See No Evil, Hear No Evil: Audio-Visual-Textual Cyberbullying Detection

DEVIN SONI, Rutgers University, USA, USA
VIVEK SINGH, Rutgers University, USA, USA

Emerging multimedia communication apps are allowing for more natural communication and richer user engagement. At the same time, they can be abused to engage in cyberbullying, which can cause significant psychological harm to those affected. Thus, with the growth in multimodal communication platforms, there is an urgent need to devise multimodal methods for cyberbullying detection and prevention. However, there are no existing approaches that use automated audio and video analysis to complement textual analysis. Based on the analysis of a human-labeled cyberbullying data-set of Vine "media sessions" (six-second videos, with audio, and corresponding text comments), we report that: 1) multiple audio and visual features are significantly associated with the occurrence of cyberbullying, and 2) audio and video features complement textual features for more accurate and earlier cyberbullying detection. These results pave the way for more effective cyberbullying detection in emerging multimodal (audio, visual, virtual reality) social interaction spaces.

CCS Concepts: • **Human-centered computing** → **Social media**; *Empirical studies in collaborative and social computing*;

Additional Key Words and Phrases: Cyberbullying; Detection; Machine Learning; Audio-visual; Multimodal

ACM Reference Format:
Devin Soni and Vivek Singh. 2018. See No Evil, Hear No Evil: Audio-Visual-Textual Cyberbullying Detection. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW, Article 164 (November 2018), 25 pages. https://doi.org/10.1145/3274433

1 INTRODUCTION

Cyberbullying is a critical socio-technical problem that seriously limits the use of online interaction spaces by multiple individuals. Dinakar et al., define cyberbullying as "*When the Internet, cellphones or other devices are used to send or post text or images intended to hurt or embarrass another person*" [20]. According to a National Crime Prevention Council report, more than 40% of teenagers in the US have reported being cyberbullied [58]. Multiple studies have highlighted the negative effects of cyberbullying [8, 82], which include deep emotional trauma, psychological and psychosomatic disorders.

1.1 Modern cyberbullying

While many researchers have worked with the effects of cyberbullying on teenagers [32, 85] and also tried to identify automated methods for cyberbullying detection [18, 20, 72], such approaches are yet to consider the dramatically changed social media landscape that the teenagers are dealing

Authors' addresses: Devin Soni, Rutgers University, USA, NJ, USA, dvs39@scarletmail.rutgers.edu; Vivek Singh, Rutgers University, USA, NJ, USA, vivek.singh@rutgers.edu.

2573-0142/2018/11-ART164 $15.00
https://doi.org/10.1145/3274433

with now compared to that of even five or ten years ago. For example, recent studies have reported that teenagers now make extensive use of image and video sharing apps (e.g., Instagram, Vine, Snapchat) for their interactions [64, 76].

Consequently, there has been significant growth in using image and video content for cyberbullying [32, 74]. In theoretical terms the Medium Theory ("Medium is the message") suggests that cyberbullying will manifest itself in very significant ways across modalities [50, 54] and in practical terms it has been argued that "cyberbullying grows bigger and meaner with photos, video" [40]. Popular social media sites (e.g. Instagram, Twitter, Facebook, Snapchat) are becoming increasingly visual, and the use of audio-based interfaces for interacting with both devices (e.g. Siri, Alexa) and other human beings (e.g. Hello, Pundit, Skype) is constantly increasing. Among the various types of cyberbullying, bullying based on video or audio clips ranks as one of the most common types, and its prevalence will only grow as more social networks place an emphasis on audiovisual content [79]. Furthermore, bullying victims rate bullying based on audiovisual content as being more severe than purely text-based bullying such as text messaging or instant messaging [53].

1.2 Cyberbullying detection

When faced with the immense volume of modern social networks, it is impossible to use an entirely manual approach to cyberbullying detection. Instead, machine learning models may be used as an initial flagging mechanism in order to significantly reduce the amount of manual inspection that must be done by content reviewers. When designing models to detect cyberbullying, it is important to connect modeling choices with the ways that people process, and are affected by, various forms of information.

The Limited Capacity Model of Mediated Motivated Message Processing (LC4MP) states that humans are inherently limited in their ability to process the various modalities of information that they encounter. They therefore respond selectively to certain channels of information in each modality, often in proportion to the intensity of the stimulus. For example, intensely negative stimuli like offensive language or violence will trigger a greater response than normal conversation or a selfie [44]. Thus, it is important that a machine learning model is able to capture a wide range of information channels, so that it is able to comprehensively process the most salient features within each modality, in each particular case of cyberbullying.

We may combine general theories of information processing with theories such as the General Aggression Model, which is a behavioral framework that can be applied to cyberbullying. It posits that the internal state of a bullying victim is a function of their cognition, affect, and arousal [42]. Content which displays provocative items such as violence or nudity is likely to evoke strong emotional response, and draw personal experiences with similar content to the viewer's mind [3]. Therefore, it is important that we design cyberbullying models that are able to process dimensions of content on the social network related to emotion and arousal. Content that is emotionally-intensive or emotion-arousing may present itself in different modalities, such as through the visuals or audio of a video clip, and a model that cannot take these into account is not capturing the full situation and is likely to suffer in performance.

1.3 Multimodal detection

While the importance of understanding multimodal content for cyberbullying detection has been widely acknowledged [40], cyberbullying detection literature is still primarily focused on (sophisticated) text processing, and its accuracy remains limited. There are as yet, few efforts that leverage the visual features and none that use automated audio and video analysis for cyberbullying detection.

Hence, with a focus on better cyberbullying detection using multiple types of signals, this work aims to systematically study the following research questions:

RQ1: Which audio and video features are associated with increased likelihood of cyberbullying occurring in a media session?
RQ2: Can audio and video analysis improve cyberbullying detection beyond that obtained by solely textual analysis, and if so, does this allow for early detection of cyberbullying?

Specifically, this work utilizes a human-labeled data-set of Vine "media sessions" (six-second videos, with audio, and corresponding text comments) and employs text, audio, and video processing techniques to automatically compute features [68]. Each media session has originally been labeled for cyberbullying by 5 crowd sourced workers. The calculated features are then analyzed using a machine learning approach to build automated detectors using the provided labels. These automated detectors could provide an initial feedback to the relevant stakeholders (e.g., the users themselves, the social network administrators, law enforcement authorities, parents, school authorities, peers) on possible cases of cyberbullying, thus allowing them to further validate the messages.

Although Vine is no longer available, this approach is applicable to multiple social media applications which support audio, video and textual content (e.g. Twitter, Instagram, Skype, Keek, Eva, Hello, and Snapchat). Additionally, we recognize that cyberbullying may occur in many forms (e.g. single bully vs. a group of bullies), and through many different mediums (e.g. calls, texts, online media) [79, 80]. While Vine may not capture all possible forms of cyberbullying, in future, we can imagine similar approaches to be used to prevent cyberbullying incidents in different online networks as well as other audio, and virtual reality-based interaction spaces.

Note that we do not expect audio and video features to replace textual features anytime soon; however, we expect them to be relevant in multiple scenarios. They could be used to complement textual features and improve overall detection performance. They could also be relevant in scenarios where audio/video posts are the first/primary posts (e.g. Vine, Viddy, SocialCam, Klip, Eva) and analyzing them could result in early detection of the posts that are likely to attract, or rather become vulnerable to, bullying posts in the future [92]. Such an early detection mechanism might be useful in prevention of cyberbullying before it occurs or at least ameliorating it to some extent.

2 RELATED WORK

Cyberbullying is an important socio-technical problem and is actively researched in multiple disciplines (e.g. education, psychology, data mining, HCI). It falls under the broader umbrella of research on negative online behavior, and there has been significant recent research on understanding, detecting, and preventing cyberbullying. Specifically, this work focuses on audio-visual-textual cyberbullying detection.

2.1 Negative online behavior

Cyberbullying falls under the gamut of negative online behavior, multiple variants of which have been studied in recent literature, such as trolling, self-harm, hate speech, rumors, and misinformation, which each have nuances that make them unique [16, 46, 60, 62, 88]. For example, a recent effort by Cheng et al. studies trolling on a popular online forum and identifies the characteristics (or rather the lack thereof) of individuals who engage in trolling [14]. Similarly, a recent paper by Chandrasekharan et al. identifies abusive content in online communities by comparing message similarity across different websites (e.g. 4chan, Reddit) [12]. Other variants of such research include those on detecting self-harm, hate speech, rumors, and misinformation (e.g. [11, 59, 75]). While each of these studies negative online behavior, each of these also has a specific focus and nuance, which make them unique. For instance, Cheng et al. note that while cyberbullying behavior is

repeated, intended to harm, and targeted at specific individuals, trolling encompasses a broader set of behaviors that may be one-off, unintentional, or untargeted [14]. In this work, we focus our attention specifically on cyberbullying detection.

2.2 Understanding cyberbullying

Cyberbullying refers to the notion of causing harm to others using technological means and comes in multiple variants, including *gossip, exclusion, impersonation, harassment, cyberstalking, flaming, outing*, and *trickery* [1, 21, 32, 43, 79, 81]. However, there is no universally accepted definition of the term [41, 52]. For instance, Smith et al., define cyberbullying as "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" [81] and Dooley et al., define cyberbullying as "Bullying via the Internet or mobile phone" [23]. Typically, the aspects of *repetition*, *intent to harm*, and *power imbalance* are frequently included in definitions of bullying; however, multiple scholars have questioned each of these aspects [41, 43, 52, 85]. This is because the notions like repetition take different meaning in cyber spaces. Same email can be forwarded to multiple recipients or the same video can be viewed or commented on repeatedly [76].

