Towards Understanding and Detecting Cyberbullying in Real-world Images

Nishant Vishwamitra*, Hongxin Hu*, Feng Luo† and Long Cheng†
*Computer Science and Engineering, University at Buffalo
†School of Computing, Clemson University
Email: *{nvishwam, hongxinh}@buffalo.edu, †{luofeng, lcheng2}@clemson.edu

***Abstract*—Cyberbullying has become widely recognized as a critical social problem plaguing today's Internet users. This problem involves perpetrators using Internet-based technologies to bully their victims by sharing cyberbullying-related content. To combat this problem, researchers have studied the factors associated with such content and proposed automatic detection techniques based on those factors. However, most of these studies have mainly focused on understanding the factors of textual content, such as comments and text messages, while largely overlooking the misuse of visual content in perpetrating cyberbullying. Recent technological advancements in the way users access the Internet have led to a new cyberbullying paradigm. Perpetrators can use visual media to bully their victims through sending and distributing images with cyberbullying content. As a first step to understand the threat of cyberbullying in images, we report in this paper a comprehensive study on the nature of images used in cyberbullying. We first collect a real-world cyberbullying images dataset with 19,300 valid images. We then analyze the images in our dataset and identify the factors related to cyberbullying images that can be used to build systems to detect cyberbullying in images. Our analysis of factors in cyberbullying images reveals that unlike traditional offensive image content (e.g., violence and nudity), the factors in cyberbullying images tend to be highly contextual. We further demonstrate the effectiveness of the factors by measuring several classifier models based on the identified factors. With respect to the cyberbullying factors identified in our work, the best classifier model based on multimodal classification achieves a mean detection accuracy of 93.36% on our cyberbullying images dataset.**

I. INTRODUCTION

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 [31]. Particularly, according to recent findings from the Pew Research Center [16], 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 [22], [79]. According to a National Crime Prevention Council report, more than 40% of teenagers in U.S. have reported being cyberbullied [61]. Dooley et al. define cyberbullying as "Bullying via the Internet or mobile phone" [39]. 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 [41].

(a) (b)

Fig. 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.

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 [9], Instagram [13] and Twitter [18]) 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 [1] and Pinterest [2], to exclusively focus

Network and Distributed Systems Security (NDSS) Symposium 2021
21-24 February 2021
ISBN 1-891562-66-5
https://dx.doi.org/10.14722/ndss.2021.24260
www.ndss-symposium.org

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 [77], [64]. This enhanced form of cyberbullying perpetrated through images now affects one of every two cyberbullying victims [6].

Figure 1 presents two examples of cyberbullying in text and in an image, respectively. Figure 1 (a) depicts a cyberbullying tweet [55] with the cyberbullying-related words shown in bold (such as 'a**', 'fat', and 'ugly'). Figure 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 [33], [36], [37], and some state-of-the-art detectors for text-based offensive[1] content detection have been developed that are sufficiently effective in combating text-based cyberbullying. For example, on running the text in Figure 1 (a) against three state-of-the-art offensive text detectors namely Google Perspective API [15], Amazon Comprehend [3], and IBM Toxic Comment Classifier [12], all of them are able to detect this text as offensive with very high 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 1 (b) through three state-of-the-art offensive image detectors namely, Google Cloud Vision API [10], Amazon Rekognition [4], and Clarifai NSFW [5], 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 [24], [57], [59]. 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 [39], [41]) 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 work, 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 [19], Clarifai NSFW, DeepAI Content Moderation API [8], and Amazon Rekognition [2]. 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 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 paper 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.,

[1]We have used the term "offensive" here to mean harassing, harmful, toxic, or hateful content.

[2]The 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.

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

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 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 work 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 paper is organized as follows. We first lay down the threat model of our work in Section II. Next, we present our cyberbullying images data collection strategy in Section III. We then present the motivation of our work in Section IV. This is followed by the details of our approach in Section V. We discuss the implementation details of the cyberbullying images classifier models and present the evaluations of those models from different perspectives in Section VI. We discuss some important aspects of our approach in Section VII. This is followed by a discussion of related work in cyberbullying defense in Section VIII. Finally, we conclude our work in Section IX.

