Getting Meta: A Multimodal Approach for Detecting Unsafe Conversations within Instagram Direct Messages of Youth

SHIZA ALI, Boston University, USA
AFSANEH RAZI, Drexel University, USA
SEUNGHYUN KIM, Georgia Institute of Technology, USA
ASHWAQ ALSOUBAI, Vanderbilt University, USA
CHEN LING, Boston University, USA
MUNMUN DE CHOUDHURY, Georgia Institute of Technology, USA
PAMELA J. WISNIEWSKI, Vanderbilt University, USA
GIANLUCA STRINGHINI, Boston University, USA

Instagram, one of the most popular social media platforms among youth, has recently come under scrutiny for potentially being harmful to the safety and well-being of our younger generations. Automated approaches for risk detection may be one way to help mitigate some of these risks if such algorithms are both accurate and contextual to the types of online harms youth face on social media platforms. However, the imminent switch by Instagram to end-to-end encryption for private conversations will limit the type of data that will be available to the platform to detect and mitigate such risks. In this paper, we investigate which indicators are most helpful in automatically detecting risk in Instagram private conversations, with an eye on high-level metadata, which will still be available in the scenario of end-to-end encryption. Toward this end, we collected Instagram data from 172 youth (ages 13-21) and asked them to identify private message conversations that made them feel uncomfortable or unsafe. Our participants risk-flagged 28,725 conversations that contained 4,181,970 direct messages, including textual posts and images. Based on this rich and multimodal dataset, we tested multiple feature sets (metadata, linguistic cues, and image features) and trained classifiers to detect risky conversations. Overall, we found that the metadata features (e.g., conversation length, a proxy for participant engagement) were the best predictors of risky conversations. However, for distinguishing between risk types, the different linguistic and media cues were the best predictors. Based on our findings, we provide design implications for AI risk detection systems in the presence of end-to-end encryption. More broadly, our work contributes to the literature on adolescent online safety by moving toward more robust solutions for risk detection that directly takes into account the lived risk experiences of youth.
Content Warning: *This paper discusses sensitive topics which may be triggering.*

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**.

Additional Key Words and Phrases: online risk detection, machine learning, ensemble models, social media, Instagram, end-to-end encryption

Authors' addresses: Shiza Ali, Boston University, Boston, Massachusetts, USA, shiza@bu.edu; Afsaneh Razi, Drexel University, Philadelphia, Pennsylvania, USA, afsaneh.razi@drexel.edu; Seunghyun Kim, Georgia Institute of Technology, Atlanta, Georgia, USA, seunghyun.kim@gatech.edu; Ashwaq Alsoubai, Vanderbilt University, Nashville, Tennessee, USA, ashwaq.alsoubai@vanderbilt.edu; Chen Ling, Boston University, Boston, Massachusetts, USA, ccling@bu.edu; Munmun De Choudhury, Georgia Institute of Technology, Atlanta, Georgia, USA, munmund@gatech.edu; Pamela J. Wisniewski, Vanderbilt University, Nashville, Tennessee, USA, pam.wisniewski@vanderbilt.edu; Gianluca Stringhini, Boston University, Boston, Massachusetts, USA, gian@bu.edu.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
2573-0142/2023/4-ART132 $15.00
https://doi.org/10.1145/3579608

ACM Reference Format:
Shiza Ali, Afsaneh Razi, Seunghyun Kim, Ashwaq Alsoubai, Chen Ling, Munmun De Choudhury, Pamela J. Wisniewski, and Gianluca Stringhini. 2023. Getting Meta: A Multimodal Approach for Detecting Unsafe Conversations within Instagram Direct Messages of Youth. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 132 (April 2023), 30 pages. https://doi.org/10.1145/3579608

1 INTRODUCTION

Instagram, a photo and video-sharing social media platform, has grown in popularity among teenagers and young adults. In fact, 72% of Americans aged 18 to 29 say that they use Instagram and at least 71% of them use it once a week [82]. One reason youth like Instagram is that it helps them connect with their friends and the community at large. According to 56% of Instagram users, "the platform makes them feel more connected to the people they know" [95]. Instagram is also a place where people share their lives with their friends and discover more about themselves; however, social media can also facilitate potential abuse, such malicious users who bully, harass, or sexually groom impressionable youth [86, 118]. Recently, Instagram came under fire after leaked internal research claimed that the parent company Facebook (now Meta) was aware of the site's negative mental health effects on youth [75]. The claim suggested that the company was aware of the ills of its platforms, such as Instagram's effect on the mental health and body image of teenage girls. According to Facebook's internal research, 13.5% of teen girls believe Instagram worsened suicidal thoughts, and 17% believe it contributed to eating disorders [13].

With the surge of malicious actors on social media, researchers in the Human-Computer Interaction (HCI) and Artificial Intelligence (AI) fields have shifted to automated risk detection being a potential way to address this societal problem; for example, by developing automated approaches to detect cyberbullying [122, 124], suicidal ideation [26], mental health issues [81], sexual solicitations [91] or substance misuse [96]. Researchers developing these automated systems, however, face two main challenges. First, despite the magnitude of hate speech, cyberbullying, and sexual solicitations on Instagram, there is an absence of ecologically valid datasets specific to youth that can help us study and prevent such attacks on youth [15]. Furthermore, there is an over focus on publicly available data [51, 91] rather than private messages which are shown to contain more risky interactions [133]. Further, risk classifiers often do not consider the victim's account of their own risk experiences [50]; instead, they often rely heavily on third-party annotators to identify risks [91]. We believe that it is critical to evaluate the opinions of youth on their experiences in order to find contextual factors that are relevant for risk detection. Lastly, the vast majority of risk detection algorithms, such as those that detect cyberbullying or harassment, use textual features [64, 107]. However, previous research has shown that risk detection algorithms should include information collected from visuals as well as contextual features to take a Human-Centered Machine Learning (HCML) approach to better detect risk in private conversations on Instagram [2, 20, 51, 91].

Second, communication on Instagram is inherently multimodal (i.e., including images, text, and videos), and therefore adopting a multimodal approach for risk detection is particularly important [20], especially for detecting risk that can be presented in different ways, such as text and images, and also because teens use more image/videos than text based communication [11]. Multimodal machine learning aims to build models that can process and relate information from multiple modalities [9] especially text and images that are very different from one another, thus it has been used to predict popularity of social media photos [46, 70], perform sentiment analysis [66, 105], perform Instagram post analysis [73] and detect false information [38].

However, combining information from text and media presents both technical and computational challenges. For example, analyzing multimodal data may not be feasible given Meta's agenda to move towards end-to-end encryption [14], which will inhibit the use of content-based features for

classification, not only for external researchers, but for the platform itself, since the only people able to see end-to-end encrypted content are senders and recipients. In preparation for this shift, it is important for the research community to investigate the feasibility of adopting detection solutions that do not take content into account and are therefore compatible with end-to-end encryption. An example of the features that could be used in this setting is metadata features like the number of messages that are exchanged in a conversation, their inter-arrival time, and the direction of messages (e.g., sent from the perpetrator to the victim or vice versa, as well as the relationship between the victim and perpetrator. In this study, we take a multimodal approach to detect risk in private conversations on Instagram specifically to the domain of youth online risk detection by analyzing metadata, textual, and image-based features that are indicative of conversational patterns and later detecting the risk posit in the conversation to address the following research questions:

- **RQ1:** *Can we identify unsafe conversations from safe ones using metadata features - or are linguistic and image features necessary?*
- **RQ2:** *Can we accurately detect the types of risk presented in an unsafe conversation?*

Answering RQ1 is useful to identify promising directions in which risk detection algorithms should move in the face of end-to-end encryption. Answering RQ2 will help researchers and platforms design detection systems that allow to take the specific type of risk into account and adopt appropriate countermeasures. To answer these questions, we collected a dataset of private Instagram conversations from 172 teens and young adults aged 13-21. Our experiments relied on a dataset of 28,725 conversations containing approximately 5 million Instagram private messages. We then created a balanced dataset of randomly chosen safe versus unsafe conversations to train a conversation-level risk ensemble classifier to help us answer RQ1. We tested several machine learning models including Decision Trees, Random Forest, and Linear SVM models and extracted the useful features. We found that while linguistic cues and image features helped in detecting risk when used in ensemble with a metadata classifier (AUC=0.83), metadata features provided an accuracy close to the ensemble (AUC=0.81). This allows us to answer RQ1 affirmatively, showing that in the presence of end-to-end encryption conversation-level characteristics could still allow platforms to detect risk and protect their users.

Next, we developed risk-type classifiers using a multi-class model approach for our three different feature sets for predicting the specific risk type in a given conversation with an accuracy (RQ2). We found that in this case, the contextual information provided by linguistic cues and image features is needed to make an accurate classification (AUC=0.80 using a weighted-vote ensemble). In light of these findings, we discuss design implications for platforms moving towards end-to-end encryption which would limit the ability for any kind of risk detection involving analysis of the actual content of the conversations. In summary, our paper makes the following contributions:

- We identified three key feature sets (i.e., metadata, linguistic cues, and image features) that can be used to detect whether a conversation is risky or not for youth. In addition, we designed an ensemble classifier using the different feature sets to predict the types of risks in a private conversation.
- We highlight that unsafe conversations can be detected by meta-level features based on the conversation structure (i.e., the number of users in the conversation, the length of the conversation, etc.); however, it is critical to consider the content of these conversations (i.e., linguistic and image features) to differentiate between different risk types.
- We provide design implications for machine learning (ML) approaches, especially in the presence of end-to-end encryption to accurately detect the various types of risks encountered by youth online.

2 RELATED WORK

In this section, we review research on online risks and the need for automated approaches to detect them. Finally, we highlight the contributions of our paper based on the identified research gaps.

2.1 The Online Risks Commonly Encountered by Youth

The rapid expansion of social networking sites among teens has raised awareness of their potential positive and negative effects on youth health and development. There have been public reports of young people being sexually approached and harassed on social networking platforms, putting them at greater risk of sexual victimization [133], as well as reports of cyberbullying [3, 18, 128], violence [39, 92], and exposure to explicit content [57, 59, 74, 76]. The prevalence of unwanted sexual solicitation and exposure to explicit content among youth has also increased, in fact, one in nine youth experience online sexual solicitation [61]. Females are also frequently subjected to unwanted nudity from strangers and struggle with how to decline sexting requests from people they know [42, 43, 90]. Furthermore, cyberbullying has emerged as a potential harm, raising questions regarding its influence on mental health [41]. Hamm et al. found that there is a consistent relationship between cyberbullying and depression among youth. There has also been research on how negative online risk events may cause post-traumatic stress disorder (PTSD) symptoms in teens [68] also while the majority of young people are generally resilient online, a susceptible minority report experiencing various harms as a result of their online activities [33]. Conversations on social media may accelerate harm on vulnerable youth, for example social media may provide a sense of community for deliberate self-harmers and offer supportive advice [31].

All these concerns have been studied in the past, however, these dangers are occasionally investigated together, but more commonly they are investigated separately [130]. We take upon a more holistic approach by first treating "risk" more generally as being any interactions that youth may find uncomfortable or unsafe. Later, we expand on risky conversations and detect different types of risks that youth encounter in their private messages on Instagram.

2.2 The Need for Robust Online Risk Detection and Mitigation

While there is a wealth of research on youth online safety, there is a lack of intervention-based approaches to mitigate online risks [83]. A major line of research in youth online safety has leveraged Artificial Intelligence (AI) based risk prevention interventions for automatically detecting risks such as suicide [24, 60, 93], depression [58, 106], online sexual risks including solicitations and unsolicited exposure [91, 115], and cyberbullying [51]. Yet, most of these AI-based risk detection approaches have been studied without considering the intertwined nature of youth risk experiences, which potentially limits the applicability and reliability of these models in a real-world setting. There is a need for viable solutions that not only empower teenagers to engage with technology, but also remain safe from undue harm. In fact, youth find automated risk detection approaches a promising way to address their online safety concerns [8]. Badillo et al. [8] found that depending on the severity of the risk, children desired varying levels of agency. In most cases, they preferred to solve the problem themselves rather than relying on their parents. Thus, by creating intelligent agents for helping youth cope with online risks, we can promote more resilience-based and self-regulatory approaches for online safety that move beyond parental controls [129] and, we can potentially shift the onus of youth protection from parents to the social media companies that put youth in harms way [15].

2.3 Gaps in Current Automated Approaches for Risk Detection

There has been a myriad of research on detecting online risks, such as hate speech [36, 98], harassment [32], and cyberbullying [19] as well as detecting abusive language [17] in online conversations, particularly in public spaces. However, there are gaps in the current automated approaches for risk detection including lack of user-informed ground truth, emphasis on public over private data, over-focus on textual data or media, rather than both, and focusing on one type of risk rather than looking at risks more holistically [91]. Many studies rely on public datasets and third-party annotators with a lack of human-centeredness in their approaches.