Numerous studies have focused on cyberbullying and also how cyberbullying differs from traditional bullying [32, 79, 87]. Cyberbullying, as opposed to in-person bullying, opens up several channels for attack by the bully. They can bully over a combination of calls, instant messages, comments, images, and multimedia content such as videos [79, 80]. With the growth of multimedia-based social networks such as Instagram, bullying based on audiovisual content is one of the most common types, and it is growing increasingly popular as more websites add multimedia content to their platforms [79]. According to danah boyd and colleagues, with the advent of online social networks such as Twitter and Facebook, cyberbullying has become more prevalent due to the inherent *persistence, searchability, and replicability*, along with the *invisibility* of the audience in such networks [10]. When compared to text-based bullying, bullying related to audiovisual content has also been shown to be more severe and harmful to victims [53]. With both the increasing prevalence and intensity of modern cyberbullying, it is therefore important that mechanisms are designed to detect and mitigate these issues.

2.3 Automated cyberbullying detection

Previous research efforts on automatic cyberbullying detection have largely focused on using (sophisticated) text-based methods for cyberbullying detection [20, 56, 72]. For instance, Reynolds et al., [72] used the number, density and value of foul words as features to determine the cyber bullying messages. Similarly, Dinakar et al. found that building individual topic-sensitive classifiers and common sense reasoning help to improve the detection of cyberbullying messages [21], [20]. Recently, Sui [83] expanded the text-based detection approach to model the use of hashtags, emotions as well as spatio-temporal spread to understand and detect cyberbullying. Zhao et al. [92] have reported the use of an embeddings-enhanced bag-of-words approach for improving textual cyberbullying detection, and Raisi and Huang [70] have suggested the use of participant-vocabulary consistency for detecting cyberbullying.

Other efforts have focused on the use of complementary information to enhance text-based cyberbullying detection. Dadvar et al. [17] presented an improved model using the user-based features, such as the history of the user's activities and demographic features. On the other hand, Nahar et al. built a cyberbullying network graph with the users who had been previously labeled as cyberbullies and victims, and then used a ranking method to identify the most active cyberbullies and victims [55], [57]. Huang et al. [35, 78] have suggested using social relations between the participants as a complementary layer of information to the text message for detecting cyberbullying.

2.4 Preventing cyberbullying

Multiple online communities have adopted the ideas of content moderation (based on flagging, up-voting, down-voting etc.) to prevent the harmful effects of cyberbullying and other anti-social behavior on their sites. For instance, websites such as Usenet suggest that the authors take their disputes outside of the forum, or designate special threads to engage in "fiery discussions" [12]. Some other websites (e.g. Facebook) have tools to report bullying and in extreme cases some websites (e.g. Reuters) have completely disabled comment sections [19, 24]. Lastly, many popular sites (e.g. YouTube, Facebook) have teams of human moderators, who manually monitor the site for offensive or malicious content. Such a human labor intensive approach is neither scalable nor feasible for a majority of social media platforms. Furthermore, while blocking comments may work for countering trolling on certain sites (e.g. a newspaper site) where social interaction is not the primary objective, it is an unfeasible solution for *social* media apps, and hence the problem of cyberbullying.

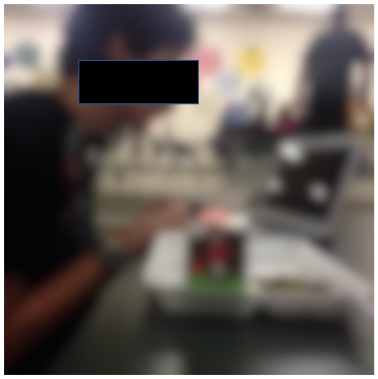
There have been a number of recent attempts at designing interfaces that may help reduce cyberbullying. For instance, Ashktorab and Vitak [6] have adopted participatory design approaches to identify app interfaces that may reduce the incidence of cyberbullying. Similarly, "reflective" interfaces as proposed by Jones encourage bullies to rethink their actions before reconfirming their decision to send out those messages [20, 37]. The strategies to encourage rethinking one's decision as suggested in literature include delayed actions, informing users of hidden consequences, links to educational material, use of normative agents, and flagging of messages [9, 47, 48, 63, 90]. While each of these aspects – understanding, detecting, and preventing cyberbullying, is important, we focus this paper on the problem of better cyberbullying *detection*, specifically audio visual textual cyberbullying.

2.5 Audiovisual cyberbullying detection

Compared to text-heavy approaches for detection, the literature on visual cyberbullying detection is relatively sparse. Even works pertaining to audio-visual platforms, such as YouTube, have thus far mainly considered textual content and meta-data, though some have broadly considered the potential architecture of an audio-visual system but have not implemented or validated their proposed systems [22, 67]. Some recent efforts, however, have started analyzing static image content for cyberbullying detection. Hosseinmardi et al. used human crowd-sourced labeling (rather than automated computational algorithms) for image analysis [33] to aid cyberbullying detection. Another effort by Zhong et al. uses custom-created deep learning modules [93] for image analysis and cyberbullying detection, and the third effort, by Singh et al., uses computer vision APIs [77]. There is no existing literature, to our knowledge, that utilizes audio content for cyberbullying detection.

However, all of these efforts focus on images rather than videos. The only existing line of work at detecting cyberbullying in video posts is by Rafiq et al. [68, 69] which employs *manual* video labeling to obtain visual features pertaining to content and emotion. This is not a scalable method, however, as it is clearly too costly to have someone manually describe the content of each and every video posted to a social network. In order for a feature to be useful in a large-scale detection platform, the features must be computed automatically, rather than manually. Additionally, Rafiq et al. do not use any audio features in their work.

Our work is therefore, to the best of our knowledge, the first attempt at automatic video analysis for cyberbullying detection, and the first attempt at using audio content for cyberbullying detection. We note that the literature on textual cyberbullying detection is far more advanced than that of audio



Speech content: this nigga like never talks ... [shrieking noises]

Sample comments:

1. He probably doesn't like talking to lame ass people
2. Lmao he go to [school name] XD
3. lol ur hilarious flipping out I love people like you gives me great laughs
4. Lol
5. Y'all are funny people are waaaayyyyyyy to sensitive now a days

Fig. 1. (warning: explicit content) Sample cyberbullying media session that was not detected using textual modeling but detected using audio, video, textual modeling.

or video based analysis. Hence, in this early effort we do not suggest *replacing* textual features with audio or video features, but rather, *combining* them for earlier and more accurate detection.

3 PROPOSED APPROACH

In this work we consider a data-set of Vine posts that has originally been labeled for cyberbullying by 5 crowd sourced workers. This data-set has been shared by the authors of [68] and each labeled "media session" includes the posted 6 second video, its associated audio, and the posted text-based comments. We identify a number of text, audio, and video based features and compute them using available APIs and libraries, such as Clarifai and Microsoft Cognitive Services. These features are included in a machine learning algorithm to develop automated classification algorithms and also identify the most important features than differentiate between bullying and non bullying classes. We choose to use APIs rather than handmade deep learning models because APIs are more accessible & standardized, and do not require extensive background knowledge to create and/or use.

We aim to identify cyberbullying cases that are not easily detected using just the textual content. This includes instances in which the bullying occurs in the video, and cases where the comments are only weakly suggestive of bullying and the video content reaffirms the presence of bullying. In Figure 1 (warning: explicit content) we provide an example of the former, in which a student bullies one of his classmates . In this media session, the bullying is clearly observable by inspecting the audio and visual content (e.g., the presence of shrieking noises), but the textual content includes roughly equal amount of both positive and negative comments thus making cyberbullying detection using just textual content more difficult. We will provide an in-depth analysis of how our method is able to detect this instance of cyberbullying in a later section.

4 FEATURES

We first identify textual, audio, and visual features relevant for cyberbullying detection. To do so, we survey the existing literature on cyberbullying detection, as well as text, audio, and video processing (e.g. [20, 68]). In order to provide a standardized basis for the features in each modality, we categorize each feature as being broadly related to one of: *channel capacity, arousal, affect*, or *cognition*, which have been identified based on an array of existing literature [3, 20, 34, 35, 44]. We summarize these features in Table 1. We acknowledge that some features may be interpreted to

fall under more than one category, but in this work we choose to limit each feature to what we believed to be the most relevant category.

These features are based on two theoretical considerations and each of the feature selected has empirical support based on past literature on cyberbullying detection. The General Aggression Model has been posited as an important comprehensive approach to understand cyberbullying [42]. Specifically, besides identifying the inputs and outcomes, it also identifies the *routes* through which cyberbullying is perpetrated. Those three routes correspond to *cognition, affect,* and *arousal.* Hence, through multiple features we try to capture clues to the cognition, affect, and arousal of the individuals engaging with the media session. Affect refers to experience of feeling or emotion and some of the features considered include the sentiment of the textual comments and the valence of the facial expressions. Arousal refers to the state of being physiologically alert, awake, and attentive. Both, low level and relatively high level features are used to capture this including the loudness of the audio and the compound arousal score obtained from facial expressions captured. Cognition refers to the mental action or process of acquiring knowledge and understanding through thought, experience, and the senses. Since much of this takes place within the minds of the users, this is the hardest category to find clues for. In this work, we ameliorate this problem in multiple ways. First, we use APIs like Microsoft Cognitive Services and Clarifai to obtain richer, deeper understanding of the text, audio, and video content to obtain at least some clues to the objects and events captured and the associated cognitive labels. Next, following Anderson and Bushman, we posit that temporary increase in the hostile scripts in one's mind may be primed by factors such as media violence [5]. Hence, we consider the combination of different modalities i.e. audio, video, text, and allow for patterns to emerge over time get some clues to this process.