II. THREAT MODEL AND SCOPE

Threat Model. In this work, 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 [8], [4], [5]. 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 work, 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.

III. 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.

A. 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 [7], 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.

B. 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 [23] to remove stop words [45] 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 -0.55 [4]). We used these words as the final keywords list to collect potential images of cyberbullying content for our dataset. Table I shows some cyberbullying story samples and the keywords extracted with our methodology.

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

Stories	Extracted Keywords
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 I: Samples of cyberbullying stories and the extracted keywords.

Fig. 2: Image samples that did not have any Regions of Interest (ROIs).

C. 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 III-B. We finally collected 117, 112 images using our data collection methodology. Next, we used an object localization tool called YOLO [71] 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 2. After this step, we were left with 19, 300 images for annotations.

1) 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 [59], [69] 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 B). 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 [47] using the Randolph's κ-measure [70], 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 $\kappa = 0.80$.

IV. 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 II. From Table II, 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 II: Precision and recall of popular offensive image detectors.

After an examination of cyberbullying images annotated by users in our dataset, we found that most of such images are context-aware. Figure 3 depicts two images without cyberbullying context (annotated as non-bullying images by

(i) (ii)

(a) Without cyberbullying context.

(i) (ii)

(b) With cyberbullying context.

Fig. 3: Image context in cyberbullying images.

Image #	Google API	Yahoo NSFW	Clarifai	Deep AI	Amazon
Figure 3a (i)	0.2	0.17	0	0.17	0
Figure 3a (ii)	0.2	0.005	0.05	0.003	0.98
Figure 3b (i)	0.2	0.008	0.01	0.008	0
Figure 3b (ii)	0.2	0.004	0	0	0.97

TABLE III: Detection scores of existing detectors on image samples in Figure 3.

participants) and two other images with cyberbullying context (annotated as bullying images by participants), respectively, from our dataset. The images in Figure 3a only show a possible factor (a demeaning hand gesture or a gun), but without any contextual information. In contrast, Figure 3b shows images that portray these factors with contextual information, such as a person deliberately showing the hand gesture in Figure 3b (i) to the viewer, or the person in Figure 3b (ii) pointing the gun at the viewer. Table III 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 3a (ii) (score = 0.98) and Figure 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 IV. From Table IV, we can first 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,

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 IV: Capabilities of existing detectors and their limitations.

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.

V. 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 V-B); (ii) extract those factors from images (Section V-C); and (iii) examine the usage of those factors in classifier models (Section V-D).

A. Approach Overview

The main components involved in our approach are depicted in Figure 4. We first collect a large dataset of cyberbullying images to study this phenomenon (Figure 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 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 4, Step 2) and *"Classifier Model Measurement"* (Figure 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

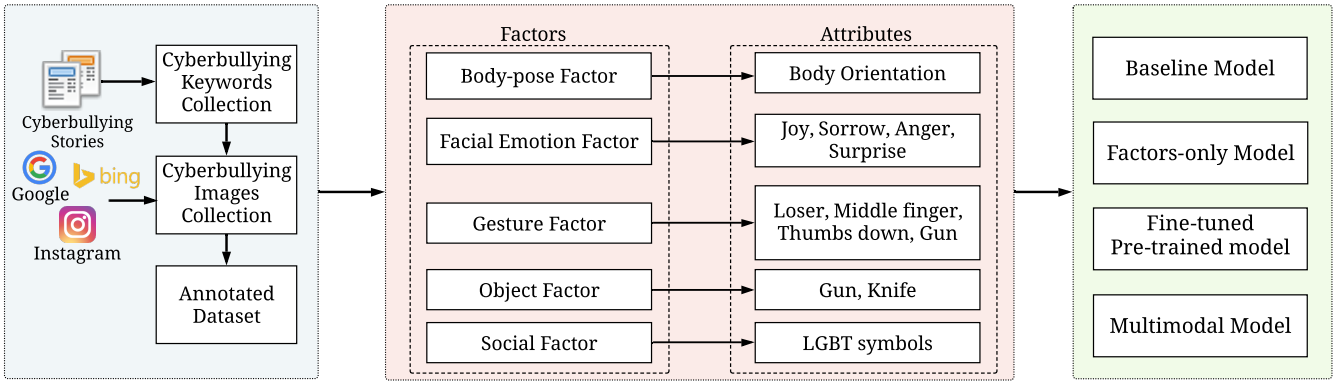


Fig. 4: Approach overview.

