ABSTRACT

GARG, VAIBHAV. Unveiling Harassment through Natural Language Processing. (Under the direction of Dr. Munindar P. Singh).

Warning: This dissertation contains some examples of harassment that may be triggering to some individuals.

Harassment is a prevailing and challenging problem in our society, leading to various types of abuse. This research studies three different scenarios of harassment: online, real-world, and a mix of both. We analyze online posts describing specific problems within each scenario and solve such problems using natural language processing.

In digital space, many apps can be misused for stalking and spying purposes. We found that app reviews reveal critical information about misused functionalities. These reviews are useful in auditing such apps and preventing further misuse. We leverage app reviews (written by victims, abusers, and third parties) to identify such problematic apps and their misused functionalities.

In the space of real-world harassment, we focus on the problem of sexual harassment and assault. Survivors of such experiences freely share their MeToo stories on platforms such as Reddit. However, due to the high character limit per post, such stories could be long and demanding to read, which may challenge prospective helpers to provide effective responses. We develop a natural-language-based tool that highlights important parts of long stories, in turn facilitating helpers in providing effective support. In the same real-world scenario, we also understand how the language of MeToo stories reflects on violations of survivors' consent. We linguistically study how consent violation stories are expressed.

In the mixed scenario, we understand an unexplored yet interesting form of speech, inciting speech. We investigate public WhatsApp messages and identify rhetorical strategies used to incite readers against a religious community. On social media, inciting speech can be disturbing to read (online harassment) and may lead to violence against its target in the real world. We understand the nature of such speech along with its prevalence on multiple platforms against multiple targets.

Overall, we employ NLP techniques to understand natural language text and build computational models addressing harassment in all three scenarios.

Unveiling Harassment through Natural Language Processing

by
Vaibhav Garg

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Computer Science

Raleigh, North Carolina
2024

APPROVED BY:

Dr. Sandeep Kuttal	Dr. Min Chi
Dr. Veronica Catete	Dr. Kama Kosenko

Dr. Munindar P. Singh
Chair of Advisory Committee

BIOGRAPHY

In 2019, Vaibhav Garg completed his Masters in Computer Science from the Indraprastha Insitute of Information Technology, Delhi (IIIT D). As a Master's student, Vaibhav discovered his interest in the application of AI. Following his interests, he interned at InfoEdge India Limited to develop AI models identifying fake profiles on a leading matrimony website. Working on such exciting projects motivated him to delve into this area and pursue a Ph.D. program at NC State University.

During his Ph.D. he identified a niche research area, *AI for Social Good*, under which he could build AI solutions having a high societal impact. Addressing the problem of harassment through NLP was one of his endeavors. During his Ph.D., he mentored both undergraduate and graduate students by providing them with specific research tasks. In spring 2023, he was nominated for *the Carla Savage Award - Awesome Ph.D. Student* by his advisor and lab members. Eventually, he won the title by a substantial number of votes. He was also fortunate to be the recipient of multiple research grants such as NSA and NSF. Apart from his academic life, he was quite passionate about meditating, Bollywood dancing, and playing badminton.

Vaibhav is grateful for all the learnings during his Ph.D. journey and is excited to start his academic career as a new chapter.

ACKNOWLEDGEMENTS

I would like to thank my parents, Mr. Rajesh and Mrs. Varsha Gupta, along with other family members for their unconditional love and support. While sitting in India, they were just a call away from me and would emotionally support to get through my tough times. I am also grateful to a few friends who believed in me and were open to discussing any problems that I was facing. Not only supportive but non-supportive friends are also worth acknowledging, who challenged me to prove them wrong.

Additionally, I would thank my lab senior, Hui Guo, who provided helpful advice on all critical Ph.D. milestones. I learned many research methods and practices by just following his footsteps. Last but not least, my advisor, Dr. Munindar P. Singh, deserves a big shout-out. He helped me to significantly improve all my papers, ranging from effective writing to conducting additional experiments. Despite multiple paper rejections, he would encourage me to dissociate from such outcomes and remain persistent in my efforts.

I am grateful to get support from the U.S. Department of Defense (Science of Security Lablet grant H98230-17-D-0080) and the National Science Foundation (grants IIS-1908374 and IIS-2116751).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER

1

INTRODUCTION

Harassment is a well-known problem in our society. According to the Center for Disease Control and Prevention (CDC), in the United States, one in five women has experienced attempted rape, and one in four men has experienced sexual violence. With the advent of the internet, the problem of harassment is not only limited to real-world scenarios (e.g., verbal, physical, or sexual harassment) but also applies to online and mixed scenarios. In totality, we divide the harassment scenarios into three types: (i) online harassment, (ii) real-world harassment, and (iii) mixed of both. *Online harassment* includes targeting individuals through disturbing text or by disseminating their personal information, in turn causing them threat, harm, and distress (Rights 2023). On the other hand, *Real-world harassment* is an unwelcome conduct that is based on attributes such as sex, race, color, and so on (EEOC 2023). In the third scenario, harassment is initiated in the online setting but may also spread to a real-world setting. In this thesis, we identify specific problems in each of these scenarios and computationally solve them.

In the online harassment scenario, we identify that the abusers misuse mobile apps to access information (such as location and call logs) of other users or bystanders (victims). Such apps are dual-use and violate victims' privacy expectations. App users should be aware of such apps and their functionalities before installing them. Moreover, app developers and platforms such as Apple App Store should take measures to ensure victims' privacy. We found that both abusers and victims report such harassment in reviews of these apps. In reviews, abusers brag about using these apps to spy on or stalk their family or friends, whereas victims share their concerns and grievances, including frustration at the loss of privacy. We leverage the rich information in reviews and apply NLP techniques to identify these apps and their problematic functionalities.

For the real-world harassment scenario, we focus on the prevailing problem of sexual harassment and assault. Survivors of such experiences may not be open about sharing it with family or friends. To seek supportive responses or advice, they leverage online platforms to anonymously post about their experiences. One such platform is Reddit which hosts multiple subreddits such as r/meToo. We found that while sharing harassment experiences on these subreddits, the survivors end up writing long posts (with mean and maximum of 1,881 and 33,432 characters, respectively). For a prospective helper on subreddits, reading long stories can be demanding and time-consuming. Moreover, while framing the response to the survivors, the helper may miss out on important parts of the long posts that should be addressed. We leverage these long MeToo stories (shared on subreddits) to identify three types of key information: (i) incident, (ii) effects on the survivor, and (iii) advice sought. We develop a natural language-based model that can highlight this key information (in long posts) for helpers to address in their responses.

In the real-world scenario, we also study MeToo stories expressing consent violations. We leverage the FRIES consent framework by Javidi et al. (2023) and investigate their violations in MeToo stories. We study the prominent violations and how linguistically they are expressed in stories.

In the third scenario, we focus on the problem of inciting speech prevalent on social media. Inciting speech is speech that can make the readers angry or urge them to act against the target. As compared to hate speech, inciting speech is subtle and hence harder to identify. On one hand, inciting speech on social media can disturb, embarrass, or threaten the target (online harassment), whereas on the other hand, it can also lead to violence against the target (real-world harassment). For concreteness, we focus on WhatsApp posts (from Indian groups discussing politics) that include inciting speech against Muslims. We found that incitement can of three forms: criticizing the target community's members and their beliefs, criticizing the target community for its imputed misdeeds, and exhortation against the target community. We leverage NLP techniques to identify inciting sentences in a social media post.

1.1 Background and Research Objectives

This thesis supports the following statement.

Each harassment scenario is structurally different and the associated NL text is linguistically distinct. Despite such differences, similar approaches involving curating datasets followed by transfer learning are effective at understanding harassment cases or identifying them in each scenario.

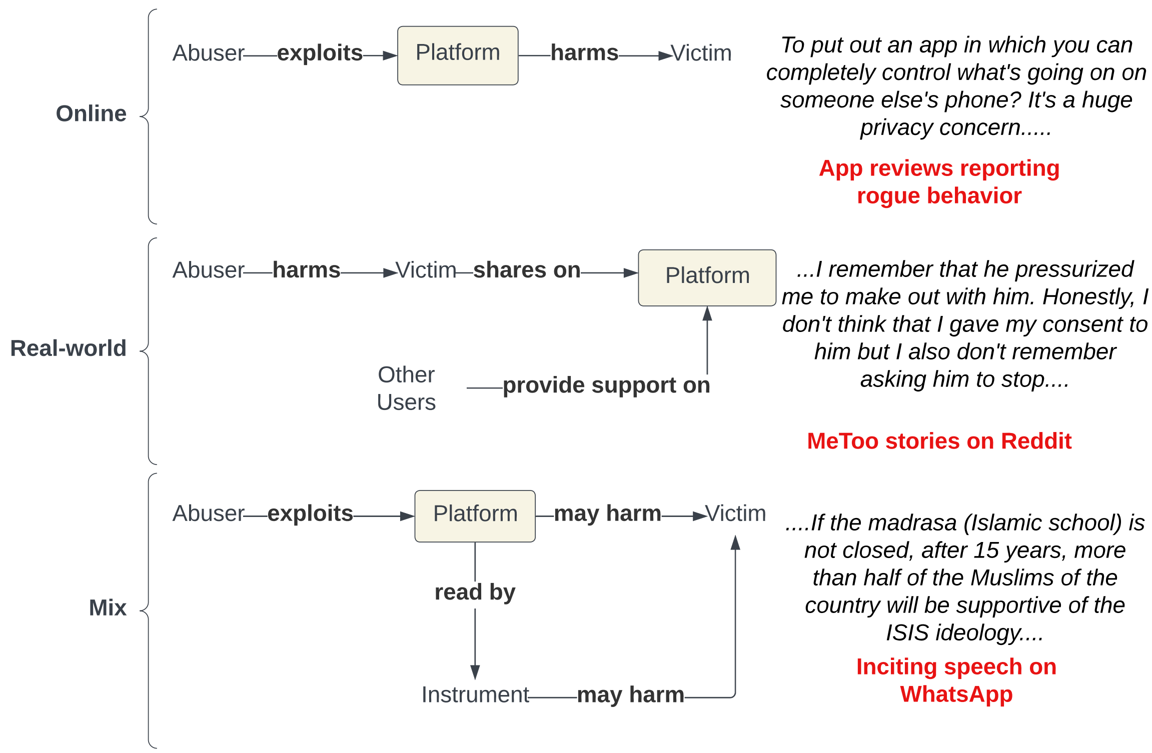


Figure 1.1: Each scenario of harassment is both structurally and linguistically different.

We explain the above thesis statement through Figure 1.1, which illustrates how each harassment scenario is distinct. Structurally, in the real-world scenario, the abuser directly harms the victim, who then leverages the platform to seek support from other users. However, in the other two scenarios, the abuser harms the victim by exploiting the platform. The sample text in the figure describes three different harassment cases, also indicating the linguistic distinction between these scenarios. Despite such structural and linguistic differences, this thesis emphasize that similar NLP approaches are effective at understanding or identifying these harassment cases. Such NLP approaches include training transfer learning models on the curated datasets.

We discuss the specific research objectives in each scenario in the sections below.

1.1.1 Online Harassment

Chatterjee et al. (2018) use app descriptions to identify intimate partner surveillance (IPS) apps, the apps that someone can use to spy on their intimate partners (spouse, boyfriend, or girlfriend). Such IPS apps are dual-use apps, that have a legitimate purpose but are misused for spying. Their work is specific to IPS, whereas we focus on the general misuse of mobile apps for information access.

The privacy expectations of an app user or a third party (i.e., a *victim*) are violated when the victim (1) is not aware of the information access (spying) or (2) may know about the access, but is uncomfortable with it. The latter case includes incidents of forced consent or when public information on apps (such as dating platforms) is accessed beyond the victim's level of comfort (stalking). We use the term *rogue behavior* to mean these two types of information access and use *rogue apps* to mean the apps that enable rogue behavior. We use the term *rogue behavior* to mean these two types of information access and use *rogue apps* to mean the apps that enable rogue behavior.

We found that app reviews contain rich information about such misuse. Hence, we aim to identify such rogue apps through their reviews and uncover their rogue functionalities. This will help in misuse audit and warn users (potential victims) about rogue behavior. Moreover, app platforms and developers should take measures to meet users' privacy expectations.

1.1.2 Real-World Harassment

For understanding experiences of sexual harassment and assault, prior studies on MeToo stories (Karlekar and Bansal 2018; Hassan et al. 2020; Khatua et al. 2018; Ghosh Chowdhury et al. 2019a) focus on classification. The expectation is that a prospective helper (on online platforms) can provide support to the survivors of the identified posts. However, merely identifying relevant posts is not enough. Prior research shows that important parts of the posts such as (i) the Incident, (ii) the effects on the survivor (Field-Springer et al. 2022), and (iii) the advice that the survivor is seeking (Andalibi et al. 2016), must be addressed to provide help. We delve into how these three components of a story can be identified. Automatically identifying and highlighting them can assist helpers in addressing important concerns in their responses.

In addition, we look for consent violations in MeToo stories from the language standpoint. This includes investigating what type of prominent violations are expressed, their language attributes, and how they are interrelated. This will help in understanding consent linguistically.

1.1.3 Mixed Harassment Scenario

Inciting speech is to be differentiated from hate speech, which is another kind of antisocial communication (Calvert 1997). Unlike hate speech (ElSherief et al. 2018), inciting speech doesn't necessarily contain derogatory language (words such as *n*gger* and *a**hole*) to attack an individual or a group. Whereas many social media studies focus on hate speech identification (Aluru et al. 2021; Das et al. 2022), identification of inciting speech has not garnered as much attention.

Our qualitative analysis revealed three rhetorical strategies used for incitement: (i) identity (attacking target members and their beliefs), (ii) imputed misdeeds (highlighting misdeeds of the target), and (iii) exhortation (calling for action). This problem is even more prominent on platforms such as WhatsApp where incitement is used against religious groups but there is no content moderation due to end-to-end encryption (Saha et al. 2021). Hence, we aim to understand such inciting speech.

1.2 Contributions and Novelty

We list contributions and novelty for each project below.

1.2.1 Understanding App Reviews for Misuse Audits

Our work's novelty lies in leveraging app reviews to identify rogue apps and their rogue functionalities. App reviews despite being a public source of information, remain unexplored in the rogue scenario. We introduce assigning alarmingness scores to reviews and rogue scores to apps, based on the reported rogue behavior. We list our contributions below.

iROGUE, an app reviews-based approach for identifying rogue apps and their functionalities. We have predicted a ranked list of 239 rogue apps, along with their alarming reviews reporting rogue behavior. Our approach can be used in real-time to identify new rogue apps as soon as their reviews arrive. We envision our approach to be incrementally updated by incorporating new reviews and keep identifying other rogue apps through the learnt behavior. Moreover, our approach is not only applicable to the Apple App Store but also other platforms such as Google Play Store.

Our qualitative analysis revealed four types of rogue functionalities being prominent. These functionalities range from profile stalking to monitoring daily phone activities. Future app users (potential victims) should be warned of the identified functionalities before installing them on phones. Moreover, app platforms such as the Apple App Store can prioritize scrutinizing the rogue apps ranked high in the list or ask their developers to work on providing user privacy. We discuss about this project in Chapter 2.

1.2.2 Identifying Incidents, Effects, and Requests for Advice from MeToo Stories

Prior research (Karlekar and Bansal 2018; Hassan et al. 2020; Khatua et al. 2018; Ghosh Chowdhury et al. 2019a) computationally identifies relevant MeToo stories from a massive stream of online text. To the best of our knowledge, we are the first ones to identify important parts of the relevant stories. In particular, we make the following contributions.

We curate METHREE, a dataset containing 8,947 sentences, labeled for the three categories. Constructing a sufficiently natural and precise dataset turns out to be nontrivial. We leverage active learning for labeling with tractable manual effort.

We investigated MeToo stories for the structure and prominence of three parts. On an average, union of incident, effects, and advice requests constitute 55.11% of the story's content. We also found a lot of variation in story structure which makes the automatic highlighting a challenging task.

We conducted an empirical study of helpers on multiple MeToo-related subreddits. We found that they find reading and responding to long stories demanding. We also tested the effectiveness of our highlighting tool through the same survey of helpers. We discuss this project in Chapter 3.

1.2.3 Understanding Consent Violations through Stories

To the best of our knowledge, this is the first study to understand consent stories linguistically. We contribute in the following manner.

MYCONSENT, a labeled dataset of stories with their consent violations. This dataset helped us to analyze linguistic attributes in stories describing violations. Moreover, while labeling, we found language cues representing each type of violation, which could be helpful for future researchers understanding consent.

Our statistical analysis revealed prominent violations expressed through language. We also investigated co-occurances of multiple violations in the same story. We found that informed and specific violations co-occur most frequently, followed by the pair of explicit and enthusiastic violations.

Our linguistic analysis revealed how time-sensitive attributes are often expressed in stories describing freely given violation. Moreover, the language in violation stories indicate low confidence, leadership, and social status as compared to non-violation stories. However, the authenticity of language is high in violation stories as opposed to non-violation ones. We discuss this project in Chapter 4.

1.2.4 Understanding Inciting Speech As New Malice

Inciting speech as a discrete category of antisocial communication has not been computationally studied. Hence, our research is inherently novel. We make the following contributions.

Through qualitative analysis, we found three distinct ways of how religious incitement is expressed: identity, imputed misdeeds, and exhortation. We call them rhetorical strategies. In identity, the inciting speech either criticizes or stereotypes the victim, indirectly inducing readers' hostility against the victim. In imputed misdeeds strategy, the abuser alleges the victim of certain misdeeds. The allegations are sometimes based on facts or intended to spread misinformation. In exhortation, the abuser calls for real-world action against the victim, either directly or through subtle hints.

We curated a dataset called INCITE containing 7,000 sentences labeled for three rhetorical strategies. INCITE helped us to understand textual signatures of each strategy in the religious context. Moreover, we could train a computational model to identify such strategies from sentences of WhatsApp, Twitter, and Gab posts.

We found that not only Islamophobic but also non-Islamophobic posts contain inciting speech. This finding emphasizes the linguistic differences between Islamophobic speech (fear speech) and inciting speech. Through a variation of our computational model, we identified the prevalence of inciting speech even against non-religious targets. This explains the dire need to address inciting speech on online platforms. We discuss this project in Chapter 5.

CHAPTER

2

UNDERSTANDING MOBILE APP REVIEWS TO GUIDE MISUSE AUDITS

Background: Software audits majorly focus on program analysis of open-source applications. However, misuse done by app users (instead of app developers) falls out of the scope of such audits. Moreover, many apps on distribution platforms (such as Apple App Store) don't have their code as public open-source, making us incapable of auditing them. **Problem:** We address this challenge in responsible computing where a *rogue* mobile app is misused by one app user (an *abuser*) against another user or bystander (*victim*). We introduce the idea of a *misuse audit* of apps as a way of determining if they are rogue. **Method:** For misuse audit, we leverage app reviews to understand the problematic rogue apps and their functionalities causing misuse. First, we build a computational model to identify alarming reviews reporting such misuse. Second, using the model, we identify rogue apps and their functionalities and validate them through manual inspection of reviews or installed apps. **Findings:** Abusers and victims mostly write stories indicating rogue acts, whereas stories from the third person dominantly represent rogue potential. Surprisingly, positive reviews, including abusers' language with high dominance, also reveal rogue behavior. Our model identified a total of 239 apps exhibiting four types of rogue functionalities. **Implications:** Our method can help in the real-time identification of rogue apps and their functionalities, in turn facilitating their audits.

2.1 Introduction

Traditional audits of mobile apps conduct a review of their source code (García et al. 2021; Xia et al. 2015). However, interpersonal misuse arising from app users (instead of app developers) goes unnoticed by such processes. We introduce *misuse audit*, a process for auditing mobile apps through the reported misuse cases.

A misuse occurs when an app user (*abuser*) exploits the app to access information of other users or third parties (*victims*). In particular, a misuse audit can identify cases when the victim: (1) doesn't know about the information access (spying) or (2) may know about the access, but is uncomfortable with it. The latter case (often missed by traditional audits) includes incidents of forced consent or when public information (such as on dating apps) is accessed beyond the victim's level of comfort. We use the term *rogue behavior* to mean these two types of information access and use *rogue apps* to mean the apps that enable rogue behavior.

Research (Chatterjee et al. 2018; Freed et al. 2019; Havron et al. 2019) shows that rogue apps may cause discomfort, fear, and potential harm to the victim. Possible ways to prevent this risk include highlighting these apps and their rogue functionalities to users, app distribution platforms, and app developers. However, identifying these apps is nontrivial as they are dual-use in nature Chatterjee et al. (2018), meaning they have a legitimate purpose but are misused by abusers. We propose an approach, iROGUE, for misuse audits that identify rogue apps and uncover rogue functionalities.

We found that the reviews of an app often describe rogue functionalities, its misuse (potential and actual), and the expectations of users. Such reviews are evidence of rogue behavior and can guide audits of such apps. Moreover, app reviews are more valuable than app metadata (such as app descriptions) because metadata indicates only the legitimate purpose of dual-use apps and not the actual misuse. In particular, we propose the following research questions:

$\mathbf{RQ_{info}}$. What information regarding misuse audit is contained in reviews?

$\mathbf{RQ_{identify}}$. How can we conduct a misuse audit through app reviews?

$\mathbf{RQ_{functionality}}$. What rogue functionalities are present in audited apps?

Example 1 shows three reviews (edited for grammar), taken from Apple's App Store Store (2023b) that are relevant to the rogue behavior. Although our study is based on Apple App Store's reviews, iROGUE can be applied to reviews from other sources, including Google's Play Store Store (2023m).

Example 1: Cases relevant to rogue behavior

Fly on the wall!
(for the AirBeam Video Surveillance app Store (2023a))
"with this app, i can spy on my family without them knowing it! it's such an awesome app!"

This app basically ruined my family to an extent
(for the Life360 app Store (2023f))
"My mother made everyone in the family get this app. She freaks out when the app doesn't do its job because of random obstacles that mess with the location accuracy. Drains the battery and makes my parents paranoid to know where I am at all times. I don't even do any bad stuff, yet years of trust building are being swept away by the ability to spy on the children of a household. If you're a parent I highly recommend you don't get this app because it is extremely uncomfortable to have and it makes parents trust their children less."

Honest
(for the 3Fun: Threesome & Swingers app Store (2023k))
"...A lot of the local people I've talked to (Male half of a couple) have been guys who are saying they're part of a couple, and in all reality are single guys just looking to collect pictures. There is no way to report that that is why you are reporting them. It's just a boilerplate report feature. I feel there should be a way for the 3Fun community to point out people for bad behavior like this."

In Example 1, the first review for AirBeam Video Store (2023a), addresses the scenario where the app assists a user in accessing a victim's information without the victim's knowledge. AirBeam Video is a surveillance app to be installed on the abuser's device. Hence, the victim may not be an app user but a bystander. The second review, for Life360 Store (2023f), complains about the problem of inappropriate access to the user's location by the user's mother. Due to the unequal power dynamics between the victim (reviewer in this case) and the abuser (mother in this case), the victim is forced to install apps that violate their privacy. The third review, from 3Fun Store (2023k), describes a story of improper access to profile pictures. Even though the profile pictures are public, the victim is uncomfortable with the access. It is common for users to upload such information (pictures in this case) on an app. When doing so, they hold expectations of how other users would access it. As shown in these three cases, information access may lead to discomfort, fear, or potential harm (Freed et al. 2019; Havron et al. 2019). Thus, despite such cases of information access being common, they should be brought to in notice of app developers and platforms.

2.1.1 Key Findings

We found that reviews contain rich information about apps which is crucial in auditing them. For example, not only negative but positive reviews also reveal an app's rogue behavior. Relevant reviews are written by abusers, victims, and third person, each having a distinct linguistic variation, which makes mining reviews challenging (Section 2.2). Moreover, while reporting rogue behavior, reviews have varying degrees of *convincingness* and *severity*, which can be leveraged in misuse audit (Section 2.3). We found rogue apps exhibit a variety of rogue functionalities, ranging from tracking location to monitoring phone activities such as accessing chats, phone contacts, call history, and so on (Section 2.4).

