

NLP AND MACHINE LEARNING TECHNIQUES TO DETECT
ONLINE HARASSMENT ON SOCIAL NETWORKING
PLATFORMS

by

Sima Sharifrad

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
August 2019

© Copyright by Sima Sharifrad, 2019

Table of Contents

List of Tables	vi
List of Figures	viii
Abstract	ix
Acknowledgements	x
Chapter 1 Introduction	1
1.1 Motivation	2
1.2 Problem	4
1.3 Contribution of Thesis	4
1.4 Outline	6
1.5 Published and Submitted Papers	7
Chapter 2 Background and Related Work	9
2.1 Online Harassment and Abuse on Twitter	9
2.1.1 Different Definitions and Categories of Online Harassment	11
2.1.2 Available Datasets	14
2.1.3 Crowdsourcing Platforms	14
2.2 Topic Modeling	15
2.3 Classifiers	17
2.4 Affect Classification	18
2.4.1 Available Datasets	21
2.5 ConceptNet and WikiData	22
2.6 Neural Networks for Text Classification	24
2.6.1 Recurrent Neural Networks	25
2.6.2 Convolutional Neural Networks	25
2.6.3 Word Vector Representation	27
2.7 Text Generation and Augmentation	28
2.8 Summary	29

Chapter 3	Dataset Construction and Classification	30
3.1	Introduction	30
3.2	A Dataset for Online Harassment	31
3.3	Pilot Study	34
3.4	Experiments	35
3.4.1	Data Annotation	35
3.4.2	Setup	37
3.4.3	Topic Modeling	38
3.4.4	Word Embedding Space	39
3.5	Results	40
3.6	Discussion	41
3.7	Summary	41
Chapter 4	Understanding Different Types of Online Harassment Using Convolutional Filters	42
4.1	Goal	42
4.2	Introduction	43
4.3	Setup	43
4.4	CNN Interpretation	44
4.4.1	Results	48
4.5	Discussion and Conclusion	49
Chapter 5	Understanding Emotion Intensity and Emotion Type in Online Harassment Tweets	51
5.1	Goal	51
5.2	Introduction	51
5.2.1	Text Preprocessing	53
5.3	Experiment	53
5.3.1	Experimental setup	53
5.4	Results	55
5.5	Discussion	58

5.6	Summary	59
Chapter 6	Boosting Text Classification Performance on Sexist Tweets by Text Augmentation and Text Generation Using a Com- bination of Knowledge Graphs	60
6.1	Goal	60
6.2	Introduction	61
6.3	Text Generation	64
6.3.1	Text Augmentation	65
6.4	Results	66
6.4.1	Experimental setup	66
6.4.2	Text generation results	67
6.5	Summary	69
Chapter 7	A Case Study: Using Attention-based Bidirectional LSTM to Identify Different Categories of Offensive Language Directed Toward Female Celebrities	70
7.1	Goal	70
7.2	Introduction	71
7.3	Data sets	73
7.3.1	Vulgar Language Types	73
7.3.2	Intensity Types	73
7.3.3	Analyzing Different Harassment Types	74
7.4	Data Pre-processing	75
7.5	Model	76
7.6	Experiment	77
7.6.1	Analyzing Different Vulgar Language Types	77
7.6.2	Analyzing the Intensity of Tweets	77
7.6.3	Analyzing the Harassment Tweets	78
7.7	Data Around Famous Female Figures	78
7.8	Conclusion	79
Chapter 8	Conclusion and Future work	81
8.1	Conclusion	81

8.2 Future work	82
Bibliography	84

List of Tables

2.1	Available Dataset for Hate Speech Detection	15
3.2	Dataset Label Distribution.	35
3.3	Details of Each Task and the Scales	36
3.4	Gender Contribution and the Confidence Score	37
3.5	The First Most Frequent LDA Topics in Each Category	39
3.6	Closest Words to Women and Men in Embedding Space	40
3.7	Classification Results on Accuracy	40
3.8	F-measure Score of Classes of Harassment Using Glove and CNN	41
4.1	Information Captured from Filter #0, Size 3 for the Indirect Harassment Class.	45
4.2	Information Captured from Filter #12, Size 3 for the Sexual Harassment	46
4.3	Information Captured from Filter #8, Size 3 for the Physical Harassment Class	47
4.4	Accuracy of Classifiers on Original and Augmented Dataset.	49
5.1	Distribution of Emotion Intensity for Anger, Joy, Sadness and Fear [87]	54
5.2	Accuracy of Methods on Emotion Intensity on Anger and Fear	55
5.3	Accuracy of Methods on Emotion Intensity on Joy and Sadness	56
5.4	Accuracy of Emotion type detection	56
5.5	Final Results of Emotion Intensity on Harassment Categories	57
5.6	Final Results of Emotion Type in Each Category	58
7.2	Description of Intensity Dataset [87]	74
7.1	Description of Different Vulgar intentions [53]	74

7.3	Description of Different Harassment Types presented in chapter 3 [119].	75
7.4	Results on Different Vulgar Language Types.	77
7.5	Results on Intensity of Dataset	77
7.6	Description of Ten Famous Female Figures	78
7.7	Experimental results of ten collected datasets from famous female figures on three available datasets	80

List of Figures

1.1	High level work in the thesis	7
3.1	Submission Rate by the Country..	36
5.1	Schema.	52
6.1	Wikidata Knowledge Graph.	63
6.2	ConceptNet Knowledge graph.	63
6.3	Combination of Wikidata and ConceptNet knowledge graphs .	64

Abstract

Social media has become an unavoidable part of our daily lives. It attracts different users with different mindsets. In particular, Twitter is a platform with a diverse audience who engage in different topics and interact with personalities from all walks of life. Even though Twitter may be considered as a mere reflection of the discourse between people, it paves a way for certain types of hostile behavior. Twitter not only amplifies marginalized voices but also harassment. Often this harassment is directed towards female users and the purpose is to silence women by threatening, insulting, ignoring or driving them away from Twitter, making it simply the newest wrinkle in that long history of exclusion from public spaces and conversations. Although there exist several recent articles focusing on toxicity, hateful speech and cyberbullying, they do not focus on women as their target. Due to the nature of Twitter, collecting tweets which represent online harassment requires specific filtering of the hashtags and the final dataset is usually scarce, valuable and imbalanced. In addition, these tweets carry different emotions and are enforced on the victims with different intensities. Therefore, understanding this language can help us flag the tweets, understand the mental state of the users and train machine learning algorithms which can help us detect and understand this language better. In this thesis, we visit the problem of toxicity and hateful speech focusing on women as the audience and trying to understand different manifestations of it using machine learning and natural language processing techniques. The problem was formulated as a classification task and a dataset was formed for this purpose. After that, methods were proposed to increase the quality of the classification and test the effectual state of the users and a real case study was examined.

Acknowledgements

I would like to express my sincere thanks and appreciation to my supervisor Professor Stan Matwin, who made all this possible by being consistently optimistic and supportive. His encouragement, concerns, ideas and support from the initial stage of this research to the ending point shed light on the way ahead. It was an honour and happiness to work with you and learn from you.

I would also like to share my gratitude with my dear and sweet parents, Batool and Faramarz, who have been sad and happy with me toward all the days and years and without calling them every morning and hearing their voice, this would not be possible. I owe you all I have.

I would like to thank my partner, Samuel, who gives me peace, love and all of his kindness every day. Who teaches me the art of meditation and self-awareness, who has always been by my side and supports me in any way possible.

I would like to thanks the Big Data Analytic' members for all good memories and discussions.

Last but not least, many thanks to the members of my examining committee.

Chapter 1

Introduction

The Machine Learning field is commonly linked with “big data”. Scientists use massive datasets in order to make better predictions in different realms such as credit card fraud detection, movie recommendations and sentiment detection in online reviews. Even though machine learning has been employed successfully in medical and commercial purposes, its usage in social sciences has yet to be well investigated. Machine learning has been used in studies to detect review sentiments[14], to classify cyberbullying and racism[1], to detect instances in the text from different text resources such as Wikipedia, Facebook, and Twitter[89]. Using machine learning algorithms can reduce the labor input in the process and reach human-level accuracy without the interference of human emotions. It can work in real time and flag different content based on the described tasks.

In very recent years, studies started addressing the problem of hateful speech on Twitter. The term “hateful” is very broad and includes any type of abusive or threatening speech or expression against a person or groups of people because of their race, religion or sexual identity. In other words, it includes any type of racism and sexism. However, there is no prior study focusing on sexism and its reflection on online social platforms such as Twitter. In addition, there has been no dataset focusing on this problem in the machine learning and natural language processing domain.

Therefore, we have decided to build this thesis to point toward efficient solutions for the problem of online harassment. We focus on different forms of online harassment, identify different categories of it and detect it by using different types of machine learning, deep learning, and natural language processing techniques.

This Thesis is organized into six main chapters, each focusing on one aspect of the language. The first chapter is about the data collection, the description of categories, examples of each category and the algorithms used to classify the tweets.

We show that a character-level convolutional neural network has a high accuracy in providing us with good feature vectors and is used successfully with Naive Bayes. In the next part, we deploy a novel method: convolutional neural networks filters. Filters followed by a threshold are capable of retrieving the informative ngrams in each category. These ngrams are useful sources of information to understand the nature of the harassing language in each category. Furthermore, the selected ngrams produced by these filters for each category are used to improve the classification task. For example, for the first category, indirect harassment, the ngrams are clustered into two classes. One of the classes is about cooking and the other is about indirect words of harassment or benevolent sexist words. In the second category, sexual harassment, ngrams are clustered into two categories. One of them focuses on the body as a sex object and the other contains more provocative words. In the third category, physical harassment, ngrams are clustered into two groups. The first cluster focuses on the physical attributes of girls and the second one focuses on threatening the physical attributes of girls.

In the next chapter, we focus on understanding the mental state of users who have written harassing tweets using the recently released dataset, “SemEval2018: Affect in Tweets” [88]. In terms of emotion type, the plurality of tweets in “indirect harassment” are categorized as being joyful, surprising and showing anticipation. In the second category, most tweets are categorized as disgust, sadness, and anger, while in the third category, physical harassment, more feelings of anger, disgust and sadness are shown. In the sixth chapter, we propose novel methods for the text generation and text augmentation using knowledge graphs in order to improve the classification task. The two knowledge graphs, ConceptNet and Wikidata are used for the task of text generation and text augmentation. The proposed methods for text generation and augmentation increase the performance of classification considerably.

1.1 Motivation

Social media is an integrated part of today’s society. Looking through the positive aspects of it, online social platforms can be used by young people as an educational tool. People usually share their advice and life experiences with their broad audience. In addition, these social networks can be used as a tool for different groups of society to

share their points of view. People are usually able to form groups and seek support from different ones such as "Disability is not ability", which tries to make society aware of the way disabled people are treated. People can take advantage of different online content and learn skills more tangibly and professionally. For example, there are many different recipes on YouTube about cooking a specific sweet in different ways which are suitable for different groups of people.

Furthermore, social media platforms serve the society in different ways. Amongst them all, Twitter has been used as a tool for abuse and harassment toward users, specifically women. Experts are worried and criticized Twitter because it has become the platform for many trolls, racism, misogynists and hate groups which can express themselves openly [132, 143]. Additionally, this platform is poorly built to handle the problems of online harassment. The CEO of Twitter has declared that "We suck at dealing with abuse and trolls on the platform and we have sucked at it for years" [132].

Harassment has widely affected women throughout different times and places, from the working environment, schools, military installations and social gatherings to online social platforms. Based on Huffington Post [54], on average one out of three women, aging between eighteen to thirty-four, has been sexually harassed at work. Online harassment usually refers to different forms of abusive behavior, including flaming: name calling or insults; doxing: revealing home address or phone number and impersonation or public shaming: destroying the person's reputation in online social platforms [10]. In addition to the previous categories, there are other types of tweets directed at women. These tweets usually deter women from having specific political or social positions as well as jobs outside their houses, undermining the confidence and determination of women and also pointing at women as sex objects. When harassment happens in online platforms, the severity and complication of the experiences usually escalates for the victims and makes them respond hard [40, 74, 73]. Even though these problems have existed for many years, now only a few of the victims are gradually speaking out.

Previous studies have focused on collecting sexist and racist datasets by using very broad definitions or focusing on only two categories of sexism such as benevolent and hostile sexism, which undermine other types of online harassment [59]. However,

there is no prior study focusing on different types of online harassment utilizing machine learning and natural language processing techniques [58, 146]. Automatically detecting the content containing sexual harassment could be the basis to remove, or flag it for human evaluation. Differentiating different types of harassment provides a means of control as a viable tool for future research. Automatically flagging and removing online harassment content requires niche category identification and description techniques. There is not yet a good definition of online harassment. In other words, while identifying and labeling online harassment is an easy task for human beings, providing a definition is hard. In this situation, machine learning algorithms can infer from the labeled data a definition to classify the contents as harassment or not. Thus, this thesis proposes to develop models which apply machine learning and deep learning techniques to automatically classify tweets into different types of online harassment. As the basic goal, this automatic classification will significantly improve the process of detecting this type of speech on social media by reducing the time and effort required by human beings.

1.2 Problem

There is a lack of research in the gender-based analysis of harassment in scientific research attracting machine learning and natural language processing communities. The closest work in the natural language processing domain to our research is the work of Waseem et al, 2016 [146]. They collected tweets and categorized them into being racist or sexist based on eighteen definitions. Their definitions could not be used to help to distinguish between different types of online sexism. What is more, at the time of our research, most tweets were removed by Twitter. In another study, tweets were categorized into benevolent or hostile sexism [58]. However, what is happening in social platforms does not fit into these two types of sexism. There is a need to conceptualize online sexual harassment since it has different manifestations.

1.3 Contribution of Thesis

The proposed solutions are as follows:

- In this thesis, inspired by social science [70, 135], we proposed three categories

of online harassment. These categories are as follows “indirect harassment”, “sexual harassment” and “physical harassment”. For this purpose, we needed a clear instruction to label the tweets. We ran a pilot study of having twelve volunteer non-activists to label fifty tweets. After the process of labeling and getting the feedback, we proceeded to label the bigger dataset which made our experiment possible. After that, we ran different word representation learning, machine learning, and standard deep learning algorithms to classify the tweets.

- Interpreting neural networks help us to trust the predictions of the models and find useful ways to improve them. Based on the literature, one-dimensional convolutional filters followed by a max pooling and a threshold can detect a family of useful ngrams to make decisions [55]. We used convolutional filters to understand different types of harassment and help us to improve the classification results. After that, we took advantage of a newly released dataset “SemEval-2018 task1: Affect in tweets”, to show the types of emotion and their intensity in each online harassment category [88]. We trained, tested and evaluated different classification methods on the SemEval-2018 dataset and chose the classifier with the highest accuracy to test each category of the tweets to know the mental and effectual states of the users. It is a good venue to explore because all the tweets are not directly sexist and they carry different emotions. Based on our best knowledge this represents a new contribution to the field and we are the first to demonstrate the power of such in-depth sentiment analysis on the tweets.
- In today’s world, in order to take advantage of deep learning techniques and training machine learning algorithms, we need a big amount of data. However, when it comes to sensitive datasets, they are usually rare and imbalanced. One of the traditional methods to generate the dataset is to use a dictionary to replace the words. However, this method is very limited in terms of application and domain. Therefore, we used ConceptNet and Wikidata to improve classification by (1) text augmentation and (2) text generation. In our text generation approach, we generated new tweets by replacing words using data acquired from ConceptNet relations in order to increase the size of our training set. In our

text augmentation approach, the number of tweets remained the same but their words were augmented (added) to make the tweets longer by the words extracted from their ConceptNet relations and Wikidata descriptions. In our text augmentation approach, the number of tweets in each class remained the same. Experiments showed that our approach improves classification significantly in all of our machine learning models. Our approach can be readily applied to any other text classification problems using any machine learning model.

- We organized the first competition in ECML PKDD conference as “SIMAH (SoCiaL Media And Harassment) Categorizing Different Types of Online Harassment Language on Social Media”. This competition included ten thousand tweets. In pursuit of the study objective, two tasks were proposed. Task A, binary classification, investigated whether a tweet is harassment or not. For task B, we categorized the harassment tweets into three categories of “Indirect harassment”, “physical harassment” and “sexual harassment”. This dataset is available to the public and it is made to bring this problem to the community making it possible for more researchers to work on it and to compare results.

1.4 Outline

The rest of this thesis is organized as follows:

In chapter 2, we present the literature review of the studies around hateful speech, sexism and online sexual harassment, the description of the concepts and tools used in the following chapters and a presentation around the social perspective of sexual harassment.

In chapter 3, we present the categories and describe the process of coming up with new categories, the process of data collection, instructions, annotations, feedbacks, gender-based ideas about the content, process of labeling, challenges and finally present the results of classification algorithms.

In chapter 4, we focus on using convolutional neural network filters to understand different types of proposed categories and use the extracted information to improve the classification task.

In chapter 5, we focus on the newly released dataset in ”SemEval 2018: Affect in

the tweet” and try to understand the different types of emotions and intensity of the tweets in each category. We present the detail of the newly available dataset and the way we approach this problem.

In chapter 6, we present the new methods for text augmentation and text generation for the imbalanced dataset and present the results of improvement.

In chapter 7, we present a case study and in chapter 8 we have conclusions and future work. Figure 1.1 shows an overview of the work done in this thesis.

1.5 Published and Submitted Papers

- Sima Sharifirad, Alon Jacov and Stan Matwin, Understanding Different Types of Online Harassment Using Convolutional Filters, Widening NLP, ACL 2019.
- Sima Sharifirad and Stan Matwin, Using Attention-based Bidirectional LSTM to Identify Different Categories of Offensive Language Directed Toward Female Celebrities, Widening NLP, ACL 2019.

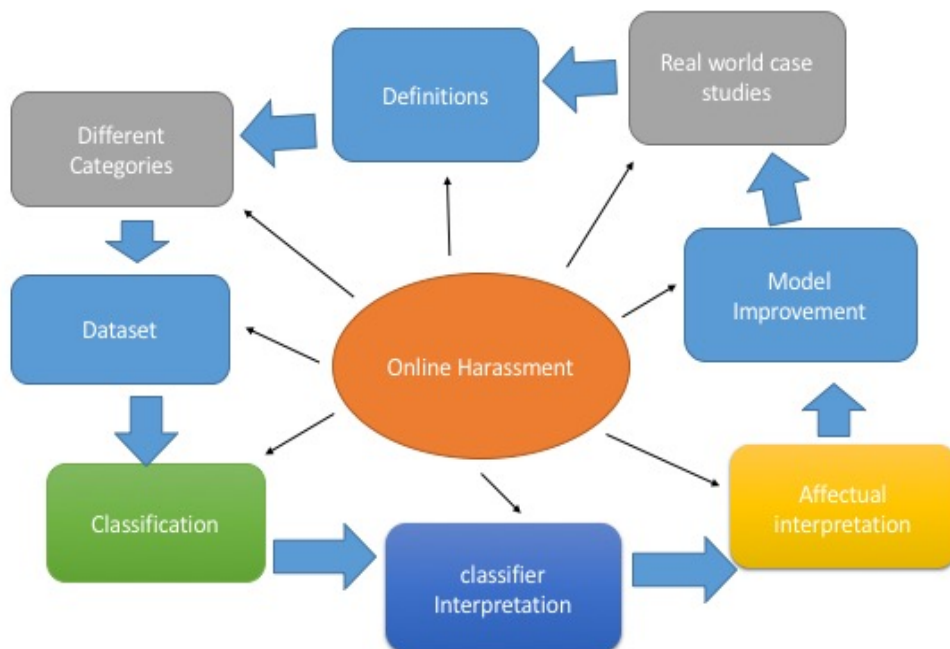


Figure 1.1: High level work in the thesis

- S. Sharifirad; B. Jafarapour and S. Matwin, Boosting Text Classification Performance on Sexist Tweets by Text Augmentation and Text Generation Using a Combination of Knowledge Graphs , ALW2, EMNLP 2018.
- Sima Sharifirad and Stan Matwin, How is Your Mood When Writing Sexist tweets? Detecting the Emotion Type and Intensity of Emotion Using Natural Language Processing Techniques, OceANS, KDD 2018.
- Sima Sharifirad and Stan Matwin, Automatic classification of sexist tweets, WiML, NeurIPS 2018.
- Sima Sharifirad and Stan Matwin, Classification of Different Types of Sexist Languages on Twitter and the Gender Footprint on Each of the Classes, CiCLing 2018.