Most of the literature relies on public posts and images for analysis; however, private conversations are very different from those that take place in the public realm. For example, people share different aspects of their life on "stories" on Instagram as compared to what they show on their feeds where anyone can comment publicly [119]. Indeed Parapar et al. [79] presented machine learning approaches to identify suspicious subjects in chat-rooms using a combination of psycholinguistic, content-based, and chat-based features to detect predation in chatrooms; however, they use publicly available dataset from Perverted Justice (PJ) (a dataset of volunteers acting as teens and convicted sex offenders) to identify predators in these chatrooms. Secondly, this research was not specific to the domain of youth online risk detection. Furthermore, analyzing private social media can provide insight into interplay between user characteristics and sentiments [35]. Gao et al. performed analysis to understand the patterns of textual sentiments and metadata in public and private Facebook posts and concluded that public posts differ significantly, specifically, the texts in public posts are more positive than those in private chats, whereas, in private conversations, people feel more free to express their sentiments.

Secondly, previous research has relied primarily on third-party annotations [91]. For example, Anderson et al. [6] and Miah et al. [71] proposed systems to identify child grooming and child exploitation in online chat conversations. However, research shows that there are low levels of agreement between manual and automated analysis [16] and low levels of agreement between the labeling performed by the researchers and the participants themselves [50]. Such systems lack ecologically valid ground truths, which is why we use annotations done by the participants themselves, adding first-person viewpoints in our detection system.

Lastly, most of the studies on detecting risk in the online space have focused on textual features [91]. However, this approach falls short of addressing the essential question of who is participating in the conversation, messaging frequency, and the media that might be shared in that conversation. For example, Zhang et al. [134] detected early signs of conversation failure. Miah et al. [71] investigated the effectiveness of text classifiers to identify Child Exploitation (CE) in chatting. Text classification has also been used to filter messages for suicide prevention [27]. However, they did not incorporate how media usage, particularly the use of images, can contribute to the overall riskiness of the conversation. It is also crucial to include human-centered elements in such detection systems, such as the relationship of individuals (i.e., our participants) with the people they are conversing with [5]. Ali et al. [2] performed a mixed-method analysis of the media files shared privately in Instagram conversations and established a need to include image features into AI-based risk detection systems.

2.4 The Novelty of Our Approach

In this paper, we present a computational approach to risk detection on a dataset that is not only tailored to youth but also incorporates their own risk perception in private online conversations. We identify and improve on the gaps for an automated risk detection system by utilizing HCML principles [91]. Our study is particularly unique because a) we perform classification on private

conversations that were labeled by the participants themselves and b) we utilize a multimodal approach for risk detection using three different feature sets (e.g., metadata, linguistic cues, and image features) to create an ensemble classifier. Our contribution is an end-to-end multimodal approach, which combines multiple heterogeneous modalities (metadata, linguistic cues, and image features) for the detection of unsafe conversations as well as the detection of risk type in the unsafe conversations. To this end, we use stacking ensemble design [34]. Ensemble classifiers often perform better than single classifiers [29]. Ensemble methods provide significant improvement of accuracy compared to individual classifiers [30, 40, 53, 63]. We also investigated which modality is most important for the detection of risk and whether a combined multimodal analysis is beneficial in contrast to monomodal processing.

3 DATASET

This section describes how we collected our data, our risk-flagging practices, and the ethical considerations that guided this work. We also provide an overview of the number of conversations and messages that we collected, together with demographic information about our participants.

3.1 Data Collection

To collect data for this study we developed a specialized infrastructure as described in our previous work [89]. We designed a social media data collection system, using a secure web-based system based on Amazon Web Services, RDS, EC2, PHP, Python, and other technologies. To be eligible as a participant, one should meet these requirements; 1) Be an English speaker between 13 to 21 years old based in the US. 2) Maintain an active Instagram account for at least three months throughout their teen years (13-17). 3) Must have at least fifteen Direct Message (DMs) conversations 4) Have at least two conversations with messages that made them feel unsafe or uncomfortable.

Guidelines to be considered eligible were clearly communicated to the participants. Participants were advised that risky encounters may include, but are not limited to, the following categories based on existing teenage online risk literature [130] and user experience on Instagram: Nudity/porn, Sexual messages/sexting or solicitations, Harassment, Hate speech, Violence/threat of violence, Sale or promotion of illegal activity, Self-injury, or Other conditions that may cause emotional or physical harm. For eligible participants who are over 18 years old, we asked for an informed consent before the study. For eligible participants who are under 18 years old, we took extra care and obtain informed consent(s) from their parents.

3.2 Data Upload, Risk-Flagging, and Data Verification Process

Participants were asked to request their data from Instagram as ".json" files according to General Data Protection Regulation [37]. Then, we asked them to upload their file to our data collection platform as we mentioned earlier. Upon successful upload of data, we presented their chats in a narrative form and required the participants to flag each chat they uploaded as "safe" or "unsafe." Although we demonstrated the viability of employing predefined risk types, we urged participants to self-assess circumstances in which they feel unsafe or uncomfortable during the message interaction. Further context was also requested for the "unsafe" interactions flagged during the assessment. We used instruction sentences like "describing why the conversations made you or someone else feel uncomfortable or unsafe." on our user interface of the data collection platform. Uploaded and flagged data was verified by trained researchers once the data was collected. Researchers verified whether the participants are attentive enough to assess their chats by checking the length of their completion throughout the flagging process and the realness of the data. Participants who passed the data qualification received a $50 Amazon gift card as compensation.

3.3 Ethical Considerations

Because our study included participants under 18 years old with private conversations on Instagram, we took extra care to protect the confidentiality and privacy of the participants. Firstly, we obtained a Certificate of Confidentiality issued by the National Institute of Health, which protects participant privacy and prohibits the data from being subpoenaed during the legal discovery process. Secondly, to avoid exchanging information with third parties, we chose not to analyze our data utilizing cloud-based services. Thirdly, we required that the researchers who study the data must complete IRB Human Subjects CITI training and are not permitted to save the data to their personal computers. Moreover, in addition to obtaining IRB approval for our study, we emphasized our federal responsibility to report child pornography to the relevant authorities and offered explicit warnings against sharing digital content depicting a minor's nudity. Furthermore, our IRB approved child mandated reporting protocol required us to report any sexual explicit image sent to a minor (i.e., under the age of 18) by an identifiable adult, where it was not possible for us to ascertain the age of the conversation partner in these conversations. Additionally, we paraphrased all quotes included in this paper to protect our participants. We believe our efforts to prevent the leakage of this complicated and sensitive data are best practices.

3.4 Data Overview

We aimed to build a comprehensive dataset using a web-based platform, as mentioned in our previous work [89] for youth (ages 13-21) to donate and annotate their Instagram data. To achieve this goal, we reached a large audience of young people in the United States by contacting more than 650 youth organizations and advertising our study on social media (i.e., Facebook, Instagram, and Twitter). We recruited 172 eligible participants with an average age of 16 years old. Our study's demographic data showed 69% female participants, 23% percent male participants, and 5% non-binary while the rest preferred not to answer. Participants were 41% White, 19% Black/African-American, 16% mixed race, 16% Asian or Pacific Islander, and 8% Hispanic/Latino. Participants came from all around the country, including Florida (12%), California (5%), Indiana (3%), and 28 other states. Participants were predominantly heterosexual or straight (48%) with a few bisexuals (28%) and 11% homosexuals while others decided not to self-identify.

Our participants contributed towards 28,725 chats, including 4,181,970 Instagram private messages. Among these chats, 17,786 of them are flagged as "safe" (15,397) or "unsafe"(3,389). For the 202,613 messages in unsafe chats, 2,947 of them were further flagged by risk type, see Table 1 for the top 5 risk types present in our dataset. Harassment is the most commonly flagged risk type by participants, followed by sexual messages/solicitation.

Table 1. Top 5 risk types annotated by participants

Risk Type	Total Messages	Total Conversations
Harassment	828	507
Sexual messages/solicitation	573	383
Nudity/porn	422	319
Hate speech	152	130
Sale or promotion of illegal activities	144	125

4 METHODS

We now present the key components of an Instagram conversation and the methods used to acquire the multimodal feature sets, and how we developed classifiers to detect different types of risk.

4.1 Characterizing Instagram Conversations

On Instagram, users can have private conversations with their followers/friends as well as strangers. An Instagram conversation comprises of one or more users exchanging messages. These messages can be text or media. We focus on private Instagram conversations because research has shown that the most concerning sexual risks and harassment occur in private, for example in instant messaging and chat rooms [91]. Figure 1 shows an example of a typical Instagram conversation.

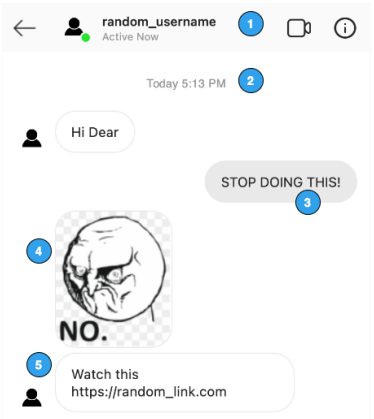

Fig. 1. An example of a conversation on Instagram:(1) username of the other person involved, (2) conversation start timestamp, (3) a text message, (4) an image message and, (5) message containing a link.

For the sake of our research we encode Instagram conversations into three components:

- **Metadata:** The recipient(s), i.e., who the message was sent. Number of people involved in the conversation. The length of the conversation (number of messages).
- **Linguistic Cues:** The actual text messages shared in the conversation.
- **Image Features:** Images shared in the conversation.

Then we designed a set of classifiers for each of the feature sets and then combined their verdicts together to obtain a holistic risk detector. We implemented a multimodal approach because one modality might not be able to characterize all types of risks in an Instagram conversations. The combination of visual and textual content builds up most modern communication today, and in this paper, we extend recent advances in risk detection by putting forward a multimodal approach that is able to predict whether a conversation as a whole is safe or unsafe. While the two modalities (text and image) may encode different information [10, 20, 54], adding metadata (including relationship and engagement factors) may further improve the accuracy of the system and therefore can complement each other. Then we detected risky conversations and risk types using our two-step pipeline system, by first detecting whether a conversation is safe or not, and then further determining for the unsafe conversations as to what risk was associated with them. Thus, we perform binary as well as multi-class classification.

4.2 Binary Classification: Detecting whether a conversation is safe or unsafe (RQ1)

To perform binary classification (that is, categorizing whether a conversation is safe or unsafe), we characterized Instagram conversations using a set of machine learning classifiers, each of which focused on a different set of features (metadata, linguistic cues, and image features) extracted from Instagram conversations. Then, we used the insights derived from this analysis to build an ensemble

classifier geared to determine whether a conversation is deemed safe or unsafe. In the next sections, we discuss each feature set in detail:

4.2.1 Metadata. We use metadata features to inspect some prominent user interaction properties, especially in the private realm, to understand users' engagement in safe vs. unsafe conversations. Building upon previous research done on public Instagram posts [10, 21, 73] and chat-based features to detect predation in chatrooms [79], we select user interaction properties such as levels of engagement of people in the conversation, timings, the number of people involved, and average conversation length as key properties of the conversation. Based on previous research, unsafe conversations are significantly shorter and youth tend to disengage when they encounter them [2]. We further these properties by looking at message and image engagement of the participant, their average response times, and the relationship the participant has had with the other persons they are conversing with, to add human centered insights into our model. Table 2 summarizes the features used for our metadata feature set. All metadata features were extracted from the conversation; however, to extract the relationship the participant had with others, we used the participant's own coded data that they provided while labeling their own dataset.

With the increasing adoption of end-to-end encryption by platforms, metadata features can become particularly important in developing risk detection systems, because while content will cease to be available to anyone who is not the sender of the recipient, high-level conversation information will still be visible to the platform. We tested several machine learning models including Decision Trees [103], Random Forest [104], and Linear SVM [104] with the metadata feature set.

Table 2. Metadata features associated with a given conversation.

Metadata Features
Average conversation length, in seconds
The number of users involved in the conversations.
The total number of messages in the conversation.
The total number of images in the conversation.
The total messages sent by the participant
The total messages received by the participant
The total images sent by the participant
The total images received by the participant
The average response time of the participant
The average response time of the other users in the conversations
The relationship the participant had with others

4.2.2 Linguistic Cues. The main way of communicating in private conversations on Instagram is through text messages. Indeed youth use text messages to reinforce existing relationships, both with friends and romantic partners [84, 114]. Also, past literature shows that the most commonly used features for risk detection systems in research were textual or lexical features [91]. Therefore, we analyze the linguistic cues to predict whether a given conversation is safe or unsafe.