2.1.2 Contributions and Novelty

Our work's novelty lies in introducing the problem of misuse audits. We leverage app reviews, which despite being a public source of information, still remain unexplored for misuse audits. We contribute to responsible computing by:

- Presenting app reviews as a viable source to mine for rogue behavior.
- iROGUE, an approach based on app reviews for auditing apps.
- A ranked list of rogue apps along with their alarming reviews revealing rogue behavior.

2.1.3 Organization

The rest of this chapter is organized as follows. Section 2.2 describes our preliminary investigation that shows that app reviews contain evidence useful for audits. Section 2.3 describes how to identify rogue apps using our misue audit approach, iROGUE. Section 2.4 shows the procedure to uncover rogue functionalities of audited apps. Section 2.5 lists related work on app reviews and mobile apps. Section 2.6 describes the supplementary materials. Section 2.7 concludes this research project.

2.2 App Reviews for Misuse Audits

We describe how we collected and investigated reviews for misuse audit, which led to several qualitative findings.

2.2.1 Seed Dataset

Chatterjee et al. Chatterjee et al. (2018) identified 2,707 iOS apps as candidates for Intimate Partner Surveillance (IPS), i.e., apps used by a person to spy on their intimate partner. IPS apps are specific to intimate partners and hence are a subset of general-purpose rogue apps. Hence, we start our analysis from Chatterjee et al.'s IPS candidate list.

During our data collection, 1,687 of these 2,707 apps had at least one review on the Apple App Store, leading to *seed dataset* containing 11.57 million reviews. Out of 1687 apps in the seed dataset, only 210 were confirmed IPS by Chatterjee et al.. We include all details of dataset collection in the supplement section (Section 2.6).

2.2.2 Investigating Reviews

Since the seed dataset contains 11.57 million reviews, it is impractical to manually check each review. To find the reviews revealing rogue behavior, we sampled them using *our keywords* (*spy*, *stalk*, and *stealth*). We describe the curation of our keywords in the supplement section (Section 2.6).

There are 5,287 reviews containing at least one of our keywords. From these 5,287 reviews, we randomly sampled 995 reviews for manual scrutiny. This sample involves 179 apps with between 1 and 237 reviews each. We checked 995 reviews for misuse audit. Out of these 995 reviews, we found 402 reviews (of 83 apps in this sample) reporting rogue behavior.

We found that such reviews show variation across two dimensions: *story* and *reviewer*. Based on the story, we found reviews of the following two types:

***Rogue act*:** Reviews describing someone performing a rogue behavior. In such reviews, the reviewer is sure about the app's rogue functionality. For example,

"This app is useless and it just helps overprotected parents spy on their sad kids"

***Rogue potential*:** Reviews express the possibility of rogue behavior. The reviewer may not be sure of rogue functionality but identifies risks with the app. For example,

"...Hate to be a hater. May work well to spy on the kids by 'accidentally' leaving iPhone in secret place."

The above two reviews are negative. However, we also found positive reviews indicating a rogue app. Example 2 shows positive reviews by *abusers*, who brag about the rogue functionalities (history tracking in this case) and sometimes express their delight in it. Other reviews are written by *victims*, who state their concerns and grievances (frustration at the loss of privacy), and some by *third persons* reporting the misuse of the app.

Example 2: Types of reviewers reporting rogue behavior

Abuser

"So much better than <other app>. I love the history feature. My kids say it's creepy and I'm being a stalker. I don't disagree but it is really nice and convenient to be able to keep track of my kids."

Victim

"I hate this app so much! My mother is always questioning me and if I delete it she will ground me … No one wants their parents to stalk them!!"

Third Person

"… I don't feel like parents should track their kids AT ALL. everyone needs a little something called trust and if you don't have it then your kids will act out and have to become sneaky … I do have this app but only with my friends and we don't stalk each other … "

Our linguistic analysis reveals that reviews by abusers are not only positive (higher valence than the other two categories) but also illustrate the abuser's dominance over other parties. Figure 2.1 shows this difference in the valence and dominance scores between all three reviewers (supplementary section describes how these scores are computed). In general, these relevant reviews show linguistic variations, making the automatic mining of reviews a daunting task.

Figure 2.2 shows the relative distribution of 402 relevant reviews (normalized by percentage) across the type of story and reviewer. The third person mostly writes rogue potential stories (73%), whereas abusers and victims dominantly describe rogue acts (99% and 100% respectively).

To sum up, reviews contain rich information about an app's rogue behavior. We found two types of stories that are relevant for auditing such apps. Moreover, such stories are written by three types of reviewers, whose language varies, making the problem of mining these reviews a difficult task.

2.3 Identifying Rogue Apps through Misuse Audit

We propose iRogue, a review-based approach for misuse audits. Figure 2.3 shows an overview of the iRogue approach. First, we collect reviews from the Apple App Store, as shown for the seed dataset in Section 2.2.1. Second, we label a subset of the collected reviews for *alarmingness* (defined below) and train a computational model (Section 2.3.1). Third, we apply our model to all reviews to identify rogue apps (Section 2.3.2). Fourth, we manually examine some of the

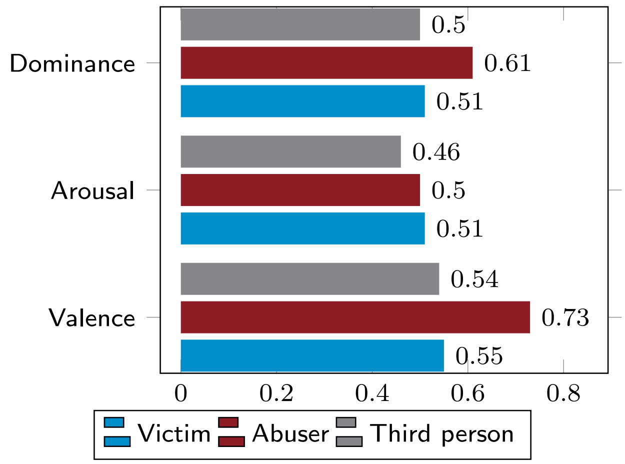


Figure 2.1: Valence, arousal, and dominance scores for three types of reviewers. Abusers' language shows high dominance and valence as compared to the language used by victims and third person. The task of mining reviews becomes challenging due to these linguistic variations.

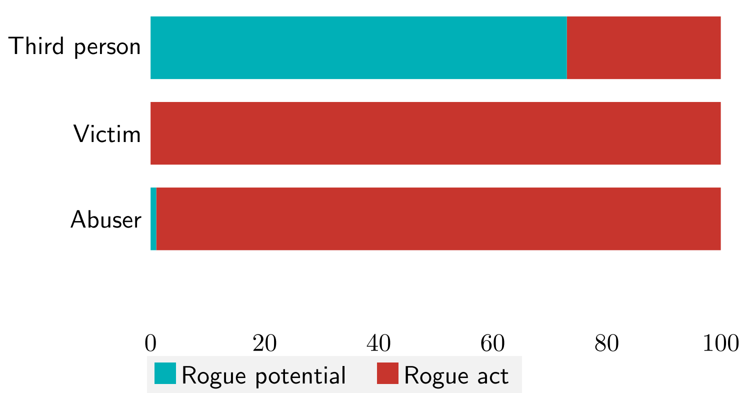


Figure 2.2: Out of 402 reviews manually identified as relevant, abusers and victims mostly write stories indicating rogue acts (~99% and 100% respectively), whereas stories from the third person dominantly represent rogue potential (~73%).

identified rogue apps to find rogue functionalities (Section 2.4).

We envision iROGUE to be incrementally updated by adding reviews of newly found rogue apps. Moreover, apps that don't have reviews, can be audited as soon as their reviews arrive.

2.3.1 Alarmingness of Reviews

We found that reviews exhibit two characteristics: (i) their claims about misuse (we call it *convincingness*) and (ii) the effect of misuse (we call it *severity*). We describe each of these

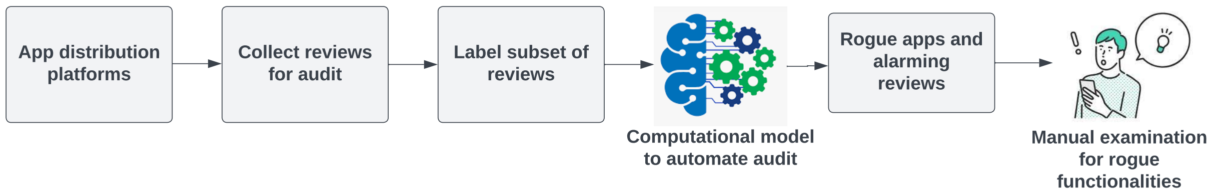


Figure 2.3: Overview of iROGUE.

characteristics below.

Convincingness depends on how strong the rogue claim is. Some reviews report detailed rogue behavior (hence convincing), whereas others are merely suspicions. In Example 3, the first review is unrelated to rogue behavior and hence is not convincing. The second review describes the reviewer's suspicion of the app, which may or may not be true (slightly convincing). The third review (by an abuser) confirms the rogue behavior but lacks details of the rogue functionalities or victims. On the contrary, other reviews (in Example 3) are extremely convincing because they confirm rogue behavior along with mentioning the location feature, how to set up devices, or the victims being stalked. Extremely convincing reviews also include cases when the app is used for positive purposes (tracking family members or pets for safety) but has the potential to be misused in the future. The reviews that are slightly, moderately, or extremely convincing are relevant to misuse audits. Assigning a convincingness score helps in ranking all reviews according to the strength of their claims.

Severity measures the effect of rogue behavior on the victim. Example 4 shows the range of reviews varying in severity. The first review is unrelated to rogue behavior. Thus, it is not severe. The second review shows that the rogue act is performed with consent, making this review a slightly severe case. The third review is written by the abuser and lacks the victim's perspective to analyze the rogue effect. We assume such acts are performed without consent and consider them moderately severe. The fourth review describes the victim's misery. The victim even says "This app has truthfully ruined my teenage years" in the review, which gives solid evidence to be an extremely severe case. Moreover, in the fifth review, the victim complains that others can see when he was last active (also known as last seen information). This is the public information on each profile, but still, the victim is uncomfortable with the access. App developers should be aware of such misuse. Since app reviews discuss these cases, we can identify and rate them as extremely severe.

Example 3: Varying degree of convincingness

1: Not convincing
"It is such a great game, love it so much!"

2: Slightly convincing
"Setup was a breeze. Quicktime 7 pro found it easily. Unfortunately, resolution seems much, much, lower than hoped. Video size can not be adjusted live. Hate to be a hater. May work well to spy on the kids by 'accidentally' leaving iPhone in secret place."

3: Moderately convincing
"This app is perfect for stalking people... "

4: Extremely convincing
"This app is awesome for our family to keep track of where everyone is at all times! (You can turn the location off too in case you want to be in stealth mode when buying Christmas presents too.) ...Even our dog knows that the alert sound when a family member arrives home means ... "

"...I use it to spy on my dogs while I'm at work; so I use it for fun, nothing fancy. My iPad is my camera, and my iPhone is my viewer. ... "

"bro this app is high key creepy. when i'm with my dad on his days my mom even mentions how she knew everything i was doing and it even made my dad creeped out. if you need this app then ngl yo wack. i don't want my mom stalking me."

Example 4: Varying degree of severity

1: Not severe
"Love the graphics so far it is a great game"

2: Slightly severe
"I love this app, just great because you can time your day accordingly, I like my girlfriend knowing where I am and I love stalking her, we have fun with it.... "

3: Moderately severe
"This app is perfect for stalking people... "

4: Extremely severe
"honestly if you want your kid to rebel against you even more, this is the app for you! This app

has truthfully ruined my teenage years all because my mother now has a way of tracking me down 24/7. I couldn't do the normal teenage things because I was being stalked all day ... "

"... i want to share my last seen just to my family and my girlfriend not others. please add new feature in privacy that i can share my last seen to no body except my family and girl friend thanks soo much !"

To capture the above two characteristics, we define *alarmingness* of a review, which is the geometric mean of its convincingness and severity. An app can receive a large number of reviews. Unlike binary classification of reviews, the alarmingness score not only identifies relevant reviews for audit but also ranks them based on rogue behavior. For an app, the higher the alarmingness of the review, the more solid evidence it provides about the rogue behavior and hence more useful in auditing.

We manually annotated a subset of reviews for alarmingness and used the annotated data for training. From the seed dataset, we selected two sets: 952 randomly selected reviews (set s_1) that contain at least one of our keywords and 932 randomly selected reviews (set s_2) that do not contain any of those keywords. Having both sets s_1 and s_2 made our training data unbiased toward any keywords.

From these 1,884 reviews, we excluded the reviewers' identifiers such as their usernames. Two authors rated these reviews for convincingness and severity on a four-point Likert scale (1: not, 2: slightly, 3: moderately, 4: extremely). The supplementary section (Section 2.6) describes the details of the annotation process. For reviews that were rated by the two annotators, we computed the average convincingness and severity scores. This annotation study possessed minimal risk and was exempted by the Institutional Review Board (IRB) of our university. We used these annotated reviews as our training data.

We extracted linguistic features of 1,884 reviews through the Universal Sentence Encoder (USE) Cer et al. (2018), a widely used approach that is proven effective for app reviews (Guo and Singh 2020; Devine et al. 2021). Using USE features, we trained various multi-target (convincingness and severity scores were targets) regression models Borchani et al. (2015) and found that the Support Vector Regressor (SVR) outperforms. The supplementary section (Section 2.6) provides all the training details and evaluation.

We leveraged trained SVR to predict convincingness and severity scores of all 11.57 million reviews in the seed dataset. The alarmingness of each review is calculated by taking the geometric mean of its predicted convincingness and severity.

2.3.2 Identifying Rogue Apps

Using statistical methods, we aggregated the alarmingness of reviews and ranked all apps according to their reported misuse (details in Supplementary Section 2.6). From the seed dataset, our model identified a total of 100 rogue apps (model evaluation in Supplementary Section 2.6). Moreover, our approach is not dependent on the choice of candidate apps and could be applied to any set of apps. To audit more apps, we applied our model on (i) a dataset of similar apps and (ii) a dataset of 100 popular apps in the utility category. We found not only IPS but many general-purpose rogue apps exhibiting multiple rogue functionalities, also described in the next section.

Similar Apps

We retrieved 896 apps (similar to 100 predicted rogue from the seed dataset), using the Apple App Store's recommendations ("You May Also Like" section). Our motivation for using Apple recommended apps was that these apps should offer functionalities similar to those in 100 predicted apps. Further, we collected 2,652,678 reviews of these 896 apps which we call *snowball dataset*.

Our model predicts 138 rogue apps from the snowball dataset. Examples 5 shows alarming reviews of one such app from the snowball dataset, Smart Family Companion App Store (2023j).

Our evaluation of the snowball dataset reveals that our model yields the best recall of 77.27% as compared to all other baselines (discussed in the Supplementary Section 2.6). In this scenario, a false negative can lead to harm through misuse, whereas a false positive just wastes effort in an unnecessary audit. Hence, high recall is more valuable than high precision.

Example 5: Rogue from snowball dataset

App: Smart Family Companion Store (2023j)
Rogue Score: 2.35/4.00

Alarming Review 1
Alarmingness: 4.00/4.00
Date of Review: 2020-03-15
"How is this even ethical? To put out an app in which you can completely control what's going on on someone else's phone? It's a huge privacy concern. To be honest, apps like this shouldn't exist. It's one thing to put control on a YOUNG CHILD'S phone (which can be done in settings easily) put to put this on an older kids phone is going to destroy trust. No parent should be able to see what their child is doing on their phone 24/7. It's borderline abusive."

Alarming Review 2
Alarmingness: 4.00/4.00
Date of Review: 2019-09-03
"... This app tracks every last thing your child does on their phone. As you can imagine, no 16-year-old wants to have their own private life constantly exposed to you. Just because you are their parent, and you live together, doesn't mean that they have to share everything with you. Location, data usage, browsing history, etc. frankly aren't your business. That's their own private information that you don't need to know. Coming from a family with control issues, there is no better way to destroy your relationship with your children. I doubt anyone would want to be around someone who is constantly monitoring and controlling them ... "

100 Popular Utility Apps

Surveillance apps that can be misused for spying, fall under the "Utilities" category, making it an important category to audit. We considered 100 popular utility apps (mentioned on the Apple App Store page Store (2023l)) and collected their reviews.

From these 100 apps, our approach, iROGUE, predicts only one app as rogue, which after reviews' scrutiny by us, comes out to be non-rogue. We also scrutinized some apps predicted as non-rogue and found their predictions true. We report all details of this analysis in the Supplementary Section 2.6.

We found that any generalized set of apps can contain a wide range of subcategories such as payment apps, calculators, and so on, which are less likely to be misused. Hence, they don't form good candidates for identification. Iteratively auditing apps (through iROGUE) similar to the already identified rogue apps (as shown in Section 2.3.2), is a feasible solution to uncover a large rogue landscape.

2.4 Uncovering Rogue Functionalities

We illustrate the idea of uncovering rogue functionalities using top 40 ranked apps (according to iROGUE) in the seed dataset. First, we analyzed an app's description to know the basic functionalities. Second, we analyzed 10 most alarming reviews (discovered by iROGUE) for identifying existing misuse and rogue functionalities. Third, we installed a few apps on a device to validate the claims in reviews. Through this process, we discovered the following types of rogue functionalities:

Monitoring phone activities. Some apps monitor a victim's phone activities, such as browsing history and text messages. Such apps are installed on the victim's device and activities can be monitored on another synced device. SaferKid Text Monitoring App Store (2023i) is an example of such an app.

Audio or video surveillance. Some apps enable audio or video surveillance without the victim's knowledge. These apps listen, view, or record a victim's voice or actions and some of them need not be installed on the victim's phone. Find My Kids: Parental control Store (2023c) contains such functionality.

Tracking location. Some Global Positioning System (GPS) apps enable tracking a victim's phone, with (forced consent) or without their knowledge. For example, Find My iPhone Store (2023d).

Profile stalking. Some apps are misused for stalking user profiles or user content (such as images), making the victim uncomfortable about the misuse. For example, Kik Messaging and Chat App Store (2023e).

Table 2.1 shows these four types of rogue functionalities, and alarming reviews reporting them. Some reviews in Table 2.1 are old (2014 or 2012), but we confirmed that similar concerns are being raised in the recent reviews of the same apps. For example, the Find My iPhone Store (2023d) app still lets its users see the location of the connected devices. Due to unequal power dynamics between abuser and victim (say in a family setting) Chatterjee et al. (2018), the victim is sometimes forced to connect to such apps and allow their device to be located. In other words, legitimate apps designed for locating loved ones can be misused for rogue behavior, hence they are dual-use Chatterjee et al. (2018).

We also illustrate the rogue behavior of the SaferKid Text Monitoring App Store (2023i) by installing it on two devices: a parent's device (iOS version 14.4.1) and a child's device (Android version 11.0). Figure 2.4 shows the rogue features present in this app such as monitoring text messages, web history, and call history. Activities on the child's device can be monitored on the synced parent's device. We verified each of these functionalities. Figure 2.5 shows the screen displaying all chats of the victim. Apps such as SaferKid are advertised as safety apps for children but can be secretly or forcefully installed on another device to monitor the user's activity. Not only parents, but any individual can misuse such apps by installing them on the victim's phone.

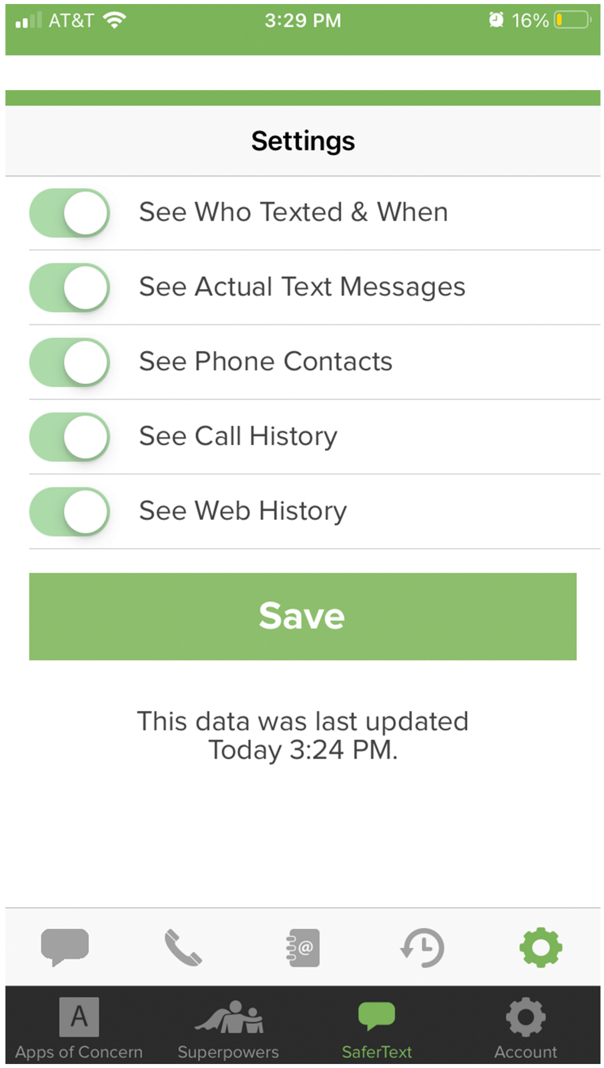


Figure 2.4: The SaferKid Store (2023i) app provides multiple ways to monitor victim's phone activities.

2.5 Related Work

We describe previous works focusing on auditing apps, app reviews, and spying through mobile apps.

2.5.1 Auditing and App Reviews

Most of the research in software audits analyzes the flow and dependencies in the source code (García et al. 2021; Xia et al. 2015; Perl et al. 2015; Pashchenko et al. 2020). Only a few audit studies focus on other metadata such as system trace Wang et al. (2015). However, all these studies miss identifying rogue apps because their analysis doesn't address cases of forced

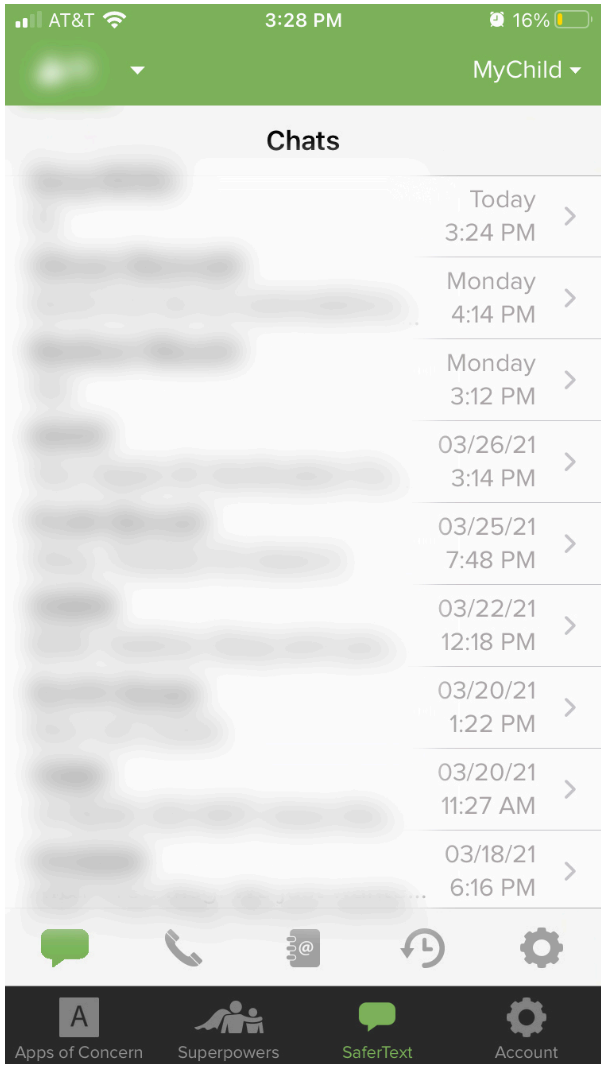


Figure 2.5: Using the SaferKid Store (2023i) app, victim's chats can be seen on the abuser's phone.

consent and inappropriate access to public information. Our work leverages app reviews (unexplored by prior studies) for misuse audits.

