne, VIC, Australia, February 11-15, 2019*, pages 339–347, 2019.

[53] A. Chouldechova, D. Benavides-Prado, O. Fialko, and R. Vaithianathan. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81, pages 134–148, 2018.

[54] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[55] R. . Compliance. Can machine learning build a better fico score? [Online]. Available from: `https://www.fico.com/blogs/risk-compliance/can-machine-learning-build-a-better-fico-score/`.

[56] S. Cook. Cyberbullying facts and statistics for 2016-2018. [Online]. Available from: `https://www.comparitech.com/internet-providers/cyberbullying-statistics/`, 2018.

[57] L. Counts. Minority homebuyers face widespread statistical lending discrimination, study finds. [Online]. Available from: `http://newsroom.haas.berkeley.edu/minority-homebuyers-face-widespread-statistical-lending-discrimination-study-finds/`.

[58] H. Cowie. Cyberbullying and its impact on young people's emotional health and well-being. *Psychiatrist*, 37:167–170, 04 2013.

[59] M. Dadvar and F. de Jong. Cyberbullying detection; a step toward a safer internet yard. 04 2012.

[60] M. Dadvar, F. de Jong, R. Ordelman, and D. Trieschnigg. Improved cyberbullying detection using gender information. *Dutch-Belgian Information Retrieval Workshop*, pages 23–25, February 2012.

[61] M. Dadvar and K. Eckert. Cyberbullying detection in social networks using deep learning based models; A reproducibility study. *CoRR*, abs/1812.08046, 2018.

[62] H. Dani, J. Li, and H. Liu. Sentiment informed cyberbullying detection in social media. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2017, Proceedings*, volume 10534 LNAI of *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pages 52–67, Germany, 1 2017. Springer Verlag.

[63] H. Dani, J. Li, and H. Liu. *Sentiment Informed Cyberbullying Detection in Social Media*, pages 52–67. 01 2017.

[64] J. Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. [Online]. Available from: `https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G`, 2018.

[65] A. Datta, M. C. Tschantz, and A. Datta. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *CoRR*, abs/1408.6491, 2014.

[66] A. Datta, M. C. Tschantz, and A. Datta. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *CoRR*, abs/1408.6491, 2014.

[67] T. Davidson, D. Warmsley, M. W. Macy, and I. Weber. Automated hate speech detection and the problem of offensive language. *CoRR*, abs/1703.04009, 2017.

[68] M. De-Arteaga, A. Romanov, H. Wallach, J. Chayes, C. Borgs, A. Chouldechova, S. Geyik, K. Kenthapadi, and A. Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. pages 120–128, 01 2019.

[69] C. DeBrusk. The risk of machine-learning bias (and how to prevent it). [Online]. Available from: `https://sloanreview.mit.edu/article/the-risk-of-machine-learning-bias-and-how-to-prevent-it/`.

[70] W. Dieterich, C. Mendoza, and T. Brennan. Compas risk scales : Demonstrating accuracy equity and predictive parity performance of the compas risk scales in broward county. 2016.

[71] 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, Sept. 2012.

[72] K. Dinakar, R. Reichart, and H. Lieberman. Modeling the detection of textual cyberbullying. *ICWSM Workshop on Social Mobile Web*, 2011.

[73] ditchthelabel.org. The annual cyberbullying survey. *http://www.ditchthelabel.org/*, 2013.

[74] L. Dixon, J. Li, J. Sorensen, N. Thain, and L. Vasserman. Measuring and mitigating unintended bias in text classification. pages 67–73, 12 2018.

[75] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati. Hate speech detection with comment embeddings. In *International Conference on World Wide Web*, pages 29–30, 2015.

[76] N. Dordolo. The role of power imbalance in cyberbullying. *Inkblot: The Undergraduate J. of Psychology*, 3, 2014.

[77] C. Dr. Katzer, D. Fetchenhauer, and F. Belschak. Cyberbullying: Who are the victims? *Journal of Media Psychology: Theories, Methods, and Applications*, 21:25–36, 01 2009.

[78] M. Duggan. Online harassment 2017. [Online]. Available from: `http://www.pewinternet.org/2017/07/11/online-harassment-2017/`, 2017.

[79] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. S. Zemel. Fairness through awareness. *CoRR*, abs/1104.3913, 2011.

[80] C. Dwork, N. Immorlica, A. T. Kalai, and M. Leiserson. Decoupled classifiers for group-fair and efficient machine learning. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pages 119–133, 2018.

[81] C. EDDY. Recognizing cultural bias in ai. [Online]. Available from: `https://peopleofcolorintech.com/articles/recognizing-cultural-bias-in-ai/`, 2017.

[82] B. Edelman and M. Luca. Digital discrimination: The case of airbnb.com. 01 2014.

[83] A. Edwards and A. Leatherman. Chatcoder: Toward the tracking and categorization of internet predators. 3, 01 2009.

[84] A. Edwards, K. Reynolds, A. Garron, and L. Edwards. Detecting cyberbullying: Query terms and techniques. pages 195–204, 05 2013.

[85] C. El Haber. *Ending the torment: tackling bullying from the schoolyard to cyberspace.* 03 2016.

