

**UNDERSTANDING AND DEFENDING AGAINST CYBERHARASSMENT
IN THE ERA OF MACHINE LEARNING**

by

Nishant Vishwamitra

June 1, 2022

A dissertation submitted to the
Faculty of the Graduate School of
the University at Buffalo, The State University of New York
in partial fulfilment of the requirements for the
degree of

Doctor of Philosophy

Department of Computer Science and Engineering

Copyright by
Nishant Vishwamitra
2022
All Rights Reserved

Acknowledgments

I would like to express my deepest gratitude to Dr. Hongxin Hu, my mentor, for his unlimited guidance, care, and support throughout my Ph.D. studies at the University at Buffalo. Dr. Hu has not only shown me how to conduct groundbreaking research, but also how to be a productive member of the research community. I wish to thank my dissertation committee members Dr. Hongxin Hu, Dr. Ziming Zhao, and Dr. Weihang Wang for their invaluable input and support. I wish to thank Dr. Daniel Pienta from Baylor University and Dr. Jason Thatcher from Temple University for their immense support and guidance during my Ph.D. career. My thanks also extend to Dr. Long Cheng and Dr. Feng Luo from Clemson University, for their generous support and encouragement during my Ph.D. journey. I am also grateful to all my labmates from CactiLab for their help on my research, especially, Keyan Guo who provided me with useful suggestions and insights on the approach and evaluation for online hate detection. Lastly, I would like to thank my parents for supporting me throughout my life.

Table of Contents

Acknowledgments	iii
List of Tables	ix
List of Figures	xi
Abstract	xiv
Chapter 1	
Introduction	1
1.1 Problem Motivation	1
1.2 Research Motivation	2
1.3 Research Objective	5
1.4 Overview of Research Tasks	6
1.5 Task1: Enable detection of visual cyberbullying in images (RQ1 and RQ2)	7
1.6 Task2: Enable detection, explanation and understanding of on- line hate (RQ3 and RQ4)	8
1.7 Task3: Enable robustness studies of multimodal models (RQ5) . .	10

Chapter 2

Background and Related Work	11
2.1 Researches on Cyberharassment Detection	11
2.1.1 Cyberbullying Detection	11
2.1.2 Online Hate Speech Understanding, Explanation and De- tection	13
2.2 Researches on Robustness of ML Models	15
2.2.1 Multimodal Learning	15
2.2.2 Unimodal Adversarial Attacks	16

Chapter 3

Understanding and Detecting Cyberbullying in Images	18
3.1 Threat Model and Scope	25
3.2 Cyberbullying Images Data Collection	26
3.2.1 Methodology	26
3.2.2 Cyberbullying Keywords Extraction	27
3.2.3 Data Collection and Annotation	28
3.2.4 Image Annotation	29
3.3 Motivation and Observation	30
3.4 Our Approach	33
3.4.1 Approach Overview	34
3.4.2 Factor Identification	35
3.4.3 Factor Extraction	39
3.4.4 Measurement of Machine Learning Models for Classifica- tion of Cyberbullying in Images	44
3.5 Implementation and Evaluation	47

3.5.1	Implementation	48
3.5.2	System Effectiveness Evaluation	48
3.5.3	Deployment and User-based Evaluation	57
3.6	Conclusion	66
Chapter 4		
	Detecting and Explaining Traditional Online Hate Speech	67
4.1	Data Collection Methodology	71
4.2	Background	73
4.3	Study Methodology	74
4.3.1	Keywords Discovery from BERT Attention Mechanism	75
4.4	Implementation and Evaluation	79
4.4.1	Implementation	79
4.4.2	Hate Speech Control with BERT Attention	80
4.4.3	Is BERT Detecting Hate Speech based on Existing Definitions of Hate?	81
4.5	Conclusion	84
Chapter 5		
	Towards Understanding and Mitigating New Waves of Online Hate	85
5.1	Examining New Waves of Online Hate	90
5.1.1	Data Collection	90
5.1.2	Hateful twitter users dataset	91
5.1.3	COVID-19-related tweets and memes dataset.	91
5.2	New Waves of Online Hate Contexts During the COVID-19 Pandemic	94
5.3	Different Representations of New Waves of Online Hate	96

5.4	Effectiveness of Existing Techniques Against New Waves of On-line Hate	97
5.5	Our Approach	99
5.5.1	Approach Overview	100
5.5.2	Data Collection	101
5.5.3	Online Hate Attributes	101
5.5.4	AZL: Attribute-based Zero Shot Classification	104
5.6	Implementation and Evaluation	106
5.6.1	Implementation	106
5.6.2	Effectiveness of AZL	106
5.7	Conclusion	107

Chapter 6

	Robustness of Cyberharassment Detection Models	108
6.1	Background	111
6.1.1	Multimodal Learning	112
6.1.2	Unimodal Adversarial Attacks	113
6.2	Threat Model	113
6.3	Our Approach	114
6.3.1	Unified View of Deep Multimodal Models	114
6.3.2	MUROAN Framework	116
6.4	Implementation and Evaluation	119
6.4.1	Implementation	120
6.4.2	Baselines	120
6.4.3	Effectiveness Evaluation	123
6.4.4	Qualitative and Quantitative Analysis	127

6.4.5	Computation Cost of MUROAN	133
6.5	Conclusion	134
Chapter 7		
	Discussion	135
7.1	Understanding and Detecting Cyberbullying in Images	135
7.2	Detecting and Explaining Traditional Online Hate Speech	139
7.3	Towards Understanding and Mitigating New Waves of Online Hate	140
7.4	Robustness of Cyberharassment Detection Models	141
Chapter 8		
	Conclusion	142
	Bibliography	145

List of Tables

3.1	Samples of cyberbullying stories and the extracted keywords. . .	27
3.2	Precision and recall of popular offensive image detectors.	30
3.3	Detection scores of existing detectors on image samples in Figure 3.3.	31
3.4	Capabilities of existing detectors and their limitations.	32
3.5	Analysis of cyberbullying factors. Higher value of cosine similarity indicates higher correlation.	35
3.6	Analysis of correlation of person with threatening object and gesture.	38
3.7	Frequencies of factors responsible for labeling an image as cyberbullying or non-cyberbullying.	49
3.8	Correlation coefficient (Spearman ρ) between visual factors and cyberbullying label. The coefficients are significant at $p < 0.001$ level.	50
3.9	Accuracy, precision and recall of classifier models.	53
3.10	Capabilities of state-of-the-art offensive image detectors with respect to cyberbullying factors.	63
3.11	Keywords used to identify factors.	64

4.1	Summarized list of sample keywords in the datasets, most attended to by BERT model.	76
4.2	Top-k (k = 10) keywords attended to in each layer of BERT model.	77
4.3	Samples of control strategy.	78
4.4	Samples of words chosen as targets. Username identifiers have been removed to preserve user identities.	82
5.1	Detection capability of existing systems and pre-trained models on evolving hate.	99
5.2	New waves of hate types.	101
5.3	Converting labels to hypothesis.	105
5.4	Evaluation of AZL model on new waves of online hate.	106
6.1	Comparison of average percentage points affected by MUROAN and CW attack.	124
6.2	Comparison of Attack Success Rate (ASR).	125

List of Figures

1.1	This dissertation consists of three major pieces of work, addressing cyberharassment by enabling cyberbullying defense using AI/ML, addressing cyberharassment by enabling online hate defense using AI/ML, and addressing robustness of AI/ML employed in cyberharassment defense. The black blocks indicate the components that will be contributed by this dissertation	6
3.1	Cyberbullying in text v.s. cyberbullying in an image. (a) shows a tweet with demeaning words and phrases. (b) shows an image of a person showing a ‘loser’ hand gesture.	19
3.2	Image samples that did not have any Regions of Interest (ROIs).	28
3.3	Image context in cyberbullying images.	31
3.4	Approach overview.	33
3.5	Cyberbullying Vs. non-cyberbullying body-pose.	41
3.6	Some hand gestures found in cyberbullying images in our dataset.	42
3.7	Multimodal model used in our approach.	46
3.8	Scree plot showing proportions of variance and cumulative proportion of variance explained by each component.	51
3.9	Factor loadings of the features across two extracted factors.	52

3.10	ROC analysis of classifier models.	53
3.11	Precision-recall graph of the multimodal model.	53
3.12	Overhead evaluation of the multimodal model integrated into an Android application.	57
3.13	Image samples from the ASL dataset.	58
3.14	User study interface: participants are provided with a free text box to enter factors on their own.	65
3.15	Interface of image annotation task.	65
4.1	Percentages of tweets collected according to date ranges. All date ranges belong to the year 2020.	73
4.2	Attention distance in the two COVID-29 datasets.	79
4.3	Attentions to Target words Vs. Non-target words in case of Asian- hate.	81
4.4	Attentions to Target words Vs. Non-target words in case of Boomer- hate.	81
5.1	New waves of online hate context.	94
5.2	Representation of new waves of online hate.	98
5.3	AZL overview.	101
6.1	By decoupling the input modalities through the removal of a few datapoints in the image via MUROAN framework, the multi- modal model predicts a wrong answer class: <i>Nothing</i> , indicating that decoupling attack can easily compromise multimodal models.	109
6.2	Overview of our approach.	114
6.3	CDF of percentage of datapoints changed.	124

6.4	Three samples depict three types of minimum coupled datapoints in the VQA and Hateful Memes dataset. In sample (a), the minimum coupled datapoints are in the image only (indicated by red circles), and it is enough to only make changes to a those datapoints to decouple the sample. In sample (b), the minimum coupled datapoints are in the text only (indicated by red font), it is enough to make changes to the text only to decouple the sample. In sample (c), the coupled datapoints consist of both image and text, therefore both need to be changed to decouple this sample. .	127
6.5	Additional Samples from the Hateful Memes dataset.	129
6.6	Additional Samples from the VQA dataset.	130
6.7	Samples from the hateful memes dataset depicting random noise-based and gradient-based perturbation types.	131
6.8	CDF of robustness of Late Fusion model and MMBT against MUROAN decoupling attack algorithm.	131
6.9	Computation cost of MUROAN	133

Abstract

With the rise of social media, cyberharassment (e.g., cyberbullying and cyberhate) has become more prevalent in daily interactions and has thus been identified as a critical social problem. It often involves inappropriate online behavior and deliberate cyber threats against individuals, or specific social groups on the grounds of characteristics such as race, sexual orientation, gender, or religious affiliation. For instance, the Cyberbullying Research Center reported that 37% of middle and high school students have been cyberbullied during their lifetime, and this number is expected to further increase as teens continue to have an increased online presence. In a recent Pew survey [1], roughly four in ten (i.e., 41%) Americans reported personally experiencing varying degrees of harassment and bullying online, and Internet users all over world (i.e., 48%) have also reported having similar experiences [2, 3]. Furthermore, the new waves of anti-Asian hate [4, 5], mask-related hate [6, 7] and vaccine-related hate [8, 9] set-off by the COVID-19 pandemic have had a devastating effect on our society globally.

Although Machine Learning (ML) has immense potential for automatic cyberharassment detection, and researchers are increasingly using ML techniques to address this important social problem, they face key challenges to effectively address cyberharassment using ML. We identify three pertinent challenges in

defending against cyberharassment in the era of ML. *First*, the representation of cyberharassment has shifted from traditional text-based cyberharassment to multimodal (i.e., both texts and images) cyberharassment, which poses new challenges to effective cyberharassment detection. *Second*, cyberharassment is a fast-evolving phenomenon, whether fueled by current events or by people looking for new ways to evade cyberharassment detection systems with techniques such as adversarial examples against ML-based models. *Third*, in spite of the recent advances in ML in cyberharassment detection, the robustness of these ML models in an adversarial setting remains poorly understood.

In this dissertation, we focus on understanding and defending against cyberharassment using ML techniques, and studying the robustness of ML models employed in cyberharassment defense. We are interested in (1) understanding and defending against visual cyberbullying (i.e. cyberbullying via images), (2) understanding, detecting and explaining online hateful content, and (3) studying the robustness of ML models employed in cyberharassment defense. In the area of visual cyberbullying defense, we discover five visual factors of cyberbullying in images, and develop a multimodal deep learning model that detects cyberbullying content in images based on those factors. In online hateful content understanding, detection and explanation, we carry out an in-depth analysis of COVID-19-related hateful tweets, and discovered new keywords used to disseminate COVID-19-related hate speech using BERT attention mechanism. Additionally, we carry out further studies to understand the nature of online hate, and discover that new waves of online hate, such as hate related to masks and vaccines witnessed during the COVID-19 pandemic are a core issue limiting the effectiveness of traditional detectors in practical applications, and propose a new ML methodology based on attribute-based Zero Shot Learning to effec-

tively address new waves of hateful content. Lastly, in the area of ML model robustness, we study the robustness of multimodal models against adversarial attacks using a novel attack algorithm based on decoupling input modalities.

Introduction

1.1 Problem Motivation

The social and economic destabilization caused by global events, such as the COVID-19 pandemic has produced a range of emotions in people, including fear, anxiety, and even hostility. Notably, instances of cyberharassment are increasingly occurring on social media that target people based on race/ethnicity, age, social class, immigration status and political ideology. For instance, Asian Americans are frequent targets of cyberharassment related to COVID-19, with derogatory terms for the disease, such as “kung flu” and “chop fluey”, shared more than 10,000 times on Twitter during March alone [10]. Meanwhile, the phrase “Boomer Remover”, a callous nickname for COVID-19 used to mock the high mortality rate among older people infected with the disease, has been shared more than 65,000 times on Twitter [11]. Moreover, a recent report on online toxicity found a 900% increase in cyberharassment towards China and Chinese people on Twitter [12], and traffic to sites and posts that target Asians over COVID-19 has skyrocketed. *Cyberharassment* thus has emerged as a critical

cybersecurity issue [2].

The techniques to address cyberharassment should be accompanied with a strong mitigation strategy so that Internet users can be deterred from posting such content online. Artificial Intelligence/Machine Learning (AI/ML) is an effective means to deter the spread of such content by automatically detecting and removing such content. Additionally, user warnings and word removal recommendations [13, 14] are also often used to obfuscate or redact such content. Thus AI/ML provides an effective means to detect cyberharassment content online, and then obfuscate or redact such content so that exposure to such content can be prevented. Although ML brings significant benefits to systematically detect cyberharassment and also point out for effective control, current ML models have several limitations that limit its applicability to address cyberharassment. The newer representations of cyberharassment consisting of multiple (a.k.a multimodal) modalities, occurrence of new waves of online hate and lack of robustness studies of ML-based multimodal cyberharassment detection methods greatly hinder the applicability of ML in cyberharassment detection.

1.2 Research Motivation

Recent reports of cyberharassment related to Asian-Americans [4, 5], mask [6, 7] and vaccine [8, 9] set-off by the COVID-19 pandemic have had a devastating effect on our society globally. As our cyberspaces move into the future consisting of advanced technologies such as Web 3.0 [15], augmented reality [16] and the Metaverse [17], cyberharassment is bound to take on new, more sinister shapes. Thus, efforts to effectively counter such new eruptions in cyberharassment must be taken immediately. Furthermore, the changing technology landscape has not

only changed the way users access OSNs, but has also changed how perpetrators express online hate. For example, in recent times, more and more online social networks (OSN) users are using images and videos to propagate hate speech that was traditionally in the textual format [18, 19]. As our cyberspaces move into the future consisting of advanced technologies such as Web 3.0 [15], augmented reality [16] and the Metaverse [17], new ways of expressing cyberharassment are bound to take on new, more sinister shapes. As an instance, although the recent wave of COVID-19-related cyberharassment has engendered studies from various domains [20, 21, 22, 23], these studies focused on the spread of COVID-19-related cyberharassment through the medium of textual data, such as tweets. Memes, a relatively new phenomenon, have emerged as a new form of expression beyond text. Memes consist of images with superimposed text, that deliver a particular message when considered in the context of both the image and text content together [24, 25]. Traditionally used as devices to induce humor, memes have recently taken a more negative turn, by being used as mediums of spreading online hate speech [26, 24]. For example, memes that portray Asian people eating dog meat [27], racist memes targeting Chinese eating habits [28], and the morbid meme “Boomer-remover” [29] against the so-called Boomer generation have been widely circulated. Since the context of memes is framed by both image and text, the hate speech propagated in memes is significantly different from text-only hate speech. While recent studies have provided many interesting insights into the nature of hate speech in textual data during COVID-19 pandemic, the role of memes in the propagation of hate speech during COVID-19 pandemic has been largely overlooked. Thus, measures to address such *multimodal* forms of cyberharassment thus need to be taken urgently.

Online cyberharassment is not a static problem. It is highly influenced by global events and the changing technological landscape. For example, recent polarizing events such as the COVID-19 pandemic [4], the 2020 presidential elections [30] and the Black Lives Matter (BLM) [31] protests have shown how emotions of fear, uncertainty, and anxiety involved in these episodes can set-off new spikes in unprecedented cyberharassment [32]. As an instance, the new waves of anti-Asian hate [4, 5], mask-related hate [6, 7] and vaccine-related hate [8, 9] set-off by the COVID-19 pandemic have had a devastating effect on our society globally. Thus, efforts to effectively counter such *new eruptions* in cyberharassment must be taken immediately.

The enormous eruptions of new waves of cyberharassment and their increasingly complex landscapes have unfortunately not induced a corresponding improvement in their detection capability, and existing online hate detection systems have consistently lagged behind in flagging down new cyberharassment content. For example, the recent waves of anti-Asian hate [4, 5], mask-related hate [6, 7] and vaccine-related hate [8, 9] encountered during the COVID-19 pandemic could not be sufficiently contained by online hate moderation tools deployed in online social networks (OSNs), as a result of which cyberharassment against minority communities and other vulnerable groups spread unabated during this period. While these same detection system seemed quite effective in controlling traditional cyberharassment such as violent extremism [33, 34] and trolling [35], they were found struggling to stop the recent, new waves of cyberharassment [36].

Adversarial attacks have been known to successfully fool AI/ML models [37, 38]. Recently, several unimodal adversarial attacks for deep unimodal models have been formulated to study their robustness. For example, unimodal ad-

versarial images [37, 39, 40, 38] and unimodal adversarial text [41] have been widely studied, which have exposed numerous vulnerabilities in deep unimodal models. However, these attacks cannot be directly employed to study the robustness of their deep multimodal counterparts. First, since these attacks can only be applied to single modalities, they do not affect the fusion mechanism that is fundamental to Deep Multimodal Models (DMMs). Second, since DMMs combine several different types of modalities (e.g., image, text, speech, etc.), a single unimodal attack cannot be used for all those modalities. We note that formulating comprehensive methods to study the *robustness* of DMMs is of utmost importance to adopting them in real-world systems.

1.3 Research Objective

In this dissertation, our research goal is to study and address visual cyberbullying using multimodal ML, detect and explain traditional online hate using ML, as well as understand and detection new waves of online hate, and study the robustness of multimodal models in adversarial settings. Through the research, we will answer the following research questions.

- Enable detection of visual cyberbullying in images
 - RQ1: What are the visual factors of cyberbullying in images that can be used to detect this problem?
 - RQ2: How can ML be used to detect cyberbullying in images using those factors?
- Enable detection, explanation and understanding of online hate

- RQ3: How can we use ML to enable the detection and explanation of traditional online hate?
- RQ4: Can we build ML techniques that can detect new waves of online hate?
- Enable robustness studies of multimodal models
 - RQ5: Can we study the robustness of multimodal models in adversarial settings?

1.4 Overview of Research Tasks

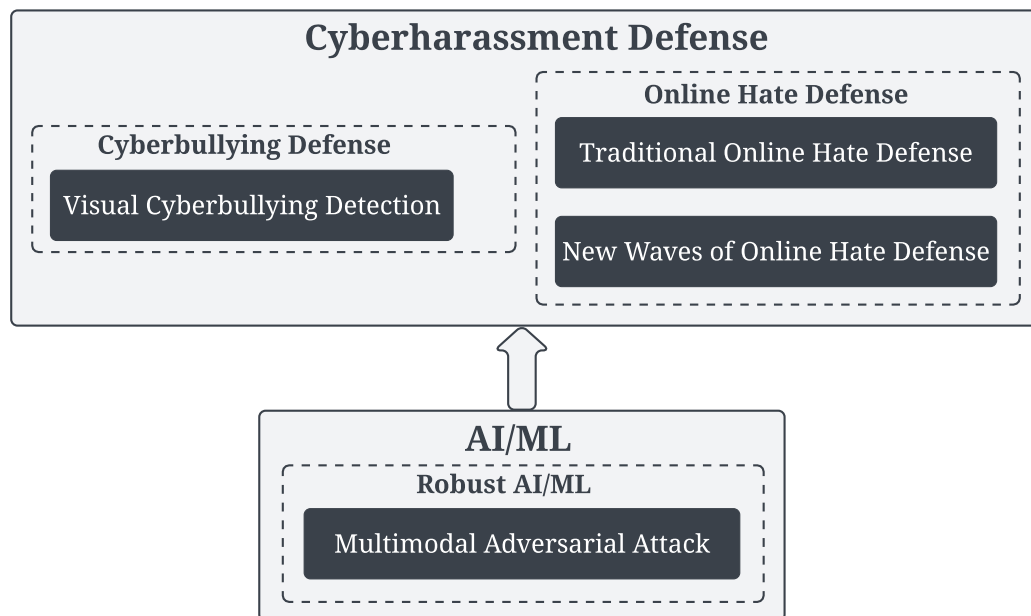


Figure 1.1: This dissertation consists of three major pieces of work, addressing cyberharassment by enabling cyberbullying defense using AI/ML, addressing cyberharassment by enabling online hate defense using AI/ML, and addressing robustness of AI/ML employed in cyberharassment defense. The black blocks indicate the components that will be contributed by this dissertation

To achieve the research goal, we propose to support the understanding and defense of cyberharassment using AI/ML techniques, and support the study of robustness of AI/ML techniques employed in cyberharassment defense, as shown in Figure 1.1.

1.5 Task1: Enable detection of visual cyberbullying in images (RQ1 and RQ2)

We first collect a large dataset of cyberbullying images labeled by online participants. We analyze the cyberbullying images in our dataset against five state-of-the-art offensive image detectors, Google Cloud Vision API, Yahoo Open NSFW [42], Clarifai NSFW, DeepAI Content Moderation API [43], and Amazon Rekognition ¹. We find that 39.32% of the cyberbullying samples can circumvent *all* of these existing detectors. Then, we study the cyberbullying images in our dataset to determine the visual factors that are associated with such images. Our study shows that cyberbullying in images is with highly contextual nature unlike traditional offensive image content (e.g., violence and nudity). We find that cyberbullying in images can be characterized by five important, high-level contextual visual factors: *body-pose*, *facial emotion*, *object*, *gesture*, and *social factors*. We then measure four classifier models (*baseline*, *factors-only*, *fine-tuned pre-trained*, and *multimodal* classifier models) to identify cyberbullying in images based on deep-learning techniques that use visual cyberbullying factors outlined by our study. Based on the identified factors, the best classifier model

¹The offensive image detectors have been selected based on their ability to detect images with certain features, such as violence, profanity, and hate symbols, which have been found in cyberbullying images.

(*multimodal* classifier model) can achieve a detection accuracy of 93.36% in classifying cyberbullying images.

1.6 Task2: Enable detection, explanation and understanding of online hate (RQ3 and RQ4)

We address Traditional online hate, wherein we propose a novel approach to discover new keywords linked to COVID-19-related hate speech and the word associations to effectively implement its control. We collect a new dataset (Boomer-hate dataset) of tweets targeting old people and supplement this dataset with an existing COVID-19 dataset (Asian-hate dataset) targeting Asian American community [21]. We then train a BERT (Bidirectional Encoder Representations from Transformers) model [44] to classify tweets as Hate Vs. Non-hate. Based on the analysis of BERT attention mechanism, a transformer model [45] based on attention, we develop an approach to discover new keywords (186 keywords targeting the Asian community and 100 keywords targeting older people) related to COVID-19. For implementing effective control, we develop a strategy based on the attention attributed to these keywords by other words in a tweet, so that all sensitive words in a tweet can be censored or reconsidered. We then undertake an exploratory analysis of COVID-19-related hate speech and find that most of such high-impact, long distance attentions are learned in the earlier layers of the BERT model (layers 2 to 7 for Asian-hate dataset) or later layers (layers 10 and 11 for Boomer-hate dataset) depending on the underlying data distribution. Our study also makes an important finding that in the case of Boomer-hate dataset, the BERT model makes predictions based on the associa-

tion of hate keywords and targeted groups or individuals, a finding that is inline with existing hate-speech research. Informed by the previous research results, as part of future work I plan to study New waves of online hate defense, and Multimodal Adversarial Attack.

We address the problem of new waves of online hate, by studying it, understanding its challenges, and formulating automatic systems that can detect it. Our intuition, informed by previous studies [5, 7, 8] and reports [46, 2], is that new waves of online hate are characterized by rapidly changed contexts. We first report a systematic study on the phenomenon of new online hate waves, by collecting a large dataset of 3312 hateful users and their 4042454 tweets on Twitter, and studying their tweeting behavior before and after the COVID-19 pandemic. We find that before the pandemic, the tweeting behavior of these hateful users were related to traditional hate contexts, which completely changed into online hate related to new contexts post pandemic. Next, we conducted a large scale study of the effectiveness of state-of-the-art, existing systems of hateful content detection such as Perspective API [47], Google Cloud Vision API [48] and MMBT [49] on datasets of COVID-19-related 1,679 tweets, and found that these detectors are severely limited (average F1 score of 0.31) against new waves of hate tweets. We then identify key challenges to the timely and effective intervention of new waves of online hate: (i) learn knowledge from traditional hate contexts and apply learned knowledge to new contexts, (ii) training with just a few samples of new hate contexts.

We introduce our framework, Attribute-based Zero-shot Learning (AZL), that can detect new waves of online hate by addressing each of those challenges. AZL uses an attribute-based learning methodology [50] to transfer important knowledge about traditional hate contexts to the detection of new hate contexts,

and uses Zero-shot learning [51] to effectively classify new hateful contexts with just a few training samples. We evaluate AZL from several different perspectives, and find that our framework achieves state-of-the-art-detection average F1 score of 0.72 on new hate contexts, such as Asian (76.52%), mask (67.47%), vaccine (70.73%) and boomer (72.34%) related hate.

1.7 Task3: Enable robustness studies of multimodal models (RQ5)

In this task, we first highlight how multimodal adversarial attacks based on decoupling the input modalities in DMMs can easily compromise these models. Then, we introduce a framework called MUROAN to study the robustness of DMMs based on decoupling of modalities, thereby revealing vulnerabilities in the fusion mechanism of existing DMMs. MUROAN uses a unified view of DMMs to expose its key vulnerability. Then, we introduce a new type of adversarial attack called decoupling attack in MUROAN, wherein the objective of its attack algorithm is to decouple the input modalities of multimodal models to induce a misclassification. As depicted in Figure 6.1, a decoupling of the image and text modalities through occlusion of a few datapoints in the image induces a misclassification. In addition, we leverage the MUROAN framework to measure several state-of-the-art DMMs. We find that the seemingly straightforward decoupling attack of MUROAN is in fact highly effective in compromising DMMs.

Background and Related Work

2.1 Researches on Cyberharassment Detection

2.1.1 Cyberbullying Detection

Cyberbullying is a critical social problem that has been actively researched, especially by the psychology, social, and behavioral science communities. Recently, cyberbullying research has also attracted attention from the computer science community, and there has been a significant amount of research dedicated to studying the detection of cyberbullying, with an emphasis on textual cyberbullying. In this work, we focus on the understanding and detection of cyberbullying in images.

There has been significant research in understanding the psychological and social aspects of cyberbullying. The study in [52] discusses early work in cyberbullying, including the nature of cyberbullying in online and social media environments. The study in [53] reveals that cyberbullying in images is especially harmful among the other types of cyberbullying discussed in this work. Methods introduced in [54] approach the problem of cyberbullying differently,

by using bystander intervention strategies in social media networks. Many works discuss the definition of cyberbullying, although there is no universally accepted definition of cyberbullying currently [55, 56]. For example, a study [57] defines cyberbullying as “an aggressive, intentional act carried out by a group or individual, using electronic forms of contact, repeatedly and over time against a victim who can not easily defend him or herself”. However, the concept of repetition is questioned by many studies [55, 56, 58] in the field of cyberbullying. A major limitation of these studies is that they do not discuss any practical methods to defend against cyberbullying online.

Several automatic methods of cyberbullying defense that target text-based cyberbullying have emerged [59, 60, 61]. The work in [59] presents a machine learning approach to detect cyberbullying using textual content such as comments and social media post descriptions. Another automatic approach to detect textual cyberbullying is presented in [62], in which the authors present topic sensitive binary classifiers to detect cyberbullying in YouTube comments. The discussion of the language factors involved in textual cyberbullying and contextual factors of cyberbullying events in social media is presented in [63]. A recent study [64] elaborates on an approach that incorporates the use of hashtags, emotions and spatio-temporal features to detect textual cyberbullying. Another recent study [65] explores the enhancement of word embedding of cyberbullying texts, by using an embedding enhanced bag-of-words features set. Other works have also suggested the use of meta data to improve the prediction of textual cyberbullying [66, 67, 68]. However, these studies only partially address the problem of cyberbullying, as cyberbullying involves several different forms of media, such as images, in addition to text.

2.1.2 Online Hate Speech Understanding, Explanation and Detection

Several recent studies have emerged in the area of hate speech detection. In [69], the authors used Reddit, which is a community with a platform that shares information in the form of posts with the ability to be up voted or down voted based on the reader's opinion towards it. They used a public data set from subreddit /r/TD to collect 16,349,287 comments about the president and the presidency. They utilized TF-IDF to identify distinct hate words towards Donald Trump and used Wikipedia articles to identify nicknames for Trump. They concluded with findings about how humans used tools like bots to keep themselves entertained, but did not focus on pinpointing removing those bots, resulting in minimal research on preventing internet trolling.

The authors of [70] used Gab (gab.com) to find out the diffusion of hate speech. For the dataset, they used a Lexicon based filter to identify racial slurs, and chose non-ambiguous words to increase accuracy. They also utilized DeGroot's model of information diffusion to identify hateful users. They focused on the diffusion characteristics of hateful users, but not how to pinpoint and remove hateful comments in general. In [71], the authors used a large dataset from Reddit and Gab and narrowed it down to hate speech by using human intervention, which is inefficient because it takes a long time to label so many tweets. It is also unreliable because there are some tweets that are incorrectly labeled. They used a survey and crowdsourcing to label all the tweets, which is not reliable, takes too much time, and adds cost. They created a dataset of hate speech and used programs like Seq2Seq and VAE. These are unreliable because it only uses an input and output tags, and does not go through multiple veri-

fications. VAE may be unreliable for such tasks because sequences are discrete (unlike continuous image signals), and does not pinpoint certain hate words.

A recent work [21] studies the spread of hate and counter-hate during the COVID-19 pandemic. The authors collect a dataset of 2,400 tweets and train a text classifier to identify hate and counterhate tweets. The authors also find that hateful users in Twitter were less engaged in anti-Asian hate speech prior to their first anti-Asian tweet, following which such tweets turned to being more aggressive and hateful. However, a proportional rise in counterhate tweets was not observed by the authors.