The second theoretical model considered in this work - LC4MP - states that human beings have limited capacity for cognitive processing of information, including when it arrives via different modalities (e.g., audio, video, text) [44]. Specifically, human beings employ shortcuts to information processing tasks that minimize the use of cognitive resources and emotional or emotion-arousing content often triggers greater cognitive, affective, and behavioral responses, potentially including those related to cyberbullying [3]. Hence, this theory suggests two kinds of features. First, it would be useful to capture the amount of signal contained in each channel e.g. the number of words in text or the number of faces in video. Second, it again suggests identifying features that capture emotion or emotion-arousing content across modalities. Since people are limited in their ability to process all aspects of media content, it is plausible that only a subset of features will be relevant in each case of cyberbullying [44]. It is therefore important to have each feature set computed for each modality, since each case of cyberbullying may manifest itself using a different subset of these features. For example, the video may display something harmless on its own, but the combined audio and comment responses may target a victim. Conversely, the text content may not contain bullying or mention bullying, even if the video's visuals and audio clearly portray it. In these situations, being able to obtain a full view of the independent modalities allows us to detect cyberbullying cases that would have otherwise gone undetected, since the different modalities do not necessarily provide signals in concordance with each other.

In order to compute some of the textual features, we pre-process the media sessions' comments such that they only contain alphanumeric characters. In order to process videos, we use OpenCV's video processing capabilities to extract visual frames, and we use FFmpeg to extract audio files [27, 61]. For features using Principal Component Analysis (PCA), we use cross-validation to determine the number of components; three happened to work best in all cases, likely due in part to the modest sample size.

Table 1. Summary of features used for detection.

Number	Modality	Feature	Channel Capacity	Affect	Arousal	Cognition
1	**Textual**	Number of words	x			
2		Sentiment		x		
3		Valence		x		
4		Density of punctuation			x	
5		Density uppercase characters			x	
6		Density of explicit content			x	
7		Arousal			x	
8		PCA of GloVe embeddings (3)				x
9	**Visual**	Number of faces	x			
10		Length of visual text	x			
11		Sentiment of visual text		x		
12		Valence of face		x		
13		Arousal of face			x	
14		Presence of gore			x	
15		Presence of explicit nudity			x	
16		Presence of drugs			x	
17		Presence of suggestive nudity			x	
18		PCA of scene labels (3)				x
19	**Audio**	Number of spoken words	x			
20		Percentage speech content	x			
21		Percentage music content	x			
22		Percentage silence content	x			
23		Sentiment of spoken content		x		
24		Valence of voice		x		
25		Loudness			x	
26		Density of explicit spoken content			x	
27		Arousal of voice			x	
28		PCA of GloVe embeddings (3)				x

4.1 Textual Features

Textual features are widely used in the detection of cyberbullying. Following [68], we treat each media session as one document by concatenating all of the user comments.

4.1.1 Channel Capacity.

- **Number of words** Prior literature suggests that cyberbullying sessions tend to contain more words than non-cyberbullying sessions [33].

4.1.2 Affect.

- **Sentiment** Cyberbullying sessions to tend use more negative language due to increased use of insults and swear words [33, 57]. We use a sentiment analyzer created by Hutto and Gilbert called VADER, which is specifically suited to handle sentiment analysis of social media text [36]. It takes into account not only the textual content of the text, but also the punctuation, capitalization, and emoticons. VADER provides a single compound sentiment score between -1.0 (negative sentiment) and +1.0 (positive sentiment).
- **Valence** Another way to measure affect is through valence, as suggested (but not yet pursued) in previous work [35]. Valence measures the positivity of a stimulus. We use a list of valence scores to obtain a score for each word in the comments section, and then average these scores to obtain the value for each media session [89].

4.1.3 *Arousal.*

- **Density of punctuation** Punctuation marks (e.g.'!', '?') are used as a way to communicate excitement on social media; and excessive or repeated use of punctuation may be considered analogous to shouting [7].
- **Density of uppercase characters** Similar to punctuation use, excessive use of uppercase characters may be considered analogous to shouting [7].
- **Density of explicit content** Cyberbullying sessions tend to contain more explicit content and swear words as cyberbullies may directly use them to assault their victims [35]. Following [49] a list of 723 English terms, including common expletives and insults, was used to identify these instances [51]. We record the percent of words in the comments that appear in this list.
- **Composite arousal score** The use of arousal as a feature has been suggested in previous work on cyberbullying detection [35]. We use a list of arousal scores to obtain a score for each word in the comments section, and then average these scores to obtain the value for each media session [89].

4.1.4 *Cognition.*

- **Word embeddings** Following [2] we explore the use of GloVe word embeddings for cyberbullying detection. These embeddings represent each word as an n-dimensional vector, and place words in a vector space in such a way that words with similar meanings are placed near each other. More formally, the embeddings are a deep-learning based latent representation of relationships among words, that often exhibit a rich structure that supports inference and visualization [66]. Specifically, we use 50-dimensional GloVe word embedding vectors trained on a Twitter corpus to interpret the semantics of the comments [66]. We first average the vectors over all of the words in the comments. Then, in order to reduce the number of dimensions, we apply Principal Component Analysis (PCA) and keep the first 3 principal components.

4.2 Visual Features

While the literature on the use of visual features is relatively sparse, there has been some recent work suggesting the value of human-labeled or automated visual analysis for cyberbullying [68, 77]. We use Microsoft Cognitive Services [45] to extract the number of faces and emotions, and Clarifai [15] to extract scene labels. We avoid the use of immutable personal characteristics such as gender or race due to ethical considerations.

4.2.1 *Channel Capacity.*

- **Number of faces** Cyberbullying often targets a specific person. Videos without people are unlikely to have visual displays of bullying and are less likely to attract bullies in the textual comments. Specifically, videos with multiple people are more likely to contain instances of cyberbullying within them as this could mean the victim and bully are in the scene together. Thus, as prior work has shown, the number of faces present in the video is likely to be a useful signal [77].
- **Length of visual text** Many videos within our data-set display text on the screen for various reasons (e.g., to display links, or a past of a slide-show). Prior research has shown that text displayed in the video correlates with cyberbullying likelihood [33].

4.2.2 *Affect.*

- **Sentiment of visual text** Prior research has found the sentiment portrayed by the textual content to be more negative in cyberbullying sessions [33, 57], hence we suspect that this

trend will also show in the visual text present in the video. We again use the VADER model's compound sentiment score for this.

- **Valence of facial expression** We use Microsoft's API to obtain scores for each second of video for 8 emotions: sadness, neutrality, contempt, disgust, anger, surprise, fear, and happiness. These are similar to those manually obtained in prior visual cyberbullying work [33]. We first average these scores across the six seconds of video. We then use the method in [29] to obtain valence scores for each of these 8 emotions. Finally, we convert these into valence scores for the video using a weighted average, where the weight is the average score given to that emotion from the API.

4.2.3 *Arousal.*

- **Explicit content** For each second of the video, we also obtain scores for each of the following categories pertaining to controversial content: gore, explicit nudity, drug, and suggestive nudity. We then average each score over each second of video. The presence of inappropriate content has been shown to correlate with the occurrence of cyberbullying [76].
- **Arousal of facial expression** We use Microsoft's API to obtain scores for each second of video for 8 emotions: sadness, neutrality, contempt, disgust, anger, surprise, fear, and happiness. These are similar to those manually obtained in prior visual cyberbullying work [33]. We first average these scores across the six seconds of video. We then use the method in [29] to obtain arousal scores for each of these 8 emotions. Finally, we convert these into arousal scores for the video using a weighted average, where the weight is the average score given to that emotion from the API.

4.2.4 *Cognition.*

- **Scene labels** We are able to obtain a set of labels for each second of the video that describe the scene of the video. These labels range from describing the overall scene content (e.g. outside vs. inside), to specific objects in the scene (e.g. computers, cars), to qualitative descriptors (e.g. dark, light). We represent the labels for each video as a bag-of-words, where we concatenate the labels for each second of video to form the list of words for each media session. We first apply the tf-idf transformation [73] to the raw document-count matrix, as some labels were considerably more telling than others. We then apply PCA to reduce the dimensionality of the feature, and keep the first 3 principal components.

4.3 Audio Features

We note that there is practically no work on audio-based cyberbullying detection. We hypothesize that audio features, like visual features, will convey unique information not present in the other modes of communication. The words spoken in the video and the emotion displayed provide us with information about the original content of the video, and frame the subsequent textual content that forms as a response. Several of these features are analogous to their counterparts within the set of textual and visual features. We use CMUSphinx [86] to extract the speech in the audio, and pyAudioAnalysis [31] to segment the audio and measure valence & arousal.

4.3.1 *Channel Capacity.*

- **Number of words** Cyberbullying media sessions tend to have more words in the comment sections, so it is possible that the speech content will also be longer in cyberbullying sessions [33].
- **Content segments** We can break down the auditory content by segmenting it into portions containing speech, music, and silence. We first classify each small (50 ms.) segment of video as being one of those three categories, labeling the segment with the most likely label in the

case where multiple may be true to varying degrees. Then, once each small segment has been labeled, we obtain the percent that each category makes up of the total length of the video.

4.3.2 *Affect.*

- **Sentiment of spoken content** Much like the textual content of cyberbullying sessions, which are typically more negative, we also suspect that the spoken content will too be more negative [33, 57]. We again use the VADER sentiment analyzer, as it is capable of understanding modern slang [36].
- **Valence of voice** We also obtain the emotional content of the spoken audio in the form of valence scores. This provides us with the speaker's affect as evident in their tone of voice and manner of speaking, which may contrast with the sentiment of the spoken content. These are computed by pyAudioAnalysis using a variety of lower-level audio signal features such as pitch and Mel-frequency cepstral coefficients (MFCCs) [31].