Prior studies show that reviews offer valuable feedback for the app's improvement. Dhinakaran et al. (2018) mine requirement-related information from app reviews and propose an active-learning framework to fetch relevant ones in a semi-supervised manner. Haque et al. (2022) leverage reviews to draw implicit and explicit comparisons between competing apps. Guo and Singh (2020) extract <action, app problem> pairs from the stories expressed in reviews, which can be useful for the developers to work on. Chen et al. (2014) showcase the most relevant reviews to developers by grouping informative ones and applying unsupervised techniques. For developers to learn from feature suggestions and bug reports, Kurtanović and

Table 2.1: Types of rogue functionalities.

Rogue Functionality	App Example	Alarming Review
Monitoring phone activities	SaferKid Text Monitoring App Store (2023i)	*… Tracking things like social media, texts, and search history is just a complete disregard of privacy. You have to have trust in your kids … Apps like these shouldn't be allowed. IF YOU TRUST YOUR KID, DONT DOWNLOAD.* (Date: 2019-12-07)
Audio or video surveillance	Find My Kids: Parental control Store (2023c)	*This app proves to have a invasion of privacy. Due to the fact if your kids was at a friends house and talking to his friends parents, this app records what is going on and is a invasion of privacy. If your child left their phone downstairs or anywhere and they are playing it can record private conversation between adult and is a unsafe …* (Date: 2019-01-16)
Tracking location	Find My iPhone Store (2023d)	*… It's supposed to be used to recover a lost phone, not to religiously stalk your children.… The fact that a mom actually installed this app onto her son's phone without his knowledge is flat out wrong. … If you're constantly monitoring your child 24/7, just imagine what your child will do when they go off to college. …* (Date: 2014-02-13)
Profile stalking	Kik Messaging and Chat App Store (2023e)	*MAKE THIS SAFE PEOPLE WANT TO USE IT BUT DON'T WANT TO BE STALKED OR ABUSED IN THE APP PROTECT THE APP OR PEOPLE WON'T USE IT MY FRIEND HAD A MAN SEND HER A BAD PICTURE IF YOU KNOW WHAT I MEAN. OVERALL THIS APP IS GREAT!!!!!"* (Date: 2015-06-24)

Maalej (2017) filter rationale-backed reviews through classification. Moreover, some studies (Nguyen et al. 2019; Besmer et al. 2020) mine reviews to know users' perceptions about apps selling their data. However, to the best of our knowledge, app reviews remain unexplored for misuse by app users, basically user-based spying or stalking scenarios.

2.5.2 Spying through Mobile Apps

Chatterjee et al. Chatterjee et al. (2018) identify IPS apps with carefully designed search queries and manual verification based on app information. They leverage information such as app descriptions and permissions. However, for dual-use apps, the actual usage deviates from the intended purpose shown in app descriptions. To identify such misuse, we focus on the evidence provided in app reviews.

Roundy et al. Roundy et al. (2020) focus on identifying apps used for phone number spoofing and message bombing, which lie outside the scope of rogue apps. Conversely, rogue apps include those that enable stalking public information, which are outside their scope. Roundy et al. use metadata such as installation data, to uncover spying apps that are installed on infected devices. However, we focus on evidence of rogue behavior present in app reviews, to uncover rogue apps. Roundy et al. rely upon Norton's security app Store (2023g) to determine which devices are infected. Thus, their approach would miss apps that a general user can leverage to spy.

All these studies along with a few others (Tseng et al. 2020; Zou et al. 2021; Tseng et al. 2021) investigate how technology is abused for spying. However, they are only limited to IPS and not the broader set of rogue apps. In particular, they do not consider cases when the victim is uncomfortable with the access (of public information) even if aware of it, along with cases involving friends and strangers as abusers.

2.6 Supplementary Materials

2.6.1 Seed Dataset

Chatterjee et al. Chatterjee et al. (2018) identified 2,707 iOS apps as potential Intimate Partner Surveillance (IPS) apps, i.e., apps used by a person to spy on their intimate partner. Out of these 2,707 apps, Chatterjee et al. confirmed 414 apps to be IPS, using semi-supervised pruning.

When we collected our data (from 2008-07-10 to 2020-01-30), 1,687 (out of 2,707) apps existed with at least one review on the Apple App Store. This yielded our *seed dataset* containing 11.57 million reviews. Out of 1687 apps in the seed dataset, only 210 were confirmed IPS by Chatterjee et al..

2.6.2 Our Keywords

For the formation of a list of keywords, we initialized a set with the words: *spy*, *stalk*, and *stealth* and queried WordNet Miller (1995) for their synonyms. We performed the query operation until we didn't find any new word in the set. The resulting set contained keywords: *spy*, *stalk*, *stealth*, *descry*, *chaff*, and *haunt*. However, *chaff* and *haunt* weren't relevant for describing rogue behavior in reviews. Also, *descry* was present in only two reviews, both of which were irrelevant to rogue behavior. To expand the set of keywords, we explored other corpora such as PyDictionary PyDictionary (2023) and Thesaurus Pythesaurus (2023) but did not find synonyms that are widely used in app reviews. Moreover, keywords used in the previous study Chatterjee et al. (2018), such as *track* and *control* bring many false positive reviews. For example, "I like tracking my distance when I walk with my dog." and "...you can also control the audio of your mac through the app ... I can control music tracks without having to touch the computer." are not relevant. Thus, our relevant set of keywords reverts to *spy*, *stalk*, and *stealth*.

2.6.3 Linguistic Analysis

To compute valence, arousal, and dominance scores, we leveraged lexicons provided by Mohammad (2018), which contains scores for about 20,000 English words. First, from each review, we extracted relevant sentences using our keywords. Second, using parts of speech tagging, we extracted adjectives mentioned in the relevant sentences. Since sentences describe the rogue behavior, adjectives in them represent the reviewer's sentiment. For example, *"This horrible (adj) app spy on me"*. Third, for each type of reviewer, we computed average scores of adjectives that are present in the lexicon.

2.6.4 Annotating Reviews and Model Training

The annotation was conducted in three steps. First, two researchers of this project rated 599 of 1,884 selected reviews according to the initial set of annotation instructions. The initial instructions included definitions (of convincingness and severity scores) and examples corresponding to each point on the Likert scale. In this step, for each annotator, we calculated the alarmingness scores of reviews using the convincingness and severity ratings. If the alarmingness scores computed for both the annotators were not at least three (median value on Likert scale), or both are not less than three, annotators resolved such cases via discussions. After discussing, the annotators produced the final set of annotation instructions. In the second step, the annotators followed the final instructions and rated 900 reviews. In this step, we computed an Intraclass Correlation Coefficient (ICC) Hallgren (2012) to measure inter-rater agreement. We found ICC(3,k) to be 0.9195 for convincingness and 0.9190 for severity, which accounts for excellent agreement. ICC is suitable for Likert scale ratings. Unlike other measures such as Cohen's Cohen (1988) kappa, which is based on (all or nothing) agreement, ICC takes into account the magnitude of agreement (or disagreement) to compute inter-rater reliability. In the third step, the remaining reviews were divided among the two annotators so that only one annotator rates each review.

Extracting Deep Learning Features from App Review

We obtained the feature vector of each app review as follows:

Combine Sentences: Remove periods in each app review and combine all its sentences to form single sentence.

Text Preprocessing: Remove all punctuation marks, stop words Uysal and Gunal (2014), and our keywords, the latter because those keywords may correlate with reviews with higher scores and could create bias in the model.

Sentence Embedding: We leverage the Universal Sentence Encoder (USE) Cer et al. (2018) to extract embeddings for each app review. USE uses Deep Averaging Network (DAN) to provide a 512-dimension embedding for a long text. USE is trained on a large variety of natural language tasks with the aim of capturing the context. In our case, USE directly provides sentence level embeddings of an app review, by keeping the context intact. However, alternatives such as GLoVe Pennington et al. (2014) and Word2Vec Mikolov et al. (2013) lose such context. We leverage the pretrained USE network by using Tensorflow Hub Flow (2023).

We evaluated the performance of three regression models: support vector regression Basak et al. (2007), random forest Smith et al. (2013), and decision tree Xu et al. (2005), by ten-fold cross-validation on our dataset. To mitigate bias of our keywords, we remove such keywords in the preprocessing step, so that the regression model learns from the the context of the review and not from specific keywords. As shown in Table 2.2, the Support Vector Regressor (SVR) yields the smallest MSE of 0.625 with 0.458 standard deviation, so we choose it for the subsequent phases of our approach.

Table 2.2: Performance of three regression models on ten-fold cross-validation.

Regression Model	Average MSE	Standard Deviation
Decision Tree	1.344	0.402
Random Forest	0.712	0.417
Support Vector	**0.625**	0.458

2.6.5 Statistical Methods for Identifying and Ranking Rogue Apps

Weighted Mean of Alarmingness: In general, for a rogue app, a small proportion of reviews report rogue behavior. Thus, we need to catch rogue apps using their few reviews that have high values on the alarmingness scale. Thus, we assign weights to reviews based on their alarmigness, as follows:

Defining score buckets: While annotating reviews, we defined levels of convincing reviews (not convincing to extremely convincing) and severe reviews (not severe to extremely severe) on a Likert scale. We also follow the same levels on the alarmingness scale (1: not alarming to 4: extremely alarming). We define a score bucket between every consecutive level of alarmingness (not alarming to slightly, slightly alarming to moderately alarming, moderately alarming to extremely alarming). Table 2.3 shows how score buckets are formed using levels of alarmingness.

Assigning weights to score buckets: We have 11.57 million reviews in the seed dataset. Based on the alarmingness computation, we calculated the probability of a review falling into a score bucket. Since the reviews reporting rogue behavior are less, probabilities in buckets 2 and 3 are less than that in bucket 1. We take the inverse of these probabilities to get the weights for each score bucket. As a result, we assign higher weights to buckets 2 and 3 than to bucket 1. Table 2.3 also shows the weight assigned to each score bucket.

Table 2.3: Score buckets for alarmingness.

Alarmingness Score Range	Alarmingness Level Range	Bucket	Bucket Weight
[1,2)	Not alarming to Slightly	1	2.29×10^{-3}
[2,3)	Slightly to Moderately	2	6.08×10^{-2}
[3,4]	Moderately to Extremely	3	9.36×10^{-1}

If a_1, a_2, ..., a_n are the alarmingness scores of an app's reviews, and w_1, w_2, ..., w_n are their respective weights (according to Table 2.3), then, $W_\text{alarmingness}$, the weighted mean of alarmingness is given by:

$$W_\text{alarmingness} = \frac{a_1 * w_1 + a_2 * w_2 + \dots a_\text{n} * w_\text{n}}{w_1 + w_2 + \dots w_\text{n}}$$

The weighted mean of alarmingness ranges from 1 to 4.

Normalized Count: The weighted mean of alarmingness does not account for the count of reviews that report rogue behavior against an app. Suppose, *app A* has 15 reviews reporting rogue behavior and *app B* has 25 reviews reporting rogue behavior. If all reviews reporting rogue behavior have the same alarmingness score, the weighted means of the two apps would be the same. But, *app B* shows more evidence of rogue behavior and should have a higher rogue score than *app A*. Thus, we also consider the count of reviews. For each app, we calculate the number of reviews in bucket 3. We tried incorporating counts of other buckets, but it led to worse performance of the approach.

The minimum possible value of the count is zero. However, in some cases, counts can be high, leading to no definite upper limit. Thus, we normalize the counts of all the apps between one and four.

We want to assign high rogue scores to apps that have high scores in both (i) weighted mean of alarmingness and (ii) normalized count. Thus, the rogue score is computed as the geometric mean of these two values.

Selecting Threshold for Prediction

For each app in the seed dataset, iROGUE computes the rogue score. All apps are ranked in decreasing order of rogue scores. The apps with a score greater than a threshold are predicted rogue. To decide the correct threshold, we follow two steps: (i) label the ground truth of rogue apps and (ii) vary a threshold between certain values and choose the threshold that gives the best performance of iROGUE.

We create our ground truth by manually scrutinizing reviews. However, scrutinizing reviews of all 1,687 apps (seed dataset) is not feasible. Thus, first, we scrutinize the 50 most alarming reviews (with a minimum score of two—at least slightly alarming) for apps with the highest 100 rogue scores. Second, we scrutinize reviews containing our keywords for these 100 apps. The first step is aligned with our approach since it checks the top alarming reviews. However, the second step is neutral because it searches for evidence for the apps that iROGUE failed to identify through top alarming reviews. This way we mitigate the threat of bias while curating ground truth for apps with the highest 100 rogue scores.

We label an app as rogue provided any of the scrutinized reviews report a rogue behavior. Table 2.4 shows the types of reviews we consider indicative or otherwise of rogue evidence. If the reviews of an app describe information access performed without the victim's knowledge or when the victim shows discomfort (first three reviews in Table 2.4), we consider the app as a rogue. Also, some rogue apps are used for positive purposes such as tracking family members for safety (fourth review in Table 2.4) but still possess a potential for future misuse. Similarly, apps used for tracking pets or other objects are considered rogue (fifth review). Reviews of some apps don't possess any misuse in the present or in the future, leading to their final label of not rogue (sixth review). Through manual inspection, we determine that of the 100 apps, 73 are rogue and 27 are not.

Table 2.4: Instructions followed for labeling rogue apps.

Type of case	Subtype of case	Example	Rogue?
Tracking people's information	Without the victim's knowledge	*"Now that I can spy on my wife I will always know when she is cheating"*	Yes
Tracking people's information	With the victim's knowledge but with discomfort	*"Ok my mom got this for me and ... it's kinda creepy that this app was made so parent could basically stalk their kids."*	Yes
Tracking people's information	Public information but the victim is uncomfortable	*"I had someone cyberstalking and harassing me. Multiple attempts in every way shape and form were made to contact app-name to block and ban the stalker's account due to a concern for my well- being."*	Yes
Tracking people's information	Positive purpose	*"I love finding my family members. Wife was in bad car wreak and I was able to find her location using this app. Thank you!"*	Yes
Tracking pets or other objects		*"Wow! Day one and I'm stalking my puppet like a soccer mom that ran out of adderall! I'm very excited to use this to interact with my puppet while I'm at work and to check in on the dog walker!"*	Yes
Not related to information accessing		*"an absolutely amazing and very helpful app. i don't know how i would keep track of prayer times without it. love the app. thank u!!!"*	No

For the 100 apps, the rogue score varies between 1.74 and 3.60. We vary the threshold from 1.73 to 3.59 in steps of 0.01. Apps with a rogue score above the threshold are predicted rogue but not rogue otherwise. At each value of the threshold, we report the recall, precision, and F1 score.

Table 2.5 shows the performance achieved at specific thresholds. As we increase the threshold, the precision increases at the cost of recall. For rogue apps, a false negative costs more than a false positive because a false negative leaves a rogue app undetected, which can harm many victims, whereas a false positive causes only wasted effort in manual scrutiny. Hence, achieving high recall is more important than achieving high precision. Thus, from Table 2.5, we choose 1.73 threshold that gives the best recall of 100% at 73% precision. Since we fine-tune the threshold on the same seed dataset, we also check iROGUE's performance (using the chosen threshold) on the other dataset (Section 2.6.6).

Table 2.5: Choosing an appropriate threshold according to the recall scores.

Threshold	Precision (%)	Recall (%)	F1 Score (%)
1.73	**73.00**	**100.00**	**84.39**
1.74	73.40	94.52	82.63
1.75	76.13	91.78	83.22
1.76	76.82	86.30	81.29
1.77	76.92	82.19	79.47
1.78	78.37	79.45	78.91
1.79	79.71	75.34	77.46
1.80	80.95	69.86	75.00
1.81	80.95	69.86	75.00
1.82	81.96	68.49	74.62
1.83	81.03	64.38	71.75
1.84	80.35	61.64	69.76
1.85	82.69	58.90	68.80
1.86	83.33	54.79	66.11
1.87	82.97	53.42	65.00

Example 6: Rogue app from seed dataset

App: Find My Family & Friends Store (2023f)

Rogue Score: 3.60/4.00

Alarming Review 1

Alarmingness: 4.00/4.00

Date of Review: 2019-11-28

"...Such a terrible thing for unaware parents to use. <u>Most parents think teens don't need privacy and they constantly need to know where they are and what they're doing and who they're with at all times.</u> This may make the parent feel at peace but what about the child?

It's selfish of parents to not take into consideration of how the teen may feel about always having this app and the parent giving them a very stalkish feeling, it's very uncomfortable."

Example 7: Rogue app from seed dataset

App: OurPact Jr. Child App Store (2023h)
Rogue Score: 2.47/4

Alarming Review 1
Alarmingness: 4.00/4.00
Date of Review: 2018-06-19
"...however this app shuts down almost everything and can see every text and website you've visited. now, i haven't done anything bad online (recently), but i find that a little creepy and honestly an invasion of privacy. no wonder this app has such crappy reviews. also, i used to have way more apps than i do now. because my parents now have the ability to restrict apps that may be "inappropriate". i already have to ask permission to download apps, so if they were inappropriate my parents wouldn't let me download them. there's too many apps like this and i think kids need a break from all this crap on their devices."

Alarming Review 2
Alarmingness: 4.00/4.00
Date of Review: 2018-08-16
"This is a useless app that no parent need to install I pray for every child who has this app installed on their electronics some parents don't understand the modern society but that's okay (but not really) I'm only given 2 hours and writing this review is using up time WHICH IS NOT FRIKEN OK!!! I hate this hate this app and I hope every child that has had their device attacked by this installment hates this app as much as me. This app should never be okay to use its inappropriate and everybody's children who have this app installed are making there children ANTI-SOCIAL AND VERY NOT COOL. I have many reasons why this app is SOOOOOOO scaring and dreadful so if your reading and thinking about installing this on ur child's device DONT INSTALL IT because that will ruin their future."

Performance of Baseline Methods

On the seed dataset, we also check the performance of the baseline methods described below.

Our Keywords on App Description. We search for the presence of one of our keywords (*spy*, *stalk*, and *stealth*) in app descriptions. Apps whose descriptions contain any of these keywords are predicted rogue, whereas other apps are predicted as not rogue.

Extended Keywords on App Description. We identify additional relevant keywords by extracting verbs through Part-Of-Speech (POS) tagging Manning (2011). POS tagging marks every word in a sentence to an appropriate part of speech (verb, noun, adjective, and so on). Applying this process on descriptions of 73 rogue apps (from the ground truth) produced 145 verbs, out of which six (*track*, *monitor*, *locate*, *control*, *stolen*, *lost*) we selected as relevant to rogue behavior. The verbs: *stolen* and *lost* are relevant because they describe the apps that are used to find a misplaced phone, which indicates an ability to track another device. We extend our keywords by adding these six verbs. Apps whose descriptions contain these keywords are predicted rogue.

T% Keyword Reviews. For each app, we compute the percentage of reviews containing our keywords. We set a threshold, T, on this percentage, above which apps are predicted rogue. In our evaluation, T takes the values of 0.3, 0.2, and 0.1, respectively.

Table 2.6 summarizes the precision, recall, and F1 scores of all baselines and our approach. Our keywords on the description predict only one rogue app, leading to 100% precision (the highest among all). However, our keywords miss 72 rogue apps, which leads to the worst recall of 1.36%. Among all the baselines, keyword search on reviews with 0.1% threshold achieves the highest recall of 65.07%, which is much lower than iROGUE's recall value. iROGUE's better performance may be due to fine-tuning iROGUE's threshold on the same seed dataset. Thus, we also compare iROGUE's performance with these baselines on the other dataset (Section 2.6.6).

Table 2.6: Performance (in %) of baseline methods and iROGUE on the seed dataset. A bold value for a metric indicates the highest score among all approaches.

Method	Recall	Precision	F1
Our keywords on app descriptions	01.36	**100.00**	02.68
Extended keywords on app descriptions	61.64	80.35	69.76
0.3% keyword reviews	46.03	96.66	62.36
0.2% keyword reviews	50.79	96.96	66.66
0.1% keyword reviews	65.07	95.34	77.35
iROGUE	**100.00**	73.00	**84.39**

Examples 6 and 7 show alarming reviews of Find My Family & Friends App Store (2023f) and OurPact Jr. Child App Store (2023h), which iROGUE correctly identifies as rogue. Both of them are dual-use apps. Find My Family & Friends is a safety app, but alarming reviews report parents misusing the tracking functionality on children, which children are uncomfortable with. Moreover, the alarming reviews of the OurPact Jr. Child App report that parents can monitor children's texts and visited websites by installing the app on the child's device.

2.6.6 Performance on Snowball Dataset

In the snowball dataset, to curate the ground truth, we follow the same labeling process as described in Section 2.6.5, for the apps with the highest 200 rogue scores. That's how we label 132 apps as rogue.

Table 2.7 shows the performance of all baseline methods and iROGUE on the snowball dataset. Our keywords when used on descriptions predict only one app as rogue, leading to the lowest recall. This is because, on Apple App Store Store (2023b), dual-use apps are not advertised using keywords: *spy*, *stalk*, and *stealth*. The same approach achieves 100% precision but high recall is desirable in the context of rogue apps.

On app descriptions, extended keywords perform better (68.18% recall at 85.71% precision) than our keywords due to commonly used words (such as *track*, *locate*) in app descriptions. Our keywords are applied on reviews (rows 3–5 in Table 2.7), and discover evidence of rogue behavior.

Table 2.7: Performance (in %) of baseline methods and iROGUE on the snowball dataset. A bold value for a metric indicates the highest score among all approaches.

Method	Recall	Precision	F1
Our keywords on app descriptions	00.75	**100.00**	01.48
Extended keywords on descriptions	68.18	85.71	**76.26**
0.3% keyword reviews	41.66	91.66	57.29
0.2% keyword reviews	44.69	92.18	60.20
0.1% keyword reviews	51.51	88.31	65.07
iROGUE	**77.27**	73.91	75.55

2.6.7 Scrutinizing Utility Apps

We consider 100 popular utility apps that are mentioned on the Apple App Store page Store (2023l). Out of 100 apps, nine are already scrutinized either in the seed or snowball dataset. For the rest 91 apps, we retrieved 392,928 reviews, over the duration of 2008-10-18 to 2022-08-04.

iROGUE predicts only one app as rogue, which after reviews' scrutiny by us, comes out to be non-rogue. We also scrutinize 10 apps with the highest rogue scores, by reading their top 50 alarming reviews and reviews containing our keywords. But, none of them are rogue.

2.7 Discussion

We proposed iROGUE, an approach to automatically analyze app reviews for misuse audits. iROGUE, first, collects reviews from the Apple App Store Store (2023b). Second, it trains a computational model to identify alarming reviews. For auditing, the identified reviews were leveraged to flag rogue apps and find their rogue functionalities. In total, iROGUE predicts 239 rogue apps from multiple sources, leading to the best recall, as compared to other baseline methods.

Below, we describe our data availability, threats to validity, and promising future directions.

2.7.1 Data Availability

During computation, we used Tensorflow Hub Flow (2023) for extracting features and Scikit-learn library Pedregosa et al. (2011) for training the regression model. We will release all alarming reviews associated with each app, as an open dataset. Our annotated reviews along with the computational model will also be shared.

2.7.2 Reporting to App Store and Developers

We informed the Apple App Store and app developers about the potential misuse of their apps as indicated in alarming reviews. Our manual search found that 23 apps are already deleted from the Apple Store, which may be because of raised privacy concerns or other unknown reasons. Moreover, we heard back from 12 developers about their concerning misuse. Eight of them confirmed that their privacy teams will take necessary steps to curb misuse, whereas two developers assumed that app users (including abusers) will never misuse their apps (non-realistic assumption) and the other two developers were not open to any feedback.

2.7.3 Threats to Validity

We now discuss the threats we identify in our work. The identified threats are of two types: (i) the threats that we mitigate; and (ii) the threats that remain.