Using attention mechanisms in natural language processing tasks such as classification, next sentence prediction, question answering and neural machine translation (NMT) were first introduced by [72] and [73], and most implementations are based on the models introduced in [74]. The use of attention mechanisms were broadly adapted to various NLP tasks, often achieving then state-of-the-art performances in tasks such as reading comprehension [75] and natural language inference [76]. Multi-headed attention was first introduced by [45] for NMT and English constituency parsing and termed the model as “transformer”, and further adopted for transfer learning [44], language modeling [77, 78], and semantic role labeling [79].

In this work, we focus on the BERT model [44], a large transformer [45] network. Transformers consist of multiple layers where each layer contains multiple attention heads. Each attention head takes as input a sequence of vectors $h = [h_1, \dots, h_n]$ corresponding to the n tokens of the input sentence. Each vector h_i is transformed into query, key, and value vectors q_i, k_i, v_i through separate linear transformations. The head computes attention weights α between all pairs of words as softmax-normalized dot products between the query and key vec-

tors. The output o of the attention head is a weighted sum of the value vectors, and α_{ij} represents a dot product between the query and key vectors, expressed in Equation 4.1 below.

$$\alpha_{ij} = \frac{\exp(q_i^T k_j)}{\sum_{l=1}^n \exp(q_i^T k_l)} \quad o_i = \sum_{j=1}^n \alpha_{ij} v_j \quad (2.1)$$

The attention weights can be interpreted as controlling the importance of every other token when learning the next representation of the current token.

BERT is trained using the “masked language modeling” strategy over billions of data samples, and more details about the training process can be found in [44]. An important detail about BERT training is that a special token [CLS] is added to the beginning of the text and another token [SEP] is added to the end, so that multiple sequence inputs can be trained together.

2.2 Researches on Robustness of ML Models

2.2.1 Multimodal Learning

The renewed interest in multimodal learning can be attributed to more powerful models [44, 45] that can learn strong fusion of input modalities and the availability of several multimodal datasets [80, 24]. These models and datasets have resulted in DMMs achieving impressive results on standard benchmarks. Much of the DMMs that have achieved impressive performances can be categorized under the following categories.

Traditional Fusion-based Models. Several DMMs have attempted to address how to effectively combine multimodal information [81]. Feature concatenation is one of the most preferred fusion techniques in these models, while

some of the models use other feature fusion techniques such as element-wise product. Since these models showed impressive performances on several multimodal benchmarks, they are considered strong baselines for many multimodal tasks.

Transformer-based Fusion Models. Recently, the BERT model [44], a type of transformer [45], has been shown to achieve state-of-the-art performance [49, 82] on multimodal benchmarks, by learning the interaction between the input modalities via self-attention over many different layers. For example the MMBT [49] model fuses image embeddings in the form of pooled filter maps from a ResNet model and word tokens as two segments of BERT [44]. As shown by these works, the transformer based DMMs outperform their unimodal counterparts in multimodal tasks by quite a large margin.

2.2.2 Unimodal Adversarial Attacks

The discovery of unimodal adversarial attacks has engendered active research in the safety and robustness of unimodal deep learning models. In this section, we discuss important unimodal adversarial attacks on images and text.

Unimodal Adversarial Image. A large body of adversarial attacks have been introduced in recent times that mainly focus towards robustness analysis of computer vision models. For example, several works, such as fast-gradient attacks [83], optimization-based methods [37, 38], and other such methods [40], have been proposed successfully. Furthermore, alarmingly critical real-world attacks such as adversarial patches [84] have been introduced recently, which cast serious questions on the safety of these vision models.

Unimodal Adversarial Text. Recently, some works have focused on uni-

modal adversarial text to study robustness of Natural Language Processing (NLP) models. While earlier works [85] effectively employed character level perturbations to perform adversarial attacks, more recent works have found word replacement strategies [41] to be largely effective in compromising these models

Understanding and Detecting Cyberbullying in Images

Today's Internet users have fully embraced the Internet for socializing and interacting with each other. It has been reported that 92% of users go online daily [86]. Particularly, according to recent findings from the Pew Research Center [87], 95% of adolescents surveyed (ages 12-17) spend time online, reflecting a high degree of user engagement, and 74% of them are "mobile Internet users" who access the Internet on cell phones, tablets, and other mobile devices at least occasionally.

The rise of social networks in the digital domain has led to new definitions of friendships, relationships, and social communications. However, one of the biggest issues of social networks is their inherent potential to engender *cyberbullying*, which has been widely recognized as a serious social problem. Multiple studies have suggested that cyberbullying can have severe negative impact on an individual's health, which include deep emotional trauma, psychological and psychosomatic disorders [88, 89]. According to a National Crime Preven-

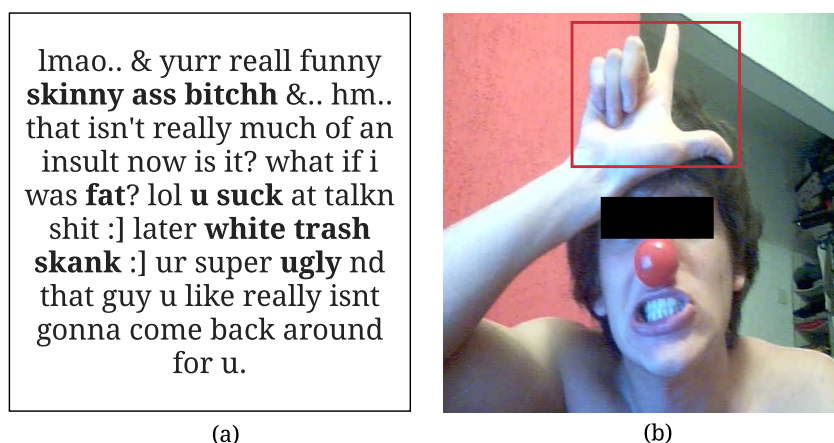


Figure 3.1: Cyberbullying in text v.s. cyberbullying in an image. (a) shows a tweet with demeaning words and phrases. (b) shows an image of a person showing a 'loser' hand gesture.

tion Council report, more than 40% of teenagers in U.S. have reported being cyberbullied [90]. Dooley et al. define cyberbullying as “Bullying via the Internet or mobile phone” [91]. Cyberbullying encompasses all acts that are aggressive, intentionally conducted by either a group or an individual in cyberspace using information and communication technologies (e.g. e-mail, text messages, chat rooms and social networks) repeatedly or over time against victims who cannot easily defend themselves [92].

Techniques used by perpetrators in cyberbullying change rapidly. For example, multimedia devices (such as mobile phones, tablets, and laptops) have now evolved from basic, single-purpose tools to high-tech multi-media devices that are fully integrated into the daily lives of millions of users. These devices introduce several new dimensions to usage of Internet services. For example, they provide on-board cameras to capture and instantly share images online. Therefore, perpetrators can use the camera-capacity of their multi-media devices to bully others through sending and distributing harmful pictures or videos to

their victims via these devices. Furthermore, the current trend for social networking websites (e.g. Facebook [93], Instagram [94] and Twitter [95]) is to provide users with options to freely share their images. Indeed, the popularity of image-sharing has seen a significant increase, thereby enabling numerous social networking websites, such as Instagram, Flickr [96] and Pinterest [97], to exclusively focus on image-sharing. These trends have introduced a shift from traditional text-based cyberbullying content like messages and tweets, to cyberbullying content that makes use of visual items to perpetrate cyberbullying behaviours among victims. Empirical evidence demonstrates that the cyberbullying in images may cause more distress for victims than do other forms of cyberbullying [53, 98]. This enhanced form of cyberbullying perpetrated through images now affects one of every two cyberbullying victims [99].

Figure 3.1 presents two examples of cyberbullying in text and in an image, respectively. Figure 3.1 (a) depicts a cyberbullying tweet [63] with the cyberbullying-related words shown in bold (such as ‘a**’, ‘fat’, and ‘ugly’). Figure 3.1 (b) depicts an image, in which a person is showing a demeaning hand sign (a ‘loser’ hand gesture) to bully his victim. We note that over the years, text-based cyberbullying detection has been a topic of in-depth study by researchers [100, 60, 62], and some state-of-the-art detectors for text-based offensive¹ content detection have been developed that are sufficiently effective in combating text-based cyberbullying. For example, on running the text in Figure 3.1 (a) against three state-of-the-art offensive text detectors namely Google Perspective API [47], Amazon Comprehend [101], and IBM Toxic Comment Classifier [102], all of them are able to detect this text as offensive with very high

¹We have used the term “offensive” here to mean harassing, harmful, toxic, or hateful content.

confidence (Google Perspective API as 92.84% likely to be offensive; Amazon Comprehend as negative sentiment with score of 0.97; and IBM Toxic Comment Classifier as offensive with score of 0.99). However, such kind of research with respect to cyberbullying in images has been largely missed, and the state-of-the-art offensive image detectors, which are very accurate on the detection of traditional offensive image content, such as nudity and violence, also do not have the capability to effectively detect cyberbullying in images. For example, on running the image in Figure 3.1 (b) through three state-of-the-art offensive image detectors namely, Google Cloud Vision API [48], Amazon Rekognition [103], and Clarifai NSFW [104], none of them could detect this image as offensive (detected by Google Cloud Vision API as “Unlikley” to cause any harm; Amazon Rekognition as no need of moderation; and Clarifai NSFW as safe for work with score of 0.67). Therefore, there is a crucial need for research that can shed more light on the phenomenon of cyberbullying in images.

The social and psychological aspects of cyberbullying in text have been the subject of intense study [105, 106, 107]. These studies have revealed that the cyberbullying in text is characterized by certain factors, such as harassing words or phrases, name-calling, and humiliating insults. However, these studies have mainly focused on its textual factors used by the perpetrators of cyberbullying with text, while largely overlooking the study of visual factors associated with cyberbullying in visual media such as images. It is a challenging task to identify the factors of cyberbullying content in images due to two reasons. First, cyberbullying in images is highly contextual and often subtle, depending on the complex interactions of several aspects of an image. Studying its factors therefore is not as straightforward as cyberbullying in text. Second, several clear definitions of cyberbullying in text are available (such as [91, 92]) and used to identify

its factors, whereas the definition of cyberbullying in images is not established, which makes the study of its factors much harder. To examine cyberbullying in images, new ways to understand its personal and situational factors should be studied.

Based on above observations and studies, we believe it is timely and important to systematically investigate cyberbullying in images and understand its factors, based on which automatic detection approaches can be formulated. In this chapter, we first collect a large dataset of cyberbullying images labeled by online participants. We analyze the cyberbullying images in our dataset against five state-of-the-art offensive image detectors, Google Cloud Vision API, Yahoo Open NSFW [42], Clarifai NSFW, DeepAI Content Moderation API [43], and Amazon Rekognition². We find that 39.32% of the cyberbullying samples can circumvent *all* of these existing detectors. Then, we study the cyberbullying images in our dataset to determine the visual factors that are associated with such images. Our study shows that cyberbullying in images is with highly contextual nature unlike traditional offensive image content (e.g., violence and nudity). We find that cyberbullying in images can be characterized by five important, high-level contextual visual factors: *body-pose*, *facial emotion*, *object*, *gesture*, and *social factors*. We then measure four classifier models (*baseline*, *factors-only*, *fine-tuned pre-trained*, and *multimodal* classifier models) to identify cyberbullying in images based on deep-learning techniques that use visual cyberbullying factors outlined by our study. Based on the identified factors, the best classifier model (*multimodal* classifier model) can achieve a detection accuracy of 93.36% in classifying cyberbullying images. Our findings about the factors of cyberbullying in

²The offensive image detectors have been selected based on their ability to detect images with certain features, such as violence, profanity, and hate symbols, which have been found in cyberbullying images.

images and the best suited classifier model for their detection can provide useful insights for existing offensive image content detection systems to integrate the detection capability of cyberbullying in images.

The key contributions of this chapter are as follows:

- **New Dataset of Cyberbullying Images.** We present a novel methodology to collect a large dataset of cyberbullying images. We first compile a set of keywords based on a collection of stories of cyberbullying provided by online users with real cyberbullying experiences. We then use these keywords to collect a large, real-world images dataset with 117,112 images crawled from online sources. The dataset with 19,300 valid images has been annotated by online participants from Amazon Mechanical Turk (MTurk) ³.
- **Measurement of State-of-the-art Offensive Image Detectors.** We present a measurement of five state-of-the-art offensive image detectors against our cyberbullying images dataset, wherein we study their effectiveness of detecting cyberbullying images. We find that these state-of-the-art detectors are not capable of effectively identifying cyberbullying in images.
- **New Factors of Cyberbullying in Images.** We analyze our dataset and identify five visual factors (i.e., *body-pose*, *facial emotion*, *object*, *gesture*, and *social factors*) of cyberbullying in images. We also find that the factors linked to cyberbullying images are highly contextual. Those factors discovered by our study play an important role towards understanding cyberbullying in images and building systems that can be used to detect

³Our dataset will be made publicly available (subject to ethical concerns, discussed in Section ??).

cyberbullying in images.

- **Extensive Evaluation of Visual Factors of Cyberbullying.** We first analyze the visual factors of cyberbullying identified in our work with exploratory factors analysis and our study reveals that the factors are associated with two underlying social constructs, which we interpret as ‘Pose Context’ and ‘Intent Context’. We then measure four classifier models based on our identified factors. We note that by including the visual factors identified in this dissertation in those classifier models, they can effectively detect cyberbullying content in images as offensive content with high accuracy. The best classifier model, which is a multimodal classifier model, can detect cyberbullying images with an accuracy of 93.36% (along with a precision and a recall of 94.27% and 96.93%, respectively).

The rest of this chapter is organized as follows. We first lay down the threat model of our work in Section 6.2. Next, we present our cyberbullying images data collection strategy in Section 3.2. We then present the motivation of our work in Section 3.3. This is followed by the details of our approach in Section 6.3.2. We discuss the implementation details of the cyberbullying images classifier models and present the evaluations of those models from different perspectives in Section 3.5. We discuss some important aspects of our approach in Section ???. This is followed by a discussion of related work in cyberbullying defense in Section ???. Finally, we conclude our work in Section ??.

3.1 Threat Model and Scope

Threat Model. In this chapter, we consider two types of users: 1) a *perpetrator* is a user who sends a cyberbullying image to other users; and 2) a *victim* is a user who receives a cyberbullying image from a perpetrator. We consider the scenario where images depicting cyberbullying are sent by a perpetrator to a victim when the perpetrator uploads such images online, posts such images on social networks or shares such images via mobile devices. The affected users are the victims viewing the photo. In our current work, we focus on addressing cyberbullying in images, and do not consider images accompanying with cyberbullying text. We also do not consider the traditional offensive image content, such as nudity, pornography, and violence, which have been deeply studied by previous work [43, 103, 104]. Besides, we do not consider cyberbullying cases with inside meaning that is only understandable to specific users. For example, a perpetrator *Alice* sends images of snakes to a victim *Bob* since *Bob* has a fear of snakes.

Problem Scope. In this chapter, our goal is to identify factors of cyberbullying in images and to demonstrate that they can be used to detect cyberbullying content in images. Our major purpose is not to design a novel classifier model that achieves the highest detection accuracy, instead we analyze several typical classifier models to demonstrate that they can effectively detect cyberbullying content in images after integrating the visual factors of cyberbullying identified by our work.

3.2 Cyberbullying Images Data Collection

To identify factors of cyberbullying in images, we need an effective mechanism to collect a large amount of cyberbullying-related visual information, which should be representative of real-world cyberbullying found in images. In our work, we introduce an approach to collect a large dataset of cyberbullying images, wherein we first extract a set of keywords and keyphrases of cyberbullying from cyberbullying stories about self-reported experiences of real victims of cyberbullying, which are then used to collect a cyberbullying images dataset. Our data collection tasks are approved by IRB. We elaborate the methodology of our approach in the following section.

3.2.1 Methodology

In this section, we discuss our pre-data collection study for collecting cyberbullying images dataset. In this study, we use the cyberbullying stories from Internet users with their own cyberbullying experiences to collect an images dataset that is representative of real-world cyberbullying in images.

We use the self-reported stories from [108], a collection of anonymized stories of cyberbullying collected from voluntary online users who have themselves experienced cyberbullying. Therefore, this corpus of cyberbullying stories and experiences is a wealth of cyberbullying related information for research in this field. We mined this corpus and compiled 265 unique stories of cyberbullying, each of which is contributed by a user. Among the users in this study, 30 users reported themselves as adults and 197 reported themselves as below the age of 18 years. A majority of users reported themselves as female (178 users), whereas a relatively smaller number of users reported themselves

as male (54 users). The rest of the users wished not to report their age or gender.

3.2.2 Cyberbullying Keywords Extraction

To extract keywords of cyberbullying in images from the cyberbullying stories, we used the following method. We first removed all identifiers from the cyberbullying stories information. Next, we used the Python NLTK library [109] to remove stop words [110] from all stories. At the end of this process, we collected 2,648 keywords. Then, we used the sentiment analyzer of the Python NLTK library to remove neutral and positive words, followed by manual verification of the words, which left us with 378 words (we used a polarity threshold of -0.55⁴). We used these words as the final keywords list to collect potential images of cyberbullying content for our dataset. Table 3.1 shows some cyberbullying story samples and the keywords extracted with our methodology.

Stories	Extracted Key-words
The oldest boy's dad is crazy and has been sending text containing verbal harm messages and even a text holding a gun and a message to the boyfriend and just wanted to know what we should do.	holding, gun, crazy, harm
I have been threatened that someone was going to kill me and told me to shut the f*ck up here is a picture.	f*ck, kill, threatened
How does it feel being the fat ugly outcast of all your pretty skinny friends why do you take a bazillion pictures of yourself.	fat, ugly
I am keep getting name called such as f*g, douche bag, small d*ck.	f*g, douche, d*ck

Table 3.1: Samples of cyberbullying stories and the extracted keywords.

⁴Polarity threshold is defined in the interval -1 to +1. More negative words have a polarity value closer to -1.



Figure 3.2: Image samples that did not have any Regions of Interest (ROIs).

3.2.3 Data Collection and Annotation

The models of cyberbullying detection in images should be capable of differentiating between images with cyberbullying content from other benign images. In addition, they should also distinguish between harmless images that do not intend to cause cyberbullying, so that false alarms are reduced. To collect a diverse dataset of images that captures important patterns of cyberbullying in images, we used multiple web sources, including web search engines (Google, Bing, and Baidu) and publicly available social media images from multiple online social media websites (Instagram, Flickr, and Facebook). We collected images using keywords and phrases compiled from our findings in Section 3.2.2. We finally collected 117, 112 images using our data collection methodology. Next, we used an object localization tool called YOLO [111] to exclude images that do not have any regions of interest (ROIs). These are images that typically do not have any content and hence, do not convey any meaning. Some samples of images that were excluded in this step are depicted in Figure 3.2. After this step, we were left with 19, 300 images for annotations.

3.2.4 Image Annotation

We used MTurk to obtain annotations for the collected images. Our objective was to annotate whether an image contains cyberbullying content or does not contain any cyberbullying content. Therefore, we referred to the definition of cyberbullying from [107, 112] as guidelines for annotation. Specifically, we focused on cyberbullying in images as “an act of online aggression carried out through images” for the participants of our study (the interface of our image annotation task can be found in Appendix 3.5.3). We displayed a warning to participants about the nature of the task in both the task title and description according to MTurk guidelines. We placed a restriction that only allows participants with an approval rating of 90% or higher and 1000 approved HITs to participate in our annotation task. We offered a \$0.05 reward for each task submission and recorded an average task completion time of 18 seconds per task. We allowed each image to be annotated by three distinct participants and chose the majority voted category as the final annotation. Finally, in our dataset, 4,719 images were annotated as cyberbullying images and 14,581 images were annotated as non-cyberbullying images.

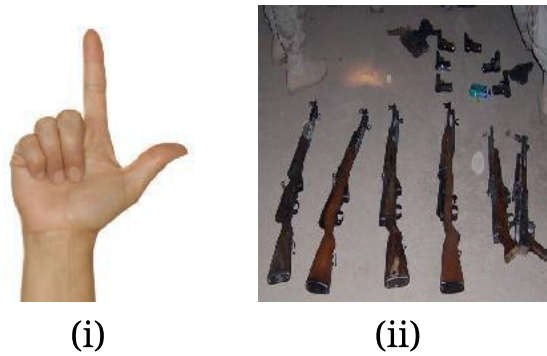
We computed the inter-rater agreement [113] using the Randolph’s κ -measure [114], a statistical measure of agreement between individuals for qualitative ratings. Note that, $\kappa < 0$ corresponds to no agreement, $\kappa = 0$ to agreement by chance, and $0 < \kappa \leq 1$ to agreement beyond chance. We measured κ on our cyberbullying images dataset, and obtained $\kappa = 0.80$.

3.3 Motivation and Observation

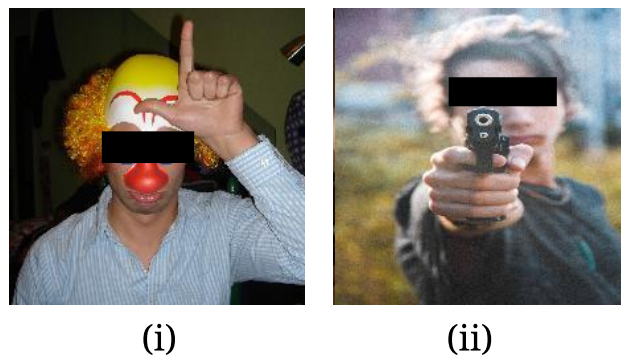
To illustrate our motivation, we first conducted a study into the detection capability of several popular offensive image detectors, including Google Cloud Vision API (Google API), Yahoo Open NSFW, Clarifai NSFW, DeepAI and Amazon Rekognition, and ran these detectors against images annotated as cyberbullying in our dataset. We chose these detectors because they have the ability to detect certain offensive attributes in images. We computed the performance of these detectors in terms of precision and recall metrics on the cyberbullying images as shown in Table 3.2. From Table 3.2, we observed that those state-of-the-art detectors have low performance in detecting cyberbullying images. Among those popular offensive image detectors, Yahoo Open NSFW (precision = 36.27%, recall = 2.82%) and Clarifai NSFW (precision = 42.94%, recall = 10.67%) offer overall lowest performance. DeepAI (precision = 69.43%, recall = 15.92%) and Amazon Rekognition (precision = 77.44%, recall = 23.55%) offer only a small improvement over the previous two detectors, although they consider a higher number of attributes. Among the popular detectors, Google API (precision = 35.65%, recall = 39.40%) achieves the best performance, although this detector also misses a large number of cyberbullying samples (60.59%). A more startling observation was that **39.32%** of the cyberbullying samples could circumvent all five popular offensive image detectors.

Detector	Precision	Recall
Google API	35.65%	39.40%
Yahoo Open NSFW	36.27%	2.82%
Clarifai NSFW	42.94%	10.67%
DeepAI	69.43%	15.92%
Amazon Rekognition	77.44%	23.55%

Table 3.2: Precision and recall of popular offensive image detectors.



(a) Without cyberbullying context.



(b) With cyberbullying context.

Figure 3.3: Image context in cyberbullying images.

Image #	Google API	Yahoo NSFW	Clarifai	Deep AI	Amazon
Figure 3.3a (i)	0.2	0.17	0	0.17	0
Figure 3.3a (ii)	0.2	0.005	0.05	0.003	0.98
Figure 3.3b (i)	0.2	0.008	0.01	0.008	0
Figure 3.3b (ii)	0.2	0.004	0	0	0.97

Table 3.3: Detection scores of existing detectors on image samples in Figure 3.3.

After an examination of cyberbullying images annotated by users in our dataset, we found that most of such images are context-aware. Figure 3.3 depicts two images without cyberbullying context (annotated as non-bullying images by participants) and two other images with cyberbullying context (annotated as

Detector	Categories of Offensive Content	Limitations
Google Cloud Vision API	Object detection, face detection, image attributes, web entities, content moderation	No offensive image detection capability
Yahoo Open NSFW	NSFW detection	Limited to only nudity detection
Clarifai NSFW	NSFW detection, content moderation concepts	Only limited types of concepts (explicit, suggestive, gore and drug)
DeepAI Content Moderation API	Content moderation	Only limited to a few objects (guns confederate flag)
Amazon Rekognition	Object and scene detection, face recognition, emotion detection, unsafe image detection	Limited categories of unsafe detection (nudity and violence)

Table 3.4: Capabilities of existing detectors and their limitations.

bullying images by participants), respectively, from our dataset. The images in Figure 3.3a only show a possible factor (a demeaning hand gesture or a gun), but without any contextual information. In contrast, Figure 3.3b shows images that portray these factors with contextual information, such as a person deliberately showing the hand gesture in Figure 3.3b (i) to the viewer, or the person in Figure 3.3b (ii) pointing the gun at the viewer. Table 3.3 depicts the scores of each popular offensive image detectors on those image samples. We observed that the Google API scores all the image samples equally, and rates them as “unlikely” to be unsafe. Yahoo NSFW, Clarifai and DeepAI seem to have very small scores for all image samples, and therefore are unable to differentiate between non-cyberbullying and cyberbullying content. Amazon Rekognition seems to only detect guns in Figure 3.3a (ii) (score = 0.98) and Figure 3.3b (ii) (score = 0.97), and naively flags down all such images. Thus, we note that the existing detectors cannot detect cyberbullying in images effectively.

We further study the capabilities and limitations of the five state-of-the-art offensive image detectors, as depicted in Table 3.4. From Table 3.4, we can first

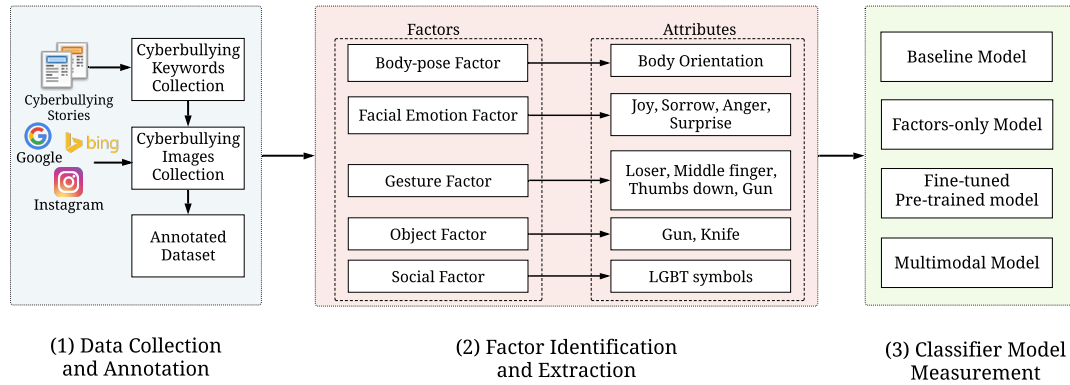


Figure 3.4: Approach overview.

observe that none of state-of-the-art detectors consider cyberbullying in images as a category of offensive content. Thus, our first motivation is that this important category of offensive content should be included by existing systems as an offensive content category. Secondly, since the factors of cyberbullying in images are unknown, the existing detectors are not capable of detecting them. Thus, we are motivated to shed light on identifying the visual factors of cyberbullying so that they can be automatically detected in images.

3.4 Our Approach

We analyse the cyberbullying images in our dataset in three steps: (i) understand and identify the factors related to cyberbullying in images (Section 3.4.2); (ii) extract those factors from images (Section 3.4.3); and (iii) examine the usage of those factors in classifier models (Section 3.4.4).

3.4.1 Approach Overview

The main components involved in our approach are depicted in Figure 3.4. We first collect a large dataset of cyberbullying images to study this phenomenon (Figure 3.4, Step 1 “*Data Collection and Annotation*”). Next, we analyze the collected data to identify factors in the way participants consider cyberbullying in images (Figure 3.4, Step 2, “*Factor Identification and Extraction*”, “*Factors*”). In this step, we identify five factors of cyberbullying in images in our dataset: body-pose, facial emotion, gesture, object and social factors. We then focus on two processes to study and address cyberbullying in images: “*Factors Identification and Extraction*”, “*Attributes*” (Figure 3.4, Step 2) and “*Classifier Model Measurement*” (Figure 3.4, Step 3). In *Factor Extraction*, our primary goal is to extract the attributes of those factors of cyberbullying in images. We use several off-the-shelf tools and techniques to extract these visual factors. In *Classifier Model Measurement*, we then use several *deep learning*-based classifiers to demonstrate that the identified factors can be used to effectively detect cyberbullying in images. To understand the importance of these factors and to study their effectiveness in detecting cyberbullying in images, we train four classifier models: baseline, factors-only, fine-tuned pre-trained, and multimodal models. During the evaluation of a new photo, we extract the factors and predict a score of cyberbullying in images using those classifier models. We discuss our methodology in more details in the following sections.

In our work, the context of cyberbullying refers to the story that an image is conveying, where the intent is to bully receivers/viewers of the image. For example, a photo with a person at a gun shop looking at various guns on display has a totally different context compared with a photo, which depicts a person

pointing a gun at viewers. Towards this end, we study this context in-depth, identify its factors in images, and design techniques that identify cyberbullying content by capturing the context.

Factor	Attribute	Cyberbullying	Non-cyberbullying	Description
Body-pose	Front pose	0.86	0.53	Pose of subject in image is towards the viewer
	Non-front pose	0.50	0.84	
Emotion	Joy	0.34	0.25	Facial emotion of subject in image
	Sorrow	0.02	0.02	
	Anger	0.09	0.04	
	Surprise	0.07	0.05	
Gesture	Hand gesture	0.71	0.32	Hand gesture made by subject in image
	No hand gesture	0.70	0.94	
Object	Threatening object	0.33	0.06	Threatening object present in image
	No threatening object	0.94	0.99	
Social	Anti-LGBT	0.45	0.06	Anti-LGBT symbols and anti-black racism in image
	Anti-black racism	0.03	0.00	

Table 3.5: Analysis of cyberbullying factors. Higher value of cosine similarity indicates higher correlation.

3.4.2 Factor Identification

Various studies [52, 55, 56] focused on text-based cyberbullying have tried to understand its nature, and revealed several personal and situational factors, such as the use of abusive or harassing words and phrases. However, no existing research has attempted to understand the factors associated with cyberbullying in images. To examine cyberbullying in images, new personal and situational factors related to image content should be studied. The identified factors can help formulate classifier models for detection, and potentially enable popular offensive content detectors (e.g., Google Cloud Vision API and Amazon Rekognition) to automatically detect cyberbullying in images as an offensive content category.

To study the factors of cyberbullying in images in our dataset, we conduct an experiment by considering all the cyberbullying and non-cyberbullying images in our dataset. In this experiment, we use existing tools to analyze the nature of

the images considering recurring visual factors we observe in the dataset, summarised in Table 3.5. We analyze the body-pose [115] of the subject in an image, as prior research [116] has shown that threatening poses are a commonly used tool in cyberbullying. We analyze hand gestures [48] as hand gestures are popular forms of sign language used to convey meaning through images. We study the facial emotion [117] of the subject in images, as facial emotions can convey several meanings to a viewer. We study the objects [111] that are used by perpetrators to threaten, or intimidate a victim. Lastly, we study social factors such as anti-LGBT (lesbian, gay, bisexual, transgender, and queer) content in images in our dataset. We use the cosine similarity [118] to compare the differences of these factors with respect to cyberbullying and non-cyberbullying images.