4.3.3 *Arousal.*

- **Loudness** The level of loudness of audio indicates arousal of the speaker and could be predictive of negative responses to the content, such as bullying [38]. In the audio domain, loudness may be analogous to the use of uppercase characters or punctuation in textual content. The pyAudioAnalysis library provides us with the average loudness in decibels, computed as an average of the loudness of successive 50 ms. audio segments.
- **Density of explicit spoken content** Much like the textual content within cyberbullying sessions tend to contain more explicit content, we suspect that the spoken content will as well, as the subjects may either be speaking negatively about their bully or the person whom they are bullying [35]. The same list of 723 terms was used to identify these instances [51], and the percent of spoken words in this list was recorded.
- **Arousal of voice** We obtain arousal scores for the spoken audio. This provides us with the speaker's arousal as evident in their tone of voice and manner of speaking, which may contrast with the sentiment of the spoken content. These are computed by pyAudioAnalysis using a variety of lower-level audio signal features such as pitch and MFCCs. [31].

4.3.4 *Cognition.*

- **Word embeddings of spoken content** We again use 50-dimensional GloVe word embedding vectors to represent the content of the spoken content [66]. We average the vectors over all of the words in the speech content, then, in order to reduce the number of dimensions, we apply PCA and keep the first 3 principal components.

5 EXPLORATORY ANALYSIS

5.1 Corpus

This work uses the Vine data-set made available by Rafiq et al. [68] that has been used in several studies of cyberbullying [68, 69]. It was created with the snowball sampling method, and only sessions with at least 15 comments were retained. The threshold of 15 posts was selected to capture enough posts where repetition patterns in bullying can be observed. For each public Instagram user, the collected profile data included the media objects (videos/images) that the user has posted and their associated comments, user id of each user followed by this user, user id of each user who follows this user, and user id of each user who commented on or liked the media objects shared by the user. Rafiq et al. consider each media object, plus its associated comments, as a "media session," which we also follow here.

Labeling data is a costly process, and therefore, in order to make the labeling of cyberbullying more manageable, Rafiq et al., tried to reduce the data-set size. To have a higher rate of cyberbullying instances, they considered media sessions with at least one profanity word in their associated comments. Note that the presence of profanity does not guarantee the presence of bullying, but increases the odds [69]. The sessions were then binned by profanity count, and equally-sized samples were taken from each bin to construct a preliminary data-set. They were then hand-labeled by five crowd-sourced annotators via CrowdFlower (now known as Figure Eight) who were instructed to label media sessions as involving cyberbullying if there were negative words and comments with intent to harm someone, and the posts include two or more instances of negativity against a victim who could not easily defend him or herself.

The labelers were given training on identifying cyberbullying instances and multiple quality checks were imposed on labeling. One such criteria was Figure Eight's provided confidence score, which is a custom metric that is a function of user trust scores on their platform, and agreement with other labelers on the task [28]. Since, there were no standardized guidelines to choose a cut-off for this metric we follow the 0.6 cut off as adopted by previous works which used this data-set [68, 69]. This resulted in a data-set of 959 labeled media sessions [68]. Each media session contains the submitted video, video meta-data, and comments. After removing the media sessions with corrupted video files, the data-set contained a total of 833 sessions, of which 265 are reported as containing cyberbullying.

Cyberbullying in multimodal social media can occur in multiple ways: (a) the subject of the video may be bullying someone else; (b) the video subject be bullied by others in subsequent comments; or (c) commenters may bully each other while ignoring the original video post completely. The labels provided by Rafiq et al. did not differentiate between these scenarios[68]. Further, the situations involving altercations between commenters who do not interact with the subject of the video can be detected with purely text-based methods.

In this work, we specifically wanted to focus on multimodal cyberbullying detection i.e. those where the cyberbullying incidents were associated with the original audio-video post. Hence, we went through a secondary round of manual labeling (undertaken by one of the co-authors), and kept only those instances in the data-set where there was either (a) bullying being demonstrated in the original audio video post; or (b) bullying undertaken in response to the content of the original audio-video post. This filtering resulted in a total of 733 sessions, of which 165 contain bullying involving the video content in some way.

5.2 Analysis

One of the goals of this work is to explore how the various textual audio video features relate to the prevalence of cyberbullying. Hence, we compute the features defined in the previous section for each media session, and identify the significant differences (confirmed using t-tests) for various features between the two classes. We then report the percent differences, calculated as $100 \times \frac{mean(bully) - mean(non-bully)}{|mean(bully)|}$ and calculate p-values using t-tests. The results are summarized in Table 2.

5.2.1 Textual features. We observed that the bullying media sessions included more words. More text has been connected with higher cyberbullying odds in past literature too [34]. Next, we observed that lower sentiment and valence were associated with cyberbullying. Again this is as suggested by prior literature connecting negative content to bullying [57].

At the same time, the density of uppercase characters and punctuation were found to be associated with lower odds of cyberbullying. This is contrary to findings in existing literature; it is possible that on Vine, which uses very informal communication, use of punctuation and capitalization is an

Modality	Feature	Difference	P-value
Textual	**Number of words**	63.5%	< 0.001
	Sentiment	-507.3%	< 0.001
	Valence of Text	-4.6%	< 0.001
	Punctuation	-23.4%	< 0.001
	Uppercase	-33.8%	< 0.001
	Explicit content	88.4%	< 0.001
	Arousal of Text	2.8%	< 0.001
	GloVe PCA Component 1	-544.8%	< 0.001
	GloVe PCA Component 2	n.s.	> 0.05
	GloVe PCA Component 3	-437.0%	< 0.001
Visual	**Number of Faces**	51.1%	0.005
	Length of Visual Text	n.s.	> 0.05
	Sentiment of Visual Text	n.s.	> 0.05
	Valence of Face	n.s.	> 0.05
	Arousal of Face	141.9%	0.049
	Gore	n.s.	> 0.05
	Explicit Nudity	45.0%	0.001
	Drugs	n.s.	> 0.05
	Suggestive Nudity	45.6%	0.001
	Labels PCA Component 1	-936.0%	< 0.001
	Labels PCA Component 2	n.s.	> 0.05
	Labels PCA Component 3	-422.6%	< 0.001
Audio	**Number of spoken words**	64.1%	< 0.001
	Percentage speech content	40.8%	< 0.001
	Percentage music content	-40.4%	< 0.001
	Percentage silence content	n.s.	> 0.05
	Sentiment of spoken content	-365.6%	0.008
	Valence of voice	-451.1%	< 0.001
	Loudness	n.s.	> 0.05
	Density of explicit content	261.3%	0.006
	Arousal of voice	65.2%	< 0.001
	GloVe PCA Component 1	3840.4%	< 0.001
	GloVe PCA Component 2	n.s.	> 0.05
	GloVe PCA Component 3	-405.3%	0.046

Table 2. Difference between the bullying and the non-bullying classes for the features with significant differences. A positive percentage means that the feature is higher in cyberbullying sessions.

indicator of formal language rather than shouting [7]. We will investigate these aspects further in our future work.

We also note that cyberbullying text tended to have higher density of explicit text and also a higher arousal score, suggesting that it is more likely to provoke or incite a response [34, 35]. Finally, we observe that two of the principal components for GloVe embeddings are significantly different between bullying and non-bullying sessions. These features are difficult to interpret directly due

Features	Model	Accuracy	Precision	Recall	F1 Score	AUROC
Text	**K-Nearest Neighbors**	0.559	0.282	0.558	0.374	0.619
	Support Vector Machine	0.586	0.2	0.25	0.222	0.519
	Gaussian Naive Bayes	0.673	0.4	0.769	0.526	0.806
	Logistic Regression	0.75	0.482	0.788	**0.599**	**0.837**
	Random Forest	0.777	0.525	0.596	0.559	0.812
Audio and Visual	**K-Nearest Neighbors**	0.577	0.302	0.615	0.405	0.588
	Support Vector Machine	0.64	0.267	0.308	0.286	0.574
	Gaussian Naive Bayes	0.608	0.289	0.462	0.356	0.614
	Logistic Regression	0.613	0.33	0.635	**0.434**	0.706
	Random Forest	0.712	0.364	0.308	0.333	**0.707**
All Features	**K-Nearest Neighbors**	0.624	0.306	0.5	0.38	0.574
	Support Vector Machine	0.752	0.167	0.019	0.034	0.549
	Gaussian Naive Bayes	0.77	0.5	0.808	0.618	0.84
	Logistic Regression	0.814	0.56	0.904	**0.691**	**0.877**
	Random Forest	0.774	0.509	0.519	0.514	0.832

Table 3. Performance of each model for each modality combination.

to the multi-step computation, the content clearly differs between the two classes and might be relevant for the classification task (described later in Section 6).

5.2.2 Visual features. We first notice that cyberbullying sessions do tend to contain more people in them, as is consistent with our expectations. Additionally, the people in cyberbullying videos tend to have a higher level of arousal, suggesting that they are responding to, or disseminating, more controversial content. We also observe that posting explicit and suggestive content was associated with higher odds of bullying occurring in the resulting media session. Plausibly, posting videos involving inappropriate content could lead to, or frame, further negative content posted on the thread, including cyberbullying. Finally, the two of the principal components derived from scene labels were significantly different between the two categories. Again, due to the complexity of these features, it is difficult to precisely interpret them directly. However, by inspection, we do note that the type of scene (for example, indoor vs. outdoor) is mainly captured in these features.

5.2.3 Audio features. Similar to our findings with textual features, the speech content in bullying sessions tends to be longer. This finding is consistent with our expectations, and show that the text and audio content follow similar trends. We also note that speech makes up a longer portion of the audio in bullying sessions, and music makes up a smaller portion. This goes alongside the aforementioned finding within audio, as a video with more spoken content is likely to be more controversial than one that simply plays music. Next, we find that the emotion of the speaker tends to have a lower valence, use more explicit content, and be more negative in bullying sessions. This was coupled with higher arousal in voice. These findings again are similar to our findings in visual and audio features, and suggest that the speaker is more agitated or distraught in bullying sessions. Lastly, the GloVe embeddings are significantly different between bullying and non-bullying sessions, suggesting the subjects in bullying videos address different, perhaps more controversial, topics than that of non-bullying videos.