Threats Mitigated

We mitigated the following threats to validity. First, reviews in the set s_1 contained our keywords. This may create a bias in the model to predict high scores for only reviews having our keywords. To mitigate this threat, we added s_2 in the training set. This helped the model to learn from the context and not from specific keywords. Second, review annotation by crowd workers could yield incorrectly rated reviews, because of their inability to understand the problem well. Thus, two researchers of this project annotated the whole training data. Third, the ground truth (of rogue apps) could be biased if it was formed only using top alarming reviews. We mitigated this bias by scrutinizing reviews containing our keywords, which can contain evidence missed by alarming reviews. Fourth, the ideal scrutiny of apps will be done after installing all on the phone and testing them, which is an onerous and time-consuming task. Moreover, many apps show problematic functionalities in their paid versions. Hence, many audit studies Pashchenko et al. (2020); García et al. (2021); Wang et al. (2015) rely on automated methods instead of installation. We rely on actual misuse reported in app reviews.

Threats Remaining

Now, we describe the threats that remain in our work. First, we investigated only a few thousand apps, which may not be representative of all apps on the Apple App Store Store (2023b). The performance of iROGUE may vary while testing it on all apps of the Apple App Store. Second, we targeted apps and their reviews only on Apple's App Store. Upon deployment on other app stores, the performance of our approach can differ. Third, if an app distribution platform does not have a similarity recommendation, iROGUE can be prioritized for the apps that are flagged by app users (victims). Fourth, some negative reviews about rogue behavior may be written by the app's competitors. Identifying fake reviews is a challenging task and out of the scope of our study.

2.7.4 Limitations and Future Directions

We identify the following limitations of this work. Each limitation gives rise to possible future work. First, iROGUE may miss some rogue apps if they do not have alarming reviews at the time of analysis, possibly because they are new apps. However, such rogue apps can be identified as soon as alarming reviews begin to arrive. By leveraging the current evidence, iROGUE helps protect future users and third parties. A possible extension for iROGUE would be to include other information sources such as privacy policies, to identify rogue apps ahead of time.

Second, uncovering rogue functionalities involves the manual effort of inspecting top alarming reviews. A possible future direction is to automate this process.

CHAPTER

3

IDENTIFYING INCIDENTS, EFFECTS, AND REQUESTS FOR ADVICE FROM METOO STORIES

Warning: This chapter may contain triggering language for some readers, especially survivors of sexual harassment and assault.

Survivors of sexual harassment and assault sometimes share their experiences on social media, revealing their feelings and emotions and seeking advice. On platforms such as Reddit, some stories can be long due to 40,000 allowed characters per post. We found that such long stories are demanding for helpers to read and respond to. Hence, we focus on automatically highlighting parts of the story that are considered important by the prior research. Highlighting those parts can draw helpers' attention toward key information and assist them while reading and responding to long stories.

For stories posted on diverse subreddits, we first investigate the prevalence of these important parts and analyze various story structures. Further, we develop a natural language-based model to highlight sentences from a story that describe any of these parts. On ten-fold cross-validation of a dataset, our model achieves a macro F1 score of 0.82. In addition, we contribute METHREE, a dataset comprising 8,947 labeled sentences for these parts from Reddit stories.

On subreddits, we conducted a survey of helpers which shows that the highlighted parts represent important information and assist them while reading and responding to long stories. We found that highlighted text significantly reduces the demandingness of long stories. Moreover, almost all helpers felt that highlighted stories are helpful and easier to read, understand, and respond to than nonhighlighted ones.

3.1 Introduction

In the United States, 81% of women and 43% of men have reported some form of sexual harassment or assault in their lifetime (NSVR 2023). Both sexual harassment and assault are unlawful acts as they include unwelcome or nonconsensual sexual advances and verbal or physical acts of a sexual nature (EEOC 2023; RAINN 2023). Due to such experiences, survivors often feel anxious, depressed, or go through post-traumatic stress disorder (Field-Springer et al. 2022; O'Neill 2018).

To help survivors know they are not alone, Tarana Burke, a civil rights activist, coined the phrase "me too," using which survivors can openly share their experiences of sexual harassment and assault. This led to the *MeToo* movement, which leveraged social media platforms to disclose sexual harassment and assault stories. While writing MeToo stories on social media, survivors can choose to remain anonymous and hence be unafraid of public shaming. Sharing such stories facilitates survivors to vent out or seek support from fellow users on the platform, called *helpers*. The guidance provided by helpers can aid survivors in various ways such as healing and reporting abuse (Andalibi et al. 2016; O'Neill 2018).

This work investigates MeToo stories on social media to assist *helpers* providing support to survivors. More specifically, we focus on Reddit, a popular social media platform that includes forums called subreddits, some of which (`r/meToo` (Reddit 2023a), `r/SexualHarassment` (Reddit 2023c), and `r/sexualassault` (Reddit 2023b)) are geared toward survivors sharing their MeToo stories.

Reddit allows long posts (of up to 40,000 characters) and survivors take advantage of this high limit by posting long stories. For instance, 5,327 relevant stories collected by us have a mean length of 1,742 and a maximum length of 33,432 characters. The high length limit is an advantage of this platform because it ensures survivors to write without length limits and share whatever information they wish to.

3.2 The Setting: MeToo Stories on Reddit

Example 8 shows a MeToo story of approximately mean length in which the survivor describes their experience and needs support about how to handle it. We include the instances of much longer stories in the Supplement Section 3.11 (Examaple 11). To be used as examples in this chapter, we have paraphrased these stories so that the traceability of the survivor and their details is reduced. This is because even though the stories are public, it may not be appropriate to draw attention to survivors' identifiable details.

Example 8: A MeToo story.

In 20xx, I went out with a group, consisting of my girlfriend and her two guy friends. She ended up canceling on us last minute so it ended up just being me and these two guys. I had hung out with those guys before so I thought they were trustworthy. We were out in the city driving around and drinking in the car and we three got wasted.

After pulling me over in a dark area, where some vehicles were parked, I found that one of them was trying to kiss me. Most of the incident was blurry to me but I remember that he pressurized me to make out with him. Honestly, I don't think that I gave my consent to him but I also don't remember asking him to stop. I have blurry memories–pieces and gaps of being laid down after getting off my clothes. I also have memories of having his friend riding on me. I woke up the next day and found myself naked. I wore my dress and asked him to take me home. This was because I thought my family was probably looking for me at this point. I was not in the state of thinking but rather wanted to get home ASAP.

Once I got home my parents were in tears and questioned me if anything had happened, she asked if I was raped? I said no promptly. How was I sure that it was rape and how could I convey the same to my parents? Once I came into my room, I saw my body and found bruises on my hands and cuts and scrapes on my legs. I kept telling myself that since I was drunk, maybe I started having intercourse with the guy. But I was also getting memories of his friend in me. He messaged me the next day to ask about my health and told me that he had done nothing wrong, which confused me. So I buried this incident and moved on. When I look back, I understand and realize that I was raped by both of them that night.

I am still getting glimpses of the assault and I can't stop thinking about it. Things are not yet clear to me. I am getting anxious after all these glimpses.

But Why now? Why am I having flashbacks and what to do about it? Do you recommend me to go to a counselor?

3.2.1 Helping Survivors

Due to sexual harassment and assault, survivors face immediate and long-lasting emotions such as anger, fear, anxiety, self-guilt, shame, and so on (Gold et al. 2008; Fortier et al. 2009). However, the literature (Rickwood et al. 2005; Andalibi et al. 2016; O'Neill 2018) suggests that seeking supportive responses from helpers can have a positive psychological impact on their mental health.

On Reddit, helpers provide support by responding to stories seeking a variety of advice, ranging from validating the experience to confronting the abuser. However, due to the high character limit on Reddit, long stories, such as those found typically pose a challenge to helpers. A survey of helpers that we conducted found that reading and responding to long stories is demanding for helpers (Section 3.6). Such a scenario may cause no or delayed responses to survivors.

The problem addressed in this project is to assist helpers in reading long stories by identifying their key parts, and thereby facilitating the helping process. We leverage advanced NLP methods to automatically highlight the key parts of MeToo stories. We evaluated its effectiveness by surveying a sample of users, who serve as helpers on MeToo-related subreddits and on average have written between four and ten supportive responses.

3.2.2 Focus of This Study

The objective of this study is two-fold. First, identify and analyze the challenges faced by helpers while providing support on social media. We do so by conducting a survey of them on subreddits and understanding their reactions to long stories. Second, build a computational approach that can assist helpers in real time. We leverage an active learning approach to curate MeThree, a dataset of sentences (labeled for three parts and None), and a computation model. We validated the effectiveness of our model through the survey which also led to several other findings.

In no way, we are restricting survivors from posting long stories. Neither, do we think negatively about such stories. Instead, we are assisting helpers in reading and responding to such stories.

3.3 Key Contributions and Findings

We have two contributions. First, we curated MeThree, a dataset containing 8,947 sentences drawn from stories in `r/meToo`, `r/sexualassault`, and `r/SexualHarassment`. Each sentence in MeThree is labeled for the above three parts as well as None. MeThree can be used for future computational research in the MeToo domain. Second, the survey of helpers we conducted that unveiled some of the following findings.

While investigating MeToo stories, we found multiple variations in story structure. This complexity in the story structure makes automatic highlighting more difficult but potentially more useful to helpers.

A survey of 29 helpers shows that helpers find long stories demanding to read and respond to. On a four-point scale of demandingness, the survey respondents provided mean ratings of 3.10 for reading and 3.13 for responding.

We conducted the Mann-Whitney U-test that shows highlights significantly reduce the demandingness of reading (statistic value = 484, p-value = 0.01) and responding (statistic value = 486, p-value = 0.01) to long stories. Moreover, almost all respondents prefer the highlighted stories as they are helpful and easier to read, understand, and respond to than nonhighlighted stories.

3.4 Organization

Section 3.5 describes the three parts of stories considered important in literature and the analysis of stories based on them. Section 3.6 discusses our findings from the survey and the benefits of using our model. Section 3.7 delves into all the technical details, including how we collected Reddit stories and leveraged active learning to curate METHREE dataset and the model. Sections 3.8, 3.9, and 3.10 describe the related work, the implications of our work, and ethical concerns, respectively. In the end, we also present a supplementary section including additional details.

3.5 Diving Deep into MeToo Stories

For the highlighting task, we take three parts of MeToo stories that are considered important in the literature. To demonstrate these parts, we highlight them in Example 9 containing the same story described in Example 8. We describe these parts below.

Incident Text describing unwelcome sexual advances, sexual behavior, requests or forcing for sexual favors, verbal or physical acts of sexual nature, offensive jokes or remarks that are either sexual or based on someone's gender (EEOC 2023; RAINN 2023; Ghosh Chowdhury et al. 2019a; Karlekar and Bansal 2018).

Example 9's text highlighted as this describes inappropriate touching, which accounts for sexual assault.

Effects on the survivor Survivors describe how they are affected by revealing their feelings and emotions that arise during or after the incident (Field-Springer et al. 2022; O'Neill 2018). Examples of effects include the survivor feeling uncomfortable due to the abuser's actions, or being angry or upset due to the incident.

Example 9's text highlighted as this is where the survivor discusses their uncomfortable feeling due to this incident.

Advice requested Text in which survivors seek advice from other platform users (Andalibi et al. 2016; Rickwood et al. 2005; O'Neill 2018). Some examples of advice requested include asking if their (i.e., the survivor's) reported experience constitutes harassment or assault, how to pursue a legal case, and how to confront the abuser.

Example 9's text highlighted as this indicates that the survivor is confused and asks if they are overthinking the incident. Even if they are confused, such situations are clearly about harassment and assault, according to the incident described. Hence, key parts of such stories must be brought to the helpers' attention.

We also analyzed 50 randomly selected stories (from 5,327 relevant stories, whose collection is described in Section 3.7.1) to see the prevalence of the three parts. Table 3.1 shows the mean length for each of these parts in MeToo stories. On average, the effects part takes 34.18% of story text, which is greater than incident (30.76%) and advice (02.65%). Because of overlaps between the three parts, overall we find that 55.11% of story text is important.

Besides the length of the stories, an underlying challenge is complexity, which also translates to how demanding a helper finds understanding and responding to a story. We validated the importance of these parts and the overall demandingness of stories through helpers on subreddits, as discussed in Section 3.6.

Table 3.1: Distribution of a story's length across three parts.

Parts of story	Average proportion of length
Incident	30.76%
Effects	34.18%
Request for advice	02.65%
Union of all parts	55.11%

Considering these three parts, we analyzed the structure of 50 randomly selected stories. As depicted in Example 9, the expected story structure is I-E-A, meaning *incident* first (I), followed by *effects* (E), and finally seeking *advice* to overcome the current situation (A). However, only nine (18%) of 50 analyzed stories preserved this structure. The remaining 82% stories showed multiple variations of ordering I, E, and A. Two of these variations are shown in Example 10. The first paraphrased story in the example starts directly with seeking advice, which may be the main reason for a survivor to share their experience. Afterward, the survivor discusses the incident and its effects, leading to the A-I-E structure. The second paraphrased story mentions an effect and its associated incident in an iterative manner (E-I-E-I-E structure). We found that some sentences belong to more than one of the three parts, leading to more complex variations in story structure. These variations make automatic highlighting a challenging task.

Some stories (like the second story in Example 10) don't show advice being requested in their structure. We found that it's because of one of two reasons: (i) advice is already sought in the title of the story, or (ii) the survivor is only venting their experience and not seeking advice. We investigated helpers' responses to the latter case and found that helpers provide sympathetic responses to make survivors feel better. To provide such responses, reading and understanding long stories remains important.

Example 9: Incident, effects, advice requested.

Parts: Incident, Effects, Advice requested

In 20xx, I went out with a group, consisting of my girlfriend and her two guy friends. She ended up canceling on us last minute so it ended up just being me and these two guys. I had hung out with those guys before so I thought they were trustworthy. We were out in the city driving around and drinking in the car and we three got wasted.

After pulling me over in a dark area, where some vehicles were parked, I found that one of them was trying to kiss me. Most of the incident was blurry to me but I remember that he pressurized me to make out with him. Honestly, I don't think that I gave my consent to him but I also don't remember asking him to stop. I have blurry memories–pieces and gaps of being laid down after getting off my clothes. I also have memories of having his friend riding on me. I woke up the next day and found myself naked. I wore my dress and asked him to take me home. This was because I thought my family was probably looking for me at this point. I was not in the state of thinking but rather wanted to get home ASAP.

Once I got home my parents were in tears and questioned me if anything had happened, she asked if I was raped? I said no promptly. How was I sure that it was rape and how could I convey the same to my parents? Once I came into my room, I saw my body and found bruises on my hands and cuts and scrapes on my legs. I kept telling myself that since I was drunk, maybe I started having intercourse with the guy. But I could also feel his friend in me. He messaged me the next day to ask about my health and told me that he had done nothing wrong, which confused me. So I buried this incident and moved on. When I look back, I understand and realize that I was raped by both of them that night.

I am still getting glimpses of the assault and I can't stop thinking about it. Things are not yet clear to me. I am getting anxious after all these glimpses.

But Why now? Why am I having flashbacks and what to do about it? Do you recommend me to go to a counselor?

Example 10: Variations in story structure

Parts: Incident, Effects, Advice requested

Story exhibiting the A-I-E structure

Through this post, I want to know one thing: was I raped? ... Once, he was pressurizing me to lift me up my clothes and show him my breast. ... I hate myself for letting it happen and for not saying a word against it ...

Story exhibiting the E-I-E-I-E structure

... I get anxious while recalling incident, but my current partner doesn't do anything that can help me ... Basically, I was raped by a female friend multiple times. ...... For all of this, I still blame myself for not saying no to the abuser ...... Another friend of mine pounced me once. ...... Due to all these experiences, I sometimes feel stuck and freeze up if someone comes close to me

We found that two of these 50 stories included TLDR: an acronym for Too Long; Do Read, a brief summary of the MeToo story optionally provided by the poster. One of the TLDRs described the incident and advice, whereas the other described all three parts, thus indicating the importance of these parts even from the survivors' perspective. Moreover, a story including TLDR is rare (2 in 50), hence automatic highlighting is required.

3.6 Empirical Study of Helpers

We conducted a survey of helpers to understand their reactions to long stories, validate the importance of the three parts, and the effectiveness of our model in generating highlights. Since our study is primarily based on Reddit, we decided to recruit respondents from MeToo-related subreddits.

Many MeToo-related subreddits (including r/SexualHarassment, r/sexualassault, and others such as r/afterthesilence and r/Molested) don't allow such postings for research surveys. To recruit suitable respondents, we were left with two subreddits, namely, `r/metoo` and `r/rape`.

Our study posed minimal risk to respondents and was exempted by the Institutional Review Board (IRB) at our university. We received 29 completed surveys that passed the attention-check questions. On average, these 29 respondents of the survey have written between four and ten helpful responses on MeToo-related subreddits. Upon survey completion, each respondent received an Amazon gift card of $20. Respondents were also provided an optional bonus questionnaire. 24 of 29 respondents answered bonus questions and received an additional $9 Amazon gift card. We provide the whole survey along with a bonus section in Section 3.11.

The initial survey questions were related to the demographics of the respondents and were optional. Out of 29 respondents, 28 chose to reveal their gender identities (19 males, 8 females, and 1 non-binary, genderqueer, or genderfluid) and 27 revealed their age (ranging in 22–38).

We discuss our findings from 29 responses in Sections 3.6.1 and 3.6.2 and from the bonus questions in Section 3.6.3.

3.6.1 Long Stories are Demanding

For the first section of the survey, we selected ten stories (from 5,327 relevant stories) longer than the mean length of the stories: 1,742 characters. The average length of these ten selected stories was 10,211 characters. Out of these ten stories, one story was randomly assigned to each respondent. Respondents were required to read the story and answer two questions: (i) how demanding was completely reading the story (Q1) and (ii) how demanding was it to construct a response (Q2). On a Likert scale of one to four, the mean rating for Q1 was 3.13 (standard deviation of 0.95) and 3.10 (standard deviation of 0.90) for Q2. This shows that helpers find such stories demanding to read and respond to.

3.6.2 Highlights are Helpful

In the second section of our survey, we showed respondents the highlighted stories (based on our computational model's predictions about the incident, effect, and advice requested) and validated: (i) if the three parts highlighted by our model provide important information about the survivor's situation and (ii) if the highlighted text assists helpers while reading and responding. We include all technical details pertaining to our dataset and model in Section 3.7.

For this human study, we selected ten stories different than the ones selected in the first section of the survey but of comparable lengths (average length was 10,135). We set up our survey in such a way that one of these ten highlighted stories was randomly assigned to each respondent. Respondents had to read the entire story containing highlights, write a supportive response addressing survivors' concerns, and answer the following survey questions:

Q3: Do highlighted sentences represent important parts of the story?

Q4: (If answered yes to Q3) On a scale of one to four, how would you rate the importance of highlighted text?

Q5: (If answered yes to Q3) What information did the highlighted text provide you?

Q6: How helpful was the highlighted text while reading and responding?

Table 3.2 summarizes the results of our study. Almost all (28 of 29) respondents answered *yes* to question Q3, indicating the importance of the three parts we considered for highlighting. Moreover, Q4 received a high rating on the importance scale. In answering Q5, respondents backed their ratings with a qualitative text about the highlights' benefits. And, in Q6, they rated the predicted highlights as being highly helpful. Overall, Table 3.2 shows the effectiveness of our model in assisting helpers in providing support.

3.6.3 Demandingness is Reduced with Highlights

In the bonus section of the survey, we asked respondents the following questions:

Q7: On a scale of one to four, how demanding was it to read and understand the highlighted story?

Q8: On a scale of one to four, how demanding was it to construct a response to the highlighted story?

Q9: Are stories with highlights easier to read and understand than stories without highlights?

Q10: Is it easier to construct a response to stories with highlights than stories without highlights?

We analyzed if highlights affect the demandingness while reading and responding to long stories. To do so, we considered respondents' ratings for Q7 and Q8 and compared them with Q1 and Q2, respectively. For this comparison, we conducted a renowned statistical test namely, the Mann-Whitney U-test, which shows that highlights significantly reduce demandingness in reading (statistic value = 484, p-value = 0.01) and responding (statistic value = 486, p-value = 0.01) to long stories.

According to Q9 and Q10 responses, we found that almost all helpers (23 of 24) prefer highlighted stories as they are easier to read, understand, and respond to, in comparison to non-highlighted stories.

3.7 Method: The METHREE Dataset and Classifier

In this section, we describe all the technical details, starting from how we collected MeToo stories to how we applied active learning for the highlighting task. The active learning process developed hand-in-hand a labeled dataset (called METHREE) and a model to automatically highlight three parts of the story.

3.7.1 Collecting MeToo Stories

We collected a total of 9,140 MeToo stories from three subreddits: `r/meToo`, `r/sexualassault`, and `r/SexualHarassment`, for the period 2016-01-01 to 2021-07-18, using Reddit's Pushshift API (Pushshift 2023). This set of 9,140 stories was the entire set of stories in that time range in those subreddits.

Filtering Relevant MeToo stories

Some MeToo stories don't share survivors' experiences but instead share news articles, seek opinions about allegations against celebrities, or promote other platforms. Such stories are irrelevant to our study. Similar to a previous study (Hassan et al. 2020), we applied the following heuristics to focus on stories containing survivors' personal experiences:

- First-person pronouns: Many survivors while describing their personal experiences, use first-person pronouns in the title of the story. For example, "**I** started to do something about **my** past assault, but instead of feeling better, it actually gets worse" and "**My** mom's boyfriend tried to get **me** to do things to him." Thus, we checked the presence of first-person pronouns: *i*, *me*, *my*, and *mine* in the title to find relevant MeToo stories.

- Advice-related keywords: We observed that survivors also use advice-related keywords in the title. For example, "Need **advice**, or support" and "pls someone read this and **help** me figure out if i was assaulted or not." We used the keyword, *advice*, as seed and queried its synonyms from the Oxford dictionary. We obtained 25 synonyms and manually filtered four of them based on their relevance to our problem. The final list contains *help*, *suggestion*, *advice*, *guide*, and *counsel*. We referred to this list of keywords as *advice keywords*. To filter relevant stories, we checked the presence of these keywords in the title. For extracting synonyms, we also explored other corpora such as WordNet (Miller 1995) but did not find synonyms that were commonly used.

- Advice-related questions: We observed that many relevant stories ask a question (related to harassment or assault) in the title. For example, questions like "Was this rape?" and "Is this sexual harassment?" are seeking advice without mentioning any of the advice keywords. Using Part-Of-Speech (POS) tagging (Manning 2011), the titles that have an interrogation form and include *rape*, *harassment*, *assault*, and *abuse* as the object, were filtered.

Stories with titles satisfying one of the above rules were chosen. We checked 50 randomly selected stories for relevancy, involving a total of 812 sentences and a total 72,130 characters. Of these 50 stories, 47 (94%) were relevant because they either sought support or advice related to their case. Among these 47 stories, we found one story written by the survivor's friend but the story still expressed the effects on the survivor and sought advice.

In total, we obtained 5,327 relevant stories using the above heuristics. Our statistics about mean and maximum length, and the distribution of the three parts pertain to this set of stories. We might have missed some relevant stories, but the objective here is to filter stories with high precision. This is because high precision (94% in our case) means we can build a dataset of relevant sentences without further pruning. Out of 5,327 filtered stories, 69.38% (3,693) are from `r/sexualassault`, followed by `r/meToo` (22.78%; 1,214) and `r/SexualHarassment` (7.84%; 420).

3.7.2 Curating Dataset and Model via Active Learning

The preparation of the dataset and the development of a classifier happen iteratively. For classification, we adopt a pool-based active learning approach which is known for training robust models while reducing manual labeling effort (Settles 2012). We tackle the problem of highlighting elements at the sentence level. Doing so gives us the ability to highlight key portions of long posts on Reddit.

We curate METHREE, a dataset comprising 8,947 labeled sentences that are sampled from 5,327 stories as explained below, and train an XLNet model on METHREE. Each sentence can either belong to None or at least one of three categories: incident, effects, and advice (being) requested. Hence, we have a multilabel classification task.