Body-pose factor. We conduct a preliminary study of the correlation of the visual factors with images that have been labeled as cyberbullying vs. non-cyberbullying by observing the cosine similarity between images depicting the visual factors (outlined in Table 3.5). We observe that images depicting persons who pose *at* the viewer (front pose) had strong correlation with cyberbullying images (cosine similarity = 0.86, 74.74% of cyberbullying images). In contrast, these images with the person posing at the viewer were observed to have a much lower correlation (cosine similarity = 0.53, 28.29% of non-cyberbullying images) with non-cyberbullying images (i.e., these images were mostly non-front pose). On examining such cyberbullying images, we observe that these images depicted subjects that are directly looking at the image viewer in order to directly engage the viewer, whereas most subjects in non-cyberbullying images had posed looking away.

Facial emotion factor. Facial emotions have been known to convey significant meaning regarding what a person is feeling. Thus, we study the correla-

tion of facial emotions (e.g., sorrow, joy, anger, and surprise) with cyberbullying images. We observe that most cyberbullying images do not have specific emotions expressed by a subject. We also observe that even in cyberbullying images, subjects do not show any strong emotions. In fact, we observe that these subjects generally showed happy emotions such as joy (cosine similarity = 0.34, 11.39% of cyberbullying images). Our preliminary observations reveal that subjects may generally depict themselves mocking the viewer by showing emotions of joy.

Hand gesture factor. Hand gestures are a popular method that Internet users use to convey meaning in images [119, 120]. We find a high correlation of hand gestures (e.g., loser, middle finger, thumbs down and gun point) with cyberbullying images (cosine similarity = 0.71, 50.6% of cyberbullying images), indicating that in cyberbullying images, hand gestures may constitute an important factor.

Object factor. Next, we discuss the correlation of threatening objects (e.g., gun, knife) with the cyberbullying images in our dataset. We also observe some correlation of threatening objects (cosine similarity = 0.33, 10.6% of cyberbullying images) with cyberbullying images, which indicates Internet users may use these objects to threaten or intimidate a viewer [121]. Although, we also observe that many cyberbullying images (cosine similarity = 0.94, 89.40% of cyberbullying images) also do not depict direct use of these objects to cyberbully their victims. This could be due to the belief that Internet users generally may use more subtle tools to perpetrate cyberbullying, rather than directly using such threatening objects, which may risk initiating action by law enforcement agencies.

Social factor. Prior works [122, 123] have shown that cyberbullying is a

deeply concerning social issue. Hence, we manually analyze the cyberbullying images in our dataset for current social-related factors, such as anti-LGBT [124] and racism [125]. We find that a small part of images consisted of anti-LGBT symbolism (cosine similarity = 0.45, 1% of cyberbullying images), and images depicting “black-face” and historical references to hanging (cosine similarity = 0.03, < 1% of cyberbullying images).

Next, we study the correlation of a person depicting a hand gesture or a threatening object with respect to cyberbullying images (Table 3.6). We observe a significant correlation of person and hand gestures in cyberbullying images (cosine similarity = 0.72, 95.31% of cyberbullying images). On further examination, we observe that many cyberbullying images depict a person directly showing a gesture towards the image viewer. We also observe that some images with only a hand gesture and no person is significantly less correlated with cyberbullying (cosine similarity = 0.10, 4.69% of cyberbullying images), which may indicate that presence of person invokes stronger context in an image, and a factor by itself may not actually convey cyberbullying. We make a similar observation involving objects and person regarding cyberbullying images (cosine similarity = 0.31, 90.4% of cyberbullying images). We observe that many photos of objects (e.g., guns and knives) alone were not labeled as cyberbullying (cosine similarity = 0.02, 9.6% of cyberbullying images), but photos depicting a person holding these objects were overwhelmingly labeled as cyberbullying.

	Cyberbullying		Non-cyberbullying	
	Person	No person	Person	No person
Object	0.31	0.09	0.02	0.07
Gesture	0.72	0.10	0.34	0.07

Table 3.6: Analysis of correlation of person with threatening object and gesture.

From our analysis, we observe that cyberbullying in images is highly *con-*

textual in nature, involving very specific factors (outlined in Table 3.5). In our work, we use these factors to train classifier models and demonstrate that they can be effectively used to detect cyberbullying in images. A crucial requirement of defense against cyberbullying in images is to accurately detect cyberbullying based on those images. The high correlation of cyberbullying with certain factors may indicate that classifier models based on these factors could potentially detect cyberbullying in images. Furthermore, popular offensive content detectors currently do not consider cyberbullying as a category of offensive content in images and hence lack the capability to detect it. One of the objectives of our work is to highlight the importance of cyberbullying in images, so that it can be included as a category of offensive content in popular offensive content detectors. In our work, we use the visual factors of cyberbullying to demonstrate that they can be used in deep learning models (such as the ones in these content detectors) to successfully detect cyberbullying in images with high accuracy.

3.4.3 Factor Extraction

Our aim is to identify a set of cyberbullying factors in images that are minimally correlated and best predict the outcome (i.e., presence of cyberbullying in images). However, cyberbullying in images is a complex problem, and such factors are not directly derivable from image data with currently available learning techniques. Therefore, we extract these factors based on our collected dataset and preliminary analysis, and catalog them as follows.

- **Body-pose factor extraction.** Regarding body pose of a person appearing in an image, there may be several aspects of the person, such as orientation, activity, and posture. Specifically in our dataset, we observe that in

cyberbullying images, the subject is predominantly oriented towards the image viewer (i.e. towards the camera). For example, Figure 3.5 shows two image samples from our dataset. Figure 3.5(i) depicts a cyberbullying sample and Figure 3.5(ii) depicts the pose of the subject. It can be observed that this pose of the subject indicates that the subject in this image is oriented directly at the viewer and pointing a threatening object (e.g., gun) at the viewer. However, this is in contrast to Figure 3.5(iii), whereas the pose depicted in Figure 3.5(iv) of a non-cyberbullying sample indicates the subjects are not oriented towards the viewer and the threatening object not pointed towards the viewer. Thus, we wish to capture these orientations related to body-pose.

We used OpenPose [115] to estimate the body-pose of a person in the image. OpenPose detects 18 regions (body joints) of a person (such as nose, ears, elbows and knees), and outputs the detected regions and their corresponding detection confidence. We use the confidence scores of the regions as the factor values as this indicates the confidence about the appearance of those regions in the image.

- **Facial emotion factor extraction.** Since cyberbullying may involve the subject in an image expressing aggression or mocking a victim, we were specifically interested in capturing facial emotions related to these expressions, as the facial emotions of subject in images may be good indicators of the intent of the person towards conveying such expressions. For example, an angry expression could indicate an intent to be aggressive or threatening to a viewer, or a happy (e.g., sneering, taunting) expression could indicate an intent to mock the viewer.

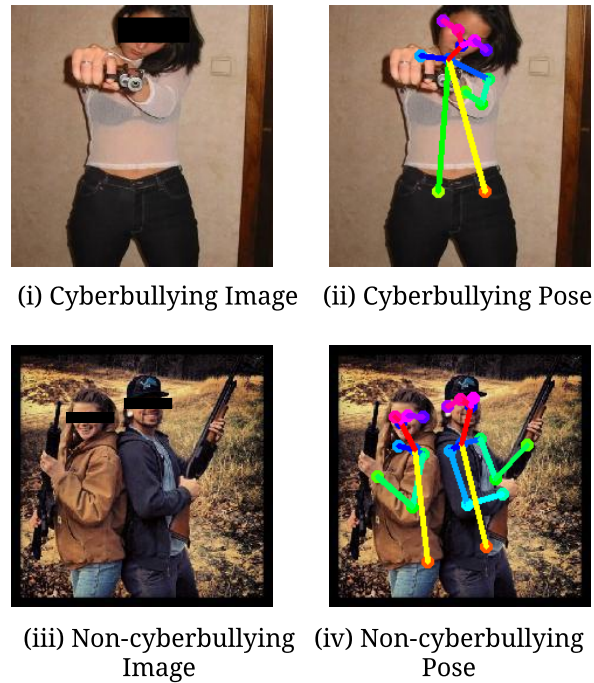


Figure 3.5: Cyberbullying Vs. non-cyberbullying body-pose.

We extract the emotions in our dataset using two sources, OpenFace [117] and Google Cloud Vision API [48]. We choose the emotion categories that are indicated with high confidence by both these sources. Overall, we use four emotion categories: joy, sorrow, anger, and surprise.

- **Gesture factor extraction.** There exist several hand gestures that subjects use in images and most of these are not harmful (e.g., the victory sign, thumbs up and OK sign). We observed that in cyberbullying images in our dataset, the hand gestures were used as tools to convey harmful intent by perpetrators of cyberbullying. Such images (e.g., Figure 3.6) depict subjects making mocking or threatening hand gestures, such as the loser gesture (Figure 3.6 (i)), middle finger (Figure 3.6 (ii)), thumbs down (Figure 3.6 (iii)), and gun gesture (Figure 3.6 (iv)). Hence, we were interested

in capturing these harmful gestures we found in cyberbullying images.



Figure 3.6: Some hand gestures found in cyberbullying images in our dataset.

We use the tag suggestions by Google Cloud Vision API [48] to indicate if an image depicts any hand gestures. The tags detected by this API do not provide fine-grained gesture categories. Therefore, we only use the presence or absence of a hand gesture as the feature indicative of hand gesture factor.

- **Object factor extraction.** Different objects depicted in an image can indicate different intents of the subject in the image. We observe that a large number of cyberbullying images portrayed the use of threatening objects, such as guns and knives, and hence we are specifically interested in capturing these objects. In cyberbullying [123, 126], perpetrators specifically

use threatening and intimidation to cyberbully their victims. Specifically, in cyberbullying in images, perpetrators can use images of themselves using such threatening objects to cyberbully the victims and hence we were interested in capturing these types of objects.

We use an open source object detection system called YOLO [111] to detect the objects in images of our dataset. YOLO outputs the object category as well as the confidence score of detection for each object depicted in an image. Since YOLO outputs a large set of categories of images, we limit the objects categories to only the categories that we are interested in (e.g., gun, knife, revolver, etc.). Then, we use the confidence scores of the subset of objects as features for this factor.

- **Social factor extraction.** We observe certain social factors in cyberbullying images that perpetrators could use to convey intent of cyberbullying. Such factors predominantly included anti-LGBT symbolism in our dataset, such as portraying certain LGBT symbols in a derogatory manner, or defacing such symbols.

Detecting such social factors in images is a complex task and currently there are no detectors that can satisfactorily detect these factors. Thus, we directly label the images that contained such symbolism in our dataset, based on online information about this topic [124, 125]. However, we note that this factor category maybe very vast, and we only consider the social factors that we observe in our collected dataset in this chapter.

In our dataset, we also find that some cyberbullying images, such as the ones depicting the social factor, do not have a person. For these images, we represent the feature vectors for these factors as zero vectors, indicating the absence of

people in these images. For example, since the body-pose factor is dependent on a person being present in the image, we represent the body-pose feature vector with the zero vector when the image does not contain a person.

3.4.4 Measurement of Machine Learning Models for Classification of Cyberbullying in Images

Feature Selection. In computer vision applications, deep neural networks (such as Convolutional Neural Networks (CNNs) have enabled the automatic selection of image features. Previous works [127] have shown that the convolutional layers of a CNN learn to identify various features, such as edges, objects, and body parts, to compute a prediction. Although this approach has yielded significantly accurate results in specific computer vision tasks (such as object detection), such an approach cannot be directly applied to a complex task, such as detection of cyberbullying in images, due to the presence of several contextual factors. Therefore, to detect cyberbullying in images, we first need to identify the factors that determine cyberbullying. In our work, we catalog five factors of cyberbullying based on the images in our dataset. Furthermore, we study the importance of each factor towards the effective detection of cyberbullying in images.

Classifier Models. To demonstrate the effectiveness of the factors identified in this chapter, we use machine learning models to predict cyberbullying vs. non-cyberbullying in images. Our main focus is to examine which of the machine learning models can achieve high accuracy of detection of cyberbullying in images. Although we demonstrate the effectiveness of the identified visual factors,

we are also interested in learning at what level of abstraction the factors have the most predictive power. Thus, we have built several classifiers at different levels of abstractions, spanning from the raw image consisting of lowest level features to the high-level factors identified in this chapter. We have evaluated all the models using 5-fold cross-validations. This study would also allow us to investigate if the classification of cyberbullying in images can be trivially solved using simple features. Below, we explain these different classifier models.

1) *Baseline model*. As a baseline model, we directly train a deep CNN with the low level image features. Our intuition behind choosing this baseline model is because we want to include use cases that are common among most of existing detectors, which are all based on CNNs. Another reason for choosing CNN is that it is still the most effective model for image-based tasks. All images were resized to 224×224 pixels and then fed into a VGG16 untrained model, which is a popular 16 layer deep CNN for computer vision tasks. This represents a model that is trained on the most concrete set of features, i.e., the raw pixel values of the images.

2) *Factors-only model*. This model that we formulate is based on a multi-layer perceptron network with only the factors identified in this chapter as inputs. Our objective is to investigate whether the factors identified alone could be used with no image features to classify images as cyberbullying vs. non-cyberbullying.

3) *Fine-tuned pre-trained model*. Fine-tuning a pre-trained model allows us to transfer the knowledge in one task to perform the task of cyberbullying classification in images. This process is analogous to how humans use knowledge learned in one task to solve new problems. We fine-tune the 16 layer VGG16 model that is trained on the object detection task using the ImageNet

dataset [128], which consists of over 14 million images. In our factors analysis, we find that certain object categories, such as person, gun, and knife, could be responsible for causing cyberbullying. This intuition leads us to choose a model trained for object detection as a baseline pre-trained model. To fine-tune this pre-trained model, we replace the final linear layer with a linear layer that outputs two values followed by the Sigmoid activation function, in order to predict cyberbullying vs. non-cyberbullying. We only train the linear layers and keep the other layers fixed as it is the norm in fine tuning.

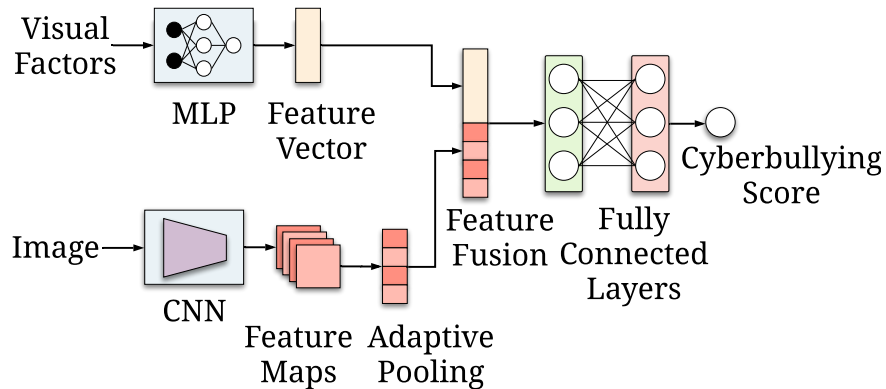


Figure 3.7: Multimodal model used in our approach.

4) *Multimodal model.* In this model, we combine the low level image features (Figure 3.7, “Image”) with the factors identified in this chapter (Figure 3.7, “Visual Factors”). To achieve this, we need a method to combine these visual factors and image features. We combine these features using feature fusion techniques, such as early and late fusion [129]. We use the VGG16 pre-trained model for image features (Figure 3.7, “CNN”) and use a multi-layer perceptron model (Figure 3.7, “MLP”) for the factors related features, and combine the feature vectors from both these models using late fusion. The VGG16 model produces an output of 512 convolutional feature maps of dimension 7×7 . We flatten the

convolutional feature maps using adaptive pooling into one-dimensional vector of 512 and fuse it (Figure 3.7, “Feature Fusion”) with the output of the MLP network. We train this model in a joint manner (Figure 3.7, “Fully Connected Layers”) to classify images as cyberbullying vs. non-cyberbullying. Ideally, we expect this model to perform the best among all models discussed, since this model is presented with low level as well as high level features (i.e., the visual factors).

3.5 Implementation and Evaluation

In this section, we first discuss the implementation of the machine learning models used in our work, followed by experiments to evaluate our approach from different perspectives. The major goals of our evaluation are summarized as follows.

- Understanding the effectiveness of factors of cyberbullying in images by using exploratory factors analysis (Section 3.5.2).
- Demonstrating the effectiveness of our factors in accurately predicting cyberbullying in images, using four classifier models (Figure 3.10 and Table 3.9).
- Studying the performance overhead of our model when integrated in mobile devices (Section 3.5.2).
- Evaluating the false positives of our model on the images depicting the American Sign Language (Section 3.5.3).

- Validation of our cyberbullying factors with a wider audience (Section 3.5.3).
- Studying the representativeness of our cyberbullying images dataset (Section 3.5.3).
- Analyzing the capabilities of the state-of-the-art offensive image detectors with respect to the cyberbullying factors (Section 3.5.3).

3.5.1 Implementation

In this section, we discuss the implementation details of the classifier models for cyberbullying in images. We use the PyTorch framework [130] to train and deploy these models. In our work, we use the VGG-16 network [131] for feature extraction in the models. We use the VGG-16 model that is pre-trained on ImageNet dataset [132] for the purpose of transfer learning. Following PyTorch naming conventions, we remove the last fully connected layer of the VGG-16 network (named “fc1”). For the multimodal model, we add a fully connected layer having 2 units for classification. Next, we add a sigmoid activation function on the output of classification. We train all the models for the same number of epochs.

3.5.2 System Effectiveness Evaluation

Understanding the Effectiveness of Cyberbullying Factors

We study in detail the factors of cyberbullying in images identified in this chapter in terms of their effectiveness in characterizing cyberbullying in images.

We first study the most frequently occurring visual factors that characterize cyberbullying images, as depicted in Table 3.7. For cyberbullying images, we note that *Body-pose* accounts for 76.91% frequency, which indicates that it is an important cyberbullying factor. *Gesture* (50.6%) is the next most frequent factor, which indicates that in cyberbullying in images, subjects may deliberately use gestures to convey harmful meaning to a viewer. Among the facial emotions, we observe that the predominant emotion in cyberbullying images is *joy* (11.41%). This is an interesting observation that indicates that subjects may be expressing joyful facial expressions to mock a viewer. The next most frequent factor is observed to be *object* (10.58%). A significant portion of the cyberbullying images involved the subject showing certain threatening objects such as guns and knives to potentially directly intimidate a viewer.

#	Factor	Cyberbullying Frequency	Non-cyberbullying Frequency
1	Body-pose	76.91%	31.41%
2	Joy	11.41%	5.97%
3	Sorrow	0.06%	0.06%
4	Anger	0.83%	0.19%
5	Surprise	0.51%	0.26%
6	Gesture	50.6%	10.76%
7	Object	10.58%	0.42%
8	Social	0.53%	0.00%

Table 3.7: Frequencies of factors responsible for labeling an image as cyberbullying or non-cyberbullying.

The factors frequencies in non-cyberbullying images are depicted in Table 3.7. In comparison to cyberbullying images, we observed that *body-pose* factor plays a significantly less important part in non-cyberbullying images (31.41%). Same observation is made about the *gesture* factor (10.76%). We observe that the gestures in non-cyberbullying images are predominantly harmless, such as the victory sign and the thumbs up sign. The *joy* facial emotion is higher than other emotions in these images too (5.97%), although it is found to be lower than in

cyberbullying images.

#	Factor	Spearman ρ
1	Body-pose	0.39
2	Joy	0.08
3	Sorrow	0.00
4	Anger	0.04
5	Surprise	0.02
6	Gesture	0.42
7	Object	0.26
8	Social	0.06

Table 3.8: Correlation coefficient (Spearman ρ) between visual factors and cyberbullying label. The coefficients are significant at $p < 0.001$ level.

Next, we conduct a study to understand the associations between human level annotations on images and the identified factors. Table 3.8 depicts the correlations (Spearman ρ) for visual factors and cyberbullying images. In Table 3.8, significant correlation coefficients suggest an association between the factors and the rationale of human annotators about cyberbullying images. A strong association of 0.39 is observed in case of the *body-pose*, indicating that annotators tend to agree that a subject in a cyberbullying image intentionally poses at a viewer. Similarly, strong association is observed for *gesture* (0.42) and *object* (0.26), indicating that annotators generally considered that photos depicting these factors are generally cyberbullying. These associations may imply that annotators may consider those images as cyberbullying, which depict clear meaning and context, as the strongly associated factors (*body-pose*, *gesture*, and *object*) imply most clear meanings among all the other factors.

In our next study, we are interested in studying those subsets of uncorrelated visual factors that are most effective in distinguishing cyberbullying images from the non-cyberbullying images. We conduct Exploratory Factor Analysis (EFA) to discover the uncorrelated factor sets. The Scree plot depicted in

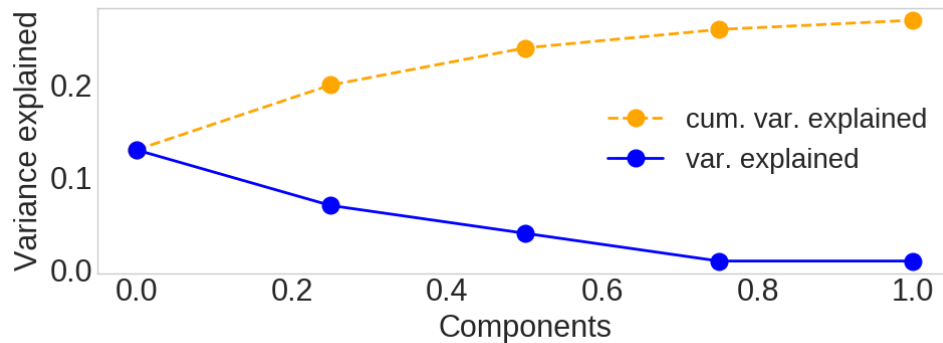


Figure 3.8: Scree plot showing proportions of variance and cumulative proportion of variance explained by each component.

Figure 3.8 suggests the number of factors ⁵ to extract. The point of inflection in the Scree plot after the second factor may suggest that two factor subsets can represent the cyberbullying in the data. Figure 3.9 exhibits the factor loadings after a ‘varimax’ rotation. We omit loadings that are too low. A feature is associated with the factor, with which it has a higher loading than the other, and also that features associated with the same factor are grouped together for certain descriptive categories. More specifically, the facial emotions *sorrow*, *surprise* and *anger* are grouped together, and characterized by lower loadings. The *object* category grouped with these emotions reveals a characteristic observation that facial expression are generally more negative when coupled with threatening object. However, the *joy* emotion is away from these indicating it is an important uncorrelated factor. *Body-pose* and *gesture* are also uncorrelated factors. From these observations, intuitively cyberbullying in images could be related to the facial expression of a person and the overall body (pose, object in hand and gesture) of a person. Thus, based on our analysis, cyberbullying in images could be intuitively characterized with two social constructs: “Pose Context”

⁵Here, “factor” refers to EFA factors and not visual factors of cyberbullying.

(pose related factors, such as pose and gesture) and “Intent Context” (e.g. an image depicts an intent using facial emotion or object).

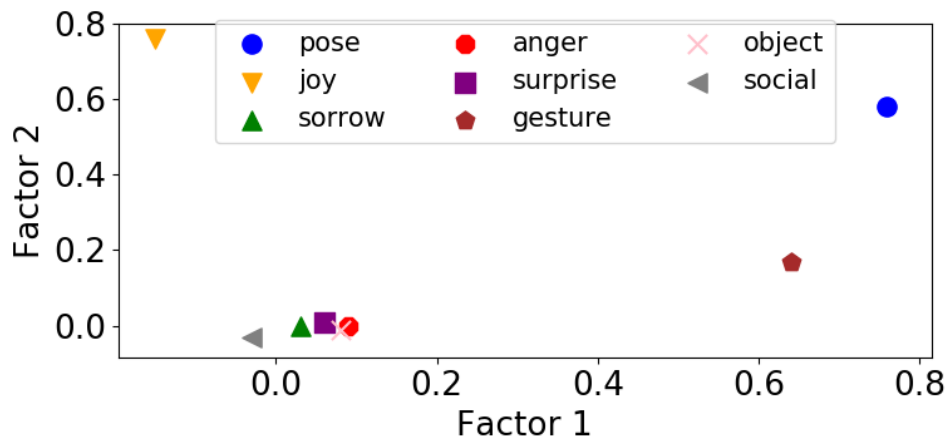


Figure 3.9: Factor loadings of the features across two extracted factors.

Effectiveness Evaluation of Classifier Models

To understand the effectiveness of the classifier models trained on high-level factors and low-level image features, we randomly select 80 percent of our dataset for training (with 5-fold cross validation) and 20 percent of the dataset for testing and we run the four types of classifiers on images from our test dataset. We perform the Receiver Operating Characteristics (ROC) [133] analysis of the classifier models for cyberbullying images prediction. The ROC analysis provides a means of reviewing the performance of a model in terms of the trade-off between False Positive Rate (FPR) and True Positive Rate (TPR) in the predictions. The ROC plot of the classifier models for cyberbullying detection in images is depicted in Figure 3.10. The Area Under the Curve (AUC) of each classifier model is depicted in the plots, which indicates the success of a model in detecting cyberbullying images.

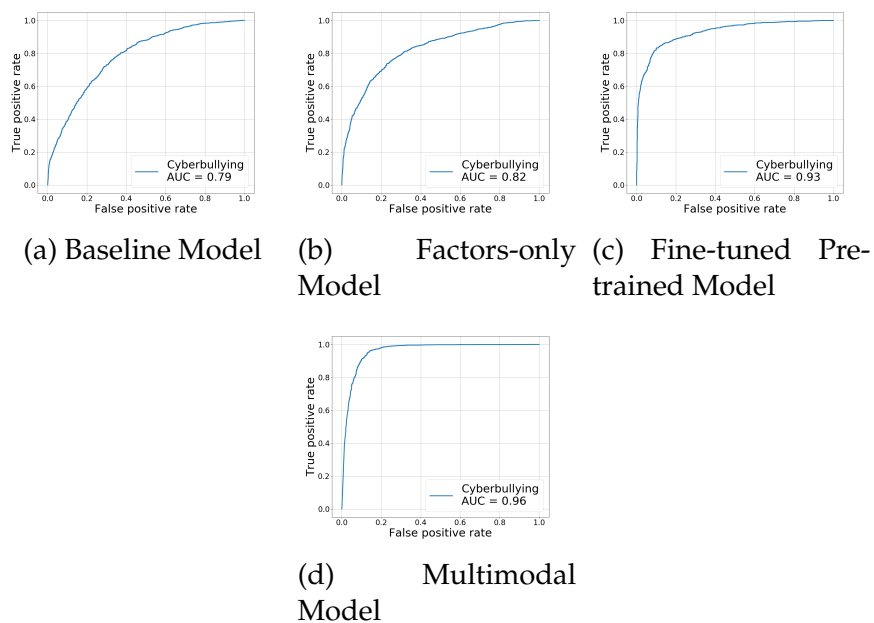


Figure 3.10: ROC analysis of classifier models.

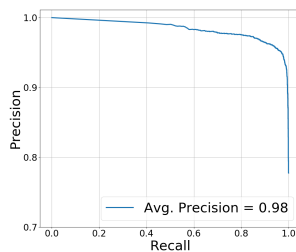


Figure 3.11: Precision-recall graph of the multimodal model.

Classifier Model	Accuracy	Precision	Recall
Baseline Model	77.25%	63.00%	29.68%
Factors-only Model	82.96%	79.34%	80.84%
Fine-tuned Pre-trained Model	88.82%	81.40%	73.70%
Multimodal Model	93.36%	94.27%	96.93%

Table 3.9: Accuracy, precision and recall of classifier models.

The TPR is a metric that represents how many correct positive results occurred among all positive samples available in the test dataset. FPR represents how many incorrect positive results occurred among all the negative samples available in the test dataset. These metrics are used in the ROC plots to analyze the performance of a model. We compute these evaluation metrics according to formulations in [133].

We find that the baseline model (Table 3.9, precision = 63.0% and recall = 29.68%) indeed has the lowest performance, indicating that cyberbullying in images is not a problem that can be trivially solved. Indeed, in our analysis, we find that cyberbullying in images is a highly contextual problem, which needs special investigation about its factors. From Figure 3.10a, a low AUC of 0.79 indicates that this model has a large number of false predictions.

Next, we investigate the factors-only model (Table 3.9, precision = 82.96% and recall = 79.34%). A better performance than the baseline model does indicate that even adding just the factors (without showing a model the original image) has quite powerful effect in classifying cyberbullying (Figure 3.10b, AUC = 0.82). Another observation we make about the factors-only model is that the recall is improved significantly, indicating that the identified visual factors do demonstrate the ability to distinguish the true positives (cyberbullying labeled images).

From our observations, the fine-tuned pre-trained model (Table 3.9, precision = 81.40% and recall = 73.70%) does not perform overall better than the factors-only model. Although the accuracy is higher, the recall of this model is significantly lower, which indicates that this model is not able to distinguish the cyberbullying images. On further examination, this model seems to be biased towards non-cyberbullying images, which could be attributed to our dataset

containing a significantly higher number of non-cyberbullying images compared to the cyberbullying images. Ideally, for good performance, we expect a model to have high precision and recall, and not just a high accuracy. We attribute the low performance of this model to the lack of the identified cyberbullying image factors. For example, a cyberbullying image portraying a person showing a gesture is interpreted by this model as just a person (since it is pre-trained). However, this model lacks the capability to distinguish that the person may be showing a gesture at the viewer.

Finally, we find that the multimodal classifier demonstrates the highest performance (Table 3.9, precision = 94.27% and recall = 96.93%) among the different classifier models. A high AUC (Figure 3.10, AUC = 0.96) is indicative of a good performance on the false positives and the false negatives. Note that this model is aware of the cyberbullying image factors identified in this chapter and also the low-level image features. A high precision and recall of this model indicates that the visual factors identified in this chapter are needed in order to distinguish especially the cyberbullying images. Due to the highly contextual nature of cyberbullying in images, the differences between such images and harmless images are very subtle. Therefore, we believe that the multimodal classifier demonstrates that our visual factors can be used to detect cyberbullying images accurately in real-world applications.

To interpret the model performance considering the unbalanced nature of our dataset, we depict the balance between the precision and recall in the case of the multimodal model in the precision-recall (PR) plot in Figure 3.11. The PR plot indicates that the multimodal model is able to correctly classify cyberbullying images with high precision.

Performance Overhead in Mobile Applications

Mobile phones play a major role in engendering cyberbullying in images, especially due to the on-board equipment, such as cameras, on these devices. Thus, our intention is that our models can be deployed on mobile devices to defend users against cyberbullying in images. To this end, we carry out an experiment to study the overhead of our model in a mobile application. We use the PyTorch Mobile framework [134] to deploy our multimodal model in an Android application, running in a Samsung Galaxy S5 mobile phone, with a memory capability of 256 megabytes. Note that we conduct this experiment on an older Android device in order to show that our model can be even run on weaker mobile devices. We are interested in measuring two types of overheads potentially introduced by running our model: (1) the model time, which is the time taken to execute a forward pass of our model; and (2) the render time, which is the time taken to resize an image according to the input dimensions needed by our model, and to render a warning message to the user if cyberbullying is detected in an image. To study the bearing of different sized photos, we measure these overheads with respect to the photo size. In this experiment, we randomly select 1000 photos from our test dataset and run them through the Android application with our model. We depict both the model time and the render time in Figure 3.12.