[86] H. Elzayn, S. Jabbari, C. Jung, M. Kearns, S. Neel, A. Roth, and Z. Schutzman. Fair algorithms for learning in allocation problems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 170–179, New York, NY, USA, 2019. ACM.

[87] M. Fekkes, F. Pijpers, A. Miranda Fredriks, T. Vogels, and S. Verloove-Vanhorick. Do bullied children get ill, or do ill children get bullied? a prospective cohort study on the relationship between bullying and health-related symptoms. *Pediatrics*, 117:1568–74, 06 2006.

[88] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 259–268, New York, NY, USA, 2015. ACM.

[89] P. Ferreira, A. M. Veiga Simo, A. Ferreira, S. Souza, and S. Francisco. Student bystander behavior and cultural issues in cyberbullying: When actions speak louder than words. *Computers in Human Behavior*, 60:301–311, 07 2016.

[90] J. L. Fleiss and J. Cohen. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33(3):613–619, 1973.

[91] S. Frenda, B. Ghanem, and M. Montes. Exploration of misogyny in spanish and english tweets. 07 2018.

[92] S. A. Friedler, C. E. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth. A comparative study of fairness-enhancing interventions in machine learning. *CoRR*, abs/1802.04422, 2019.

[93] E. FRIEDMAN. Florida teen live-streams his suicide online. [Online]. Available from: `https://abcnews.go.com/Technology/MindMoodNews/story?id=6306126&page=1`, 2008.

[94] P. Galn-Garca, J. Puerta, C. Laorden, I. Santos, and P. Bringas. *Supervised Machine Learning for the Detection of Troll Profiles in Twitter Social Network: Application to a Real Case of Cyberbullying*, volume 239, pages 419–428. 01 2014.

[95] S. Garg, V. Perot, N. Limtiaco, A. Taly, E. H. Chi, and A. Beutel. Counterfactual fairness in text classification through robustness. *CoRR*, abs/1809.10610, 2018.

[96] R. Goldman. Teens indicted after allegedly taunting girl who hanged herself. [Online]. Available from: `https://abcnews.go.com/Technology/TheLaw/teens-charged-bullying-mass-girl-kill/story?id=10231357`, 2010.

[97] H. Gómez-Adorno, G. B. Enguix, G. Sierra, O. Sánchez, and D. Quezada. A machine learning approach for detecting aggressive tweets in spanish. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018.*, pages 102–107, 2018.

[98] B. Green and Y. Chen. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 90–99, New York, NY, USA, 2019. ACM.

[99] A. Grover and J. Leskovec. node2vec: Scalable feature learning for networks. *CoRR*, abs/1607.00653, 2016.

[100] V. H. Wright, J. J. Burnham, C. T. Inman, and H. N. Ogorchock. Cyberbullying: Using virtual scenarios to educate and raise awareness. *Journal of Computing in Teacher Education*, 26, 01 2009.

[101] B. Haidar, C. Maroun, and A. Serhrouchni. Arabic cyberbullying detection: Using deep learning. pages 284–289, 09 2018.

[102] Y. Halpern, S. Horng, and D. Sontag. Clinical tagging with joint probabilistic models. In F. Doshi-Velez, J. Fackler, D. Kale, B. Wallace, and J. Weins, editors, *Proceedings of the 1st Machine Learning for Healthcare Conference*, volume 56 of *Proceedings of Machine Learning Research*, pages 209–225, Northeastern University, Boston, MA, USA, 18–19 Aug 2016. PMLR.

[103] K. Hao. Ai is sending people to jailand getting it wrong. [Online]. Available from: `https://www.technologyreview.com/s/612775/algorithms-criminal-justice-ai/`, 2019.

[104] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. *CoRR*, abs/1610.02413, 2016.

[105] H. Heidari, M. Loi, K. P. Gummadi, and A. Krause. A moral framework for understanding fair ml through economic models of equality of opportunity. pages 181–190, 01 2019.

[106] S. Hinduja and J. Patchin. Offline consequences of online victimization: School violence and delinquency. *Journal of School Violence*, 9, 01 2007.

[107] S. Hinduja and J. W. Patchin. *Bullying Beyond the Schoolyard: Preventing and Responding to Cyberbullying.* Corwin Press, Incorporated, 2008.

[108] R. Hoffmann, C. Zhang, X. Ling, L. Zettlemoyer, and D. S. Weld. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 541–550, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[109] H. Hosseinmardi, A. Ghasemianlangroodi, R. Han, Q. Lv, and S. Mishra. Towards understanding cyberbullying behavior in a semi-anonymous social network. *IEEE/ACM International Conf. on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 244–252, August 2014.

[110] H. Hosseinmardi, S. Li, Z. Yang, Q. Lv, R. I. Rafiq, R. Han, and S. Mishra. A comparison of common users across Instagram and Ask.fm to better understand cyberbullying. *IEEE Intl. Conf. on Big Data and Cloud Computing*, 2014.

[111] H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra. Analyzing labeled cyberbullying incidents on the Instagram social network. In *Intl. Conf. on Social Informatics*, pages 49–66, 2015.

[112] H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra. Detection of cyberbullying incidents on the Instagram social network. *Association for the Advancement of Artificial Intelligence*, 2015.