Pool-based active learning (Settles 2012) starts with an initial dataset (denoted by L) including labeled sentences, most of which were selected based on certain keywords (Section 3.7.3).

In active learning, the four steps shown in Figure 3.1 are repeated multiple times. First, a model (denoted by M) is trained on the set L. To do so, we compared the performance of multiple models and chose the best-performing one (XLNet) as model M (Section 3.7.4). Second, an unlabeled dataset U is labeled by the predictions from the trained model M. In our case, since most of the sentences in L contained certain keywords, to avoid bias, we selected U from sentences without those keywords. Third, from U, data points having a high chance of being misclassified are queried and manually labeled (Section 3.7.5). Fourth, U is added to L (Section 3.7.6). We repeated the active learning cycle five times to curate METHREE, the final dataset of 8,947 labeled sentences, and our computational model.

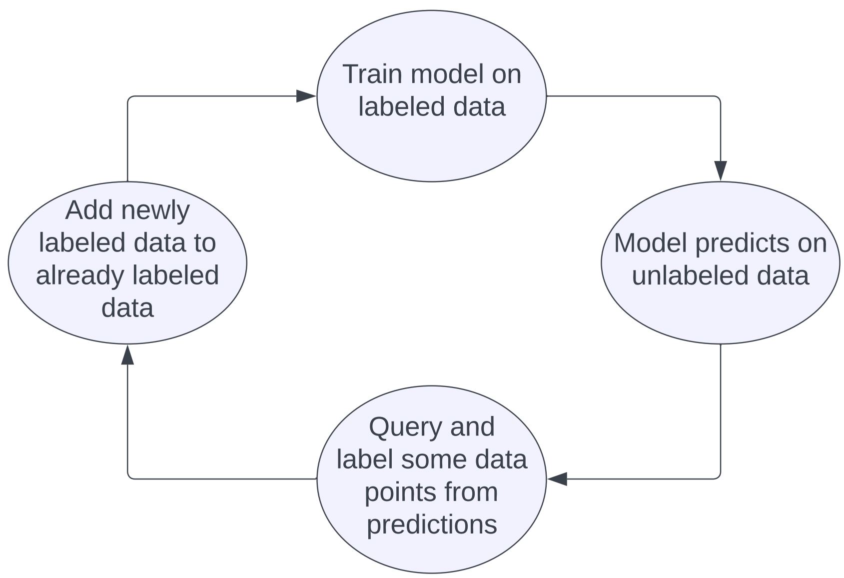


Figure 3.1: Active learning cycle involving four steps.

3.7.3 Initial Training Data for Active Learning

To form our initial training data (set L), we leverage 5,327 relevant stories. First, for each category, we found candidate sentences from these stories. Second, we labeled a sample of candidate sentences along with other sentences.

Finding Candidate Sentences

We split each of the 5,327 relevant stories into 102,204 sentences, using Natural Language Toolkit (NLTK) library (NLTK 2023). However, a random sample of these sentences was inefficient in getting sentences that describe incidents, their effects, or advice being requested. Thus, we first found candidate sentences of each category using the following keywords:

- Incident: Hassan et al. (2020) create a list of 27 MeToo-related verbs (such as *molest*, *touch*, *rape*, *masturbate*). We expanded the list by retrieving synonyms of these verbs from the Oxford English dictionary. The resulting list contained 652 verbs. We manually checked them and found 539 relevant verbs, of which only 313 were unique. We called this final set of 313 verbs *incident keywords*. We found 30,927 sentences having one or more incident keywords.

- Effects: In the NRC word emotion lexicon (Mohammad and Turney 2013, 2010), we considered four emotions: *anger*, *disgust*, *fear*, and *sadness*, that a survivor can express, and used lexicons associated with them. In this process, we found 37,271 emotional sentences. Next, we identified 14 synonyms of the word, *feel*, that are highlighted from the Oxford English dictionary. Out of these, eight are relevant to the survivor's feelings. We referred to this set (*feel*, *perceive*, *sense*, *experience*, *undergo*, *bear*, *endure*, *suffer*) as the *feelings keywords*, and found 8,617 candidate sentences containing one or more of them.

- Advice requested: We observed that many questions in MeToo stories are advice seeking. For example, "Was it actually just a mistake and should I forgive him?" and "Am I blowing it out of proportion?" Hence, we considered all questions as candidates for advice-seeking sentences. We found 6,354 such candidates. Next, we leveraged advice keywords to find additional 2,678 candidates.

To find synonyms, we used the Oxford English dictionary as Corpora such as WordNet and PyDictionary were not useful. For example, PyDictionary produced no synonyms representing feelings, and WordNet produced one word, *palpate*, which was uncommon to describe feelings.

Labeling Sentences

Due to the presence of keywords (such as incident, feelings, and so on), candidate sentences are likely to be relevant to the three categories (incident, effects, and advice). Hence, for set L, we took a random sample of 6,900 sentences from all of the above sources: incident keywords, feelings keywords, and so on. However, the training set (set L) would be biased toward these keywords. To mitigate that risk, at this step of labeling, we also included random 500 sentences not having any keywords. After discarding duplicates, we were left with a total of 5,947 sentences.

Since a majority of these 5,947 sentences contained chosen keywords, labeling them would produce a biased dataset. Note that this was the initial training data (set L). In subsequent steps of active learning, to mitigate bias, we kept including sentences without any keywords (set U), as described in Section 3.7.6.

For these 5,947 sentences, three researchers in this project were the annotators. Before labeling, they were aware of the uncomfortable and disturbing text present in these sentences. For each sentence, the annotators answered the following questions:

1. Does this sentence describe sexual harassment or assault?

2. Does this sentence describe the effects of the incident on the survivor?

3. Does this sentence ask for any advice?

The annotators read each sentence and answered the above questions as yes or no. Initially, two annotators labeled 200 sentences based on their understanding of the problem statement. Later, they discussed their disagreements and defined the final labeling instructions for all the annotators. The brief definitions are shown in Section 3.5, whereas the detailed labeling instructions (along with examples) are shown in the appendix.

All 5,947 sentences were divided among the three annotators (let's denote them by A_1, A_2, and A_3) such that each sentence was labeled by two of them. After labeling all the sentences, we obtained Cohen's kappa scores (Cohen 1960) of 0.772 (for incident), 0.774 (for effects), and 0.865 (for advice requested). These scores indicate that we achieved substantial agreement for two categories: incident and effects, and almost perfect agreement for the advice requested category. Table A.2 also shows Cohen's kappa scores for each pair of annotators. Finally, the first author resolved all the disagreements. The labeled 5,947 sentences form the initial training data (set L) for active learning.

3.7.4 Initial Model to Highlight Sentences

Once set L is curated, the next step is to train model M. We consider our problem a multilabel classification task in which each sentence is an input to the model and the output has three binary labels (one label for each category). We trained and evaluated multiple methods on the 5,947 labeled sentences (set L) as described below.

For each of these sentences, we computed embeddings such as Sentence-BERT (Reimers and Gurevych 2019), TF-IDF (Cahyani and Patasik 2021), Stanford's GloVe (Pennington et al. 2014), Word2Vec trained on the Google News (Mikolov et al. 2013), and Universal Sentence Encoder (USE) (Cer et al. 2018). For each embedding, we used the sentence vector as an input to a multilabel classifier. For GloVe and Word2Vec, we averaged word vectors to form the sentence vector. For classification, we tried Logistic Regression (LR) (Dreiseitl and Ohno-Machado 2002), Support Vector Machine (SVM) (Cervantes et al. 2020), and Random Forest (RF), and report the best method.

Besides embedding-based methods, we applied transformer-based approaches such as RoBERTa (Liu et al. 2019b) and XLNet (Yang et al. 2019). We fine-tuned RoBERTa and XLNet on set L by adding an output layer containing three units, one dedicated to each category. Both models minimized binary cross entropy over five epochs. The training batch size and tokenizer length were set to 32 and 256, respectively.

Table 3.3 reports the mean F1, precision, and recall scores for the above approaches over ten-folds of set L, as well as category-wise keywords (those used in Section 3.7.3). A sentence containing at least one keyword was predicted as 1 for that category and 0 otherwise. For the embeddings-based approaches, Table 3.3 reports the results only for their best-performing classifiers.

TF-IDF, GloVe, Word2Vec, Keyword search, and USE underperform as compared to other methods.

Sentence-BERT followed by SVM achieves the highest macro precision (0.84). However, it shows lower macro recall (0.66) than RoBERTa (0.84) and XLNet (0.87). Overall, XlNet outperforms all other methods by achieving the highest macro F1 score (0.82). Thus, we choose XLNet as model M in our method.

3.7.5 Predicting and Querying from U

After model M is trained, it's time to make predictions on the set U and label it. To mitigate the risk of a biased dataset, we chose the set U to be a random sample of 500 sentences not containing any keywords and applied the model M to it. We selected potentially misclassified sentences (in U) for manual labeling, as described below.

Uncertainty sampling (Culotta and McCallum 2005; Dagan and Engelson 1995) queries data points (sentences) for labeling where the model is uncertain. However, uncertainty sampling methods (such as least confidence and entropy) did not work well in our case. This is because, in the first active learning cycle, model M (XLNet trained on 5,947 sentences; Section 3.7.4) predicted low probabilities on sentences that lack the keywords (set U). We validated this by predicting on 100 such sentences, where the mean prediction probability was 0.08 (standard deviation of 0.23) for incident, 0.10 (standard deviation of 0.27) for effects, and 0.05 (standard deviation of 0.21) for advice requested. Due to most of the probabilities being low, uncertainty sampling methods could not discriminate between misclassified and other sentences.

To query the misclassified sentences from U for manual labeling, we found a threshold on the prediction probability. To find that threshold, we used a set U', another random sample of

500 sentences not having any keywords. On U', we plotted the Receiver Operating Characteristic (ROC) curve and computed Youden's J-statistic (Youden 1950), as described below. This plot was based on the first active learning cycle and the model M referred to below is XLNet trained on the 5,947 sentences described in Section 3.7.4.

1. The first author labeled the set U'.

   One annotator was adequate because of the high inter-rater agreement that we achieved on the earlier sample of 5,947 sentences

2. The model M made predictions on the set U'.

3. We split U' into 400 sentences (set V) and 100 sentences (set T).

4. We used V to fine-tune the threshold above which we would select sentences for manual labeling. Out of the above 400 sentences, model M misclassified 28 sentences for the incident category, 38 for effects, and 9 for advice requested. Since we wish to retrieve these sentences, we consider misclassification as positive and plot its true positive rate on the Y-axis of the ROC curve for each category, as Figures 3.3a–3.3c show. From each ROC curve, we found the best threshold by maximizing Youden's J-statistic (Youden 1950), which equals the height above the chance line.

   For incidents, effects, and advice, we identified thresholds of 0.038,177, 0.008,476, and 0.007,874, respectively. We selected sentences in the classes if they exceeded the respective threshold.

5. For each category, to ensure that we did not miss out on the misclassified sentences, we additionally selected 30 sentences below the threshold for every 100 predictions. For example, the model M predicted on the set V that has $4 \times 100 = 400$ sentences. Therefore, we selected $4{\times}30 = 120$ sentences for each category below the threshold.

   In brief, for each category, query (1) sentences with prediction probability above or equal to the threshold, and (2) 30 sentences below the threshold for every 100 sentences.

   Through this query strategy, we could identify most of the misclassified sentences in V: 89.28% for the incident, 92.10% for effects, and 100% for advice requested.

6. We used the set T to test our query method. In T, M misclassified 10 sentences in the incident, four in effects, and three in advice. Our thresholds identified nine (90%) misclassified incident sentences and all misclassified sentences (100%) for effects and advice.

We used the above method in each iteration to retrieve potential misclassified sentences from U and manually labeled retrieved sentences. For manual labeling, each of the three annotators (researchers in this project) labeled the sentences retrieved for a category.

3.7.6 Completing Active Learning Cycles

We completed the first active learning cycle (Figure 3.1) by adding labeled U to L.

We iterated the cycle four more times, to add 2,500 labeled sentences (each time U having 500 sentences without keywords) to the initially 5,947 labeled ones. As a result, the final L includes 8,447 sentences. As discussed in Section 3.7.5, we labeled U', an additional 500 sentences without keywords. After adding U', METHREE becomes a dataset of size 8,947 labeled sentences.

In METHREE, there are 4,331 (48.4%) sentences that belong to at least one category, and 4,616 (51.6%) to None. Figure 3.2 shows the Venn diagram of the 4,331 sentences over the three categories.

Finally, we trained XLNet on METHREE which is used to highlight sentences from long stories. Over ten cross validation of METHREE, the model achieved 0.82 macro F1 score (0.78 for incident, 0.79 for effects, and 0.89 for advice requested), 0.86 macro recall (0.82 for incident, 0.83 for effects, and 0.92 for advice requested), and 0.78 macro precision (0.74 for incident, 0.76 for effects, and 0.85 for advice requested).

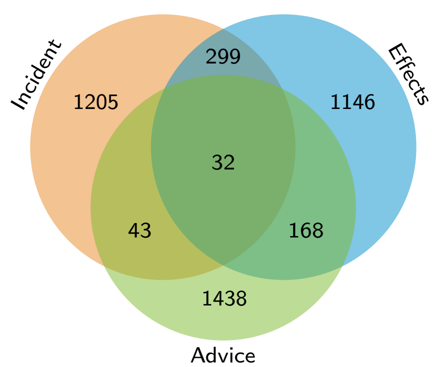
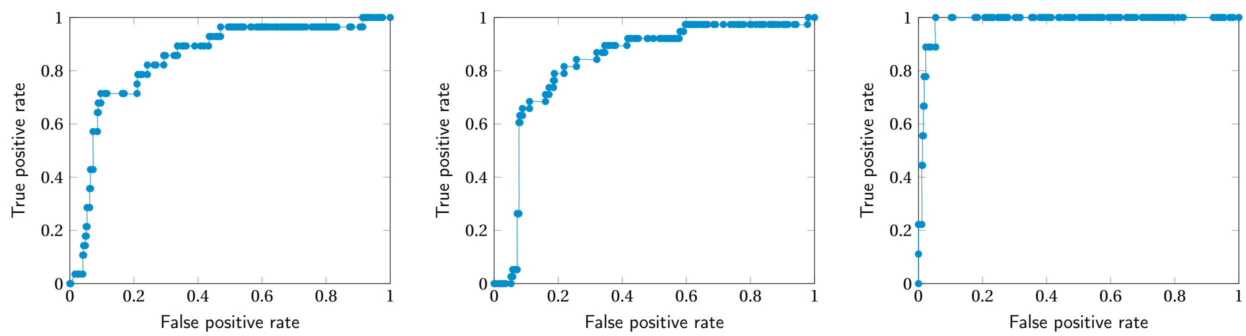


Figure 3.2: Distribution of sentences in METHREE across the three categories.



(a) Incident category. (b) Effects category. (c) Advice requested category.

Figure 3.3: ROC curves showing true positive and false positive rates, while considering misclassified sentences under positive class.

3.8 Related Work

Existing literature in psychology and healthcare focus on survivors' mental health (Gold et al. 2008; Fortier et al. 2009) and discuss how it can be improved (Rickwood et al. 2005; O'Neill 2018) through support seeking. Moreover, there has been extensive research in analyzing the stories seeking support and their reactions on online platforms (Manikonda et al. 2018; Deal et al. 2020; Field et al. 2019; Reyes-Menendez et al. 2020). However, these works don't primarily focus on classification tasks unlike us.

A few studies have focused on classifying MeToo stories, whereas we focus on sentence-level classification. Karlekar and Bansal (2018) collect 9,892 MeToo stories from the SafeCity website (Safecity 2023) and classify them for one of the harassment types: (i) groping or touching, (ii) staring or ogling, and (iii) commenting. Yan et al. (2019) improve the performance of this classification by proposing a quantum-inspired density matrix encoder. Liu et al. (2019a) leverage the same dataset and annotate it for attributes such as the abuser's age (below 30 or older), the abuser's relationship with the survivor (for example, relative or teacher), the location of the harassment (for example, park or street). They propose a framework to identify these attributes from a MeToo experience. Bauer et al. (2020) also use the SafeCity dataset and build a chatbot system to help survivors. The SafeCity dataset contains concise experiences (typically three or four sentences long) and is ill-suited to highlighting sentences, as in our approach.

Hassan et al. (2020) train a model on 520,761 tweets with the #MeToo hashtag to identify tweet-level attributes, such as the category of sexual violence reported, the survivor's identity (tweeter or not), and the survivor's gender. They achieve overall 80.4% precision and 83.4% recall in identifying sexual violence reports. Ghosh Chowdhury et al. (2019b) label 5,119 tweets for types: (i) disclosure of a survivor's personal experience and (ii) nondisclosure. Out of these

5,119 tweets, they find 1,126 (22%) indicate disclosure. They propose a language model to classify tweets into these types. Studies by Khatua et al. (2018) and Ghosh Chowdhury et al. (2019a) also focus on similar classification tasks.

These studies identify relevant MeToo stories from a massive stream of social media text. The expectation is that a helper on these platforms can provide support to the survivor of the identified story. However, merely identifying relevant stories is not enough. In order to provide support, a helper should be able to read and understand those stories. To facilitate that, we generate sentence-level highlights covering incidents, their effects on the survivor, and the advice requested.

Traditional text summarization works (Jadhav and Rajan 2018; Cheng and Lapata 2016; See et al. 2017; Li et al. 2011; Zhang et al. 2012) are trained or evaluated on other domains such as news datasets but are not built for the MeToo context. Our work is the first attempt to highlight text from long MeToo stories to the best of our knowledge.

3.9 Discussion

We now discuss our conclusion, our work's limitations, and possible future directions.

3.9.1 Conclusion

Our study addresses a crucial aspect of online support communities by harnessing NLP to identify vital elements of MeToo stories. Our aim is to facilitate a more streamlined and effective helping process within these online spaces. The computational model developed is not meant to alter or diminish the authenticity of survivors' narratives, nor is it intended to impose a singular narrative structure. Rather, the model is designed to spotlight three pivotal components of survivors' stories: the description of the incident, the effects associated with it, and the specific advice sought from fellow forum members (called helpers).

The results of our work are encouraging, hinting at the potential benefits of utilizing such an NLP model. The analysis of our sample demonstrated that stories, where these critical components were highlighted, exhibited a significant reduction in cognitive demand for helpers. This implies that the model succeeded in enhancing the readability and comprehensibility of survivor stories, potentially enabling helpers to respond more empathetically and efficiently. With the introduction of this model, we aim to complement and amplify the existing dynamics of support, fostering a more nuanced and empathetic dialogue regarding non-consensual and traumatic sexual experiences. However, further research and refinement of the model are essential to determine its long-term effectiveness and the broader implications it might hold for online support communities.

3.9.2 Limitations and Future Work

Our work suffers from some limitations, which motivate future improvements.

One limitation is that our model is trained on the sentences collected from only three subreddits. We expect the nature of sentences in MeThree to be similar to sentences present on other MeToo-related subreddits. However, we plan to fine-tune the model before applying it to the other subreddits.

We can extend our work to generate an automated response based on the highlighted incidents, effects, and advice requested. The automated response after slight corrections by a human helper can offer support and advice to the survivor. Moreover, the similarity between the advice-seeking sentences and users' responses can help assess how relevant each response is. This way, we could improve social media platforms to show more relevant responses above the less helpful ones.

3.10 Broader Perspective and Ethical Considerations

Although we highlighted text from long stories to help survivors, we acknowledge some limitations and possible misinterpretations.

1. **Consent:** Our data was collected from public Reddit stories. Hence, we did not take the consent of the survivors writing such stories. As stated by Ghosh Chowdhury et al. (2019b), some survivors may wish not to be contacted.

2. **Anonymity:** From stories, we did not save survivors' personal information such as usernames, or log of their stories. We paraphrased the stories and sentences presented in this chapter as examples to remove potentially identifying content, such as the survivor's age, job title, and location.

3. **Labeling disturbing text:** Some sentences from MeToo stories can be disturbing, especially for people who have gone through a similar experience. Therefore, we didn't hire crowd workers or volunteers for any labeling task. Instead, three researchers of this project did the labeling.

4. **Potential misinterpretation:** We were extremely aware of the sensitivity of this research before labeling sentences. However, we may have misinterpreted some MeToo experiences. That's why we don't claim that our labeling is fully accurate.

5. **Reproducibility and research:** We are not posting this dataset openly to avoid the risk of drawing attention to specific posts. Upon request, we will share anonymized METHREE with accredited researchers to accelerate future research.

3.11 Supplementary Materials

3.11.1 Survey Details

In our survey, respondents filled the consent form followed by answering the following questions:

- **Demographics**
- How many responses do you think you have written in total on MeToo-related subreddits? Please take your best guess.
  - No responses
  - Between 1 and 5
  - Between 6 and 10
  - Between 11 and 30
  - More than 30
- What is your age? (optional to answer)
- What is your gender? (optional to answer)
- **First section**
- Please read the following real story from a MeToo-related subreddit, and write a helpful or supportive response for the survivor.