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.

B. Factor Identification

Various studies [68], [56], [63] 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 V. We analyze the body-pose [25] of the subject in an image, as prior research [82] has shown that threatening poses are a commonly used tool in cyberbullying. We analyze hand gestures [10] as hand gestures are popular forms of sign language used to convey meaning through images. We study the facial emotion [21] of the subject in images, as facial emotions can convey several meanings to a viewer. We study the objects [71] 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 [48] 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 V). 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 correlation 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 [54], [80]. 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 threat-

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

ening 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 [52]. 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 [30], [51] 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 [14] and racism [11]. 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 VI). 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.

From our analysis, we observe that cyberbullying in images is highly *contextual* in nature, involving very specific factors (outlined in Table V). In our work, we use these factors to train classifier models and demonstrate that they can be

	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 VI: Analysis of correlation of person with threatening object and gesture.

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.

C. 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 5 shows two image samples from our dataset. Figure 5(i) depicts a cyberbullying sample and Figure 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

(i) Cyberbullying Image (ii) Cyberbullying Pose

(iii) Non-cyberbullying Image (iv) Non-cyberbullying Pose

Fig. 5: Cyberbullying Vs. non-cyberbullying body-pose.

threatening object (e.g., gun) at the viewer. However, this is in contrast to Figure 5(iii), whereas the pose depicted in Figure 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 [25] 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.

  We extract the emotions in our dataset using two sources, OpenFace [21] and Google Cloud Vision API [10]. 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 6) depict subjects making mocking or threatening hand gestures, such as the loser gesture (Figure 6 (i)), middle finger (Figure 6 (ii)), thumbs down (Figure 6 (iii)), and gun gesture (Figure 6 (iv)). Hence, we were interested in capturing these harmful gestures we found in cyberbullying images.

(i) Loser (ii) Middle Finger

(iii) Thumbs Down (iv) Finger Gun

Fig. 6: Some hand gestures found in cyberbullying images in our dataset.

We use the tag suggestions by Google Cloud Vision API [10] 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 [51], [50], 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 [71] 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 [14], [11]. 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 work.

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.

D. 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 [87] 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 work, 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 work. 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 $\times$ 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 work 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 [34], 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.

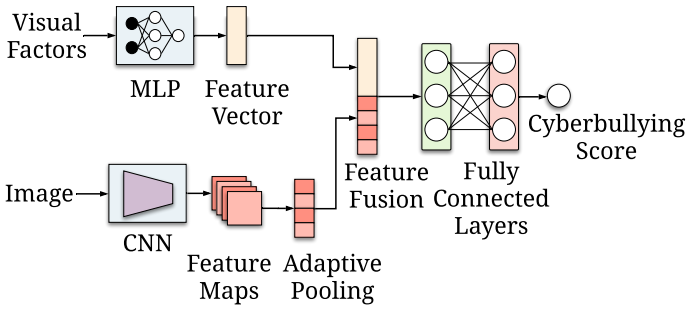


Fig. 7: Multimodal model used in our approach.

4) Multimodal model. In this model, we combine the low level image features (Figure 7, "Image") with the factors identified in this work (Figure 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 [65]. We use the VGG16 pre-trained model for image features (Figure 7, "CNN") and use a multi-layer perceptron model (Figure 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 $\times$ 7. We flatten the convolutional feature maps using adaptive pooling into one-dimensional vector of 512 and fuse it (Figure 7, "Feature Fusion") with the output of the MLP network. We train this model in a joint manner

(Figure 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).