From Figure 3.12, we first observe that both the model time and the render time are mostly within the millisecond range, showing that it is indeed practical to adopt our models in mobile devices. We note that the size of the photo does not have any significant bearing over the model time and the render time, as we do not notice any effect of the size of image on the performance. We observe that

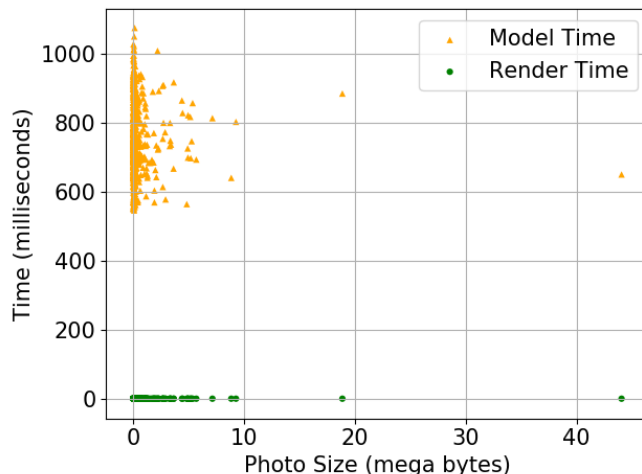


Figure 3.12: Overhead evaluation of the multimodal model integrated into an Android application.

the average model time is 753 milliseconds and the average render time is 0.06 milliseconds, both of which are sufficiently small. Thus, using the multimodal model in mobile devices only cause a minor overhead on the devices.

3.5.3 Deployment and User-based Evaluation

False Positives Evaluation on American Sign Language Dataset

Our analysis of cyberbullying factors in images reveal that hand gestures play a major role in carrying out cyberbullying. However, many harmless hand gestures, such as those used in the American Sign Language (ASL), are quite ubiquitous, and a concern with a cyberbullying model is that it may flag down such benign images as cyberbullying images. In this experiment, our objective is to conduct a false positive evaluation of our model on images from a publicly available ASL dataset [135]. Figure 3.13 depicts two samples from this dataset.

We run the multimodal model on all the test images of the ASL dataset (the ASL test dataset consists of 479 images). Our multimodal model correctly de-



Figure 3.13: Image samples from the ASL dataset.

tects all 479 images as non-cyberbullying images. This indicates that our model has learned to identify the harmful cyberbullying hand gestures, while the other hand gestures, such as the ones in the ASL dataset, are precisely detected as non-cyberbullying.

Validation of Cyberbullying Factors with a Wider Audience

In our work we introduce new factors of cyberbullying in images, as discussed in Section 3.4.2 We compile these factors by carefully observing the images labeled as cyberbullying by participants who take part in our data collection task. In this evaluation, we carry out a study to validate these factors with a wider audience. A sample of our study is depicted in Figure 3.14 in Appendix 3.5.3. In our study, we first show each participant, randomly selected image samples depicting a factor of cyberbullying, and ask the participant to input the factors, due to which the image samples have been reported as cyberbullying, in a free text box. By providing a free text box, we ensure that participants are not biased in any way by the factors compiled by us. Furthermore, we also provide participants the option to choose the images as non-cyberbullying thereby further

reducing any bias effects. We collect the free text responses for several cyberbullying images depicting different attributes of the cyberbullying factors. Asking participants to enter factors on their own allows the participants to think of factors by themselves without any bias and also allows us to validate our factors from a larger audience.

Our study was approved by our institution's IRB. We recruited 104 participants from Amazon MTurk for this study. Each task took about 10 minutes on average, and we paid a reward of \$2 for task completion. Three participants failed our attention check questions and two participants had entered the exact same text for all the images, and failed the attention check questions. After filtering out these five participants, we were left with 99 total participants in our study.

Next, we have to determine the factors from the free text entries that were entered by our participants. We identified the cyberbullying factors from participants' entries by mining them for text keywords and phrases pertaining to individual factors. For example, we used the words/phrases such as "pointed", "directed at me" and "aimed at me" to interpret that a participant is indicating that the body-pose of the person in the image is the cause of cyberbullying, and keywords like "gun", "pistol" and "firearm" to interpret that a participant is indicating that a threatening object, such as a gun, in the image is the cause of cyberbullying. We provide a full list of these words and phrases in Table 3.11 of Appendix 3.5.3. In the following, we discuss our findings from this study.

From the results of our study, the overall χ^2 [136] shows significant variation ($\chi^2(11) = 308.84, p < .0001$) among the 12 conditions (e.g., body-pose, gun, knife, middle finger, etc.) for the identified factors from participants' entries, indicating that different factors affected cyberbullying perception differently.

For the body-pose factor, we presented two samples to each participant. The first sample showed a person posing directly towards the viewer with a threatening object (e.g., Figure 3.14 in Appendix 3.5.3). The second sample showed a person posing away from the viewer with a similar threatening object. For the image sample with the person directly posing towards the viewer, 84.61% of participants who found this image as cyberbullying identified the factor to be the body-pose of the person in the image. For the image sample with the person posing away from the viewer, 72.41% of the participants found it to be non-cyberbullying, and none of the participants identified the body-pose of the person for this image sample. We think it is possible that the few participants who chose this image sample as cyberbullying could base their opinions on the threatening objects in this sample, although the body-pose of the person in the image is not correctly identified as a factor by all the participants. From the participants' entries, we found that they were most concerned that the image with the person posing towards the viewer is directly threatening the viewer by this pose, from responses such as *"Someone holding a gun and pointing it at the camera could be a direct threat to you"* and *"She is aiming a gun and when I look at the image it seems to be pointed directly at me"*. Thus, the participants have identified body-pose as a factor in the cyberbullying image.

Next, we discuss the results about the facial emotion factor in our study. In our study, each participant was shown an image sample based on facial emotions of joy, sorrow, anger and surprise. Overall only 9.43% of participants mentioned the facial emotion as a factor of cyberbullying, which is consistent with our finding in Section 3.5.2 that the facial emotion does not have a significant effect over cyberbullying in images. Thus, we believe that the facial emotion by itself is not a strong factor of cyberbullying images.

We then discuss the results about the hand gesture factor in our study. We showed each participant an image sample of a person showing the middle-finger, loser sign, and thumbs down hand gesture, all belonging to the hand gesture factor category. Overall 80.4% of participants discussed these hand gestures as factors of cyberbullying, with 97% of participants specifically mentioning the loser hand sign and 82.7% of the participants specifically mentioning the middle-finger sign as factors of cyberbullying in images. Thus, the participants have captured the hand gesture as an effective cyberbullying factor in images.

For the threatening object factor, we showed each participant image samples depicting gun, knife, and noose, which belong to the threatening object factor category. 88.29% participants discussed these threatening objects as the factor of cyberbullying. We conclude that the participants have rightly identified threatening objects as a strong factor of cyberbullying in images.

Lastly, we discuss the results of the social factor of cyberbullying in images. In this factor category, we showed an image sample of an anti-LGBT symbol. 89% of the participants identified this social factor for causing cyberbullying in images. We could observe that most participants consider this factor as a strong factor of cyberbullying in images. From this user experiment, we observed when the participants were provided free text boxes so that they can enter the cyberbullying factors by themselves, these factors identified by the participants were in agreement with the factors that we chose in our analysis.

Representativeness of Cyberbullying Images Dataset.

Cyberbullying in images is a complex phenomenon, and currently there are limited datasets available to study such a problem. Our cyberbullying images dataset takes a step closer towards understanding this phenomenon. In order

to make our dataset representative of real-world cyberbullying in images, we have asked participants to label cyberbullying images based on a very general guideline (Section 3.2.4, cyberbullying is “an act of online aggression carried out through images”). We carried out another study to compare the representativeness of the cyberbullying images in our dataset with another set of cyberbullying images [137]. The authors of [137] have shared their dataset of cyberbullying images with us. This dataset is composed of Instagram posts consisting of images and the associated comments, and the posts (i.e., the images and the associated comments together) are labeled by participants as cyberbullying or non-cyberbullying. We first filtered those cyberbullying posts, which were labeled as cyberbullying due to the content of images, so that we could filter out those posts that are only cyberbullying due to the associated comments. This left us with 316 images. Next, we used the same guidelines as used by us to label the images of the posts as cyberbullying. We recruited participants with the same criteria as in our annotations task from Amazon MTurk for this task, and used the same criterion for determining an image as cyberbullying. Overall, 31 images from their dataset were labeled as cyberbullying on their own. We conclude that their dataset predominantly needs the associated comments along with the images to be considered as cyberbullying, and the images on their own are mostly non-cyberbullying in nature. In contrast, our dataset contains a large number of images that are, on their own capable of causing cyberbullying, which indicates the images in our dataset are more representative cyberbullying images in the real world.

Capability Analysis of Existing Offensive Image Detectors

In this study, we focus on a deep analysis of the capabilities of state-of-the-art offensive image detectors with respect to the cyberbullying factors. Table 3.10 summarizes the capabilities of these detectors pertaining to the cyberbullying factors. In the following, we discuss in more detail about the capabilities of each detector and some observations related to the cyberbullying factors.

Factor	Google API	Yahoo Open NSFW	Clarifai NSFW	DeepAI	Amazon Rekognition
Body-pose	X	X	X	X	✓
Facial emotion	✓	X	X	X	✓
Hand gesture	✓	X	X	X	X
Threatening object	✓	X	X	✓	✓
Social	X	X	X	X	X

Table 3.10: Capabilities of state-of-the-art offensive image detectors with respect to cyberbullying factors.

We find that only Amazon Rekognition has the capability to detect body-pose. For example, it can indicate whether the person in an image is turned towards the viewer or at several angles from the viewer. Next, we find that both Google Cloud Vision API and Amazon Rekognition can detect the facial emotions of people in an image. The hand gesture factor is found to be detectable only by the Google Cloud Vision API. Although Google Cloud Vision API has this capability, we find that it only points out 40.61% of the cyberbullying images due to hand gestures as likely offensive. On a closer look, we find that the Google Cloud Vision API can not detect certain kinds of hand gestures, such as the loser sign that are prevalent in the cyberbullying images, as offensive.

We also find that Google Cloud Vision API, DeepAI, and Amazon Rekognition are capable of detecting threatening objects, such as guns and knives. We further study the detection capability of Google Cloud Vision API on two threatening objects, i.e., guns and knives. We observe that although Google

Cloud Vision API detects these objects in images, it flags down only certain such images as unsafe or offensive (42.58% of cyberbullying images with guns and 43.09% of cyberbullying images with knives). To analyze this observation further, we inspect the labels produced by Google Cloud Vision API on images with these objects. We observe that only images that had blood, wounds, or gore accompanied with an object are labeled as likely offensive by this detector. However, images with a visual cyberbullying object directly pointed at the viewer or a subject in an image, or the object brandished in a threatening fashion are missed by this detector. Besides, we find that all the existing offensive image detectors do not have the capability to detect the social factor of cyberbullying. Overall, we surmise that the detection capabilities of those existing offensive image detectors can be expanded based on the findings of our work.

User Study Interface and Keywords

#	Factor	Keywords
1	Body-pose	'point', 'direct', 'at me', 'at viewer', 'tell me', 'recipient', 'toward', 'aim', 'stance', 'posture'
2	Facial emotion	'joy', 'happy', 'smile', 'laugh', 'sad', 'sorrow', 'unhappy', 'angry', 'scary', 'mean', 'menacing', 'intimidating', 'shock', 'surprise'
3	Hand gesture	'middle finger', 'flip', 'flick', 'f*ck off', 'loser', 'L sign', 'thumbs down', 'gesture', 'hand sign'
4	Threatening object	'gun', 'firearm', 'pistol', 'knife', 'noose', 'rope', 'weapon'
5	Social	'lgbt', 'symbol', 'anti-pride', 'gay'

Table 3.11: Keywords used to identify factors.

The following photo has been reported as containing cyberbullying content by some Internet users. Observe the image carefully. Then in the box below, describe why you think it has been reported as containing cyberbullying content.





Image content that could be responsible for cyberbullying.

I don't think this image contains any cyberbullying content.

Figure 3.14: User study interface: participants are provided with a free text box to enter factors on their own.

In this study, Cyberbullying is "an act of online aggression or harassment carried out through images". If you think that the depicted image fits this description, categorize it as cyberbullying, otherwise categorize it as non-cyberbullying. Imagine the scenario where the depicted image is sent to you on your mobile device. Consciously put yourself in the scenario, and categorize the image as cyberbullying or non-cyberbullying.



Cyberbullying

Non- Cyberbullying

Figure 3.15: Interface of image annotation task.

Image Annotation Task Interface

3.6 Conclusion

In this task, we first analyzed the state-of-the-art offensive image detectors and found them to be inadequate for cyberbullying images detection. We collected a real-world cyberbullying images dataset that is representative of cyberbullying faced by social media users. We discovered five visual factors of cyberbullying in our dataset, and formulated a multimodal model based on those visual factors. Our evaluation of our model shows that our model effectively detect visual cyberbullying in images.

Chapter 4

Detecting and Explaining Traditional Online Hate Speech

The social and economic destabilization caused by COVID-19 has produced a range of emotions in people, including fear, anxiety, and even hostility. Notably, COVID-19-related hate speech is increasingly occurring on social media that target people based on race/ethnicity, age, social class, immigration status and political ideology. For instance, Asian Americans are frequent targets of hate speech related to COVID-19, with derogatory terms for the disease, such as “kung flu” and “chop fluey”, shared more than 10,000 times on Twitter during March alone [10]. Meanwhile, the phrase “Boomer Remover”, a callous nickname for COVID-19 used to mock the high mortality rate among older people infected with the disease, has been shared more than 65,000 times on Twitter [11]. Moreover, a recent report on online toxicity found a 900% increase in hate speech towards China and Chinese people on Twitter [12], and traffic to sites and posts that target Asians over COVID-19 has skyrocketed.

This recent wave of COVID-19-related hate speech has given rise to novel vo-

cabularies and jargon that are used by Internet users to specifically target certain communities. While current social media platforms such as Twitter and Facebook are quite well equipped to detect hate-speech concerning traditional issues [138], they are not capable of addressing the new jargon related to COVID-19. Thus, there is a need to discover these novel jargon with respect to COVID-19-related hate speech. However, Internet users often find innovative ways to use such jargon [139, 140], in order to hide their true meaning (e.g., “xinpigs”, “thankschina”), due to which they cannot be discovered in a straightforward manner. Thus, new strategies based on deep analysis of such texts need to be formulated to summarize such jargon by discovering the keywords that are related to them.

The detection of online hate speech should be accompanied with a strong control strategy so that Internet users can be deterred from posting such texts. User warnings and word removal recommendations [13, 14] are often used to implement such control mechanisms. However, merely asking users to remove hate-related keywords is not a strong enough control strategy, as users often come up with alternate ways to post such texts by surpassing the detection mechanisms. Moreover, the other words in a text that are semantically related to such keywords (such as names of individuals or group) can still significantly harm the targeted individuals or groups. Therefore, a control strategy that can systematically point out these semantically related words is very important for effectively controlling these instances of hate speech.

The new wave of hate speech related to COVID-19 is unique because, unlike traditional forms of hate speech that are typically rooted in deep-seated animosity, hate speech linked to the COVID-19 outbreak is spontaneous, induced by fear, anxiety, and stress resultant of a rapidly-changing reality. Previ-

ously, to understand why identity-based hate speech is becoming increasingly common online [141], sociologists and criminologists have explored the roles of strain and threat in fostering such attacks. While some works [142] theorize that deviant behavior stems from a disjuncture between culturally-valued goals, others show that financial strain, such as strain caused by unemployment/underemployment and low wages, can indeed engender harassing behavior towards immigration groups [143, 144, 145]. While fear prompted by the pandemic might trigger long-held prejudice towards certain groups, such as Asian Americans or immigrants, it is unlikely that hate-speech based on age or socio-economic status is similarly an expression of embedded bias. Thus, more information on COVID-19-related hate speech is needed to better understand its impetuses.

In this chapter, we propose a novel approach to discover new keywords linked to COVID-19-related hate speech and the word associations to effectively implement its control. We collect a new dataset (Boomer-hate dataset) of tweets targeting old people and supplement this dataset with an existing COVID-19 dataset (Asian-hate dataset) targeting Asian American community [21]. We then train a BERT (Bidirectional Encoder Representations from Transformers) model [44] to classify tweets as Hate Vs. Non-hate. Based on the analysis of BERT attention mechanism, a transformer model [45] based on attention, we develop an approach to discover new keywords (186 keywords targeting the Asian community and 100 keywords targeting older people) related to COVID-19. For implementing effective control, we develop a strategy based on the attention attributed to these keywords by other words in a tweet, so that all sensitive words in a tweet can be censored or reconsidered. We then undertake an exploratory analysis of COVID-19-related hate speech and find that most

of such high-impact, long distance attentions are learned in the earlier layers of the BERT model (layers 2 to 7 for Asian-hate dataset) or later layers (layers 10 and 11 for Boomer-hate dataset) depending on the underlying data distribution. Our study also makes an important finding that in the case of Boomer-hate dataset, the BERT model makes predictions based on the association of hate keywords and targeted groups or individuals, a finding that is inline with existing hate-speech research. Our finding paves the way for deep analysis of BERT for detection of hate-speech as well as explaining BERT (known as BERTology), a largely unexplored research area concerning BERT.

Our contributions are summarized as follows:

- **New Dataset of COVID-19-related Hate Speech Against Old People.** We collect a new dataset of COVID-19-related hate speech against old people. Our Boomer-hate dataset consists of 388 hate tweets and 1358 non-hate tweets from 1401 Twitter users. We will make our dataset publicly available for further research. In our work, we supplement our own dataset with another publicly available dataset [21] pertaining to COVID-19-related Asian hate, so that our study covers a broad spectrum of hate speech witnessed during COVID-19.
- **COVID-19-related Hate Speech Keywords Discovery.** We first train a BERT model on the datasets to learn Hate Vs. Non-hate speech. We then develop an approach based on BERT attention mechanism, to discover the most attended-to keywords that are responsible for causing hate in hateful tweets. We discover 186 keywords related to Asian-hate and 100 keywords related to Boomer-hate using our approach. For effective control of hate speech, we use our approach to find the words that significantly attend

to the hate keywords so that they can be presented to users for removal or reconsideration. The new keywords discovered by our approach are an important resource for further hate-speech research, and we plan to submit them to a popular online hate keywords repository ¹.

- **Exploratory Findings About COVID-19-related Hate Speech.** Our exploratory findings specifically concerning BERT and hate-speech detection sheds light on the inner-workings of the BERT model, using which we can identify if the model uses specific word associations only to detect hate speech, or uses a more complex association of words. We find that the high impact attentions regarding hate speech are learned in the earlier layers of the BERT model in case of Asian-hate and later layers in case of Boomer-hate, and that BERT seems to be associating hate-related keywords and groups or individuals for hate-speech predictions for Boomer-hate.

4.1 Data Collection Methodology

In our study, we collect a timely dataset of tweets from Twitter related to COVID-19-related hate speech against old people. We then supplement this dataset with an existing dataset [141] of COVID-19-related hate speech against Asian American community. We use this combined dataset to study online hate speech associated with COVID-19 on Twitter.

Collection Methodology. We adopted a keyword-based approach to collect COVID-19 tweets against old people using an online Twitter data collection tool ². We used the keywords “boomer” with COVID-19 related keywords such

¹<https://hatebase.org/>

²<https://github.com/Jefferson-Henrique/GetOldTweets-python>

as “Coronavirus” and “Covid-19” to search for such tweets. We restricted the tweet collection to English language only. Using these keywords, we collected 28,827 tweets between December 2019 and June 2020 from 1401 Twitter users. Figure 4.1 shows the percentage of tweets related to COVID-19 hate speech against older people and the date ranges they were searched in. Since the date ranges prior to Feb 24, 2020 yielded very low tweets, we have ignored those date ranges. It can be seen in Figure 4.1 that the majority of the tweets linked to COVID-19-related hate speech against old people were found in March, 2020. We note that this may be the time, during which the adverse effects of the pandemic on older individuals were brought to light that could have triggered the spike in the hate-related tweets during this time.

Boomer-Hate Dataset. Since there are no ground truth labels of COVID-19-related anti old people hate tweets, we asked two experts in our research team to label the collected tweets. We first cleaned the tweets based on sentiment polarity and removed the tweets that are neutral sentiment using Python NLTK library ³. Existing studies of hate speech from the social science literature [146, 147] have shown that hate speech is directed at an individual or group based on “an arbitrary or normatively irrelevant feature”, and that it casts the target as an “undesirable presence and a legitimate object of hostility.” We used a similar definition for our annotation task: (a) has one or more COVID-19-related keywords, (b) is directed towards an individual or a group of older people (Boomers), and (c) is abusive or derogatory.

The two experts labeled all the tweets in the dataset, which results in 388 hate-speech related tweets and 1358 non-hate-related or neutral tweets.

Asian-Hate Dataset. We used a publicly available dataset [141] of tweets

³<https://www.nltk.org/>

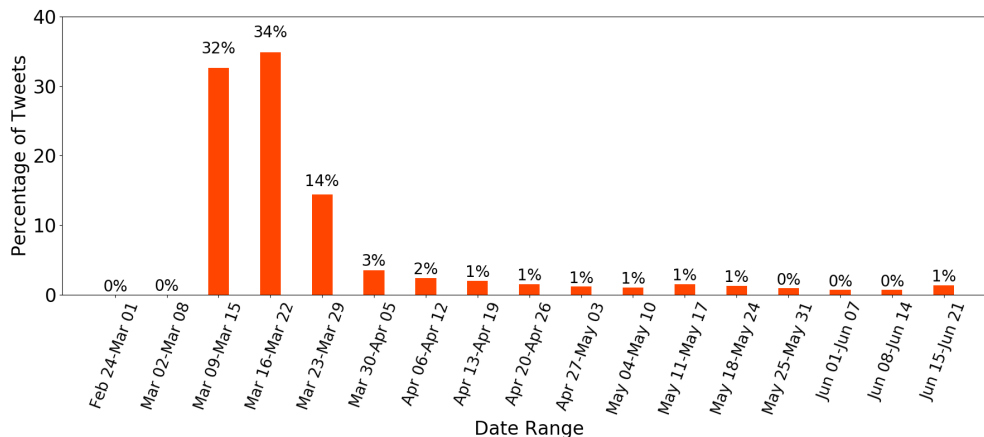


Figure 4.1: Percentages of tweets collected according to date ranges. All date ranges belong to the year 2020.

aimed at COVID-19-related hate speech against the Asian American community. This dataset contains 2,319 labeled tweets, with 678 of them labeled as hateful tweets.

4.2 Background

In this chapter, we focus on the BERT model [44], a large transformer [45] network. Transformers consist of multiple layers where each layer contains multiple attention heads. Each attention head takes as input a sequence of vectors $h = [h_1, \dots, h_n]$ corresponding to the n tokens of the input sentence. Each vector h_i is transformed into query, key, and value vectors q_i, k_i, v_i through separate linear transformations. The head computes attention weights α between all pairs of words as softmax-normalized dot products between the query and key vectors. The output o of the attention head is a weighted sum of the value vectors, and α_{ij} represents a dot product between the query and key vectors, expressed in Equation 4.1 below.

$$\alpha_{ij} = \frac{\exp(q_i^T k_j)}{\sum_{l=1}^n \exp(q_i^T k_l)} \quad o_i = \sum_{j=1}^n \alpha_{ij} v_j \quad (4.1)$$

The attention weights can be interpreted as controlling the importance of every other token when learning the next representation of the current token.

BERT is trained using the “masked language modeling” strategy over billions of data samples, and more details about the training process can be found in [44]. An important detail about BERT training is that a special token [CLS] is added to the beginning of the text and another token [SEP] is added to the end, so that multiple sequence inputs can be trained together.

4.3 Study Methodology

On a high level, our study is focused on studying the attention mechanism of BERT models to find important patterns about COVID-19-related hateful tweets. Since BERT is based on attention mechanism, the model learns the attentions between different tokens in all the tokens of an input sequences. This provides us a powerful tool to analyze linguistic associations in the dataset that BERT is trained on. Our work leans on the exploratory research side of BERT (known as “BERTology” [148, 149]). We first train a 12 layer, 12 attention heads “bert-base-uncased” model [45] on our dataset (we use 90% for training and 10% for testing). In the following sections, we analyze the BERT model trained on the hate datasets, spanning several layers and attention heads to formulate hate-speech control strategies and draw important observations about how BERT detects hate speech.

4.3.1 Keywords Discovery from BERT Attention Mechanism

The first objective of our work is to find new keywords of hate-speech from the two datasets (Asian-Hate and Boomer-hate datasets). In this section, we discuss our approach for discovering these keywords and our findings regarding the keywords found in the two datasets. In this experiment, we evaluate the words that are most attended to, by the fine-tuned BERT model in each layer. To achieve this, we aggregate the attention on each token of an input sequence by all attention heads in each layer, as given below in Equation 4.2.

$$Aggr^l(o_i) = \sum_{h \in H} o_i^h \quad (4.2)$$

In the Equation 4.2, H refers to the attention heads in each layer of BERT model and o_i refers to the attention weight of a token in an input sequence. For each layer, we take the top-k ($k = 5$) tokens as potential keywords. We do not consider tokens that are not split by the BERT word-piece tokenizer to reduce words normally occurring in English dictionary. We further remove those words that are not part of a sentence ⁴. A summarized list of discovered keywords are depicted in Table 4.1.

In our analysis of Table 4.1, we found several new keywords used to propagate hate speech with respect to COVID-19-related Asian-hate and Boomer-hate. In the Asian-hate dataset, we found that BERT attributes the most attention to keywords that are a combination of word-pieces related to Asian community (e.g., “chin”) and word-pieces related to the COVID-19 pandemic (e.g., “virus”), giving rise to keywords such as “chinkvirus” and “wuhanflu”. In the Boomer-hate dataset, we found that certain keywords followed a similar pat-

⁴We use Python NLTK library’s POS tags

Table 4.1: Summarized list of sample keywords in the datasets, most attended to by BERT model.

Dataset	Top Keywords
Asian-hate Dataset	chinkvirus, wuhanflu, chinesebioterrorism, chineseviruscorona, chinaliedhdeperiencedied, wholiedpeople, chinamustexplain, nochinainfluenceonamerica, wuhanhealthorganisation, abioweaponslab, fuckchina, chinesebiologicalchemical, ccpvirus, prisonplanet, makechinapay, neverforgetneverforgive
Boomer-hate Dataset	boomerremover, gaslighters, corbid, 60sfolks, boomerdeath, karen, hitler , headassery, thankyouboomer, yoof, deletus, boomermoover, michiganders, entomber, boomerentomber, komekko, doubledowndonnie, boomerdoomer, coronachan, socialistremover, oldaf, immunocompromised, thintheherd

tern of word-pieces related to older people (e.g., “boomer”) and word-pieces related to derogatory terms (e.g., “remover”), giving rise to keywords such as “boomerremover”, but certain keywords did not necessarily follow any particular pattern, but seemed to be more contextual in nature (e.g., “karen”, “oldaf” and “deletus”). We also found some keywords that were completely new, that were simply derogatory to older individuals (e.g., “yoof” refers to the way an older person may pronounce “youth”). These findings may indicate that while users follow a particular pattern in the Asian-hate tweets, on the other hand users seem to adopt more complex and varied techniques in the Boomer-hate tweets.

Next, in order to study how these keywords are learned in each BERT layer, we analyze the attention given to these keywords by each layer of the BERT model. We recall that the BERT model used in this chapter has 12 layers of multi-headed attentions. In this study, we analyze the keywords that are most

attended to in each BERT layer. The Table 4.2 shows the top-k (k=10) most attended keywords in each BERT layer, normalized across all attention heads. We did not find any apparent pattern which indicated that particular keywords could be receiving more attention in certain layers. Existing research in BERTology such as [149] suggest that certain layers of BERT may be focusing on different word associations. Therefore, we further analyzed the layers from this perspective. We focused on long-distance attentions in each layer based on the attention on multiple tokens, as given by Equation 4.3.

Table 4.2: Top-k (k = 10) keywords attended to in each layer of BERT model.

Layer #	Top-k Keywords
Layer 1	coronavirus, chinesevirus, wuhanvirus, chinavirus, ccpvirus, wuhancoronavirus, chinesevirus19, chinese coronavirus, coronavirusoutbreak, chinaliedpeopledied
Layer 2	coronavirus, covid19, chinavirus, chinesevirus, wuhanvirus, chinaliedpeopledied, realdonaldtrump, covid2019, xijipingvirus, chinesevirus19
Layer 3	chinaliedpeopledied, chinaliedpeopledie, fuckchina, covid19, coronavirus, wuhanvirus, chinesevirus, chinese, racismisavirus, chinavirus
Layer 4	chinaliedpeopledied, coronavirus, covid19, fuckchina, chinesevirus, chinaliedpeopledie, wuhanvirus, chinavirus, ccpvirus, chinesevirus19
Layer 5	covid19, chinaliedpeopledied, chinesevirus, coronavirus, chinavirus, wuhanvirus, chinesevirus19, ccpvirus, fuckchina, covid2019
Layer 6	chinaliedpeopledied, chinesevirus, coronavirus, chinavirus, covid19, wuhanvirus, chinaliedpeopledie, ccpvirus, fuckchina, chinesevirus19
Layer 7	chinesevirus, coronavirus, chinaliedpeopledied, wuhanvirus, chinavirus, covid19, fuckchina, ccpvirus, wuhancoronavirus, chinaliedpeopledie
Layer 8	coronavirus, chinesevirus, chinaliedpeopledied, wuhanvirus, fuckchina, chinavirus, covid19, ccpvirus, wuhancoronavirus, chinaliedpeopledie
Layer 9	chinaliedpeopledied, coronavirus, chinesevirus, fuckchina, wuhanvirus, chinavirus, covid19, ccpvirus, chinaliedpeopledie, racismisavirus
Layer 10	chinaliedpeopledied, coronavirus, fuckchina, covid19, chinesevirus, chinavirus, chinese, chinaliedpeopledie, racismisavirus, chinesevirus19
Layer 11	coronavirus, covid19, chinaliedpeopledied, fuckchina, chinesevirus, chinavirus, wuhanvirus, chinese, ccpvirus, chinaliedpeopledie
Layer 12	chinesevirus, coronavirus, chinaliedpeopledied, covid19, wuhanvirus, chinavirus, ccpvirus, chinaliedpeopledie, racismisavirus, chinesevirus19

$$D = \frac{\sum_{i=1}^N \sum_{j=1}^i \alpha_{ij}(x) \times (i - j)}{\sum_{i=1}^N \sum_{j=1}^i \alpha_{ij}(x)} \quad (4.3)$$

The Equation 4.3 determines attention spanning across tokens, normalized by their distances (i and j are indices). Therefore, higher attention tokens farther

#	Original Tweet	Keywords
1	some chinese are horrible as fuck chinaliedpeopledie boycotchina wuhanvirus	chinese, chinaliedpeopledie, boycotchina, wuhanvirus
2	itsing6 spokespersonchn fuck ccpvirus chinesevirus	fuck, ccpvirus, chinesevirus
3	h*****f ****l s*****d fuck off commie chinaliedpeopledied fucktheccp	fuck off, commie, chinaliedpeopledied, fucktheccp
4	5g does fuck u ask the kungflu	fuck, kungflu
5	it'll be the only party left come november boomerremover	boomerremover
6	magkovid unta it incompetent NA senators they called the virus a boomer remover for a reason	magkovid, boomer, remover

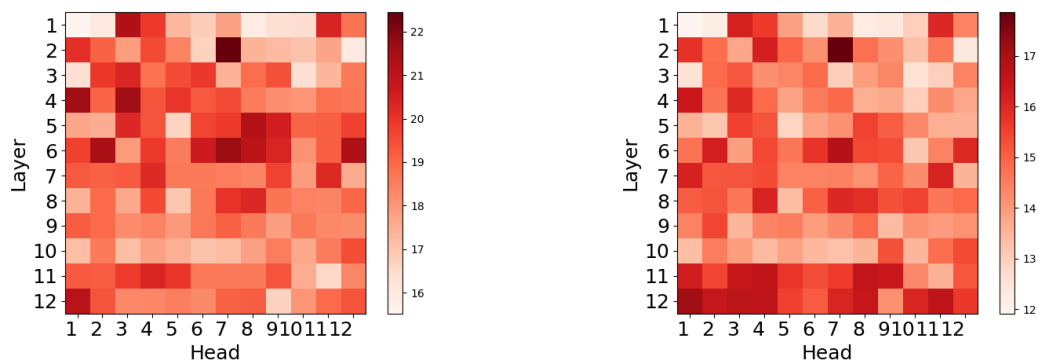
Table 4.3: Samples of control strategy.

apart would have higher distance attention. We computed this metric for each attention head in a layer and the result is depicted in Figure 4.2, which depicts a heat-map of the attention distance for each head in each layer for the two datasets.