Chapter 2

Background and Related Work

In this chapter, we review some of the background and similar studies in the natural language processing and machine learning domain. In the first section, 2.1, we have a quick glance at the works and facts related to online harassment and abuse on Twitter. Then, in the first subsection, we present different definitions, categories and related concepts of online harassment. After that, in the subsequent subsection, we present available datasets. In Section 2.2, we explore topic modelling algorithms and models. In Section 2.3, we continue the chapter by reviewing recent works in affect classification. In subsection 2.3.1, we mention available datasets related to sentiment analysis (SA), emotion detection (ED) and opinion mining (OM). In section 2.4, we present related works of two knowledge graphs, ConceptNet and Wikidata.

In this thesis, different computational methods such as word embedding and neural network are borrowed for text classification. To present enough background for the following chapters, we provide the details of these methods in section 2.5. Then, in section 2.6, we focus on text generation and augmentation methods. Finally, in section 2.7, we present the summary.

2.1 Online Harassment and Abuse on Twitter

The internet brings people together from different worlds. It makes different relationships possible and is an unavoidable part of our lives. As the internet becomes common for human experience, it becomes critical for women to feel safe online. Their lack of safety online acts as an impediment to their freedom and their human right. Based on Sulleyman [127], online harassment is the amplified version of the real world's ideas about gender discrimination and equality and as users switch between offline and online worlds, it is necessary to focus on both spaces [127].

Often, when we talk about violence or harassment against women, online harassment is ignored. However, online violence is a clear manifestation of real violence and

can tighten the boundaries of freedom, expression, safety, and privacy. Based on a study done by APC, Association for Progressive Communication, in seven countries, women and girls between 18 and 30 are the most targeted population and it might be mainly because of their access to the internet and social media. Often times, online harassment and violence are ignored by the victimized peers or families and results in the termination of the Internet or social media use[39].

Women get harassed three times more than men and there are some differences between online harassment towards the two genders [57]. Online harassment towards women is usually more violent and frequent and the purpose of it is to humiliate or threaten them. However, online harassment toward men is less frequent and it makes them feel ashamed or embarrassed. Often on social media, women are the main targets of different types of severe harassment such as receiving rape videos, extortion, doxing for the purpose of harm, stalking and pornography.

Online harassment is often not one dimensional. Women from different social groups get harassed because of their race, their ethnicity, sexuality, religion and even disability [156]. Women often get involved in a flood of hate and harassment and experience name calling such as “prostitute”, “whore”, “hooker” or “sex slaves” because harassers want them to be silent. Based on Trollbuster [57], forty percent of female journalists tend not to write about sensitive or controversial stories because of online harassment and thirty percent of them would like to stop their profession because of online abuse. Based on the research in 2016, eight out of ten journalists abused were women and the other two were black men [43]. There is some research focusing on the reasons for online harassment and trolling. Based on research by Bianca Fileborn, harassers normally feel entertained to write abusive contents and would like to impose their power over the victims[38], .

Different steps can be taken in macro and micro levels to deal with online harassment. In the micro level, detecting and analyzing harassment content can open new avenues for social media companies and critics to identify different social trends. In the macro level, dealing with online harassment requires a global movement from countries, governments and individuals to set rules and regulations to combat online harassment without restricting women’s freedom. In fact, social media should be safe for all age groups and genders.

2.1.1 Different Definitions and Categories of Online Harassment

There have been different studies focus on negative behaviours online. They categorize these behaviour into different forms such as hateful speech, online harassment and cyber-bullying. These negative forms of behaviour share some features but they characterize their own definitions.

Pew Research Center defines online harassment as a language which includes offensive name calling, embarrassing words, stalking, sexually harassing, physically threatening in a continuous way [32]. In another study, Yin et al. [162] defines online harassment as an intentional annoying communication on social platforms.

Dadvar et al.,[22] defines Cyberbullying as language which carries different pieces such as having a victim which cannot be supported and defended receiving an intentional negative language repeatedly over a period of time from a person or a groups of people online [36]. Patchin et al,[100] define cyberbullying as receiving unwanted language such as name calling, sexual related language, ignoring a person out of community or threatening them.

Hateful speech has defined as a speech that shows or express any types of hatred against a person or group of people [82]. Saleem Mohammad et al.,[114] defines hateful language as a speech towards another person or group of people based on their group identity.

The Women’s Media Centre (WMC) [156] identifies different types of harassment and presents a list of them; we will limit all the categories to the ones we used in our research. The first in the list is ”flaming” which is a big wave of violent, threatening messages towards a person. The second in the category is gender-based names and name calling such as “bitch”, “slut”, “whore” and “cunt” or any other specific words used to refer to women.

Glick and Fiske [96] address two different types of sexism including benevolent sexism (BS) and hostile sexism (HS) and argue that these types of sexism compliment each other in human history. They argue that hostile sexism usually stems from the feelings of man’s dominance and tries to reinforce the power in different ways. Glick and Fiske [96] believe that gender distinction, power, control, and sexuality are often the major components of HS and usually become harsh when the traditional beliefs about women or the power of men are underestimated by women. These

contents are clearly very harsh, negative and violent. After that, in 1997, Glick and Fiske [96] present three features: “dominative paternalism”, “derogatory beliefs” and “heterosexual hostility” as the main features of HS. The first feature, dominative paternalism, expresses the concept that women should be controlled by men. The second feature, derogatory beliefs, considers women as a sexual object for men to be exploited. Heterosexual hostility usually assumes women as sexual objects and believes women can use their sexual attraction to gain control over men.

On the other hand, benevolent sexism (BS) is about a general concept postulating that women are weak creatures and need support from men. It does not contain any violence or harsh words but indirectly imposes traditional beliefs about women and their abilities in a positive way, for example, “you play as good as a boy”.

Jessica Megarray [80] focused on the problem of online sexual harassment. She conducted a case study by collecting tweets using hashtag *#mencallmesomething*, a hashtag in which women used to write about their experience of abuse, violence, and harassment. These tweets clearly show the opinions around women in a patriarchal society. One of its manifestations is calling women certain names such as “Evil” or “hysteric” throughout history [125, 104]. The second manifestation is violence or aggression focusing on physical attributes of women or comments about wishing them to get diseases or cancers with some comments “a sexually transmitted disease or vagina cancer” [93]. The third manifestation is commenting on the lack of attractiveness of women, such as: “I would have to rape her with 3 Popsicle sticks taped to my flaccid wang” [72]. The fourth manifestation is all types of death threats such as “tape a plastic bag on [her] head [and] kill [herself] live on a webcam”. The last manifestation is cyberstalking in which the personal information of women is published publicly such as the street address.

In 2009, Vitis and Gilmour proposed a definition for online sexual harassment; they considered content to be sexually harassing if it has three features: the harassment is directed towards women and it is threatening [135]. These contents can contain negative comments [40], unwanted distribution of pictures in pornography sites and can have the form of cyberstalking [8] and violence and gender-based threats.

Frances Shaw [121] focused on harassment and trolling on Instagram. She defined harassment as a behavior which is threatening and is directed to a female, and defines

trolling as publishing inflammatory or vague posts in order to provoke individuals or groups of people to respond aggressively. Ruth Lewis [70] reported that sixty percent of regular users get harassed while eighty-eight percent of them are female users on Twitter. These users usually publish feminist posts on the internet. She divided online harassment into ten major types of abuse and argues that women can experience some of these types together. These types are as follows: harassment, sexual harassment, physical threats and violence, sexual threats and violence, stalking, flaming, trolling, electronic sabotage, impersonation, and defamation. She presented a brief meaning for each.

Harassment is a frequent part of communication without user consent or is the invasion of users privacy. Indeed, sexual harassment is unwanted communication which contains sexual content such as sexual images, extracting personal information and using them to threaten, harass or intimidate users. In her study, she presented real tweets received by women. Women shared their stories on Twitter using the hashtag “#everydaysexism”.

Sara Mills[84], in her book, language, and sexism, divided sexism into two categories of *overt sexism* and “indirect sexism”. She mentioned overt or direct sexism as a way which directly targets women and results in discrimination through traditional opinions. On the other hand, “indirect sexism”, known as “benevolent sexism”, is mixed with sarcasm, irony, and humor. This type of sexism does not have any slur words, known as leading words, so their identification is difficult.

Zeerak Waseem et al. suggested eleven definitions and collected around 136,052 tweets [146]. These definitions are as follows: using a sexist or racial slur, attacking a minority, criticizing a minority, indirect use of hate speech, misrepresenting the truth and distorting the view of the minority group, using specific hashtags such as #immigrant, #nigger, #sjw, #WomenAgainstFeminism and so forth. Having used these definitions, they then grouped the tweets into three categories: sexist, racist or neither. In the following research, Akshita and Radhika [58] examined the tweets collected around the hashtag #mkr, my kitchen rule, an Australian game show, to classify tweets into hostile and benevolent sexism. They used Waseem dataset as their hostile class[146]. For their benevolent class, they used the Twitter search API to collect tweets for the benevolent class. They collected about 95,292 tweets

and manually label 7205 tweets as benevolent tweets. After that, they used different classifiers to classify tweets into three categories of hostile sexism, benevolent sexism, and neither[96].

Alessandro et al. [75] tried to automatically detect hate speech including sexist words along with other types of taboo words by focusing on just words and combinations of nouns, verbs, adjectives, and adverbs. Badjatiya et al. [5] took advantage of deep neural network models and used the Waseem dataset [146]. They used different types of convolutional and recurrent neural networks along with different representation learning. Long-Short-Term-Memory (LSTM) was combined with random embedding and among them, gradient boosted decision trees had the best performance. Park and Fung [99] approached the Waseem dataset and considered the task of hateful speech detection as a two-step classification task: First, if the tweet is abusive or not and second if it is sexist or racist. They proposed a hybrid Convolutional Neural Network (CNN) by combining character level and word level CNN. In another study, Clarke and Grieve [18] also used the Waseem dataset. In their study, they proposed a Multi-Dimensional Analysis (MDA), examined and categorized tweets as being interactive, antagonistic and attitudinal, comparing the classes of sexist and racist tweets.

2.1.2 Available Datasets

In this section, available datasets are introduced with their sources, their names, the types of repositories, years, number of classes and language.

2.1.3 Crowdsourcing Platforms

One of the best ways to categorize tweets into different categories is by using online crowdsourcing platforms. Online crowdsourcing platforms are considered an effective tool to annotate the data. Amazon Mechanical Turk (MTurk) is among the first which is perceived as a reliable and cost-effective tool to collect high-quality data for different research purposes. However, in MTurk, participants can take part in the process of annotation more than once and it reduces the credibility of the answers. In this way, there is a possibility that each sentence is annotated by one person more than once. After MTruk, other crowdsourcing platforms such as CrowdFlower(CF)

Table 2.1: Available Dataset for Hate Speech Detection

Source	Name	Distribution	Year	Number of Instances	Classes used	Language	link
Cornell University	Hate speech detection	GitHub repository	2017	24802	hate speech, offensive language, neither	English	[23]
University of Copenhagen	Hate speech	GitHub repository	2016	16914	sexist, racist	English	[146]
StackOverflow	offensive comments	Available with NDA form	-	-	-	English	[144]
Yahoo News Dataset of User Comments	Abusive language	Available by sending email	2016	951,736	abusive and clean	English	[145]
User-Centred Social Media	German Hate-speech Refugees	available under liscnece	2016	470	hate speech and not hate speech	German	[112]
Hate base	Hate-base	available	2017	-	-	universal	[49]
Greek News Data Set	not available	2015	-	-	-	-	[102]

and Prolic Academia (ProA) are used for academic purposes. A useful and detailed discussion about different purposes of different online platforms is presented in the article by Peer et al. [103].

2.2 Topic Modeling

There are different topic modeling algorithms, the main ones are LDA (Latent Dirichlet Allocation), LSA (Latent Semantic Analysis) or LSI (Latent Semantic Indexing)

and NMF (Nonnegative Matrix Factorization). They are used for text classification, recommender systems, news sources and blogs [139, 90, 91]. In addition, they are used successfully for bias detection in the news and blogs [2] and depression detection on Tweets [111]. All these methods have some similarities such as the fact that these algorithms cannot infer the number of topics from a corpus, the number of topics is usually considered as the input parameter for these algorithms, and the second input of these algorithms is the matrix of word-document. Below we will describe them briefly.

Latent Dirichlet allocation (LDA) is a generative and graphical probabilistic model focusing on a corpus, a collection of documents. The name Dirichlet stems from generating topics from a Dirichlet prior. In the past, it was believed that documents are associated with one topic. However, the main concept behind topic modeling is that documents are a mixture of latent topics and each topic is characterized by a distribution over words. In other words, the main goal of LDA is to calculate the documents generated by the topics. It is an iterative model and it comprises of two main steps. In the first step, each word is considered to come from a random topic. In the second step, it goes through each word and iteratively reassigns the word to a topic with two metrics, the probability of the word coming from a topic and the probability of the generated document by a topic. After conducting topic modeling, there are usually some specific points to consider: larger topics are often the most common ones in the corpus, topics which are closer are usually similar to each other in comparison to the topics which are far from each other, each topic is comprised of different words and topics are the most representative words that can be identified. In addition, this model, LDA, also has been used for clustering the documents based on their topic.

One of the fundamental works in topic modeling is Latent Semantic Analysis (LSA) [26]. It maps documents and terms to a latent semantic space. It is initially explored for producing a low-dimensional representation for a document in the domain of information retrieval. The main idea behind this model is to have a matrix of input, term-document matrix, and decompose it into two different matrices of document topic and topic-term matrix. This method usually takes the matrix as an input

and implements a dimensionality reduction method such as Singular Value Decomposition (SVD), and the main reason is that the similarities between documents can be estimated better in the reduced latent space representation in comparison to the original space. In this way, if documents do not have any shared words but have some co-occurred terms, they will have a similar representation. This feature is useful for noise reduction and detection of synonyms. However, this method has its own pros and cons. It is fast and easy to implement and gives robust results, but because of its linear nature, it might not perform well on all the datasets. In addition, It considers dimensionality reduction such as SVD as a part of the model and is usually computationally expensive.

Non-negative matrix factorization (NMF) is another common method for topic modeling. It has been widely used in image processing to extract facial features; in text mining for topic modeling and document classification; in hyper-spectral imaging for identifying endmembers; and in pixel classification, air emission control, single-channel source separation, and community detection [97, 68, 140]. NMF is a matrix factorization method where the matrices should be non-negative. Conceptually, assume we have matrix X and it is divided into two matrices of W and H so that we can approximate $X \approx WH$. Even though there is no guarantee to recover the original matrix, we will try to approximate it as close as possible. It is an NP-hard problem, so we target to find the best local minima possible.

In the above formula, W is the term-topic matrix and H is the topic document matrix. When W and H are updated iteratively they have a high level of convergence. In each iteration, W and H are multiplied by a variable depending on the final approximate value. NMF has a general positive feature. This positive feature is the correlation between latent topics and can produce lower dimension factors [4].

2.3 Classifiers

One-vs-rest (OVR): we trained and evaluated K independent binary classifiers for each class separately for our classification tasks. We considered all the samples in that class positive and the rest negative using LinearSVC with linear Kernel in the Sklearn Python package.

Support Vector Machine (SVM) is used as a supervised model for the classification[3]. The main idea of the algorithm is to maximize the minimum distance from the hyperplane, which separates the samples to the nearest sample. To classify the tweets, we used different word vectors and we considered labels as one hot encoding.

Naive Bayes (NB) the main idea behind naive Bayes is to maximize the posterior probability (MAP). Based on the number of classes known as prior probabilities, a class label is assigned to the new sample with its features, in order to maximize the posterior probability given the new sample. In fact, it computes the class conditional probabilities of the features having the available classes.

K-Nearest Neighbour (KNN): This algorithm is considered as a non-parametric classification algorithm. This algorithm usually calculates the Euclidean distance between the new sample and every other training example. The k smallest distance along with the most represented class are considered as the class label.

Multilayer perceptron (MLP): This algorithm is in the category of supervised algorithms. It is made in a way that can be changed easily for the classification task. In the case of classification, the word vectors (input) are multiplied by different weight vectors to calculate the activation function for each specific data point. The weight vector, which produces the highest activation, will be the class data that the sample belongs to. It requires multiple training iterations to learn the data.

2.4 Affect Classification

Human beings use language and text to communicate. These communications carry feelings and emotions and are expressed with different intensity. With the emerging of online social platforms, it has become an opportunity and challenge for users to understand the feelings, opinions, and intents of users who write a short piece of text. With the emergence of studies and activities in the opinion mining and sentiment analysis domain, new tools and systems are developed.

Over the past years, there have been many definitions for feelings, emotions, sentiments, and opinions and they all focus on the subjectivity of the human beings as it is a characteristic of the human mind. Shouse [122] proposes a meaning for *feeling*: “sensation that has been checked against previous experiences and labeled”. It is considered as a person-specific feature and is different from one person to another. In

the same vein, there have been many definitions of *emotions*. Shouse [122] argues that human beings genuinely show their emotions and these emotions can be expressed in public society. In addition, the sentiment is defined to be a reaction to an emotion which is relatively permanent and is provoked in correspondence to a value or an object [14]. Eric Shous presents a nice description for feeling “An emotion is the projection/display of a feeling. Unlike feelings, the display of emotion can be either genuine or feigned. The distinction between feelings and emotions was highlighted by an experiment conducted by Paul Ekman who videotaped American and Japanese subjects as they watched films depicting facial surgery. When they watched alone, both groups displayed similar expressions. When they watched in groups, the expressions were different. We broadcast emotion to the world; sometimes that broadcast is an expression of our internal state and other times it is contrived in order to fulfill social expectations. Infants display emotions although they do not have the biography nor language skills to experience feelings. The emotions of the infant are direct expressions of affect” [122, 34].

Based on Larousse Dictionary, emotion is a “sudden trouble, transient agitation caused by an acute experience of fear, surprise, joy, etc.” or “mental feeling or affection (e.g. pain, desire, hope, etc.) as distinct from cognitions or volition” based on Oxford English Dictionary[92]. Cannon defines emotion as a mental state and a mere somatic response [13]. Eric Shous also describes emotion as “the projection/display of a feeling. Unlike feelings, the display of emotion can be either genuine or feigned. The distinction between feelings and emotions was highlighted by an experiment conducted by Paul Ekman who videotaped American and Japanese subjects as they watched films depicting facial surgery. When they watched alone, both groups displayed similar expressions. When they watched in groups, the expressions were different. We broadcast emotion to the world; sometimes that broadcast is an expression of our internal state and other times it is contrived in order to fulfill social expectations. Infants display emotions although they do not have the biography nor language skills to experience feelings. The emotions of the infant are direct expressions of affect.” [122, 34]

In the same vein, affect is used interchangeably with emotion. However, it seems it carries “tactile, sensuous, and perhaps also involuntary connotations” [79]. Eric

Shous describes affect as “a non-conscious experience of intensity; it is a moment of unformed and unstructured potential. Of the three central terms in this essay À feeling, emotion, and affect - affect is the most abstract because affect cannot be fully realised in language, and because affect is always prior to and/or outside of consciousness. Affect is the body’s way of preparing itself for action in a given circumstance by adding a quantitative dimension of intensity to the quality of an experience.” [122, 77]

A sample of sentiments can be hate, love and other types of sentiments on online platforms. Definitions about the term *opinion* are around the way humans think or perceive an idea, a concept or an object and it includes topics, opinion holders, claims and sentiments. In summary, feelings are a conscious person’s specific features, emotions are more social phenomena which are influenced by culture and effect which seems to come first in the list. Also, sentiments are affected by time and society while opinion is a person-centered adaptation of facts and truth that may not have the element of emotion.

Sentiment Analysis (SA) and Opinion Mining (OM) seem to be one definition or to be complementary to one another. However, scientists argue that opinion mining is the art of understanding an opinion around an entity in a sentence while sentiment analysis seems to be around finding an opinion around an entity and trying to understand the feeling of the user in that opinion [94]. SA can be applied in a different level of extension, in the sentence level, document level, aspect level, and concept level. A task is categorized into three categories of positive, neutral and negative and machine learning algorithms are used to classify them. In the literature, generally, there are three types of a broad range of methods for sentiment analysis such as machine learning methods, lexicon based methods, and hybrid methods [78].

In the machine learning approach, machine learning models and algorithms are used for sentiment classification. The lexicon approach is a dictionary-based approach in which seed words are recognized in the sentences. In a corpus-based approach, a list of opinion words is used to find other opinion words. Hybrid methods are the combinations of the other two methods and are very common in the SA domain. Another similar interesting domain of work in sentiment analysis is Emotion Detection(ED). Calvo and D’Mello [12] consider some primary categories of happiness, sadness, fear,

surprise, anger, and disgust. After that, Plutchik [106] proposes eight primary emotions such as joy, sadness, anger, fear, trust, disgust, surprise and anticipation.