VI. 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 VI-B).
- Demonstrating the effectiveness of our factors in accurately predicting cyberbullying in images, using four classifier models (Figure 10 and Table IX).
- Studying the performance overhead of our model when integrated in mobile devices (Section VI-D).
- Evaluating the false positives of our model on the images depicting the American Sign Language (Section VI-E).
- Validation of our cyberbullying factors with a wider audience (Section VI-F).
- Studying the representativeness of our cyberbullying images dataset (Section VI-G).
- Analyzing the capabilities of the state-of-the-art offensive image detectors with respect to the cyberbullying factors (Section VI-H).

A. Implementation

In this section, we discuss the implementation details of the classifier models for cyberbullying in images. We use the PyTorch framework [67] to train and deploy these models. In our work, we use the VGG-16 network [75] for feature extraction in the models. We use the VGG-16 model that is pre-trained on ImageNet dataset [60] 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.

B. Understanding the Effectiveness of Cyberbullying Factors

In this section, we study in detail the factors of cyberbullying in images identified in this work 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 VII. 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 VII: Frequencies of factors responsible for labeling an image as cyberbullying or non-cyberbullying.

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

The factors frequencies in non-cyberbullying images are depicted in Table VII. 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.

Next, we conduct a study to understand the associations between human level annotations on images and the identified factors. Table VIII depicts the correlations (Spearman ρ) for visual factors and cyberbullying images. In Table VIII, 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.

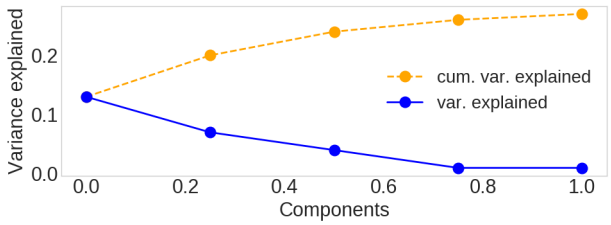


Fig. 8: Scree plot showing proportions of variance and cumulative proportion of variance explained by each component.

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

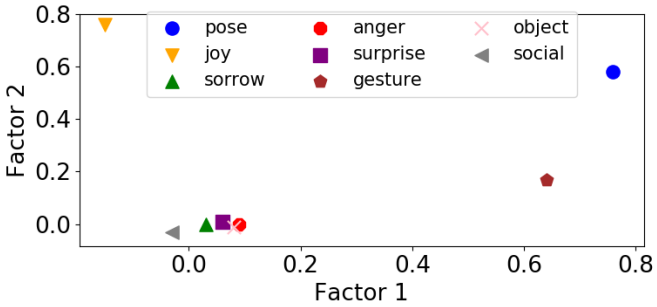
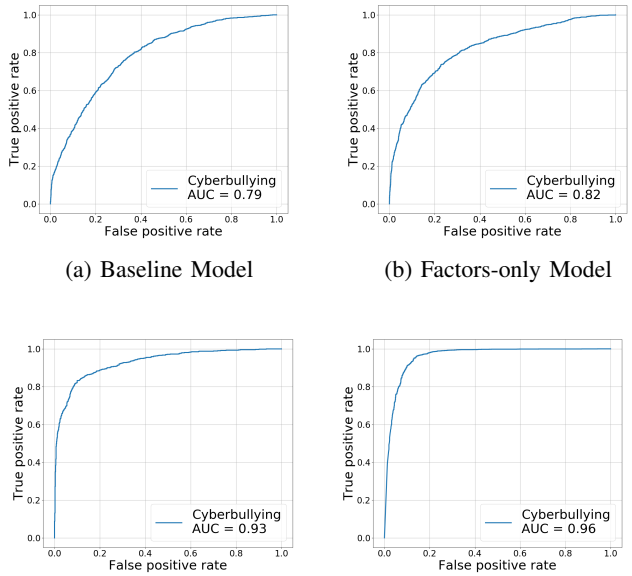


Fig. 9: Factor loadings of the features across two extracted factors.

C. 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) [42] 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 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.