From Figure 4.2a which shows the results for Asian-hate dataset, we can observe that the attention distance in earlier layers (layers 2 to 7) are higher (depicted by darker color). This could indicate that the hate-related attentions for Asian-hate spanning across tokens are predominantly learned in the earlier layers of the BERT model.

On analyzing the Figure 4.2b which depicts the results of this experiment for Boomer-hate dataset, we observed a different result, which may indicate that in this case, the long distance attentions are learned in later layers of the BERT model, with layers 11 and 12 showing overall higher mean attention distances. This observation could be due to the fact that the hateful tweets in the Boomer-hate dataset seems to be significantly correlated to a few, specific keywords (e.g. “boomer” and “remover”). Another explanation of this observation could be that the BERT model may be dynamically learning these associations according to the underlying distribution of the training data.

We observed that in the later layers, most attention is given to certain words



(a) Attention distance by layer and head in the Asian-hate dataset.

(b) Attention distance by layer and head in the Boomer-hate dataset.

Figure 4.2: Attention distance in the two COVID-29 datasets.

or phrases, and also to the start and end tokens (“[CLS]” and “[SEP]”) of the BERT tokenizer. Therefore, in COVID-19 related hate tweets, the attentions in earlier or later layers can be studied to understand the word associations in such tweets, depending on the distribution of the training data.

4.4 Implementation and Evaluation

In the following, we discuss the implementation of our approach and evaluate it by running it on the Asian-hate and Boomer-hate datasets to perform control, and examine if BERT detect in detecting hate speech based on existing definitions of hate.

4.4.1 Implementation

Our approach has been implemented as a Pytorch [130] model. We use the pre-trained BERT model provided by Huggingface [150]. Experiments have been performed on NVIDIA V100 GPUs.

4.4.2 Hate Speech Control with BERT Attention

We utilize the results of the previous section to formulate a control strategy for COVID-19-related hate-speech using BERT attention mechanism. We use the attentions given to the keywords discovered in Section 4.3.1 by other words in a sequence, in the layers found to have long distance word associations (from Figure 4.2a and Figure 4.2b). Since these other words contribute to the hateful context in an input sequence, these words must also be pointed out for removal or re-consideration. We then propose to a user to re-consider sending such words or changing these words.

Existing studies on BERT attention mechanism [149, 148] suggest that the attention formulation in Equation 4.1 prioritizes tokens with higher dot product vectors. Hence, the attention mechanism of BERT can be used to find other words in a tweet, that attend to the hateful keywords. In this chapter, we use this phenomenon to find the top other words that attend the most to the hateful keywords. Table 4.3 depicts randomly selected samples from the hate datasets with hateful words and keywords highlighted.

In a real-world system, we propose a control strategy in which a tweet posted by a user is run through our model to detect any hate content. If any hate content is detected in the tweet, keywords discovered in our work can be searched in the tweet. If any of the keywords are found, our strategy of finding other words that significantly attend to these keywords can be presented to the user for removal or reconsideration, along with the hateful keywords.

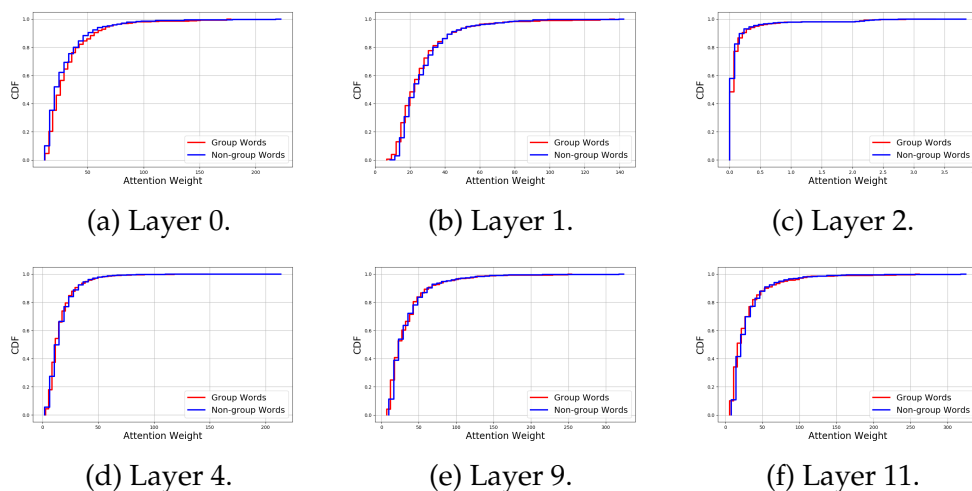


Figure 4.3: Attentions to Target words Vs. Non-target words in case of Asian-hate.

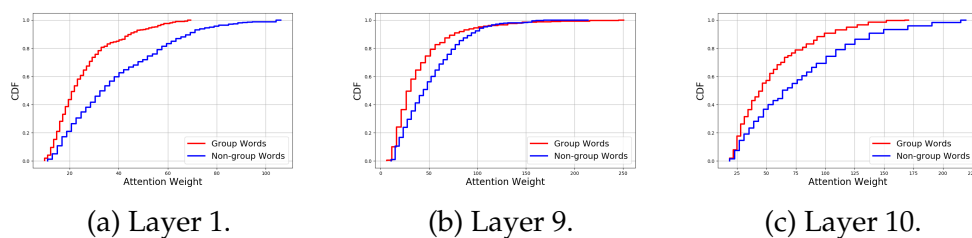


Figure 4.4: Attentions to Target words Vs. Non-target words in case of Boomer-hate.

4.4.3 Is BERT Detecting Hate Speech based on Existing Definitions of Hate?

Several existing studies [151, 152, 153] suggest that hate-speech targets disadvantaged social groups in a manner that is potentially harmful to them. From a broader perspective, these disadvantaged groups could also be individuals, who could be targets of hate speech. Our objective in this experiment is to study whether the BERT model implicitly detects hate-speech based on such existing definitions of hate-speech from literature.

We first identify the words that pertain to the targets of hate-speech in both the COVID-19 datasets. We consider both groups (e.g. “Chinese”, “Seniors”) and individuals (e.g. “Xi Jinping”) as targets for this experiment. Some samples of the chosen target words are depicted in Table 4.4.

Target	Samples
Groups	han, chinese, chinese-tourists, taiwanese, libs, babyboomers, magats, muslim, jews, asians, koreans, african, africans, christians, indians
Individuals	spokespersonchn, jinping, trump, jackma, pompeo, boris, potus, chr*****, m*****, g*****8

Table 4.4: Samples of words chosen as targets. Username identifiers have been removed to preserve user identities.

Our objective is to study to what extent BERT model may be using associations between hateful keywords and such targets words to detect hate-speech. We base our study on the attention that these keywords may be attributing to these target words. If the model is learning to pay higher attention to the target words from the keywords (corresponding to higher attention weights) than the non-target words in a tweet, this could indicate that the BERT model strongly uses these associations to detect hate-speech. For each tweet in both the COVID-19 datasets, we capture the attention weights from the the hateful keywords to the target words such as the ones in Table 4.4. We then plot the CDF of such attention weights for certain layers for both the Asian-hate and the Boomer-hate datasets. Our results are presented in Figure 4.3 and Figure 4.4, respectively for the Asian-hate dataset and Boomer-hate dataset.

In the Asian-hate dataset results depicted in Figure 4.3, we plot the CDF for layers 0, 1, 2, 4, 9 and 11 for target words (depicted by red curve). We chose

these layers so that we have good representation from all depth levels and also from our result from Section 4.4.2 that for this dataset, longer distance association may be formed in the earlier layers. For comparison, we also plot the CDF for non-targets words (depicted by blue curve) occurring in the tweets, which are ordinary words. We found that for this dataset, the BERT model seems to pay similar attention for keywords and target/non-target words. While preliminarily this may indicate that BERT does not learn well to associate keywords with target words, we found that BERT learns the subtle differences between hate and non-hate tweets (e.g., “chinese get out” and “stop telling chinese to get out”), based on associations between keywords and both target words and non-target words. Our analysis of the Asian-hate dataset led to the observation that although the keywords and target words are themselves not hateful, their associations could be hateful in hate tweets. In order to make correct detection, the BERT model seems to learn the associations between these two kinds of words in conjunction with the other non-target words in the tweet to make accurate predictions. Thus, we observed that BERT does form association between hate keywords and target words, however it does not only depend on these associations to make predictions, which may be the reason why BERT is found to be more powerful than other sequence models such as recurrent neural networks.

Next, we analyze the Boomer-hate dataset using the same procedure described above. The results of our experiment on Boomer-hate dataset is depicted in Figure 4.4. We found the results on this dataset to be quite different from the results in the case of Asian-hate dataset. In this case, the BERT model seemed to be associating more strongly between the hateful keywords and the target words (depicted by red curve), when compared to the non-target words (depicted by blue curve). For example, in Figures 4.4a and 4.4c, we can see clearly,

the observation that association between target words and hateful keywords are given a lot more attention than the non-target words. Even in Figure 4.4b (a later layer with more distance associations, Section 4.4.2), this trend seems to be visible.

Upon further investigation, we observed that this behavior could be due to the reason that the Boomer-hate dataset is more sparsely containing hateful keywords and the target keywords. For example, in the case of Asian hate, we observed a lot of different targets ranging from groups (e.g., “chinese”, “taiwanese”, “asians”) and keywords (e.g., “kungflu”, “wuflu”, “wuhanvirus”). However, in the case of Boomer-hate we found relatively fewer number of such words, as the target is mostly singular (older people only) and the hate keywords therefore, are also quite limited. Hence, we observed that in such cases, where a less varied patterns need to be learned by BERT model, it depends more on learning association between certain words than learn more subtle and varied associations.

4.5 Conclusion

In this task, we focused on the COVID-19 related hate speech detection and control using the attention mechanism of BERT. We discovered several novel keywords of online hate speech in Asian-hate and Boomer-hate datasets, and identified important findings about how BERT detects Asian-hate and Boomer-hate. We then used our approach to control hate speech online, by pinpointing the exact words or phrases that are responsible. Our evaluation shows that our approach is able to effectively detect and control online hate speech.

Chapter 5

Towards Understanding and Mitigating New Waves of Online Hate

Online hate negatively transforms our online and offline societies. While hate speech has existed as a critical social issue from quite some time, recent advancements in Internet and social media platforms have led to a massive rise in online hate. In a recent Pew survey [1], roughly four in ten (i.e., 41%) Americans reported personally experiencing varying degrees of harassment and bullying online, and Internet users all over world (i.e., 48%) have also reported having similar experiences [2, 3]. Furthermore, the odds of users experiencing abuse have increased by 1.3 times over the past three years and young adults aged 18–24 and vulnerable communities such as LGBTQ+ reportedly face heightened levels of risk [2].

Online hate is not a static problem. It is highly influenced by global events and the changing technological landscape. For example, recent polarizing events

such as the COVID-19 pandemic [4], the 2020 presidential elections [30] and the Black Lives Matter (BLM) [31] protests have shown how emotions of fear, uncertainty, and anxiety involved in these episodes can set-off new spikes in unprecedented online hate [32]. As an instance, the new waves of anti-Asian hate [4, 5], mask-related hate [6, 7] and vaccine-related hate [8, 9] set-off by the COVID-19 pandemic have had a devastating effect on our society globally. As our cyberspaces move into the future consisting of advanced technologies such as Web 3.0 [15], augmented reality [16] and the Metaverse [17], online hate is bound to take on new, more sinister shapes. Thus, efforts to effectively counter such new eruptions in online hate must be taken immediately.

The enormous eruptions of new online hate waves and their increasingly complex landscapes have unfortunately not induced a corresponding improvement in their detection capability, and existing online hate detection systems have consistently lagged behind in flagging down new hateful content. For example, the recent waves of anti-Asian hate [4, 5], mask-related hate [6, 7] and vaccine-related hate [8, 9] encountered during the COVID-19 pandemic could not be sufficiently contained by online hate moderation tools deployed in online social networks (OSNs), as a result of which online hate against minority communities and other vulnerable groups spread unabated during this period. While these same detection system seemed quite effective in controlling traditional online hate such as violent extremism [33, 34] and trolling [35], they were found struggling to stop the recent, new waves of online hate [36].

To understand why existing detection systems have not been able to keep pace with the problem of new waves of online hate, the gap between the detection paradigm employed by these systems and the new online hate paradigms should be contemplated. Existing detection systems [48, 103, 43, 104, 42] are

largely based on supervised artificial intelligence and machine learning (AI/ML) models that are trained on large datasets [154, 153, 155] of hate speech collected from OSNs, that are traditionally text-based. *First*, a limitation of this paradigm is that these models are static, i.e., they are applicable to only traditional contexts of online hate. However, the context of new waves of online hate rapidly changes. To address new waves of online hate, we need new systems that can learn from the traditional contexts of online hate, and apply the learned knowledge to new contexts of online hate. *Second*, the existing detection systems need large datasets to be trained sufficiently, since they use supervised learning paradigm. However, large datasets of new waves of hate are unavailable and it is not feasible to collect large datasets in a timely manner. Thus, new learning paradigms that can sufficiently address this problem with a few training samples need to be investigated. *Third*, since perpetrators represent new hate waves in many different formats such as text, images and videos, existing methods that are based on text alone cannot be used for other representations. New approaches that can incorporate different representations of hate need to be investigated.

In this chapter, we aim to practically address the problem of new waves of online hate, by studying it, understanding its challenges, and formulating automatic systems that can detect it. Our intuition, informed by previous studies [5, 7, 8] and reports [46, 2], is that new waves of online hate are characterized by rapidly changed contexts. We first report a systematic study on the phenomenon of new online hate waves, by collecting a large dataset of 3312 hateful users and their 4042454 tweets on Twitter, and studying their tweeting behavior before and after the COVID-19 pandemic. We find that before the pandemic, the tweeting behavior of these hateful users were related to traditional hate con-

texts, which completely changed into online hate related to new contexts post pandemic. We also found that these users were increasingly using newer representation techniques, such as images and memes to spread hateful content. Next, we conducted a large scale study of the effectiveness of state-of-the-art, existing systems of hateful content detection such as Perspective API [47], Google Cloud Vision API [48] and MMBT [49] on datasets of COVID-19-related 1,679 tweets, and found that these detectors are severely limited (average F1 score of 0.31) against new waves of hate tweets. We then identify key challenges to the timely and effective intervention of new waves of online hate: (i) learn knowledge from traditional hate contexts and apply learned knowledge to new contexts, (ii) training with just a few samples of new hate contexts, and (iii) the need to support multiple representations of online hate.

We introduce our framework, Attribute-based Zero-shot Multimodal Learning (AZL), that can detect new waves of online hate by addressing each of those challenges. AZL uses an attribute-based learning methodology [50] to transfer important knowledge about traditional hate contexts to the detection of new hate contexts, uses Zero-shot learning [51] to effectively classify new hateful contexts with just a few training samples, and uses Multimodal representation techniques [156] to incorporate different representations of hateful content. We evaluate AZL from several different perspectives, and found that our framework achieves state-of-the-art-detection average F1 score of 0.72 on new hate contexts, such as Asian (76.52%), mask (67.47%), vaccine (70.73%) and boomer (72.34%) related hate..

Our paper makes the following contributions:

- **New understanding about the nature of new waves of online hate.** We

report the first systematic study on the nature of new waves of online hate, and the effectiveness of existing hateful content detection systems on new hate contexts. Our study, focusing on the COVID-19 pandemic as a case-study, sheds light on how nature of hate rapidly changes, with rapidly changing contexts, and changing representations. Our study shows how these new forms completely evade existing, state-of-the-art techniques of hateful content detection that are used in real-world systems for hate moderation. Our studies highlight the gap between these new waves of online hate and the detection capabilities of existing systems. Furthermore, our studies emphasize the need for a paradigm shift in the way we approach the issue of practically addressing hateful content moderation.

- **New framework for detecting new waves of online hate.** We developed a novel framework called AZL to detect new waves of online hate. AZL is designed to address the challenges of detecting new hate waves. Our framework uses attribute-based learning to use transferable knowledge from traditional hate contexts to detect new hate contexts, zero-shot learning to detect new hate contexts using only a few samples, and multi-modal representation learning to address different representations of online hate. Our framework takes a first step toward more effective control of the emerging threat of new waves of online hate.
- **Multi-faceted evaluation of AZL on 4 new hate contexts and 2 different representations.** We evaluate our framework on four new contexts of hate encountered during the COVID-19 pandemic, such as anti-Asian hate, hate related to mask, vaccine-related hate and hate towards older individuals (a.k.a “Boomer” hate) and two different representations, i.e.,

tweets and memes. Furthermore, we also evaluate our framework from several different perspectives, such as ablation studies based on the different components of our framework, performance in real-world settings and extensive comparisons against existing real-world systems. Results demonstrate AZL to be highly effective on new waves of online hate.

5.1 Examining New Waves of Online Hate

The phenomenon of hate is not new. Earlier generations have known this phenomenon in other forms, such as hateful speeches and news articles. With the coming of the digital age, hate found a new platform, in terms of the Internet, and online hate emerged has a critical issue. As the society and technological innovation evolves, online hate takes new shapes. Recently, online hate was a major damaging effect of the COVID-19 pandemic globally, highlighted by several media organizations [10, 12] and research works [4, 157]. In this section, we studied the nature of the news waves of online hate considering the COVID-19 pandemic as a case-study. In the following, we present our study that illustrates how new waves of online hate contexts emerge after traditional ones, and how new representations could play a major role in these news waves. Following the study, we scrutinize the performance of existing state-of-the-art detectors, and discovered that they have serious limitations when used on new waves of online hate.

5.1.1 Data Collection

To examine new waves of online hate, we carried out two data collection tasks.

5.1.2 Hateful twitter users dataset

To collect this dataset, we started with a set of 11 seed users, who have been reported [158] to post hateful content on Twitter. Then, we augmented the seed users set with their followers, and we used the Tweepy [159] framework to collect these followers. We collected a total of 3312 users whose accounts were not suspended, deleted or made private. We then proceeded to collect all the historical tweets of these users. We collected a total of 4042454 tweets, out of which 506505 were made before the first reporting (i.e., December 2019 [160]) of the pandemic, and 3535949 tweets were made after the first reporting of the pandemic.

5.1.3 COVID-19-related tweets and memes dataset.

To collect tweets and memes related to COVID-19, we compiled a set of 195 hashtags that we found to be prevalent during the COVID-19 pandemic [161, 162, 163, 164], which consisted of diverse COVID-19-related hashtags such as *covidiots*, *ChinaVirus*, *americafirst*, *WearAMask*, *trump2020* and *COVID19Vaccine*.

COVID-19-related Tweets Collection. Twitter Streaming API can only be used to collect real-time tweets, and Twitter Search API can only collect tweets in the past seven days. Therefore, we used *snsrape* [165] to collect tweets during the period from January 1, 2020 to September 30, 2020, and Twitter Streaming API from October 1, 2020 to June 30, 2021 (i.e., 18 months) based on the hashtags. In total, we obtained 507 million tweets published by 38 million users after removing all tweets not in English and retweets. Based on different definitions of hate speech in the literature [166] and on Facebook [167], we present a more specific definition related to COVID-19 used in our data labelling. We defined hate

speech as texts/comments in tweets used to attack a person or a group based on their social category, such as race, sex, sexual orientation, gender, national origin, religion, disability, occupational status, or political belief. More specifically, text that promotes/incites violence, contains dehumanizing comparisons, tries to segregate/exclude, harass with/without racial epithet, expresses inferiority and contains profanity/offensive language were all considered offensive speech. We first utilized the Perspective API [47] to obtain the initial subset of tweets. Perspective API scores texts based on how toxic they are and gives a score between 0 and 1. We randomly sampled 36,000 tweets from our dataset and used Perspective API to score each tweet. 1,235 tweets having a toxic score greater than 0.9 were retained. To remove potential bias that could result from Perspective API, another 450 tweets with toxic scores lower than 0.9 were also retained. Finally, three internal annotators labeled 1,679 tweets after removing duplicates. Of the 1,679 tweets labeled, 554 were labeled as hate speech and 1,125 as non-hate speech. If two annotators had the same label which is different from the third annotator, the label of the two annotators was adopted unless the third annotator provided a clarification based on our definition.

COVID-19-related Memes Collection. Next, we proceeded to use the hashtags to search for memes on Twitter. We only searched for tweets with image content (i.e., potential memes) based on the list of compiled hashtags for the period of February 2020 to April 2021. To control the size of the collection to a usable limit, we randomly selected one month from each quarter in the collection period. Using our COVID-19-related hashtags, we collected a total of 1,025,702 potential memes from February 2020 to April 2021.

Next, we used certain criteria to exclude memes that are invalid. First, we restricted our dataset to consist of only those memes that are in English. Since we

focus on multimodal memes [24] (i.e., images with superimposed text), we first removed memes that did not have any text in them. We also removed memes with very long text (> 30 words) to exclude screenshots and news articles, using an open-source tool Tesseract [168]. Then, we excluded those memes that did not have any image-based content (or Regions of Interest) in them (i.e., just plain background images) using the YOLO object detector [169]. Next, we removed duplicated memes. Finally, we were left with 114,064 valid memes in our dataset.

We annotated a subset of memes in our dataset. We developed a rigorous annotation process to establish the ground truth of memes based on the meme's content. In our annotation scheme, we annotated any meme as hateful, that is: (1) directed towards an individual or a group of people, organization or country, and (2) attacks victims using violent or dehumanizing speech, scandalization of personal appearance, propagates harmful stereotypes or misrepresentation, makes statements of inferiority, expressions of contempt, disgust or dismissal, cursing and calls for exclusion or segregation [49, 167].

We annotated the memes in two rounds. In the first round, the annotators independently labeled a set of randomly selected 200 memes. After independent labeling, the annotators resolved disagreements and updated the labeling guidelines based on the discussions. This resulted in the final annotation criteria presented above. In the second round, a subset of randomly selected 5000 memes were annotated independently by all annotators. This round led to near-perfect inter-rater agreement. Overall, 1,341 memes were annotated as hateful and 3659 memes were annotated as non-hateful.

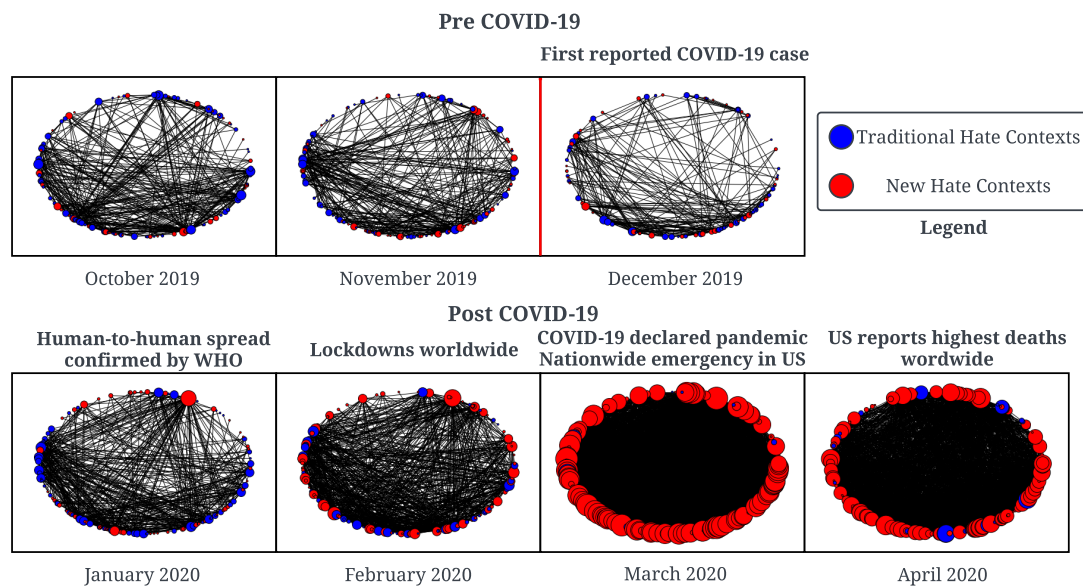


Figure 5.1: New waves of online hate context.

5.2 New Waves of Online Hate Contexts During the COVID-19 Pandemic

In this study, we wanted to find out whether global events spiral off changes in the online hate contexts in hateful users' tweets. To study the change in contexts of online hate, we used the dataset of hateful Twitter users and their tweets (Section 5.1.2), and conducted a large-scale study of these users and their tweets pre and post the COVID-19 pandemic. We studied the tweet behavior of these users regarding traditional and new contexts of online hate. The traditional contexts were related to traditional online hate, such as hate against ethnic groups (e.g. Mexican and African-American), gender-related hate (e.g. women), and hate against religious minorities (e.g. antisemitism and anti-Muslim sentiment). The new contexts were related to recent reports of online hate during the COVID-19 pandemic, such as anti-Asian hate (e.g. terms like "Chinese virus" and "Wuhan

flu” used to refer to COVID-19), hate related to mask and vaccine mandates, and Ageism-related hate (e.g. COVID-19 referred using terms like “Boomer-remover”). To determine the context of the tweet, we matched those terms to the tweet text. Next, to examine the tweeting behavior of these hateful users regarding traditional and new contexts of online hate pre and post pandemic, we constructed their networks, wherein each node represents a user’s correlation with traditional or new hate contexts, and edges represent the similarity of the tweets contexts between each user. We visualized the tweeting behavior of these users from June 2019 to June 2020, i.e., 6 months before and after the first reporting of the COVID-19 pandemic [160], and correlated them with the development of the COVID-19 pandemic [160], depicted in Figure 5.1.

To determine the statistical significance of the evolution of hate context pre and post the pandemic, we used the Wilcoxon hypothesis test [170], which is a non-parametric alternative to the dependent samples t-test. The Wilcoxon test was used since the data does not approximate a normal distribution (an assumption of the dependent samples t-test), but satisfied the assumptions of the Wilcoxon test. For the significance test, we considered the users that had tweet activity at least 6 months before and 6 months after the first reported case of the pandemic. We found that users had significantly changed the hate context in the months after the first pandemic report ($Z = 7971.0, p < 0.001$), with XX% users changing their tweet behavior from traditional contexts to new contexts of online hate. In the tweets posted before the pandemic by these users, a majority of the tweets were focused on traditional hate contexts. However, after the pandemic, a clear change towards new hate contexts can be observed in Figure 5.1 (e.g. in March 2020 and April 2020). A strong similarity between the users’ tweeting behavior is also observed, depicted by the density of edges. We also

observed a correlation between the hate context and the timeline of the COVID-19 pandemic [160], with increasingly new online hate being disseminated with the increasingly severe developments of the pandemic.

What these results indicate is that the contexts of online hate changes with global events, and new hate contexts emerge. For example during the COVID-19 pandemic, the context of hate changed from traditional hate-related contexts, to new contexts such as masks, vaccines and anti-Asian hate. This rapid change in context is problematic for existing online hate detection systems which evidently could not effectively detect the new hate contexts, as evidenced by the wave of new hate contexts following the first pandemic report in Figure 5.1. Thus, detection systems for addressing new contexts of online hate waves should address two challenges: (1) learn to effectively use knowledge about traditional hate contexts to detect new hate waves, (2) be trained in a timely manner with few data samples.

5.3 Different Representations of New Waves of Online Hate

Next, we wanted to study the role of different representations of online hate, other than traditional text-based representations in the new waves of online hate. To this end, we conducted a study about the visual content in the full set of tweets collected from the hateful users. In this study, we examined the proportion of visual media in the dissemination of online hate, pre and post the pandemic (i.e. in traditional vs. new contexts). We counted the number of tweets having visual media (i.e., images and videos) for each month, from

June 2019 to June 2020. Figure 5.2 depicts the results of our study. We used the Wilcoxon hypothesis test to determine the statistical significance of the role of different representations of online hate pre and post the first report of the pandemic (red line in Figure 5.2). We found that users were found to have significantly used different representations of online hate in the months after the first pandemic report ($Z = 124706.0, p < 0.001$), wherein an increase of **XX%** was observed in the usage of different representations post pandemic first report. We found that before the pandemic, the role of visual media in spread of online hate (i.e., blue bars in Figure 5.2) is not significant. However, in the months following the pandemic, we observed a rise in the use of visual media in the spread of online hate (i.e., red bars in Figure 5.2). A closer inspection of these tweets revealed a significant usage of image-based representations such as memes to convey the hateful meaning about vaccine, mask and Asian-hate normally conveyed via traditional, textual means. What this study indicates is that detecting new waves of online hate in the form of different representations is an important challenge for detection systems.

5.4 Effectiveness of Existing Techniques Against New Waves of Online Hate

Thus, preliminary evidence presented in the previous sections indicates that new waves of online hate pose significant challenges in terms of the need for transferring knowledge from traditional hate to detection of new waves of online hate, availability of only a few samples of new contexts of online hate, and different representations of new waves of online hate. Next, we aimed to find

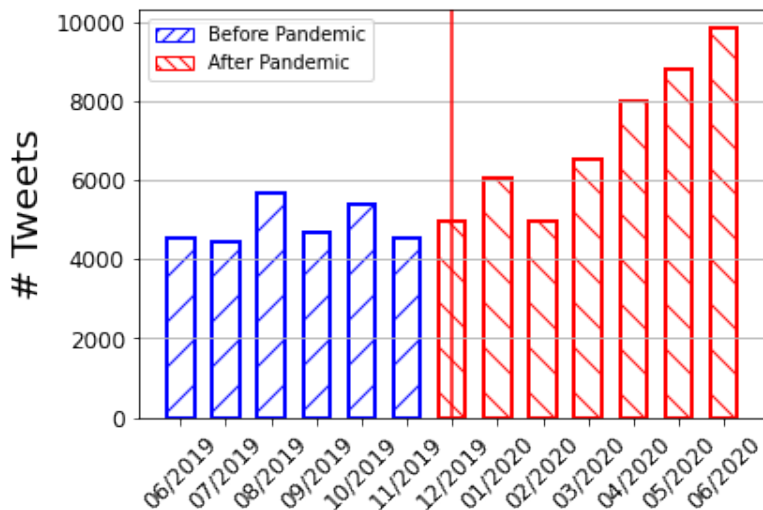


Figure 5.2: Representation of new waves of online hate.

out how effective the state-of-the-art systems and AI/ML models are against the new waves of online hate.