[113] L. Hu, N. Immorlica, and J. W. Vaughan. The disparate effects of strategic manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 259–268, New York, NY, USA, 2019. ACM.

[114] M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, pages 168–177, New York, NY, USA, 2004. ACM.

[115] Q. Huang and V. K. Singh. Cyber bullying detection using social and textual analysis. *Proceedings of the International Workshop on Socially-Aware Multimedia*, pages 3–6, 2014.

[116] S. Hunter, J. Boyle, and D. Warden. Perceptions and correlates of peer-victimization and bullying. *The British journal of educational psychology*, 77:797–810, 12 2007.

[117] B. Hutchinson and M. Mitchell. 50 years of test (un)fairness: Lessons for machine learning. *CoRR*, abs/1811.10104, 2018.

[118] O. Irsoy and C. Cardie. Deep recursive neural networks for compositionality in language. In *NIPS*, 2014.

[119] A. C. Islam, J. J. Bryson, and A. Narayanan. Semantics derived automatically from language corpora necessarily contain human biases. *CoRR*, abs/1608.07187, 2016.

[120] R. Jain. When recommendation systems go bad. [Online]. Available from: `https://cds.nyu.edu/recommendation-systems-go-bad-%E2%80%A8/`, 2016.

[121] J. Juvonen and E. F Gross. Extending the school grounds?-bullying experiences in cyberspace. *The Journal of school health*, 78:496–505, 10 2008.

[122] J. Juvonen and S. Graham. Bullying in schools: The power of bullies and the plight of victims. *Annual review of psychology*, 65, 08 2013.

[123] J. Juvonen and S. A. Graham. Bullying in schools: the power of bullies and the plight of victims. *Annual review of psychology*, 65:159–85, 2014.

[124] J. Juvonen and E. F. Gross. Extending the school grounds?bullying experiences in cyberspace. *Journal of School Health*, 78(9):496–505, 2008. Exported from https://app.dimensions.ai on 2019/02/17.

[125] N. Kallus and A. Zhou. Residual unfairness in fair machine learning from prejudiced data. In *ICML*, 2018.

[126] A. Kalyuzhnaya, N. Nikitin, N. Butakov, and D. Nasonov. *Precedent-Based Approach for the Identification of Deviant Behavior in Social Media*, pages 846–852. 2018.

[127] S. Kannan, A. Roth, and J. Ziani. Downstream effects of affirmative action. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 240–248, New York, NY, USA, 2019. ACM.

[128] A. Kaplan and M. Haenlein. Users of the world, unite! the challenges and opportunities of social media. *Business Horizons*, 53:59–68, 02 2010.

[129] M. Kay, C. Matuszek, and S. A. Munson. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 3819–3828, New York, NY, USA, 2015. ACM.

[130] N. Kilbertus, M. Rojas-Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Scholkopf. Avoiding discrimination through causal reasoning. *CoRR*, abs/1706.02744, 2017.

[131] M. P. Kim, A. Ghorbani, and J. Y. Zou. Multiaccuracy: Black-box post-processing for fairness in classification. *CoRR*, abs/1805.12317, 2018.

[132] S. K. Kim and N. S. Kim. The role of the pediatrician in youth violence prevention. *Korean J Pediatr*, 56(1):1–7, Jan 2013. 23390438[pmid].

[133] W. Kirwin. Implicit recommender systems: Biased matrix factorization. [Online]. Available from: `http://activisiongamescience.github.io/2016/01/11/Implicit-Recommender-Systems-Biased-Matrix-Factorization/`, 2016.

[134] K. H. Kopasz and P. R. Smokowski. Bullying in School: An Overview of Types, Effects, Family Characteristics, and Intervention Strategies. *Children and Schools*, 27(2):101–110, 04 2005.

[135] R. M. Kowalski, S. P. Limber, and P. W. Agatston. *Cyberbullying: Bullying in the Digital Age*. John Wiley & Sons, 2012.

[136] R. Kumar, A. N. Reganti, A. Bhatia, and T. Maheshwari. Aggression-annotated corpus of hindi-english code-mixed data. *CoRR*, abs/1803.09402, 2018.

[137] D. L. Hoff and S. N. Mitchell. Cyberbullying: Causes, effects, and remedies. *Journal of Educational Administration*, 47:652–665, 08 2009.

[138] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

[139] V. Lavrenko and W. B. Croft. Relevance based language models. In *Proc. of the International ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 120–127, 2001.

[140] F. Law. Cyberbullying and social media. [Online]. Available from: `https://education.findlaw.com/student-conduct-and-discipline/cyberbullying-and-social-media.html`.

[141] Q. Le and T. Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1188–1196, 2014.

[142] T. Lei, H. Joshi, R. Barzilay, T. S. Jaakkola, K. Tymoshenko, A. Moschitti, and L. M. i Villodre. Denoising bodies to titles: Retrieving similar questions with recurrent convolutional models. *CoRR*, abs/1512.05726, 2015.

[143] D. Lessne and C. Yanez. Student reports of bullying: Results from the 2015 school crime supplement to the national crime victimization survey. [Online]. Available from: `https://nces.ed.gov/pubs2017/2017015.pdf`, 2016.