2.4.1 Available Datasets

The best source for sentiment analysis can be found in product reviews. These reviews play a role in corporate decision making and product priority settings. Users are generally interested in knowing the best and worst aspects of a product, service or an idea. One of the openly available datasets is the HoteExpedia dataset, originally containing 381,941 reviews from 6030 hotels [94].

Another useful dataset for emotion detection task is "SemEval-2018 Task1: Affect in Tweets" proposed by Saif Mohammad [88]. He published a dataset about the intensity of emotion (E), the intensity of sentiment (V, valence) and the emotion types. This dataset is available in three languages namely English, Arabic and Spanish [87]. He proposed different subtasks to be solved by using regression or classification.

Another useful dataset for the task of emotional intensity is the dataset proposed by Mohammad and Bravo-Marquez [88]. They released a dataset of 7100 tweets published in the 2017 WASSA Shared Task on Emotion Intensity [86]. For example, the first task, detecting the intensity of emotion is to specify for a tweet a score of zero or one in which one shows the highest degree of emotion and zero shows the lowest amount of emotion in four categories of anger, fear, joy or sadness in a regression problem. The other task is about categorizing tweets into four categories of 0 to 3, which 0 shows no emotions and 3 represents high emotions. The second subtask, the sentiment intensity or valence, is to determine if a tweet is in two categories of one or zero. Zero represents a positive state of mind and one represents a negative mental state. The last subtask is about just classifying the tweets into seven categories labeled from 3 to -3 , which 3 represents a very positive and -3 shows a very negative state of mind. The last subtask is classifying the tweets into eleven emotion types including anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise and trust.

2.5 ConceptNet and WikiData

There are various definitions for knowledge graphs. From its name, one can understand that a knowledge graph provides knowledge by using a graph. Paillhiem [101] describes a knowledge graph as a large network or graph which is extracted from a web page comprising of facts, real-world entities, documents, and datasets, as well as their properties and relations including many domains. These entities are interrelated and can be extracted from different paths in a graph, that is the reason why they are called knowledge graphs [65].

In simple words, it is a graph of interrelated descriptions of entities, events or abstract concepts which have two properties. First, they have a formal structure making it efficient and easy to understand for both people and human beings. Second, it can play the role of a database because knowledge can be extracted from these graphs via queries. It is a graph which carries different features of a graph and can be analyzed in the same way. In fact, it acts as a knowledge base because nodes in this graph carry the formal semantic information and descriptions and they can be used for interpreting the data and reasoning for new facts. In other words, knowledge graphs help us broaden our understanding of the way knowledge can be presented on the web [126].

Based on Zhang [163], there are three features and prerequisites for a graph to be considered as a knowledge graph. First, it should have a structure, second, the graph entities should not be ambiguous and third, this graph should have limited relations. There have been over 37 knowledge graphs. Among them, there have been some knowledge graphs for academic purposes. For example, DBpedia is a type of knowledge graph which is considered as a transformation of Wikipedia, it is constructed from the infobox which existed in the Wikipedia pages and as a result contains a lot of relations. A ConceptNet is a knowledge graph from the information which existed between people and is expressed in various forms. The Wikidata is another type of knowledge graph proposed by the Wikimedia Foundation in order to provide structured data for Wikipedia. It has a nice feature which saves users which making changes in the entities and facts in the knowledge graph [137].

In this thesis, two knowledge graphs, ConceptNet and Wikidata, are used for the task of text generation and text augmentation. In the following section, we will

explore these two knowledge graphs: The *ConceptNet* is a big knowledge graph which is made collaboratively from a web of knowledge. It includes common-sense about different sorts of facts and entities such as spatial, physical, temporal and physiological aspects of ordinary life [137]. ConceptNet is used in different domains and has been used in machine translation [25] and speech recognition [124].

In this knowledge graph, nodes are the concepts, either in the form of words or phrases, and edges are the relations. There are more than thirty relations in the knowledge graphs. The relations and concepts at a high level are usually presented in the form of a triplet, two concepts and the relation between the two. An example of relations in the ConceptNet is “A kind of”, “Similar to” and “Is used for”. As an example, the word *woman* has different types of relationships in ConceptNet such as “Related terms”, “Synonyms”, “Derived terms”, “Root Words”, “Antonyms” and so forth. They are presented in the form of a $\langle woman, Antonym, man \rangle$ or $\langle woman, synonyms, ladies \rangle$.

One of the challenges of using ConceptNet is that each word can carry different senses, depending on the context; therefore, the right sense should be used. The second challenge is the inconsistency in some of the concepts in the knowledge graph, looking into the relations and concepts. For example, replacing a word by the relation type *Synonyms*, might change the meaning of the sentences and add ambiguity to it.

Wikidata, on the other hand, is a part of the Wikimedia projects and is fundamental for other resources such as Wikipedia, the Wiktionary and the Wikibooks. It comprises more than 46 million data items and works as an ideal database for Wikipedia in more than 288 languages. It is a good source of structured data and contains text, image, coordinates, dates and so forth. In Wikidata, each entity or word has five properties. In other words, each item can have a label and description. For example, the word, *woman*, has different translated *labels* in different languages and has the *description* of “female adult human”, and is known as “female human”. In addition, it has different *statements*: subclass of, part of, has quality and so forth.

Wikidata serves different features, it can be used for making infoboxes in Wikipedia in different languages, it is used to list generation and link suggestions. Wikidata is an open-domain knowledge base. Also, both humans and machines are able to edit

it collaboratively. In addition, it keeps facts that are inconsistent, ambiguous or contradictory to present a different aspect of existed knowledge about that topic [138]. In 2018, Wikidata has had 50 million items, and it is developing by users and its references keep changing and checking periodically. On the flip side, Wikidata does not have a mathematical formula based on Scharpf et al.,[115].

2.6 Neural Networks for Text Classification

Artificial Neural Networks (ANN) are mainly inspired by human nervous systems and are computational processing systems. Their main component is interconnected nodes which are called neurons. These highly connected neurons learn an input in order to produce an output in an optimized manner. Inputs are usually multidimensional vectors used by the input layer and then get distributed in the hidden layers. Hidden layers will decide based on the result of previous layers and make a stochastic change in their own layers in order to improve the output.

Text classification is one of the subtasks in natural language processing (NLP), in which we try to allocate predetermined categories to free-text documents. During the past decade, Neural Networks (NN) have attracted a lot of attention due to their high capability in different domains and applications including web search, information retrieval, ranking and document classification [26, 98], sentiment analysis [123], machine translation [129], text generation [128], part-of-speech tagging [30], named entity recognition [19], question answering [50], semantic role labeling [166] and automatic parsing [69]. Accessing a large reservoir of images, texts, speech and videos along with advances in parallel computing architecture provide a good background for neural networks. Deep Neural Network (DNN) is the stacked neural networks on top of each other in different layers. Each layer usually handles a nonlinear process of information and has a set of parameters along with the inputs and outputs. DNN usually requires a big corpus to train due to their large number of parameters, and it is normally difficult to generalize well when the number of data points is limited.

One of the most used deep learning models for text classification is the big families of Convolutional and Recurrent Neural Networks. Convolutional Neural Network (CNN) is a deep artificial neural network which is primarily used to classify images for a wide range of applications from image segmentation to face recognition. The

main feature of CNN, which makes them different from other categories, is the pooling layer.

Recurrent Neural Networks are a family of neural networks perceived to be powerful for processing different texts with different lengths. As the location of each word plays an important role in the overall meanings and semantics of the sentences, these algorithms fit well since they can remember the information which comes before each word. In the following sections, we will elaborate on these two types of Neural Networks.

2.6.1 Recurrent Neural Networks

Recurrent Neural Networks (RNN) are first proposed by Elman [35] and are able to process a sequence of words with different lengths by considering a transition function to the internal hidden states of a vector iteratively. These networks are proven to have a time complexity of $O(N)$. These models dissect a piece of text word by word and keep the semantics of all the previous texts in a hidden layer with specific lengths. This allows the model to store the semantic of long documents. These models memorize the processed information in the hidden vectors and share the hidden states as a factor of time. The problems with RNN are: *exploding* or *vanishing gradients*. Over long sequences, the gradient vector either grows or decays exponentially. Long short-term memory network (LSTM) was first proposed by Jurgen Schmidhuber [52]. This algorithm specifically solves the problem of memorizing long-term dependencies. It has a separate memory cell inside in order to take care of necessary updates.

2.6.2 Convolutional Neural Networks

One of the famous Artificial Neural Networks which has been explored and used widely, specifically for image processing and pattern recognition tasks, is Convolutional Neural Network (CNN). It is successfully utilized for natural language processing tasks [20, 63] such as semantic parsing [161], sentence modeling [62], text generation (GAN), hateful speech classification [146], online harassment detection [120] and effectual detection of harassment tweets [117]. Different facts about applying convolutional neural networks on the text were presented by Kim [64]. He shows convolutional neural be very with a single layer. These networks are able to

self-optimize to reach the desired output. From the input, an image, to the output, each neuron will receive its own input to perform an operation. The last layer usually specifies the type of loss function which is selected to work with a specific number and type of classes.

Usually, during the first learning process, the performance might not be desired, in this case, different tips and tricks should usually be considered to reach a satisfactory result. There are two noticeable differences between the CNN and the ANN. First, the CNN is mainly used for specific applications in image processing and pattern recognition domains and second, these types of networks seem to handle a high volume of colored images properly. For example, in the MNIST benchmark, black and white datasets, images have small dimensions and processing them is easy for most ANN models. However, when images become colorful, they add another dimension, length, width, and height, and the number of weights increases in the first layer dramatically. In terms of architecture, CNN usually comprises of three types of layers: a convolutional layer, a pooling layer, and a fully-connected layer. In the simplest forms, these layers are stacked on top of each other to make a simple CNN model; in a more complicated model, different layers are added and connected differently.

The input layer is responsible for keeping the pixel values of an image. The second layer, the convolutional layer, identifies the resulted output of the neurons which shows the local image regions by calculating the product between the weights and the specific region it is connected to [95]. The pooling layer acts as a downsampling technique which tries to reduce the volume of parameters by using activations. In the last step, in the fully connected layer, the scores for the classes are produced by using activations and are used for the purpose of classification.

The input of a CNN is usually a three-dimensional array of pixel values in the form of $32 \times 32 \times 3$. Then, CNN uses different small filters, for example, in the dimension of $5 \times 5 \times 3$ and covers the pixels from left to right then top to bottom. As the filters move on, the pixels, the values in the filters, are multiplied by the value of the pixel and then all are summed up. This process is repeated and for every location in which filters slide, a new number generated. There are also other parameters such as depth, stride, and zero-padding which can play an important role in optimizing the parameters and decreasing the complexity of models. In this context, depth is

the number of neurons in the output layer specifically to the region of input. Stride mainly presents how the filter convolves around the input. For instance, if the stride is one, it mainly means there will be one shift to the right or to the bottom at a time. The bigger the stride is, the smaller the dimension of the output usually is. Last but not least is zero-padding, which is an effective method to control the dimensionality of the input by placing zero around the frame of the picture.

2.6.3 Word Vector Representation

Traditionally, NLP approaches used one-hot encoding or bag-of-words models to represent a vector for each word. One-hot encoding is a binary vector with a lot of zeros and ones which shows the vector of the word. This method is simple to implement and it does not have an order, and as a result, the context is ignored and due to the inherent binary characteristic of the vectors, the frequency of the words is not considered. In addition, this model does not capture the syntactic (structure) and semantic (meaning) across the words. On the other hand, Bag of the Word (BOW) is a way to extract useful information from a text and works by considering the occurrence of the word within a context. It is usually comprised of two steps, choosing vocabulary or words, and calculating their presence in the text. In this method, the order of the words is not considered. This method can have a problem of high complexity since it depends on the number of words in the vocabulary and the score which is considered. This model seems to have its own drawbacks. For instance, designing and selecting the right vocabulary is a hard task because it impacts the sparsity of the vectors. Secondly, handling a small amount of information with high sparsity is not computationally optimal. Third, this method ignores the order of the words and as a result the contextual closeness in the occurrence of the words.

In contrast, word embedding considers each word vector as continuous floating point numbers whose values in the word vector simply shows the word's distributed weight across the dimensions. In fact, they are low-dimensional vectors used to show the words in the geometrical and meaningful space [9]. In word embedding, words which have similar meanings have semantically similar word vectors. Another advantage of word embedding is their capability to be learned in both supervised and

unsupervised space from a large corpus [83][108]. One of the word vector representations is Word2Vec. It is a two-layer neural network which takes the text as an input and produces a vector as a feature vector which machines can understand. Word2vec is very efficient in order to keep similar word vectors together in the vector space. An interesting point about Word2vec is the fact that it is not predicted but learned. In other words, providing the context, the vector of the word is learned to maximize the quality of representation in order to minimize the loss. It is an iterative model and the purpose is to gradually change the representation of the words. Therefore, the Word2vec model tries to predict the probability of a random word in the corpus through having a closed position to the input words.

2.7 Text Generation and Augmentation

Natural language generation is an art. There are many metrics which should be taken into account in order to have a good text generation process. First, a generator should construct the best possible option based on the given situation, knowledge resources, content and textual shape. A good language generation system should decide about what type of information to present, when to mention that information and which structures are ideal choices to generate. It is important to consider an overall framework or skeleton in which each sentence is produced [109].

There are different methods to generate text. In the simplest scenario, a word can be selected based on the context and is added to the end of the text iteratively. The main piece is the way each word is selected for a generation. One of the methods to choose the word is sampling. In this method, the most probable word is selected from the conditional word probability distribution. The second method is the greedy method; as the name presents, the words with the highest probability are picked each time. The third method, beam search, takes into account different probabilities for choosing the right word in each step [159]. In addition to the previous methods, Recurrent Neural Networks (RNNs) have been used for different applications such as generative sequence models, sequence labeling, machine translation, and text classification [47, 60, 85]. Often, these models generate text by sampling from a distribution conditioned on the previous text along with a hidden state which keeps the word representation of the generated text. The main disadvantage of these models is that

their hidden states are unpredictable because these models are usually forced to condition on a text which is not in the training dataset. GANs (Generative Adversarial Networks) [45] are proposed as a solution for the previous problem. This framework consists of a generator and a discriminator. The generator tries to produce images and fools discriminator which had been trained on the real dataset and generated images in an adversarial setup. This method is originally proposed for the images due to the discrete nature of the text. This method has been successfully applied to the text by using Reinforcement Learning [46]. However, these models need a high volume of data for text generation and when the number of data points is not considerable and they are imbalanced, they do not have the desired performance.

2.8 Summary

In this chapter, we went through different related works in our research. We started with an overview of online harassment and sexism and different definitions of harassment online. Then, we continued by exploring sentiment and affect classification, ConceptNet, and Wikidata. As we described topic modeling, we went through exploring different neural network families, Convolutional Neural Network and Recurrent Neural Network and text generation and text augmentation methods.

In the next chapters, we describe the definitions which are used to categorize the dataset, the creation of the new dataset, the process of annotating it, the feedbacks from annotators and the initial classification task.

Chapter 3

Dataset Construction and Classification

3.1 Introduction

In cross-disciplinary research, an important element is having high-quality datasets which are useful and valuable and not only make studies possible but also are a great contribution to the field. There is a gap in the social science and machine learning disciplines. Lack of datasets which can connect these two areas has made research impossible. Often proposing a dataset can bridge the gap. As a result, we decided to make a dataset addressing online harassment. Based on the following publications, the contributions of this chapter are as follows:

- **Created a new online harassment dataset:** We created the first and only dataset of tweets labeled for different online harassment types. More than 3000 tweets are annotated for whether the tweet is "physical harassment", "sexual harassment", "indirect harassment" or "neither of the previous categories".
- **Examine Different Classifiers and Propose the Best Classifier:** We framed the problem of online harassment as a classification task, trying to categorize the tweets into four categories of online harassment. We deployed different classification algorithms on the dataset and identified the most suitable one for the classification task.

The construction of the dataset is described in the following article. In addition, this chapter is largely based on:

- S. Sharifirad and S. Matwin, Classification of Different Types of Sexist Languages on Twitter and the Gender Footprint on Each of the Classes, CiCLing 2018.

3.2 A Dataset for Online Harassment

Inspired by Ruth Lewis [70] and Sara Mills [84] studies, initially we came up with eight categories. For each category, we came up with the category name, definitions, points, hints and examples. Besides the positive examples, we decided to provide some false examples because some categories were close to each other. Below we present the initial categories with their information.

- Benevolent sexism (#1)

Definition:

Stereotypical beliefs about women or enforcing restricted roles that are framed in such a way that they can appear to represent positive beliefs.

Points:

These tweets are indirectly sexist and they do not contain any swear words.

Examples:

-A wise woman builds her house but a foolish woman tears it down with her hands.

- It's less of #adaywithoutwomen and more of a day without feminists, which, to be quite honest, sounds lovely.

- Lucky is the man who is the first love of a woman, but luckier is the woman who is the last love of a man.

- Physical threats (#2)

Definition:

Attacking or threatening a female biology or commenting on the female physical attributes in a violent and aggressive manner and wishing them to become diseased.

Examples:

-I was told I deserved to die a painful death.

-How about you get cancer.

-Just putting it out there, you deserve all those deaths you are getting.

- Sexual threats (#3)

Definition:

Harassment of women online as a kind of pornographic invitation which perpetuates hierarchical gender norms and incites others to display threatening sexual behaviour towards the victim.

Points:

Tweets related to forcing women to have sex.

Example:

-F..k that cunt, I would with my fist.

False example:

-Hope one of those bitches falls over and breaks her leg (this tweet is more about breaking and falling so its right category is number 2)

- Body harassment (#4)

Definition:

Harassment related to the body or beauty of the women in particularly reliant on positioning women as a sex object. Judging female physical attributes such as weight, breast size, hotness, their lack of hotness or humiliating their clothes or beauty.

Example:

-I cannot stop looking at Nikki's dreadful black crooked bra.

- Masculine harassment (#5)

Definition:

Tweets which show the control of men over women and are related to masculinity or showing the superiority of the men over women in a patriarchal society.

Points:

Tweets against working women, or lack of their physical or mental ability.

Example:

- The menus look like they were made by a 5-year-old little girl...in this case just the mental age of a 5 year old girl I guess

- Yes, we get it. You're pretty. Tone down the self promo and just cook.

- Lack of attractiveness harassment (#6)

Definition:

Tweets which show the feeling of being ignorant or not being sexy.

Example:

-Don't pull the gender card. You're not being harassed because you're a woman. It is because you are an ignorant cunt.

- Stop calling yourselves pretty and hot. You're not and saying it a million times doesn't make you either.

- Why do they keep saying that they are so pretty.

- Stalking (#7)

Definition:

Compiling information about female users and using it to harass, threaten and/or intimidate them. These behaviours include repeated pursuit, premeditation, repetition, and obsession with the victim.

Points:

-A women's street address.

Examples:

-I have all your names on ledger. 5000\$ a month or I torch you all over the web, forever.

- Impersonation (#8)

Definition:

Stealing the identity of the user.

Examples:

-If you are seriously gonna try to change gaming, I will hack your account and put gay porn everywhere.

3.3 Pilot Study

We ran our first pilot study giving participants instructions: we asked one male and 12 female non-activists¹ to label 50 tweets into eight categories. We collected their feedback on the clarity of the instruction, the clarity of tweets and also the level of task hardship. We calculated the inter-annotator measurement between raters using Fleiss' kappa score [116]. The score was 0.30 percent which was a poor level of agreement between the raters. After that, based on the feedback, we combined categories into four general categories and repeated the pilot study again. This time the calculated score was 0.70 which was a good score of agreement showing that the categories are well defined and instruction is clear enough to label more tweets.

Raters addressed some challenges and feedback as follows: First, there were some questions from the raters if they were supposed to label the tweets as sexist or if they should identify users as sexist. Second, labeling some tweets without their context was a confusing task. Raters wanted to know if the tweets were either a reply to the previous tweet or were the original tweet. Third, raters wanted to know the gender of the twitter users. Fourth, some tweets were not clear if they were directed toward a male or a female. In addition, there were some abbreviations in the tweets that made labeling difficult for them.

The male rater found some tweets were generally hateful towards women or feminists but did not specifically find them sexist. There was a difference between what was in their mind about the definition of online sexism and the definitions which were presented in the instruction. Lastly, they found hashtags and hints provided in the instruction useful as the context of the tweets.

We combined high overlap categories together and proposed three categories mentioned in Table 3.1.

¹In line with gender study policy for the raters, we exclude female activist mainly because of the bias they would bring to the labeling process.

3.4 Experiments

After coming up with the right instruction, we collected more than three thousand tweets². We used Figure Eight [33] to label the tweets. We also prepared 150 test questions, known as gold questions. These gold questions were, in fact, the tweets for which we already knew the labels and which were already labeled by an author and one female non-activist. These questions were used as a test for raters in the initial phase in order to check their eligibility to label other tweets. The distribution of dataset is in Table 3.2.