Example 11: Sample story presented in survey

At <age>, I was studying in <school-name> when I was also dating my first boyfriend. He was <age> at that time. We never made out and had no experience even with kissing or other things. None in my group used to like him. He was very bad at his studies and failed all subjects in school (he even repeated some years) but was good at sports. he visited me quite a few times when I was in boarding school. I could see some red flags in him but I did not bother. This was because he was too affectionate and touchy-feely, he would put his hand above mine and I would move away, he held my hand without consent so I pulled out mine, and he asked if he could hug me, then hugged me without me answering anything. I felt uncomfortable even if the incident was not sexual. I used to ignore all this but later realized I should have kept my boundaries. Once I was taking the train from a group outing to <place> for a family occasion. He used to live near the place and suddenly appeared to see me after the function. Then he gave an excuse that he wanted something to buy nearby, hence he boarded the train with me. Despite me refusing multiple times, he was waiting for me while I was meeting my dad at the station. I was seriously feeling uncomfortable but he wouldn't leave until my dad showed up and he said hi to my dad, my dad even questioned me who he was to which I replied that he is a friend. These incidents are before we started dating. After a few months into the relationship, when we used to spend time and eventually took a nap, he would touch me inappropriately on my private parts with my clothes on. It was not in an aggressive way, but he was wrong in his assumption that I was sleeping. I didn't have courage to confront at that moment as it was shocking and confusing experience, but later on I did confront him through chats to which he denied all accusations (and me putting lot of pressure, he finally felt apologetic and convey that he was sorry for being opportunist at that moment). By nature, he was also clingy on and possessive, and whenever I had to to go somewhere (such as to and from my workplace) he would check on me by coming along. Initially, I considered these incidents as his sweet gestures, such that he was investing all his time and effort to keep seeing me and worries for my well-being, but later on I realized he was going overboard and didn't like being around him all the time. After a few months into our relationship, I was almost <minor's age> (and him almost <age>), he used to bring up the topic of having sex and kept persuading me for the same. I refused his sexual offers a few times so he didn't do much physically. But one day I finally gave up on his insistence and we had sex on my bed in one of my rooms. He was quite considerate and was very cautious to make sure I wouldn't get pregnant (used contraceptives) but didn't know anything about foreplay so it was not there. And gosh did it hurt! After that, we started having sex quite often, almost every time when he came to visit me. I turned 17 by then. Slowly my pain was reducing and each time he used contraceptives but there was still little to no foreplay. Once he suggested me to go to a hotel with him for having sex and I did not refuse, which shows how dumb I am. We ended up going to a hotel multiple times, all during the initial parts of the day except once when it was a special occasion and we made out at night at that time. To my surprise, we did sex multiple times in the same night. After the first few times of having sex, he stopped asking for consent, I never convey my yes to him for the following "n" times, he just assumed it and inserted his penis every time without asking. I would admit that I did not struggle or yell, but each of the times it was physically hurting and he would not stop until he finishes. I never confronted even once with him. At one point during my final years of study, I was highly anxious as I might have gone pregnant because contraceptives also have a failure rate. Moreover, once when he pulled his penis out after he finished and he was deflating so the contraceptive didn't come out with his penis, hence he used his hand to pull it out. Since I am an overthinker and worry a lot, so I started becoming anxious. I took 2 pregnancy tests and both were negative. My private part was also bleeding a bit so I decided to consult the doctor. The doctor said that this case is of implantation bleeding which generally happens after a few days of period. Due to God's grace, I was not pregnant at that time, hence I also received my next period the next month. The next year of our relationship went into a lot of arguments, we argued in the worst possible ways and I initiated to break this tie with him. I met him at <place> and brought all the gifts and other things that he gave me but he accepted none of them. He was extremely angry at me and I was afraid if he would yell at me in public, so I took him to the roof of that place. he then shouted at me and threw all the stuff around. He also bought some liquid food items for eating together but he took one cup and poured it all over his clothes and body. After this incident, I was extremely shocked, taken aback, and afraid at the same time. As a result, I just froze. By the time he went to one of the restrooms to clean his clothes and found me huddled on the ground after coming back. After seeing me this way, he hugged me and apologized for his behavior. Prior to this incident, he would threaten to hurt himself during any argument. Once, he threatened to jump off the roof (ran towards the edge and then stopped), took the scissors and a nail clipper and acted to hurt his fingers and hands, and even took a knife and pretended to hurt himself when I visited him at his home. As a result, I was worried and concerned about him and took all potential-hurting equipments away from him, but it was that concern he wanted to see in me, to confirm my care towards him. By the way, he also used to show care for me and was always there for me as my boyfriend, especially in investing some time just to see that I am okay. He also bought me some food items such as snacks and drinks because he knew that I like eating all of them, he would always ask for my heavy bag whenever I am carrying it, always opened doors for me, and was always with me whenever I took public transport anywhere. On the day of <occasion>, I told him not to come because my family would be present and there were only two people allowed to accompany the candidate. When I went back to the dorms, he was already there waiting for me. Then he wouldn't allow me to have lunch in the cafeteria because he himself was hungry and brought some food for both of us to eat in my room. He was already angry with me because I did not allow him to attend the <occasion>. I was still wearing my outfit for the occasion, and suddenly I see him lifting my outfit and inserting his penis inside me. I was not in the mood and refused to have sex at that time. However, he did not listen and kept going until he ejaculated. At that moment, I never thought that he raped me in my own room and that too in an outfit. He even said I was ungrateful because I did not thank him enough for the food that he ordered for both of us. I generally say thanks, hence he was expecting it this time. I told him that he was the one who restricted me from eating at the cafeteria. After I graduated, I went to a different country for my higher education. Even though he was not with me, we were still in touch with each other through online chats. During winter break, I flew back to visit my home and met family and friends. He came to pick me up at the airport and helped with my luggage which was nice of him. But when we arrived at my home, he wanted to come in just to see my family and interact with them. I was very confident that I do not want him to know which floor I lived so I asked him to go up to my front door where I literally shouted at him and asked him to leave. During the break, I visited his home once. He asked me to come to his room when his family including his parents were not home. I was happy and surprised that he had bought cloth for me that I liked, so I even thanked him for that. Then he instructed me to lie down on his bed, I thought we were going to cuddle so I did lie down. He climbed into bed after me and suddenly started taking my pants off. I was very perplexed by his reaction and put my hand out to stop him but he did not resist and continued taking my pants off. After a point, he disappeared from the bed for a while and I wondered where he ran to. Still oblivious to the danger I was in. When he came back he had contraceptives with him, forced me down on the bed, and inserted his penis suddenly. I shouted at him and told him it hurts. He said it's hurting because we have not done it in the past few months just because I went abroad for my studies. He then continued until he got the satisfaction that he required. After a few months, I finally broke up our relationship because I found he cheated on me since I flew back abroad. At first, he refused all allegations but then got desperate because I was breaking up with him and admitted that he cheated on me quite a few times. Still, I was very confident that I would never go back to him again and even told him that. I blocked him from all of my social media accounts. He sent me 3 parcels in which he was saying sorry with small gifts inside and I briefly unblocked him to tell him to stop sending me things or bothering me and my social circle before blocking him again. When I flew back home during another vacation, I bumped into him twice in public and felt extremely uncomfortable. Throughout those times he had been desperately trying to make contact through some new and fake social media accounts and again trying to connect with me and my friends. He texted some of my friends who would get back to me regarding this. One time he even went to my friend's place, called her and pushed her to meet him. She was scared and went down to meet him, where she gave him some ideas of what to gift me to apologize again. He even reached out to my mom pushing to meet her for food and wanting to buy me a flight to fly back due to covid era. I was absolutely stressed and sad that he was reaching out to my family and multiple friends and I couldn't do anything about it except guide them on how to handle him. There were many more wrong things he did, these were just the main things I felt that forced me to have an opinion against him. After a whole year, he was still reaching out to talk to me, demanding why I blocked him and trying to prove he did nothing and was not even a cheater. He turned the tables on me saying that he said that he cheated just to see my reaction, to which he found me going out with a guy friend for a drink once. Whenever I think of all this, I feel like I was so screwed up during all those years I spent with him. I never ate healthy food, hardly studied anything, and was constantly under pressure and anxiety. Now, I feel I am safe and in a much better place in all aspects: physically, emotionally, mentally, spiritually, and psychologically. Recently, I am dating a guy in school and he is also the only person who knows everything about my past assault. I know that if I told my parents about my ex, they would be highly stressed and restrict me even for going out. I don't get support from my friend. They think that I was still with my ex even when I did not like him. But they do not know the whole matter. I m very fortunate that I found someone now who understands me, connects well, guide and love me. Thanks to god!

- How demanding was completely reading the story?
  - Not much demanding
  - A little bit demanding
  - Quite a bit demanding
  - Significant demanding
- How demanding was it to construct responses to the given story?
  - Not much demanding
  - A little bit demanding
  - Quite a bit demanding
  - Significant demanding
- **Second section**
- Now, you will do the same thing you just did. Please read the following story and then write a helpful or supportive response for the survivor.

- <One MeToo story with highlights is shown>
- The post you just read contained highlighted sentences. Do these sentences represent some important parts of the story?
  - Yes
  - No
- (If yes to the last question) How would you rate the importance of the highlighted text?
  - Not much important
  - A little bit important
  - Quite a bit important
  - Significant important
- (If yes to the second last question) What information did the highlighted text provide you?
- How helpful was the highlighted text in reading and responding?
  - Not much helpful
  - A little bit helpful
  - Quite a bit helpful
  - Significant helpful
- **Bonus section**
- How demanding was it to read and understand the highlighted story?
  - Not much demanding
  - A little bit demanding
  - Quite a bit demanding
  - Significant demanding
- How demanding was it to construct a response to the highlighted story?
  - Not much demanding
  - A little bit demanding

- Quite a bit demanding
- Significant demanding

- Are stories with highlights easier to read and understand than stories without highlights?
  - Yes
  - No
- Is it easier to respond to stories with highlights than to stories without highlights?
  - Yes
  - No

Table 3.2: Key questions in our survey and summary of responses received. The complete survey is provided in the supplement section.

First section			
Q1	[1,4]	How demanding was completely reading the story?	Mean rating: 3.13, standard deviation: 0.95
Q2	[1,4]	How demanding was it to construct responses to the given story?	Mean rating: 3.10, standard deviation: 0.90
Second section			
Q3	Y/N	Do highlighted sentences represent important parts of the story?	28 of 29 responded Yes
Q4	[1, 4]	(If yes on Q3) How would you rate the importance of the highlighted text?	Mean rating: 3.34, standard deviation: 0.76
Q5	Text	(If yes on Q3) What information did the highlighted text provide you?	"It gives me a quick idea of what the point of the article is" "Grabbed my attention" "Help me to focus on the keywords and better understand the content of this story"
Q6	[1, 4]	How helpful was the highlighted text while reading and responding?	Mean rating: 3.37, standard deviation: 0.62
Bonus section			
Q7	[1, 4]	How demanding was it to read and understand the highlighted story?	Mean rating: 2.41, standard deviation: 1.01
Q8	[1, 4]	How demanding was it to construct a response to the highlighted story?	Mean rating: 2.41, standard deviation: 0.97
Q9	Y/N	Are stories with highlights easier to read and understand than stories without highlights?	23 of 24 responded Yes
Q10	Y/N	Is it easier to construct responses to stories with highlights than to stories without highlights?	23 of 24 responded Yes

Table 3.3: Comparing performance of multiple trained models. Bold indicates the highest score for a metric.

Category	Metric	TF-IDF+ LR	GloVe + RF	Word2Vec + RF	Keyword search	USE + SVM	Sentence-BERT + SVM	RoBERTa	XLNet
Incident	F1 Score	0.64	0.43	0.50	0.47	0.69	0.70	**0.77**	**0.77**
	Recall	0.56	0.31	0.33	0.79	0.63	0.61	0.80	**0.82**
	Precision	0.75	0.71	0.73	0.34	0.77	**0.83**	0.74	0.74
Effects on the survivor	F1 Score	0.66	0.38	0.39	0.49	0.69	0.70	0.78	**0.80**
	Recall	0.60	0.25	0.26	0.82	0.65	0.61	0.80	**0.86**
	Precision	0.73	0.76	0.79	0.35	0.75	**0.84**	0.77	0.74
Advice requested	F1 Score	0.74	0.63	0.64	0.67	0.74	0.80	**0.89**	**0.89**
	Recall	0.70	0.58	0.59	0.93	0.71	0.75	0.91	**0.94**
	Precision	0.78	0.73	0.73	0.53	0.77	0.86	**0.87**	0.84
Macro	F1 Score	0.68	0.48	0.51	0.54	0.71	0.73	0.81	**0.82**
	Recall	0.62	0.38	0.39	0.85	0.66	0.66	0.84	**0.87**
	Precision	0.75	0.73	0.75	0.41	0.76	**0.84**	0.80	0.77

CHAPTER

4

UNDERSTANDING STORIES DESCRIBING CONSENT VIOLATIONS

In Chapter 3, we discussed how natural language can be used to highlight important parts of the MeToo story. Such MeToo stories discuss the abusers' unwelcome behavior, which may be considered as the violation of the survivor's consent (especially in assault cases). According to Singh (2022), understanding consent is complex due to the many cognitive and performative aspects involved.

In particular, sexual affirmative consent is defined as a mutual agreement between all parties (indulging in sexual activity) that is clearly communicated, enthusiastic, and voluntary (Jozkowski 2015; Parenthood 2021). Javidi et al. (2023) extend this consent formulation by articulating it as *FRIES* framework which stands for *Freely given*, *Reversible*, *Informed*, *Enthusiastic*, *Explicit*, and *Specific*. We explain these six criteria of consent below.

Freely given: Participants should have a choice to make without pressure, manipulation, or under the influence of drugs or alcohol. Moreover, minors are not able to consent

Reversible: During sexual activity, participants can change their minds anytime and be free to stop the sexual act

Informed: Participants should be well informed about the whole sexual activity before indulging in it. For example, agreement about using or not using condoms and types of sexual gestures

Enthusiastic: Participants should only do stuff they want to do, not things that they feel they're expected to do.

Explicit: Consent should be clearly communicated between parties before indulging in any sexual activity

Specific: Saying yes to one thing (like going to the bedroom) doesn't mean they have said yes to others (like having sex).

We found that MeToo stories indicate how these criteria of the survivors' consent are violated. Example 12 illustrates the violation of three FRIES criteria (underlined) in a story. As described in the story, the survivor was gaslighted and convinced to have sex, suggesting that consent was not freely given. The survivor said "no" (violation of Explicit) and was not completely aroused or interested during sex (violation of Enthusiastic).

Understanding consent and its violations has been a challenging task. Some researchers argue that providing consent is a mental action, however, others claim it to be a communicative act (Singh 2022; Schnüriger 2018; Hohfeld 1919). Moreover, the dynamics between multiple parties (abusers and survivors) make consent a complex concept (Singh 2022). We understand the stories indicating consent violations by analyzing the language expressed in them. To this end, we leverage FRIES framework (explained above) to label violations expressed in 989 MeToo stories. We call this labeled dataset MyConsent. Each story in MyConsent expresses one or multiple or sometimes no violations (non-physical harassment), making it a rich multilabel dataset. Based on the final labels in MyConsent, we analyze violation stories on multiple linguistic aspects such as LIWC attributes, emotions, named-entity mentions, and so on. We found that a popular large language model is incapable of understanding such violation stories, making this problem a challenging task. To the best of our knowledge, this is the first computational study to understand violation stories from a linguistic perspective.

Example 12: Consent violations expressed in MeToo story

Opinions on my story

I want to know your opinions on my assault and hence decided to post it here. I was 18 and my boyfriend was 16. He was very toxic to me and I was heavily abused by him, both emotionally and verbally. One night he gaslighted and convinced me to indulge in sexual intercourse with him, even after I told him not to do that. As a result, I suffered through lots of pain (not sexually aroused) and bleeding. He reacted to my bleeding as something gross. I have told some people about this incident and they think it's not an assault as he was my boyfriend and also younger than me. I feel disheartened by hearing all these opinions. Hence, I am seeking an unbiased opinion about my case here.

We organize this chapter as follows. Related work on consent is listed in Section 4.1. Section 4.2 explains the curation of our consent dataset, MYCONSENT, based on MeToo stories. While building the dataset, we consider the same ethical guidelines, as described in the previous chapter (Section 3.10). We analyzed a popular large language model for its understanding of violation stories, as described in Section 4.3. The linguistic findings obtained after analyzing such stories are described in Section 4.4. Moreover, Section 4.5 includes our main contributions and limitations along with future directions.

4.1 Related Work

Prior studies focus on formulating the theoretical model of consent. Some studies consider consent to be a mental choice, whereas some consider it to be a communicative act (Hohfeld 1919; Koch 2018; Singh 2022). In the area of computing, there have been multiple frameworks of consent (Singh 2022; Im et al. 2021) covering both objective (communicative act), subjective (mental act), and practical aspects (power dynamics and so on) of consent. However, these approaches don't target the setting of sexual consent.

Sexual consent is defined as an agreement between all participants engaging in a sexual act to be clearly communicated, mutual, enthusiastic, ongoing, and voluntary (Jozkowski 2015; Parenthood 2021). Taking this idea forward, Javidi et al. (2023) designed an informative program to teach adolescents about sexual consent. They use the catch acronym, FRIES, to educate their subjects about all aspects of consent: *freely given, reversible, informed, enthusiastic, explicit, and specific*. Using this framework, they obtained high efficacy when tested on 833 U.S. adolescents coming from various backgrounds. This lends credence to the FRIES framework to be used in future research involving sexual consent.

We adopt the FRIES framework to study consent and stories expressing its violations. Leveraging natural language can uncover how such stories are expressed. As described above, most of the prior studies formulate the consent theoretically, whereas we understand its violations through MeToo stories and NLP techniques.

4.2 Curation of MyConsent

In Chapter 3, we describe how we collected 5,327 relevant stories from three subreddits. For studying consent, reading and labeling vivid experiences of sexual violence is disturbing. Hence, we selected 1,000 stories shorter than the mean length (1,742 characters) for our analysis. In particular, the length of selected stories was between 500 and 1,100 characters.

The task was to read a complete story and label "yes" or "no" for the categories: (i) violation of Freely given, (ii) violation of Reversible, (iii) violation of Informed, (iv) violation of Enthusiastic, (v) violation of Explicit, (vi) violation of Specific, (vii) unknown violation, and (viii) none. If the story describes verbal or other types of harassment (not involving physical touch), it is labeled "yes" for None and "no" for others. Moreover, stories briefly mentioning the incident but focusing more on the survivors' trauma or seeking advice are labeled "yes" for "unknown violation" and "no" for others. All other cases with vivid descriptions are labeled "yes" for one or more FRIES violations. The detailed labeling process is described in the appendix, which led to a total of 1,000 labeled stories for violation of consent. After removing duplicate stories, we were left with 989 of them in MyConsent.

4.3 Incapability of Large Language Models

We analyze if large language models are capable of understanding stories describing consent violations. To begin with, we used ChatGPT, one of the popular large language models that is trained on online text to generate answers to queries. Our analysis goes as follows. First, we described the FRIES framework by explaining definitions of each criterion to the ChatGPT. Second, we randomly selected 10 stories from MyConsent and showed them to ChatGPT. Third, for each story, we asked ChatGPT about all consent violations expressed in language. Fourth, it responded with violated criteria and the explanation for how they were violated.

Our qualitative analysis found that ChatGPT is unable to understand multiple violations through stories. For example, it identifies reversible violation from many stories, which may not hold as per the story description. The following excerpt from a story shows how a survivor could stop the incident while being abused, indicating that reversibility was respected. However, ChatGPT states that the survivor was not given the chance to stop their participation in the sexual interaction.

"... We did not discuss much about sexual interaction and suddenly I was choked and slapped. I could not digest what just happened. Eventually, I got the chance to speak and told him to stop. He did stop and went to bed. However, I was shaken and very perplexed about the situation as I initiated the sexual gesture ..."

Particularly, ChatGPT was confused between explicit and freely given violations. As a result, it considered many cases of pressure and coercion as explicit violation, despite having communication between participants. We illustrate this confusion with an example below. The survivor was gaslighted and pressured to be involved in sex, which is against their consent. Under such circumstances, they made a conscious decision to be sexually involved ("ended up giving in"). Hence, we consider it as a freely given violation but not explicit. We found that ChatGPT could not spot the difference between the two violations.

"He gaslighted me to believe that if I was the one withholding sex from him then I have an unfair power over him. Every week, I felt so pressured and guilty about it that I ended up giving in. I never used to enjoy the process as it was an obligation for me, but technically it was consent because I made a conscious decision?…"

Moreover, for most of the stories possessing informed or specific violations, ChatGPT could not identify such violations. The below story shows how the survivor was not informed and did not consent to the abuser for specifically touching their private parts. However, the abuser still did so. In many cases, ChatGPT could not judge if the survivor was well-informed about sexual interaction and if they consented to a specific sexual gesture.

"…She kept asking for a kiss although I was not sure. I did not expect that she would also touch my private parts without even asking me. A part of me liked this but I always wanted to say no. Am I wrong for liking it? I don't know"

Despite being trained on massive online data, popular large language models are ineffective at understanding violation stories. Our research aims to understand such stories from the language perspective. We build MyConsent dataset and analyze its stories for multiple linguistic attributes, as described in the following sections.

4.4 Findings

4.4.1 Occurrences of Violations

Table 4.1 shows the distribution of categories in the final MyConsent dataset. We found that Freely Given is the most violated criterion as indicated by stories on MeToo-related subreddits. 51.6% stories in MyConsent express this violation. The majority of such cases account for assaulting a minor or someone under the influence, or sometimes due to unequal power dynamics between the abuser and survivor. The criteria of Enthusiastic (18.1%) and Explicit (14.5%) are also prominently expressed to be violated in these stories. Moreover, 27.1% stories lack enough information to identify the correct violation (unknown violation), whereas 6% stories describe verbal or non-physical harassment (None).

Table 4.1: Violations found in MeToo stories of MyConsent.

Violation type	Number of stories found (Percentage)
Freely given	510 (51.6%)
Enthusiastic	179 (18.1%)
Explicit	143 (14.5%)
Specific	78 (7.9%)
Reversible	77 (7.8%)
Informed	63 (6.4%)
Unknown	268 (27.1%)
None	59 (6.0%)

To find frequent co-occurrences of multiple violations, we visualized the correlation as a heatmap, as shown in Figure 4.1. We found that informed and specific violations have a high correlation (0.68) indicating their strong co-occurrence in multiple stories. Such cases generally appear when two participants (survivor and abuser) consent over one act but end up doing another act. The survivor was neither well informed of nor did they consent to the specific act performed. The following excerpts of two such stories (from MyConsent) are provided below. Note that such cases take place not only in intimate relationships but also in professional settings. For example, a patient consents to the gynecologist to check private parts but not to inappropriately touch them.

"…I agreed to have slow and soft sex. But later he went rough and hard. I was in a lot of pain so I wanted him to stop …"

"I went to see a gynecologist some days ago… it was weird that toward the end of the medical examination, he took off his gloves and fingered me …"

Enthusiastic and explicit violations also have a significant correlation (0.43). Such cases are those when the survivors don't say "yes" clearly but are still urged to participate, which makes them feel violated. Even in the absence of a clear "yes", abusers assumed that survivors did consent and went ahead with sexual activity. As a result, instead of being enthusiastic, survivors became uncomfortable, as shown in the two excerpts below.

"My partner and I were watching something and he asked me if I wanted to make out several times and I kept declining it. Maybe I moved over in a way that indicated him my interest so he

directly kissed me. I kissed him back because I thought it's rude not to, but honestly I didn't really want to..."

"*...He asked for my consent and I said I don't think so. I think it may be a fault to not clearly say no. After a few minutes of making out, it just happened. I just froze and became clueless. I couldn't utter a word...*"

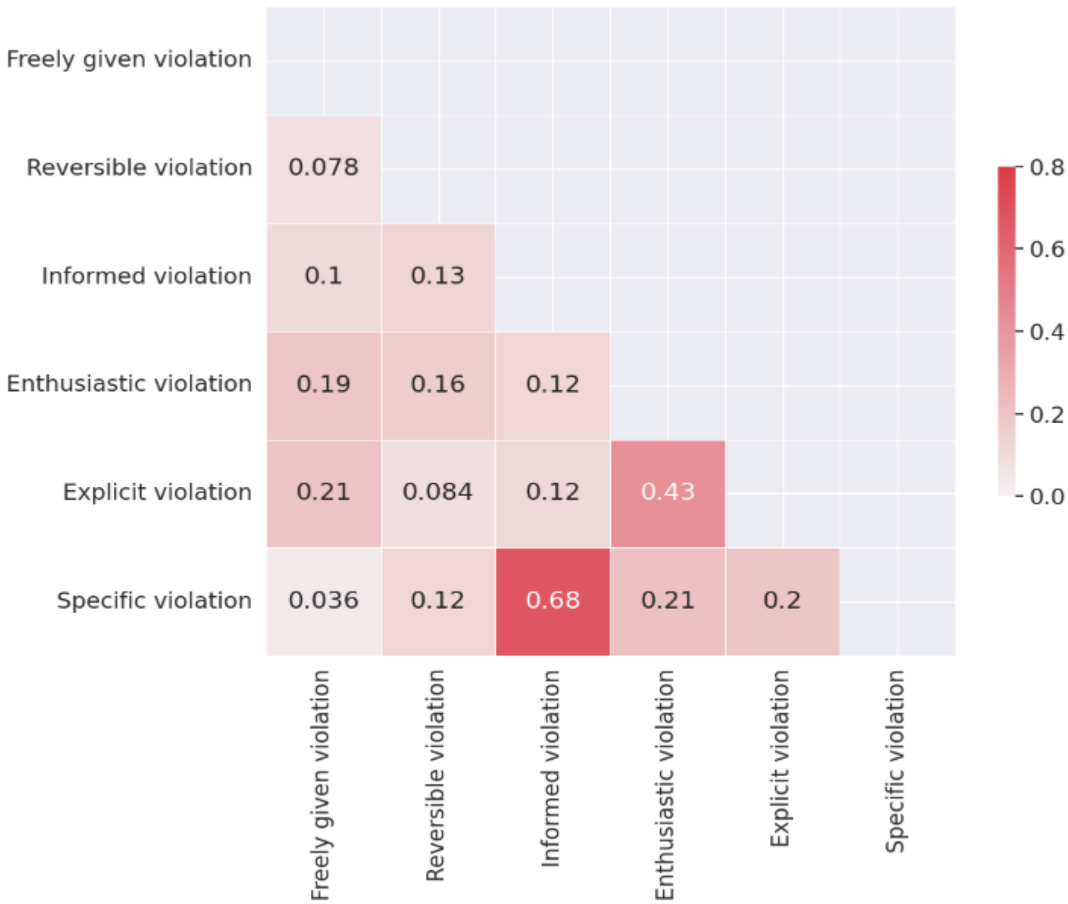


Figure 4.1: Heatmap illustrating correlations found between distinct violations.