[144] S. Levin. A beauty contest was judged by ai and the robots didn't like dark skin. [Online]. Available from: `https://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people`, 2016.

[145] Q. Li. New bottle but old wine: A research of cyberbullying in schools. computers in human bahavior, 23. 1777-1191. *Computers in Human Behavior*, 23:1777–1791, 07 2007.

[146] Q. Li. A cross-cultural comparison of adolescents experience related to cyberbullying. *Educational Research - EDUC RES*, 50:223–234, 09 2008.

[147] Z. Li, J. Kawamoto, Y. Feng, and K. Sakurai. Cyberbullying detection using parent-child relationship between comments. pages 325–334, 11 2016.

[148] S. Livingstone, L. Haddon, J. Vincent, G. Mascheroni, and K. Olafsson. Net children go mobile: The uk report. [Online]. Available from: `http://netchildrengomobile.eu/ncgm/wp-content/uploads/2013/07/NCGM_UK-Report_FINAL-VERSION.pdf`, 2014.

[149] N. MacBryde. Rail death teen threatened online. [Online]. Available from: `https://www.bbc.com/news/uk-england-hereford-worcester-14239702`, 2011.

[150] D. Madras, E. Creager, T. Pitassi, and R. Zemel. Fairness through causal awareness: Learning causal latent-variable models for biased data. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 349–358, New York, NY, USA, 2019. ACM.

[151] A. Mahendiran, W. Wang, J. Arredondo, B. Huang, L. Getoor, D. Mares, and N. Ramakrishnan. Discovering evolving political vocabulary in social media. In *Intl. Conf. on Behavioral, Economic, and Socio-Cultural Computing*, 2014.

[152] D. Maher. Cyberbullying: An ethnographic case study of one australian upper primary school class. *Youth Studies Australia*, 27, 01 2008.

[153] P. Maitra and R. Sarkhel. A k-competitive autoencoder for aggression detection in social media text. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 80–89. Association for Computational Linguistics, 2018.

[154] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval.* Cambridge University Press, 2008.

[155] H. Margono, X. Yi, and G. K. Raikundalia. Mining Indonesian cyber bullying patterns in social networks. *Proc. of the Australasian Computer Science Conference*, 147, January 2014.

[156] B. Martin. Mobbing: Emotional abuse in the american workplace. *Journal of Organizational Change Management*, 13:401–446, 08 2000.

[157] K. Massoudi, M. Tsagkias, M. de Rijke, and W. Weerkamp. Incorporating query expansion and quality indicators in searching microblog posts. *Proc. of the European Conference on Advances in Information Retrieval*, 15(5):362–367, November 2011.

[158] I. McGhee, J. Bayzick, A. Kontostathis, L. Edwards, A. McBride, and E. Jakubowski. Learning to identify internet sexual predation. *Intl. J. of Electronic Commerce*, 15(3):103–122, 2011.

[159] E. Menesini and A. Nocentini. Cyberbullying definition and measurement: Some critical considerations. *Zeitschrift Fur Psychologie-journal of Psychology - Z PSYCHOL*, 217:230–232, 01 2009.

[160] A. K. Menon and R. C. Williamson. The cost of fairness in binary classification. In *Proc. of the Conf. on Fairness, Accountability and Transparency*, pages 107–118, 2018.

[161] D. Meyer. The gentle neoliberalism of modern anti-bullying texts: Surveillance, intervention, and bystanders in contemporary bullying discourse. *Sexuality Research and Social Policy*, 13, 05 2016.

[162] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

[163] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119, 2013.

[164] T. Mikolov, W. tau Yih, and G. Zweig. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, 2013.

[165] S. Milli, J. Miller, A. D. Dragan, and M. Hardt. The social cost of strategic classification. *CoRR*, abs/1808.08460, 2019.

[166] M. Mintz, S. Bills, R. Snow, and D. Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 1003–1011, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[167] T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. Platanios, A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, and J. Welling. Never-ending learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 2302–2310. AAAI Press, 2015.

[168] M. A. Moreno. CyberbullyingJAMA Pediatrics Patient Page. *JAMA Pediatrics*, 168(5):500–500, 05 2014.

[169] H. Mubarak, K. Darwish, and W. Magdy. Abusive language detection on arabic social media. pages 52–56, 2017.

[170] M. Munezero, C. Suero Montero, T. Kakkonen, E. Sutinen, M. Mozgovoy, and V. Klyuev. Automatic detection of antisocial behaviour in texts. *Informatica (Slovenia)*, 38:3–10, 03 2014.

[171] V. Nahar, S. AL Maskari, X. Li, and C. Pang. Semi-supervised learning for cyberbullying detection in social networks. pages 160–171, 07 2014.

[172] V. Nahar, X. Li, and C. Pang. An effective approach for cyberbullying detection. *Communications in Information Science and Management Engineering*, 3(5):238–247, May 2013.

[173] V. Nahar, S. Unankard, X. Li, and C. Pang. Sentiment analysis for effective detection of cyber bullying. pages 767–774, 04 2012.

[174] G. Nalinipriya and M. Asswini. A dynamic cognitive system for automatic detection and prevention of cyber-bullying attacks. *ARPN Journal of Engineering and Applied Sciences*, 10:4618–4626, 01 2015.