Table 3.2: Dataset Label Distribution.

Category	Number of tweets
<i>Indirect harassment</i>	260
<i>Sexual harassment</i>	417
<i>Physical harassment</i>	123
<i>Not sexist</i>	2440

3.4.1 Data Annotation

Using Figure Eight, customers do not have control over the number of raters and their demographic information. In our task, 349 raters contributed. There were 290 female and 58 male raters, thus 83 percent of the total raters were females³⁴. At the end of the study, a report is given to the customers besides the labels which show the clarity of instruction, ease of job and other metrics. Table 3.3 shows the details of this information. After running the results, we got a .csv file showing the country where the raters come from, the judgment counts, their gender and their submission rate.

The highest judgment count was 440 from Ukraine with a submission rate of 621. Raters who participated to label the tweets were mainly from nine countries: United State, India, Vietnam, Italy, Venezuela, Russia, Serbia, Poland, and Egypt.

²We used specific hashtags to collect the tweets for example: mkr, my kitchen rule, is one of the important hashtags which leads us to other co-occurrence hashtags.

³It could be interesting to know the age intervals of the raters. This information was not collected during the time of study

⁴While in the pilot study the gender of raters is very imbalanced in the crowd-sourcing phase the contribution of the two genders is less imbalanced

We selected the top ten raters based on their submission rate. Figure 3.1 shows the submission rate for each country.

Table 3.3: Details of Each Task and the Scales

Metrics	Scale(/5)
Instruction Clear	4
Test question fair	4.1
Ease of job	3.8
Pay	4.2
Overall	4

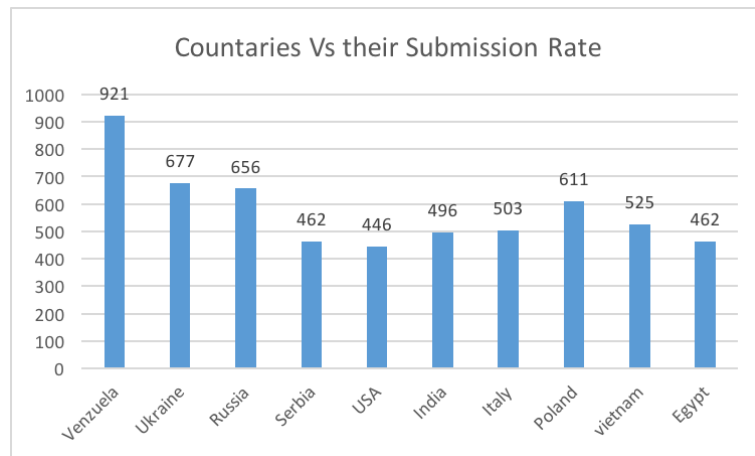


Figure 3.1: Submission Rate by the Country..

We were provided with the confidence of each rater and their gender by the Figure Eight platform. By simple calculations, we get Table 3.4 which shows the contribution of each gender in each category along with the total confidence in each category. The highest confidence score was for classifying the tweets belonging to the fourth category, not-sexist. Looking into the final labels, we understood a wide range of sexual harassment tweets were misclassified as not sexist when in fact they were indirect harassment.

Perhaps distinguishing between indirect harassment and not sexist tweets is a difficult task because it depends more on the culture or maybe because it does not have any leading keywords such as insulting words. In addition, raters might be biased toward different concepts, words and sentences. Based on their culture and

their life experience, they might label each tweet differently from their peers and bring their bias to the labeling. It is an interesting future line of work to examine the bias and discrimination in the process of labeling.

The second highest confidence is related to the second category, sexual harassment. One of the reasons is that this type of sexism is a clearly defined type of sexism culturally and it has a lot of keywords to help to classify this category. The least confidence is *physical harassment*. We believe the reason for this is because distinguishing between sexual harassment and physical harassment is a difficult task. It needs a lot of background information about this topic and the tweets need to have clear keywords.

Table 3.4: Gender Contribution and the Confidence Score

Categories	Contribution of Female Raters	Contribution of Male Raters	Total confidence of each category
Indirect harassment	3573 (84%)	653(15%)	0.57
Sexual harassment	2614 (86%)	407(13%)	0.65
Physical harassment	1739(82%)	373(17%)	0.44
Not sexist	9373(82%)	2039(17%)	0.80

3.4.2 Setup

For the pre-processing task, we removed stopwords, words were lemmatized and stemmed using Gensim and NLTK libraries⁵. For each category, we made the tfidf matrix and apply LDA using Gensim library.

For the classification task, we prepared three different corpuses as follows: Bag of words, tfidf and Glove. The bag-of-words model was made considering only the ten thousand most frequent words. In the bag-of-words method, we consider the count of each word as the features. For the TFIDF method (term-frequency inverse-document-frequency), we consider the counts as the term-frequency and the inverse-document-frequency is the logarithm of the division between total number of samples and the words in the training subset. The resulting values are normalized by dividing the largest value.

⁵We adopt these pre-processing tasks such as stemming, from previous literature on Twitter data analysis[3]

We convert each word to its embedding using Glove pre-training word representation on Twitter [108]. This model is proposed by Stanford NLP team and they apply co-occurrence probability for making the embedding. In this model, if words occur together frequently, they would have more similar word embedding. It is pre-trained on 2B tweets, 27B tokens, 1.2M vocab and they have four versions of word vectors with dimensions (25, 50, 100, 200). For our task, we chose word vectors of length 200. Maximum length of 10 words is considered for each tweet. Vector of the tweets is made by concatenating the vectors of the vector of the words in the tweets.

After transforming each tweet to a vector using BoW, tfidf and Glove, we feed these vectors for each method separately to the RandomOverSampler library in sklearn to balance the dataset.

We used Long Short Term Memory (LSTM) using Keras, 8 batches of size 30 each, sequence-length of 15, a learning rate of 0.001 and size of 128. For CNN, we used 2 ConvNets each having 9 layers deep with 6 convolutional layer and 3 fully-connected layers to regularize. We also used two dropouts with a ratio of 0.5. We partially use the code presented in Xiang Zhang et al. GitHub [157].

3.4.3 Topic Modeling

Topic modeling is a method to extract the hidden topics from the corpus of the documents. They are also a good tool for data exploration, when there is a new corpus and we would like to explore different types of structures which could be possibly found. In this research, we used topic modelling to explore the newly built dataset and identify the latent topics in each category.

Initially, one topic has been extracted from the documents but later this assumption was expanded to extracting the mixture of the topics from the documents. One of the most prevalent methods in topic modeling is Latent Dirichlet allocation (LDA). We combined the tweets of each category and deployed topic modeling to report the most frequent content topics using LDA. Table 3.5 shows the result of LDA in each category⁶. The first category, indirect harassment, has words describing women who cannot do anything and they are mentally weak. The second category, sexual harassment, has mostly swear words regarding hate and sex. The third category has words

⁶We ran LDA on separate categories to extract the topics in each category

which are more about the appearance of the women.

In NLP tasks, the input is usually a sequence of words which has a semantic and a meaning. In tfidf method, we ignore the position of each word in a sentence and as a result the semantic and meaning of the sentence is lost. Therefore, we use word vector representation. This representation allocates a real value vector to each vector and are used as feature vectors in NLP applications. An advantage of this method is keeping the context of the words without human intervention.

Table 3.5: The First Most Frequent LDA Topics in Each Category

Categories	LDA Results
Indirect harassment	Women, arrogant, cant, like, think, sassy, dumb
Sexual harassment	Blonde, pigs, bitch, make, sex, hate
Physical harassment	Pretty, ugly, sane, without, gang, look, attention, smash
Not sexist	Women, girls, blessed, empowering, lucky, feel

3.4.4 Word Embedding Space

Word embeddings are low-dimensional vectors used to represent words in the meaning space [9]. The purpose of using embedding is to be able to work in a space where semantically similar vectors are close. Word vectors have been shown to be very beneficial for natural language processing since they can represent a vector for each word along with syntactic and semantic differences in the properties of a word in comparison to others. They can learn in an unsupervised manner and it brings a great advantage for different purposes. One of those advantages is the word similarity task.

In this thesis, we tried to find the closest words in the embedding space using cosine similarity to the two categories of the words, *women*, *girls* and *men* using Word2vec. Table 3.6 shows the list of words. In the first category, indirect sexism, the most similar words around *women* and *girls* are *annoying* and *cooking* while the similar words to *men* are *being better*, *have the ability to think* or *having the power*. The second category, sexual harassment, the similar words to the *women and girls* are *porn*, *nude* and *f.cking* and to the *men* are more about their *sexual features*. The third category, physical harassment, the most similar words to *women and girl* are more about the appearance of the women such as *face*, *their look* and *the colour of*

their skin. The most similar word to *men* is *criticizing*.

Table 3.6: Closest Words to Women and Men in Embedding Space

Categories	Most similar words to Women, girls	Most similar words to men
Indirect Harassment	Tell, annoying, speak, cooking, cant, hate	Better, more, power, think
Sexual harassment	Bitch, porn, fucking, nude, slut	Shit, c..ck, can, girls, bitch
Physical harassment	Bitch, black, old, looking, face, see	Babe, honest, watch, unattractive
Not sexist	Good, girls, like, can, see	Blonde, jaws, game

3.5 Results

We defined the task as a classification task including four categories of "indirect harassment", "physical harassment", "sexual harassment" and "not sexist". Table 3.7 shows the result of three classifiers namely Naive Bayes, LSTM and CNN on three word vector models. Generally the convolutional neural network has high performance in comparison to LSTM or Naive Bayes considering different vector representation models. In the second place, LSTM achieves the higher score in comparison to Naive Bays.

Table 3.7: Classification Results on Accuracy

Corpus	Naive Bayes	LSTM	CNN
Bag-Of-Words	0.70	0.78	0.80
TF-IDF	0.73	0.89	0.90
GLOVE	0.79	0.91	0.93

In addition to the accuracy score, we present the F1-measure scores for each class separately using glove and CNN and the average of them in Table 3.8. Since we balanced the dataset using oversampling, interpretation of F1-measure for the classes is hard and using accuracy is more reasonable.

Table 3.8: F-measure Score of Classes of Harassment Using Glove and CNN

Categories	CNN
Indirect Harassment	0.88
Sexual Harassment	0.83
Physical Harassment	0.86
Not Harassment	0.89
Average F-score	0.86

3.6 Discussion

In our experiment, the highest accuracy score is for using Glove as word embedding method and cnn and it is mainly because of using a pre-trained word embedding model trained on Twitter dataset and second because tweets have a lot of abbreviated words which are not a complete words and considering the computation on the character level extracts more information for us. Another reason is that the convolutional neural network works better on the text which is less curated such as the reviews or the tweets. In our experiment, the convolutional neural network has superior performance with naive Bayes and it is compatible with Zhang et al. study [164]. In their study, they mentioned that ConvNets has a good performance with user-generated data and real-world generated data.

3.7 Summary

In this study, we presented new categories of online harassment on social media. After making the right instruction, we collected and labeled the data. Then, we tried to understand each category in lexicon level using topic modeling and trying to find the closest words in the embedding space. We addressed the initial classification results on accuracy after balancing the dataset.

Chapter 4

Understanding Different Types of Online Harassment Using Convolutional Filters

4.1 Goal

Deep learning algorithms have become a strong tool for producing impressive results in many classification tasks. They have achieved high predictive accuracy in different applications. It is necessary to understand for a specific task if accuracy stems from the use of suitable problem representation. In the previous chapter, we noticed that a character-level convolutional neural network with Naive Bayes has a high accuracy relative to other methods. In this chapter, we try to analyze the linguistic characteristics of online harassment tweets in each category by interpreting a convolutional neural network. We interpret a CNN model trained to classify harassment in order to understand these categories, by detecting semantic categories of ngrams and clustering them. Then, these salient ngrams in each category are used to improve the performance of the classification task. The main contributions of this chapter are as follows: we investigate the application of a trained CNN model on different online harassment tweets, such that we can derive the ngrams processed by the model for each category. We use the clustering to understand and reveal different patterns in these ngrams by category, we use this information to improve the classification task when the data is imbalanced and scarce and finally we show we can improve the predictive power of simple models such as Naive Bayes by injecting useful features from a model with higher accuracy.

In the previous chapter, we used Word2vec as the word representations but in this chapter we will use pre-trained glove on Twitter data. The same setup and hyper-parameters will be used for Naive Bayes, LSTM and CNN.

4.2 Introduction

Convolutional Neural Networks (CNNs) have been successfully applied to classify hateful speech [42]. Inherent non-interpretability of CNN is a challenge for users of sexism detection, the reason being that classifying a given speech instance with regard to sexism is difficult at a glance from a CNN. However, recent research has developed interpretable CNN filters for text data [55]. In a CNN, filters followed by different activation patterns along with global max-pooling can help us tease apart the most important ngrams from the rest. Clustering these useful ngrams can reveal different aspects of a dataset and hidden meanings of it. In a recent study, researchers proposed a new dataset focusing on different types of online harassment. This dataset contains three classes of “indirect harassment”, “physical harassment” and “sexual harassment” [120].

Convolutional neural networks were originally invented for the computer vision field and subsequently successfully utilized for natural language processing tasks [20, 63] such as semantic parsing [161], sentence modeling [62], text generation (GAN), hateful speech classification [146], online harassment detection [120] and effectual detection of harassment tweets [118, 117]. In more related studies, CNN was used for hateful speech classification [15]. In another study, a combination of CNN with gated recurrent unit network was used to classify thousands of hateful speech instances [165].

4.3 Setup

Inputs were embedded using pre-trained GloVe Wikipedia 2014 50-dimensional vectors [108]. The convolutional layer uses 15 filters of size 3 (fine-tuned from combinations of filters of sizes $\{2,3,4\}$). The convolutional layer calculates the inner product between each filter and each n-gram (in our case, 3 words) assigning a score to each n-gram. These scores are then fed into a max-pooling layer which selects the top-scoring ngram per filter and a ReLU activation. The score vector is then classified by a linear classifier. For optimization, we use the Adam optimizer. The dataset is divided into 85 percent training samples and 15 percent testing. The model achieved an 87% accuracy on the test set.

4.4 CNN Interpretation

After running the model and fine-tuning, we proceeded to interpret the model in accordance with methods introduced by Jacovi [55]. First, we derived the identity class for each filter, based on the respective weights of the filter in the final linear classifier of the model. In essence, the highest identity score shows the class for which the ngrams chosen by the filter provide evidence for the classification. For example, a filter of identity for indirect harassment class means that when this filter outputs a strong activation (supported by a specific ngram in the input), then this is evidenced by the model towards classifying the input as indirect harassment.

Next, we investigated informative features in the input: ngrams whose activation passed a certain threshold value (calculated heuristically per previous work) were deemed as *informative ngrams*, while the rest were deemed *uninformative ngrams* which passed the max-pooling layer in the model in spite of their low value in the classifying task. For each filter, we gathered the informative ngrams and discarded the non-informative ones as noise. We detailed the ngrams that received the strongest activation in the dataset (i.e. the ngrams that serve as the strongest evidence for the classification of the filter’s identity class). These ngrams are representative of the semantic meaning that the filter detects. After that, we clustered this group of biggest ngrams using Mean Shift Clustering [15] based on the activation of each word in the ngram.

Among all the filters, we have chosen three that are strong indications for classes "indirect Harassment", "Sexual harassment" and "Physical harassment" respectively. Filter #1 has the highest identity number to predict class "Indirect harassment". This means that ngrams chosen by this filter serve as strong evidence towards an Indirect harassment classification of the input tweet. Table 4.1 shows the information captured for the first class, indirect harassment. This information includes the biggest ngrams(three gram) of the filter, the clustering result which shows the number of clusters of the biggest ngrams (two in all of our reported filters), the biggest ngrams in each cluster (representative of the cluster) and a sample of the tweets which resulted in the chosen ngrams for the cluster.

Looking into Table 4.1, *indirect harassment*, we understand that the words are more about cooking, and it is mainly because "#mkr" was one of the main hashtags

Table 4.1: Information Captured from Filter #0, Size 3 for the Indirect Harassment Class.

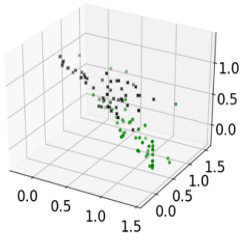
Results	values
<i>Indirect harassment biggest n-grams</i>	Kitchen rules eat, food will smell, nah nigga, rice server who, your not cooking, peanuts amp biscotti
<i>Clustering figure</i>	
<i>Words in the first cluster (57/107)</i>	Kitchen rules eat, food will smell, burning foods grapefruit, food and eating, meat and peanut ,preparation amp cooking, cows and pigs.
<i>Sample of tweets containing above words</i>	maybe you girls should go back to cutting up cows and pigs dont quit your day job. i cant even watch my kitchen rules without cringing at these girls
<i>Words in the second cluster (50/107)</i>	Nah nigga, wheat amp, cows don't make, youre unreliable that's, coworker ive run, blankets amp pillows.
<i>Sample of tweets containing the above words</i>	girls think they can trap you with a baby bruh nigga i will end my life. i am so sick amp tired of this attitude oh wow youre smart for a girl.

Table 4.2: Information Captured from Filter #12, Size 3 for the Sexual Harassment

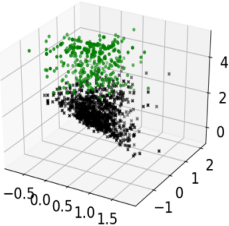
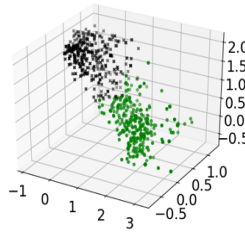
Results	values
<i>Sexual harassment biggest n-grams</i>	Black porn bitch, a massive cunt, or acid bitch, dm your bitch, girl nude pics, a messy bitch
<i>Clustering figure</i>	
<i>Words in the first cluster (1104/1473)</i>	Their own asses, ugly girls nakes, girls wearing sexy, balck girls naked, live hard nipples.
<i>Sample of tweets containing above words</i>	-You know that has zero to do with left right amp you know Don walked in on naked teen girls. she is live Hard nipple F new College
<i>Words in the second cluster (369/1473)</i>	girl free bitch, fake arse bitch, lil gangsta bitch, little spoiled bitch, amessy bitch
<i>Sample of tweets containing the above words</i>	the blonde in red has breeders hips in that dress model my arse. hate these blonde bitches already

Table 4.3: Information Captured from Filter #8, Size 3 for the Physical Harassment Class

Results	values
<i>Physical harassment biggest n-grams</i>	Blondes with pretty, blondes and pretty, unattractive than girls, blondes as they, sexy hot girls, sexy fotos nude.
<i>Clustering figure</i>	
<i>Words in the first cluster (378/719)</i>	Unattractive than girls, naked paintball girls, ugly promo girls, naked cuban girls, naked black girls, sexy wet girls.
<i>Sample of tweets containing above words</i>	elegant and beautifulcheap and trashynothing more unattractive than girls banging on about how hot hey are. i love pretty girls and my friendly ass will look at you and smile but if you wanna be stuck up then bitch you re ugly
<i>Words in the second cluster (341/719)</i>	Bloneds kill the, unapologetically attacked her, bloned fucked with, bitch without my
<i>Sample of tweets containing the above words</i>	-awful child you know youre killing me youre. can t be a basic bitch without my starbucks pic

that were used for collecting the tweets in a study conducted by Sharifirad and Matwin [120]. It is the abbreviation of “my kitchen rule”, an online cooking TV show in which the audience can write tweets about the way they are cooking. If participants in the competition are women, these tweets are often indirect harassment or physical and sexual harassment. We extracted about 107 informative ngrams for this filter. We cluster the ngrams into two clusters. Based on the words in each cluster, it becomes clear that words in the first cluster are about cooking and the other group is more about words to call girls stupid and so forth.

Table 4.2 shows the results in the sexual harassment category. In total, there have been about 1473 informative ngrams detected. Most of the ngrams are clustered in the first cluster, mentioning the sexual attributes of the girls, and the second cluster detects all the ngrams which contain the word bitch and it is the most frequent word. The third category, physical harassment (Table 4.3), the biggest ngrams are divided into two clusters. In the first cluster, words are around the physical and appearance of the girls for example if they are pretty, the second cluster is more about threatening words ¹.