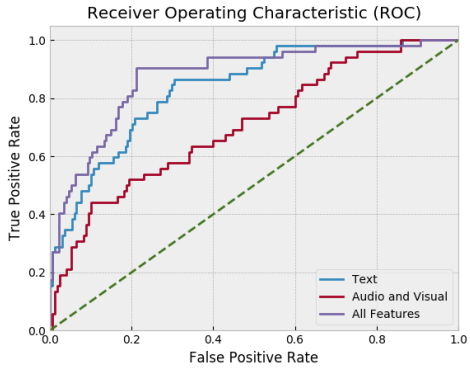


Fig. 2. ROC curves for the best classifier of each modality set.

6 CLASSIFICATION

6.1 Methodology

We now attempt to build an automatic cyberbullying classifier using machine learning with the discussed features.

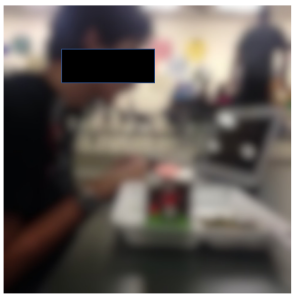
We try three modality sets: text, audio + visual, and text + audio + visual. We use a 70/30 train/test split to evaluate the classifier's performance, and repeat this 100 times in order to reduce variance in the results. We utilize SMOTE (Synthetic Minority Oversampling Technique) in order to balance the training set in each iteration [13]. SMOTE balances data-sets by oversampling the minority class and undersampling the majority class. However, rather than simply including duplicate minority examples, SMOTE creates new synthetic examples by creating convex combinations of existing data points [13]. After applying SMOTE, we obtain a training set of 400 bullying and 400 non-bullying instances.

Given our modest sample size, we choose to test relatively simple classification models, as more complex models would likely overfit. We specifically use scikit-learn's implementation of K-Nearest Neighbors, Support Vector Machine, Gaussian Naive Bayes, Logistic Regression, and Random Forest [65]. For each classifier, we select hyper-parameters (regularization strengths for Logistic Regression & Support Vector Machine chosen from $\{0.01, 0.1, 1.0, 10.0, 100.0\}$, and the number of neighbors for K-Nearest Neighbors chosen from $\{1, 2, 3, 4, 5\}$) based on 5-fold cross-validation within the training set before applying the SMOTE transformation. For all modality sets, Logistic Regression and Support Vector Machine ultimately end up performing best with regularization strengths of 0.1 and 1.0 respectively, and K-Nearest-Neighbors performed best using 3 neighbors.

In order to measure the performance of our classifiers, we consider multiple well-known metrics like accuracy, precision, recall, F1 score, and area under the ROC (Receiver Operating Characteristic) curve (AUROC). We choose to consider metrics beyond accuracy due to the class imbalance present in the data-set; these additional metrics are more informative in cases in which the class of interest (cyberbullying) constitutes a minority [35]. Specifically, when optimizing hyper-parameters and choosing the best models, we choose to optimize for F1-score, as it is the most sensitive to poor performance in the minority class, and is a function of both precision and recall.

6.2 Results

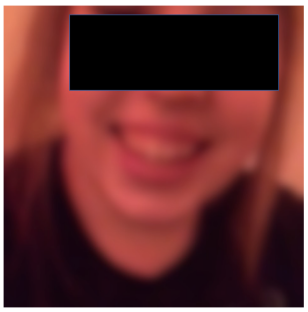
We provide the results for each tested modality combination in Table 3. Optimizing for F1 score, we find Logistic Regression to be the best performer in all of the feature combinations. In the



Speech content: this nigga like never talks ... [shrieking noises]

Sample comments:
1. He probably doesn't like talking to lame ass people
2. Lmao he go to [school name] XD
3. lol ur hilarious flipping out I love people like you gives me great laughs
4. Lol
5. Y'all are funny people are waaaayyyyyyy to sensitive now a days

(a) Bullying within the video



Speech content: I sucked [person]'s dick are you proud

Sample comments:
1. shut up fatass stop hating
2. Yeah obviously because you payed
3. I don't care what you say bc I know I'm not 12
4. Hahahah love you guys
5. Trailer trash

(b) Bullying targeting the video's subject

Fig. 3. (warning: explicit content) Examples of bullying sessions identified only by the multimodal approach.

following discussion we therefore consider only the performance of Logistic Regression for each modality combination. Figure 2 shows the ROC curves for Logistic Regression in each feature set. We specifically note that the classifier for all features is able to obtain a true positive rate of around 90% with a nearly 30% lower false positive rate than the classifier for text features.

As prior work has shown [77], the non-textual features are relatively weak features on their own. These additional modalities instead provide supporting signals that, in conjunction with textual features, can help the classifier make correct decisions in borderline cases. This is shown by the large increase going from text to text + visual + audio features, as we see a noticeable performance increase from an F1 score of 0.599 to 0.691, a percent increase of 16.92%. We confirm that this increase is statistically significant with a p-value < 0.001. The increase in AUROC is also significant with a p-value < 0.001.

In Figure 3 we show two examples of media sessions in which the multimodal approach identifies a bullying session that the purely textual approach was unable to find (warning: explicit content). Each example shows a random frame from the session's video, the speech content, and a random sample of the textual content. The two examples respectively showcase the two main types of sessions that our multimodal approach is better suited to detect: bullying directly in the video, and bullying of the video's subject by commenters. This work advocates the use of the "gestalt principle" to combine multiple weak detectors to collectively generate more confidence for classification of such bullying situations, and provide a holistic view of the media session [30, 39]. By using modalities other than text, we gain unique information that, together, better frames the context and content of the overall media session than any single modality does. Note that Vine, and other social media platforms, have different standards for language, particularly in terms of the use of

swear words. So while a media session's comments may include swear words, this is not necessarily indicative of cyberbullying or cyberaggression, as casual language use on Vine tends to involve a larger than average amount of swear words regardless of the context.

In the first example (shown earlier), the bully records one of his peers as he is eating, against his will. The bully makes fun of his reserved personality and the victim appears uncomfortable but is unable to retaliate and stop the bully from recording him. The bully then makes a series of loud shrieking noises to bother the victim on video. The textual content does not follow several of the common cyberbullying trends; it contains a short comment section, and has comments that are not very negative and that do not use many swear words (relative to other Vine comments). The comments seems more like reactions to a funny video than to an instance of cyberbullying. However, the audio-visual content offers several important clues that increase the odds of cyberbullying detection. The video contains two faces, one of which (not shown) shows a high level of arousal. The speech content is fairly long, negative, and uses a swear word. The audio is entirely speech, and displays a very high level of arousal. These clues, together, allow the model to accurately label this video as containing cyberbullying, in conjunction with the textual features.

In the second example, a girl speaks with an exaggeratedly high-pitched voice and claims to have had sexual relations with someone. From the context of the text comments, it seems like the person she is referring to is a Vine personality, not someone she personally knows. The comments repeatedly make fun of her age, appearance, and way of speaking. However, many of these comments are not explicitly negative (i.e. do not use many swear words and do not have negative sentiment), and instead are part of a lengthy argument with the subject of the video, who is unable to fend off the detractors. The proposed multimodal approach is able to catch this by combining the knowledge of the textual content – primarily the long length of the text – with the knowledge of the video's content, which displays her high level of spoken and visual arousal, the dominance of the video's audio by speech, and the use of explicit terms in the speech content. Together, these paint a picture of a comment section in which users repeatedly comment on the subject of the video based on her speech.

7 EARLY DETECTION

Given the improvements obtained in automatic detection using audio-visual features, we now see if this improvement is also present in the context of early detection. There has been little work on early detection of cyberbullying, and none that have used audio features [93]. We define the task of early detection as identifying cyberbullying using the audio content, video content, and textual features computed using *only the first 5 comments*[1]. This provides a good approximation of the textual content without having to wait for the entire comment section to unfold – potentially allowing the system to preemptively respond when the first signs of bullying appear. To test the performance we use the same features as mentioned in the previous section and the same classification methodology.

In the context of early detection, we only need to consider results for the text, and text + audio + visual feature sets, as we have already reported the performance of the audio + visual feature set. In Table 4 we show the classification performance in the early detection context.

Interestingly, we note that in the purely textual approach, there was no performance decrease by only using the first 5 comments. After statistical significance testing, we confirm that the difference in performance in early detection is statistically insignificant using t-tests when compared to that obtained by using all of the comments. One way to interpret these results is that the first few posts often *set the tone* for the conversation in a thread. A recent study by Cheng et al. [14] on

[1]We have also considered the thresholds as first 10 and first 15 comments and the results follow a similar pattern.

Features	Model	Accuracy	Precision	Recall	F1 Score	AUROC
Text	**K-Nearest Neighbors**	0.559	0.282	0.558	0.374	0.619
	Support Vector Machine	0.582	0.197	0.25	0.22	0.515
	Gaussian Naive Bayes	0.686	0.411	0.75	0.531	0.808
	Logistic Regression	0.755	0.488	0.808	**0.609**	**0.834**
	Random Forest	0.768	0.508	0.577	0.541	0.804
All Features	**K-Nearest Neighbors**	0.605	0.308	0.538	0.392	0.613
	Support Vector Machine	0.723	0.263	0.096	0.141	0.623
	Gaussian Naive Bayes	0.732	0.459	0.75	0.569	0.819
	Logistic Regression	0.8	0.549	0.865	**0.672**	**0.865**
	Random Forest	0.786	0.54	0.654	0.591	0.83

Table 4. Performance of each model for each modality combination in early detection.

trolling behavior analyzed the first five posts of a thread and found that the odds of a comment being flagged for trolling rose consistently depending on whether the previous comments were flagged for trolling. In other words, once trolling starts in a thread, it often continues and becomes worse, and this is often obvious in the first five posts themselves. In the context of cyberbullying the aspects of repetition and power imbalance appear to be relevant these cases. A victim being bullied in the comments would likely experience many harmful comments, and may have bullies who immediately respond to their content in order to exert their power over the victim and frame future comments. This aspect is clearly interesting and motivates research questions on the effects of first few posts in a thread in cyberbullying context and we plan to study this in further detail in our future work.