[175] E. Nalisnick, B. Mitra, N. Craswell, and R. Caruana. Improving document ranking with dual word embeddings. In *WWW'16*. WWW - World Wide Web Consortium (W3C), April 2016.

[176] national bullying center in UK. Bullying uk national bullying survey. [Online]. Available from: `https://www.bullying.co.uk/anti-bullying-week/bullying-uk-national-survey-2014/`.

[177] K. Nigam, A. Mccallum, and T. M. Mitchell. Semi-supervised text classification using em. 09 2006.

[178] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang. Abusive language detection in online user content. In *Proceedings Intl. Conf. on World Wide Web*, pages 145–153, 2016.

[179] noswearing.com. List of swear words & curse words. *http://www.noswearing.com/dictionary*, 2016.

[180] Z. Obermeyer and S. Mullainathan. Dissecting racial bias in an algorithm that guides health decisions for 70 million people. pages 89–89, 01 2019.

[181] D. Olweus. *Bullying at School: What We Know and What We Can Do*. Understanding Children's Worlds. Wiley, 1993.

[182] D. Olweus. Cyberbullying: An overrated phenomenon? *European Journal of Developmental Psychology - EUR J DEV PSYCHOL*, 9:1–19, 09 2012.

[183] J. W. Patchin. Summary of our cyberbullying research (2004-2016). [Online]. Available from: `https://cyberbullying.org/summary-of-our-cyberbullying-research`, 2016.

[184] J. W. Patchin and S. Hinduja. *Cyberbullying Prevention and Response: Expert Perspectives*. Routledge, 2012.

[185] D. U. Patton, K. McKeown, O. Rambow, and J. Macbeth. Using natural language processing and qualitative analysis to intervene in gang violence. *arXiv preprint arXiv:1609.08779*, 2016.

[186] J. W. Pennebaker, M. E. Francis, and R. J. Booth. Linguistic inquiry and word count: LIWC. *Mahway: Lawrence Erlbaum Associates*, 2001.

[187] K. Petrasic and B. Saul. Algorithms and bias: What lenders need to know. [Online]. Available from: `https://www.whitecase.com/publications/insight/algorithms-and-bias-what-lenders-need-know`.

[188] J. PHELPS. First lady melania trump speaks out against cyberbullying. [Online]. Available from: `https://abcnews.go.com/Politics/lady-melania-trump-speaks-cyberbullying/story?id=57284988`, 2018.

[189] G. K. Pitsilis, H. Ramampiaro, and H. Langseth. Detecting Offensive Language in Tweets Using Deep Learning. *ArXiv e-prints*, Jan. 2018.

[190] E. A. Platanios, H. Poon, T. M. Mitchell, and E. Horvitz. Estimating accuracy from unlabeled data: A probabilistic logic approach. *CoRR*, abs/1705.07086, 2017.

[191] B. Plomion. Does artificial intelligence discriminate? [Online]. Available from: `https://www.forbes.com/sites/forbescommunicationscouncil/2017/05/02/does-artificial-intelligence-discriminate/#6fb0fbb430bc`, 2017.

[192] V. Polonski. Mitigating algorithmic bias in predictive justice: 4 design principles for ai fairness. [Online]. Available from: `https://towardsdatascience.com/mitigating-algorithmic-bias-in-predictive-justice-ux-design-principles-for-ai-fairness-machine-learning-d2227ce28099`, 2018.

[193] R. Price. Microsoft is deleting its ai chatbot's incredibly racist tweets. [Online]. Available from: `https://www.businessinsider.com/microsoft-deletes-racist-genocidal-tweets-from-ai-chatbot-tay-2016-3`, 2016.

[194] T. E. G. D. Protection. The eu general data protection. [Online]. Available from: https://eugdpr.org/, 2018.

[195] M. Ptaszynski, P. Dybala, T. Matsuba, F. Masui, R. Rzepka, and K. Araki. Machine learning and affect analysis against cyber-bullying. In *Linguistic and Cognitive Approaches to Dialog Agents Symposium*, pages 7–16, 2010.

[196] R. I. Rafiq, H. Hosseinmardi, R. Han, Q. Lv, and S. Mishra. Scalable and timely detection of cyberbullying in online social networks. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, SAC '18, pages 1738–1747, New York, NY, USA, 2018. ACM.

[197] E. Raisi and B. Huang. Cyberbullying detection with weakly supervised machine learning. In *Proceedings of the IEEE/ACM International Conference on Social Networks Analysis and Mining*, 2017.

[198] N. Ramakrishnan, P. Butler, N. Self, R. Khandpur, P. Saraf, W. Wang, J. Cadena, A. Vullikanti, G. Korkmaz, C. Kuhlman, A. Marathe, L. Zhao, H. Ting, B. Huang, A. Srinivasan, K. Trinh, L. Getoor, G. Katz, A. Doyle, C. Ackermann, I. Zavorin, J. Ford, K. Summers, Y. Fayed, J. Arredondo, D. Gupta, and D. Mares. 'Beating the news' with EMBERS: Forecasting civil unrest using open source indicators. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1799–1808, 2014.