(a) Baseline Model (b) Factors-only Model

(c) Fine-tuned Pre-trained Model (d) Multimodal Model

Fig. 10: ROC analysis of classifier models.

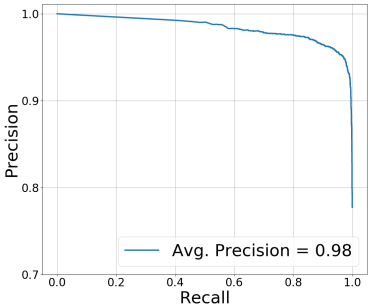


Fig. 11: Precision-recall graph of the multimodal model.

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%
Mutimodal Model	93.36%	94.27%	96.93%

TABLE IX: 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

[5] Here, "factor" refers to EFA factors and not visual factors of cyberbullying.

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 [42].

We find that the baseline model (Table IX, precision = 63.0% and recall = 29.68%) indeed has the lowest performance, indicating that cyberbullying in images is not a problem that can be trivally 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 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 IX, 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 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 IX, precision = 81.40% and recall = 73.70%) does not perform overall better than the 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 IX, precision = 94.27% and recall = 96.93%) among the different classifier models. A high AUC (Figure 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 work and also the low-level image features. A high precision and recall of this model indicates that the visual factors identified in this work 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 11. The PR plot indicates that the multimodal model is able to correctly classify cyberbullying images with high precision.

D. 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 [17] 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 12.

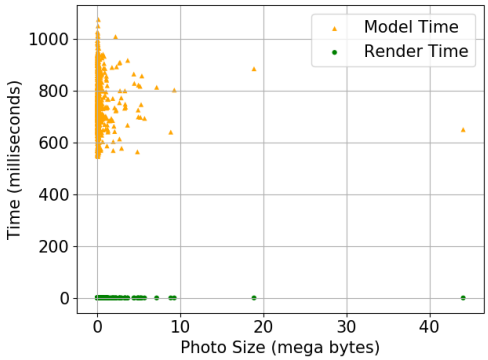


Fig. 12: Overhead evaluation of the multimodal model integrated into an Android application.

From Figure 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 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.

E. 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 [46]. Figure 13 depicts two samples from this dataset.

Fig. 13: Image samples from the ASL 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 detects 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.

F. Validation of Cyberbullying Factors with a Wider Audience

In our work we introduce new factors of cyberbullying in images, as discussed in Section V-B 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 14 in Appendix A. 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 XI of Appendix A. In the following, we discuss our findings from this study.

From the results of our study, the overall χ^2 [62] 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 14 in Appendix A). 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 VI-B 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% par-

ticipants 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.

G. 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 III-C1, 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 [86]. The authors of [86] 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.

H. 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 X 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.

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.

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

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

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.

VII. DISCUSSION

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 [73], [43] 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 [29] 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 IV) 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 [27], [26], [38], [40], pornography [84], grooming [74] and trolling [44] 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 [28]. 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 misclassify 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 [20], [66]. 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 paper, 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.

VIII. RELATED WORK

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 [68] discusses early work in cyberbullying, including the nature of cyberbullying in online and social media environments. The study in [77] reveals that cyberbullying in images is especially harmful among the other types of cyberbullying discussed in this work. Methods introduced in [35] 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 [56], [63]. For example, a study [78] 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 of herself". However, the concept of repetition is questioned by many studies [56], [63], [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 [83], [36], [72]. The work in [83] 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 [37], 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 [55]. A recent study [81] elaborates on an approach that incorporates the use of hashtags, emotions and spatio-temporal features to detect textual cyberbullying. Another recent study [85] 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 [32], [53], [76]. However, these studies only partially addresses the problem of cyberbullying, as cyberbullying involves several different forms of media, such as images, in addition to text.

Factors analyses and correlations in societal problems (such as privacy and security) has gained momentum very recently. For example, a work [49] that studies the identification of bystanders in an image identifies several factors, such as pose, comfort, and replaceability, so that machine learning models can be built based on them to protect users' privacy.