The text + audio + visual feature set did, however, have a statistically significant difference when compared to that obtained by using all comments ($p < 0.01$) using t-tests. Overall, we find that our multimodal approach performs nearly as well as it did using all of the comments, suffering a percent decrease in F1 score of only 2.75%. This shows that the textual signals can be computed in such a way, using only the first few comments, that is still able to properly complement signals from other modalities. Early detection is therefore a viable task to pursue under the context of multimodal detection in future work.

8 DISCUSSION

Our first research question asked which audio and video features are associated with increased prevalence of cyberbullying. Table 2 has identified multiple such features that were found to be significantly associated with cyberbullying. Many of the features found significantly associated with cyberbullying were in agreement with the existing literature.

The General Aggression Model (GAM) adopted by Kowalksi et al., [42] to study cyberbullying states that there are three routes to cyberbullying, those based on cognition, affect, and arousal. In this work a number of features found to be significant belong to affect and arousal, which is consistent with the GAM. In general more cyberbullying tends to occur in the presence of negative content and also one which arouses the users in a significant way. From a cognition perspective, GAM states that some input variables influence aggressive behavior by increasing the relative accessibility of aggressive concepts in memory and a host of factors, such as media violence, can prime aggressive thoughts [5]. Taken together, these routes (cognition, affect, and arousal) can be

considered factors that interfere with inhibition of aggression. For example, with high levels of stress or arousal, the individuals are unable to inhibit their aggression, and engage in cyberbullying more often. This is also consistent with the General Strain Theory as adapted to study cyberbullying, which suggests that individuals who experience significant strain will develop anger and frustration in response, which places them at a higher risk for engaging in deviant behavior [25].

The findings of this work can also be interpreted based on the advancements in the field of media psychology. For example, one way to analyze media sessions is as those where the comments are a response to the original poster's audio-video content uploaded. Zillmann's theory of "Excitation transfer" suggests that viewer's are physiologically aroused when they watch aggressive media. After watching an aggressive scene, an individual could become aggressive due to the arousal from the scene. The comments generated after the audio-video content is posted can be considered to be primed by the original post. Hence, a more arousing video content is more likely to be followed by a more explicit comments section- thereby increasing the odds of cyberbullying. At the same time, each modality has its own peculiarities and one cannot expect the different modalities to become perfect replicas of each other. This is espoused under the Medium Theory and in fact, Marshal McLuhan famously argued that "Medium is the message" [50, 54].

The combination of different modalities to convey messages and yield cyberbullying can also be interpreted based on the Limited Capacity Model of Mediated Motivated Message Processing (LC4MP), which investigates the real-time processing of mediated messages [44]. Some of the core beliefs of LC4MP include that humans have a limited cognitive capacity and often take shortcuts to information processing. Hence, they often respond to different channels in proportion to the intensity stimulus received. Hence, more emotional and emotion-arousing content can again be understood to yield more emotional response from the viewers potentially including those involving cyberbullying [3].

One aspect which was found to be different in this work compared to the existing literature was the negative association between upper character and punctuation usage and cyberbullying. We notice that while negative sentiment is positively associated with cyberbullying, the use of uppercase characters and punctuation marks is *not* positively associated with cyberbullying. This suggests that the use of punctuation marks and upper characters is not a proxy for negative content as posited in our initial discussion. Based on observing a few of the text samples, we find that the lack of uppercase characters and punctuation marks also happens when the users choose to be casual or careless with their use of language. For instance, they may not use the full stops and commas at the right places and not capitalize the first characters in sentences. We consider this to be an interesting finding and plan to investigate this aspect in more detail in our future work.

We also notice a general consistency, but not an exact match, in the direction of associations between cyberbullying and the corresponding features across the three modalities. If we consider Zillmann's theory of "Excitation transfer" to be the only explanation, we would have expected an exact replication of the associations across modalities. On the other hand, the Medium Theory would have suggested very stark differences across modalities. In reality, we find the associations to paint a more nuanced picture. The associations followed a general consistency across modalities rather than being replications of each other.

Our second research question asked if audio and video analysis help improve cyberbullying detection in social media beyond that obtained by text analysis, and if these features could be used in the context of early detection. Based on the discussion in the previous sections, we see the clear value of using audio and video signals as complementary signals to text signals to automatically detect cyberbullying. We notice a significant jump in the performance of the classification algorithms in terms of multiple metrics like accuracy, ROC area, and F-score. We also found that our method

generalized well to the early detection problem, and maintained a similar level of performance using only a small portion of the textual content.

Note that the results do not close the doors on improving the results further using more sophisticated text analysis. Rather, the results motivate opening new doors in terms of automated audio & video analysis and early detection, which may be relevant in many emerging trends in online social interaction. Additionally, it may be useful in future works to investigate better methods of multimodal decision systems such as late-fusion [78].

8.1 Design implications

Tackling cyberbullying and other types of negative content has been a top priority for multiple online social networks. For instance, Youtube has an explicit Harassment and Cyberbullying policy [91] and Facebook maintains a Bullying Prevention Hub [26]. Facebook CEO Mark Zuckerberg stated in the recent F8 annual conference that "We need to do more to keep people safe and we will" [84]. Facebook already employs 15,000 human moderators to screen and remove offensive content, and it plans to hire another 5,000 by the end of this year, Zuckerberg said in the recent testimony to the US Congress [71].

However, these problems are likely to only get exacerbated with the growth in audio-video content on these platforms [71]. Currently, there are very few empirical insights on the video and audio features that are highly associated with cyberbullying. The findings from this paper could be used by online platforms like Facebook, Instagram and Youtube to design better automated detectors for cyberbullying. These detectors would empower community members to identify cyberbullying content at different stages in the lifetime of user-generated content on their platforms. Following the life-cycle of a Youtube video, for example – videos could initially be vetted and pre-screened based on the visual and audio content, and then if a video passes that stage, the platform and community members could then identify cyberbullying in the comment section.

We note that although the performance of these detectors is increased through the use of audio-visual features, they are still not at the point where they could work autonomously. This would run the risk of misclassifications, in which innocuous content could be falsely labeled as cyberbullying. A detector at this level of performance would be best used in conjunction with a human reviewer, such that the detector flags potential bullying content for human review [12]. Considering the extremely large volume of content that social networks process, increased detector performance could significantly lessen the load on human reviewers. Additionally, these detectors could also be integrated into systems that trigger reflective mechanisms at the time of submission [20, 37]. A detector such as ours that is capable of early detection would be a valuable tool in identifying high-risk posts [92]. The more accurate and consistent such an early detector is, the more likely users are to heed its reflective messages or warnings.

8.2 Theoretical implications

This paper adds to the scientific knowledge about the phenomena of cyberbullying. Just as research contrasting cyberbullying with traditional bullying led to enhanced understanding and suggestions to prevent or reduce it, an exploration of differences between "traditional" cyberbullying (i.e., text-based messages on web 2.0 sites), and emerging challenges of mobile, "appcentric," audio-visual-textual cyberbullying is a vital first step towards mitigating its effects.

This work provides empirical evidence for multiple theories related to cyberbullying. The findings connecting emotional and emotion-arousing content with cyberbullying were consistent with the General Aggression Model, LC4MP, and the General Strain Theory. At the same time, these findings provide partial support for the Medium Theory and the "Excitation Transfer" theories. The associations followed a general consistency across modalities rather than implying a set percent transfer of excitation across modalities.

There are as yet very few studies that have empirically studied the interconnections between audio and video features and cyberbullying. Hence, this work sheds some initial light on these aspects while encouraging future work to look into these aspects in detail. A more theoretically inclined researcher could re-examine the empirical evidence obtained here to extend for instance, the General Aggression Model across modalities, or combine the predictions of Medium Theory and Excitation Transfer theories into a more comprehensive cyberbullying theory in the future.

8.3 Ethical considerations

Cyberbullying has multiple negative implications for those affected and hence its automatic detection has some clear benefits. However, identifying individuals as both victims and bullies can have negative consequences. For instance, identified victims may be targeted for further bullying and identified bullies may face administrative or even legal action. Given, the above considerations, we choose to not disclose any directly identifying information about the individuals in the dataset. Further, we do not try to identify who is the bully in this work but rather focus on whether there is bullying present in the media session. This work recognizes bullying as a behavior rather than identity and does not consider "bully" and "victim" to be static labels. There could also be some negative feelings aroused among readers of this work. We include explicit warnings before presenting any of the examples. A further more comprehensive set of guidelines on how exactly to research and share information among CHI/CSCW researchers is an important avenue for further work in its own right [4].

8.4 Limitations and opportunities for future work

We also note certain limitations of the current study. First, we note that the results are based on a single modest-sized data-set and the considered Vine platform is no longer actively used for posting videos (it was bought by Twitter and ultimately closed down). However, many similar social networks such as Snapchat and Instagram also now support video posts. Furthermore, there exist other similar audio and video-based social network platforms like Clips, Prisma, and Boomerang, which are increasingly becoming popular. We expect this trend to continue and become even more prominent with the recent launch of IGTV by Facebook. Hence, the proposed approach may be applicable to a wide variety of emerging scenarios that reflect the trend in social media platforms toward video-based content. Many of the insights gained in this work by analyzing the Vine data-set (e.g., the design of audio and visual features, their effect directions, approach for their automated computation, and the overall multimodal approach for better detection) will likely be transferable when considering cyberbullying cases on these adjacent platforms.