To do this, we employed an end-to-end Convolutional Neural Network (CNN) from the architecture of Kim et al. [52]. Based on our previous work [88] on comparing text classifiers (including Linear Support Vector Machine (SVM), Random Forest (RF), and Logistic Regression (LR), and CNN) for sexual risk detection on the same dataset, we found that CNN outperformed the traditional models. This CNN model configuration was optimized by grid search for hyperparameter tuning.

As input to this CNN model, we used the Keras Tokenizer Python library [1] to convert conversations to vectors of the tokens. The CNN model architecture is a modified version of the architecture by Collobert et al. [23], including a convolutional layer with multiple filter widths and feature maps, a max-over-time pooling operation [23], and a fully connected layer with dropout and softmax output. To train and assess this classifier, we used the labels specified by our participants.

4.2.3 Images. Instagram is predominantly an image- and video-sharing platform, which is why focusing on images is crucial. Images are an interesting way to attract people to a conversation and research has shown that youth rely heavily on media to converse [2, 72, 108]. However, some nefarious individuals use social media to spread offensive images and symbols to certain groups of people [28, 49, 56]. As it is relevant how predators can send unsafe/triggering images [62, 109], we next elaborate on how we extracted meaningful information from them. We first divided the images into two sets, photos and screenshots, based on the research findings of Ali et al. [2]. Ali et al. observed that participants often shared screenshots of other conversations with their friends, including screenshots of unsafe conversations that made them uncomfortable. Photos and screenshots provide different information, taking screenshots is a common way of capturing screen content to share it with others or save it for later, for example a conversation, an error on a website, or a meme [99], whereas, a photograph is personal and can be explained using a caption telling who or what is in the image.

We first calculated the number of screenshots shared in a conversation and used this as a feature for the model. Then we employed an Optical Character Recognition (OCR) tool Pytesseract [110] to extract the text from screenshots. It helps to identify and "read" the text embedded in images. For the photos, we used the MSCOCO deep learning model [123], which, given an image, generates a textual caption. The key idea was to train a network that recognizes elements that appear in the images and then generates an adequate caption. MSCOCO has been built using an extensive dataset with over 300,000 images. The output of the system is a meaningful caption of the image given as input. For each image in the conversation, we generated the caption that best represents it.

We extracted the words that are most representative of the images and the screenshots shared in both safe and unsafe conversations by applying Term Frequency/Inverse Document Frequency (TF-IDF) on the captions and OCR text. These captions were extracted to act as a feature of the images shared in safe and unsafe conversations. TF-IDF is a product of two metrics, namely Term Frequency (TF) and Inverse Document Frequency (IDF). We also included another image feature, which was the number of screenshots in the conversation. Our current implementation of the image feature set supports Decision Tree, Random Forest, and Linear Support Vector Machine (SVM).

We used these three independent feature sets (i.e., metadata, linguistic cues, and image features) to estimate the likelihood of a conversation being unsafe for youth. These are built to operate independently, possibly when a new conversation takes place. Each classifier is designed to model traits from the three aspects of the conversation. Available decisions of all models are later combined to provide one unified output.

4.2.4 Ensemble Classifier. A high-level description of our detection system is presented in Figure 2. We trained our dataset using a set of prediction models $M = (M_1, M_2, M_3)$ that output a probability $M_i(\sigma_1, ..., \sigma_n) = P_i[c = unsafe|(\delta_1, ..., \delta_n)]$ for each conversation c being annotated as unsafe given a feature vector $(\sigma_1, ..., \sigma_n)$ obtained from different elements i of the conversation. We refer to these models as *individual classifiers*. Then, we created various ensemble models that combine all predictions in M, where each of the model M_i is weighted by w_i based on the performance obtained. The final classifier outputs a decision based on its voting algorithm.

[1] keras - https://keras.io/api/preprocessing/text/

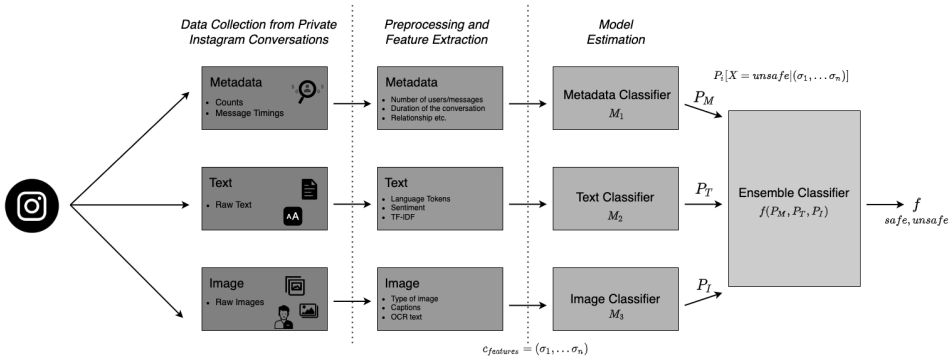

Fig. 2. Architecture of our unsafe conversation detection system.

The goal of the ensemble method is to combine the predictions of the base models. Each classifier individually models the likelihood that a conversation be labeled as unsafe by youth based on its set of features. The ensemble takes all of these judgments (based on metadata, text, and images as described previously) into account to generate a more accurate prediction. This enables for more reliable predictions by having at least one model that is capable of making an informed decision.

Our classifier is based on a stacking ensemble design, which has been shown to outperform individual classifiers [34]. The premise of our stacking design is to use predictions of machine learning models from the previous level as input variables for models on the next level. Our ensemble approach is set up to take a weighted vote from all available predictions. We estimate a function f that takes each of the individual probabilities as input and outputs the aggregated prediction. Formally, $f(P_M, P_T, P_I) = safe, unsafe$.

Majority-Vote Ensemble In this system we gave equal weight to all w_i (i.e. $w_i = 1$) creating an equal vote among the base models. We refer to this non-weighted voting system as *majority-vote*.

Average-Vote Ensemble In the *average-prediction system* we combined each individual prediction using the arithmetic mean of the output probabilities. Note that the parameters in both majority-vote and average-prediction are fixed and do not require calibration. Thus, the validation set is not used in these two modes.

Weighted-Vote Ensemble We define are weighted model as follows:

$$f(M) = \sum w_i \cdot M_i \tag{1}$$

This combines all predictions in M, where each of the model M_i is weighted by w_i based on the accuracy obtained on a validation set. This set is different from the training set, which is used to build the individual probabilities and the testing set, which is used to evaluate the models. To ease presentation, we refer to the model presented in (1) as weighted-vote. This model can be simplified by giving equal weight to all w_i (for example, $w_i = 1$) and obtaining a nominal value for P_i before voting. The process of weighting the individual predictors also serves as a way to calibrate the output of the probabilities. The final classifier then provides a decision-based output based on its voting algorithm.

4.3 Multi-Class Classification: Detecting the Major Risk Type of a Conversation (RQ2)

After determining whether a conversation is safe or unsafe, we move on to specified risk type(s) in the conversation based on the annotations provided by our participants. The participants annotated the risk in unsafe conversations based on individual messages using the following categories:

- **Nudity/porn:** Images of naked or partly naked people.
- **Sexual messages or Solicitations:** Sending or receiving a sexual communication ("Sexting").
- **Harassment:** Messages containing credible threats, attempting to humiliate or shame someone, including personal information in order to blackmail or harass someone, or threatening to post nude images of someone.
- **Hate speech:** Messages that incite violence or target individuals based on their ethnicity.
- **Violence/Threat of violence:** Messages, images, or films depicting severe violence, or encouraging violence, or attacking someone based on their religious, ethnic, or sexual heritage.
- **Sale or promotion of illegal activities:** Messages encouraging the use or distribution of illegal substances such as narcotics.
- **Self-injury:** Messages supporting or promoting self-injury, such as suicidal ideation, cutting, and/or eating disorders, are examples of self-injury.
- **Other:** Other situations that could potentially lead to harm as deemed appropriate by the participant.

For this study, we picked the top five most frequently occurring risks in our dataset, i.e., 'harassment', 'sexual messages/solicitation', 'nudity/porn', 'hate speech', and 'sale or promotion of illegal activities' labeled by the participants. Then we used the same three feature sets i.e., metadata, linguistic cues, and image features that we used for binary classification, to categorize these most prominent risks in the conversation using multi-class detection for each of the conversations. We also used the probabilities obtained from these models to design our ensemble classifier for risk-type detection.

5 RESULTS AND EVALUATION

In this section, we first present the results of the binary-classifiers that predict whether or not a conversation is safe, with a particular focus on whether metadata features can be used in isolation to detect risk (RQ1). Next, we present the results of risk-type classifiers that predict the presence of different risks (i.e., multi-class classifiers) (RQ2). An analysis of the top features that contributed to each classifier's best accuracy performance (RQ2) is also presented.

5.1 Data Pre-processing

We had a total of 17,786 conversations labeled by the participants as safe or unsafe. Out of these 15,397 were marked as safe conversations whereas 2,389 conversations were unsafe. This distribution is in line with risky interactions' occurrence in the real world, where we expect the vast majority of conversations to be safe [2, 65]. Training risk detection classifiers on this unbalanced dataset, however, incurs the risk of creating models that predict the class that occurs more frequently in the data set as compared to the other minority class [7], favoring false negatives (i.e., unsafe interactions classified as safe) over false positives (i.e., safe interactions classified as unsafe). Since in this paper we want to understand the most representative features that distinguish risky from non-risky conversations, this imbalance is not ideal. Moreover, our goal in this formative research is to establish the technical feasibility of building a robust machine learning based detection model, for which a balanced dataset is advisable. For these reasons, we choose to extract a balanced dataset for our experiments. To create a balance between the two classes we randomly selected 2,389 safe conversations, and we used undersampling using random selection of safe conversations as resampling technique for our experiments [97]. For completeness, we also performed experiments on the unbalanced dataset in Section 5.2, and discuss the implications of applying our results in the wild, where safe conversations greatly outnumber safe ones, in Section 7.4.

Train, Test, and Validate Splits. We split our datasets into three sets: two for training and tuning parameters of the ensemble (training and validation) and one for testing, and reporting performance metrics on the latter. The total number of conversations in each split is proportionally sampled depending on the less populated class, assigning splits of 60%, 20%, and 20% to the training, validation, and test sets. This procedure is repeated ten times and the results are averaged over the ten different rounds. Table 3 summarizes all settings in our experiments, along with the number of samples used for binary classification.

Table 3. The number of samples used in our experiments for binary risk classification. The sets are balanced as there are the same amount of samples per class (safe samples + unsafe samples).

	Model	Training	Validation	Test
M1	Metadata	2 x 1,433 = 2,866	2 x 478 = 956	2 x 478 = 956
M2	Text	2 x 1,433 = 2,866	2 x 478 = 956	2 x 478 = 956
M3	Image	2 x 1,433 = 2,866	2 x 478 = 956	2 x 478 = 956

We followed the same process for multi-class risk-type detection as well. Each message in the unsafe conversations had been labeled by the participants for *risk-type*. The risk category *Sale or promotion of illegal activities* had the least number of conversations i.e., 125, and therefore we chose 125 random conversations in the rest of the categories as well, summary of our data division for risk-type samples is displayed in 4.

Table 4. Number of samples used in our experiments for multi-class risk-type detection. The sets are balanced as there is the same amount of samples per each class (harassment samples + sexual messages/solicitation samples + nudity/porn samples + hate speech samples + Sale or promotion of illegal activities samples).

	Model	Training	Validation	Test
M1	Metadata	5 x 75 = 375	5 x 25 = 125	5 x 25 = 125
M2	Text	5 x 75 = 375	5 x 25 = 125	5 x 25 = 125
M3	Image	5 x 75 = 375	5 x 25 = 125	5 x 25 = 125

We used the same metadata, text, and image classifiers and finally the ensemble classifiers to perform risk type detection on the conversations as well.

Evaluation Metrics. To evaluate our models, we employed average model accuracy, standard deviation of model accuracy, F1-measure, area under the curve (AUC), as well as class-specific precision and recall. While the accuracy and F1 values offer a broad overview of the models' performance, the precision and recall scores for each class provide more specific information.

5.2 RQ1: Identifying Unsafe Conversations Using Binary Classification

We implemented and evaluated multiple binary-classifiers for detecting unsafe conversations using three different feature sets (i.e., metadata, linguistic cues, and image features). In the following sections, we explain the results of each of the feature sets in detail.

5.2.1 Metadata Classifier: We investigated different models for creating the metadata classifier as explained in Section 4 and have summarized the performance metrics using the metadata features in Table 5. Overall, we found that the Random Forest model outperformed other traditional classifiers with an AUC = 0.81 and accuracy = 0.83. To analyze the effectiveness of each feature as an indicator for whether a conversation is safe or unsafe, we calculated the importance score of each feature.