4.4.1 Results

After getting the informative n-grams in each category, we added them to the tweets of that class and repeated the classification task. We used Glove to get the embedding vector for each word of length 200 pre-trained on Twitter dataset. For each tweet, we concatenated all the word vectors together. We considered Support Vector Machine (SVM), Naive Bayes and CNN as the classifier to compare the performance of the classifiers and see if adding these informative ngrams can improve the performance of the classification. After having the most informative ngrams, we added them to the end of each tweet in the class the ngrams belong to. We repeated the previous method for getting the tweet vectors and re-ran the classifiers. It is worth mentioning, we balanced the dataset for both original dataset and augmented dataset using the oversampling explained in the previous chapter. Results are presented in Table 4.4.

¹It would have been possible to use TF-IDFs as a filter methodology for feature selection. However, Using CNN for feature selection is similar to wrapper methods, which learns features by the interaction with the learning model, wrappers are usually a better approach for feature selection in comparison to Filter[134]

Looking at Table 4.4, the highest accuracy is for the convolutional neural network on the augmented text. So far we have deployed two types of convolutional neural network. In the previous chapter, we used a character level convolutional neural network and in this chapter we used a shallow cnn. In comparison to the proposed methods in the previous chapter, augmenting useful information to the text has better impact in comparison to having a more complicated model.

4.5 Discussion and Conclusion

Deep learning algorithms can not only improve the performance of the classification but also can help us understand the dataset better and finally improve the classification. We tested convolutional filters on a dataset related to different types of online harassment to test if the filters can help us understand the nature of the dataset where classes share the same words. The filtering of uninformative ngrams and clustering helps to understand what the model considers important in the input space of text tweets for the harassment classification task. Each cluster of the most informative ngrams in each category refer to a different definition of that category. For example, filter #zero (Table 4.1) which presents the Indirect harassment category shows two explicit categories of ngrams, those which are related to cooking and those which have an indirect weight of harassing, even though these ngrams in each category separately do not present information but pairing them together can make a large number of indirect harassment tweets. In the second category, sexual harassment (Table 4.2), the ngrams were divided into being insulting or focusing on provoking words. These two categories of ngrams together can make a large number of sexual harassment tweets. The third category, physical harassment (Table 4.3), has two clusters of information, one focused on the physical attribute of the girls while the second class focused on threatening the girls physically. These findings can help us understand the nature

Table 4.4: Accuracy of Classifiers on Original and Augmented Dataset.

	Original dataset accuracy	augmented dataset
<i>SVM</i>	0.76	0.80
<i>Naive Bayes</i>	0.79	0.83
<i>CNN</i>	0.87	0.95

of these types of online harassment better. In addition, these findings would help us obtain a semantically higher level of representation in which an interpretation of the profiles of tweets in each class becomes possible.

As future work, we would like to expand the technical contribution of the method used to other larger abusive datasets and analyze the implications of the uncovered informative features on the task. In addition, while adding informative ngrams as features to models like Naive Bayes and SVM improves the performance, the improvement on the CNN model appears circular (since these features were derived from the CNN itself). It is worth trying to look at the informative ngrams of CNN again after adding the previous n-grams.

Chapter 5

Understanding Emotion Intensity and Emotion Type in Online Harassment Tweets

5.1 Goal

A vital part of the available information in social media is to perceive what is the opinion of people around different topics and what is their feeling. It is important to apprehend the drivers behind the publicly available sentiments using the connection between emotions and lexical sentiments. These sentiments can reveal the mental state of the users. In the previous chapters, we tried to make a novel dataset around online harassment and identify the important features and keywords in this dataset. In this chapter, we would like to understand the reasons which lie behind such statements in social media. In other words, we would like to understand the feeling and mental state of those who write online harassment tweets. For this purpose, different semantic representation embeddings are considered with different classifiers. First, these models are trained on two sub-tasks of the SemEval 2018 dataset, emotion intensity detection and emotion type detection [87]. Then the model with the highest accuracy is tested on the harassment dataset to identify the emotion intensity and emotion type in each category of harassment, Figure 5.1 shows the schema.

In this chapter, we use Glove and word2vec from chapter three and four and add Fattext as another word representation. In addition, we use the same algorithms with the same hyper-parameters and setup such as Naive Bayes, LSTM and CNN from the two previous chapters, and add other classifiers such as SVM, KNN and MLP.

5.2 Introduction

There has been an increase in the use of social media to harass, threaten or humiliate other users. Social science identifies the crucial components and features of harassment comments but ignores other aspects such as the emotion of this type of speech

and potentially harmful consequences to the victim’s emotion.

In this chapter, we took advantage of a newly released dataset “SemEval-2018 task1: Affect in tweets” [88], to show the type of emotion and intensity of emotion in online harassment categories. We trained, tested and evaluated different classification methods on the SemEval- 2018 dataset and chose the classifier with the highest accuracy. Then, we tested the pre-trained model on each online harassment category to know the mental state and the effective state of the user who writes harassing tweets. It is a nice venue to explore because not all the tweets are directly sexist and they carry different emotions. Based on our best knowledge they are all new contributions to the field; we are the first to demonstrate the power of such in-depth sentiment analysis on the online harassment tweets.

In this chapter, we focused mainly on two datasets. The first dataset is related to two tasks of emotion detection type and emotional intensity level, Table 5.1 shows the detailed number of tweets in each category. The second dataset is related to online harassment dataset; it is comprised of the four categories mentioned in the

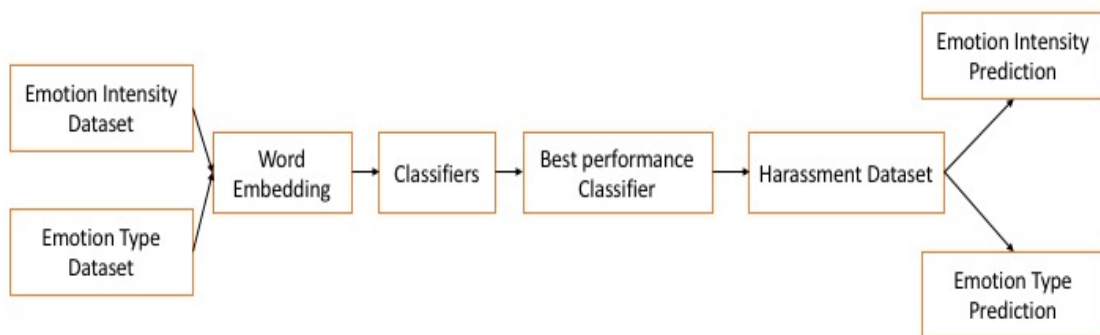


Figure 5.1: Schema.

first chapter. We trained classification algorithms on all the initial tasks (detecting emotion type, the intensity of sentiment) chose the one with the highest accuracy and then tested it with the sexist tweets to know the category of emotion and intensity of emotion in each category.

The contributions of this chapter are as follows:

- We trained, validated and tested different classification methods on the shared task dataset “SemEval 2019: Affect in tweets” [88]. We chose the best method on the Semeval2018 dataset and tested it on each online harassment categories. It helps us understand the **emotion type** of users who post harassing tweets.
- We trained, validates and tested different classification methods on the shared task dataset “SemEval 2019: Affect in tweets”. Then, we chose the best method on the Semeval2018 dataset and tested it on each online harassment categories to know the **emotional intensity** of users who post harassing tweets.

5.2.1 Text Preprocessing

Preprocessing of the tweets involves removal of the punctuation, hyperlinks/URLs, emoji and tags. Stop words were not removed because some stop words like ”not” were important for the sentiment of the sentence. Before training the classification models, WordNet lemmatization was applied to all the tweets. We set the maximum size of each tweet to 40 words and padded the tweets of shorter length with zeroes. Next, tweets were converted into the vectors using Word2vec [83], Glove [108], FastText[61] all with the length 300.

5.3 Experiment

5.3.1 Experimental setup

We experimented and trained the classifiers on emotion type detection and emotional intensity datasets. We used different representation learning for the words. For the embedding based methods, we used Word2vec [83]. Word vectors are trained from ten million English tweets from the Edinburgh Twitter Corpus [113]. The Word2vec parameters were window size of two and length of 300. We also used Glove, embedding

size of 300; its embedding has been pre-trained on around 2 billion tweets. The other embedding was FastText; these embeddings were trained on Wikipedia pages, considering the default mode, and embedding size of 300. After getting the vector for each word, we concatenated them to get a vector for each tweet. For the out of words vocabulary, we considered the vector of each character in the word and concatenated them to get the same length word vector. We trained the classifiers on each training set, then validated with the validation set and finally reported the accuracy on the test set.

Table 5.1: Distribution of Emotion Intensity for Anger, Joy, Sadness and Fear [87]

Categories	Number (train)	Number (test)	Number (validation)
No anger can be inferred(#0)	445	465	186
Low amount of anger can be inferred(#1)	322	148	54
Moderate amount of anger can be inferred(#2)	507	243	97
High amount of anger can be inferred(#3)	427	146	51
No joy can be inferred(#0)	548	194	55
Low amount of joy can be inferred(#1)	362	333	95
Moderate amount of joy can be inferred(#2)	346	360	89
High amount of joy can be inferred(#3)	359	218	51
No sadness can be inferred(#0)	594	398	170
Low amount of sadness can be inferred(#1)	260	193	88
Moderate amount of sadness can be inferred(#2)	364	255	87
High amount of sadness can be inferred(#3)	315	129	52
No fear can be inferred(#0)	1490	633	186
Low amount of fear can be inferred(#1)	320	124	54
Moderate amount of fear can be inferred(#2)	249	158	97
High amount of fear can be inferred(#3)	193	71	51

5.4 Results

Initially, we tried different methods on the SemEval 2018 dataset; Table 5.2 shows the accuracy of the methods on two tasks of emotional intensity on anger and fear and Table 5.3 shows the results on the other two datasets namely joy and sadness. In addition, Table 5.4 shows the accuracy of algorithms on the emotion type detection. We picked the method with the highest accuracy to test it on the three online sexual harassment categories to know the emotion type and intensity of emotion in each category. Starting from the emotional intensity, the first category is “anger”; the highest accuracy is about 93 percent using FastText as the embedding vector and CNN as the classification algorithm. The second category of emotional intensity was “fear”; the highest accuracy is about 91 percent with the embedding of FastText and the CNN as the learning algorithm. Coming to the third category as “joy”, the highest accuracy is about 90 percent using FastText as the embedding vector and CNN as the classification algorithm. For our task, fine-tuning of CNN and using a high number of epochs for learning was effective. In comparison to the baselines, deep learning methods take advantage of multiple learnings and when the number of samples is not significant, using a big number of epochs is helpful. The next task was related to the emotion classification; this multi-label classification was not an easy task and the accuracy was not as good as the previous task. The highest performance is related to FastText as the embedding vector and the CNN as the classification method with the accuracy of 0.85 percent. Finally, the classification of the sentiment intensity or valence, the highest accuracy relates to the FastText embedding and CNN for the choice of classification.

Table 5.2: Accuracy of Methods on Emotion Intensity on Anger and Fear

Algorithms	Word vectors	Anger	Fear
One Vs all	W2v/Glove/ FastText	0.53/0.54/0.55	0.60/0.63/0.65
SVM	W2v/Glove/ FastText	0.62/0.64/0.67	0.67/0.69/0.70
NB	W2v/Glove/ FastText	0.64/0.66/0.68	0.73/0.75/0.77
KNN	W2v/Glove/ FastText	0.65/0.66/0.68	0.75/0.77/0.79
MLP	W2v/Glove/ FastText	0.82/0.83/0.85	0.82/0.84/0.86
LSTM	W2v/Glove/ FastText	0.86/0.86/0.88	0.84/0.85/0.89
CNN	W2v/Glove/ FastText	0.88/0.89/0.93	0.88/0.89/0.91

After choosing our best choice of classifier and word vector, we tested the algorithm on the online harassment dataset to see how each category is different from the other one in terms of emotion type and intensity of emotion.

The first task was related to the emotional intensity of “anger”, “fear”, “joy” and “sadness”, which is shown in Table 5.5. Tweets can have a range from 0, implying no anger, to 3, showing high anger. Out of 260 total tweets, about 240 tweets were categorized as showing no anger, 89 tweets showed slight anger in the indirect harassment tweets. In terms of the emotional intensity of fear, considering the same range, about 240 tweets were categorized as showing no fear while a small number showed slight fear. In terms of the emotional intensity of joy, around 128 tweets were categorized in the moderate level of joy. This shows that the tweets in this category imply sarcastic characteristics of the users, they tweet a positive sentence but in a sarcastic way. The

Table 5.3: Accuracy of Methods on Emotion Intensity on Joy and Sadness

Algorithms	Word vectors	Joy	Sadness
One Vs all	W2v/Glove/ FastText	0.53/0.54/0.57	0.66/0.75/0.67
SVM	W2v/Glove/ FastText	0.62/0.63/0.66	0.68/0.69/0.70
NB	W2v/Glove/ FastText	0.67/0.67/0.69	0.72/0.74/0.76
KNN	W2v/Glove/ FastText	0.65/0.66/0.68	0.68/0.68/0.71
MLP	W2v/Glove/ FastText	0.79/0.79/0.82	0.78/0.79/0.82
LSTM	W2v/Glove/ FastText	0.79/0.83/0.86	0.81/0.83/0.86
CNN	W2v/Glove/ FastText	0.87/0.89/0.90	0.87/0.88/0.93

Table 5.4: Accuracy of Emotion type detection

Table4.Accuracy on Emotion type detection		
One Vs all	W2v/Glove/ FastText	0.33/0.35/0.41
SVM	W2v/Glove/ FastText	0.51/0.56/0.59
NB	W2v/Glove/ FastText	0.53/0.55/0.57
KNN	W2v/Glove/ FastText	0.52/0.54/0.55
MLP	W2v/Glove/ FastText	0.73/0.76/0.77
LSTM	W2v/Glove/ FastText	0.79/0.83/0.84
CNN	W2v/Glove/ FastText	0.80/0.84/0.85

last category pertains to the emotional intensity of sadness. The majority of tweets are in categories showing slight sadness in the tweets. The second task was about the emotion type detection task; we classified the tweets into eleven categories. Tweets in the indirect harassment category, mostly carry optimistic feelings, anger, and joy. Table 5.6 represents the results.

In terms of intensity of sentiment, sexual harassment category had the highest amount of tweets as being very negative or slightly negative. It had the highest amount of anger in the tweets. It classified almost no sense of fear. The tweets in this category also had the highest intensity of joy and a moderate intensity of sadness. In terms of emotion type classification, most tweets carried disgust, joy, and sadness. It makes this story in the mind that users, who send sexual harassment tweets, become angry from a tweet and enjoy harassing the opposite sex by showing deep disgust towards her.

The third category shown in Table 5.5 is physical harassment. In terms of intensity of emotion, shown in the table, the highest intensity of sentiment is in this category. About 119 tweets show the high intensity of anger, no intensity of fear, high intensity of joy and high intensity of sadness. It has high emotion of anger, disgust, and sadness. It seems there is a lot of similarity in terms of the type of emotion and intensity of emotions in these two categories.

Table 5.5: Final Results of Emotion Intensity on Harassment Categories

IndirectH	EI of anger(0/1/2/3)	170/89/1/0
IndirectH	EI of fear(0/1/2/3)	240/20/0/0
IndirectH	EI of joy (0/1/2/3)	12/0/128/120
IndirectH	EI of sadness (0/1/2/3)	90/140/30/0
SexualH	EI of anger(0/1/2/3)	0/7/30/380
SexualH	EI of fear(0/1/2/3)	370/47/0/0
SexualH	EI of joy (0/1/2/3)	0/0/27/390
SexualH	EI of sadness (0/1/2/3)	0/63/183/171
PhysicalH	EI of anger(0/1/2/3)	0/0/4/119
PhysicalH	EI of fear(0/1/2/3)	114/9/0/0
PhysicalH	EI of joy (0/1/2/3)	0/8/10/105
PhysicalH	EI of sadness (0/1/2/3)	0/12/23/88

Table 5.6: Final Results of Emotion Type in Each Category

Table6.Final results of emotion type in each category	
Categories	Emotion(anger/anticipation/disgust/fear/joy/love/optimism/pessimism/sadness/surprise/trust)
Indirect Harassment(260)	10/18/0/0/120/0/1/0/0/80/0
Sexual Harassment(417)	64/0/229/0/83/0/0/0/41/0/0
Physical Harassment(123)	55/0/43/0/0/0/0/4/21/0/0

5.5 Discussion

Overall, FastText and CNN have the best performances on the datasets. Interestingly enough, the emotion type and emotional intensity have a lot to say in each online harassment category. Starting from the indirect harassment, based on the original description of this type of harassment, the tweets are not directly violent. However, they indirectly show a kind of superiority of the men over women. Based on the results, in terms of emotional intensity, in the first place, they show no level of anger, fear or sadness except for joy. In addition, the intensity of joy is high in indirect harassment.

In the sexual harassment category, the emotional intensity of anger, joy, and sadness is high but the intensity of fear is not high. It shows that users who tweet sexual harassment tweets are usually angry, sad and enjoy writing these types of tweets. However, there is no high intensity of fear in these tweets. It presents that users have no fear to tweet these kinds of tweets. In the physical harassment category, there is a high intensity of anger, joy, and sadness but a low intensity of fear. The similarity of the results in the two categories of sexual harassment and physical harassment is expected because they share many words and semantics.

Another discussion is related to emotion type. Indirect harassment has the biggest number of tweets categorized in joy, then surprise and anticipation. It is no surprise that joy has the highest number of tweets and it is in line with the previous results. Indirect harassment contains other complimentary definitions such as when males expect women to behave in a certain way or they show their surprise in a sarcastic way in the tweets. For example, the tweet, "she plays as good as a boy", shows the surprise of the user when noticing a female plays well or the tweet, "she should come back to the kitchen", shows the anticipation of male users about females.

The second category, sexual harassment, has the highest number of tweets categorized as disgust, sadness, and anger. These results are in line with the previous results. In addition, disgust in this category is complementary to other emotions. In the physical harassment category, the highest number of tweets are categorized as anger, disgust, and sadness.

5.6 Summary

In this chapter, we tried to focus on a subject which has attracted a lot of attention these days. We used "Semeval task1: affect in tweets" and tried to understand the emotion type and intensity of emotion in each category of online harassment. After training the algorithms, we picked the algorithm with the highest accuracy and tested it on the online harassment tweets. This work is the first work of its type and shows there are some similarities in the physical and sexual harassment categories. Indirect harassment, known as benevolent harassment, inhabits a mild range of intensity while the other two have a very high intensity of disgust, anger, sadness, and even joy.

Chapter 6

Boosting Text Classification Performance on Sexist Tweets by Text Augmentation and Text Generation Using a Combination of Knowledge Graphs

6.1 Goal

Text classification models have been heavily utilized for a slew of interesting natural language processing problems. Like any other machine learning model, these classifiers are very dependent on the size and quality of the training dataset. Insufficient and imbalanced datasets will lead to poor performance. In previous chapters, we classified different types of online harassments and deployed techniques to understand this type of speech more clearly at the semantic level and also to understand the semantics of these types of speech. Unfortunately, online harassment like other sensitive datasets is hard to collect and annotate; and it is imbalanced. In this chapter, we addressed this problem and proposed a solution by using knowledge graphs such as ConceptNet and Wikidata.

We used ConceptNet [21] and Wikidata [147] as two linguistic resources to improve sexist tweet classification by (1) text augmentation and (2) text generation. In our text generation approach, we generated new tweets by replacing words using data acquired from ConceptNet relations in order to increase the size of our training set. In our text augmentation approach, the number of tweets remained the same but their words were augmented (concatenation) with words extracted from their ConceptNet relations and their Wikidata description. In our text augmentation approach, the number of tweets in each class remained the same. An interesting solution to low-quality datasets is to take advantage of the world knowledge in the form of knowledge graphs to augment our training data. Our experiments show that our approach improves sexist tweet classification significantly in all of our machine learning models. Our approach can be readily applied to any other text classification problem using

any machine learning model.

In this chapter, we use word2vec from chapter three as the word vector representation and other classifiers such as SVM, Naive Bayes, LSTM and CNN with the same hyper-parameters and setup as in previous chapters.

6.2 Introduction

Sentences contain different keywords and concepts. One way of understanding these concepts and getting more information about them is by using linked data and knowledge graphs. The popularity of the internet and advancements in linked data research lead to the development of internet-scale public domain knowledge graphs such as FreeBase [41], DBPedia [24], ConceptNet [21] and Wikidata [147]. Knowledge in popular knowledge graphs is usually mined from available online resources such as Wikipedia using natural language understanding techniques or harvested by crowd-sourcing or a combination of both.

Knowledge graphs are used to represent concepts and their relationships in a computer-understandable format. There are a wide range of applications of knowledge graphs in the text analysis domain such as question answering [158] query expansion [160], recommendation engines [136] and many more.