We carried out a measurement of several state-of-the-art existing systems (i.e., Clarifai Text Moderation [104], Perspective API [47], Azure Text Moderation [171], IBM Toxic Comment Classifier [102], Google Cloud Vision API [48], Clarifai NSFW [104], and DeepAI [43]) and pre-trained AI/ML models (i.e., MMBT [49], ViLBERT [172], VisualBERT [82], and VisualBERT COCO [82]) against the samples of COVID-19-related tweets and memes dataset (Section 5.1.3). We chose these systems and models so that we can cover both the hate context and representation. Our objective in this measurement experiment was to study the capability of these existing models on the new waves of hate only, and we do not propose that these systems and models are not effective against traditional hate. We depict the results of this measurement experiment in terms of precision, recall and F1-score in Table 5.1. We found that the existing systems are consistently deficient in addressing new online hate contexts, in both text and

Detection System/ Pre-trained Model	Input Type	Precision	Recall	F1-score
Clarifai Text Moderation	Text	0.69	0.16	0.27
Perspective API		0.49	0.31	0.38
Azure Text Moderation		0.54	0.21	0.31
IBM Toxic Comment Classifier		0.69	0.15	0.25
Google Cloud Vision API	Multimodal	0.31	0.03	0.06
Amazon Rekognition		0.41	0.01	0.02
Clarifai NSFW		0	0	0
DeepAI		0.28	0	0.01
MMBT [49]		0.25	0.27	0.30
ViLBERT		0.33	0.30	0.32
VisualBERT		0.35	0.13	0.19
VisualBERT COCO		0.47	0.02	0.04

Table 5.1: Detection capability of existing systems and pre-trained models on evolving hate.

other representations such as memes, observed from the low F1-scores reported by these systems and models. In fact, the highest F1-score was found to be just 0.38 (Perspective API), which is not sufficient for practical use.

While these existing systems address traditional hate quite sufficiently, they are quite limited in case of new waves of hate. New systems, that address the challenges of new waves of hate need to be formulated.

5.5 Our Approach

Informed by the findings about the new waves of online hate, we design an approach that is based on two key observations:

- Online hate is characterized by certain attributes. These attributes can be learned from traditional online hate datasets.
- Since the datasets of new waves of online hate are small, they cannot be used for supervised learning. Alternative ML techniques that can be used with small datasets should be used for new waves of online hate.

5.5.1 Approach Overview

The main components of our approach are depicted in Figure 5.3. We first collected small datasets of online hate witnessed during the recent pandemic. Specifically, we collected samples related Asian, mask, vaccine and Boomer (i.e., Ageism during COVID-19) hate. In our work, we used the generalized zero shot learning paradigm [51]. We used the new waves of hate as inference datasets, and used samples from traditional online hate datasets [153, 154] as training datasets. We first extracted attributes of online hate in samples from all the datasets. Then, we constructed entailment labels for all samples based on hate or non-hate label and the online hate attributes. Next, we trained our attribute-based zero shot learning model, AZL, on the traditional online hate samples. We then ran our trained model on the new waves of hate based on entailment labels. In the following, we provide in-depth discussions on each of the components involved in AZL.

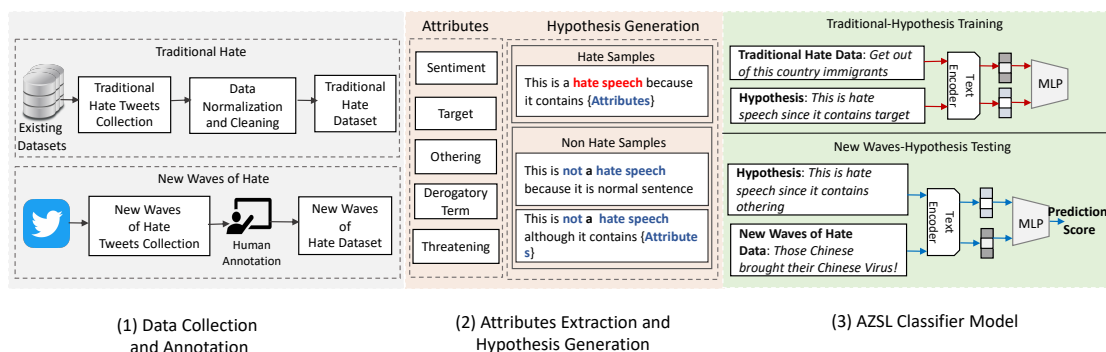


Figure 5.3: AZL overview.

New Wave of Hate Type	Number of Samples
Asian-hate	223
Mask	221
Vaccine	62
Boomer (i.e., Ageism during COVID-19)	177

Table 5.2: New waves of hate types.

5.5.2 Data Collection

5.5.3 Online Hate Attributes

Although traditional online hate and new waves of hate differ significantly in their compositions, such as the target of hate, or the subject of hate, they have some attributes that are generic to online hate. These online hate attributes can characterize online hate in general, whether the samples belong to traditional hate or new waves of hate. In our framework, we use these attributes to build classifier model that is trained to understand these attributes from traditional online hate, and then use these attributes to make decisions on new waves of hate. In this way, our classifier model learns to apply knowledge about online hate by learning it on traditional online hate and applying it to detect new waves of hate. In our work, we use five attributes that have been compiled based on existing work in the psychology [173], computer science [154, 174, 175] and social sciences [176, 177] domains. We do not claim that these five attributes are

a complete list of online hate attributes, but we use them to demonstrate the effectiveness of attribute-based learning in detecting new waves of online hate.

- **Sentiment.** Sentiment of a text depicts the polarity of the text towards a positive or negative feeling. Online hate in general is associated with negative sentiment [175, 174], since hate itself might originate from negativity towards an entity. The sentiment polarity contrast between regular content and hateful content is a useful attribute in distinguishing online hate and non hate. Furthermore, sentiment is a general attribute of online hate, and thus characterizes both traditional as well as new waves of hate.
- **Target.** Online hate is characterized by the presence of certain actors, including hate speech instigators and targets [178, 179]. Targeting a community or individual because of their immutable or prominent characteristics is a known tactic of perpetrators. Thus, the presence of a target is an important characteristic of online hate. For example, in traditional contexts, vulnerable groups such as *women* and *African-Americans* have been targets of online hate. In recent times, other groups such as *Asians* and *older people* have been targets of online hate, especially during the pandemic. The presence of a target is thus a generalized attribute of online hate.
- **Othering.** “Othering is a phenomenon in which some individuals or groups are defined and labeled as not fitting in within the norms of a social group. It is an effect that influences how people perceive and treat those who are viewed as being part of the in-group versus those who are seen as being part of the out-group” [180, 173]. In simple terms, othering is the expression of an “us vs. them” feeling. In social media, othering is prevalent in posts containing online hate [173, 175]. Othering thus, is found in online

hate speech in general, and therefore used as an attribute of online hate in our work.

- **Derogatory terms.** Derogatory terms are words or phrases used to demean, humiliate or belittle targets in a piece of text. Such terms are not only relevant to online hate, but are highly prevalent in online hate. For example, in traditional contexts, derogatory terms such as *n**ger*, *b**ch* and *chi*k* have been used to demean certain minority groups. More recently, in the new waves of hate witnessed during the COVID-19 pandemic, derogatory terms such as “kung flu” and “chop fluey” were used to target the Asian community, and the phrase “Boomer Remover” was used to mock the high mortality rate among older people. Derogatory terms are in general attributes of online hate.
- **Threatening terms.** Threatening speech is “the speech, when heard or seen by its target, would result in serious apprehension of danger, at the hands of either the speaker or a third party who responds to the speech.” [181]. Threatening terms are a general aspect of online hate, often found in extreme hate speech wherein perpetrators use such terms to depict their intent.

In our work, we extracted these attributes using text mining and other NLP techniques. We extracted sentiment using the Google sentiment analyzer [182], and each sample was allocated a sentiment between 0 and 1. We extracted the target using named entities, wherein we used Python NLTK tokenization and Spacy [183], and used the entities *PERSON*, *GPE*, *ORG*, *NORP* as targets. We extracted othering if samples contained othering words or phrases such as *they*, *them*, *their*, *you people*, etc. We extracted derogatory terms in samples using

Profanity-check [184]. Lastly, we used the Perspective API to point out threatening terms [138].

5.5.4 AZL: Attribute-based Zero Shot Classification

AZL model. We propose to detect new waves of online hate, by approaching this classification problem as a textual entailment problem, in an attribute-based zero-shot learning setting. This is inspired by: (i) Entailment allows us to create hypothesis based on attributes of online hate. Since we want to use traditional online hate to learn generalized attributes to make prediction about new waves of hate samples, we need efficient ways to encode these attribute so that a classification model can learn to detect such samples using the attributes. Entailment allows us to convert simple binary labels (i.e., *hate* and *non-hate*) into hypothesis statements (e.g., *this text is hateful since it contains othering*). (ii) Zero-shot learning paradigm is suitable when a new wave of hate occurs, wherein large datasets are definitely not available for a conventional supervised learning paradigm. Therefore, exploring detection of new waves of hate as a textual entailment problem in an attribute-based zero-shot learning paradigm is a reasonable way to achieve generalization on unseen samples of erupting new waves of hate.

Converting labels into hypothesis. The first step of AZL is to convert binary labels into hypotheses. To this end, we convert each label into a hypothesis statement consisting of the label (i.e., *hate* or *non-hate*) and the attributes that are contained in the text. Table 5.3 lists some examples for converting binary hate labels into hypotheses statements containing attributes.

Converting classification data into entailment data. Typically for a data split

Text Sample	Label	Attributes	Example Hypothesis
“get out of this country, filthy immigrants”	Hate	negative sentiment, derogatory terms	“this is hate speech since it contains negative sentiment and derogatory terms”
“the f***ing democrats and these Chinese planned China Virus”	Hate	othering, derogatory terms, target	“this is hate speech since it contains othering, derogatory terms and target”
“I lost my job due COVID-19”	Non-hate	-	“this is not hate speech”

Table 5.3: Converting labels to hypothesis.

(i.e., train, dev and test), each input sample, acting as the premise, has a positive hypothesis corresponding to the positive label and negative hypothesis corresponding to the negative labels. We convert both the traditional online hate dataset and the new waves datasets into entailment data. During training, we only use the traditional online hate samples and during the inference, we use the trained model on the new waves of hate and consider them as test dataset.

AZL model learning and inference. In this chapter, we make use of the RoBERTa-large pre-trained on ANLI [185], MNLI [186], and SNLI [187] tasks. We fine-tune this model on the traditional online hate dataset. For entailment, we use the cosine similarity loss as given by the equation below.

$$\text{Cosine}(x, y) = \frac{x \cdot y}{|x||y|} \quad (5.1)$$

In the inference time, we entail an input sample (i.e., a new wave of hate sample) with all the hate and attribute combination hypotheses, as well as the non-hate hypotheses. We consider a sample as hateful, if the entailment score with any of the hate and attribute combination hypotheses is greater than the

non-hate hypotheses.

$$Prediction = \begin{cases} Hate, & \text{if } f\{x, y\}, \forall y \in \{\text{hate hypotheses}\} \\ & > \\ & f\{x, y\}, \forall y \in \{\text{non-hate hypotheses}\} \\ Non - hate, & \text{otherwise} \end{cases}$$

In Equation 5.5.4, y is a hypothesis statement, and $f(x, y)$ depicts the entailment of an input sample x with a hypothesis.

5.6 Implementation and Evaluation

5.6.1 Implementation

We evaluated AZL using two metrics - accuracy and weighted average F1 score. We compute both metrics for four types of new online hate waves: Asian, mask, vaccine and boomer hate.

New Wave of Hate	Precision	Recall	Weighted Avg. F1 Score
Asian	86.27%	68.75%	76.52%
Mask	55.83%	79.13%	67.47%
Vaccine	61.7%	82.86%	70.73%
Boomer	77.27%	68.0%	72.34%

Table 5.4: Evaluation of AZL model on new waves of online hate.

5.6.2 Effectiveness of AZL

By the comparing the metrics in Table 5.4 to the performance of current systems in Table 5.1, it can be clearly observed that our model based on attribute-based

zero shot learning paradigm and entailment achieves significant performance improvements on new waves of online hate. It should be noted that weighted-averaged F1 score is calculated by taking the mean of all per-class F1 scores while considering each class's support. In our work, we consider two classes, hate and non-hate. The significantly higher weighted average F1 scores in our model shows that it can significantly outperform the current systems on new waves of online hate in a balanced manner.

5.7 Conclusion

In this task, we analyzed in-depth, the problem of new waves of online hate. We collected a novel datasets of hateful Twitter users, and found that the users perpetrate new waves of online hate based on crisis events. We also found that the representation of hate evolves from traditional text based representations to visual-based representations involving images, such as hateful memes. We then introduced a novel approach based on attribute-based zero shot learning, trained on textual entailment paradigm. Our evaluation of our detection framework shows that it can achieve new state-of-the-art results in detecting new waves of online hate.

Robustness of Cyberharassment Detection Models

Multimodal learning has been gradually gaining focus of the research community over the past few years. The approaches for multimodal learning have come a long way from simple models re-purposed for multimodal tasks, to deep learning-based models that are specifically designed for multimodal tasks (referred to as Deep Multimodal Models or DMMs throughout this chapter). For example, recent advances in this field have led to several state-of-the-art DMMs, such as VisualBERT [82], MMBT [49], and Pythia [188], while also engendering the collection of several multimodal datasets, such as Hateful Memes [24], and Visual Question Answering (VQA) [80]. Due to the success of these DMMs on standard benchmarks, there have been many encouraging attempts to adopt them to real-world and safety-critical scenarios, such as autonomous driving, assistance to blind people [189], and hate-speech moderation on social media [24]. However, in spite of the recent advances, the robustness of DMMs remains poorly understood.

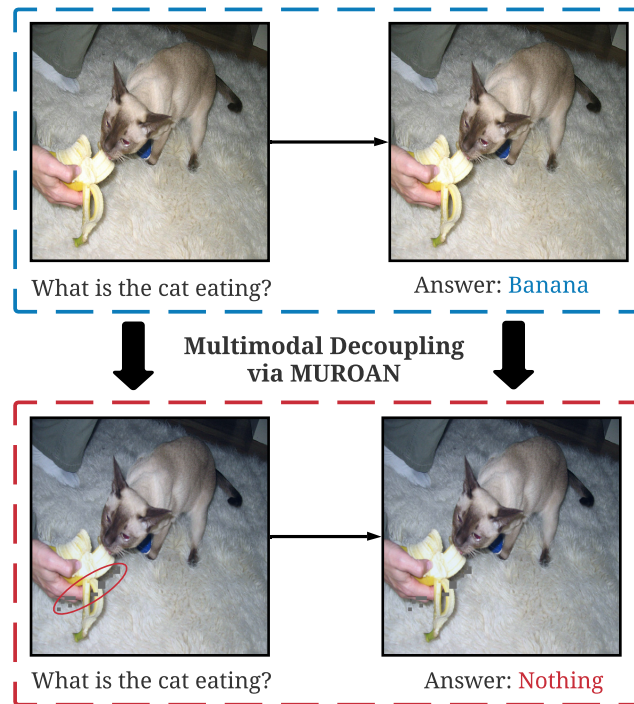


Figure 6.1: By decoupling the input modalities through the removal of a few datapoints in the image via MUROAN framework, the multimodal model predicts a wrong answer class: *Nothing*, indicating that decoupling attack can easily compromise multimodal models.

A significant difference between DMMs and their unimodal counterparts is the *fusion* mechanism in DMMs. This fusion mechanism fuses multiple input modalities to learn their joint representation, which is then processed by several fully connected layers to predict classification scores depending on the nature of the corresponding downstream tasks. Different DMMs [49, 82, 188] employ different strategies to learn strong fusion embeddings of their input modalities. This fusion mechanism presents new challenges towards studying the adversarial robustness of these models.

Recently, several unimodal adversarial attacks for deep unimodal models have been formulated to study their robustness. For example, unimodal ad-

versarial images [37, 39, 40, 38] and unimodal adversarial text [41] have been widely studied, which have exposed numerous vulnerabilities in deep unimodal models. However, these attacks cannot be directly employed to study the robustness of their deep multimodal counterparts. First, since these attacks can only be applied to single modalities, they do not affect the fusion mechanism that is fundamental to DMMs. Second, since DMMs combine several different types of modalities (e.g., image, text, speech, etc.), a single unimodal attack cannot be used for all those modalities. We note that formulating comprehensive methods to study the robustness of DMMs is of utmost importance to adopting them in real-world systems.

To address these challenges, in this chapter, we first highlight how multimodal adversarial attacks based on decoupling the input modalities in DMMs can easily compromise these models. Then, we introduce a framework called MUROAN to study the robustness of DMMs based on decoupling of modalities, thereby revealing vulnerabilities in the fusion mechanism of existing DMMs. MUROAN uses a unified view of DMMs to expose its key vulnerability. Then, we introduce a new type of adversarial attack called decoupling attack in MUROAN, wherein the objective of its attack algorithm is to decouple the input modalities of multimodal models to induce a misclassification. As depicted in Figure 6.1, a decoupling of the image and text modalities through occlusion of a few datapoints in the image induces a misclassification. In addition, we leverage the MUROAN framework to measure several state-of-the-art DMMs. We find that the seemingly straightforward decoupling attack of MUROAN is in fact highly effective in compromising DMMs.

Our contributions in this chapter are as follows.

- We present a unified view of DMMs to explore their vulnerabilities, and identify the fusion mechanism of these models as a critical component for their robustness analysis.
- We propose a novel framework called MUROAN that consists of the unified view to exploit the fusion mechanism and a decoupling attack algorithm for comprehensively studying the adversarial robustness of DMMs. MUROAN directly focuses on the fusion mechanism of DMMs by decoupling the input modalities that are fused together.
- We use MUROAN for a comprehensive robustness analysis of state-of-the-art DMMs under several dataset and model settings. Our experiments show that, in the worst case, the decoupling attack in MUROAN can achieve an attack success rate of 100% after decoupling of 1.16% of input modalities of DMMs.

We are open-sourcing our code to encourage research in training DMMs robust to decoupling attacks: <http://github.com/SecurityAndPrivacyResearch/mda>.

6.1 Background

In the following, we give an overview of the field of multimodal learning as well as the state-of-the-art unimodal adversarial attacks used for the robustness analysis of unimodal models.

6.1.1 Multimodal Learning

The renewed interest in multimodal learning can be attributed to more powerful models [44, 45] that can learn strong fusion of input modalities and the availability of several multimodal datasets [80, 24]. These models and datasets have resulted in DMMs achieving impressive results on standard benchmarks. Much of the DMMs that have achieved impressive performances can be categorized under the following categories.

Traditional Fusion-based Models. Several DMMs have attempted to address how to effectively combine multimodal information [81]. Feature concatenation is one of the most preferred fusion techniques in these models, while some of the models use other feature fusion techniques such as element-wise product. Since these models showed impressive performances on several multimodal benchmarks, they are considered strong baselines for many multimodal tasks.

Transformer-based Fusion Models. Recently, the BERT model [44], a type of transformer [45], has been shown to achieve state-of-the-art performance [49, 82] on multimodal benchmarks, by learning the interaction between the input modalities via self-attention over many different layers. For example the MMBT [49] model fuses image embeddings in the form of pooled filter maps from a ResNet model and word tokens as two segments of BERT [44]. As shown by these works, the transformer based DMMs outperform their unimodal counterparts in multimodal tasks by quite a large margin.

6.1.2 Unimodal Adversarial Attacks

The discovery of unimodal adversarial attacks has engendered active research in the safety and robustness of unimodal deep learning models. In this section, we discuss important unimodal adversarial attacks on images and text.

Unimodal Adversarial Image. A large body of adversarial attacks have been introduced in recent times that mainly focus towards robustness analysis of computer vision models. For example, several works, such as fast-gradient attacks [83], optimization-based methods [37, 38], and other such methods [40], have been proposed successfully. Furthermore, alarmingly critical real-world attacks such as adversarial patches [84] have been introduced recently, which cast serious questions on the safety of these vision models.

Unimodal Adversarial Text. Recently, some works have focused on unimodal adversarial text to study robustness of Natural Language Processing (NLP) models. While earlier works [85] effectively employed character level perturbations to perform adversarial attacks, more recent works have found word replacement strategies [41] to be largely effective in compromising these models.

6.2 Threat Model

In this section, we enumerate the goals and capabilities of an adversary in the multimodal learning attacks domain. We consider an adversary of a DMM system whose goal is to provide an adversarial multimodal input x' that results in an incorrect output classification. We consider the adversary whose objective is to cause both untargeted and targeted misclassification. We assume that the ad-

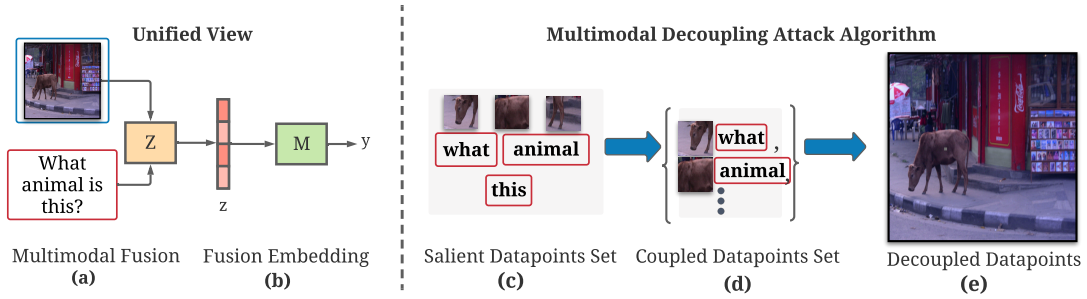


Figure 6.2: Overview of our approach.

versary has complete knowledge about the targeted DMM including the DMM model architecture parameters.

6.3 Our Approach

In this section, we discuss our approach for the robustness analysis of DMMs via MUROAN framework. In this regard, we first discuss a unified view of DMMs to explore the vulnerabilities of the fusion mechanism of DMMs, and then introduce our algorithm to decouple the fused modalities of DMMs. The overview of our approach is depicted in Figure 6.2. A threat model regarding our approach is presented in Appendix 6.2.

6.3.1 Unified View of Deep Multimodal Models

We consider a DMM $D : X \rightarrow Y$ to be a function that maps a domain X to a co-domain Y . An input is a set of vectors of different modalities $x = \{x_0^1 \dots x_n^1, x_0^2 \dots x_m^2, \dots\}$ (Figure 6.2, Step (a)). We consider Y to be the set of possible classes for a multimodal input $x \in X$. The output of the DMM for a multimodal input x is considered to be $D(x) = y$, for some $y \in Y$. We denote the confidence of the DMM for a multimodal classification probability on input

x and class y as $D_y(x)$. Lastly, we denote the cardinality of a set as $|\cdot|$, which represents the number of elements in the set.

Since DMMs have several different architectural configurations, we need a unified view (or representation) of them for a uniform vulnerability analysis of all these different multimodal architectures. To achieve this, we unify these different architectural approaches into a single view, in which we consider a DMM as a generator of the fusion embedding of multiple input modalities (Figure 6.2, Step (b)), followed by several fully connected layers that are specific for downstream tasks. In other words, we break down a DMM into two functions: the first generates a latent representation (i.e., the fusion embedding) of the multimodal inputs and the second performs classification based on the fusion embedding. We consider the fusion embedding of a multimodal input x as $Z(x) = z$, where z is the d -dimensional fusion embedding vector. Next, we consider $y = M(z)$ to represent classification based on the fusion embedding from fully connected layers that are specific to downstream tasks. Therefore, the original DMM is broken down into two functions, represented as $M(Z(x))$. We further discuss this process for two typical DMM architectures: traditional architectures and transformer-based architectures.

Traditional Multimodal Architectures. Traditional DMM architectures are composed of separate neural networks that are specific to each input modality, whose outputs are combined using fusion techniques such as element-wise multiplication, addition or concatenation. For example, the Pythia [188] architecture is composed of a convolutional neural network that learns the embedding of the image modality, and a recurrent network that learns the embedding of the text modality, which are then combined using element-wise multiplication. This combination represents the fusion embedding.

Transformer-based Multimodal Architectures. These architectures use the transformer [45] for learning a strong fusion embedding of the input modalities. The input modalities are first converted into embeddings, which are then combined using the transformer, which performs several self-attentions across many layers. The first token embedding then constitutes the fusion embedding, which is subsequently processed by fully connected layers for classification.

6.3.2 MUROAN Framework

The traditional methods of adversarial attacks are not suitable for DMMs for two specific reasons. First, most key methods of crafting adversarial attacks use either the l_∞ or l_2 norm¹. Optimization with respect to these kinds of manipulations induces a perturbation in all (or almost all) of the datapoints of an input modality by a small value $\pm\epsilon$. This is not suitable in case of multimodal inputs because different modalities have different compositions, and not all modalities support this type of manipulation. For example, image-based inputs are continuous and thus suitable for such manipulations, but text-based inputs are discrete, thus not suitable for such manipulations. Second, for DMMs, such adversarial manipulations are not suitable for robustness analysis processes since the core weaknesses of these models should be examined in the fusion mechanism of these models, which is not achieved by these manipulations. Since we are interested in studying the effect of decoupling fused modalities, we employ l_0 -norm-based optimization attack algorithm, wherein an l_0 -norm attack optimizes for the number of changes made to the inputs for a successful decoupling attack.

¹We note that an l_0 -norm attack [40] exists for unimodal models. However, the vast majority of attacks used for robustness studies use l_∞ or l_2 norm attacks.

Removal of salient datapoints from inputs has been shown to be an important factor for considering the robustness and safety of a decision model [190]. However, the key difference between the traditional unimodal domains and the multimodal domain is that such datapoints are in fact parts of separate modalities that are coupled together by the multimodal fusion mechanism. Thus, it is imperative to study the cases, in which some parts of the input modalities are removed, so as to render this fusion as unsuccessful.

For a multimodal input x , we consider coupled datapoints as some $x' \subset x$. Our objective is to find the minimum subset via the following optimization.

$$x' \subset x (|x| - |x'|) \text{ s.t. } D(x) \neq D(x') \quad (6.1)$$

However, it is impractical to solve the optimization in Equation 6.1, due to a large number of such datapoints in the multimodal input space. Thus, to solve this optimization, we use the notion of the fusion embedding to compute a salient points set first, S_n (Figure 6.2, Step (c)). We use the salient datapoints set to study the weaknesses of DMMs, by defining it as follows.

$$S_n^x = \{x_i \in x \mid Z(x/x_i) \neq Z(x)\} \quad (6.2)$$

In Equation 6.2, the salient datapoints set contains those datapoints that affect the fusion embedding upon removal (where x/x_i denotes removal of a datapoint). For example in the transformer-based DMMs, a datapoint $x_i \in S_n^x$ if $\forall i \neq j, z_i \geq z_j$ due to the transformer pooling layer. Next, we find the set of coupled datapoints (Figure 6.2, Step (d)) from the salient datapoints set. Let $[n] = \{1, 2, \dots, n\}$ denote the collection of all subsets of size $\{1, 2, \dots, n\}$ from the salient datapoints set, and $P(X)$ represent a set of all subsets of X , then coupled

Algorithm 1: MUROAN Decoupling Attack Algorithm

Input: $x, y, D, \Theta, f, \text{maxitr}$
Output: x'

```

1 Initialization:  $x' \leftarrow x$ 
2 while  $f(D, x, x', y)$  or maxitr do
3    $S_n^x \leftarrow \text{GetSalientSet}(D, x')$ 
4    $C_n^x \leftarrow \text{GetCoupledSet}(S_n^x)$ 
5   for  $x_i \in C_n^x$  do
6     if  $D(x') \neq y$  then
7       break
8     end
9      $x' \leftarrow x' / x_i$ 
10    if  $D(x') \neq y$  and  $f(D, x, x', y) \neq \text{True}$  then
11       $x \leftarrow x$ 
12    end
13  end
14 end
15 return  $x'$ 

```

datapoints set is the permutations of all datapoints of a maximum size equal to the size of the salient datapoints set, defined in Equation 6.3.

$$\{C_n \in P([n])\} \quad (6.3)$$

Now that we have computed the coupled datapoints set, we propose the MUROAN Decoupling Attack Algorithm (Algorithm 1) to iteratively refine the decoupling attack. In our algorithm, first the salient datapoints set is computed based on the process described in Equation 6.2. Then, the GetCoupledSet procedure is called, which performs two functions. First, the coupled datapoints are computed as described in Equation 6.3. Then, they are ordered based on the size of the datapoints, so as to satisfy Equation 6.1. We encode the termination of our algorithm as a boolean function f , to support multiple adversarial requirements. For example, adversarial requirements for crafting untargeted at-

tacks ($D(x') \neq y$) or targeted attacks ($D(x') = y'$) can be supported (Figure 6.2, Step (e)). Lastly, we propose the following theorem to use our decoupling attack algorithm as a robustness verification technique to find adversarial examples in DMMs if one exists.

For a multimodal model D that satisfies our unified view and a given multimodal input x , the MUROAN decoupling attack algorithm will find the optimum adversarial example that satisfies Equation 6.1.

Proof. If an adversarial example exists for input x , it can be found by an exhaustive search of the input space. The `GetCoupledSet` function returns all possible permutations of the coupled datapoints and the f function and `maxitr` can be set such that the algorithm does not terminate until a satisfactory adversarial permutation is found. Furthermore, since the permutations in the coupled datapoints set are ordered, thus, a permutation that is found by our algorithm to be adversarial is minimal.

6.4 Implementation and Evaluation

In this section, we first summarize the DMMs, datasets, and unimodal adversarial baselines that are used in our experiments. We then use MUROAN to analyze the robustness of state-of-the-art DMMs trained on popular multimodal datasets to show how decoupling attack can easily compromise these models, thereby enabling us to understand their robustness. We also consider some unimodal adversarial attack baselines in our evaluation only to show how easily decoupling attack can compromise DMMs. Our objective is not to make a direct comparison of our approach against these existing attacks, but to highlight how decoupling of input modalities can be easily used to attack the fusion mecha-

nism of DMMs. We also conduct adversarial training to study potential defense against our attack. Our findings highlight the need for rigorous safety analysis of DMMs against decoupling attacks, and lay down important groundwork for their deployment in real-world applications.

6.4.1 Implementation

We have implemented our attack using the PyTorch library. For the VQA dataset we used 1000 samples and for Hateful Memes dataset, we used 250 samples to conduct our experiments. We used pretrained models published by the original authors for all the DMMs that we have evaluated in our experiments. In the MUROAN decoupling attack algorithm, we used a maximum iteration limit of 500 epochs, post which we report the attack as unsuccessful. We ran our experiments on a single NVIDIA V100 GPU enabled eight core machine.

6.4.2 Baselines

Deep Multimodal Models

- **Pythia.** The Pythia [188] is a state-of-the-art model in the VQA challenge task. This model is composed of a convolutional network to compute an image embedding and a recurrent network to compute a sentence embedding, which are fused using element-wise multiplication.
- **Late Fusion.** We consider the late-fusion architecture based DMM in [191] as a strong baseline model. In this model, image embeddings from a convolutional neural network and text embeddings from a recurrent network are fused using element-wise sum, and then the fusion embedding is

processed through multiple classification layers to generate a probability score.