Feature Selection: We performed a feature selection study to determine the most important features for classification using the metadata classifier. Feature Importance is a score assigned to each feature of a model to determine how much it contributes to the model's prediction [101]. We examine feature's importance using Gini importance [102]. The higher the value of importance the more significant the feature is for prediction. We found that "Total Response Time" was the biggest predictor with a score of 0.53, and the next best predictor was the relationship the participant had with the other person(s) (0.11).

Table 5. Classification scores for the task of predicting whether a conversation is safe or unsafe using metadata features.

Model	Class	Prec.	Rec.	F1	AUC	Accr.
Decision Tree	Safe	0.79±0.55	0.66±0.14	0.72±0.39	0.81±0.07	0.81±0.10
	Unsafe	0.73±0.67	0.84±0.16	0.78±0.42		
Random Forest	Safe	0.85±0.14	0.78±0.12	0.81±0.13	**0.81±0.05**	**0.83±0.04**
	Unsafe	0.81±0.13	0.87±0.03	0.84±0.02		
Linear SVM	Safe	0.75±0.74	0.70±0.03	0.72±0.64	0.80±0.15	0.83±0.12
	Unsafe	0.76±0.70	0.81±0.04	0.78±0.68		

5.2.2 Text Classifier: We used a Convolutional Neural Network (CNN) with 4,778 conversations in total (safe + unsafe) as explained in section 4. Table 6 summarizes the classification scores for the CNN model predicting whether a conversation is safe or unsafe using textual features. Our text classifier gives an accuracy of 0.82 with an AUC of 0.77.

Table 6. Classification scores for the task of predicting whether a conversation is safe or unsafe using text features.

Classifier	Class	Prec.	Rec.	F1	AUC	Accr.
CNN	Safe	0.83±0.61	0.78±0.14	0.81±0.42	**0.77±0.12**	**0.82±0.05**
	Unsafe	0.80±0.43	0.84±0.13	0.82±0.22		

5.2.3 Image Classifiers: We built a prediction model based on the features extracted from the captions of the images such that $P_T(X = unsafe)$. The architecture of our system is highly flexible and accepts a wide range of classifiers as explained in section 4. Table 7 summarizes the performance metrics of the different machine learning models. Overall, we found that the Linear SVM model outperformed other traditional classifiers with an AUC = 0.65 and accuracy = 0.70. SVM performs well when the number of samples is substantially more than the number of features. Furthermore, it is less susceptible to data outliers—rather than decreasing the local error, SVM aims to lower the upper bound on the generalization error [113]. We also found that "Number of Screenshots" shared in a conversation was the biggest predictor with a score of 0.65.

Summary of Findings. Looking at these results, we concluded that the metadata classifier works best in detecting unsafe conversations. We found that the metadata classifier detects most of the unsafe conversations accurately (87%), however it misclassified conversations that are longer and whose relationship with the participant is not stranger. On the other hand, for such conversations, the text classifier works efficiently. For example, the use of explicit or abusive words is a sign of riskiness in a conversation that is detected efficiently by the text classifier. These results show

Table 7. Classification scores for the task of predicting whether a conversation is safe or unsafe using image features.

Classifier	Class	Prec.	Rec.	F1	AUC	Accr.
Decision Tree	Safe	0.59±0.74	0.56±0.47	0.52±0.62	0.54±0.17	0.54±0.18
	Unsafe	0.53±0.75	0.54±0.47	0.58±0.64		
Random Forest	Safe	0.67±0.49	0.57±0.32	0.62±0.40	0.54±0.19	0.59±0.20
	Unsafe	0.52±0.54	0.62±0.37	0.67±0.49		
Linear SVM	Safe	0.65±0.01	0.83±0.04	0.73±0.03	**0.65±0.04**	**0.70±0.12**
	Unsafe	0.78±0.12	0.57±0.14	0.66±0.15		

that metadata features are the most important information when it comes to detecting risk, and a classifier based upon them can achieve high accuracy. This has important implications in the way in which social media platforms should approach designing risk detection systems that protect user privacy with end-to-end encryption.

Performance Evaluation Using Unbalanced Dataset. While the goal of this paper is to identify important features in distinguishing between safe and unsafe conversations, it is nonetheless important to understand how this real-world imbalance could affect the performance of our system. For this purpose, we used the total 17,786 conversations labeled by the participants containing 15,397 safe conversations and 2,389 unsafe conversations. We trained our models using 60% (10,671 conversations) of the total conversations and tested using 20% of the total conversations (3,557 conversations). The rest of the conversations were used for the validation set. Our results showed that in this setting the accuracy of all models increased, for example the accuracy of the metadata classifier was 0.95 ± 0.03, text classifier was 0.94 ± 0.12, while the accuracy of the image classifier was 0.80 ± 0.17. The F1 score for the metadata classifier in the safe category was 0.73 ± 0.03 and was 0.66 ± 0.12 for the unsafe category. Similarly, the F1 score for the text classifier in the safe category was 0.93 ± 0.01 and was 0.46 ± 0.04 in the unsafe category. Finally, the F1 score for the image classifier in the safe category was 0.83 ± 0.12 and in the unsafe category was 0.57 ± 0.15. These results show that the F1 scores decrease while using an unbalanced dataset, and the number of false positives and false negatives increase. While analyzing the results manually, we found that the model adapted to the imbalanced case by favoring the majority class (i.e., safe conversations) over the minority one (i.e., unsafe conversations). This is expected in an imbalanced case like this one [7], and it is something that online platform operators like Instagram should consider when deploying similar risk detection models. In Section 7.4 we discuss how this imbalance affects the deployment of similar risk detection systems in the wild.

In Section 7.3 we discuss these design implications in detail. In the next section we show how an ensemble classifier can build upon the different results of the individual classifiers.

5.2.4 Ensemble Classifier. Table 8 summarizes the classification scores of the different ensemble classifiers we designed. The best ensemble classifier was the *Weighted-Vote Classifier* with an AUC of 0.83 and Accuracy of 0.85. We show an increase in performance by using the Weighted-Vote Classifier than using the individual classifiers separately. This indicates that putting together the three feature sets gives better results than looking at a single one.

Table 8. Classification scores for the task of predicting whether a conversation is safe or unsafe using different ensemble classifiers.

	Classifier	Class	Prec.	Rec.	F1	AUC	Accr.
E1	Majority-Vote	Safe	0.88±0.14	0.75±0.14	0.81±0.14	0.81±0.13	0.83±0.04
		Unsafe	0.79±0.24	0.90±0.33	0.84±0.28		
E2	Average-Vote	Safe	0.80±0.04	0.79±0.04	0.79±0.04	0.80±0.03	0.80±0.03
		Unsafe	0.80±0.25	0.81±0.34	0.81±0.29		
E3	**Weighted-Vote**	Safe	0.89±0.06	0.80±0.06	0.74±0.06	**0.83±0.06**	**0.85±0.03**
		Unsafe	0.82±0.08	0.90±0.07	0.86±0.07		

5.3 RQ2: Detecting Risk Types in Unsafe Conversations Using Multi-class Classification

Our next step was to check whether we can accurately distinguish the different risk types within unsafe conversations. We aim to understand what kind of indicators work well in detecting different types of risk, and where independent classifiers (i.e., metadata alone) fall short. For this purpose we again divided the conversations into 3 feature sets i.e., metadata, linguistic cues and image features. Table 9 summarizes the results of the classification scores for the task of predicting the types of risk that may be in the conversation. Similar to the binary case, we also test our models on the unbalanced dataset, finding that the accuracy of all the classifiers increases, but with significant increase in the number of false positives for each category.

5.3.1 Risk Type Analysis. Next, we manually examine conversations in each risk category to understand which feature set was able to detect what kind of risk more accurately. We identify the different characteristics of each conversation risk type, and our results are as follows:

Harassment: We found that for this risk type the metadata classifier performed better. Most of the conversations containing messages that were labeled as harassment were shorter and participants were less likely to respond to these conversational attacks. Also, almost 70% of these conversations were with strangers.

Sexual messages/solicitation: For this category the text classifier performed better. For these conversations, most of the textual messages contained words that were sexually explicit and focused on asking sexual favors, for example, this message was sent to a 17-year-old male, by another male asking for sexual favors.:

Other: *'Daddy I need your cummies, mine are all in the bank'*

In conversations containing sexual messages/solicitation, the participant engaged occasionally and later stopped engaging. Also, the number of messages sent by the other person was much more than the number of messages sent by the participant themselves, for example, the below message was sent to a 16-year-old female, by a stranger asking for sexual favors. When she refused, they asked/called the participant gay, and the participant then stopped engaging with the said person.

Other: *'I'll fuck the dog shit outa u'*

Nudity/porn: For this category, the image and metadata classifiers performed best. On manual inspection, we found that a number of suspicious links were shared in these conversations, mostly pornographic websites. Also, the participants themselves hardly participated in such conversations (very few messages by the participant themselves). For example, in one of the group conversations, a "dick pic" was shared, while the participant did not send any message.

Hate speech: For hate speech, we found that the text and metadata classifier worked best. Approximately 60% of these conversations were mostly with acquaintances and friends (i.e., people the

Table 9. Classification scores for the task of predicting the risk types of unsafe conversations.

	Model	Risk Types	Prec.	Rec.	F1	AUC	Accr.
M1	Metadata	harassment	0.85±0.34	0.87±0.12	0.86±0.23	0.71±0.25	0.72±0.23
		sexual messages	0.63±0.34	0.67±0.11	0.65±0.23		
		nudity/porn	0.73±0.33	0.74±0.13	0.73±0.23		
		hate speech	0.65±0.33	0.65±0.14	0.65±0.24		
		sale of illegal activities	0.72±0.32	0.65±0.13	0.68±0.23		
M2	Text	harassment	0.68±0.03	0.65±0.04	0.66±0.04	0.77±0.05	0.79±0.03
		sexual messages	0.86±0.01	0.89±0.01	0.87±0.01		
		nudity/porn	0.77±0.04	0.78±0.02	0.77±0.03		
		hate speech	0.79±0.05	0.79±0.04	0.79±0.05		
		sale of illegal activities	0.85±0.01	0.84±0.01	0.84±0.01		
M3	Image	harassment	0.53±0.34	0.57±0.10	0.55±0.22	0.64±0.29	0.66±0.24
		sexual messages	0.67±0.33	0.68±0.14	0.67±0.24		
		nudity/porn	0.88±0.36	0.72±0.12	0.79±0.24		
		hate speech	0.69±0.35	0.76±0.13	0.72±0.24		
		sale of illegal activities	0.53±0.34	0.57±0.19	0.55±0.27		
E1	Majority Vote	harassment	0.65±0.24	0.63±0.24	0.64±0.24	0.71±0.34	0.72±0.25
		sexual messages	0.79±0.21	0.79±0.21	0.79±0.21		
		nudity/porn	0.75±0.26	0.77±0.31	0.76±0.27		
		hate speech	0.74±0.29	0.75±0.43	0.74±0.31		
		sale of illegal activities	0.65±0.26	0.64±0.25	0.64±0.26		
E2	Average Vote	harassment	0.65±0.11	0.63±0.15	0.64±0.13	0.71±0.23	0.70±0.16
		sexual messages	0.69±0.22	0.65±0.12	0.67±0.17		
		nudity/porn	0.75±0.22	0.78±0.22	0.76±0.22		
		hate speech	0.71±0.14	0.69±0.10	0.70±0.12		
		sale of illegal activities	0.69±0.13	0.74±0.19	0.71±0.16		
E3	**Weighted Vote**	harassment	0.79±0.02	0.79±0.04	0.79±0.03	**0.80±0.13**	**0.82±0.05**
		sexual messages	0.85±0.02	0.88±0.02	0.86±0.02		
		nudity/porn	0.82±0.04	0.78±0.03	0.80±0.04		
		hate speech	0.78±0.05	0.82±0.04	0.80±0.05		
		sale of illegal activities	0.85±0.11	0.82±0.08	0.83±0.10		

participants knew- online or offline). For example, the conversation snippet below was part of a group conversation where a participant (aged 17, female) tried to convey their message logically, but was later slut shamed due to an unfortunate typo.