ConceptNet is a common sense knowledge graph that represents everyday concepts. It has approximately 21 million concepts and uses one of the 36 existing relationships such as IsA (e.g. jack IsA first name), UsedFor (e.g. car UsedFor driving) or PartOf (wheel PartOf car) [7].

Each fact in ConceptNet has a weight value which shows a degree of strength between two concepts. Higher values mean more confidence or more strength in which two concepts are connected in the graph. In other words, it shows the closeness of the concepts to each other. Wikidata is a wiki project that is used to crowdsource structured data which is consumable both for humans and machines. Wikidata contains 4400 types of relationships between more than 45 million concepts.

ConceptNet and Wikidata are far from perfectly consistent and complete. Therefore, we use both of these knowledge graphs in our approach for better coverage of the word knowledge with more consistency. An interesting source of information in Wikidata is the concept's descriptions. We used these descriptions for augmenting tweets.

For the text generation task, we replaced words in each tweet by words that they are connected to in ConceptNet using some of its 19 relations such as *IsA*, *RelatedTo*, *HasA*, *HasProperty*, etc.

Our contribution in this chapter is using ConceptNet, Wikidata and a combination of both for text generation and augmentation in order to improve sexist tweet classification. Even though we have used our approach to sexist tweet classification, it can be readily applied to other text classification problems using any of the existing text classification models. The rest of the chapter is as follows: In the next section, we will discuss the prior work on sexism, text generation, and text augmentation. Then, in the experiment part, we will go through the dataset, text pre-processing, classification algorithms, the detailed method of text generation and text augmentation. In the result, we will show the result of text generation and text augmentation and finally the conclusion.

In this chapter, we suggest text augmentation by adding concepts from ConceptNet and Wikidata and description deprived of Wikidata. The detail of the methods is in the following sections. In addition, we argue that the relations and the concepts in the ConceptNet are not complete and their combination with relation and concepts from Wikidata are more useful and complete for this process. For this purpose, we present figures showing the complementary aspect of these two knowledge graphs. We present the knowledge graphs of Wikidata, ConceptNet and both and limit the number of nodes (concepts), to 10, and the number of relations for the purpose of clarity. We chose the word *bitch* because it was the most frequent non-stop-word in our corpus. Fig 6.1 shows the image of Wikidata knowledge graph around this word, it shows the related concepts to *bitch* are related to the other two concepts, *profanity* and *insult*, with the relation, *IsA*. Fig 6.2 shows a ConceptNet knowledge graph, it has more relations such as *IsA*, *Synonym*, *relatedTo* and *CapableOf* to the words such as *sugar-baby*, *cunt*, *canine*, *difficulty* and *backbite*. Fig 6.3 shows the combination of the two knowledge graphs.

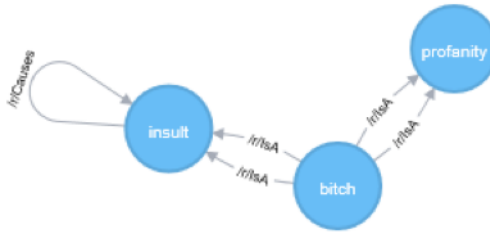


Figure 6.1: Wikidata Knowledge Graph.



Figure 6.2: ConceptNet Knowledge graph.



Figure 6.3: Combination of Wikidata and ConceptNet knowledge graphs

6.3 Text Generation

We generated new tweets using ConceptNet in order to improve coverage of our classes using three methods and compared their performance of classification using machine learning models. In the first approach which we call “All Words Replacement (AWR)”, tweets were tokenized and then each token (except for stop words) regardless of its grammatical role was replaced with all their FromOf and IsA relationships target in ConceptNet whose weight is greater than 1.0. We started from the first token and went forward until a specific number of new tweets had been generated. Relationships other than the ones listed in table 3 will lead to meaningless tweets that do not represent the original tweet. As an example, the output of ConceptNet for the query of the word “girl” is as follows: [relationship: IsA, target: woman, weight: 2.0, relationship: IsA, target: female person, weight: 1.0, relationship: IsA, target: female young human, weight: 1.0]. We then replaced the word girl with the words women, female person, and female young human. The second method which we call “Verb Replacement (VR)” was to first tokenize the tweet and replace the verb by its synonyms in the ConceptNet. In the third method, called “Noun Replacement (NR)”, all the process are the same as the second approach, VR, but with the difference that

we replaced only the nouns with the concepts coming from the ConceptNet. Table 5 shows the summary of the relation, the selected words, and the generated sentence. For each sentence, we show only one example of the newly generated tweet out of many. For each class, we used the original tweets to generate a balanced corpus of each category having fix number 1200 tweets. Table 6 shows the result of classification algorithms on the generated tweets.

	Sentence sample	Types of relation	Generated sentence
AWR	"Kathy you bitch need to slap your daughter "	FromOf IsA	"Kathy you cunt want to hit your mom "
VR	"Kathy you bitch need to slap your daughter"	Synonym	Kathy you bitch want to smack your daughter
NR	"Kathy you bitch need to slap your daughter "	Synonym RelatedTo	Kathy you bitch need to slap your mother

6.3.1 Text Augmentation

For text augmentation, we added the concepts from ConceptNet for the first proposed method. In the second method, we considered the concepts from the ConceptNet and Wikidata in a smart procedure. The number of tweets in the categories stayed the same. The first method is based on adding the related concepts to the original tweets. We tokenized each tweet, then considered "IsA" as the relation and chose the top ten related concepts based on their weight from ConceptNet and added it to the end of the tweet, even though the number of the tweets remained the same but the length of the tweets increased to the length of a paragraph. Table 4 shows an example of text augmentation. In the second approach, in addition to the augmentation of tweets using ConceptNet, we augmented tweets by the definition of their tokens in Wikidata. We tokenized the sentence, then added the top related concepts from the ConceptNet based on the sorted weight. After that, we combined ConceptNet with Wikidata. The output of the Wikidata around the word query "girls" is 39 tuples, we mention 4 of them as follows: ['q1.description': 'painting by Lisa Milroy',

'relationship': 'IsA', 'target': 'painting', 'q1.description': 'painting by Henri-Jean-Guillaume Martin', 'relationship': 'IsA', 'target': 'painting', , 'q1.description': 'young female human', 'relationship': 'IsA', 'target': 'female', 'q1.description': 'young female human', etc. Of all these concepts in Wikidata, only one of them pertained to the concept of a girl in ConceptNet. To choose the right concept from Wikidata, we first chose the top 10 concepts sorted by weight, then calculated the cosine similarity between the averaged word vectors of these concepts using Word2vec and the averaged vector of the words in the description from Wikidata. After sorting the descriptions based on the similarity score, we added the most similar description to that tweet¹.

Original tweet	Augmented tweet
"local girls near you that are down to fuck what links do yall keep clicking on to get hacked?"	local girls near you that are down to fuck rt what links do yall keep clicking on to get hacked public transport local organization smaller than national agent non geographical animanga character area unit passive verb feather hair highland strike get better of direction turn soft feather from goose hair feather mood landscape semisolid sexual intercourse rude word television station dehydrated may rehydrated right best human ear good all-purpose life but seeing difference film television show situation software solfa syllable travel create proceed carry through musical artist record confine have store stronghold grow lodge protect stay sound emission communicate destroy make buy return catch annoy touch hit seize get'?

6.4 Results

6.4.1 Experimental setup

Considering the dataset label distribution in Table 3.2, for the first task, text generation, we generated 1200 tweets in each category. We chose 1200 because it was the maximum number of tweets which could be produced from the category with the least

¹In this chapter, we considered the union of ConceptNet and Wikidata in order to take advantage of more relations and complementary information. However, one might be able to consider the intersections of relations for another study

number of tweets and we wanted to keep the dataset balanced. For the second task, text augmentation, the number of tweets remained the same as the initial distribution but the length of them changed as we add information to end of the tweets².

We considered each word as a token using python NLTK package and changed them into vectors using word2vec. For the word2vec, we used the genism library trained using CBOW and concatenated the vectors of length 300 to get a vector for each tweet. We used multi-class Naive Bayes in Scikit learn python, multiclass LSTM and multi-class CNN using Keras for the choice of classifiers. We divided the dataset into 70 percent train and 30 percent test. For each tweet, we made the labels in the form of one-hot encoding of length four and we used the same labels for all the classification process. We applied a CNN-based approach to automatically learn and classify the tweets into one of the four categories. During the evaluation, a grid search was applied to get the optimal number of filters and filter sizes. Also, we tried with multiple configurations of convolutional layers of 2, 4 and 6. The best performance consisted of two convolutional layers of each followed by a max pooling layer. Convolutional of size 256 with filter size 5 applied for all the convolutional layers. The dropout rate of 0.5 was implemented to avoid overfitting. In addition, a fully connected layer was used with a length of 128 followed by a second dropout function. This was followed by a dense layer with a size of 4 to represent the number of classification classes using the Softmax function. Our implementation is similar to the model presented in [42]. We trained a simple LSTM model including one hidden layer containing 256 nodes and rectifier activation on the hidden layer, sigmoid activation on the output layer ADAM gradient descent, and a negative likelihood loss function. We created 300 epochs and batch sizes of 5. Table 5 shows the results of the text generation.

6.4.2 Text generation results

Our first classification experiment was over the original dataset with three classes, since in the original datasets, the second class, indirect harassment, had only 6 tweets and in comparison to the other classes, it did not have enough tweets, we removed this

²While these two methods might duplicate some of the words, their use is justified by the goal to stay within the the i.i.d. framework

class and performed our classification algorithm on the rest of three classes. Our second classification approach, verb replacement (VR), was based on the four balanced classes each having about 996 tweets, coming from the first text generation method, all word replacement (AWR). The third classification experiment, noun replacement (NR), was on the four balanced dataset coming from the second method of text generation, each class having about one thousand data points and the last experiment coming from the third approach for text generation, each class having the same number of tweets. We used five classification algorithms, the one-versus-all algorithm as the baseline, Naive Bayes and SVM as more traditional classification algorithms and then two artificial neural network approaches, Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN).

On the original dataset, the highest accuracy, we achieved 75 percent using CNN (for more results please see Table 6). We believe this poor performance is due to poor coverage of our dataset and the imbalanced nature of the dataset. We aimed to alleviate these issues using ConceptNet. Experimenting on the second dataset, which was the generated data with all word replacement (AWR), shows considerably higher performance in comparison to the original dataset. In this dataset, all the four classes were balanced. LSTM and CNN both had the same and high performance in the classes. The second performance was related to the SVM and the last one was related to one versus all classification algorithm. In the third generated dataset, noun replacement, the highest performance relates to the LSTM and the second highest relates to the CNN with a very small margin. The highest performance was related to the LSTM for the third method of generated data, followed by CNN and the SVM and Naive Bayes. We ran different text generated methods to know the best to increase the

	OVR	SVM	Naive Bayes	LSTM	CNN
The original data	0.52	0.68	0.60	0.74	0.75
AWR	0.79	0.94	0.92	0.98	0.98
NR	0.77	0.83	0.85	0.92	0.91
VR	0.82	0.88	0.88	0.97	0.95

number of tweets in each class and balance it. It seemed all word replacement (AWR) of the sentence elements with specific relations from ConceptNet in combination with neural network yields the highest performance boost. As mentioned in table 5, all the generated methods had better performance in comparison to the raw data. VR had better performance in comparison to the NR. The best performance for the text generation method was AWR (all Word Replacement) using ConceptNet as the generation method and LSTM and CNN as the classification method. In addition, all the augmentation methods had better performance in comparison to the original data. The method in which we augmented the concepts from Wikidata and ConceptNet along with the description from Wikidata had better performance in comparison to the augmentation with ConceptNet. However, it did not have performance as good as the text generation method.

Methods	OVR	SVM	Naive Bayes	LSTM	CNN
Augmented with ConceptNet	0.55	0.65	0.64	0.90	0.88
Augmentation with ConceptNet and Wikidata	0.60	0.69	0.70	0.93	0.91

6.5 Summary

In this chapter, we introduced simple but effective methods for text generation and text augmentation using general purpose knowledge graphs. For text generation we solely used ConceptNet and for text augmentation, we used both ConceptNet separately and both ConceptNet and Wikidata. Since there was no mapping between ConceptNet and Wikidata, we used the cosine similarity of word vectors of related concepts in ConceptNet and words in the description of Wikidata in order to establish a mapping between their concepts. Application of our method to the problem of sexist tweet classification shows drastic improvements in classification results. Our approach can be applied without any modifications to any other text classification problem.

Chapter 7

A Case Study: Using Attention-based Bidirectional LSTM to Identify Different Categories of Offensive Language Directed Toward Female Celebrities

7.1 Goal

Social media posts reflect the emotions, intentions and mental state of the users. Twitter users who harass famous female figures may do so with different intentions and intensities. Recent studies have published datasets focusing on different types of online harassment, vulgar language and emotional intensities [27]. We trained, validated and tested our proposed model, attention-based bidirectional neural network, on the three datasets: “online harassment”, “vulgar language” and “valance” and achieved state of the art performance in all three datasets. We report the F1 score for each dataset separately along with the final precision, recall and macro-averaged F1 score. In addition, we identify ten female figures from different professions and racial backgrounds who have experienced harassment on Twitter.

We tested the trained models on ten collected corpuses each related to one famous female figure to predict the type of harassing language, the type of vulgar language and the degree of intensity of language occurring on their social platforms. Interestingly, the achieved results show different patterns of linguistic use targeting different racial backgrounds and occupations. The contribution of this study is two-fold. From the technical perspective, our proposed methodology is shown to be effective with a good margin in comparison to the previous state-of-the-art results in the three datasets. From the social perspective, we introduce a methodology which can unlock facts about the nature of offensive language targeting women on online social platforms. The collected dataset will be shared publicly for further investigation.

In this chapter, we use Bert, Bidirectional Encoder Representations from Transformers, for getting the word representation embedding. The main reason is that Bert

is a context dependent algorithm and presents different vectors for different senses of each word. Bert uses the bidirectional training of transformer and delivers the state of the art performance in many natural language processing tasks.

7.2 Introduction

Vulgarity is a common linguistic and psychological phenomenon and understanding it is beneficial for both domains. It occurs in five to seven percent of our daily conversation and is even more common on Twitter [56, 81]. Vulgar language usually carries different emotions with different intensities and serves different functions on online social platforms. This type of language can play different roles. For example, it can be used as informal or slang speech, it can be used as part of hateful speech toward users or topics or it can be used as a tool to show the intensity of expressed emotion [11]. This is especially important in hateful speech detection or online harassment detection since explicit vulgar words can have different implicit meanings. Thus, understanding this language is beneficial in the computer science community which can aim to model vulgarity using NLP applications. Accessing a large amount of information on social media makes the analysis of sentiment of the online content possible. Sentiment analysis is one of the common research areas and it is mainly about identifying the sentiment of a sentence either positive or negative. In fact, valence or sentiment represents the polarity of a sentence as being on a range from very negative to very positive. In this research, we refer to valence as the intensity of sentiment and use valence as a tool to detect the mental state of the users who write harassing tweets.

In the literature, the word *vulgarity* is usually replaced by words such as *profanity* or *swear/curse words*. These words usually serve different purposes based on the context in which they occur. Vulgar language is a tool for users to express the feelings and emotions with different intensities [140]. For example, "she is stupid as f*k" has a relatively stronger emotion in comparison to "she is stupid". Even though the word "stupid" is a negative vulgar word and it has different intensities in different contexts. Recently, there have been studies trying to identify the role of vulgar language in hateful speech [76]. In a study, vulgar words are divided into five categories as follows: abusive, cathartic, dysphemistic, emphatic and idiomatic [105].

After that, four main functions of vulgar language were proposed. These functions are as follows: expressing emotion, showing empathy, referring to the identity of a group or showing aggression [141] .

Recently, with the introduction and development of neural networks, there have been great improvements in NLP tasks such as text classification [66], machine translation [16] and sentiment analysis [71]. There are interesting works applying neural networks on sentiment analysis such as Target-Connection LSTM [130], Target-Dependant LSTM [131], attention-based LSTM [142] and bidirectional-based attention networks [48]. A key feature of attention-based neural networks is having an effective mechanism to identify the importance of each word in a sentence.

The goal of this case study is to present a comprehensive and multi-faceted experiment on the types of online harassment, vulgar language and intensity of the tweets happening around famous female figures on social media using a bidirectional attention-based neural network. To this end, the contribution of the chapter is as follows:

- Training and testing bidirectional LSTM with an attention layer on three public datasets: "online harassment" [119], "vulgar language"[11] and "valance"[87] and achieving state-of-the-art results with good margins in all the dataset.
- Using the python Twitter API to collect five hundred tweets directed to each of the ten selected female figures over a period of one month.
- Testing the three pre-trained models on each of the ten datasets to understand the online harassment types, the vulgar language types and the intensity of emotions.
- Demonstrating the classification results of the proposed model directed to famous female figures on ten datasets and addressing the impact of occupation and racial background on the results.

7.3 Data sets

7.3.1 Vulgar Language Types

Cachola and Li [11] labeled sentences including vulgar language into three categories of sentiment: positive, negative and neutral. Later, they used a bidirectional long-short-term memory network (LSTM) to classify their dataset. After that, a new dataset comprised of 7,800 tweets was proposed by Holgate and Li [53]. They focused specifically on different functions of vulgar words. They detected six categories of intention for vulgar language mentioned as follows: “Expressing aggression”, “expressing emotion”, “emphasizing”, “having an auxiliary role”, “signal group identity” and “not-vulgar”. In their research, they try to find a model which can best predict the usage of a vulgar word based on the context in which it occurs. In this study, we focused on this dataset. This dataset is related to detecting six types of vulgar language [53]. Their dataset presents different intentions of vulgar language as the label for the tweets. Table 7.1 shows the distribution and samples of each category.

The first category is expressing anger, in this category, vulgar words are used to hurt the person to whom the tweet is directed. The second category, expressing emotion, as the name reveals is about an explicit emotion in the tweet such as exclamations, internal states or attitudes toward an object. The third category, emphasis, is used to add intensity to a feeling or idea. The fourth category is auxiliary; this category reveals a general emotion or idea. The fifth group is about signalling group identity and tweets in this group show the identity of the user as being part of a group. The last category is related to tweets which are not vulgar and are not in the previous categories.

7.3.2 Intensity Types

Detecting different intensities of sentiment in a text can be considered as a typical text classification task in which texts are the data and their intensity is the label. A new annotated dataset comprised of news articles for detecting different emotions in sentence level was proposed by Strapparava and Mihalcea, (2017). After that, a phrase-based corpus for the emotional intensity of the phrases was introduced by Aman and Szpakowicz, (2007). A new dataset was introduced in SemEval 2018 with

different sub-tasks [87]. One of the sub-tasks is related to the ordinal classification of detecting valence. The task is to classify tweets into seven classes which show the mental state of the user who wrote the tweet. In this task, there are seven classes ranging from 3 to -3 where 3 represents a very positive mental state of the user and -3 shows a very negative mental state of the user. In this study, we focused on this dataset since it has a wide range of intensities [44, 67].

Table 7.2: Description of Intensity Dataset [87]

<i>Number of tweets in the train set</i>	1181
<i>Number of tweets in validation set</i>	449
<i>Number of tweets in test set</i>	937

7.3.3 Analyzing Different Harassment Types

For analyzing different harassment type we use the dataset which we made in the previous chapters. This dataset was proposed by S. Sharifirad and S. Matwin, (2018) and contains three categories of "indirect harassment", "physical harassment" and "sexual harassment". Table 7.3 shows the number of data points and an example of each class. From its definition, indirect harassment is a type of harassment which does not have any slur or swear words directly but indirectly tries to compare women with men unfairly or underestimate the mental or physical capabilities of women. The second category, physical harassment, contains slur words and tries to show females as sex objects or threatens them physically. The third class is sexual harassment.

Table 7.1: Description of Different Vulgar intentions [53]

Category	Number of tweets	Sample of Tweet
<i>Express aggression</i>	1293	"You are an ass Your industry is full of assholes and you do nothing to improve"
<i>Express Emotion</i>	2108	"There are so many things I want to do, But investing in equipment is a pain in the ass"
<i>Emphasis</i>	2561	"today is a good ass day"
<i>Auxiliary</i>	1456	"Wish <USER>could save my ass on these exams like he used to"
<i>Signal Group Identity</i>	399	"Now this is a group of ass kickers!"
<i>Non-vulgar</i>	707	"Kick Ass 2 - Red Band Trailer <URL>"

These tweets contain insulting words, name calling, words of anger, violence and sexual humiliation.

Table 7.3: Description of Different Harassment Types presented in chapter 3 [119].