4.4.2 Time-Sensitive Attributes

We observed that some stories frequently mention time-sensitive information such as *"she did this when I was 5 year old"*, *"2 years ago my abuse took place"*, *"I was young when all this happened"*, and "He repeated this abuse for the second time we met". Such information defines the time for the incident or the parties involved (ages of survivor and abuser). Using the Named-Entity-Recognition (NER) model (of the spacy module) and synonym keywords (of "young" and

Table 4.2: Violations that are significantly different in the number of time-sensitive attributes.

First violation	Second violation	P-value
Freely given	Reversible	0.021268
Freely given	Informed	0.001181
Freely given	Explicit	0.007368
Freely given	Enthusiastic	0.000968
Freely given	Specific	0.000065

"minor"), we identified attributes in stories revealing time-sensitive information. We computed the number of such attributes for each type of violation and compared these values using the T-test. We report the p-values of the T-test in Table 4.2.

The p-values < 0.05 show significant differences between the two types of violations in the number of such attributes used. To conclude, stories indicating freely given violations mention time-sensitive attributes much more frequently than all other violations.

4.4.3 LIWC Analysis

LIWC-22 toolkit (Boyd et al. 2022) contains 100 in-build dictionaries, where each dictionary consists of lexicons, emoticons, and other verbal constructs to identify the psychological aspects from a story. We applied LIWC-22 on MyConsent and compared the stories on different types of scores. We show our analysis below.

Sometimes, consent violations are expressed through negations in the language. Through the LIWC-22 toolkit, we identified the percentage of negation words (such as "no", "couldn't", "shouldn't", and so on) in each story. Table 4.3 indicates the average percentage of such words in each type of story. We found that negation words are dominantly present in stories of (i) reversible violation to show the victims' inability to reverse the situation, (ii) enthusiastic violation to show the victims' lack of interest, and (iii) explicit violation illustrating to show denial of sexual proposal from the victims' side. An example of the story including negations for two of these three violations is shown below. The first mention of negation indicates the inability to reverse the act, whereas the second shows explicit violation.

"I wanted to stop, but he covered my mouth so I couldn't say anything. I was terrified during that time because I had no control over me. It all started when he asked me if I wanted to which I said no . . . "

For a story, LIWC-22 (Boyd et al. 2022) yielded four types of summary scores: *analytic* (depicting analytical and logical thinking patterns), *clout* (depicting social status, confidence, or leadership), *authentic* (depicting how much people reveal about themselves, without any

Table 4.3: Average mentions of negations in each type of story.

Violation type	Average score
Reversible	3.19
Enthusiastic	3.11
Explicit	3.04
Unknown	2.79
Specific	2.74
Freely given	2.73
Informed	2.67
None	2.23

Table 4.4: Average clout score for each violation type.

Violation type	Average clout score
Freely given	10.35
Reversible	07.41
Informed	12.49
Enthusiastic	07.91
Explicit	08.42
Specific	12.54
Unknown	10.74
None	21.93

filter), and tone (depicting the emotional tone). On MYCONSENT, we computed these summary scores (using the LIWC-22 toolkit) for each story. Further, for each category, we averaged these scores. Table 4.4 shows category-wise average clout scores. All violation categories (including unknown) have low average scores for clout, as compared to the None category[1] (which includes non-physical harassment cases). This indicates less leadership, confidence, and social status in the stories describing assault than the ones describing non-physical harassment. This is because the assault cases may be more severe than verbal or non-physical harassment cases.

On the contrary, we found cases with violations to have high authenticity as compared to None, as shown in Table 4.5. Violation stories are shared without any filter. Such stories are open to reveal survivors and how they felt due to the incident. We did not find any conclusive insight from the LIWC-22's fourth summary variable.

[1]P-values indicate significant difference

Table 4.5: Average authenticity score for each type of violation.

Violation type	Average authenticity
Freely given	80.70
Reversible	83.83
Informed	83.15
Enthusiastic	85.57
Explicit	86.81
Specific	83.25
Unknown	80.39
None	65.01

4.5 Discussion

We investigate MeToo stories describing sexual harassment and assault for multiple categories of consent. Stories describing violations are hard to understand even through popular large language models. First, we develop a dataset, MyConsent, containing 989 stories labeled for FRIES violations. Our dataset can be used by future research in this domain. Second, we linguistically analyze stories in our dataset through tools such as NER and LIWC-22. We found that freely given is the most prominent violation. Stories expressing this violation mention time-sensitive attributes more frequently as compared to other stories. Moreover, there is a significant interplay between (informed and specific) and (explicit and enthusiastic) violations. LIWC-22 analysis (Boyd et al. 2022) helped us to infer how the language in violation stories indicates low confidence but high authenticity, as compared to other stories. Moreover, all these stories are emotional and generally reference male entities.

4.5.1 Limitations and Future Directions

Our work has a few limitations which give rise to future directions. First, our dataset contains stories from the Reddit platform; however, in the future, stories on other platforms can also be explored for consent violations. Second, it would be interesting to deeply understand the stories in the “unknown” category for possible violations. This could be done by having a large database of MeToo stories and predicting possible violations (for the “unknown” category) using the limited information given in such stories. Third, many stories indicate that survivors are confused about whether the incident was an assault. Building computational tools predicting consent violations will help in automatically answering such questions based on the violated criteria.

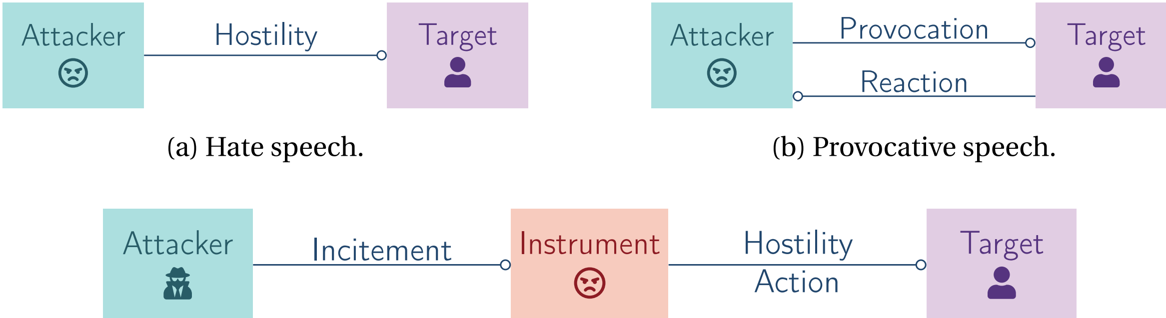
CHAPTER

5

UNDERSTANDING INCITING SPEECH AS NEW MALICE

Trigger warning: This chapter includes examples of inciting posts that may be disturbing to some readers.

Background: Inciting speech seeks readers to instill hostility or anger, or motivate them to take action, especially against a religious group, potentially leading to violence. Whereas hate speech in social media has garnered much attention from researchers, inciting speech has not been well studied. **Problem:** We address two aspects of religious incitement: (1) What rhetorical strategies are used in it? (2) Do the same strategies apply across disparate social contexts and targets? **Data and method:** We first identify inciting speech against Muslims but demonstrate the generality of the construct vis à vis other targets. We adopt existing datasets of Islamophobic WhatsApp posts and hateful and offensive posts (Twitter and Gab) against other targets. Our methods include (1) qualitative analysis revealing rhetorical strategies, (2) an iterative process to label the data, resulting in a computational tool to detect incitement. **Findings:** We found three rhetorical strategies in incitement that focused respectively on the target group's identity, their imputed misdeeds, and an exhortation to act against them. We showed that the three rhetorical strategies carry distinct textual signatures. We showed by applying our tool (with additional verification) that inciting sentences appear in non-Islamophobic posts and in other



(a) Hate speech.

(b) Provocative speech.

(c) Inciting speech.

Figure 5.1: Three kinds of malice on social media. Hate speech involves an attacker expressing hatred toward a target. Provocative speech involves an attacker provoking a target to elicit a reaction from them (such a reaction may have the target lose face or otherwise advance the attacker's agenda). Inciting speech involves the attacker riling up an intermediary as an instrument to express hostility or act maliciously against the target. We focus on inciting speech.

contexts (e.g., posts against certain gender identities), indicating the generality of incitement as a concept. **Implications:** Characterizing incitement reveals a large swath of malicious speech omitted from traditional analyses. Understanding and identifying inciting speech can facilitate online moderation and, in turn, reduce concomitant harm in real life.

5.1 Introduction

We investigate *inciting speech* as a distinct type of malicious communication on social media. Previous works (Łubinski 2017; Alexander 2001) on inciting speech focus on violence against a target. However, our characterization respects the broad definition given by the International Covenant on Civil and Political Rights (ICCPR 2023). Specifically, we consider inciting speech as speech that stimulates its readers' hostility or anger or urges them to act (e.g., discrimination or violence) against a target.

Interestingly, inciting speech on social media has not been intensively studied, being largely overshadowed by *hate speech* (Calvert 1997). We posit that inciting speech is a distinct category of malicious speech. In comparison to other malicious speech, inciting speech is different based on at least one of the two factors: (1) *structure* and (2) *language*.

Figure 5.1 shows hate (UN 2023) and provocative (Boudana and Segev 2017) speech are *structurally* different because they involve only two parties (an *attacker* hating or provoking a *target*), whereas in incitement the attacker incites an *instrument* (a reader not from the target group) to elicit the instrument's anger, hostility, or action against the target.

Based on *language*, inciting speech is different from others in terms of the words or nature of language (direct or indirect) used in it. For example, hate speech is typically understood and computationally recognized using derogatory language such as *n**ger* and *a**hole* (ElSherief et al. 2018). Inciting speech doesn't necessarily contain such derogatory or abusive words, which makes its computational detection harder than hate speech. Our analysis digs deep into the language aspect and draws a comparison between inciting and other malicious speech.

The past few years have witnessed harassment and violence against multiple religious groups such as Christians, Hindus, and Muslims (Pew 2023; HAF 2023). Inciting speech on social media may lead to an increase in such antisocial behavior in the real world. Hence, we primarily focus on incitement targeting a religious group. In addition, we find the applicability of incitement constructs in other social domains.

Saha et al. (2021) conduct a study of "fear speech" on WhatsApp posts discussing politics. Their dataset identifies some posts as Islamophobic (i.e., indicating fear speech as these posts instill fear or threat of Muslims in their readers) and others as non-Islamophobic. Example 13 shows an Islamophobic and a non-Islamophobic post from Saha et al.'s dataset. Importantly, *both* illustrate inciting speech. The underlined snippets either typecast Muslims or highlight their alleged practices as misdeeds, seeking to engender hostility toward them.

Example 13: Inciting speech against Muslims

Islamophobic post

"Leave chatting and read this post or else all your life will be left in chatting. In 1378, a part was separated from India, became an Islamic nation - named Iran…and now Uttar Pradesh, Assam and Kerala are on the verge of becoming an Islamic state…<u>People who do love jihad is a Muslim. Those who think of ruining the country - Every single one of them is a Muslim !!!! Everyone who does not share this message forward should be a Muslim…</u>"

Non-Islamophobic post

"…Sharemarket's rise and fall is not the right indicator of the actual economy. The unprecedented rise in Index is not going to happen soon -the real rate of return will be around 12–15%. Govt next agenda besides solving the Article 370 problem is to bring a uniform civil code and it will solve a lot of issues in India in the long run. *Saving economy in the context of triple talaq bill* <u>Indirectly due to triple talaq, the hidden agenda of mullahs was flourishing– 2–3 kids from the first wife, 2–3 kids from the second wife. *How much it is gonna drain our strained economy/resources * or have been draining the economy in the past. These were sinister agendas of Mullahs in India funded by a few neighboring countries to change the demographic pattern of India.</u> Due to vote bank appeasement politics none of the political parties thought of its economic constraints ... ”

Further, we analyze inciting speech against other targets such as gender, sexual orientation, and so on. We adduce evidence throughout this chapter that inciting speech is a distinct category from hate and fear speech. In particular, inciting posts in Example 13 don't use derogatory language (such as abusive words) and thus are not classified as hate speech (ElSherief et al. 2018). Also, one of these posts is labeled non-Islamophobic (non-fearful) by Saha et al. We defer additional discussion of the literature on malicious speech (such as Islamophobia, provocation, hate speech, and fear speech) to Section 5.7.

5.2 Dataset and Main Steps

Saha et al. (2021) studied Islamophobia on public WhatsApp groups in India. They found that groups focused on politics also included Islamophobic posts. They shared a dataset of 4,782 WhatsApp posts, out of which they identified 1,142 posts as Islamophobic.

We introduce the notion of inciting posts and build on Saha et al.'s public dataset. First, through qualitative analysis of their posts, we identify three rhetorical strategies used in inciting speech. Second, we provide INCITE, a dataset of 7,000 sentences labeled for the three rhetorical strategies and None. These sentences are extracted from some of the WhatsApp posts collected by Saha et al. (2021). We train a computational model for classifying sentences according to these strategies. Third, we apply our computational model on all 4,782 posts in Saha et al.'s dataset. Interestingly, we find that inciting speech occurs not only in Islamophobic but also in non-Islamophobic posts, indicating that the established notion of fear speech *misses* an important swath of malicious speech in social media. Our model is trained on incitement against Muslims, however, it can be fine-tuned to identify such speech against other religious communities such as Hindus. Fourth, we study the prevalence of the identified strategies outside the religious context. To this end, we generalize and apply our computational model to the HateXplain dataset, which includes hate or offensive speech from Twitter and Gab (i.e., not limited to India or Islamophobia). Throughout, we manually spot-check the results of our computational model.

Most importantly, the evidence shows that incitement appears in non-hateful or non-offensive text, indicating that inciting speech is a valuable construct missed by current analyses. We describe all technical details, including curation and training, and evaluation metrics in the Supplementary Section 5.9.

5.3 Rhetorical Strategies in Incitement

We adopted thematic analysis (coding reliability) (Nowell et al. 2017) to understand how inciting speech is expressed on social media. To this end, we analyzed more than 100 Islamophobic posts (from Saha et al.'s Islamophobia dataset) and identified three prevalent rhetorical strategies in such posts. We observed that a post may exhibit multiple strategies but a sentence typically has no more than one. Therefore, we focus on individual sentences to make our analysis more robust; a post can be considered inciting if it contains at least one inciting sentence. We describe the three rhetorical strategies we found below.

Identity Sentences that criticize or stereotype the target group, based on the group's beliefs (including their scriptures), pratices, or leaders. Example 14 illustrates this rhetorical strategy, which may be realized through explicit defamation or implicature through disingenuous rhetorical questions.

Example 14: Rhetorical strategies: Identity

"...Hindu-Muslim unity is impossible because the Muslim Quran does not tolerate Hinduism as a friend."

"Now you can think that whom will these Muslims serve as IAS ...Islam or country?"

"The testimony of Kafir (non-Muslims) in Sharia court, ie Qazi court, is not valid"

"...Which superstitious religions are in India * * Yes Islam is the anti-human religion of superstition * * 1) If Allah , God is equally powerful * * There is God and he hates Kafiro * * And Kafiro does not have the right to live"

Imputed misdeeds Sentences highlighting past or current misdeeds such as oppression or violence (allegedly) perpetrated by the target group. Example 15 illustrates this rhetorical strategy, which may be realized through accusation or indirectly presenting fake or genuine incidents as "facts."

Example 15: Rhetorical strategies: Imputed misdeeds

"...Only Hindu temples are destroyed and business and other activities are being increased in their place and Hindus are also being attacked!"

"Entered, broke the Shivling into pieces, and acquired as much property as he could in the solution"

“Muslim doctor silently sterilized 4000 Buddhist women … ”

Exhortation Sentences that urge readers to act against the target. Such actions include violence, discrimination, or boycotts. Example 16 illustrates this rhetorical strategy, which may be realized through explicit commands or subtle hints about what must be done.

Example 16: Rhetorical strategies: Exhortation

“Hindu society should stop worshiping these tombstones, tombs, pirs.”

“Be violent for religion”

“If the madrasa (Islamic school) is not closed, after 15 years, more than half of the Muslims of the country will be supportive of the ISIS ideology”

While curating INCITE, our dataset containing 7,000 sentences, we validated that incitement follows only the above three strategies. While labeling these sentences for incitement strategies (and None), we considered the context by also reading their preceding and succeeding sentences in the posts. All details about this labeling and data curation are described below.

5.3.1 Curation of INCITE

As shown in Figure 5.2, our quantitative method includes two phases. First, we used NLTK (2023), a natural language processing toolkit, to split Saha et al.’s Islamophobic posts into sentences. We curated INCITE, a dataset of a sample of 7,000 inciting sentences. Second, we trained and evaluated multiple embeddings-based and transformer-based models over five-fold cross-validation of INCITE. Then, we chose the best-performing model for the identification of incitement sentences from posts. We describe the first phase of labeling INCITE in the appendix, whereas the details of the training computational model are provided in the Supplementary Section 5.9.

Out of 7,000 labeled sentences, 1,065 (15.21%) belonged to identity, and 1,579 (22.55%) represented imputed misdeeds, and 433 (6.19%) represented exhortation. Moreover, 3,923 (56.05%) sentences did not belong to any of these strategies (None). This is because the proportion of malicious speech is generally low in a sample, also corroborated by other studies (Yin et al. 2023; Saha et al. 2021). Our computational model (technical details in the Supplementary Section 5.9) trained on INCITE is leveraged for further identification and analysis of inciting speech, as described later in this chapter (Sections 5.4.3 & 5.5).

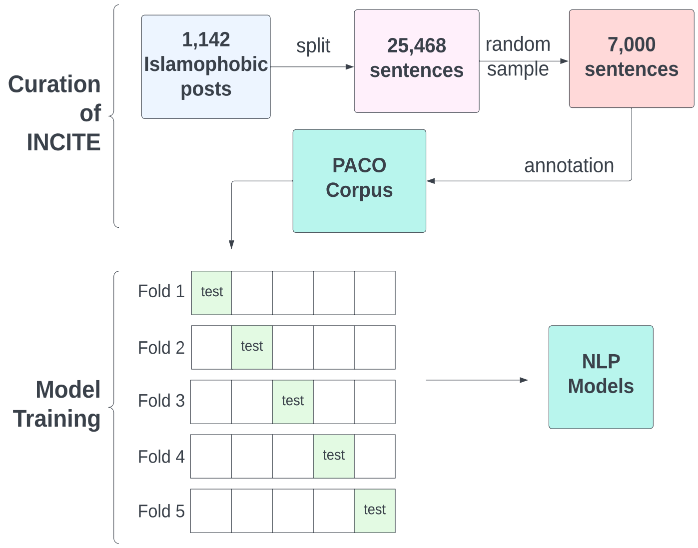


Figure 5.2: Overview of our method. Curating a dataset of inciting sentences followed by training multiple models and choosing the best one.

5.4 Analyzing Incitement Strategies

We now take a closer look at the above three incitement strategies to establish their robustness as constructs.

5.4.1 Generality of the Incitement Strategies

Although our study primarily focuses on religious incitement, we also find the prevalence of three rhetorical strategies in other domains. As a result, such strategies are generic and apply to a variety of targets. To show their general applicability on social media, we examined Mathew et al.'s HateXplain dataset, which includes Twitter and Gab posts along with a label of whether each post is hateful or offensive speech and whom it targets. We found many inciting posts in HateXplain, addressing a variety of targets, of which Example 17 shows three instances. The first post in Example 17 exemplifies two strategies: identity and imputed misdeeds. The second post exemplifies all three strategies. The third post exemplifies the identity and exhortation strategies.

Example 17: The strategies apply against multiple targets

Targeting whites (Identity, imputed misdeeds)
"<number> year old white hat boys are the worst breed. All they do is recycle their same old holocaust racist sexist and homophobic jokes to hide their insecurities. I got news for you buddy it not working"

Targeting zionists (Identity, exhortation, and imputed misdeeds)
"Zionism is the epitome of evil. We must demand that US congress ceases the funding of the atrocities committed by zionists against Palestinian Gazans Lebanese. Any other Muslim country or people of Iran will never bow to Israel or Saudi Arabia the world will not either."

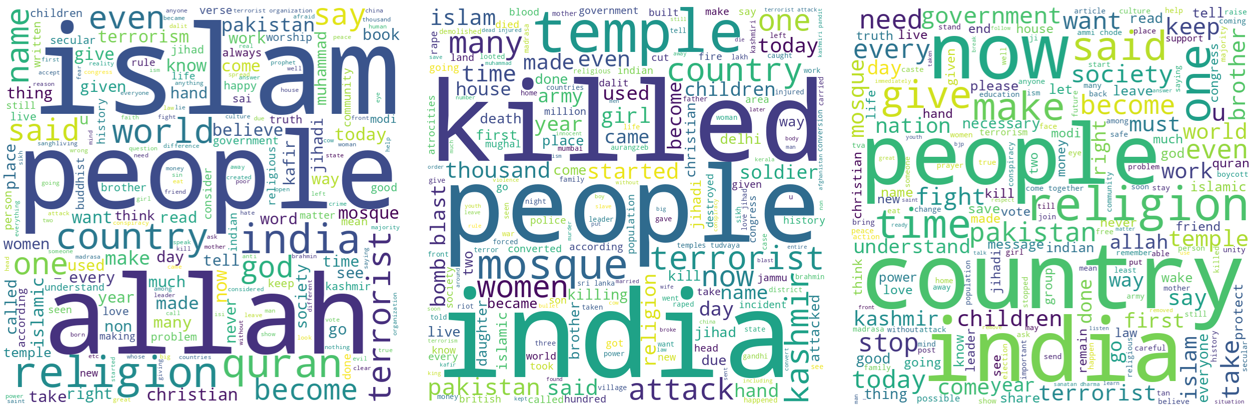
Targeting nonwhites (Identity, exhortation)
"Studies show non-whites view freedom completely different than whites, for example, <number> of Hispanics view hate speech as an act of violence which means they would ban it ... their view of hate speech could be anything, wake up your family and friends"

Importantly, the posts in Example 17 do not fall under hate or offensive speech (as corroborated by Mathew et al.'s annotations), again demonstrating inciting speech as a distinct category.

5.4.2 Textual Signatures of Incitement Strategies

Linguistically, each rhetorical strategy can be characterized by its most frequent words, which we call *textual signatures*. We obtain these textual signatures for sentences in INCITE. First, to be context natural, we removed prominent words such as *Hindus*, *Muslims*, *India*, *people*, and *country* from these sentences. Second, we visualized the word clouds for each rhetorical strategy, as shown in Figure 5.3.

For the identity strategy (Figure 5.3a), words such as *allah*, *Islam*, *quran*, and *mosque* indicate the target community's beliefs and practices along with pejoratives such as *terrorist* and *jihadi*. For imputed misdeeds (Figure 5.3b), we found verbs showing violence such as *killed*, *looted*, *rape*, and *attack*. Nouns such as *children*, *daughter*, *women*, and *temple* indicate the victims or sites of the alleged misdeeds. For exhortation, action verbs such as *make*, *stop*, *fight*, *take*, *support*, and *boycott* are prominent. Moreover, words such as *now* and *today* convey an urgency to take action. This shows that the textual signatures are majorly distinct (despite some common words) across three strategies.



(a) Identity. (b) Imputed misdeeds. (c) Exhortation.

Figure 5.3: Prominent words found for each rhetorical strategy.

5.4.3 Distributions of the Incitement Strategies

We first applied our computational model (model details in the Supplementary Section 5.9) to identify inciting speech from all posts in Saha et al.'s dataset. To this end, we took their entire set of 4,782 posts and split them into 108,319 sentences. Our model predicted the three rhetorical strategies or None for each sentence. A total of 29,755 inciting sentences were identified (9,374 for identity, 15,849 for imputed misdeeds, and 4,532 for exhortation); the remaining 78,564 being None. As corroborated by other studies (Yin et al. 2023; Saha et al. 2021), the proportion of malicious instances is found low, which makes the detection task even more challenging and important.