> ***Participant:*** *'Okay first of all women are not combat ready. That's why there's no woman soldiers that are currently fighting. They aren't physically capable of fighting men. They can't compete in male sports hoe do u expect them to fight in war?'*
> ***Other:*** *'Since you don't seem to understand that your a hoe I guess I'll just call you a thot'*

In these types of conversations, the participant tried to correct/put forward their opinion but later ended the conversation, when they realized that they were being bullied. For example, in this conversation the participant (aged 16, female) tried to correct the other person but later stopped replying due to their harsh response:

> ***Other:*** *'Rot in hell'*
> ***Participant:*** *'Stop. She apologized and so did I so you need to drop it.'*
> ***Other:*** *'You trick as bitch'*

Sale or promotion of illegal activities For sales and promotion of illegal activities we found that the text classifier worked better. For most of these conversations, the participants pointed out that the messages were from someone whose account was hacked. The participants reported concern that these types of messages made them extremely uncomfortable, for example, a participant (aged 17, male) was sent the following message:

> ***Other:*** *'[participant name] OMG your actually on here, [removed username] , your number 15! its really messed up'*

In most of these conversations monetary advantage in the form of money or gift card was promised to the participant, for example:

> ***Other:*** *'[participant name] you won't believe this! I just received a $1,000 GiftCardVisa yesterday and all I did was participate a little and they actually sent me it. I used it to get a new phone I have been eyeing LOL!! I thought u would want one also since your my follower. Hurry tho theres only a few of them left! Click'*

A lot of these messages were trying to give rewards to participants for following them or taking other actions such as clicking on links. for example:

> ***Other:*** *'Hi! Here is your code ####-@@@@-**** I'll send the last four digits but first follow EVERYONE I follow! Once you did message me back what type of giftcard you'd want! Also let me know the value on the giftcard :)'*
>
> ***Other:*** *'Congratulations!!! Taylor Swift Has Chosen You To Be Among Our Latest Winners, You Have Won $10,000.00 In the Covid19 Assistance Fans Promo. Your Name Has Been Shortlisted, You Are Truly Lucky And Thank You For Being A Fan.'*

Overall, we presented the results of the individual metadata, text, and image classifiers and the ensemble approach of the combination of those classifiers. The weighted-Vote performed the best for classifying safe/unsafe conversations and also the risk type.

6 ERROR ANALYSIS OF MISCLASSIFICATIONS IN RQ1

In this section, we examine individual prediction cases to gain a better understanding of the factors that lead to misclassifications by our individual binary classifiers and those that could only be detected using the Ensemble Classifier. We looked more closely at each model and qualitatively examine the characteristics of the unsafe conversations that were correctly detected by the model specifically but not by others. We discuss our qualitative observations below.

6.1 Correctly Identifying Conversations Using Individual Feature Sets

Each feature set provides us with different information about the conversation, for example, metadata features provide details regarding participants' interest in the conversations and who they are conversing with. Linguistic cues, and image features provide details about what the conversation is about. In the following sections, we provide examples of conversations where one feature set triumphed over the other and instances where the others failed.

6.1.1 Correctly Classified by Metadata, Misclassified by Linguistic Cues, and Image Features. We determined that the metadata classifier was able to detect 87% of all unsafe conversations. Upon manual analysis of the unsafe conversations detected by the metadata classifier, we saw that they were shorter and that the participant stopped engaging with other(s). Below, we share two examples of conversation snippets correctly detected by the metadata classifier but misclassified by the text and image classifiers:

> ***Other:*** *'Where are you staying at?'*
>
> ...

Other: *'Oh fr that's what's up. I'm here till Saturday you should come meet me at the beach one day or somethin'*
Participant: *'Sweet but I have practice tomorrow and Saturday aha are you just here for vacay.'*
...
Other: *'You have a snapchat?'*

In this conversation, a 19 year old female was harassed by a stranger who kept asking for her Snapchat ID and her location. The participant engaged very less in the beginning and then stopped engaging abruptly. The conversation was short (total 12 messages), and was initiated by the other person. Furthermore, no media files were shared in this conversation. The text classifier also failed to detect this classifier because the language used was not predatory.

Other: *'A contact named the group My hot photos!'*
Other: *[Instagram User sent an attachment.]*

This was a short group conversation in which the participant (aged 19, gender female) was randomly added by a stranger. The participant did not engage in this conversation, and all messages were sent by other people in the group. In this conversation, one link possibly to a pornography website was also shared, however, no image files were shared in the said conversation.

Most of the conversations detected using the metadata feature set had the unique characteristic of being short (avg 20 messages) and 37% of these conversations were group conversations with 76% of these conversations being with strangers.

6.1.2 Correctly Classified by Linguistic Features, Misclassified by Metadata and Image Features. Our text classifier was particularly helpful in detecting conversations that are longer, and contain highly sexual and offensive language. Below we share two examples of conversations correctly detected using linguistic cues, but were not detected using the metadata and image feature sets.

Other 1: *'you're invalid'*
Other 1: *'i said validate me you queer'*
Other 2: *'finna sass the fuck out of [name]'*
Participant: *'Would you please stop it!!!'*
Other 1: *'your dick get broke'*

This was a group conversation where the LGBTQIA+ community was targeted. Here the participant (age = 17, gender = male), actively participated, however, tried to call out the toxic behavior of his peers. The relationship as stated by the participant himself was friends. Interestingly, some of the accounts involved in this conversation were marked as deleted. The conversation was demeaning towards women and the LGBTQIA+ community, using sexual and explicit language.

Other: *'Hey I have a foot fetish and I was wondering if you could send me pictures of your feet'*
Other: *'Are you there'*
Other: *'I don't have any money right now is there anything I can do to get feet pictures for free'*
Participant: *'Just google feet pics'*
Other: *'I would but your way hotter then the girls'*

In this conversation, the participant aged 20, female, was asked for private pictures, i.e., "feet pics" in exchange for money that they would send at a later date. The participant denied to give any such favors, but the other person was persistent. Interestingly, no media files were exchanged in this conversation but the participant also did not quit engaging with the other person.

As can be seen from these examples, conversations that are longer, with more textual context are detected by the text classifier, however, since these conversations do not contain images and are usually with a person(s) whom the participant knows, therefore they were misclassified by the metadata and image classifiers.

6.1.3 Correctly Classified by Image Features, Misclassified by Metadata and Linguistic Cues. Most of the conversations that were correctly classified by image features and misclassified by metadata and text classifier contained suggestive images with females in them. We share an example of such a conversation:

Other: *[Shared picture]*
Other: *'Because I got through your picture and see something in you'*
Other: *'You are beautiful like the angels in heaven'*
Participant: *'You have stop this okay'*
Other: *'Am falling in love with you'*
Participant: *'i have a boyfriend'*
Other: *'Is he going to Mary you'*
Other: *'send me his picture'*
...

In this conversation, the participant aged 17, gender female was harassed by the other person, by showing her personal picture containing a clothed female showing some cleavage and asking for sexual favors. She replied that she had a boyfriend and he needed to stop messaging her. The participant also mentioned that she was a minor and later went on to block the other person.

The image feature set was helpful in identifying unsafe conversations that shared images of people, for example, selfies of females. The conversations had varied lengths.

6.2 Correctly Identifying Conversations Using Weighted Ensemble Classifier

Some of the conversations were only detected using the Weighted Ensemble Classifier, using the probabilities of all three classifiers (metadata, linguistic cues, and image features). For example, the following conversations were detected only using the weighted ensemble model:

Other: *'All lives matter'*
Participant: *'all lives can't matter until black lives matter'*
Other: *'So your saying my life doesn't matter'*
Participant: *'are you white ?'*
Other: *'Yeah'*
Participant: *'exactly . your life already matter s'*

In this conversation, the participant aged 15, female was trying to convince the other person, of the importance of the "Black Lives Matter"(BLM) movement, whereas the other person was trying to argue back that all lives matter and BLM implies that their life doesn't matter since they are white. The other person left the conversation themselves.

Other: *'Please watch that. I hope you get help for your gender dysphoria'*
Other: *[Send Video]*

In this conversation, the participant (aged 14, gender unspecified) was being harassed due to being transgender, by a stranger. The other person sent her a video link asking her to watch it to "fix her gender dysphoria". The participant did not engage in this conversation and blocked them.

Our Weighted Ensemble Classifier helped detect most of the unsafe conversations, especially those in which 'risk' is not visible directly, but were labeled by the participant as unsafe. Since it

uses the probabilities from all three feature sets (metadata, linguistic cues, and image features) it was able to identify all the nuances of unsafe conversations.

7 DISCUSSION

In this paper, we presented a systematic approach to the detection of unsafe conversations on Instagram encountered by youth. With the impending push toward end-to-end encryption on Meta platforms, we make the important finding that meta-level data inferred from conversational patterns can be used to detect risks in conversation-level data. This is an important implication for risk detection and the accountability of social media platforms to protect users, even if end-to-end encryption is implemented. However, if these platforms want to take the extra step of mitigating online risks, they may be limited in being able to respond to specific risk types appropriately without contextual information about the type of risk users are encountering. To dig deeper into the specific risk of the conversation we needed to look at other features such as linguistic cues and image features. Therefore, this may be a reason for Meta and other social media companies to reconsider the push towards end-to-end encryption as it limits their ability to protect users, particularly young people, from harms that are committed on their platforms. Below, we discuss the human-centered insights and limitations for deploying our system into the real world.

7.1 Meta-level Data as a Key Risk Indicator (RQ1)

Our analysis highlighted that risky conversations, regardless of risk type, presented similar high level characteristics. Compared to safe conversations, risky conversations are shorter and more one sided, as victims tend to disengage early on in the conversation. This is in line with Ali et al.'s finding when analyzing meta data and media files on Instagram [2], as well as with research that studied conversational patterns and interactions in online fraud [44] and spam [48]. However, our results go beyond these prior studies to show that meta-level data is a *better* predictor of conversational-level risk than contextual data contained within the messages themselves. Our metadata classifier used features that were directly available from the conversation, except for the relationship the participant had with the person they were conversing with, which may be inferred from topography of the social network (e.g., reciprocal connection, unidirectional following, or no connection). Based on our feature selection study, metadata features such as engagement of the participants and the response time of participants are key features that can be used to distinguish between conversations that made the users uncomfortable or otherwise. Other features, like the history of the other people involved or whether they were suspended previously for malicious activity, can be useful. However, this information is accessible only to the platform. For our study, we used data directly annotated by the participants; however, companies like Meta could extract similar features from profile-level characteristics of users, as well as through social network analysis.

Overall, our findings open up interesting opportunities for future research and implications for the industry as a whole. First, performing risk detection based on metadata features alone allows for lightweight detection methods that do not require the expensive computation involved in analyzing text and images. Second, developing systems that do not analyze content eases some of the privacy and ethical issues that arise in this space [12, 69], ensuring user protection. This can also potentially enable the creation of privacy-preserving datasets to be used to train risk detection systems, where the data shared only contains metadata (e.g., anonymous IDs of who sent a message, timestamps, etc.). Finally, with the switch of online platforms like Meta to end-to-end encryption, analyzing metadata is a promising way in which social media platforms could detect risk without intruding on their users' privacy.

7.2 Additional Context is Needed for Determining Risk Type and Designing Appropriate Intervention Strategies (RQ2)

While metadata helps distinguishing risky from safe conversations, our findings indicate that these high level features are not nuanced enough to distinguish between different risk types. To detect the specific type of risk involved in a conversation, the information provided by linguistic cues and image data is paramount. In particular, we found that some kinds of risk, such as sexual messages/solicitation and sales or promotion of illegal activities are easier to detect using textual features, while risks such as nudity/porn can be better detected using image features. A notable exception is harassment, where the repetitive and one-sided nature of communication was consistently enough to distinguish it from other types of risk, and this can be effectively detected through metadata features. These findings have several implications for future research. First, our work highlights the importance of existing NLP and computer vision work in detecting online risk [78, 94, 100]. While these techniques are more heavyweight than metadata ones, they are necessary in performing fine-grained risk detection. An opportunity in this space would be combining the two approaches, using metadata as a first layer filter for all conversations and applying deeper (and more expensive) textual and image analysis only to those conversations that are flagged as suspicious by the metadata filter. This type of multi-level approach has been successfully used in other security and safety contexts, like spam detection [85, 112] This approach would enable more effective defenses, especially for risk types where context is important to make decisions on how or when to intervene.

Secondly, risk mitigation especially for youth requires a user-centered and tailored approach in order for it to be successful [1, 4, 131]. Risk prevention programs, in general, are more effective when tailored to context [121], which is why we use contextual information. For example, the relationship a participant had with the other person is an important context to determine whether a conversation is risky. Metadata can provide high-level cues about conversations that are unsafe for youth; however, the detection and response to the specific type of risk require the use of linguistic cues and image data. This finding raises important philosophical and ethical questions in light of Meta's recent push towards end-to-end encryption as such contextual cues would be useful for well-designed risk mitigation systems that leverage AI. We discuss these implications in more detail in the next section.