Next, we acknowledge that data-set used in this work is somewhat small in size. The modest data-set size is in part attributable to the careful human screening required, typically by multiple, validated users, to create such data-sets [33]. The data-set is, however, similar in size to other recent cyberbullying detection efforts [34, 77, 92]. The high cost of manual detection, in fact, motivates more research on automatic cyberbullying detection. Additionally, many multimodal social networks, such as Instagram, have recently restricted API access, and some, such as Snapchat, do not provide a public-facing API at all.

Despite these limitations, this work has multiple implications for social computing research. Making cyberspaces safe and accessible for all users is an important research priority. With the growth curves in cameras, phones, and multimodal content, it is extremely important to automatically detect and prevent cyberbullying instances on multimodal platforms. This work marks the first concerted effort at utilizing *automated* audio and video content analysis for cyberbullying detection. The results obtained, and more importantly the groundwork laid, pave the path for significant advancements in multimodal cyberbullying detection.

9 CONCLUSION

This work tackles the problem of cyberbullying detection in multimodal social media environments. It surveys the existing literature on cyberbullying detection to identify multiple textual, audio, and visual features for cyberbullying detection. These features are evaluated using multiple emerging APIs and combined to create multimodal cyberbullying detectors. The results identify a number of audio-visual features that are found to be associated with cyberbullying. They also suggest that audio-visual features can help improve the performance of purely textual cyberbullying detectors, and can facilitate early detection of cyberbullying. These results pave the way for further research on multimodal cyberbullying detection, which could improve the quality of life of users, and even save lives.

ACKNOWLEDGMENTS

This material is in part based upon work supported by the National Science Foundation under Grant No. 1464287.

REFERENCES

[1] Denise E Agosto, Andrea Forte, and Rathe Magee. 2012. Cyberbullying and teens: what YA librarians can do to help. *Young Adult Library Services* 10, 2 (2012), 38.

[2] Sweta Agrawal and Amit Awekar. 2018. Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms. *arXiv preprint arXiv:1801.06482* (2018).

[3] Saleem Alhabash, Jong-hwan Baek, Carie Cunningham, and Amy Hagerstrom. 2015. To comment or not to comment?: How virality, arousal level, and commenting behavior on YouTube videos affect civic behavioral intentions. *Computers in human behavior* 51 (2015), 520–531.

[4] Nazanin Andalibi, Pinar Öztürk, and Andrea Forte. 2017. Sensitive Self-disclosures, Responses, and Social Support on Instagram: The Case of# Depression.. In *CSCW*. 1485–1500.

[5] Craig A Anderson and Brad J Bushman. 2002. Human aggression. *Annual review of psychology* 53 (2002).

[6] Zahra Ashktorab and Jessica Vitak. 2016. Designing Cyberbullying Mitigation and Prevention Solutions through Participatory Design With Teenagers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 3895–3905.

[7] Alexandra Balahur. 2013. Sentiment Analysis in Social Media Texts.. In *WASSA@ NAACL-HLT*. 120–128.

[8] Linda Beckman, Curt Hagquist, and Lisa Hellström. 2012. Does the association with psychosomatic health problems differ between cyberbullying and traditional bullying? *Emotional and behavioural difficulties* 17, 3-4 (2012), 421–434.

[9] Tibor Bosse and Sven Stam. 2011. A normative agent system to prevent cyberbullying. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2011 IEEE/WIC/ACM International Conference on*, Vol. 2. IEEE, 425–430.

[10] Danah Boyd, Alice Marwick, Parry Aftab, and Maeve Koeltl. 2009. The conundrum of visibility: Youth safety and the Internet. *Journal of Children and Media* 3, 4 (2009), 410–419.

[11] Stevie Chancellor, Yannis Kalantidis, Jessica A Pater, Munmun De Choudhury, and David A Shamma. 2017. Multimodal Classification of Moderated Online Pro-Eating Disorder Content. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 3213–3226.

[12] Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. 2017. The Bag of Communities: Identifying Abusive Behavior Online with Preexisting Internet Data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 3175–3187.

[13] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Int. Res.* 16, 1 (June 2002), 321–357. http://dl.acm.org/citation.cfm?id=1622407.1622416

[14] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, New York, NY, USA, 1217–1230.

[15] Clarifai. 2017. General Model. https://www.clarifai.com/models/general-image-recognition-model/aaa03c23b3724a16a56b629203edc62c [Online; accessed 29-August-2017].

[16] Niall J Conroy, Victoria L Rubin, and Yimin Chen. 2015. Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology* 52, 1 (2015), 1–4.

[17] M. Dadvar, Franciska M.G. de Jong, Roeland J.F. Ordelman, and Rudolf Berend Trieschnigg. 2012. *Improved cyberbullying detection using gender information.* Ghent University, 23–25.

[18] Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. 2013. Improving cyberbullying detection with user context. In *Advances in Information Retrieval*. Springer, 693–696.
[19] Nicholas A Diakopoulos. 2015. The Editor's Eye: Curation and Comment Relevance on the New York Times. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 1153–1157.
[20] Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. 2012. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 2, 3 (2012), 18.
[21] Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of Textual Cyberbullying. In *The Social Mobile Web*.
[22] Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of Textual Cyberbullying. *The Social Mobile Web* 11, 02 (2011), 11–17.
[23] Julian J Dooley, Therese Shaw, and Donna Cross. 2012. The association between the mental health and behavioural problems of students and their reactions to cyber-victimization. *European Journal of Developmental Psychology* 9, 2 (2012), 275–289.
[24] Justin Ellis. 2015. What happened after 7 news sites got rid of reader comments. http://www.niemanlab.org/2015/09/what-happened-after-7-news-sites-got-rid-of-reader-comments/. [Online; accessed 19-Sep-2017].
[25] Dorothy L Espelage, Mrinalini A Rao, and Rhonda G Craven. 2012. Theories of cyberbullying. *Principles of cyberbullying research: Definitions, measures, and methodology* (2012), 49–67.
[26] Facebook. [n. d.]. Bullying Prevention Hub. https://www.facebook.com/safety/bullying/. Accessed: 2018-06-26.
[27] FFmpeg. 2017. FFmpeg. https://www.ffmpeg.org/ [Online; accessed 29-August-2017].
[28] Figure Eight. 2018. How to Calculate a Confidence Score. https://success.figure-eight.com/hc/en-us/articles/201855939-How-to-Calculate-a-Confidence-Score, Accessed: 2018-03-01.
[29] Johnny R.J. Fontaine, Klaus R. Scherer, Etienne B. Roesch, and Phoebe C. Ellsworth. 2007. The World of Emotions is not Two-Dimensional. *Psychological Science* 18, 12 (2007), 1050–1057. https://doi.org/10.1111/j.1467-9280.2007.02024.x arXiv:https://doi.org/10.1111/j.1467-9280.2007.02024.x PMID: 18031411.
[30] Gerald Friedland and Ramesh Jain. 2014. *Multimedia Computing*. Cambridge University Press.
[31] Theodoros Giannakopoulos. 2015. pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis. *PloS one* 10, 12 (2015).
[32] Sameer Hinduja and Justin W Patchin. 2012. *School climate 2.0: Preventing cyberbullying and sexting one classroom at a time*. Corwin Press.
[33] Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2015. Analyzing Labeled Cyberbullying Incidents on the Instagram Social Network. In *International Conference on Social Informatics*. Springer, 49–66.
[34] Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2015. Detection of cyberbullying incidents on the instagram social network. *arXiv preprint arXiv:1503.03909* (2015).
[35] Qianjia Huang, Vivek Kumar Singh, and Pradeep Kumar Atrey. 2014. Cyber bullying detection using social and textual analysis. In *Proc. Int. Workshop on Socially-Aware Multimedia*. ACM, 3–6.
[36] C.J. Hutto and Eric Gilbert. 2015. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. In *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*.
[37] Birago Birago Korayga Jones. 2012. *Reflective interfaces: Assisting teens with stressful situations online*. Ph.D. Dissertation. Massachusetts Institute of Technology.
[38] James J. Kellaris and Ronald C. Rice. 1993. The influence of tempo, loudness, and gender of listener on responses to music. *Psychology and Marketing* 10, 1 (1993), 15–29. https://doi.org/10.1002/mar.4220100103
[39] Kurt Koffka. 2013. *Principles of Gestalt psychology*. Vol. 44. Routledge.
[40] Janet Kornblum. 2008. Cyberbullying grows bigger and meaner with photos, video. http://usatoday30.usatoday.com/tech/webguide/internetlife/2008-07-14-cyberbullying_N.htm. *USA Today* (2008).
[41] Rajitha Kota, Shari Schoohs, Meghan Benson, and Megan A Moreno. 2014. Characterizing cyberbullying among college students: Hacking, dirty laundry, and mocking. *Societies* 4, 4 (2014), 549–560.
[42] Robin M Kowalski, Gary W Giumetti, Amber N Schroeder, and Micah R Lattanner. 2014. Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth. *Psychological bulletin* 140, 4 (2014), 1073.
[43] Robin M Kowalski and Susan P Limber. 2013. Psychological, physical, and academic correlates of cyberbullying and traditional bullying. *Journal of Adolescent Health* 53, 1 (2013), S13–S20.
[44] Annie Lang. 2009. The limited capacity model of motivated mediated message processing. *The SAGE handbook of media processes and effects* (2009), 193–204.
[45] Ingo Lutkebohle. 2016. Recognize Emotions in Images. https://www.microsoft.com/cognitive-services/en-us/emotion-api/. [Online; accessed 19-July-2016].