- **MMBT.** The MMBT model [49] is a state-of-the-art DMM that utilizes the BERT [44] to learn multimodal embeddings by the implicit alignment of image and text features with the self-attention mechanism of transformers [45], for a wide range of visual-linguistic tasks. The query vector of this model, which is treated as the fusion embedding, is processed through a classifier head for downstream tasks.

Multimodal Datasets

- **Hateful Memes.** The Hateful Memes [24] dataset consists of image and text pairs pertaining to hateful memes, a recent phenomenon that poses a serious societal threat in today’s day and age. The objective is classification into two categories: “hateful” or “non-hateful”.
- **Visual Question Answering (VQA).** The VQA dataset [191] consists of images with multiple associated natural language questions. Each image and question pair expects a list of answers. The objective is to predict the best answer from the list of answers for each image-question pair.

Unimodal Adversarial Baselines

We considered two image-based unimodal adversarial baselines (i.e., Carlini and Wagner attack [38] and Projected Gradient Descent attack [39]), and two text-based unimodal adversarial baselines (i.e., Genetic Attack [192] and TextFooler [41]). Please refer to Appendix 6.4.2 for more details about these unimodal baseline

attacks. Furthermore, the configuration details of these baselines can be found in Appendix 6.4.2.

Unimodal Adversarial Baseline Details

- **CW Attack.** We use the Carlini and Wagner [38] attack algorithm as baseline for unimodal adversarial images for image-based modality.
- **PGD Attack.** We have also used the Projected Gradient Descent [39] attack algorithm which is a popular image-based attack baseline in our work.
- **Genetic Attack.** We use the Genetic Attack [192] algorithm (referred to as “Genetic” in this chapter) as baseline for unimodal adversarial text for text-based modality.
- **TextFooler.** TextFooler [41] is a greedy word substitution based adversarial attack algorithm specifically designed to attack text-based models. We use this algorithm as an additional text-based unimodal attack baseline.

Baseline Configuration

In this section, we provide the configuration of all the baseline models and techniques used in our paper.

PGD. We have used the L_∞ norm PGD with a perturbation budget of $8/255$, a step size of 0.01, and the number of iterations as 40.

CW. We have used the L_∞ norm CW attack with a step size of 0.10, learning rate of 0.01 and number of iterations of 500.

Genetic Attack. We have used a population size of 20 and maximum generations as 20.

TextFooler. We set the threshold to choose important word as -1, the threshold for selecting sentences of high semantic similarity as 0.5, and the size in score module as 15.

Adversarial Training. We followed the adversarial training settings in the original paper [39]. We used CW loss function (see above for CW configuration), and training procedure using SGD optimizer, with a learning rate of 0.0001 and momentum of 0.9.

6.4.3 Effectiveness Evaluation

Robustness Analysis

In this section, we used our framework to analyze the robustness of state-of-the-art DMMs under various attack conditions to show that the robustness of these DMMs are largely overestimated.

Adversarial robustness of an AI/ML-based model refers to how robust it is to (test time) perturbations of its inputs by an adversary intent on fooling the model. A reliable way to analyze robustness of a model is to study the average perturbations (i.e., inputs changed) for a successful attack. We studied the percentage of average points changed by MUROAN decoupling attack algorithm in comparison with the CW attack and PGD attack for a successful misclassification. We used the same cutoff of 500 epochs for both the algorithms in all the tests, post which we reported a failure. We have depicted the results of this experiment in Figure 6.3.

Figure 6.3 depicts the CDF of the average percentage of datapoints changed by the attacks under consideration. We found that the unimodal adversarial attacks (i.e., the CW attack and PGD attack) needed to change significantly more

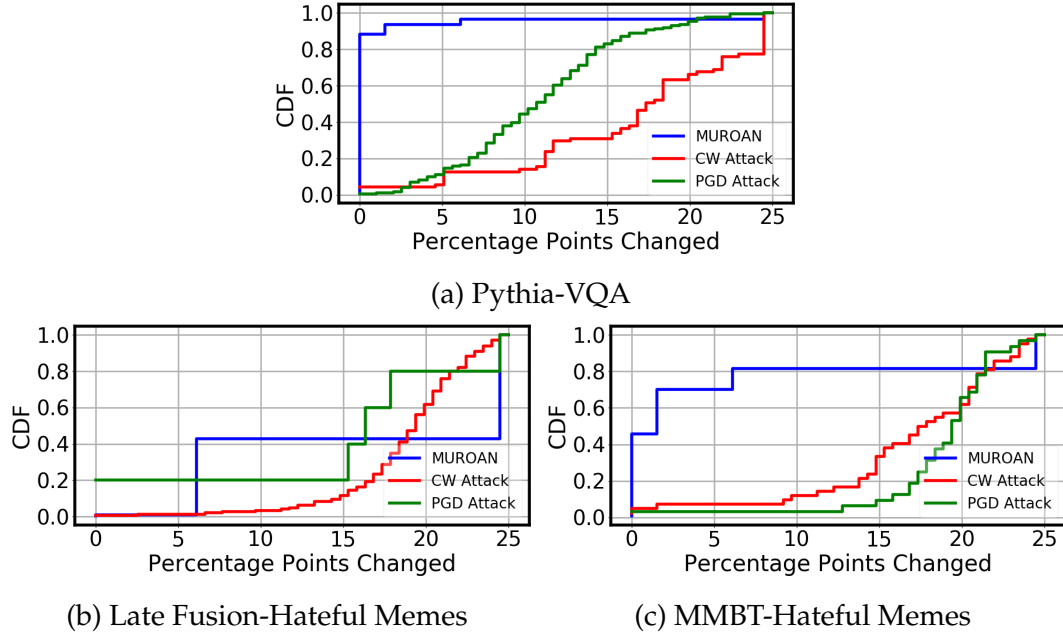


Figure 6.3: CDF of percentage of datapoints changed.

Model-Dataset	Average Points Changed - MUROAN	Average Points Changed - CW	Average Points Changed - PGD
Pythia-VQA	1.16%	93.99%	96.47%
Late Fusion-Hateful Memes	16.93%	99.86%	98.28%
MMBT-Hateful Memes	5.73%	94.92%	98.33%

Table 6.1: Comparison of average percentage points affected by MUROAN and CW attack.

datapoints to induce successful attacks in comparison to MUROAN. For instance from Table 6.1, for the Pythia-VQA, it was observed that the CW attack changed 93.99% of the input datapoints and the PGD attack changed 96.47% of input datapoints, whereas MUROAN decoupling attack algorithm changed 1.16% of input datapoints. The observation that the unimodal attacks changed a large number of datapoints but our attack changed significantly fewer datapoints for successful attack indicates that these attacks cannot be used to sufficiently study the robustness of DMMs. This finding may have important implications on using DMMs such as VQA models in real-world applications, such as

Model-Dataset	ASR-MUROAN	ASR-CW	ASR-Genetic	ASR-CW+Genetic	ASR-PGD	ASR-TextFooler	ASR-PGD+TextFooler
Pythia-VQA	100%	79.77%	49.30%	86.45%	86.51%	0%	86.51%
Late Fusion-Hateful Memes	97.25%	59.11%	0%	59.11%	96.12%	0%	96.12%
MMBT-Hateful Memes	83.33%	47.19%	0%	47.19%	75.19%	0%	75.19%

Table 6.2: Comparison of Attack Success Rate (ASR).

visual question answering for the blind [189]. Next, we discuss another important application domain, namely Hateful Memes. For the Late Fusion-Hateful Memes model, it was again observed that the unimodal baselines (i.e., CW attack and PGD attack) changed significantly more datapoints (e.g. 99.86% in case of CW), whereas MUROAN decoupling attack algorithm changed an average of 16.93% of input datapoints. For the MMBT-Hateful Memes model, it was observed for instance, that the CW attack changed 94.92% of datapoints, whereas MUROAN decoupling attack algorithm changed an average of 5.73% of input datapoints.

Next, we compared the Attack Success Rate (ASR) of MUROAN decoupling attack algorithm with respect to the unimodal adversarial images and text baselines, namely the CW [38] attack, Genetic [192] attack, PGD [39] and TextFooler [41] attack respectively. The results of this experiment have been depicted in Table 6.2. We first discuss the impact of the unimodal adversarial images on the DMMs. In all the three DMMs, we found that although the unimodal adversarial images could affect these DMMs, they were not sufficiently effective when compared to the ASRs of MUROAN decoupling attack algorithm. For the Pythia-VQA model, the CW attack for instance achieved an ASR of 79.77%, although the ASR achieved by MUROAN decoupling attack algorithm was 100%. For the two DMMs for hateful memes (i.e., Late Fusion-

Hateful Memes and MMBT-Hateful Memes), a similar observation was made, although the CW attack achieved significantly lower ASR for both DMMs. Next, we took a closer look at the impact of the unimodal adversarial text (i.e., Genetic attack and TextFooler) on the DMMs, in comparison with MUROAN. For the Pythia-VQA, it was observed that the Genetic attack for instance has little effect when compared to MUROAN decoupling attack algorithm, and even to the CW attack, wherein both these attacks outperformed the unimodal adversarial text baselines by a large margin. In case of the hateful memes DMMs (i.e., Late Fusion-Hateful Memes and MMBT-Hateful Memes) this margin was found to be even larger. It was observed that the unimodal adversarial text had no significant effect on the DMMs for hateful memes.

Thus, we observed that the safety and robustness of these DMMs need to be deeply examined, specifically from the perspective of decoupling attacks. In this regard, our experiments indicate that our attack exposes the vulnerabilities in the fusion mechanism of DMMs, and the robustness of this mechanism needs significant improvement, especially if DMMs are to be deployed in real-world systems.

Adversarial Training

Our experiments in Section 6.4.3 raise an important question: how can we defend against decoupling attacks? We performed a preliminary experiment to see if adversarial training [83], a popular technique to improve adversarial robustness, can be used to reduce the attack success rate. The configuration details of the adversarial training procedure can be found in Appendix 6.4.2. We performed adversarial training using the MMBT model for the hateful memes classification. We generated 247 adversarial examples via MUROAN framework

and trained the model on these samples combined with the original dataset from scratch. We observed that the adversarial trained DMM was still vulnerable to newly crafted decoupled samples, despite the model achieving near 100% accuracy classifying adversarial examples included in the training set. These results demonstrate the difficulty in defending against decoupling attacks using traditional adversarial training. We hope these results inspire further work in increasing the robustness of DMMs.

6.4.4 Qualitative and Quantitative Analysis

Qualitative Analysis of MUROAN



Figure 6.4: Three samples depict three types of minimum coupled datapoints in the VQA and Hateful Memes dataset. In sample (a), the minimum coupled datapoints are in the image only (indicated by red circles), and it is enough to only make changes to a those datapoints to decouple the sample. In sample (b), the minimum coupled datapoints are in the text only (indicated by red font), it is enough to make changes to the text only to decouple the sample. In sample (c), the coupled datapoints consist of both image and text, therefore both need to be changed to decouple this sample.

In this section, we provide a qualitative analysis of the decoupled samples

that MUROAN decoupling attack algorithm generated. Upon observation of such samples in the two baseline datasets (i.e., VQA and Hateful Memes), we discuss certain aspects of the nature of decoupling pertaining to our observations. In Figure 6.4², we depict three samples from our robustness analysis experiments. Figure 6.4 (a) is from the VQA dataset, and Figures 6.4 (b) and (c) are from the Hateful Memes dataset. These three samples represent the three levels of decoupling we observed in our experiments. In Figure 6.4 (a), the minimum coupled datapoints were found in the image only, therefore it is sufficient to decouple just the single image modality. In the VQA dataset, since questions are asked about certain parts of an image, this observation is intuitive since it should be sufficient to only affect the relevant parts of the image. In Figure 6.4 (b), the minimum coupled datapoints were only found in the text modality, since intuitively we cannot see why this sample could be a hateful meme from the image alone. In Figure 6.4 (c), the minimum coupled datapoints consist of both the image and the text modalities. In this case, both the input modalities need to be affected for decoupling this fusion. Therefore, we note that vulnerabilities in the DMMs are of a very different nature when compared to their unimodal counterparts.

Additional Qualitative Results Based on Perturbation Types

In this section, we provide additional qualitative examples of our attack against the MMBT-Hateful Memes model and the Pythia-VQA model in Figure 6.5 and Figure 6.6, respectively. In Section 6.4.4, we discussed a few samples from MUROAN from the Hateful Memes dataset. We further discuss more samples

²Note: samples (b) and (c) are from the Hateful Memes dataset [24], which some readers may find distressing.

from the VQA dataset in addition to some samples from the Hateful Memes dataset in this section.




	who knew that this country is full of white trash	Non-hateful
	islam is a religion of peace stop criticizing my religion	Non-hateful
	told girlfriend that mom is deaf so speak loud and slow told mom that girlfriend is retarded	Non-hateful

Figure 6.5: Additional Samples from the Hateful Memes dataset.

Figure 6.5 depicts three samples from the MMBT-Hateful Memes baseline. The first sample depicts the case where only the text is needed to be manipulated to decouple the input modalities in a sample. The second example depicts the case where only a part of the image needs to be manipulated to decouple the modalities in a sample. The third example depicts the case where both image and the text need to be manipulated to decouple the modalities in a sample.

Figure 6.6 depicts three samples from the Pythia-VQA baseline. In this case, the objective is to fool the DMM so as to output a wrong answer (as opposed to a wrong label in the Hateful Memes case). We observed a similar trend in case of VQA as well, as noted in Section 6.4.4. In some cases (such as the first sample and the second sample in Figure 6.6), it was sufficient to only manipulate one of the input modalities to decouple the input modalities in a sample. In some

	what object is depicted in this picture?	Original Answer: Stop sign Adverarial answer: House
	what sports team is mentioned?	Original Answer: Red sox Adverarial answer: Yes
	what are the color of the lines on the court?	Original Answer: Red and blue Adverarial answer: Red

Figure 6.6: Additional Samples from the VQA dataset.

cases though, both modalities had to be manipulated for decoupling them (such as the third sample in Figure 6.6).

Furthermore, to study the versatility of MUROAN regarding the perturbation type, we performed another experiment wherein we used two additional perturbation types apart from the occlusion-based perturbation used throughout the paper: (1) Random noise-based perturbation, and (2) Gradient-based perturbation. Samples from this experiment are depicted in Figure 6.7. It can be observed that the perturbation types can be chosen in MUROAN based on the application domain.

Quantitative Robustness Analysis of DMMs

We have discussed in Section 6.4.3 about how our attack can be used to study the robustness of several DMMs. In this section, we use our attack to study and compare the robustness of two baseline DMMs, Late Fusion and MMBT,

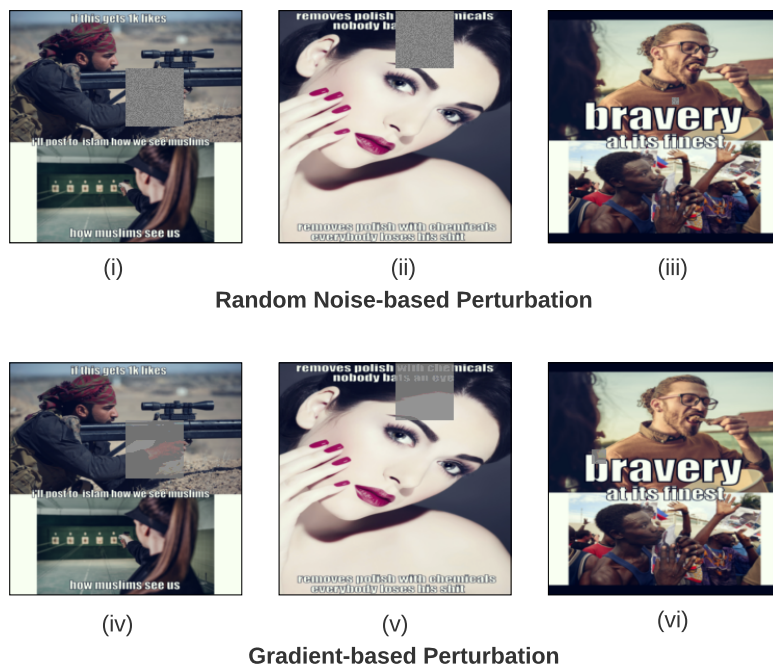


Figure 6.7: Samples from the hateful memes dataset depicting random noise-based and gradient-based perturbation types.

discussed in our paper.

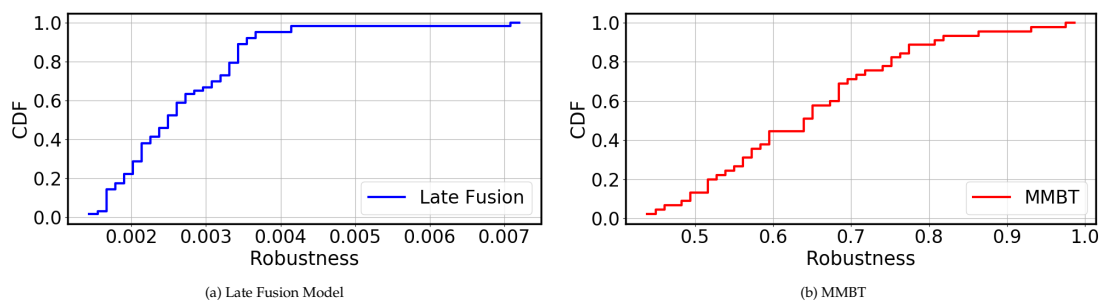


Figure 6.8: CDF of robustness of Late Fusion model and MMBT against MUROAN decoupling attack algorithm.

In this experiment, our objective is to compare the two DMMs that are trained for the same task to determine which DMM is more robust against our attack. In this way, we can use MUROAN to additionally compare DMMs in terms of

their robustness. We study and compare the robustness of the two DMMs both trained on the Hateful Memes dataset based on the *robustness metric* ψ [193]. Model robustness is defined as follows.

$$\psi(x) = \frac{1}{\max_{\delta \in \text{set}} D_{KL}(P(x), P(x + \delta))} \quad (6.4)$$

Equation 6.4 uses the Kullback–Leibler divergence loss (D_{KL}) to depict the divergence between the probability distributions of the original samples and the adversarial samples generated by MUROAN decoupling attack algorithm. In other words, the D_{KL} is higher for a model, for which the adversarial samples are further from the original distribution, indicating stronger robustness. In this experiment, we compared the robustness of the MMBT model to the Late Fusion model, where both DMMs were trained on the same Hateful Memes dataset. The distribution of the robustness the two DMMs as calculated by Equation 6.4 based on our attack is depicted in Figures 6.8a and 6.8b, respectively. We found that the MMBT model is significantly more robust than the Late Fusion model, as can be observed from the Figure 6.8. The mean robustness of the MMBT model was found to be $\psi = 0.65$ and the mean robustness of the Late Fusion model was found to be $\psi = 0.003$. The higher robustness of the MMBT model could be attributed to the way the fusion is achieved in this DMM, using the more sophisticated self-attention mechanism of the transformer [45], while the Late Fusion model uses the element-wise addition. Thus, the robustness metric in this experiment could also indicate the strength of the fusion mechanism. In this way, the robustness of the state-of-the-art DMMs can be quantitatively measured using MUROAN.

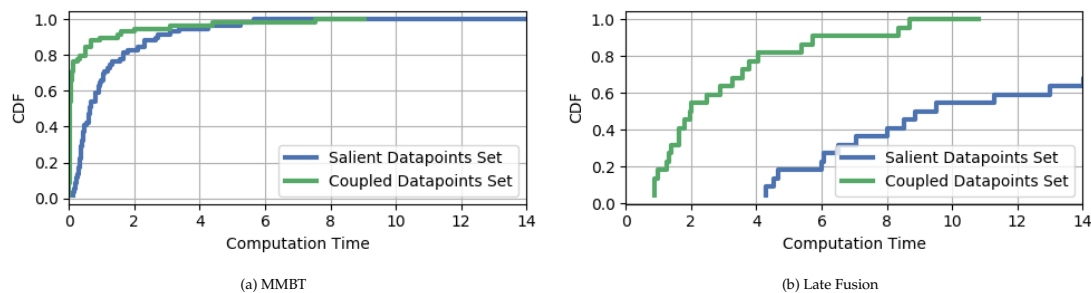


Figure 6.9: Computation cost of MUROAN

6.4.5 Computation Cost of MUROAN

In this section, we provide details about the computation cost of MUROAN. We computed the cost of two primary operations in MUROAN, i.e., computing the salient datapoints set (Equation 6.2) and computing the coupled datapoints set (Equation 6.3).

We study the time taken by MUROAN to compute S_n and C_n for all the samples in test dataset of the Hateful Memes dataset, for the MMBT and late fusion DMMs. Figure 6.9 depicts the CDFs of the computation cost in terms of time in seconds. For MMBT model, the mean S_n time is 1.36 seconds and mean C_n time is 0.59 seconds. For late fusion model, the mean S_n time is 13.93 seconds and mean C_n time is 3.40 seconds. We note that in both DMMs, the cost to compute the coupled datapoints set is significantly lower than the cost to compute the salient datapoints set, as the search in the latter case is on a smaller subset of datapoints. For end-to-end models such as MMBT, the overall computation cost for MUROAN is significantly low for the more traditional architecture based on late fusion methodology.

6.5 Conclusion

In this task, we have studied the robustness of DMMs against multimodal decoupling attacks that are aimed at compromising the fusion mechanism of DMMs. We have introduced a new framework called MUROAN for studying the robustness of DMMs, which consists of a unified view of the DMMs that exposes the fusion embedding, and an algorithm for decoupling the input modalities. Our experiment shows that MUROAN is very effective in attacking several existing multimodal models successfully. MUROAN paves the way for studying the robustness of DMMs via decoupling input modalities in the future.

Discussion

In this chapter, we will discuss potential limitations of the research conducted in this dissertation and point out the promising directions for this research.

7.1 Understanding and Detecting Cyberbullying in Images

In this section, we discuss some limitations and potential enhancements of our work. It should be noted that this work represents the first step towards understanding and identifying the visual factors of cyberbullying in images, and demonstrate that it can be effectively detected based on these factors.

Known Biases in MTurk Surveys. We have used Amazon MTurk as the platform to annotate images in our dataset and to carry out our user studies. Although MTurk provides a convenient method for researchers to enlist high-quality participants online, it also has certain well-known issues that may affect the data collected through it. In the following, we discuss these issues along with how they may have affected the studies in this work. As MTurk is quite

convenient, it follows convenience sampling techniques [194, 195] to enlist participants. Therefore, some participants may not fully representative of the entire population that uses the Internet and hence may not have encountered real-world cyberbullying. In our data collection, MTurk may have introduced some bias towards US-based participants. Common method bias [196] could also be introduced in MTurk studies, wherein self-reported responses may lead to spurious effects. Besides, participants in our study may have some inaccurate knowledge of cyberbullying, which may have caused additional bias in their responses towards our data collection and user experiments.

Different Contexts of Cyberbullying. Cyberbullying is a complex issue, having different contexts. The conventional context of cyberbullying is text-based cyberbullying, which has been well studied and its factors have been extensively cataloged by existing work. A step ahead from this conventional context of cyberbullying is the context of cyberbullying in images, which is the focus of this work. More complex contexts of cyberbullying involve cyberbullying scenarios associated with both images and text. Further contexts of cyberbullying involves videos (i.e., image streams and speech), where we believe our work could also be useful for addressing cyberbullying in the visual part of the video context. As part of our future work, we plan to study those more complicated cyberbullying contexts.

Broadening of Social Factor. In our work, we found attributes, such as anti-LGBT symbols, under the social factor were used for cyberbullying in images. Especially, we found that many images that depicted the anti-LGBT attribute portrayed defacement of the pride symbol. While anti-LGBT is an important attribute of the social factor, we note that there are other attributes under this factor too, such as hate symbols and memes portraying racism against Black

and Asian communities, sexism against women, and religious bigotry. In our dataset, we could not find images portraying these other attributes of the social factor. As part of our future work, we plan to carry out a new study wherein we will broaden attributes of the social factor, and study their effects on cyberbullying in images.

Enabling Existing Detectors to Detect Cyberbullying in Images. We have discussed our finding (in Section 3.3) that the existing state-of-the-art offensive image detectors (e.g. Google Cloud Vision API, Amazon Rekognition, and Clarifai NSFW) cannot effectively detect cyberbullying in images. Through our work, we aim to provide insights into the phenomenon of cyberbullying in images and potentially facilitate those existing offensive image detectors to offer the capability for detecting cyberbullying in images. In this regard, we would suggest two possible ways for building such a capability: (1) training detection models based on new cyberbullying image datasets (like the dataset we have created); and (2) adopting multimodal classifiers with respect to the visual cyberbullying factors (as we have identified in this work) for the detection of cyberbullying in images, since we found that the multimodal classifier is the most effective classifier for detecting cyberbullying in images based on our measurement.

Adoption and Deployment. Current techniques of preventing cyberbullying in social networks, especially cyberbullying in images is limited to reporting and flagging down such images and posts by social network users themselves. In addition to cyberbullying, other online crimes such as online hate [197, 198, 199, 200], pornography [201], grooming [202] and trolling [153] have been identified as dangerous threats. Preliminary research in the automatic detection of these threats have gained momentum in recent times. The multimodal classifier

model explored in our work can be combined with systems that defend against these other threats to provide an overall safer online environment. Additionally, the multimodal classifier can be deployed as a mobile app in mobile devices.

Multi-faceted Detection of Cyberbullying in Images. Many online social networks (such as Facebook and Instagram) support multi-faceted information content, such as textual content accompanying with visual content. In this work, we have only focused on cyberbullying image factors identification and classification. In our future work, we intend to augment the cyberbullying factors with textual information and study the role of the combination of visual and textual cyberbullying. We also intend to study the cyberbullying incidents involving a combination of images and texts in a sequential fashion, so that timely intervention can be possible. In this direction, we intend to discover new factors of cyberbullying involving both textual and visual information. Another future direction that we plan on studying is the issue of revenge-porn [203]. This issue involves a perpetrator who shares revealing or sexually explicit images or videos of a victim online. Due to its offensive and harassing nature, revenge-porn is emerging as a new image-based cyberbullying issue. This issue may be characterized by specific factors that are different from traditional pornography, due to which current offensive content detectors may mis-classify images with this issue. As future work, we intend to study this issue and discover its factors, so that the existing offensive content detectors can be made capable of detecting it in online images.

Adversarial Manipulation of Predictions. Another direction that we intend to explore is the protection of deep-learning based classifiers from adversarial attacks [204, 40]. These attacks are specifically crafted to “fool” deep learning based systems into outputting erroneous predictions. Specifically, we intend

to further explore adversarial manipulations that are aimed at compromising multimodal classifier-based systems. Since our current work and future work would use multimodal machine learning for detecting cyberbullying in images and for intervention, we believe it is highly important to make such models more resistant to such attacks.

Ethical Issues. Our deep learning models have been trained on our dataset of cyberbullying images and our data collection task has been approved by IRB. We intend to make our dataset publicly available. However, we have also found that our dataset may contain some potentially extremely sensitive images, such as images with great violence against children. Therefore, we plan to exclude such extremely sensitive images from our shared dataset. Furthermore, in this chapter, we have attached a few samples of cyberbullying images to illustrate certain concepts so that readers can better understand our paper. We have applied masks over the human subjects' eyes in all attached images to protect their privacy. We do not intend to distribute any sensitive images or leak the human subjects' privacy.

7.2 Detecting and Explaining Traditional Online Hate Speech

In this chapter, we have studied the recent phenomena of hate speech triggered by the COVID-19 pandemic. We have focused our study on the hate-speech in Twitter against Asian community and old people. We have trained a BERT-based model to detect hate-speech based on the datasets in this chapter and used the multi-headed attention mechanism of BERT to discover novel key-

words (186 keywords targeting the Asian community and 100 keywords targeting older people) using our strategy. Further, we have discussed how BERT could be learning longer distance attentions based on the underlying distribution of training data, and found that such attentions are learned in the earlier layers for the Asian-hate dataset and later layers for the Boomer-hate dataset. We have introduced a strategy to study whether BERT is learning hate-speech detection based on existing definitions of hate-speech. We have learned that in the case of Asian-hate dataset, BERT focuses on varied attention between several words, whereas in the case of the Boomer-hate dataset, BERT focuses on certain word associations to detect hate-speech.

7.3 Towards Understanding and Mitigating New Waves of Online Hate

In this chapter, we have studied in-depth, the nature of the new waves of online hate. We discovered that there are eruptions of new waves of online hate with large-scale events, which are of a different nature than traditional online hate. We also discovered that the new waves of online hate consist of different representations, such as image and text based hateful memes. Our measurement analysis of existing systems and models of online hate detection reveals that they are vastly limited against new waves of online hate. Informed by our findings, we introduced our framework, AZL, for the detection of new waves of online hate. Our framework is based on an attribute-based, zero shot learning paradigm using entailment to detect new waves of online hate. We train our detection model on traditional hate and run inference on four different types of

new waves of online hate, i.e., Asian-hate, mask, vaccine and boomer hate. Our evaluation shows that our framework achieves a huge improvement in the detection of new waves of online hate over the existing systems, with a weighted average F1 score of 76.52% for Asian-hate.

7.4 Robustness of Cyberharassment Detection Models

In this chapter, we have focused on DMMs that mainly operate on image and text modalities as inputs. We chose this type of DMMs since it could represent different compositions of inputs (i.e., a continuous input and a discrete input). Our approach however can be generalized to incorporate any other types of DMMs, considering compositions of other inputs including speech and video modalities.

In conclusion, we have studied the robustness of DMMs against multimodal decoupling attacks that are aimed at compromising the fusion mechanism of DMMs. We have introduced a new framework called MUROAN for studying the robustness of DMMs, which consists of a unified view of the DMMs that exposes the fusion embedding, and an algorithm for decoupling the input modalities. Our experiment regarding adversarial training shows that it does not improve the robustness against our decoupling attacks. MUROAN paves the way for studying the robustness of DMMs via decoupling input modalities in the future.

Conclusion

This dissertation devotes to exploring how AI/ML can be used to address on-line cyberharassment, and how robustness of such cyberharassment detection models can be studied and improved.

We focused on two critical issues related to cyberharassment in today's day and age - visual cyberbullying and online hate. In the area of visual cyberbullying, we conducted a large-scale analysis of existing state-of-the-art offensive image detectors against cyberbullying images from our dataset, and found that they are significantly limited in detecting cyberbullying images. We then analyze our dataset and catalog five important factors if visual cyberbullying. We formulated a multimodal model to detect visual cyberbullying, and our model achieves state-of-the-art accuracy of 93.36% in detection of cyberbullying images.

In the area of traditional online hate, we focused on detecting this traditional hate tweets online, as well as explaining the reason for the hate, by pinpointing the exact words and phrases involved in causing the hate. We used the BERT attention model for detection and introduce a novel mechanism to explain and

pinpoint the words based on attention mechanism. Our evaluation results imply that our detection and explanation methods are very effective in detecting as well as explaining and thus controlling traditional online hate.