Category	Number of tweets	Sample of Tweet
<i>Indirect harassment</i>	260	"she plays as good as a boy"
<i>Sexual harassment</i>	417	"Damn girls, you are fine"
<i>Physical harassment</i>	123	"Just putting it out there, you deserve all those deaths you are getting"
<i>Not sexist</i>	2440	"lets go and clebrate the feeling of freedom"

7.4 Data Pre-processing

We used the same dataset which was introduced in chapter 3. We pre-processed the tweets by removing any punctuation, hyperlinks/URLs, emojis and tags. Stop words were removed except the stop word "No" because of its impact on the sentiment of the sentence. We used pre-trained BERT-Base-Cased, Bidirectional Encoder Representations for the word embeddings of all the dataset we used such as "Harassment dataset, proposed in chapter 3", "vulgar language dataset"[53], "Emotion Intensity Dataset"[87]. We consider 25 as the input sentence length, batch-size = 256 and num-workers=4 and use the final embeddings of 768 as output.

This model is trained on BooksCorpus and English Wikipedia, totally having 3300 million words to get the word vectors[29]. BERT is a new language model for extracting features from text without fine tuning it on our datasets. This word vector has the advantage of considering the context of the words. In this study, we used Bert-Base because of its size and used bert-as-service to get the word vectors of length 300 (bert-as-service, 2018). In recent experiments, it has been shown that after fine tuning, BERT has a superior performance in comparison to other word embedding on Twitter data and in the domain of abuse detection [28].

7.5 Model

In text classification, we try to determine the categories or labels of the text based on the provided text content. Recently, Recurrent Neural Network (RNN) have been shown to have a high capability of learning in many natural language processing tasks. Both short and long sentences can be learned using RNNs. Unfortunately, *exploding* or *vanishing gradients* are two main drawbacks of RNNs happening in long sequences in which the gradient vector either grows or decays exponentially. Long short-term memory network (LSTM) was first proposed by Hochreiter and Schmidhuber (1997). This algorithm specifically solves the problem of memorizing long-term dependencies. It has a separated memory cell inside in order to take care of necessary updates.

For many natural language processing tasks, it is useful to have access to the past and future context of the words. LSTM can only consider past information and simply ignores future information. As a solution, Bidirectional LSTMs have been proposed. These networks have two small sub-networks: the left direction is the forward pass and the right direction is the backward pass. In other words, the forward RNN function reads the input sequence respectively and calculates a value for hidden state's units while the backward RNN reads the sequence in reverse and calculates the values of backward hidden states [107].

Attention neural networks were first proposed for machine translation tasks. They have been shown to perform successfully on different tasks such as machine translation, question answering and speech recognition [51, 6, 17, 158]. In this network, attention allows the model to put weight on certain words in a sentence for doing different tasks. In this study, we used bidirectional RNN with attention layers for the classification task.

In this paper, we used bidirectional RNN because of its unique ability to memorize the content of a word both in future and past. We accompanied this model with an attention layer to learn a weight for each word. This weight usually shows the importance of the word and considers each input tweet in the form of x_1, x_2, x_3, x_{t1} and x_t . The bidirectional RNN has both forward and backward directions. In this model, the hidden state for each word such as x_i is the concatenation of the hidden state of forward and backward direction together. In addition, we used an attention layer. We calculated a weight with respect to a word by the attention mechanism

and then weighted the final summation of hidden states of all of the words based on the weights. The result is another textual feature in the calculation [31].

After extracting the textual features, the features were fed into a softmax classifier for classification. We used drop out and consider Adam as the optimization algorithm. In this model, we optimized the parameters of bidirectional RNN such as W and b , weight and bias term, and the parameters of the attention layer such as M , the transformed hidden state. We used a grid search to determine the best parameters for dimensions of hidden layers from $\{500, 1000, 2000\}$, drop out ratio of 0.5, learning rate from $\{1, 0.1, 0.01, 0.001\}$ for each dataset. The model was implemented in Keras inspired by Androidk, (2018).

7.6 Experiment

7.6.1 Analyzing Different Vulgar Language Types

We split each of the three datasets into train, validation and test set. We report the final precision, recall and F1 score and report F1-score for each class separately. Our predictive model F1 score for each of the vulgar classes is as follows: Aggression-85.6, Emotion-82.3, Emphasize-86.1, Signal Group Identity-80.5, Auxiliary-84.7, Not vulgar-88.1. Our model achieves a macro averaged F1 score of 84.5 across the six classes. Table 3 shows the final precision, recall and F1 score over the six classes. Our results in table 7.4 show a big improvement in comparison to the baseline set by Holgat and Li [53].

Table 7.4: Results on Different Vulgar Language Types.

Method	Precision	Recall	F1
<i>Bidirectional attention LSTM</i>	84.1	83.2	84.5
<i>Baseline [53]</i>	68.8	66.4	67.4

7.6.2 Analyzing the Intensity of Tweets

Table 7.5: Results on Intensity of Dataset

Method	Pearson score
<i>Bidirectional attention LSTM</i>	90.2
<i>Baseline cite</i>	83.6

7.6.3 Analyzing the Harassment Tweets

This dataset is quite imbalanced; we first balanced the dataset using Synthetic Minority Oversampling Technique (SMOTE). After fine tuning our model, the F1 score for the three categories is as follows: Indirect harassment-90.8, Physical harassment-92.1, sexual harassment- 93.8 and not-sexist-94.3. The macro averaged F1-score is 92.7 and accuracy of 94.1.

7.7 Data Around Famous Female Figures

In this case study, we consider ten female figures of different racial backgrounds and professions who have experienced online harassment. We select these female figures using Wikipedia and Google search. The main idea was if knowing about race, occupation and other attributes of these female celebrities can help us understand the type of harassment generated toward a specific profile of the harassed woman. We collected five hundred tweets over the duration of one month using the Python Twitter API. These tweets have been released in author GitHub for further investigation. Table 7.6 shows the names and occupations of the selected female figures.

Table 7.6: Description of Ten Famous Female Figures

Name	description	Colour
<i>Brianna Wu</i>	She is an American computer programmer and game developer. In 2014, she posted tweets about Gamergate fans and later received rape threats such as revealing her address [60].	White
<i>Gabrielle Douglas</i>	She is an American gymnast who was champion in 2012 and a silver medalist in 2015 [149].	Black
<i>Julie DiCaro</i>	Julie Dicaro is a sport host and culture writer. She studies journalism and French [110].	White
<i>Kimberle Crenshaw</i>	She is an American civil rights advocate and a famous scholar for critical race theory. She is a full time professor and specializes in race and gender [150].	Black
<i>Sarah Colby Spain</i>	Sarah Colby Spain is a sports reporter. She reports for ESPN and is a columnist in espnw.com [153].	White
<i>Serena Jameka Williams</i>	Serena Jameka Williams is a famous tennis player. She is ranked number one based on The Women Tennis Association(WTA) [154].	Black
<i>Pamela Merritt</i>	Pamela Merritt is an American writer and activist focusing on woman's rights [152].	Black
<i>Mary Beard</i>	Mary Beard is an English scholar and classicist. She is a professor in the University of Cambridge [151].	White
<i>Zelda Rae Williams</i>	Zelda Rae Williams is an actress [155].	White
<i>Ava Marie DuVernay</i>	Ava Marie DuVernay is an American film director, film marketer and producer [148].	Black

After training, validating and testing three separate models on three datasets, we tested these three models on ten collected datasets, each around one female figure. Looking into the table of labels related to famous figures revealed very interesting findings for us.

Generally, black female figures on average receive more aggressive tweets with high negative valence. In addition, they receive more of the tweets which show signal group

identity which means some users write tweets which focused on the black community and try to provoke other members in their group to write toxic comments. In addition, they have more tweets in the category of auxiliary vulgar language in comparison to other racial backgrounds. White female figures also received aggressive and emotional tweets. However, there were a few tweets about signal group identity and auxiliary.

Female athletes irrespective of their racial background receive a high number of physical harassment, rape and death threat tweets. Interestingly enough, they receive indirect harassment as well, which shows there are users who harass female athletes by indirectly questioning their way of playing or their competency in sports in comparison to their male colleagues.

Those female figures who are actresses receive two main types of harassment: physical and sexual. Usually, in social media they do not get compared to their male peers but are harassed, are threatened with death or rape or receive tweets which view them as sex objects.

Those female figures who are activists or are civil rights speakers receive tweets with a lot of aggression, negative emotions and a high valence of physical or sexual harassment. If they receive tweets which were moderately negative, they are tweets which invite them to be silent or condemn their speech but these tweets don't contain any swear words. These findings are not only in line with Delisle et al.,(2018) research but also provides more information on different occupations.

7.8 Conclusion

In this research, we tried to understand the trend of toxic tweets happening in social media around famous female figures in different occupations such as speakers, athletes, actresses, activists and programmers with different racial backgrounds. We used three available datasets of vulgar language, harassment language and valence. We trained, validated and tested our proposed model on these three datasets and reached the state of art performance on all three. Then, we tested the trained models on ten collected datasets around ten famous female figures and reported the results. The outcome and results revealed very interesting patterns. Based on our best knowledge, this is the first study of its kind which considers different occupations and different racial backgrounds. It is worth mentioning that this is a case study and we cannot

generalize this to the general social media. we are hoping this case study can be a good starting point for other research to collect more datasets, for example about one hundred females, and present a more in-depth study. In addition, talking about the racial background of people is a sensitive task, it should be done carefully, respectfully and fair.

Table 7.7: Experimental results of ten collected datasets from famous female figures on three available datasets

	Brianna Wu	Gabrielle Douglas	Julie Di-Caro	Kimberly Crenshaw	Sarah Spain	Serena Williams	Pamela Merritt	Mary Beard	Zelda Williams	Ava DuVernay
<i>Aggression</i>	392	150	377	170	281	201	164	347	247	248
<i>Emotion</i>	68	0	102	2	210	7	0	141	233	0
<i>Emphasize</i>	12	32	0	28	0	120	112	10	12	0
<i>Signal Group Identity</i>	2	278	0	243	0	101	178	0	1	201
<i>Auxiliary</i>	6	40	9	57	5	71	43	2	7	51
<i>Not-Vulgar</i>	20	0	12	0	4	0	0	0	0	0
<i>Indirect Harassment</i>	0	9	0	0	11	77	0	0	30	0
<i>Physical Harassment</i>	310	371	208	310	223	302	323	220	157	7
<i>Sexual Harassment</i>	190	120	292	190	266	121	0	0	0	89
<i>very positive</i>	0	0	0	0	0	0	0	0	0	0
<i>Moderately positive</i>	1	0	0	0	0	0	0	0	0	0
<i>Slightly positive</i>	3	0	0	0	0	0	0	0	0	0
<i>Neutral</i>	7	0	0	0	0	0	0	0	0	0
<i>Slightly Negative</i>	27	16	281	17	43	11	0	3	33	21
<i>Moderately negative</i>	81	83	129	94	323	70	212	334	201	27
<i>Very negative</i>	381	401	90	389	134	419	288	163	266	262

Chapter 8

Conclusion and Future work

8.1 Conclusion

In this thesis, we focused specifically on the problem of online harassment in social media. Even though harassment can be toward both female and male genders, we focused on female genders since it is more common and the manifestations are more elaborate. Inspired from social science work, we came up with initial categories and ran a pilot study to estimate the task hardship and the clarity of instruction. Then, based on the inter-annotate metrics such as Fleiss Kappa score, we combined the first categories into three main categories and ran the pilot study again. After having a good score of agreement between the raters, we used different hashtags to collect more than three thousand tweets and used crowd sourcing platforms to label the tweets.

We formed the problem as a classification task. We used different semantic representation of the word vectors with different classifiers for categorizing different types of harassment and reported the best model. Then, we picked the model and tried to interpret it. Interpreting the model then led us to different clusters of the words being informative and not informative. We then augmented the dataset with the most informative words and repeated the classification. We reported the results on both original and augmented datasets and reported an increase in the performance of the classifier.

We also tried to show the effectual state of users who write abusive language in the form of detecting the intensity of the feelings and the type of feelings in each category of harassment. To improve the classification task, we need to have a large amount of data and an imbalanced dataset. Usually, for the specific datasets, the dataset is scarce and imbalanced. We proposed new methods for text generation and text augmentation and showed an improvement in the classification task. Furthermore, we ran an experiment as a case study on famous female celebrities to identify the effect of race and occupation on the type of abusive language they receive on online

social platforms.

Our experiments suffered from some limitations as well. First, annotating a sensitive content such as harassment is a hard task. Usually raters come from different cultural backgrounds, some know English as their second language and bring their own biases to the dataset. In addition, for some annotation tasks raters require training to better understand the given task. Furthermore, balancing the dataset using oversampling techniques might not be the best way to do it. Also, there exists some vagueness between hateful speech, toxicity and online harassment. As a broader application of this work, can be formulated in order to warn users if their tweets are sexist

8.2 Future work

Different lines of research can be explored in the future.

Considering Emojis as A Feature: Pictographs also commonly known as "emojis" have been very prevalent to improve online communication. After its introduction in the late 1990's, emojis have been widely used to present sentiment, emotion, and sarcasm in the content. In addition to text, they carry important information for users. The capability for processing and capturing the meaning and interpreting emojis co-occurring with text automatically is necessary as online social platforms embrace this type of communication. Considering emojis in the text helps to understand the social, cultural, communicative and linguistic aspect of the contents. In online social platforms, emojis take over emotions on Twitter and happen to be very prevalent in harassment tweets. It would be interesting to analyze the corporation of emojis as an additional feature for text classification or to investigate the level of agreement between the meaning in which the tweets transfer and the emojis in the tweets.

Text Generation Methods Recently, there have been proposed many methods for text generation. Different methods have been used utilizing deep learning techniques. Among them, Generative Adversarial Networks(GAN) [45] has attracted a lot of attention. Originally, GANs were proposed for the task of image generation. After that, William Fedus [37] proposed a method to generate text using GANs. However, all these methods require a large amount of trained data and when the data is scarce,

they don't have the desired performance. In our study, we used knowledge graphs to replace each word with another word based on their specific relations. It is worth mentioning that there could be other methods to try as future work, for example, replacing each word based on the similarity to another word in an embedding space.

Multi-Modal Classification: Users start to share all sorts of information on online social platforms. Twitter is one of those platforms which contains a variety of information. One of the key directions as future research is to combine two sources of information, text and image, to predict the sentiment or improve the classification task.

Alternative Classification Approaches: Combining ConceptNet, Wikidata and Emoji ontology for considering all useful information in the tweets. Replacing the concepts based on their similarity with another word in their embedding space. Working on transfer learning algorithms and zero shot learning along with different character and word ngrams can also be considered.

Transfer learning: One of the approaches which enables us to use knowledge from one task into another task is transfer learning [133]. Usually transferring prior knowledge from one task to another helps us understand more about the target domain. It would be interesting to apply different transfer learning techniques on available toxic language and hateful speech to the task of online harassment and see the differences.

Considering Context of the Tweets: One of the good approaches is to consider context of the harassment tweets and using their history including re-tweeting and chains of discussions. In addition, it might be useful to target some users and collect the tweets from them for further investigation.

Bibliography

- [1] Sweta Agrawal and Amit Awekar. Deep learning for detecting cyberbullying across multiple social media platforms. *Advances in Information Retrieval*, pages 141–153, 2018.
- [2] Amr Ahmed and Eric P Xing. Supervised and semi-supervised multi-view topical analysis of ideological perspective. *In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing.*, pages 1140–1150, 2010.
- [3] Dr. Hussein K. Al-Khafaji and Areej Tarief Habeeb. Efficient algorithms for preprocessing and stemming of tweets in a sentiment analysis system. *Journal of Computer Engineering (IOSR-JCE)*, 2017.
- [4] Sanjeev Arora, Rong Ge, Ravi Kannan, and Ankur Moitra. Computing a non-negative matrix factorization-provably. *Proc. of the 44th Symp*, pages 145–162, 2012.
- [5] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep learning for hate speech detection in tweets. *In Proceedings of the 26th International Conference on World Wide Web Companion.*, pages 759–760, 2017.
- [6] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
- [7] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *ICLR*, 2015.
- [8] Azy Barak. Sexual harassment on the internet. *Social Science Computer Review*, 23(1):77–92, 2005.
- [9] Yoshua Bengio, Rejean Ducharme, and Pascal Vincent. A neural probabilistic language model. *In Advances in Neural Information Processing Systems*, 2001.
- [10] "Lindsay Blackwell, Tianying Chen, Sarita Schoenebeck, and Cliff Lampe". When online harassment is perceived as justified. *Proceedings of the Twelfth International AAAI Conference on Web and Social Media (ICWSM)*, 2018.
- [11] Isabel Cachola, Eric Holgate, Daniel PreotĂbiuc-Pietro Ăband iuc Pietro, and Junyi Jessy L. Expressively vulgar: The socio-dynamics of vulgarity and its effects on sentiment analysis in social media. *COLING*, pages 2927–2938, 2018.
- [12] Rafael A. Calvo and Sidney D’Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. *Affective Computing, IEEE Transactions on*, 1:18–37, 2010.

- [13] Walter B. Cannon. The james-Åillange theory of emotions: a critical examination and an alternation. *Am. J. Psychol.*, 39, 1927.
- [14] Raymon B. Cattell. Sentiment or attitude? the core of a terminology problem in personality research. *Character Personality; A Quarterly for Psychodiagnostic Allied Studies*, 9:6–17, 2006.
- [15] Yizong Cheng. Mean shift, mode seeking, and clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 17(8):790–799, 1995.
- [16] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. page arXiv preprint arXiv:1406.1078, 2014.
- [17] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recognition. *In Advances in Neural Information Processing Systems*, pages 577–585, 2015.
- [18] Isobelle Clarke and Jack Grieve. Dimensions of abusive language on twitter. *In Proceedings of the First Workshop on Abusive Language Online*, pages 1–10, 2017.
- [19] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. *In Proceedings of the International Conference on Machine LearningL*, pages 160–167, 2008.
- [20] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel P. Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537, 2011.
- [21] ConceptNet. <http://conceptnet.io>.
- [22] Maral Dadvar, DolfTrieschnigg, Roeland Ordelman, and Franciska de Jong. Improving cyberbullying detection with user context. *European Conference on Information Retrieval*, 2013.
- [23] Thomas Davidson, Dana Warmusley, Michael W. Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. *In Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017.*, pages 512–515. AAAI Press, 2017.
- [24] dbpedia. <https://wiki.dbpedia.org>.
- [25] Helena de Medeiros Caseli, Bruno Akio Sugiyama, and Junia Coutinho Anacleto. Using common sense to generate culturally contextualized machine translation. *Proceedings of the NAACL HLT*, 57(10):24–31, 2010.

- [26] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391, 1990.
- [27] Laure Delisle, Alfredo Kalaitzis, Krzysztof Majewski, Archy de Berker, Milena Marin, and Julien Cornebise. A large-scale crowdsourced analysis of abuse against women journalists and politicians on twitter. *Workshop on AI for Social Good, 32nd Conference on Neural Information Processing Systems (NIPS 2018)*, 2018.
- [28] Laure Delisle, Alfredo Kalaitzis, Krzysztof Majewski, Archy de Berker, Milena Marin, and Julien Cornebise. Troll patrol methodology note. 2018.
- [29] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2019.
- [30] Cicero Nogueira dos Santos and Bianca Zadrozny. Learning character-level representations for part-of-speech tagging. *In Proceedings of the 31th International Conference on Machine Learning ICML*, pages 1818–1826, 2014.
- [31] Changshun Du and Lei Huang. Text classification research with attention-based recurrent neural networks. *International Journal of Computers Communications Control*, page 50–61, 2018.
- [32] M. Duggan. Online harassment. *Pew Research Center*, 2014.
- [33] Figure Eight. <https://www.figure-eight.com>.
- [34] Paul Ekman. Universal and cultural differences in facial expression of emotion. *Nebraska Symposium on Motivation*. Ed. J. R. Cole. Lincoln: U of Nebraska P, 1972.
- [35] Jeffrey L Elman. Finding structure in time. *Cognitive science*, pages 14(2):179–211, 1990.
- [36] Dorothy L. Espelage and Susan M. Swearer. Research on school bullying and victimization: What have we learned and where do we go from here? *School Psychology Review*, 32(3), 365–383 (2003).
- [37] William Fedus, Ian Goodfellow, and Andrew M. Dai. Maskgan: Better text generation via filling in the—. *ICLR*, 2018.
- [38] Bianca Fileborn. Online harassment of women driven by misogyny, fear and a need for power. <https://www.abc.net.au/news/2018-04-18/why-men-abuse-women-online/9666900>, 2018. [Online: accessed 2016-04-17].