IX. CONCLUSION

In this paper, we study the phenomenon of cyberbullying in images, specifically its factors and classification based on those factors. We have discussed how images can have cyberbullying content due to the highly contextual visual factors. We have introduced our approach for the identification of visual cyberbullying factors in images. We have found that visual cyberbullying involves five factors, body-pose, facial emotion, gesture, object and social factors. We have examined four classifier models that can detect visual cyberbullying based on the identified factors at different levels of abstractions. Among these four classifier models, the multimodal classifier performed the best, since it is based on both images features and visual factors based features. We have evaluated the effectiveness of the identified visual factors and conducted studies to examine the performance of the four classifier models. Our analysis, which demonstrates that multimodal classification approach is best suited for detecting cyberbullying in images, is an important finding to achieve detection capability for cyberbullying in images.

ACKNOWLEDGMENT

This work is supported in part by the National Science Foundation (NSF) under the Grant No. 2031002, 1846291, 1642143, and 1700499.

REFERENCES

[1] Flickr. https://www.flickr.com.

[2] Pinterest. https://www.pinterest.com/.

[3] Amazon Comprehend, 2020. https://aws.amazon.com/comprehend/.

[4] Amazon Rekognition, 2020. https://aws.amazon.com/rekognition/.

[5] Clarifai, 2020. https://www.clarifai.com/.

[6] Cyberbullying: one in two victims suffer from the distribution of embarrassing photos and videos, 2020. www.sciencedaily.com/releases/2012/07/120725090048.htm.

[7] Cyberbullying Stories, 2020. https://cyberbullying.org/stories.

[8] DeepAI, 2020. https://deepai.org/.

[9] Facebook, 2020. https://www.facebook.com.

[10] Google Cloud Vision API, 2020. https://cloud.google.com/vision/.

[11] Hate on Display Hate Symbols Database, 2020. https://www.adl.org/hate-symbols?cat_id%5B146%5D=146.

[12] IBM Toxic Comment Classifier, 2020. https://developer.ibm.com/technologies/artificial-intelligence/models/max-toxic-comment-classifier/.

[13] Instagram, 2020. https://www.instagram.com/.

[14] LGBTQ Pride Symbols and Icons, 2020. https://algbtical.org/2%20SYMBOLS.htm.

[15] Perspective API, 2020. https://www.perspectiveapi.com.

[16] Pew Research Center, 2020. http://www.pewresearch.org/.

[17] Pytorch mobile. https://pytorch.org/mobile/home, 2020.

[18] Twitter, 2020. https://twitter.com.

[19] Yahoo NSFW, 2020. https://github.com/yahoo/open_nsfw.

[20] Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6:14410–14430, 2018.

[21] 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.

[22] 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.

[23] 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.

[24] Marilyn A Campbell. Cyber bullying: An old problem in a new guise?. *Australian journal of Guidance and Counselling*, 15(01):68–76, 2005.

[25] 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.

[26] 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.

[27] 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 Confernece on Social Computing (SocialCom)*, pages 71–80. IEEE, 2012.

[28] Danielle Keats Citron and Mary Anne Franks. Criminalizing revenge porn. *Wake Forest L. Rev.*, 49:345, 2014.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

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

[34] 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.

[35] Dominic DiFranzo, Samuel Hardman Taylor, Franccesca 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.

[36] 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.

[37] Karthik Dinakar, Roi Reichart, and Henry Lieberman. Modeling the detection of textual cyberbullying. *The Social Mobile Web*, 11(02):11–17, 2011.

[38] 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. ACM, 2015.

[39] 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.

[40] Mai ElSherief, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna, and Elizabeth Belding. Peer to peer hate: Hate speech instigators and their targets. *arXiv preprint arXiv:1804.04649*, 2018.

[41] 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.

[42] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[43] 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.

[44] 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. *arXiv preprint arXiv:1802.00393*, 2018.

[45] Christopher Fox. A stop list for general text. In *Acm sigir forum*, volume 24, pages 19–21. ACM, 1989.

[46] 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.

[47] Kilem L Gwet. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC, 2014.

[48] 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.