[199] S. Ravi and Q. Diao. Large scale distributed semi-supervised learning using streaming approximation. *CoRR*, abs/1512.01752, 2015.

[200] M. REIDSMA. Algorithmic bias in library discovery systems. [Online]. Available from: https://matthew.reidsrow.com/articles/173, 2016.

[201] ReportLinker. For young generations, standing up to the cyberbully is now a life skill. [Online]. Available from: https://www.reportlinker.com/insight/americas-youth-cyberbully-life-skill.html, 2017.

[202] K. Reynolds, A. Kontostathis, and L. Edwards. Using machine learning to detect cyberbullying. *Intl. Conf. on Machine Learning and Applications and Workshops (ICMLA)*, 2:241–244, 2011.

[203] M. Rezvan, S. Shekarpour, L. Balasuriya, K. Thirunarayan, V. L. Shalin, and A. Sheth. A quality type-aware annotated corpus and lexicon for harassment research. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '18, pages 33–36, New York, NY, USA, 2018. ACM.

[204] M. Rezvan, S. Shekarpour, K. Thirunarayan, V. L. Shalin, and A. P. Sheth. Analyzing and learning the language for different types of harassment. *CoRR*, abs/1811.00644, 2018.

[205] M. H. Ribeiro, P. H. Calais, Y. A. Santos, V. A. F. Almeida, and W. M. Jr. "like sheep among wolves": Characterizing hateful users on twitter. *ArXiv e-prints*, Jan. 2018.

[206] S. Riedel, L. Yao, and A. McCallum. Modeling relations and their mentions without labeled text. In *Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases: Part III*, ECML PKDD'10, pages 148–163, Berlin, Heidelberg, 2010. Springer-Verlag.

[207] R. Rieland. Artificial intelligence is now used to predict crime. but is it biased? [Online]. Available from: `https://www.smithsonianmag.com/innovation/artificial-intelligence-is-now-used-predict-crime-is-it-biased-180968337/`, 2018.

[208] H. Rosa, J. Carvalho, P. Calado, B. Martins, R. Ribeiro, and L. Coheur. Using fuzzy fingerprints for cyberbullying detection in social networks. pages 1–7, 07 2018.

[209] C. Russell. Efficient search for diverse coherent explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 20–28, New York, NY, USA, 2019. ACM.

[210] H. J. Ryu, M. Mitchell, and H. Adam. Improving smiling detection with race and gender diversity. *CoRR*, abs/1712.00193, 2017.

[211] S. Salawu, Y. He, and J. Lumsden. Approaches to automated detection of cyberbullying: A survey. *IEEE Transactions on Affective Computing*, PP:1–1, 10 2017.

[212] H. Sanchez and S. Kumar. Twitter bullying detection. 02 2019.

[213] J. M. Schneider, R. Roller, P. Bourgonje, S. Hegele, and G. Rehm. Towards the automatic classification of offensive language and related phenomena in german tweets. In J. Ruppenhofer, M. Siegel, and M. Wiegand, editors, *Proceedings of the GermEval 2018 Workshop*. Online, 2018.

[214] S. Serra and H. Venter. Mobile cyber-bullying: A proposal for a pre-emptive approach to risk mitigation by employing digital forensic readiness. pages 1 – 5, 09 2011.

[215] E. K. Shriver. Focus on children's mental health research at the nichd. [Online]. Available from: `http://www.nichd.nih.gov/news/resources/spotlight/060112-childrens-mental-health`.

[216] E. K. Shriver. Taking a stand against bullying. [Online]. Available from: `http://www.nichd.nih.gov/news/resources/spotlight/092110-taking-stand-against-bullying`.

[217] T. H. Silva, P. O. de Melo, J. M. Almeida, J. Salles, and A. A. Loureiro. A picture of Instagram is worth more than a thousand words: Workload characterization and application. *DCOSS*, pages 123–132, 2013.

[218] C. Sinders. Toxicity and tone are not the same thing: analyzing the new Google API on toxicity, PerspectiveAPI. https://medium.com/@carolinesinders/toxicity-and-tone-are-not-the-same-thing-analyzing-the-new-google-api-on-toxicity-perspectiveapi-14abe4e728b3.

[219] V. K. Singh, Q. Huang, and P. K. Atrey. Cyberbullying detection using probabilistic socio-textual information fusion. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, volume 00, pages 884–887, Aug. 2016.

[220] R. Slonje and P. K. Smith. Cyberbullying: another main type of bullying? *Scandinavian journal of psychology*, 49 2:147–54, 2008.

[221] R. Slonje, P. K. Smith, and A. FriséN. The nature of cyberbullying, and strategies for prevention. *Comput. Hum. Behav.*, 29(1):26–32, Jan. 2013.

[222] K. Smith. 122 amazing social media statistics and facts. [Online]. Available from: `https://www.brandwatch.com/blog/amazing-social-media-statistics-and-facts/`, 2019.

[223] M. SMITH, D. PATIL, and C. MUOZ. Big risks, big opportunities: the intersection of big data and civil rights. [Online]. Available from: `https://obamawhitehouse.archives.gov/blog/2016/05/04/big-risks-big-opportunities-intersection-big-data-and-civil-rights`, 2016.