7.3 Implications for the Design of AI Risk Detection Systems

The cybersecurity and safety communities have highlighted the inherent trade-off between privacy and security, which has been studied and discussed for years [132]. Many online platforms are pushing towards adopting end-to-end encryption for their messaging services: applications such as Telegram, Meest, Signal, Line, and WhatsApp have end-to-end encryption enabled and implemented, allowing users to communicate more securely [55]. The purported benefit of end-to-end encryption is increased user privacy. With end-to-end encryption, messages are encrypted by the sender and decrypted by the receiver; and therefore, the platform is unable to see what content is being exchanged. Since end-to-end encryption allows only the sender and receiver to view the message's contents it is considered to be one of the most effective ways to ensure security of the content [126]. While this practice may dramatically improve user privacy [22], it also limits what platforms can do with respect to detecting risky content exchanged in conversations [14]. For instance, textual and image features cannot be leveraged for the detection of risk in unsafe conversations, as we have demonstrated is possible in this paper. It is in this way that enhanced end user privacy through end-to-end encryption could act as a double-edged sword that can hide abuse, protect malicious users, and create safety issues for more vulnerable users [67, 117, 120]. Thus, balancing end user

privacy and safety becomes a complex problem, especially when dealing with vulnerable users, such as minors [125].

Meta, the parent company of Instagram, is also moving towards building and implementing end-to-end encryption across Messenger and Instagram DMs [25]. In fact, Meta has been involved in an on-going political and legal struggle over the use of end-to-end encryption on its services, with a special emphasis on child protection. The UK Home Secretary Priti Patel has organized an international effort to persuade Meta to abandon plans to integrate its messaging applications and encrypt all user conversations, claiming that such plans make it more difficult to protect minors [45]. The main criticism from parties interested in employing encryption to keep messages secure is that any weakening of encryption for the advantage of third-parties also aids those with more sinister purposes. For instance, they believe that allowing law enforcement access through a "backdoor" creates a vulnerability in security systems that criminals can exploit [77]. The unintended consequence providing a built-in way to access private information can also be used by cybercriminals to exploit children, commit acts of terrorism, or perpetrate human trafficking [111]. Conversely, end-to-end encryption could also serve to protect cybercriminals when committing these heinous acts and prevent critical evidentiary data from being discovered during legal proceedings if they are caught [127].

And, when it comes to the ability for the platform itself to take an active role in youth risk prevention, there remains additional considerations. While many social media platforms and researchers are exploring the use of improved parental controls [47] and the use of age verification [80], using AI risk detection to mitigate some of the harms youth are exposed to in real-time can also be advantageous. However, with end-to-end encryption becoming widespread, researchers and platforms might have to adapt their detection mechanisms to use features that are still available in this setting. Our results highlight that metadata only features, which respect user privacy by not revealing any sensitive information, are in many cases effective in distinguishing safe conversations from unsafe ones on Instagram. These findings can be used as a guideline by platforms when designing AI risk detection systems: based on the level of granularity at which they wish to detect user risk and on the level of privacy that they want to guarantee to their users, platforms can choose to based their detection on high-level features such at the metadata ones or perform more heavyweight and privacy-invasive analysis.

As such, our results lead to important considerations for the design of AI risk detection systems in light of the industry-wide push towards end-to-end encryption. One design recommendation could be to give users (or parents) the option to turn on end-to-end encryption and/or risk detection, so that they can negotiate the trade-offs between privacy [87] and safety themselves. Another option would be to take a "Child Safety by Design" [116] approach of requiring social media companies at a policy-level to include safeguards for minors, which may preclude them from using end-to-end encryption. Alternatively, our work demonstrates the potential for AI systems to develop meta-level data risk detection as a way to create some safeguards for youth while allowing for privacy protection through end-to-end encryption. As such, companies that opt to move toward end-to-end encryption still have viable options for developing AI-driven safeguards for youth within their platforms, rather than absolving themselves of the responsibility for youth protection.

7.4 Limitations & Future Research

The findings of this study highlight the need to improve the overall quality and accuracy of risk detection systems, with significant implications for using various feature sets (for example, metadata, linguistic cues, and image features) for risk prevention and intervention, particularly in the private realm on social media. The dataset we employed for our study is modest sized (N = 172) and considers data from Instagram alone. Furthermore, our data set is complex and exhibits

considerable variations due to its multimodal nature. For this reason, we utilize a multimodal approach to risk detection. Metadata, linguistic cues, and image features represent different sources of information, which can suggest different information. Our method allows each of the modalities to provide individual predictions, without the interference of the other, which are then merged using an ensemble model to get a final prediction. Secondly, as researchers collecting this data, there were certain metadata level characteristics that we could not examine, such as the length of an account, the activity, and their previous history, for example, whether or not they harassed somebody in the past or were reported/suspended. These additional features can be used by companies like Meta to further risk mitigation especially by partnering with researchers these indicators can be explored in more depth for the purpose of protecting users on the platform.

The system we present is a key first step in developing automated tools that are able to assist both social media companies and end users identify conversations that make youth uncomfortable. However, there are risks in deploying such automated systems, principally the risk of labeling safe conversations as unsafe. If we want a more strict classification of unsafe conversations, it may cause some safe conversations to be marked as unsafe, causing users to be frustrated and may end up denying the service to legitimate users. Another potential risk of an incorrect detection is that it reduces the responsiveness of such a tool, causing users to disregard them altogether. Our system is more inclined towards detecting as many unsafe conversations as possible, given that our target audience is youth and young adults. For our best working ensemble classifier, i.e., *Weighted-Vote*, we saw the highest recall for unsafe conversations. It reached the highest performance and was able to identify the most number of unsafe conversations(90%). For our target audience, i.e., young adults and youth we believe it is necessary to detect all risky conversations even if we may need some stricter measures. Therefore, having high recall is more important, because for our case it is okay to classify a safe conversation as unsafe (false positive) and alerting the participant involved, but it is critical not to miss identifying risky conversations (false negative).

To be able to elicit the most representative features that distinguish safe from unsafe conversations, we balanced our dataset. However, when applying our approach in the wild online platform operators would have to deal with an unbalanced dataset where safe conversations far outnumber unsafe ones. Our experiments showed that when training and testing a classifier on an imbalanced dataset the overall accuracy increases, but the fraction of false positives for the safe category also increases, with a decrease in the F1 score. This is not ideal, as generating too many unjustified warnings would make users ignore them. Online platforms should develop strategies to deal with this. For example, they could use our automated system to flag potentially risky conversations and have a human moderator make the final decision. Lastly, there is always a chance of predators or malicious users changing their behaviors so as to bypass the system. From the previous section, we saw that there are some distinct metadata characteristics of risky conversations such as lack of participant engagement, shorter conversations and lesser images being shared in conversations which can be bypassed by malicious users or scammers by deliberately engaging the participant more and sharing images in such conversations. However our system relies on a wide range of features that are more difficult to evade, such as linguistic cues which analyze the actual content of the conversation.

Our analysis provides an important first step to enable automated (machine learning based) detection of online risk behavior going forward. Our system is based on reactive characteristics of the conversation however our research also paves the way for more proactive approaches to risk detection which are likely to be more translatable in the real world given their rich ecological validity. Since the conversations that we studied are actually annotated by the participants themselves, we can extrapolate the key points from this research to proactive risk detection to other datasets on other platforms too that show similar trajectories.

8 CONCLUSION

In this study, we employed several feature sets to detect multimodal risk in private Instagram interactions. We found that features based on metadata information work well when used to develop risk classifiers, and this type of features could be used by platforms to developing risk detection systems when textual and image information becomes unavailable due to the adoption of end-to-end encryption. We then focused on detecting the specific type or risk in conversations, and found that in this case the contextual information provided by text and images is key to develop accurate classifiers. This work can inform future research on developing risk detection systems by both the CSCW community and the platforms themselves.

ACKNOWLEDGMENTS

This research is supported in part by the U.S. National Science Foundation under grants #IIP-1827700, #IIS-1844881, #CNS-1942610 and by the William T. Grant Foundation grant #187941. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the research sponsors. We would also like to thank all the participants who donated their data and contributed towards our research.

REFERENCES

[1] Zainab Agha, Neeraj Chatlani, Afsaneh Razi, and Pamela Wisniewski. 2020. Towards conducting responsible research with teens and parents regarding online risks. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–8.

[2] Shiza Ali, Afsaneh Razi, Seunghyun Kim, Ashwaq Alsoubai, Joshua Gracie, Munmun De Choudhury, Pamela J Wisniewski, and Gianluca Stringhini. 2022. Understanding the Digital Lives of Youth: Analyzing Media Shared within Safe Versus Unsafe Private Conversations on Instagram. (2022), 1–14.

[3] Nasiru Ishola Aliyu, Abdulrahaman Musbau Dogo, Fatimah Olajumoke Ajibade, and Tosho Abdurauf. 2020. Analysis of cyber bullying on facebook using text mining. *Journal of Applied Artificial Intelligence* 1, 1 (2020), 1–12.

[4] Ashwaq Alsoubai, Xavier V. Caddle, Ryan Doherty, Alexandra Taylor Koehler, Estefania Sanchez, Munmun De Choudhury, and Pamela J. Wisniewski. 2022. MOSafely, Is That Sus? A Youth-Centric Online Risk Assessment Dashboard. In *Companion Publication of the 2022 Conference on Computer Supported Cooperative Work and Social Computing* (Virtual Event, Taiwan) *(CSCW'22 Companion)*. Association for Computing Machinery, New York, NY, USA, 197–200. https://doi.org/10.1145/3500868.3559710

[5] Ashwaq Alsoubai, Jihye Song, Afsaneh Razi, Nurun Naher, Munmun De Choudhury, and Pamela J. Wisniewski. 2022. From 'Friends with Benefits' to 'Sextortion:' A Nuanced Investigation of Adolescents' Online Sexual Risk Experiences. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 411 (nov 2022), 32 pages. https://doi.org/10.1145/3555136

[6] Philip Anderson, Zheming Zuo, Longzhi Yang, and Yanpeng Qu. 2019. An Intelligent Online Grooming Detection System Using AI Technologies. (2019), 1–6. https://doi.org/10.1109/FUZZ-IEEE.2019.8858973

[7] Sumaira Ashraf and Toqeer Ahmed. 2020. Machine Learning Shrewd Approach For An Imbalanced Dataset Conversion Samples. *Journal of Engineering and Technology* 11 (2020).

[8] Karla Badillo-Urquiola, Diva Smriti, Brenna McNally, Evan Golub, Elizabeth Bonsignore, and Pamela J Wisniewski. 2019. Stranger danger! social media app features co-designed with children to keep them safe online. (2019), 394–406.

[9] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 2 (2019), 423–443. https://doi.org/10.1109/TPAMI.2018.2798607

[10] Francesco Barbieri, Miguel Ballesteros, Francesco Ronzano, and Horacio Saggion. 2018. Multimodal Emoji Prediction. (2018).

[11] Jessica Baron. 2019. The key to gen Z is video content. *Forbes* (Jul 2019). https://www.forbes.com/sites/jessicabaron/2019/07/03/the-key-to-gen-z-is-video-content/?sh=e92cf1534848

[12] Nadine Barrett-Maitland and Jenice Lynch. 2020. Social media, ethics and the privacy paradox. *Security and privacy from a legal, ethical, and technical perspective* (2020).

[13] Shannon Bond and Bobby Allyn. 2021. Facebook whistleblower tells Congress products hurt kids and weaken democracy NPR. (2021). https://www.npr.org/2021/10/05/1043207218/whistleblower-to-congress-facebook-products-harm-children-and-weaken-democracy

[14] Timothy Buck. 2022. Updates to end-to-end encrypted chats on Messenger. *Meta* (Jan 2022). https://about.fb.com/news/2022/01/updates-to-end-to-end-encrypted-chats-messenger/

[15] Xavier V Caddle, Afsaneh Razi, Seunghyun Kim, Shiza Ali, Temi Popo, Gianluca Stringhini, Munmun De Choudhury, and Pamela J Wisniewski. 2021. MOSafely: Building an Open-Source HCAI Community to Make the Internet a Safer Place for Youth. (2021), 315–318.

[16] Ana Isabel Canhoto and Yuvraj Padmanabhan. 2015. 'We (don't) know how you feel'–a comparative study of automated vs. manual analysis of social media conversations. *Journal of Marketing Management* 31, 9-10 (2015), 1141–1157.

[17] Noé Cecillon, Vincent Labatut, Richard Dufour, and Georges Linarès. 2019. Abusive language detection in online conversations by combining content-and graph-based features. *Frontiers in big Data* 2 (2019), 8.

[18] Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, Athena Vakali, and Nicolas Kourtellis. 2019. Detecting Cyberbullying and Cyberaggression in Social Media. *ACM Transactions on the Web (TWEB)* 13 (2019), 1 – 51.