[49] Rakibul Hasan, David Crandall, Mario Fritz, and Apu Kapadia. Automatically detecting bystanders in photos to reduce privacy risks. In *2020 IEEE Symposium on Security and Privacy (SP)*, 2020.

[50] Sameer Hinduja and Justin W Patchin. Bullying, cyberbullying, and suicide. *Archives of Suicide Research*, 14(3):206–221, 2010.

[51] 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.

[52] Sameer Hinduja and Justin W Patchin. State cyberbullying laws. *Cyberbullying Research Center*, 2012.

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

[54] Adam Kendon. Do gestures communicate? a review. *Research on language and social interaction*, 27(3):175–200, 1994.

[55] 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.

[56] 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.

[57] 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.

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

[60] 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.

[61] 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.

[62] Mary L McHugh. The chi-square test of independence. *Biochemia medica: Biochemia medica*, 23(2):143–149, 2013.

[63] Ersilia Menesini and Annalaura Nocentini. Cyberbullying definition and measurement: Some critical considerations. *Zeitschrift für Psychologie/Journal of Psychology*, 217(4):230–232, 2009.

[64] 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.

[65] 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.

[66] 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.

[67] 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.

[68] 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.

[69] Justin W Patchin and Sameer Hinduja. *Cyberbullying prevention and response: Expert perspectives*. Routledge, 2012.

[70] Justus J Randolph. Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa. *Online submission*, 2005.

[71] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[72] 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.

[73] Oliver C Robinson. Sampling in interview-based qualitative research: A theoretical and practical guide. *Qualitative research in psychology*, 11(1):25–41, 2014.

[74] 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.

[75] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[76] 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.

[77] Robert Slonje and Peter K Smith. Cyberbullying: Another main type of bullying? *Scandinavian journal of psychology*, 49(2):147–154, 2008.

[78] 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.

[79] 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.

[80] Jürgen Streeck. Gesture as communication i: Its coordination with gaze and speech. *Communications Monographs*, 60(4):275–299, 1993.

[81] Junming Sui. *Understanding and fighting bullying with machine learning*. PhD thesis, Ph. D. dissertation, The Univ. of Wisconsin-Madison, WI, USA, 2015.

[82] Nancy E Willard. *Cyberbullying and cyberthreats: Responding to the challenge of online social aggression, threats, and distress*. Research press, 2007.

[83] 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.

[84] 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.

[85] 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.

[86] 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.

[87] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. *arXiv preprint arXiv:1412.6856*, 2014.

APPENDIX A
USER STUDY INTERFACE AND KEYWORDS

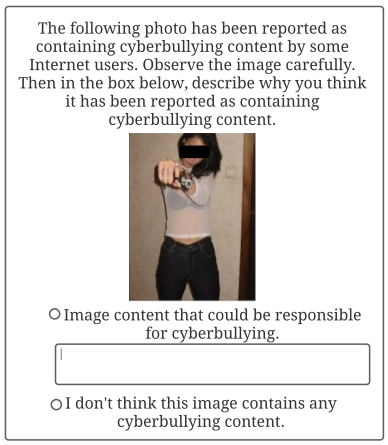


Fig. 14: User study interface: participants are provided with a free text box to enter factors on their own.

#	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 XI: Keywords used to identify factors.

APPENDIX B
IMAGE ANNOTATION TASK INTERFACE

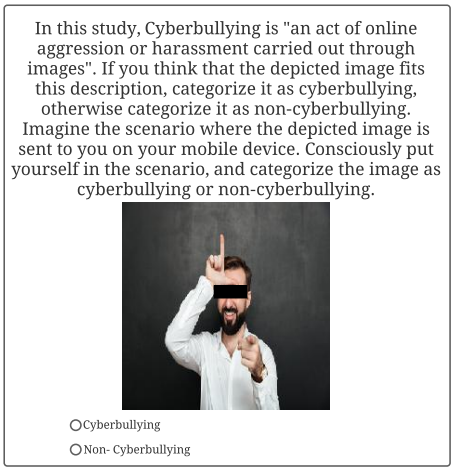


Fig. 15: Interface of image annotation task.