- [39] Global Fund for Women. Online violence: Just because its virtual does not make it any less real. <https://www.globalfundforwomen.org/online-violence-just-because-its-virtual-doesnt-make-it-any-less-real/#.XEtZSq0ZPok>, 2018.
- [40] Mary Anne Franks. Sexual harassment 2.0. *Maryland Law Review*, 71:655–704, 2012.
- [41] freebase. <https://developers.google.com/freebase/>.
- [42] Björn Gambäck and Utpal Kumar Sikdar. Using convolutional neural networks to classify hate-speech. In *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90. Association for Computational Linguistics, 2017.
- [43] Becky Gardiner, Mahana Mansfield, Ian Anderson, Josh Holder, Daan Louter, and Monica Ulmanu. The dark side of Guardian comments. <https://www.theguardian.com/technology/2016/apr/12/the-dark-side-of-guardian-comments>, 2016. [Online: accessed 2016-04-12].
- [44] Pranav Goel, Devang Kulshreshtha, Prayas Jain, and Kaushal Kumar Shukla. An ensemble of deep neural architectures for emotion intensity prediction in tweets. *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA), Copenhagen, Denmark, 8 September*, pages 58–65, 2017.
- [45] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, Sherjil Ozair David Warde-Farley, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
- [46] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Maskgan: Better text generation via filling in the –. *ICLR*, 2018.
- [47] Alex Graves. supervised sequence labelling with recurrent neural networks. *Studies in Computational Intelligence*. Springer, 2012.
- [48] Shuqin Gu, Lipeng Zhang, Yuexian Hou, and Yin Song. A position-aware bidirectional attention network for aspect-level sentiment analysis. pages *Proceedings of the twenty-seventh international conference on computational linguistic*. 774–784, 2018.
- [49] Hatebase.org. . <https://hatebase.org>, 2019.
- [50] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701, 2015.

- [51] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, page 1684–1692, 2015.
- [52] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, page 9(8):1735–1780, 1997.
- [53] Eric Holgate, Isabel Cachola, Daniel Preot, Iñac Pietto, and Junyi Jessy Li. Why swear? analyzing and inferring the intentions of vulgar expressions. *EMNLP*, pages 4405–4414, 2018.
- [54] Huffman. <https://www.huffingtonpost.com.au>, 2018.
- [55] Alon Jacovi, Oren Sar Shalom, and Yoav Goldberg. Understanding convolutional neural networks for text classification. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 56–65. Association for Computational Linguistics, 2018.
- [56] Timothy Jay. The utility and ubiquity of taboo words. *Perspectives on Psychological Science*, 4(2), 2009.
- [57] Al Jazeera. Trolls and threats: Online harassment of female journalists. <https://www.aljazeera.com/programmes/listeningpost/2018/10/trolls-threats-online-harassment-female-journalists-181006101141463.html>, 2018. [Online; accessed 2018-11-12].
- [58] A. Jha and Mamidi R. When does a compliment become sexist: Analysis and classification of ambivalent sexism using twitter data. In *2nd Workshop on NLP and Computational Social Science*, pages 7–16, 2017.
- [59] Akshita Jha and Radhika Mamidi. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 7–16. Association for Computational Linguistics, 2017.
- [60] Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. Exploring the limits of language modeling. *CoRR abs/1602.02410*, 2016.
- [61] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jegou, and T. Mikolov. Fasttext.zip: Compressing text classification models. 2016.
- [62] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, Baltimore, MD, USA, June 22-27, 2014.*, pages 655–665, 2014.

- [63] Yoon Kim. Convolutional neural networks for sentence classification. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1746–1751. ACL, 2014.
- [64] Yoon Kim. Convolutional neural networks for sentence classification. *EMNLP*, 2014.
- [65] M. Kroetsch and G. Weikum. Special issue on knowledge graphs. *Journal of Web Semantics*, <http://www.websemanticsjournal.org/index.php/ps/announcement/view/19>, 2016.
- [66] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. Recurrent convolutional neural networks for text classification. pages In AAAI, volume 333, 2267–2273, 2015.
- [67] Egor Lakomkin, Chandrakant Bothe, and Stefan Wermter. Gradascenatemoi- 2017: Character and word level recurrent neural network models for tweet emotion intensity detection. in *Proc. of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis at the Conference EMNLP*. ACL, pages 169–174, 2017.
- [68] Augustin Lefevre. Dictionary learning methods for single-channel source separations. *Ph.D. thesis, Ecole Nor-male Supérieure de Cachan*, 2012.
- [69] Joel Legrand and Ronan Collobert. Syntactic parsing of morphologically rich languages using deep neural networks. *Technical report, Idiap*, 2015.
- [70] Ruth Lewis, Michael Rowe, and Clare Wiper. Online abuse of feminists as an emerging form of violence against women and girls. *British Journal of Criminology*, pages 1–20, 2016.
- [71] Baiyan Liu, Xiangdong An, and Jimmy Xiangji Huang. Using term location information to enhance probabilistic information retrieval. in international acm sigir conference on research development in information retrieval. page 883–886, 2015.
- [72] Jessica Luther. "rt@shakestweetz: [tw]"id have to rape her with 3 popsicle sticks taped to my flaccid wang. <http://t.co/feyed8jn> mencallmethings". page 7 November, 2011.
- [73] L. A. Mainiero and K. J. Jones. Workplace romance 2.0: Developing a communication ethics model to address potential sexual harassment from inappropriate social media contacts between coworkers. *Journal of Business Ethics*, 114:367–379, 2013.

- [74] L. A. Mainiero and K. J. Jones. Sexual harassment versus workplace romance: Social media spillover and textual harassment in the workplace. *The Academy of Management Perspectives*, 27:187–203, 2013.
- [75] Alessandro Maisto, Serena Pelosi, Simonetta Vietri, and Pierluigi Vitale. Mining offensive language on social media. *Proceedings of CLiC-it 2017 4th Italian Conference on Computational Linguistics*, 2017.
- [76] Shervin Malmasi and Marcos Zampieri. Challenges in discriminating profanity from hate speech. *Journal of Experimental Theoretical Artificial Intelligence*, pages 30(2):187–202, 2018.
- [77] Brian Massumi. Notes on the translation and acknowledgements. In *Gilles Deleuze and Felix Guattari, A Thousand Plateaus*. Minneapolis: U of Minnesota P, 1987.
- [78] Diana Maynard and Adam Funk. Automatic detection of political opinions in tweets. *Proceedings of the 8th international conference on the semantic web, ESWC'11*, pages 88–99, 2011.
- [79] William Mazzarella. Enchantments of modernity empire, nation, globalization, affect: What is it good for? *Taylor and Francis group*, 2012.
- [80] Jessica Megarry. Online incivility or sexual harassment? conceptualising women’s experiences in the digital age. *Women’s Studies International Forum*, 47:46–55, 2014.
- [81] Matthias R Mehl, Simine Vazire, Nairan RamirezEsparza, Richard B Slatcher, and James W Pennebaker. Are women really more talkative than men? *Science*, 317(5834):82–82, 2007.
- [82] R. Mendel, E. Traut-Mattausch, E. Jonas, S. Leucht, J. M. Kane, K. Maino, W. Kissling, and J. Hamann. Confirmation bias: why psychiatrists stick to wrong preliminary diagnoses. *Psychological Medicine (2011)*, 41, 2651–2659., 2011.
- [83] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR.*, 2013.
- [84] Sara Mills. Language and sexism. *New York, NY: Cambridge University Press*, 2008.
- [85] Takeru Miyato, Andrew M Dai, and Ian Goodfellow. Virtual adversarial training for semi-supervised text classification. In *International Conference on Learning Representations*, 1050:25, 2017.

- [86] Saif M. Mohammad and Felipe Bravo-Marquez. Emotion intensities in tweets. *In Proceedings of the sixth joint conference on lexical and computational semantics (*Sem), Vancouver, Canada, 2017.*
- [87] Saif M. Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. SemEval-2018 Task 1: Affect in Tweets. <https://competitions.codalab.org/competitions/17751>, 2018.
- [88] Saif M. Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. Semeval-2018 task 1: Affect in tweets. *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*, pages 1–17, 2018.
- [89] Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. Stance and sentiment in tweets. page Special Section of the ACM Transactions on Internet Technology on Argumentation in Social Media, 2016b.
- [90] David Newman, Chaitanya Chemudugunta, Padhraic Smyth, and Mark Steyvers. Analyzing entities and topics in news articles using statistical topic models. springer. *In Intelligence and Security Informatics, Lecture Notes in Computer Science.*, pages 93–104, 2006.
- [91] Viet-An Nguyen, Jordan L Boyd-Graber, and Philip Resnik. Lexical and hierarchical topic regression. *In Advances in Neural Information Processing Systems*, pages 1106–1114, 2013.
- [92] Keith Oatly. Emotion. *The Blackwell Dictionary of Cognitive Psychology*, Blackwell, Oxford, 1994.
- [93] Meaghan O’Connell. hope you catch a sexually transmitted disease or vagina cancer, cuntwit.mencallmethings. page 7 November, 2011.
- [94] Sattam Almatarneh * OrcID and Pablo GamalloOrcID. Comparing supervised machine learning strategies and linguistic features to search for very negative opinions. *Journal of Information*, 10(1):16, 2019.
- [95] R. O’Shea, K.and Nash. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458.*, 2015.
- [96] Glick P1, Fiske ST, Mladinic A, Saiz JL, Abrams D, Masser B, Adetoun B, Osagie JE, Akande A, Alao A, Brunner A, Willemsen TM, Chipeta K, Dardenne B, Dijksterhuis A, Wigboldus D, Eckes T, Six-Materna I, Exposito F, Moya M, Foddy M, Kim HJ, Lameiras M, Sotelo MJ, Mucchi-Faina A, Romani M, Sakalli N, Udegbe B, Yamamoto M, Ui M, Ferreira MC, and Lopez Lopez W. The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. *Journal of personality and social psychology*, 70(3):491, 1996.

- [97] Pentti Paatero and Unto Tapper. Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5:111–126, 1994.
- [98] Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 115–124, 2005.
- [99] Jo Ho Park and Pascale Fung. One-step and two-step classification for abusive language detection on twitter. In *ALW1: 1st Workshop on Abusive Language Online, Vancouver, Canada, Association for Computational Linguistics.*, 2017.
- [100] Justin W. Patchin and Sameer Hinduja. Bullies move beyond the schoolyard a preliminary look at cyberbullying. *Youth Violence and Juvenile Justice* 4(2):148-169, 2006.
- [101] H. Paulheim. Automatic knowledge graph refinement: A survey of approaches and evaluation methods. *Semant. Web*, 8:489–508, 2015.
- [102] J. Pavlopoulos, P. Malakasiotis, and Androutsopoulos I. Deeper attention to abusive user content moderation. pages In EMNLP, Copenhagen, Denmark, 2017b.
- [103] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. Beyond the turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70:153–163, 2017.
- [104] Julia penelope. Speaking freely: Unlearning the lies of the fathers’ tongues. *New York: Pergamon Press.*, 1990.
- [105] Steven Pinker. The stuff of thought: Language as a window into human nature. *Penguin*, 2007.
- [106] Robert Plutchik. A general psychoevolutionary theory of emotion. *Emotion: Theory Res Exp*, 1:3–33, 1980.
- [107] Jean Pouget-Abadie, Dzmitry Bahdanau, Bart van Merriënboer, Kyunghyun Cho, and Yoshua Bengio. Overcoming the curse of sentence length for neural machine translation using automatic segmentation. In *Proc. of SSSST-8, Doha, Qatar*, 2014.
- [108] Jeffrey Pennington R. and C Manning. Glove: Global vectors for word representation. 2014.
- [109] McKeown K R. Text generation. *Cambridge University Press*, 1985.
- [110] Radio.com. <https://670thescore.radio.com/hosts/julie-dicaro>. 2018.

- [111] Philip Resnik, William Armstrong, Leonardo Claudino, Thang Nguyen, Viet-An Nguyen, and Jordan Boyd-Graber. Beyond lda: Exploring supervised topic modeling for depressionrelated language in twitter. *In Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology (CLPsych)*., 2015.
- [112] Bjorn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. Measuring the reliability of hate speech annotations: the case of the european refugee crisis. pages Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication, Bochumer Linguistische Arbeitsberichte, Bochum, vol. 17, pp. 6–9, 2016.
- [113] Petrovic. S, Osborne. M, and V Lavrenko. The edinburgh twitter corpus. *Proceedings of the NAACL HLT 2010 Workshop.*, 2010.
- [114] Haji Mohammad Saleem, Kelly P Dillon, Susan Benesch, and Derek Ruths. A web of hate: Tackling hateful speech in online social spaces. *In TA-COS*, 2016.
- [115] Philipp Scharpf, Moritz Schubotz, and Bela Gip. Representing mathematical formulae in content mathml using wikidata. *In Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2018.
- [116] Kappa score.
- [117] Sima Sharifirad, Borna Jafarpour, and Stan Matwin. Boosting text classification performance on sexist tweets by text augmentation and text generation using a combination of knowledge graphs. *ALW2, EMNLP*, 2018.
- [118] Sima Sharifirad, Borna Jafarpour, and Stan Matwin. How is your mood when writing sexist tweets? detecting the emotion type and intensity of emotion using natural language processing techniques. *Oceans, KDD*, 2018.
- [119] Sima Sharifirad and Stan Matwin. Different types of sexist language on twitter and the gender footprint. *CiCLing*, 2018.
- [120] Sima Sharifirad and Stan Matwin. Classification of different types of sexist languages on twitter and the gender footprint on each of the classes. *CiCLING*, 2018.
- [121] F. Shaw. The bye felipe campaign and discursive activism in mobile dating apps. *Social Media + Society*, pages 1–10, 2016.
- [122] Eric Shouse. Feeling, emotion, affect. *Media Culture J.*, 8(6):1, 2005.

- [123] Richard Socher, Jeffrey Pennington, Eric Huang, Andrew Y. Ng, and Christopher D. Manning. Semi-supervised recursive autoencoders for predicting sentiment distributions. *In Conference on Empirical Methods in Natural Language Processing*, 2011.
- [124] Robert Speer, Jayant Krishnamurthy, Catherine Havasi, Dustin Smith, Henry Lieberman, and Kenneth Arnold. An interface for targeted collection of common sense knowledge using a mixture model. *Proceedings of the 14th international conference on Intelligent user interfaces.*, pages 137–146, 2009.
- [125] Dale Spender. Man made language. *Pandora Press; 4th Revised edition edition*, 1990.
- [126] Jo Stichbury. WTF is a knowledge graph? <https://hackernoon.com/wtf-is-a-knowledge-graph-a16603a1a25f>, 2017.
- [127] Doreen Raheena Sulleyman. Online abuse: A monster that silences and curtails women’s rights online. <https://www.apc.org/en/blog/online-abuse-monster-silences-and-curtails-womens-rights-online>, 2018. [Online; accessed 2018-11-20].
- [128] Ilya Sutskever, James Martens, and Geoffrey E Hinton. Generating text with recurrent neural networks. *In Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 1017–1024, 2011.
- [129] Ilya Sutskever, Oriol Vinyals, and Quoc VV Le. Sequence to sequence learning with neural networks. *In Advances in neural information processing systems*, pages 3104–3112, 2014.
- [130] Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. Effective lstms for target-dependent sentiment classification. page arXiv preprint arXiv:1512.01100, 2015.
- [131] Duyu Tang, Bing Qin, and Ting Liu. Aspect level sentiment classification with deep memory network. page arXiv preprint arXiv:1605.08900, 2016.
- [132] Nitasha Tiku and Casey Newton. NTwitter CEO: ”We suck at dealing with abuse”. <https://www.theverge.com/2015/2/4/7982099/twitter-ceo-sent-memo-taking-personal-responsibility-for-the>, 2015.
- [133] Lisa Torrey and Jude Shavlik. Transfer learning. *Handbook of Research on Machine Learning Applications, published by IGI Global, edited by E. Soria, J. Martin, R. Magdalena, M. Martinez and A. Serrano*, 2008.
- [134] B. Venkatesh and J. Anuradha. A review of feature selection and its methods. *CYBERNETICS AND INFORMATION TECHNOLOGIES ÔÇØVolume19, No1*, 2019.

- [135] Laura Vitis and Fairleigh Gilmour. Dick pics on blast: A woman’s resistance to online sexual harassment using humour, art and instagram. *Crime, Media, Culture: An International Journal*, 13(3):335–355, 2017.
- [136] E Voorhees. Query expansion using lexical-semantic relations. In *Proceedings of A CM SIGIR International Conference on Research and Development in Information Retrieval*, pages 61–69, 1994.
- [137] D. Vrandecic and M. Kzotzs. Wikidata. *Communications of 24 the A*, 57(10):78–85, 2014.
- [138] Denny Vrandecic. Wikidata: a new platform for collaborative data collection. In *Alain Mille, Fabien Gandon, Jacques Misselis, Michael Rabinovich, and Steffen Staab, editors, Proceedings of the 21st World Wide Web Conference, WWW*, pages 16–20, 2012.
- [139] Hanna M Wallach. Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd international conference on Machine learning. ACM.*, pages 977–984, 2006.
- [140] F. Wang, T. Li, X. Wang, S. Zhu, and C. Ding. Community discovery using nonnegative matrix factorization. *Data Min. Knowl. Disc.a*, 22(3):493–521, 2011.
- [141] Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P Sheth. Cursing in english on twitter. In *Proceedings of the 17th ACM conference on Computer Supported Cooperative Work Social Computing, CSCW*, pages 415–425, 2014.
- [142] Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. Attention-based lstm for aspect-level sentiment classification. pages In EMNLP,606–615, 2016.
- [143] Charlie Warzel. ”A Honey-pot For Assholes”: Inside Twitter’s 10-Year Failure To Stop Harassment. <https://www.buzzfeed.com/charliewarzel/a-honey-pot-for-assholes-inside-twit-ters-10-year-failure-to-s/>, 2016. [Online; accessed 2018-01-10].
- [144] Zeerak Waseem. <https://sites.google.com/view/alw3/resources/stackoverflow-dataset>. 2019.
- [145] Zeerak Waseem, Thomas Davidson, Dana Warmusley, and Ingmar Weber. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the ACL-Workshop on Abusive Language Online. Media, ACL, 2017, vancouver, BC, Canada.*, pages 78–84, 2017.
- [146] Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. volume 1, pages 88–93. San Diego, California, USA, June. Association for Computational Linguistics., 2016.

- [147] Wikidata. https://www.wikidata.org/wiki/Wikidata:Main_Page.
- [148] Wikipedia. <https://en.wikipedia.org/wiki/ava-duvernay>. 2018.
- [149] Wikipedia. <https://en.wikipedia.org/wiki/gabby-douglas>. 2018.
- [150] Wikipedia. <https://en.wikipedia.org/wiki/kimberle-williams-crenshaw>. 2018.
- [151] Wikipedia. <https://en.wikipedia.org/wiki/mary-beard>. 2018.
- [152] Wikipedia. <https://en.wikipedia.org/wiki/pamela-merritt>. 2018.
- [153] Wikipedia. <https://en.wikipedia.org/wiki/sarah-spain>. 2018.
- [154] Wikipedia. <https://en.wikipedia.org/wiki/serena-williams>. 2018.
- [155] Wikipedia. <https://en.wikipedia.org/wiki/zelda-williams>. 2018.
- [156] Women’s Media Center (WMC). Online Abuse 101. <http://www.womensmediacenter.com/speech-project/online-abuse-101/>, 2019.
- [157] Zhang X., Zhao J., and LeCun Y. Character-level convolutional networks for text classification. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2015.
- [158] Q. Xu, Z. Qin, and T Wan. Generative cooperative net for image generation and data augmentation. pages 1037–1043, 2017.
- [159] Neil Yager. Neural text generation: How to generate text using conditional language models. <https://medium.com/phrassee/neural-text-generation-generating-text-using-conditional-language-models>, 2018.
- [160] Xuchen Yao and Benjamin Van Durme. Information extraction over structured data: Question answering with freebase. *In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, Maryland, USA*, pages 956–966, 2014.
- [161] Wen-tau Yih, Kristina Toutanova, John C. Platt, and Christopher Meek. Learning discriminative projections for text similarity measures. *In Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 247–256. Association for Computational Linguistics, 2011.
- [162] Dawei Yin, Zhenzhen Xue, Liangjie Hong, Brian D. Davison, April Kostathis, and Lynne Edwards. Detection of harassment on web 2.0. *Content Analysis in the Web 2.0 Workshop*, 2009.
- [163] Lei Zhang. Knowledge graph theory and structural parsing. *Twente University Press*, 2002.

- [164] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657, 2015.
- [165] Ziqi Zhang, David Robinson, and Jonathan Tepper. Detecting hate speech on twitter using a convolution- gru based deep neural network. In *Lecture Notes in Computer Science. Springer Verlag*, 2018.
- [166] Jie Zhou and Wei Xu. End-to-end learning of semantic role labeling using recurrent neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 1127–1137, 2015.