[224] P. Smith, C. Barrio, and R. Tokunaga. Definitions of bullying and cyberbullying: How useful are the terms? *Principles of cyberbullying research: Definitions, measures, and methodology*, pages 26–40, 01 2013.

[225] P. Smith and S. Sharp. *School Bullying: Insights and Perspectives.* Routledge, 1994.

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

[227] L. Smith-Spark. Hanna Smith suicide fuels calls for action on ask.fm cyberbullying. *http://www.cnn.com/2013/08/07/world /europe/uk-social-media-bullying/*, 2013.

[228] D. Soni and V. K. Singh. See no evil, hear no evil: Audio-visual-textual cyberbullying detection. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW):164:1–164:26, 2018.

[229] A. Sourander, A. Brunstein Klomek, M. Ikonen, J. Lindroos, T. Luntamo, M. Koskelainen, T. Ristkari, and H. Helenius. Psychosocial risk factors associated with cyberbullying among adolescents: A population-based study. *Archives of general psychiatry*, 67:720–8, 07 2010.

[230] V. Southerland. With ai and criminal justice, the devil is in the data. [Online]. Available from: `https://www.aclu.org/issues/privacy-technology/surveillance-technologies/ai-and-criminal-justice-devil-data`, 2018.

[231] B. Spice. Questioning the fairness of targeting ads online. [Online]. Available from: `https://www.cmu.edu/news/stories/archives/2015/july/online-ads-research.html`, 2015.

[232] A. Squicciarini, S. Rajtmajer, Y. Liu, and C. Griffin. Identification and characterization of cyberbullying dynamics in an online social network. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, ASONAM '15, pages 280–285, New York, NY, USA, 2015. ACM.

[233] B. Sri Nandhini and S. Immanuvelrajkumar. Online social network bullying detection using intelligence techniques. *Procedia Computer Science*, 45:485–492, 12 2015.

[234] L. Sweeney. Discrimination in online ad delivery. *Commun. ACM*, 56(5):44–54, May 2013.

[235] N. Tahmasbi and E. Rastegari. A socio-contextual approach in automated detection of public cyberbullying on twitter. *Trans. Soc. Comput.*, 1(4):15:1–15:22, Dec. 2018.

[236] F. Tapia, C. Aguinaga, and R. Luje. Detection of behavior patterns through social networks like twitter, using data mining techniques as a method to detect cyberbullying. pages 111–118, 10 2018.

[237] D. the Label. The annual bullying survey 2013. [Online]. Available from: `http://www.ditchthelabel.org/annual-cyber-bullying-survey-cyber-bullying-statistics/`, 2013.

[238] D. the Label. The annual bullying survey 2017. [Online]. Available from: `https://www.ditchthelabel.org/wp-content/uploads/2017/`, 2017.

[239] O. o. t. P. S. The White House. Background on white house conference on bullying prevention. [Online]. Available from: `https://obamawhitehouse.archives.gov/the-press-office/2011/03/10/background-white-house-conference-bullying-prevention`.

[240] R. Tokunaga. Following you home from school: A critical review and synthesis of research on cyberbullying victimization. *Computers in Human Behavior*, 26:277–287, 05 2010.

[241] R. S. Tokunaga. Following you home from school: A critical review and synthesis of research on cyberbullying victimization. *Computers in Human Behavior*, 26(3):277–287, 2010.

[242] S. Tomkins, L. Getoor, Y. Chen, and Y. Zhang. Detecting cyber-bullying from sparse data and inconsistent labels. In *NeurIPS Workshop on Learning with Limited Labeled Data*, 2017.

[243] S. Tomkins, L. Getoor, Y. Chen, and Y. Zhang. A socio-linguistic model for cyberbullying detection. In *Intl. Conf. on Advances in Social Networks Analysis and Mining*, 2018.

[244] M. Tosik, C. L. Hansen, G. Goossen, and M. Rotaru. Word embeddings vs word types for sequence labeling: the curious case of cv parsing. In *VS@HLT-NAACL*, 2015.

[245] V. Tsintzou, E. Pitoura, and P. Tsaparas. Bias disparity in recommendation systems, 11 2018.

[246] B. Ustun, A. Spangher, and Y. Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 10–19, New York, NY, USA, 2019. ACM.

[247] J. van der Zwaan, V. Dignum, and C. Jonker. Simulating peer support for victims of cyberbullying. *New Phytologist - NEW PHYTOL*, 01 2010.

[248] C. Van Hee, G. Jacobs, C. Emmery, B. Desmet, E. Lefever, B. Verhoeven, G. De Pauw, W. Daelemans, and V. Hoste. Automatic Detection of Cyberbullying in Social Media Text. *ArXiv e-prints*, Jan. 2018.

[249] C. Van Hee, E. Lefever, B. Verhoeven, J. Mennes, B. Desmet, G. De Pauw, W. Daelemans, and V. Hoste. Automatic detection and prevention of cyberbullying. 10 2015.

[250] K. Van Royen, K. Poels, W. Daelemans, and H. Vandebosch. Automatic monitoring of cyberbullying on social networking sites: From technological feasibility to desirability. *Telematics and Informatics*, 32, 01 2014.