[19] Vikas S Chavan and Shylaja S S. 2015. Machine learning approach for detection of cyber-aggressive comments by peers on social media network. (2015), 2354–2358. https://doi.org/10.1109/ICACCI.2015.7275970

[20] Ying-Yu Chen and Shukai Hsieh. 2020. An Analysis of Multimodal Document Intent in Instagram Posts. (2020).

[21] Chun-Yueh Chiu, Hsien-Yuan Lane, Jia-Ling Koh, and Arbee L. P. Chen. 2020. Multimodal depression detection on instagram considering time interval of posts. *Journal of Intelligent Information Systems* 56 (2020), 25–47.

[22] Miriam Cihodariu. 2022. Best encrypted messaging apps of 2021 and Why you should use them. *Heimdal Security Blog* (Jun 2022). https://heimdalsecurity.com/blog/the-best-encrypted-messaging-apps/

[23] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of machine learning research* 12, ARTICLE (2011), 2493–2537.

[24] Glen A. Coppersmith, Ryan Leary, Patrick Crutchley, and Alex B. Fine. 2018. Natural Language Processing of Social Media as Screening for Suicide Risk. *Biomedical Informatics Insights* 10 (2018).

[25] Antigone Davis. 2021. Our approach to safer private messaging. *Meta* (Nov 2021). https://about.fb.com/news/2021/12/metas-approach-to-safer-private-messaging/

[26] Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering Shifts to Suicidal Ideation from Mental Health Content in Social Media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 2098–2110. https://doi.org/10.1145/2858036.2858207

[27] Bart Desmet, Kirsten Pauwels, and Veronique Hoste. 2015. Online suicide risk detection using automatic text classification. (2015).

[28] Rebecca A DiBennardo. 2018. Ideal victims and monstrous offenders: How the news media represent sexual predators. *Socius* 4 (2018).

[29] Thomas G Dietterich. 2000. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*. Springer, 1–15.

[30] Rakkrit Duangsoithong and Terry Windeatt. 2009. Relevance and redundancy analysis for ensemble classifiers. (2009), 206–220.

[31] Michele P. Dyson, Lisa Hartling, Jocelyn Shulhan, Annabritt Chisholm, Andrea Milne, Purnima Sundar, Shannon D. Scott, and Amanda S. Newton. 2016. A Systematic Review of Social Media Use to Discuss and View Deliberate Self-Harm Acts. *PLoS ONE* 11 (2016).

[32] Venkatesh Edupuganti. 2017. Harassment detection on twitter using conversations. (2017).

[33] Aiman El Asam and Adrienne Katz. 2018. Vulnerable young people and their experience of online risks. *Human–Computer Interaction* 33, 4 (2018), 281–304.

[34] Isvani Frías-Blanco, Alberto Verdecia-Cabrera, Agustín Ortiz-Díaz, and Andre Carvalho. 2016. Fast adaptive stacking of ensembles. (2016), 929–934.

[35] Bo Gao, Bettina Berendt, and Joaquin Vanschoren. 2015. Who is more positive in private? Analyzing sentiment differences across privacy levels and demographic factors in Facebook chats and posts. *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (2015), 605–610.

[36] Joshua Garland, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne, and Mirta Galesic. 2020. Countering hate on social media: Large scale classification of hate and counter speech. *arXiv preprint arXiv:2006.01974* (2020).

[37] General Data Protection Regulation (GDPR). 2021. Art. 20 GDPR – Right to data portability | General Data Protection Regulation (GDPR). (2021). https://gdpr-info.eu/art-20-gdpr/

[38] Anastasia Giahanou, Guobiao Zhang, and Paolo Rosso. 2020. Multimodal Fake News Detection with Textual, Visual and Semantic Information. (2020).

[39] Theodoros Giannakopoulos, Aggelos Pikrakis, and Sergios Theodoridis. 2010. A Multimodal Approach to Violence Detection in Video Sharing Sites. *2010 20th International Conference on Pattern Recognition* (2010), 3244–3247.
[40] Muthukumarasamy Govindarajan. 2015. Comparative study of ensemble classifiers for direct marketing. *Intell. Decis. Technol.* 9 (2015), 141–152.
[41] Michele P. Hamm, Amanda S. Newton, Annabritt Chisholm, Jocelyn Shulhan, Andrea Milne, Purnima Sundar, Heather Ennis, Shannon D. Scott, and Lisa Hartling. 2015. Prevalence and Effect of Cyberbullying on Children and Young People: A Scoping Review of Social Media Studies. *JAMA pediatrics* 169 8 (2015), 770–7.
[42] Heidi Hartikainen, Afsaneh Razi, and Pamela Wisniewski. 2021. Safe Sexting: The Advice and Support Adolescents Receive from Peers regarding Online Sexual Risks. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–31.
[43] Heidi Hartikainen, Afsaneh Razi, and Pamela Wisniewski. 2021. 'If You Care About Me, You'll Send Me a Pic'-Examining the Role of Peer Pressure in Adolescent Sexting. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*. 67–71.
[44] Cormac Herley. 2012. Why do nigerian scammers say they are from nigeria? (2012).
[45] Alex Hern. 2021. Priti Patel v facebook is the latest in a 30-year fight over encryption. *The Guardian* (Apr 2021). https://www.theguardian.com/technology/2021/apr/19/priti-patel-v-facebook-is-the-latest-in-a-30-year-fight-over-encryption
[46] Jiani Hu, Toshihiko Yamasaki, and Kiyoharu Aizawa. 2016. Multimodal learning for image popularity prediction on social media. (2016), 1–2. https://doi.org/10.1109/ICCE-TW.2016.7521017
[47] Zainab Iftikhar, Osama Younus, Taha Sardar, Hammad Arif, Mobin Javed, Suleman Shahid, et al. 2021. Designing Parental Monitoring and Control Technology: A Systematic Review. In *IFIP Conference on Human-Computer Interaction*. Springer, 676–700.
[48] Chris Kanich, Christian Kreibich, Kirill Levchenko, Brandon Enright, Geoffrey M Voelker, Vern Paxson, and Stefan Savage. 2008. Spamalytics: An empirical analysis of spam marketing conversion. (2008), 3–14.
[49] Mona Kasra. 2017. Vigilantism, public shaming, and social media hegemony: The role of digital-networked images in humiliation and sociopolitical control. *The Communication Review* 20 (2017), 172 – 188.
[50] Seunghyun Kim, Afsaneh Razi, Gianluca Stringhini, Pamela Wisniewski, and Munmun De Choudhury. 2021. You Don't Know How I Feel: Insider-Outsider Perspective Gaps in Cyberbullying Risk Detection. (2021).
[51] Seunghyun Kim, Afsaneh Razi, Gianluca Stringhini, Pamela J Wisniewski, and Munmun De Choudhury. 2021. A Human-Centered Systematic Literature Review of Cyberbullying Detection Algorithms. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–34.
[52] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. (2014).
[53] Sotiris B. Kotsiantis. 2011. An incremental ensemble of classifiers. *Artificial Intelligence Review* 36 (2011), 249–266.
[54] Julia Kruk, Jonah Lubin, Karan Sikka, Xiaoyu Lin, Dan Jurafsky, and Ajay Divakaran. 2019. Integrating Text and Image: Determining Multimodal Document Intent in Instagram Posts. *ArXiv* abs/1904.09073 (2019).
[55] Shiv Balak Kumar. 2021. Best encrypted social media platforms in 2021. *LinkedIn* (Sep 2021). https://www.linkedin.com/pulse/best-encrypted-social-media-platforms-2021-shiv-balak-kumar
[56] Kirti Kumari, Jyoti Prakash Singh, Yogesh K. Dwivedi, and Nripendra P. Rana. 2019. Aggressive Social Media Post Detection System Containing Symbolic Images. (2019).
[57] Larissa Lewis, Julie Mooney Somers, Rebecca J. Guy, Lucy Watchirs-Smith, and S Rachel Skinner. 2018. 'I see it everywhere': young Australians unintended exposure to sexual content online. *Sexual health* 15 4 (2018), 335–341.
[58] Chenhao Lin, Pengwei Hu, Hui Su, Shaochun Li, Jing Mei, Jie Zhou, and Henry Leung. 2020. SenseMood: Depression Detection on Social Media. *Proceedings of the 2020 International Conference on Multimedia Retrieval* (2020).
[59] Chen Ling, Utkucan Balcı, Jeremy Blackburn, and Gianluca Stringhini. 2021. A first look at zoombombing. In *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 1452–1467.
[60] Meizhen Lv, Ang Li, Tianli Liu, and Tingshao Zhu. 2015. Creating a Chinese suicide dictionary for identifying suicide risk on social media. *PeerJ* 3 (2015).
[61] Sheri Madigan, Vanessa C. Villani, Corry Azzopardi, Danae Laut, Tanya D. Smith, Jeff R. Temple, Dillon Thomas Browne, and Gina Dimitropoulos. 2018. The Prevalence of Unwanted Online Sexual Exposure and Solicitation Among Youth: A Meta-Analysis. *The Journal of adolescent health : official publication of the Society for Adolescent Medicine* 63 2 (2018), 133–141.
[62] Catherine D. Marcum. 2007. Interpreting the Intentions of Internet Predators: An Examination of Online Predatory Behavior. *Journal of Child Sexual Abuse* 16 (2007), 114 – 99.
[63] Enrico Mariconti, Guillermo Suarez-Tangil, Jeremy Blackburn, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Jordi Luque Serrano, and Gianluca Stringhini. 2019. "You Know What to Do": Proactive Detection of YouTube Videos Targeted by Coordinated Hate Attacks. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–21.

[64] Tolba Marwa, Ouadfel Salima, and Meshoul Souham. 2018. Deep learning for online harassment detection in tweets. (12 2018). https://doi.org/10.1109/PAIS.2018.8598530
[65] Lincy Meera Mathews and Seetha Hari. 2019. Learning From Imbalanced Data. *Advances in Computer and Electrical Engineering* (2019).
[66] Diana Maynard, David Dupplaw, and Jonathon S. Hare. 2013. Multimodal Sentiment Analysis of Social Media. (2013).
[67] Nora McDonald and Andrea Forte. 2022. Privacy and Vulnerable Populations. In *Modern Socio-Technical Perspectives on Privacy*. Springer, Cham, 337–363.
[68] Bridget Christine McHugh, Pamela Wisniewski, Mary Beth Rosson, and John M Carroll. 2018. When social media traumatizes teens. *Internet Research* (2018).
[69] Sam McNeilly, Luke Hutton, and Tristan Henderson. 2013. Understanding ethical concerns in social media privacy studies. In *Proceedings of the ACM CSCW Workshop on Measuring Networked Social Privacy*.
[70] Mayank Meghawat, Satyendra Yadav, Debanjan Mahata, Yifang Yin, Rajiv Ratn Shah, and Roger Zimmermann. 2018. A Multimodal Approach to Predict Social Media Popularity. *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)* (2018), 190–195.
[71] Md. Waliur Rahman Miah, John Yearwood, and Siddhivinayak Kulkarni. 2011. Detection of child exploiting chats from a mixed chat dataset as a text classification task. (2011).
[72] Kimberly J Mitchell, David Finkelhor, Lisa M Jones, and Janis Wolak. 2012. Prevalence and characteristics of youth sexting: A national study. *Pediatrics* 129, 1 (2012), 13–20.
[73] Vatsala Mittal, Aastha Kaul, Santoshi Sen Gupta, and Anuja Arora. 2017. Multivariate Features Based Instagram Post Analysis to Enrich User Experience. (2017).
[74] Mainack Mondal, Leandro Araújo Silva, Denzil Correa, and Fabrício Benevenuto. 2018. Characterizing usage of explicit hate expressions in social media. *New Review of Hypermedia and Multimedia* 24 (2018), 110 – 130.
[75] NewYork-Times. 2021. Whistle-Blower Says Facebook 'Chooses Profits Over Safety' - The New York Times. (2021). https://www.nytimes.com/2021/10/03/technology/whistle-blower-facebook-frances-haugen.html
[76] Laura Louise Nicklin, Emma Swain, and Joanne Lloyd. 2020. Reactions to Unsolicited Violent, and Sexual, Explicit Media Content Shared over Social Media: Gender Differences and Links with Prior Exposure. *International Journal of Environmental Research and Public Health* 17 (2020).
[77] Adrien Ogee and Marco Pineda. 2019. Encryption is under threat this is how it affects you. *World Economic Forum* (2019). https://www.weforum.org/agenda/2019/12/encryption-cybersecurity-privacy-explainer/
[78] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. 2017. Towards a visual privacy advisor: Understanding and predicting privacy risks in images. In *Proceedings of the IEEE international conference on computer vision*. 3686–3695.
[79] Javier Parapar, David E. Losada, and Álvaro Barreiro. 2014. Combining Psycho-linguistic, Content-based and Chat-based Features to Detect Predation in Chatrooms. *J. Univers. Comput. Sci.* 20 (2014), 213–239.
[80] Liliana Pasquale, Paola Zippo, Cliona Curley, Brian O'Neill, and Marina Mongiello. 2020. Digital age of consent and age verification: Can they protect children? *IEEE Software* (2020).
[81] Sachin R Pendse, Daniel Nkemelu, Nicola J. Bidwell, Sushrut Jadhav, Soumitra Pathare, Munmun De Choudhury, and Neha Kumar. 2022. From Treatment to Healing:Envisioning a Decolonial Digital Mental Health. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 548, 23 pages. https://doi.org/10.1145/3491102.3501982
[82] Pew-Research. 2021. 7 facts about Americans and Instagram | Pew Research Center. (2021). https://www.pewresearch.org/fact-tank/2021/10/07/7-facts-about-americans-and-instagram/
[83] Anthony T Pinter, Pamela J Wisniewski, Heng Xu, Mary Beth Rosson, and Jack M Caroll. 2017. Adolescent online safety: Moving beyond formative evaluations to designing solutions for the future. (2017), 352–357.
[84] Suzanne L. Porath. 2011. Text Messaging and Teenagers: A Review of the Literature. *Journal of the Research Center for Educational Technology* 7 (2011), 86–99.
[85] Calton Pu, Steve Webb, Oleg Kolesnikov, Wenke Lee, and Richard Lipton. 2006. Towards the integration of diverse spam filtering techniques. (2006), 17–20.
[86] Sukma Ari Ragil Putri and AAI Prihandari Satvikadewi. 2017. A critical discourse analysis study of cyberbullying in LGBTQ's Instagram account. 33 (2017).
[87] Afsaneh Razi, Zainab Agha, Neeraj Chatlani, and Pamela Wisniewski. 2020. Privacy Challenges for Adolescents as a Vulnerable Population. In *Networked Privacy Workshop of the 2020 CHI Conference on Human Factors in Computing Systems*.
[88] Afsaneh Razi, Ashwaq AlSoubai, Seunghyun Kim, Shiza Ali, Gianluca Stringhini, Munmun Choudhury, and Pamela J. Wisniewski. 2023. Sliding into My DMs: Detecting Uncomfortable or Unsafe Sexual Risk Experiences within Instagram Direct Messages Grounded in the Perspective of Youth. 28.
[89] Afsaneh Razi, Ashwaq Alsoubai, Seunghyun Kim, Nurun Naher, Shiza Ali, Gianluca Stringhini, Munmun De Choudhury, and Pamela J. Wisniewski. 2022. Instagram Data Donation: A Case Study on Collecting Ecologically Valid Social