[46] Jing Ma, Wei Gao, Zhongyu Wei, Yueming Lu, and Kam-Fai Wong. 2015. Detect rumors using time series of social context information on microblogging websites. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, 1751–1754.
[47] Paul E Madlock and David Westerman. 2011. Hurtful Cyber-Teasing and Violence WhoÃćÂĂÂŹs Laughing Out Loud? *Journal of interpersonal violence* 26, 17 (2011), 3542–3560.
[48] Brendan Maher. 2016. Can a video game company tame toxic behaviour? *Nature* 531, 7596 (2016), 568–571.
[49] Massimo Marchiori. 2017. The secure mobile teen: Looking at the secret world of children. In *2017 IEEE 13th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*. IEEE, 341–348.
[50] Marshall McLuhan and Quentin Fiore. 1967. The medium is the message. *New York* 123 (1967), 126–128.
[51] FrontGate Media. 2017. A LIST OF 723 BAD WORDS TO BLACKLIST & HOW TO USE FACEBOOK'S MODERATION TOOL. http://www.frontgatemedia.com/new/wp-content/uploads/2014/03/Terms-to-Block.csv [Online; accessed 29-August-2017].
[52] Ersilia Menesini and Annalaura Nocentini. 2009. Cyberbullying definition and measurement: Some critical considerations. *Zeitschrift für Psychologie/Journal of Psychology* 217, 4 (2009), 230–232.
[53] Ersilia Menesini, Annalaura Nocentini, and Pamela Calussi. 2011. The measurement of cyberbullying: Dimensional structure and relative item severity and discrimination. *Cyberpsychology, Behavior, and Social Networking* 14, 5 (2011), 267–274.
[54] Joshua Meyrowitz. 2008. Medium theory. *The international encyclopedia of communication* (2008).
[55] Vinita Nahar, Xue Li, and Chaoyi Pang. 2013. An Effective Approach for Cyberbullying Detection. *Communications in Information Science and Management Engineering* 3, 5 (2013), 238–247.
[56] Vinita Nahar, Xue Li, Chaoyi Pang, and Yang Zhang. 2013. Cyberbullying detection based on text-stream classification. In *The 11th Australasian Data Mining Conference (AusDM 2013)*.
[57] Vinita Nahar, Sayan Unankard, Xue Li, and Chaoyi Pang. 2012. Sentiment analysis for effective detection of cyber bullying. In *Web Technologies and Applications*. Springer, 767–774.
[58] National Crime Prevention Council. 2014. Stop bullying before it starts. http://www.ncpc.org/resources/files/pdf/bullying/cyberbullying.pdf. Accessed: 2017-06-10.
[59] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 145–153.
[60] Bridianne O'Dea, Stephen Wan, Philip J Batterham, Alison L Calear, Cecile Paris, and Helen Christensen. 2015. Detecting suicidality on Twitter. *Internet Interventions* 2, 2 (2015), 183–188.
[61] OpenCV. 2017. Reading and Writing Images and Video. http://docs.opencv.org/2.4/modules/highgui/doc/reading_and_writing_images_and_video.html [Online; accessed 29-August-2017].
[62] F Javier Ortega, José A Troyano, FermíN L Cruz, Carlos G Vallejo, and Fernando EnríQuez. 2012. Propagation of trust and distrust for the detection of trolls in a social network. *Computer Networks* 56, 12 (2012), 2884–2895.
[63] Justin W Patchin and Sameer Hinduja. 2012. *Cyberbullying prevention and response: Expert perspectives*. Routledge.
[64] Jessica A Pater, Andrew D Miller, and Elizabeth D Mynatt. 2015. This Digital Life: A Neighborhood-Based Study of Adolescents' Lives Online. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2305–2314.
[65] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
[66] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543. http://www.aclweb.org/anthology/D14-1162
[67] T Pradheep, JI Sheeba, T Yogeshwaran, and S Pradeep Devaneyan. 2017. Automatic Multi Model Cyber Bullying Detection from Social Networks. (2017).
[68] R. I. Rafiq, H. Hosseinmardi, R. Han, Q. Lv, S. Mishra, and S. A. Mattso. 2015. Careful what you share in six seconds: Detecting cyberbullying instances in Vine. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 617–622. https://doi.org/10.1145/2808797.2809381
[69] Rahat Ibn Rafiq, Homa Hosseinmardi, Sabrina Arredondo Mattson, Richard Han, Qin Lv, and Shivakant Mishra. 2016. Analysis and detection of labeled cyberbullying instances in Vine, a video-based social network. *Social Network Analysis and Mining* 6, 1 (2016), 88.
[70] Elaheh Raisi and Bert Huang. 2016. Cyberbullying identification using participant-vocabulary consistency. *arXiv preprint arXiv:1606.08084* (2016).
[71] MIT Technology Review. 2018. Intelligent Machines - Three problems with Facebook's plan to kill hate speech using AI. https://www.technologyreview.com/s/610860/three-problems-with-facebooks-plan-to-kill-hate-speech-using-ai/. Accessed: 2018-07-06.

[72] Kelly Reynolds, April Kontostathis, and Lynne Edwards. 2011. Using machine learning to detect cyberbullying. In *Proc. Int. Conf. Machine Learning and Applications and Workshops (ICMLA)*, Vol. 2. IEEE, 241–244.
[73] Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management* 24, 5 (1988), 513–523.
[74] Steven J Seiler and Jordana N Navarro. 2014. Bullying on the pixel playground: Investigating risk factors of cyberbullying at the intersection of childrenÃćÂĂÂŹs online-offline social lives. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace* 8, 4 (2014).
[75] Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. 2016. Hoaxy: A platform for tracking online misinformation. In *Proceedings of the 25th International Conference Companion on World Wide Web*. International World Wide Web Conferences Steering Committee, 745–750.
[76] Vivek Singh, Marie Radford, Huang Qianjia, and Susan Furrer. 2017. "They basically like destroyed the school one day": On Newer App Features and Cyberbullying in Schools. In *Proceedings of the international conference on Computer Supported Collaborative Work and Social Computing (CSCW)*. ACM, 1210–1216.
[77] Vivek K Singh, Souvick Ghosh, and Christin Jose. 2017. Toward Multimodal Cyberbullying Detection. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 2090–2099.
[78] Vivek Kumar Singh, Qianjia Huang, and Pradeep Kumar Atrey. 2016. Cyberbullying Detection Using Probabilistic Socio-Textual Information Fusion. In *Proc. IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ACM.
[79] Robert Slonje and Peter K Smith. 2008. Cyberbullying: Another main type of bullying? *Scandinavian journal of psychology* 49, 2 (2008), 147–154.
[80] Peter K Smith, Jess Mahdavi, Manuel Carvalho, Sonja Fisher, Shanette Russell, and Neil Tippett. 2008. Cyberbullying: Its nature and impact in secondary school pupils. *Journal of child psychology and psychiatry* 49, 4 (2008), 376–385.
[81] Peter K Smith, Jess Mahdavi, Manuel Carvalho, and Neil Tippett. 2006. 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).
[82] Andre Sourander, Anat Brunstein Klomek, Maria Ikonen, Jarna Lindroos, Terhi Luntamo, Merja Koskelainen, Terja Ristkari, and Hans Helenius. 2010. Psychosocial risk factors associated with cyberbullying among adolescents: A population-based study. *Archives of general psychiatry* 67, 7 (2010), 720–728.
[83] Junming Sui. 2015. *Understanding and fighting bullying with machine learning*. Ph.D. Dissertation. UNIVERSITY OF WISCONSIN–MADISON.
[84] USA Today. 2018. Mark Zuckerberg pledges Facebook will put 'people first,' avoid past mistakes. https://www.usatoday.com/story/tech/news/2018/05/01/mark-zuckerberg-pledges-facebook-put-people-first-avoid-past-mistakes/564474002/. Accessed: 2018-07-06.
[85] Robert S Tokunaga. 2010. Following you home from school: A critical review and synthesis of research on cyberbullying victimization. *Computers in human behavior* 26, 3 (2010), 277–287.
[86] Carnegie Mellon University. 2017. CMU Sphinx. https://cmusphinx.github.io/ [Online; accessed 29-August-2017].
[87] Kris Varjas, Christopher C Henrich, and Joel Meyers. 2009. Urban middle school students' perceptions of bullying, cyberbullying, and school safety. *Journal of School Violence* 8, 2 (2009), 159–176.
[88] William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*. Association for Computational Linguistics, 19–26.
[89] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods* 45, 4 (01 Dec 2013), 1191–1207. https://doi.org/10.3758/s13428-012-0314-x
[90] Ralf Wölfer, Anja Schultze-Krumbholz, Pavle Zagorscak, Anne Jäkel, Kristin Göbel, and Herbert Scheithauer. 2014. Prevention 2.0: Targeting cyberbullying@ school. *Prevention Science* 15, 6 (2014), 879–887.
[91] Youtube. 2016. Harassment and cyberbullying policy. https://support.google.com/youtube/answer/2802268?hl=en&ref_topic=2803176. Accessed: 2018-06-26.
[92] Rui Zhao, Anna Zhou, and Kezhi Mao. 2016. Automatic detection of cyberbullying on social networks based on bullying features. In *Proceedings of the 17th International Conference on Distributed Computing and Networking*. ACM, 43.
[93] Haoti Zhong, Hao Li, Anna Squicciarini, Sarah Rajtmajer, Christopher Griffin, David Miller, and Cornelia Caragea. 2016. Content-driven Detection of Cyberbullying on the Instagram Social Network. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI'16)*. AAAI Press, 3952–3958. http://dl.acm.org/citation.cfm?id=3061053.3061172

Received April 2018; revised July 2018; accepted September 2018.