[251] H. Vandebosch and K. Cleemput. Cyberbullying among youngsters: Prevalence and profile of bullies and victims. *New Media and Society*, 11:1–23, 01 2009.

[252] H. Vandebosch and K. V. Cleemput. Cyberbullying among youngsters: profiles of bullies and victims. *New Media & Society*, 11(8):1349–1371, 2009.

[253] S. Wachs, W. Schubarth, A. Seidel, and E. Piskunova. *Detecting and Interfering in Cyberbullying Among Young People (Foundations and Results of German Case-Study): Third International Conference, DTGS 2018, St. Petersburg, Russia, May 30 June 2, 2018, Revised Selected Papers, Part II*, pages 277–285. 05 2018.

[254] K. WADDELL. How algorithms can bring down minorities' credit scores. [Online]. Available from: `https://www.theatlantic.com/technology/archive/2016/12/how-algorithms-can-bring-down-minorities-credit-scores/509333/`.

[255] C. Wagner, D. García, M. Jadidi, and M. Strohmaier. It's a man's wikipedia? assessing gender inequality in an online encyclopedia. *CoRR*, abs/1501.06307, 2015.

[256] J. Wakefield. Microsoft chatbot is taught to swear on twitter. [Online]. Available from: `https://www.bbc.com/news/technology-35890188`, 2016.

[257] J. Wang, R. J. Iannotti, and T. R. Nansel. School bullying among adolescents in the united states: physical, verbal, relational, and cyber. *J Adolesc Health*, 45(4):368–375, Oct 2009. 19766941[pmid].

[258] W. Warner and J. Hirschberg. Detecting hate speech on the world wide web. In *Workshop on Language in Social Media*, pages 19–26, 2012.

[259] K. Weinberger, A. Dasgupta, J. Langford, A. Smola, and J. Attenberg. Feature hashing for large scale multitask learning. In *Proc. of the Intl. Conf. on Machine Learning*, pages 1113–1120, 2009.

[260] N. E. Willard. *Cyberbullying and Cyberthreats: Responding to the Challenge of Online Social Aggression, Threats, and Distress.* Research Publishers LLC, 2nd edition, 2007.

[261] M. F. Wright. Cyberbullying in cultural context. *Journal of Cross-Cultural Psychology*, 48(8):1136–1137, 2017.

[262] C. Xu, D. Tao, and C. Xu. A survey on multi-view learning. *CoRR*, abs/1304.5634, 2013.

[263] C. Xu, D. Tao, and C. Xu. Multi-view learning with incomplete views. *IEEE Transactions on Image Processing*, 24:5812–5825, 2015.

[264] J. Xu, A. Schwing, and R. Urtasun. *Tell me what you see and i will show you where it is*, pages 3190–3197. IEEE Computer Society, 9 2014.

[265] J.-M. Xu, K.-S. Jun, X. Zhu, and A. 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*, NAACL HLT '12, pages 656–666, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[266] J.-M. Xu, X. Zhu, and A. Bellmore. Fast learning for sentiment analysis on bullying. In *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining*, WISDOM '12, pages 10:1–10:6, New York, NY, USA, 2012. ACM.

[267] L. Yao, S. Riedel, and A. McCallum. Collective cross-document relation extraction without labelled data. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1013–1023, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[268] M. Yao, C. Chelmis, and D.-S. Zois. Cyberbullying detection on instagram with optimal online feature selection. In *IEEE/ACM International Conf. on Advances in Social Networks Analysis and Mining*, pages 401–408, 2018.

[269] S. Yao and B. Huang. Beyond parity: Fairness objectives for collaborative filtering. *CoRR*, abs/1705.08804, 2017.

[270] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards. Detection of harassment on Web 2.0. *Content Analysis in the WEB 2.0*, 2009.

[271] M. Young, L. Rodriguez, E. Keller, F. Sun, B. Sa, and J. Whittington. Beyond open vs. closed: Balancing individual privacy and public accountability in data sharing. In *FAT* '19*, 2019.

[272] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment &#38; disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pages 1171–1180, Republic and Canton of Geneva, Switzerland, 2017. International World Wide Web Conferences Steering Committee.

[273] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In S. Dasgupta and D. McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 325–333, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[274] B. H. Zhang, B. Lemoine, and M. Mitchell. Mitigating unwanted biases with adversarial learning. *CoRR*, abs/1801.07593, 2018.

[275] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K. Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *CoRR*, abs/1707.09457, 2017.

[276] J. Zhao, X. Xie, X. Xu, and S. Sun. Multi-view learning overview: Recent progress and new challenges. *Information Fusion*, 38:43–54, 2017.

[277] R. Zhao and K. Mao. Cyberbullying detection based on semantic-enhanced marginalized denoising auto-encoder. *IEEE Transactions on Affective Computing*, PP:1–1, 02 2016.

[278] J. Zhu, J. Mao, and A. L. Yuille. Learning from weakly supervised data by the expectation loss svm (e-svm) algorithm. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1125–1133. Curran Associates, Inc., 2014.

[279] D.-S. Zois, A. Kapodistria, M. Yao, and C. Chelmis. Optimal online cyberbullying detection. *2018 IEEE Intl. Conf. on Acoustics, Speech and Sig. Proc.*, pages 2017–2021, 2018.