Furthermore in the area of online hate, we then addressed the problem of new waves of online hate. We carried out an in-depth analysis of online hate witnessed during the COVID-19 pandemic, and made three important findings - (i) large-scale events give rise to new waves of online hate, (ii) these new waves contain new representation in the way online hate is expressed, and (iii) existing online hate detection systems are significantly ineffective in detecting new waves of online hate. Informed by our findings, we introduced our framework to detect new waves of online hate, which is based on an attribute-based zero shot learning paradigm, using textual entailment for training and inference. Our model achieves new state-of-the-art results on four totally unseen new waves of online hate, and vastly outperform the existing systems.

We finally focused on studying the robustness of AI/ML models that are used in cyberharassment detection, and especially considered multimodal models. We identified the fusion mechanism of these models are a core component, and introduced our custom attacks that focus on decoupling this fusion. Our attacks achieve state-of-the-art performance in compromising multimodal models, compared to traditional unimodal attacks.

In summary, the contributions made by this dissertation are as follows.

- Developing a multimodal model for visual cyberbullying detection based on visual factors and evaluating its performance on real-world cyberbullying images
- Developing a BERT-based detection and explanation model for detection

and control of traditional cyberbullying and evaluating its performance on traditional online hate datasets

- Developing an attribute-based zero shot learning using textual entailment for detecting new waves of online hate and evaluating it on four new waves datasets
- Developing multimodal attacks for studying the robustness of multimodal models and evaluating its effectiveness on state-of-the-art multimodal models and comparisons with traditional unimodal attacks

We expect our cyberharassment detection technologies to provide suitable protections against current cyberharassment issues such as cyberbullying and online hate, which are widely propagated in today's Internet. We also envision that our systems can offer effective mitigation against new issues of cyberharassment that are yet unseen. Finally, we hope that our success in studying the robustness of cyberharassment-related AI/ML models can improve the resilience of these models in real-world scenarios.

Bibliography

- [1] Pew research center, july, 2017, “online harassment 2017”. <https://www.pewresearch.org/internet/2017/07/11/online-harassment-2017/>. Accessed: 2022-03-03.
- [2] Kurt Thomas, Devdatta Akhawe, Michael Bailey, Dan Boneh, Elie Bursztein, Sunny Consolvo, Nicola Dell, Zakir Durumeric, Patrick Gage Kelley, Deepak Kumar, et al. Sok: Hate, harassment, and the changing landscape of online abuse. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 247–267. IEEE, 2021.
- [3] Microsoft, “civility, safety & interaction online”. <https://news.microsoft.com/wp-content/uploads/prod/sites/421/2020/02/Digital-Civility-2020-Global-Report.pdf>. Accessed: 2022-03-03.
- [4] Bing He, Caleb Ziems, Sandeep Soni, Naren Ramakrishnan, Diyi Yang, and Srijan Kumar. Racism is a virus: anti-asian hate and counterspeech in social media during the covid-19 crisis. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 90–94, 2021.
- [5] Pamela P Chiang. Anti-asian racism, responses, and the impact on asian americans’ lives: a social-ecological perspective. In *COVID-19*, pages 215–229. Routledge, 2020.
- [6] Hee An Choi and Othelia EunKyoung Lee. To mask or to unmask, that is the question: Facemasks and anti-asian violence during covid-19. *Journal of human rights and social work*, 6(3):237–245, 2021.
- [7] Shihan Wang, Marijn Schraagen, Erik Tjong Kim Sang, and Mehdi Dastani. Public sentiment on governmental covid-19 measures in dutch social media. 2020.

- [8] Claire Wardle and Eric Singerman. Too little, too late: social media companies' failure to tackle vaccine misinformation poses a real threat. *bmj*, 372, 2021.
- [9] David A Broniatowski, Mark Dredze, and John W Ayers. "first do no harm": Effective communication about covid-19 vaccines, 2021.
- [10] Macguire, E., Anti-Asian Hate Continues to Spread Online Amid COVID-19 Pandemic, in Al-Jazeera. <https://www.aljazeera.com/news/2020/04/anti-asian-hate-continues-spread-online-covid-19-pandemic-200405063015286.html>, 2020.
- [11] Whalen, A., What is Boomer Remover and Why is it Making People So Angry?, in Newsweek. <https://www.newsweek.com/boomer-remover-meme-trends-virus-coronavirus-social-media-covid-19-baby-boomers-1492190>, 2020.
- [12] Mehta, I., Twitter Sees 900% Increase in Hate Speech Towards China – because Coronavirus, in The Next Web. <https://thenextweb.com/world/2020/03/27/twitter-sees-900-increase-in-hate-speech-towards-china-because-coronavirus>, 2020.
- [13] Karmen Erjavec and Melita Poler Kovačič. "you don't understand, this is a new war!" analysis of hate speech in news web sites' comments. *Mass Communication and Society*, 15(6):899–920, 2012.
- [14] Jennifer L Lambe. Who wants to censor pornography and hate speech? *Mass Communication & Society*, 7(3):279–299, 2004.
- [15] Jim Hendler. Web 3.0 emerging. *Computer*, 42(1):111–113, 2009.
- [16] Ronald T Azuma. A survey of augmented reality. *Presence: teleoperators & virtual environments*, 6(4):355–385, 1997.
- [17] Metaverse. <https://about.facebook.com/meta/>. Accessed: 2022-03-13.
- [18] Georgios K Pitsilis, Heri Ramampiaro, and Helge Langseth. Effective hate-speech detection in twitter data using recurrent neural networks. *Applied Intelligence*, 48(12):4730–4742, 2018.
- [19] Areej Al-Hassan and Hmood Al-Dossari. Detection of hate speech in social networks: a survey on multilingual corpus. In *6th international conference on computer science and information technology*, volume 10, 2019.
- [20] Joshua Uyheng and Kathleen M Carley. Bots and online hate during the covid-19 pandemic: case studies in the united states and the philippines. *Journal of computational social science*, 3(2):445–468, 2020.

- [21] He Bing, Soni Sandeep, Ziem, Caleb and Kumar Srijan. Racism is a virus: Anti-asian hate and counterhate in social media during the covid-19 crisis. *arXiv preprint arXiv:2005.12423*, 2020.
- [22] Nicolás Velásquez, R Leahy, N Johnson Restrepo, Yonatan Lupu, R Sear, N Gabriel, Omkant Jha, B Goldberg, and NF Johnson. Hate multiverse spreads malicious covid-19 content online beyond individual platform control. *arXiv preprint arXiv:2004.00673*, 2020.
- [23] Emilio Ferrara, Stefano Cresci, and Luca Luceri. Misinformation, manipulation, and abuse on social media in the era of covid-19. *Journal of Computational Social Science*, 3(2):271–277, 2020.
- [24] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *arXiv preprint arXiv:2005.04790*, 2020.
- [25] Yuhao Du, Muhammad Aamir Masood, and Kenneth Joseph. Understanding visual memes: An empirical analysis of text superimposed on memes shared on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 153–164, 2020.
- [26] Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. On the origins of memes by means of fringe web communities. In *Proceedings of the Internet Measurement Conference 2018*, pages 188–202, 2018.
- [27] How anti-Asian activity online set the stage for real-world violence. <https://www.independent.co.uk/news/world/americas/anti-asian-online-hate-crime-real-world-b1820031.html>, 2021.
- [28] Don't Blame Bat Soup for the Coronavirus. <https://foreignpolicy.com/2020/01/27/coronavirus-covid19-dont-blame-bat-soup-for-the-virus/>, 2020.
- [29] Morbid 'boomer remover' coronavirus meme only makes millennials seem more awful. <https://nypost.com/2020/03/19/morbid-boomer-remover-coronavirus-meme-only-makes-millennials-seem-more-awful/>, 2020.
- [30] Hate crimes surge during presidential elections. so far 2020 isn't any different. <https://www.sandiegouniontribune.com/news/public-safety/story/2020-10-31/hate-crimes-surge-presidential-elections>. Accessed: 2022-03-13.

- [31] Bristol hate crime reports spike after black lives matter protests. <https://www.bbc.com/news/uk-england-bristol-54467002>. Accessed: 2022-03-13.
- [32] ‘a tsunami of hate’: The covid-19 hate speech pandemic. <https://www.humanrightspulse.com/mastercontentblog/a-tsunami-of-hate-the-covid-19-hate-speech-pandemic>. Accessed: 2022-03-14.
- [33] Twitter deletes 125,000 isis accounts and expands anti-terror teams. <https://www.theguardian.com/technology/2016/feb/05/twitter-deletes-isis-accounts-terrorism-online>. Accessed: 2022-03-03.
- [34] Twitter reportedly won’t use an algorithm to crack down on white supremacists because some gop politicians could end up getting barred too. <https://www.businessinsider.com/twitter-algorithm-crackdown-white-supremacy-gop-politicians-report-2019-4>. Accessed: 2022-03-03.
- [35] Facebook is pretty good at catching nudity and trolls. it’s still struggling to stop hate speech. <https://slate.com/technology/2018/05/facebook-is-pretty-good-at-catching-and-deleting-graphic-content-and-nudity-hate-speech-not-so-much.html>. Accessed: 2022-03-03.
- [36] Twitter, facebook and others are failing to stop anti-asian hate. <https://www.cnet.com/news/politics/twitter-facebook-and-others-are-failing-to-stop-anti-asian-hate/>. Accessed: 2022-03-13.
- [37] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.
- [38] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE, 2017.
- [39] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.
- [40] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, pages 372–387. IEEE, 2016.

- [41] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. *arXiv*, pages arXiv–1907, 2019.
- [42] Yahoo NSFW, 2020. https://github.com/yahoo/open_nsfw.
- [43] DeepAI, 2020. <https://deepai.org/>.
- [44] 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*, 2018.
- [45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.
- [46] Harmful content can evolve quickly. our new ai system adapts to tackle it. <https://ai.facebook.com/blog/harmful-content-can-evolve-quickly-our-new-ai-system-adapts-to-tackle-it/>. Accessed: 2022-03-13.
- [47] Perspective API, 2020. <https://www.perspectiveapi.com>.
- [48] Google Cloud Vision API, 2020. <https://cloud.google.com/vision/>.
- [49] Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, and Davide Testuggine. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*, 2019.
- [50] Berkan Demirel, Ramazan Gokberk Cinbis, and Nazli Ikizler-Cinbis. Attributes2classname: A discriminative model for attribute-based unsupervised zero-shot learning. In *Proceedings of the IEEE international conference on computer vision*, pages 1232–1241, 2017.
- [51] Wei Wang, Vincent W Zheng, Han Yu, and Chunyan Miao. A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–37, 2019.
- [52] 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.
- [53] Robert Slonje and Peter K Smith. Cyberbullying: Another main type of bullying? *Scandinavian journal of psychology*, 49(2):147–154, 2008.

- [54] Dominic DiFranzo, Samuel Hardman Taylor, Francesca Kazerooni, Olivia D Wherry, and Natalya N Bazarova. Upstanding by design: Bystander intervention in cyberbullying. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 2018.
- [55] Rajitha Kota, Shari Schoohs, Meghan Benson, and Megan A Moreno. Characterizing cyberbullying among college students: Hacking, dirty laundry, and mocking. *Societies*, 4(4):549–560, 2014.
- [56] Ersilia Menesini and Annalaura Nocentini. Cyberbullying definition and measurement: Some critical considerations. *Zeitschrift für Psychologie/Journal of Psychology*, 217(4):230–232, 2009.
- [57] Peter K Smith, Jess Mahdavi, Manuel Carvalho, and Neil Tippett. An investigation into cyberbullying, its forms, awareness and impact, and the relationship between age and gender in cyberbullying. *Research Brief No. RBX03-06*. London: DfES, 2006.
- [58] Robin M Kowalski and Susan P Limber. Psychological, physical, and academic correlates of cyberbullying and traditional bullying. *Journal of Adolescent Health*, 53(1):S13–S20, 2013.
- [59] Dawei Yin, Zhenzhen Xue, Liangjie Hong, Brian D Davison, April Kontostathis, and Lynne Edwards. Detection of harassment on web 2.0. *Proceedings of the Content Analysis in the WEB*, 2:1–7, 2009.
- [60] Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(3):18, 2012.
- [61] Kelly Reynolds, April Kontostathis, and Lynne Edwards. Using machine learning to detect cyberbullying. In *Machine learning and applications and workshops (ICMLA), 2011 10th International Conference on*, volume 2, pages 241–244. IEEE, 2011.
- [62] Karthik Dinakar, Roi Reichart, and Henry Lieberman. Modeling the detection of textual cyberbullying. *The Social Mobile Web*, 11(02):11–17, 2011.
- [63] April Kontostathis, Kelly Reynolds, Andy Garron, and Lynne Edwards. Detecting cyberbullying: query terms and techniques. In *Proceedings of the 5th annual acm web science conference*, pages 195–204. ACM, 2013.
- [64] Junming Sui. *Understanding and fighting bullying with machine learning*. PhD thesis, Ph. D. dissertation, The Univ. of Wisconsin-Madison, WI, USA, 2015.

- [65] Rui Zhao, Anna Zhou, and Kezhi Mao. Automatic detection of cyberbullying on social networks based on bullying features. In *Proceedings of the 17th international conference on distributed computing and networking*, page 43. ACM, 2016.
- [66] Maral Dadvar, de FMG Jong, Roeland Ordelman, and Dolf Trieschnigg. Improved cyberbullying detection using gender information. In *Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012)*. University of Ghent, 2012.
- [67] Qianjia Huang, Vivek Kumar Singh, and Pradeep Kumar Atrey. Cyberbullying detection using social and textual analysis. In *Proceedings of the 3rd International Workshop on Socially-Aware Multimedia*, pages 3–6. ACM, 2014.
- [68] Vivek K Singh, Qianjia Huang, and Pradeep K Atrey. Cyberbullying detection using probabilistic socio-textual information fusion. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 884–887. IEEE Press, 2016.
- [69] Claudia I Flores-Saviaga, Brian C Keegan, and Saiph Savage. Mobilizing the trump train: Understanding collective action in a political trolling community. In *Twelfth International AAAI Conference on Web and Social Media*, 2018.
- [70] Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. Spread of hate speech in online social media. In *Proceedings of the 10th ACM Conference on Web Science*, pages 173–182. ACM, 2019.
- [71] Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. A benchmark dataset for learning to intervene in online hate speech. *arXiv preprint arXiv:1909.04251*, 2019.
- [72] Jan Chorowski, Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. End-to-end continuous speech recognition using attention-based recurrent nn: First results. *arXiv preprint arXiv:1412.1602*, 2014.
- [73] 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.
- [74] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

- [75] Jianpeng Cheng, Li Dong, and Mirella Lapata. Long short-term memory-networks for machine reading. *arXiv preprint arXiv:1601.06733*, 2016.
- [76] Ankur P Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. *arXiv preprint arXiv:1606.01933*, 2016.
- [77] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.
- [78] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.
- [79] Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. Linguistically-informed self-attention for semantic role labeling. *arXiv preprint arXiv:1804.08199*, 2018.
- [80] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [81] Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. Combining language and vision with a multimodal skip-gram model. *arXiv preprint arXiv:1501.02598*, 2015.
- [82] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.
- [83] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.
- [84] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017.
- [85] Ji Gao, Jack Lanchantin, Mary Lou Soffa, and Yanjun Qi. Black-box generation of adversarial text sequences to evade deep learning classifiers. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 50–56. IEEE, 2018.
- [86] Caitlin R Costello and Danielle E Ramo. Social media and substance use: what should we be recommending to teens and their parents? *Journal of Adolescent Health*, 60(6):629–630, 2017.

- [87] Pew Research Center, 2020. <http://www.pewresearch.org/>.
- [88] Linda Beckman, Curt Hagquist, and Lisa Hellström. Does the association with psychosomatic health problems differ between cyberbullying and traditional bullying? *Emotional and behavioural difficulties*, 17(3-4):421–434, 2012.
- [89] Andre Sourander, Anat Brunstein Klomek, Maria Ikonen, Jarna Lindroos, Terhi Luntamo, Merja Koskelainen, Terja Ristkari, and Hans Helenius. Psychosocial risk factors associated with cyberbullying among adolescents: A population-based study. *Archives of general psychiatry*, 67(7):720–728, 2010.
- [90] Amanda Lenhart, Mary Madden, Aaron Smith, Kristen Purcell, Kathryn Zickuhr, and Lee Rainie. Teens, kindness and cruelty on social network sites: How american teens navigate the new world of “digital citizenship”. *Pew Internet & American Life Project*, 2011.
- [91] Julian J Dooley, Therese Shaw, and Donna Cross. The association between the mental health and behavioural problems of students and their reactions to cyber-victimization. *European Journal of Developmental Psychology*, 9(2):275–289, 2012.
- [92] 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*, pages 365–383, 2013.
- [93] Facebook, 2020. <https://www.facebook.com>.
- [94] Instagram, 2020. <https://www.instagram.com/>.
- [95] Twitter, 2020. <https://twitter.com>.
- [96] Flickr. <https://www.flickr.com>.
- [97] Pinterest. <https://www.pinterest.com/>.
- [98] Ersilia Menesini, Annalaura Nocentini, and Pamela Calussi. The measurement of cyberbullying: Dimensional structure and relative item severity and discrimination. *Cyberpsychology, Behavior, and Social Networking*, 14(5):267–274, 2011.
- [99] Cyberbullying: one in two victims suffer from the distribution of embarrassing photos and videos, 2020. www.sciencedaily.com/releases/2012/07/120725090048.htm.

- [100] Maral Dadvar, Roeland Ordelman, Franciska de Jong, and Dolf Trietschnigg. Towards user modelling in the combat against cyberbullying. In *Natural Language Processing and Information Systems*, pages 277–283. Springer, 2012.
- [101] Amazon Comprehend, 2020. <https://aws.amazon.com/comprehend/>.
- [102] IBM Toxic Comment Classifier, 2020. <https://developer.ibm.com/technologies/artificial-intelligence/models/max-toxic-comment-classifier/>.
- [103] Amazon Rekognition, 2020. <https://aws.amazon.com/rekognition/>.
- [104] Clarifai, 2020. <https://www.clarifai.com/>.
- [105] Marilyn A Campbell. Cyber bullying: An old problem in a new guise?. *Australian journal of Guidance and Counselling*, 15(01):68–76, 2005.
- [106] Robin M Kowalski, Gary W Giumetti, Amber N Schroeder, and Micah R Lattanner. Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth. 2014.
- [107] Robin M Kowalski, Susan P Limber, Sue Limber, and Patricia W Agatston. *Cyberbullying: Bullying in the digital age*. John Wiley & Sons, 2012.
- [108] Cyberbullying Stories, 2020. <https://cyberbullying.org/stories>.
- [109] Steven Bird and Edward Loper. Nltk: the natural language toolkit. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, page 31. Association for Computational Linguistics, 2004.
- [110] Christopher Fox. A stop list for general text. In *Acm sigir forum*, volume 24, pages 19–21. ACM, 1989.
- [111] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.
- [112] Justin W Patchin and Sameer Hinduja. *Cyberbullying prevention and response: Expert perspectives*. Routledge, 2012.
- [113] Kilem L Gwet. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC, 2014.
- [114] Justus J Randolph. Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss’ fixed-marginal multirater kappa. *Online submission*, 2005.

- [115] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1812.08008*, 2018.
- [116] Nancy E Willard. *Cyberbullying and cyberthreats: Responding to the challenge of online social aggression, threats, and distress*. Research press, 2007.
- [117] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10. IEEE, 2016.
- [118] Lieve Hamers et al. Similarity measures in scientometric research: The jaccard index versus salton’s cosine formula. *Information Processing and Management*, 25(3):315–18, 1989.
- [119] Adam Kendon. Do gestures communicate? a review. *Research on language and social interaction*, 27(3):175–200, 1994.
- [120] Jürgen Streeck. Gesture as communication i: Its coordination with gaze and speech. *Communications Monographs*, 60(4):275–299, 1993.
- [121] Sameer Hinduja and Justin W Patchin. State cyberbullying laws. *Cyberbullying Research Center*, 2012.
- [122] Robyn M Cooper and Warren J Blumenfeld. Responses to cyberbullying: A descriptive analysis of the frequency of and impact on lgbt and allied youth. *Journal of LGBT Youth*, 9(2):153–177, 2012.
- [123] Sameer Hinduja and Justin W Patchin. Cyberbullying research summary: Bullying, cyberbullying, and sexual orientation. *Cyberbullying Research Center*: http://cyberbullying.org/cyberbullying_sexual_orientation_fact_sheet.pdf, 2011.
- [124] LGBTQ Pride Symbols and Icons, 2020. <https://algbtical.org/20SYMBOLS.htm>.
- [125] Hate on Display Hate Symbols Database, 2020. https://www.adl.org/hate-symbols?cat_id%5B146%5D=146.
- [126] Sameer Hinduja and Justin W Patchin. Bullying, cyberbullying, and suicide. *Archives of Suicide Research*, 14(3):206–221, 2010.
- [127] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. *arXiv preprint arXiv:1412.6856*, 2014.

- [128] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. Ieee, 2009.
- [129] Aditya Mogadala, Marimuthu Kalimuthu, and Dietrich Klakow. Trends in integration of vision and language research: A survey of tasks, datasets, and methods. *arXiv preprint arXiv:1907.09358*, 2019.
- [130] Adam Paszke, Sam Gross, Soumith Chintala, and Gregory Chanan. Pytorch: Tensors and dynamic neural networks in python with strong gpu acceleration. *PyTorch: Tensors and dynamic neural networks in Python with strong GPU acceleration*, 6, 2017.
- [131] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [132] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [133] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.
- [134] Pytorch mobile. <https://pytorch.org/mobile/home>, 2020.
- [135] Srujana Gattupalli, Amir Ghaderi, and Vassilis Athitsos. Evaluation of deep learning based pose estimation for sign language recognition. In *Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, pages 1–7, 2016.
- [136] Mary L McHugh. The chi-square test of independence. *Biochemia medica: Biochemia medica*, 23(2):143–149, 2013.
- [137] Haoti Zhong, Hao Li, Anna Cinzia Squicciarini, Sarah Michele Rajtmajer, Christopher Griffin, David J Miller, and Cornelia Caragea. Content-driven detection of cyberbullying on the instagram social network. In *IJCAI*, pages 3952–3958, 2016.
- [138] Perspective API. <https://www.perspectiveapi.com/#/home>, 2020.
- [139] Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Pooven-dran. Deceiving google’s perspective api built for detecting toxic comments. *arXiv preprint arXiv:1702.08138*, 2017.

- [140] Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. Hotflip: White-box adversarial examples for text classification. *arXiv preprint arXiv:1712.06751*, 2017.
- [141] Ashley Reichelmann, James Hawdon, Matt Costello, John Ryan, Catherine Blaya, Vicente Llorent, Atte Oksanen, Pekka Räsänen, and Izabela Zych. Hate knows no boundaries: Online hate in six nations. *Deviant Behavior*, pages 1–12, 2020.
- [142] Robert King Merton and Robert C Merton. *Social theory and social structure*. Simon and Schuster, 1968.
- [143] Donald P Green, Dara Z Strolovitch, and Janelle S Wong. Defended neighborhoods, integration, and racially motivated crime. *American journal of sociology*, 104(2):372–403, 1998.
- [144] Larry J Ray and David Smith. Hate crime, violence and cultures of racism. 2002.
- [145] David Gadd, Bill Dixon, and Tony Jefferson. Why do they do it? racial harassment in north staffordshire. *Centre for Criminological Research, Keele University*, 2005.
- [146] Bhikhu Parekh et al. Is there a case for banning hate speech? *The content and context of hate speech: Rethinking regulation and responses*, pages 37–56, 2012.
- [147] Paula Fortuna and Sérgio Nunes. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30, 2018.
- [148] Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, pages 14014–14024, 2019.
- [149] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does bert look at? an analysis of bert’s attention. *arXiv preprint arXiv:1906.04341*, 2019.
- [150] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.
- [151] James B Jacobs, Kimberly Potter, et al. *Hate crimes: Criminal law & identity politics*. Oxford University Press on Demand, 1998.

- [152] Samuel Walker. *Hate speech: The history of an American controversy*. U of Nebraska Press, 1994.
- [153] Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*, 2018.
- [154] Thomas Davidson, Dana Warmusley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Eleventh international aai conference on web and social media*, 2017.
- [155] Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. Hate speech dataset from a white supremacy forum. *arXiv preprint arXiv:1809.04444*, 2018.
- [156] Jing Gao, Peng Li, Zhikui Chen, and Jianing Zhang. A survey on deep learning for multimodal data fusion. *Neural Computation*, 32(5):829–864, 2020.
- [157] Nishant Vishwamitra, Ruijia Roger Hu, Feng Luo, Long Cheng, Matthew Costello, and Yin Yang. On analyzing covid-19-related hate speech using bert attention. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 669–676. IEEE, 2020.
- [158] Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio AF Almeida, and Wagner Meira Jr. Auditing radicalization pathways on youtube. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 131–141, 2020.
- [159] Tweepy. <https://www.tweepy.org>. Accessed: 2022-03-13.
- [160] Cdc museum covid-19 timeline. <https://www.cdc.gov/museum/timeline/covid19.html>. Accessed: 2022-03-03.
- [161] 2020-time-capsule-5-the-chinese-virus. <https://www.theatlantic.com/notes/2020/03/2020-time-capsule-5-the-chinese-virus/608260/>, 2020.
- [162] Urban dictionary has a new word for coronavirus screw-ups: Covidiot. <https://nypost.com/2020/03/24/urban-dictionary-has-a-new-word-for-coronavirus-screw-ups-covidiot/>, 2020.
- [163] Coronavirus outbreak: What is “covidots” trending on twitter? <https://www.financialexpress.com/lifestyle/coronavirus-outbreak-what-is-covidots-trending-on-twitter/1907432/>, 2020.

- [164] Trump plans to suspend immigration to u.s. <https://www.nytimes.com/2020/04/20/us/politics/trump-immigration.html>, 2020.
- [165] Snsrape. <https://github.com/JustAnotherArchivist/snsrape>. Accessed: 2022-03-15.
- [166] Joni Salminen, Maximilian Hopf, Shammur A Chowdhury, Soon-gyo Jung, Hind Almerexhi, and Bernard J Jansen. Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences*, 10(1):1–34, 2020.
- [167] Hate Speech. https://transparency.fb.com/policies/community-standards/hate-speech/?from=https%3A%2F%2Fm.facebook.com%2Fcommunitystandards%2Fhate_speech%2F&refsrc=deprecated, 2021.
- [168] Anthony Kay. Tesseract: an open-source optical character recognition engine. *Linux Journal*, 2007(159):2, 2007.
- [169] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv*, 2018.
- [170] Statistics Solutions. How to conduct the wilcoxon sign test, 2020.
- [171] Azure Content Moderator, 2022. <https://azure.microsoft.com/en-us/services/cognitive-services/content-moderator/>.
- [172] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.
- [173] Pete Burnap and Matthew L Williams. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & internet*, 7(2):223–242, 2015.
- [174] Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230, 2015.
- [175] Rohan Badlani, Nishit Asnani, and Manan Rai. Disambiguating sentiment: An ensemble of humour, sarcasm, and hate speech features for sentiment classification. *W-NUT*, 2019:337–345, 2019.
- [176] Rosalynd Southern and Emily Harmer. Twitter, incivility and “everyday” gendered othering: An analysis of tweets sent to uk members of parliament. *Social Science Computer Review*, 39(2):259–275, 2021.

- [177] Emily Harmer and Karen Lumsden. Online othering: An introduction. In *Online othering*, pages 1–33. Springer, 2019.
- [178] Mark S Hamm. Conceptualizing hate crime in a global context. *Hate crime: International perspectives on causes and control*, pages 173–194, 1994.
- [179] Jack Levin and Jack MacDevitt. *Hate crimes: The rising tide of bigotry and bloodshed*. Springer, 2013.
- [180] What Is Othering? <https://www.verywellmind.com/what-is-othering-5084425>, 2022.
- [181] Scott Hammack. The internet loophole: Why threatening speech on-line requires a modification of the courts’ approach to true threats and incitement. *Colum. JL & Soc. Probs.*, 36:65, 2002.
- [182] Sentiment Analyzer. <https://cloud.google.com/natural-language/docs/analyzing-sentiment>, 2022.
- [183] Edward Loper and Steven Bird. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*, 2002.
- [184] Profanity Check. <https://github.com/vzhou842/profanity-check>, 2022.
- [185] Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial nli: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599*, 2019.
- [186] Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*, 2017.
- [187] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015.
- [188] Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Pythia v0. 1: the winning entry to the vqa challenge 2018. *arXiv preprint arXiv:1807.09956*, 2018.
- [189] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3608–3617, 2018.

- [190] Junhyug Noh, Soochan Lee, Beomsu Kim, and Gunhee Kim. Improving occlusion and hard negative handling for single-stage pedestrian detectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 966–974, 2018.
- [191] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.
- [192] Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. Generating natural language adversarial examples. *arXiv preprint arXiv:1804.07998*, 2018.
- [193] Fuxun Yu, Zhuwei Qin, Chenchen Liu, Liang Zhao, Yanzhi Wang, and Xiang Chen. Interpreting and evaluating neural network robustness. *arXiv preprint arXiv:1905.04270*, 2019.
- [194] Oliver C Robinson. Sampling in interview-based qualitative research: A theoretical and practical guide. *Qualitative research in psychology*, 11(1):25–41, 2014.
- [195] Avi Fleischer, Alan D Mead, and Jialin Huang. Inattentive responding in mturk and other online samples. *Industrial and Organizational Psychology*, 8(2):196–202, 2015.
- [196] James M Conway and Charles E Lance. What reviewers should expect from authors regarding common method bias in organizational research. *Journal of Business and Psychology*, 25(3):325–334, 2010.
- [197] Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. Detecting offensive language in social media to protect adolescent online safety. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom)*, pages 71–80. IEEE, 2012.
- [198] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. Mean birds: Detecting aggression and bullying on twitter. In *Proceedings of the 2017 ACM on web science conference*, pages 13–22. ACM, 2017.
- [199] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*, pages 29–30, 2015.

- [200] Mai ElSherief, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna, and Elizabeth Belding. Peer to peer hate: Hate speech instigators and their targets. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12, 2018.
- [201] Kan Yuan, Di Tang, Xiaojing Liao, XiaoFeng Wang, Xuan Feng, Yi Chen, Menghan Sun, Haoran Lu, and Kehuan Zhang. Stealthy porn: Understanding real-world adversarial images for illicit online promotion. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 952–966. IEEE, 2019.
- [202] Marlies Rybnicek, Rainer Poisel, and Simon Tjoa. Facebook watchdog: a research agenda for detecting online grooming and bullying activities. In *Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on*, pages 2854–2859. IEEE, 2013.
- [203] Danielle Keats Citron and Mary Anne Franks. Criminalizing revenge porn. *Wake Forest L. Rev.*, 49:345, 2014.
- [204] Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6:14410–14430, 2018.

ProQuest Number: 29211815

INFORMATION TO ALL USERS

The quality and completeness of this reproduction is dependent on the quality and completeness of the copy made available to ProQuest.



Distributed by ProQuest LLC (2022).

Copyright of the Dissertation is held by the Author unless otherwise noted.

This work may be used in accordance with the terms of the Creative Commons license or other rights statement, as indicated in the copyright statement or in the metadata associated with this work. Unless otherwise specified in the copyright statement or the metadata, all rights are reserved by the copyright holder.

This work is protected against unauthorized copying under Title 17, United States Code and other applicable copyright laws.

Microform Edition where available © ProQuest LLC. No reproduction or digitization of the Microform Edition is authorized without permission of ProQuest LLC.

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