Media Data for the Purpose of Adolescent Online Risk Detection. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 39, 9 pages. https://doi.org/10.1145/3491101.3503569

[90] Afsaneh Razi, Karla Badillo-Urquiola, and Pamela J. Wisniewski. 2020. Let's Talk about Sext: How Adolescents Seek Support and Advice about Their Online Sexual Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376400

[91] Afsaneh Razi, Seunghyun Kim, Ashwaq Alsoubai, Gianluca Stringhini, Thamar Solorio, Munmun De Choudhury, and Pamela J Wisniewski. 2021. A Human-Centered Systematic Literature Review of the Computational Approaches for Online Sexual Risk Detection. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–38.

[92] Amal Rekik, Salma Jamoussi, and Abdelmajid Ben Hamadou. 2019. Violent Vocabulary Extraction Methodology: Application to the Radicalism Detection on Social Media. (2019).

[93] Simon M. Rice, Jo Robinson, Sarah A. Bendall, Sarah Elisabeth Hetrick, Georgina Cox, Eleanor Bailey, John F. M. Gleeson, and Mario Alvarez-Jimenez. 2016. Online and Social Media Suicide Prevention Interventions for Young People: A Focus on Implementation and Moderation. *Journal of the Canadian Academy of Child and Adolescent Psychiatry = Journal de l'Academie canadienne de psychiatrie de l'enfant et de l'adolescent* 25 2 (2016), 80–6.

[94] Daniel R Richards and Bige Tunçer. 2018. Using image recognition to automate assessment of cultural ecosystem services from social media photographs. *Ecosystem services* 31 (2018), 318–325.

[95] Peter Roesler. 2021. Study Shows Why Teens and Young Adults Love Instagram - Web Marketing Pros. (2021). https://www.webmarketingpros.com/study-shows-why-teens-and-young-adults-love-instagram/

[96] Arpita Roy, Anamika Paul, Hamed Pirsiavash, and Shimei Pan. 2017. Automated Detection of Substance Use-Related Social Media Posts Based on Image and Text Analysis. *2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)* (2017), 772–779.

[97] Zainab Saad Rubaidi, Boulbaba Ben Ammar, and Mohamed Ben Aouicha. 2022. Fraud Detection Using Large-scale Imbalance Dataset. *International Journal on Artificial Intelligence Tools* (2022).

[98] Joni Salminen, Maximilian Hopf, Shammur A Chowdhury, Soon-gyo Jung, Hind Almerekhi, and Bernard J Jansen. 2020. Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences* 10, 1 (2020), 1–34.

[99] Andreas Schmid, Thomas Fischer, Alexander Weichart, Alexander Hartmann, and Raphael Wimmer. 2021. Demonstrating ScreenshotMatcher: Taking Smartphone Photos to Capture Screenshots. *Mensch und Computer 2021* (2021).

[100] Carsten Schwemmer, Saïd Unger, and Raphael Heiberger. 2022. Automated Image Analysis for Studying Online Behaviour. (2022).

[101] Scikit. 2022. Feature Importances With A Forest Of Trees. https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html

[102] Scikit. 2022. Permutation Feature Importance. https://scikit-learn.org/stable/modules/permutation_importance.html?highlight=gini%2Bimportance

[103] scikitlearn. 2021. Decision Trees. (2021). https://scikit-learn.org/stable/modules/tree.html

[104] scikitlearn. 2021. Ensemble Methods. (2021). https://scikit-learn.org/stable/modules/ensemble.html

[105] Rajiv Ratn Shah. 2016. Multimodal Analysis of User-Generated Content in Support of Social Media Applications. *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval* (2016).

[106] Guangyao Shen, Jia Jia, Liqiang Nie, Fuli Feng, Cunjun Zhang, Tianrui Hu, Tat-Seng Chua, and Wenwu Zhu. 2017. Depression Detection via Harvesting Social Media: A Multimodal Dictionary Learning Solution. (2017).

[107] Shruthi and Prof Mangala C. 2017. A Framework for Automatic Detection and Prevention of Cyberbullying in Social Media. *International Journal of Innovative Research in Computer and Communication Engineering* 5, 6 (2017), 86–90. www.ijircce.com

[108] Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. 2018. Image chat: Engaging grounded conversations. *arXiv preprint arXiv:1811.00945* (2018).

[109] Shubham Singh, Rishabh Kaushal, Arun Balaji Buduru, and Ponnurangam Kumaraguru. 2019. KidsGUARD: fine grained approach for child unsafe video representation and detection. *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing* (2019).

[110] Ray Smith. 2007. An overview of the Tesseract OCR engine. 2 (2007), 629–633.

[111] Amie Stepanovich. 2016. *A human rights response to government hacking* (2016). https://www.accessnow.org/cms/assets/uploads/2016/09/Gov-Hacking-Three-Pager.pdf

[112] Gianluca Stringhini, Manuel Egele, Apostolis Zarras, Thorsten Holz, Christopher Kruegel, and Giovanni Vigna. 2012. Babel: Leveraging Email Delivery for Spam Mitigation. (2012), 16–32.

[113] Guillermo Suarez-Tangil, Matthew Edwards, Claudia Peersman, Gianluca Stringhini, Awais Rashid, and Monica Whitty. 2019. Automatically dismantling online dating fraud. *IEEE Transactions on Information Forensics and Security*

15 (2019), 1128–1137.
[114] Kaveri Subrahmanyam and Patricia Greenfield. 2008. Online communication and adolescent relationships. *The future of children* (2008), 119–146.
[115] Muhammad Uzair Tariq, Afsaneh Razi, Karla Badillo-Urquiola, and Pamela Wisniewski. 2019. A Review of the Gaps and Opportunities of Nudity and Skin Detection Algorithmic Research for the Purpose of Combating Adolescent Sexting Behaviors. (2019), 90–108.
[116] terre des hommes. 2022. Child safety by design research paper. *Child safety by design Research paper* (May 2022). https://www.terredeshommes.nl/en/publications/research-paper-child-safety-by-design
[117] Kurt Thomas, Patrick Gage Kelley, Sunny Consolvo, Patrawat Samermit, and Elie Bursztein. 2022. "It's common and a part of being a content creator": Understanding How Creators Experience and Cope with Hate and Harassment Online. In *CHI Conference on Human Factors in Computing Systems*. 1–15.
[118] Niklas Torstensson and Tarja Susi. 2015. Online sexual grooming and offender tactics - : What can we learn from social media dialogues? (2015).
[119] Penny Trieu and Nancy K Baym. 2020. Private responses for public sharing: understanding self-presentation and relational maintenance via stories in social media. (2020), 1–13.
[120] Emily Tseng, Mehrnaz Sabet, Rosanna Bellini, Harkiran Kaur Sodhi, Thomas Ristenpart, and Nicola Dell. 2022. Care Infrastructures for Digital Security in Intimate Partner Violence. In *CHI Conference on Human Factors in Computing Systems*. 1–20.
[121] Thomas W. Valente, Anamara Ritt-Olson, Alan Stacy, Jennifer B. Unger, Janet Okamoto, and Steve Sussman. 2007. Peer acceleration: effects of a social network tailored substance abuse prevention program among high-risk adolescents. *Addiction* 102, 11 (2007), 1804–1815. https://doi.org/10.1111/j.1360-0443.2007.01992.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1360-0443.2007.01992.x
[122] Cynthia Van Hee, Gilles Jacobs, Chris Emmery, Bart Desmet, Els Lefever, Ben Verhoeven, Guy De Pauw, Walter Daelemans, and Véronique Hoste. 2018. Automatic detection of cyberbullying in social media text. *PloS one* 13, 10 (2018), e0203794.
[123] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2017. Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 4 (2017), 652–663. https://doi.org/10.1109/TPAMI.2016.2587640
[124] Nishant Vishwamitra, Hongxin Hu, Feng Luo, and Long Cheng. 2021. Towards Understanding and Detecting Cyberbullying in Real-world Images. (2021).
[125] Ashley Marie Walker, Yaxing Yao, Christine Geeng, Roberto Hoyle, and Pamela Wisniewski. 2019. Moving beyond'one size fits all' research considerations for working with vulnerable populations. *Interactions* 26, 6 (2019), 34–39.
[126] Dale Walker. 2021. What is end-to-end encryption and why is everyone fighting over it? *IT PRO* (Oct 2021). https://www.itpro.com/security/encryption/359943/what-is-end-to-end-encryption-and-why-is-everyone-fighting-over-it
[127] Murdoch Watney. 2020. Law Enforcement Access to End-to-End Encrypted Social Media Communications. In *7th European Conference on Social Media ECSM 2020*. 322.
[128] Elizabeth Whittaker and Robin M. Kowalski. 2015. Cyberbullying Via Social Media. *Journal of School Violence* 14 (2015), 11 – 29.
[129] Pamela Wisniewski, Arup Kumar Ghosh, Heng Xu, Mary Beth Rosson, and John M Carroll. 2017. Parental Control vs. Teen Self-Regulation: Is there a middle ground for mobile online safety? (2017), 51–69.
[130] Pamela Wisniewski, Heng Xu, Mary Beth Rosson, Daniel F Perkins, and John M Carroll. 2016. Dear diary: Teens reflect on their weekly online risk experiences. (2016), 3919–3930.
[131] Pamela J Wisniewski and Xinru Page. 2022. Privacy theories and frameworks. In *Modern Socio-Technical Perspectives on Privacy*. Springer, Cham, 15–41.
[132] Rebecca N Wright, L Jean Camp, Ian Goldberg, Ronald L Rivest, and Graham Wood. 2002. Privacy tradeoffs: myth or reality? (2002), 147–151.
[133] Michele L. Ybarra and Kimberly J. Mitchell. 2008. How Risky Are Social Networking Sites? A Comparison of Places Online Where Youth Sexual Solicitation and Harassment Occurs. *Pediatrics* 121 (2008), e350 – e357.
[134] Justine Zhang, Jonathan P Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Nithum Thain, and Dario Taraborelli. 2018. Conversations gone awry: Detecting early signs of conversational failure. *arXiv preprint arXiv:1805.05345* (2018).

Received July 2022; revised October 2022; accepted January 2023