Figure 5.4 shows the relative distributions of these incitement sentences in both Islamophobic and non-Islamophobic posts. For quality, we spot-checked the model's predictions as described later in this section.

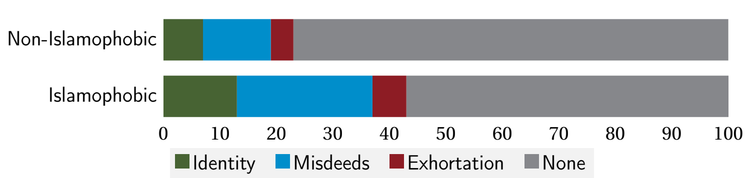


Figure 5.4: Relative distribution of incitement sentences from Islamophobic and non-Islamophobic posts in Saha et al.'s dataset. Though, Islamophobic posts show a higher prevalence of inciting sentences, non-Islamophobic posts also show a substantial (∼23%) proportion of such sentences.

We define inciting posts as those with at least one inciting sentence. We found a total of 3,972 inciting posts (1,102 Islamophobic and 2,870 non-Islamophobic), out of all 4,782 posts (1,142 Islamophobic and 3,640 non-Islamophobic) in Saha et al.'s dataset. Example 13 shows a non-Islamophobic but inciting post. Saha et al. don't label such posts as Islamophobic because such posts don't portray Muslims (or Islamic traditions) as a threat. However, these posts seek to elicit hostility or action against the target. For example, the underlined sentences in the second part of Example 13 criticize Muslims (through the use of "mullah" as a pejorative) for having large families (with a side attack on their allegedly poor values in terms of divorce) and impute responsibility on them for draining the country's economy (along with another side attack based on their purportedly high relative population growth). The attacker even describes this behavior of the target as a "sinister agenda." Since such posts show hostility toward the target, they are inciting in nature.

We conducted a spot-check of 100 inciting posts, 50 randomly selected from both Islamophobic and non-Islamophobic posts. We reviewed these 100 posts and determined that 49 Islamophobic and 39 non-Islamophobic posts were inciting: they elicit hostility or urge readers to act against Muslims.

For each post in Saha et al.'s dataset, we computed the percentage of identified inciting sentences in it, as Figure 5.5 shows. Although the prevalence of inciting sentences is higher in Islamophobic posts, what is surprising is how high it is for non-Islamophobic posts—again indicating that inciting speech differs from fear speech.

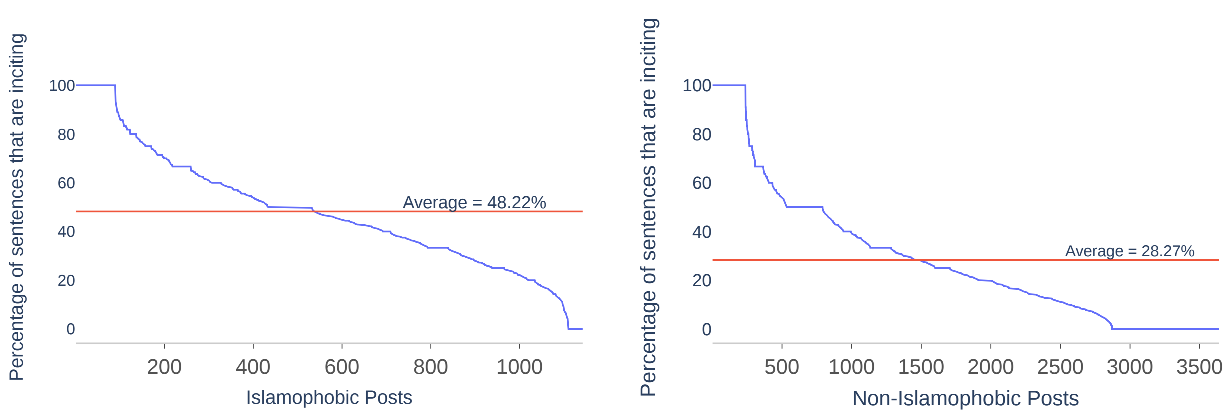
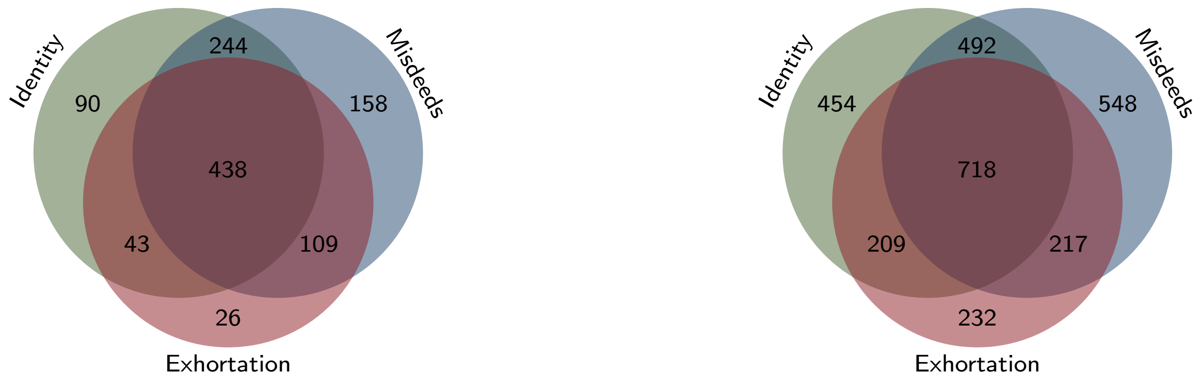


Figure 5.5: Percentage of sentences that are inciting (in descending order) found in Islamophobic and non-Islamophobic posts. On average, our model identified 48.22% of the sentences in Islamophobic posts and 28.27% of the sentences in non-Islamophobic posts as inciting.

Further, many inciting posts include two or more inciting sentences that instantiate distinct rhetorical strategies. Figure 5.6 shows that the rhetorical strategies frequently co-occur. Many inciting posts leave little to no chance by applying multiple strategies. Basically, over 75% Islamophobic posts contain at least two, and 39% contain all three rhetorical strategies. On the other hand, 57% non-Islamophobic posts have at least two and 25% have all three strategies.



(a) Over 75% (438+244+43+109) of the 1,108 Islamophobic posts involve two or more strategies and 39% (438) involve all three.

(b) Over 57% (718+492+209+217) of the 2,870 non-Islamophobic posts involve two or more strategies and 25% (718) involve all three.

Figure 5.6: Distributions of inciting posts (identified by our model) across rhetorical strategies.

5.5 Inciting Speech Against Other Targets

We tweaked our computational model to check the prevalence of inciting speech against other targets. We describe details of this model variant in the Supplementary Section 5.9. We apply this variant on HateXplain (Mathew et al. 2021), a dataset rich in hate and offensive speech extracted from Twitter and Gab that covers multiple targets. In Figure 5.1, we discussed how hate speech is structurally different from inciting speech, whereas, in this section, we focus on the language aspect of malicious speech.

HateXplain contains 20,148 posts. After removing posts with undecided labels (posts for which Mathew et al.'s annotators did not agree) and infrequent targets, we were left with 15,905 posts. Out of these 15,905 posts, our model predicted 7,102 posts to be inciting. The evaluation of our model on the HateXplain dataset is described in the Supplementary Section 5.9. Figure 5.7 shows the relative distribution of incitement sentences for hateful or offensive and other posts.

Targets for these inciting posts are based on ethnicity, gender, sexual orientation, and so on.

Of these 7,102 posts predicted as inciting, 2,409 posts do not contain hate or offensive speech (according to Mathew et al.), corroborating our finding that inciting speech is a distinct category of malicious speech. For example, "Well this gives more credence to the hypothesis that Zionism is behind the Muslim invasion to add strength and credence to Zionism and Israel expansionism" incites readers against Zionists but fails to be identified by traditional malicious speech (Mathew et al. 2021). The other 4,693 posts are both considered inciting and hateful or offensive. For example, "guilty until proven innocent unless it's a jew or a n**ger [redacted], or a lover of both" shows hatred toward Jews and Blacks (through abusive words) and incites readers by criticizing the alleged advantage taken by both groups.

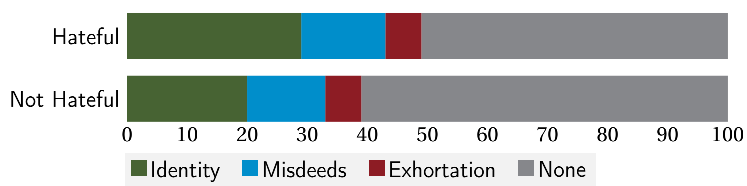


Figure 5.7: Relative distributions of incitement sentences for hateful (including offensive) posts and others. Not surprisingly, Hateful or offensive posts show a higher prevalence of incitement sentences, yet other posts also show a substantial (~39%) proportion of incitement sentences.

HateXplain identifies the target group for hateful or offensive posts. Figure 5.8 shows the distribution of targets across posts that are both inciting and hateful or offensive.

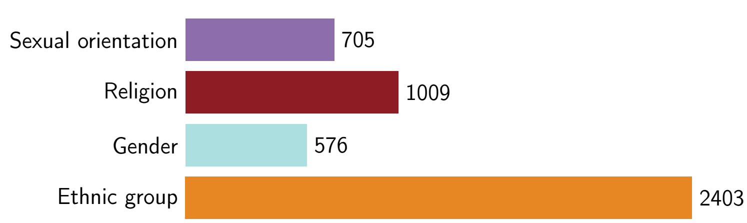


Figure 5.8: Distribution of targets across 4,693 posts that are *both* inciting and deemed hateful or offensive by Mathew et al. (2021). Recall that HateXplain provides targets only for posts deemed hateful or offensive.

5.6 Contributions and Novelty

We identified incitement as an under-studied but important form of antisocial communication prevalent on social media and conducted the first computational study of it. Using qualitative analysis, we found that three rhetorical strategies are used for religious incitement, also showing that they are illustrated in diverse ways. Despite that, these strategies are distinct in their textual signatures. Our dataset INCITE can be leveraged for future research on religious incitement. Overall, our study contributes to the realm of malicious speech used in a religious context (Saha et al. 2021; Chandra et al. 2021).

Using our computational tool, we identified these strategies from WhatsApp posts. Moreover, the tool enabled us to show that leading datasets having non-Islamophobic and non-hateful or non-offensive posts contain a substantial amount of inciting speech. As a result, we could demonstrate the prevalence of inciting speech toward other targets (such as ethnic groups, gender, and sexual orientation), which may go unnoticed by hate-speech detectors.

5.7 Related Work

In this section, we discuss existing studies related to various malicious speech such as hate, provocation, dangerous, Islamophobia, and fear speech. We also show how our study is different from these studies.

5.7.1 Hate Speech

There is no all-encompassing definition of hate speech (Benesch 2014). Typically, hate speech is considered abusive speech and a direct serious attack on an individual or a group based on attributes such as race, ethnicity, religion, and sexual orientation (Sellars 2016; ElSherief et al. 2018). Due to its direct and abusive attacks, hate speech often contains derogatory words such as *n*gger* and *a**hole* that are used against the target (ElSherief et al. 2018). However, inciting speech doesn't necessarily have derogatory words. For example, *"Question- What is non-violence. . . ?? Answer: Bakra Eid"* uses sarcastic language to incite readers against the Muslim tradition of sacrificing goats ("Bakra" in Hindi) during the Eid-ul-Adha celebration. As a result, this sentence falls under the identity strategy of incitement but not hate speech.

Many studies (Yin et al. 2023; Chiril et al. 2022; Aluru et al. 2021; Das et al. 2022; Bohra et al. 2018; Gambäck and Sikdar 2017; Zhang et al. 2018; Nobata et al. 2016; Park and Fung 2017; Del Vigna et al. 2017; Vitiugin et al. 2021) aim to understand or identify hate speech from text but inciting speech is not well-studied.

Moreover, hate speech (UN 2023) is defined between two parties, *attacker* and *target*. However, inciting speech involves an instrument (non-target) being hostile, angry, or taking action against the target. In the same way, incitement differs from other studies involving two-party situations such as cyberbullying (Huang et al. 2018; Singh and Hofenbitzer 2019).

5.7.2 Provocative Speech

Boudana and Segev (2017) conceptualize provocation as a communication strategy, "a speech that elicits a nonviolent, beneficial, or a violent response from the target." The literature (Boudana and Segev 2017; de Castro et al. 2012; Champion and Clay 2007) on provocation emphasizes its definition in a two-party environment, unlike inciting speech as introduced here.

Moreover, to the best of our knowledge, no prior computational study seeks to understand rhetorical strategies in a speech that elicits a response.

5.7.3 Dangerous Speech

Dangerous speech refers to the text that can invoke violence against a group (Speech 2023). Such violence-invoking cases overlap with the exhortation strategy. For example, *"Someday Hindus should be ready to fight against Muslims."* falls into both dangerous speech and exhortation. However, unlike dangerous speech, exhortation includes not just violence but also supporting groups opposing a target and discriminating against the target's businesses or possessions. For example, *"You buy only from Hindus like all your festivals, etc."* asks readers to boycott other businesses. Such sentences indirectly urge readers to discriminate (potentially against Muslims) without being dangerous speech.

5.7.4 Islamophobia and Fear Speech

The dictionary meaning of Islamophobia is fear of, dislike or hate toward, and discrimination against Islam and Muslims (Webster 2023). In computational social science, most studies on Islamophobia consider only dislike or hate (Sindoni 2018; Vidgen and Yasseri 2020; Chandra et al. 2021). Such studies cover the cases overlapping with the identity strategy, however, still miss the other two incitement strategies.

Other studies on Islamophobia focus on fear or threat aspects (Saha et al. 2021, 2023). In particular, Saha et al. (2021) study Islamophobic (also called fearful) messages posted on WhatsApp. They curate 27,000 posts from Indian public WhatsApp groups and find ~8,000 of them to be Islamophobic through manual annotation and similarity hashing (Gionis et al.

1999). We adopt their dataset to find inciting sentences. Sometimes the same sentence may induce hostility and fear (or threat). However, many inciting sentences such as *"Muslims call Jinnah great ... They should be ashamed of something"* predominantly instill hostility and don't overlap with fear speech. In addition, fear speech doesn't cover exhortation.

5.8 Discussion

We identify *inciting speech* as a form of malicious speech that is structurally distinct from hate speech and provocative speech. We demonstrate that inciting speech is prevalent on social media and yet linguistically undetected by hate and fear speech tools: specifically, in leading social media datasets. This chapter doesn't emphasize that a religious group always targets another, however, we acknowledge the prevalence of such inciting speech against all religious groups (HAF 2023; Pew 2023). A potential benefit of identifying incitement is that doing so can help in online moderation and resist readers from incoming inciting messages. We now describe the conclusion and limitations of our study and propose future directions of research.

5.8.1 Conclusion

Previous studies focus on identifying hate and fear speech. To the best of our knowledge, there is no existing study on understanding strategies used in incitement and computationally identifying them. We found these strategies in the religious context and their prevalence against many other targets, indicating inciting speech as a distinct category.

We develop INCITE, a dataset of 7,000 incitement sentences for the three rhetorical strategies. Further, we used INCITE to train and evaluate multiple NLP models (embedding-based and transformer-based) over five-fold cross-validation (Supplementary Section 5.9). Our best-performing model, RoBERTa achieves an average of 0.851 AUC score. The automatic identification of inciting sentences can prevent the spread of religious incitement and, in turn, prevent possible violence against the target.

We will release our dataset using FAIR guidelines Wilkinson et al. (2016), by registering it online, receiving its unique identifier and license, and sharing files for interoperability.

5.8.2 Limitations and Future Work

We primarily focus on religious incitement and train a computational model to identify it. We also found the prevalence of the same rhetorical strategies against other targets. However, building a cross-domain model for all targets is out of the scope of this work. Moreover, there can be other rhetorical strategies specific to non-religious domains, which can be explored in future work. We envision our study as the first step toward understanding inciting speech and hence can accelerate many future research in other social domains. Further, NLP models can be built with a broad vision of identifying all three: inciting speech, fear speech, and hate speech. Such models will serve as a one-stop solution to identify malicious text on social media. Moreover, it will be interesting to conduct an affective analysis of all types of malicious speech on multiple platforms.

5.8.3 Ethics Note and Broader Perspectives

We labeled the sentences present in posts from Saha et al. (2021). Such posts did not include any private information of the WhatsApp users who wrote them. Moreover, since these posts were part of public WhatsApp groups, neither Saha et al. (2021) nor we were required to obtain the consent of WhatsApp users. We acknowledge that the nature of the text can be disturbing. That's why we did not hire crowd workers to annotate sentences. Instead, two researchers in this project, who were aware of the nature of the text, completed the annotation.

Broadly, we envision a positive impact of this research, particularly accelerating future research on inciting speech, using our dataset and rhetorical strategies found.

5.9 Supplementary Materials

5.9.1 Model Training and Evaluation

Our computational model is geared toward identifying incitement against Muslims. We considered the problem of identifying inciting sentences a multiclass classification task, for which we explored multiple training approaches. From INCITE, each sentence was input to a model and the output was one of the four classes: three rhetorical strategies and None. We discuss multiple approaches and compare their performances below.

TF-IDF weighs each word in the corpus according to its Term Frequency (TF) and Inverse Document Frequency (IDF) (Cahyani and Patasik 2021). Based on the number of unique tokens in INCITE, TF-IDF yielded a 11,441-dimensional embedding for each sentence. We provided these embeddings as input to multiple classifiers, including Logistic Regression (LR) (Dreiseitl and Ohno-Machado 2002), Random Forest (RF) (Zakariah 2014), and Support Vector Machine (SVM) (Cervantes et al. 2020), and compared their performance.

Word2Vec converted each word in INCITE into a 300-dimensional embedding (Mikolov et al. 2013). We obtained such word embeddings using the Word2Vec model pre-trained on the Google News dataset. To obtain a sentence embedding, we averaged Word2Vec embeddings for all words present in that sentence. Further, sentence embeddings were provided as inputs to LR, RF, and SVM.

GloVe yields an embedding for each word in the corpus (Pennington et al. 2014). We used Stanford's GloVe model, which is trained on the Wikipedia dataset and returns a 100-dimensional word embedding. We averaged these word embeddings in the same way as we did in Word2Vec. Finally, we trained the above three classifiers on GloVe sentence embeddings.

Universal Sentence Encoder (USE) applies a Deep Averaging Network (DAN) to extract 512-dimensional embeddings for each sentence (Cer et al. 2018). We used these embeddings as features of the LR, RF, and SVM classifiers.

Transformers: We applied modern transformers such as BERT (Devlin et al. 2019), RoBERTa (Liu et al. 2019b), and XLNet (Yang et al. 2019). We fine-tuned these models on our dataset by adding a layer (with softmax activation) in the forward direction, containing four output units (one for each class). These models were trained using a batch size of 32, a maximum sequence length of 256, and five epochs to minimize the cross-entropy loss. Trained each transformer on an online GPU having 16 GB memory took only between seven and ten minutes.

The above approaches were evaluated on INCITE for five-fold cross-validation. INCITE was divided into five folds in a stratified manner, leading to the same class distribution in each fold. For evaluation, prior works on hate and fear speech (Pereira-Kohatsu et al. 2019; Salminen et al. 2020; Saha et al. 2021) adopted the AUC-ROC metric because it measures the goodness of fit, especially appropriate for imbalanced datasets. Since INCITE is also imbalanced, we also leveraged the AUC-ROC score for evaluation.

Table 5.1 shows the AUC-ROC score (obtained by the one-versus-one method (Sklearn 2023)) achieved by the above approaches in five folds. Among embeddings-based methods, TF-IDF with LR achieved 0.818 as the average AUC-ROC score, followed by Word2Vec with SVM (0.815), USE with SVM (0.812), and GloVe with SVM (0.800). However, all transformer-based approaches such as BERT (0.832 average AUC-ROC), RoBERTa (0.851 average AUC-ROC), and XLNet (0.845 average AUC-ROC) outperformed embedding-based approaches. Overall, RoBERTa achieved the highest average AUC-ROC score and was chosen as the best model for identifying inciting speech.

Table 5.1: AUC-ROC score for each of the five folds. Bold indicates the highest average AUC-ROC score.

Approach	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
TF-IDF + LR	0.810	0.825	0.832	0.805	0.820	0.818
TF-IDF + RF	0.790	0.793	0.812	0.786	0.803	0.780
TF-IDF + SVM	0.810	0.818	0.832	0.801	0.812	0.814
Word2Vec + LR	0.810	0.800	0.788	0.792	0.784	0.794
Word2Vec + RF	0.756	0.752	0.747	0.745	0.758	0.751
Word2Vec + SVM	0.823	0.820	0.805	0.816	0.814	0.815
GloVe + LR	0.784	0.785	0.775	0.755	0.766	0.773
GloVe + RF	0.760	0.758	0.752	0.760	0.758	0.757
GloVe + SVM	0.809	0.812	0.793	0.794	0.795	0.800
USE + LR	0.812	0.825	0.801	0.786	0.807	0.806
USE + RF	0.768	0.773	0.760	0.754	0.757	0.762
USE + SVM	0.821	0.829	0.811	0.798	0.803	0.812
BERT	0.838	0.826	0.839	0.819	0.838	0.832
RoBERTa	0.851	0.849	0.854	0.845	0.856	**0.851**
XLNet	0.843	0.850	0.856	0.833	0.847	0.845

5.9.2 Model for the HateXplain Dataset

Besides INCITE, we also applied our model on the HateXplain dataset (Mathew et al. 2021).

To check the prevalence of inciting speech against multiple targets, e.g., gender identity, we tweaked our computational model. To this end, we replaced religion-specific or India-specific words in INCITE with generalized tokens—i.e., *Muslims* by *<target>*, *India* by *<country>*, and *Quran*, *Allah*, and Islam by *<belief>*. Second, we retrained our model on the modified corpus. The intuition is that the retrained model learns how inciting speech is expressed without over-fitting to Islamophobia or the Indian setting.

Since we don't have the ground truth of inciting speech for the HateXplain dataset, we spot-checked of our model's predictions. From 15,905 posts in HateXplain (Mathew et al. 2021), our model predicted 7,102 to be inciting and others non-inciting.

For spot-checking, we first randomly selected and checked 50 posts predicted as inciting and found 37 of them actually inciting (true positive percentage = 74%). Second, to analyze true negatives, we checked 50 randomly selected posts predicted as non-inciting and found 45 of them to be non-inciting, leading to 90% as a percentage of true negatives. Third, we randomly selected 100 posts regardless of their predictions and analyzed them for recall. Our analysis revealed 37 posts as inciting, out of which 27 were identified by our model, leading to a recall of 72.9%.

CHAPTER

6 CONCLUSIONS

This thesis delves into the interesting fact that similar NLP techniques (involving transfer learning) can be leveraged to address diverse types of harassment. By investigating online text where harassment is either described or caused, we gain several useful insights such as how rogue functionalities are described, understanding violations of consent, and identifying multiple strategies of incitement. Our research method includes utilizing large amounts of natural language data, spanning across diverse contexts from multiple platforms: app reviews on the Apple App Store, MeToo stories on Reddit, and inciting posts on WhatsApp. Our work not only understands and tackles specific problems on these platforms but also sheds light on the prevalence of similar issues on other platforms. This paves the way for future researchers (intending to combat harassment) in multiple areas such as computational social science and responsible computing. In this chapter, we discuss the summary of our findings (Section 6.1), possible social impact (Section 6.2), and some future directions (Section 6.3).

6.1 Summary of Findings

In the realm of online harassment, an abuser (app user) exploits a mobile app to violate the privacy expectations of other app users or third parties (called victims). We found that app reviews reveal critical information about an app's rogue functionalities, and hence are crucial in auditing such apps. Relevant reviews were written by all parties (abusers, victims, and third person), each expressing distinct linguistic variations. To our surprise, not only negative but positive reviews (by abusers) also reported rogue behavior. While describing a rogue behavior, we found that reviews vary in their claims (*convincingness*) and in describing the severity of the incident. Using our computational method, we found a total of 239 rogue apps and four prominent types of rogue functionalities.

In the real-world scenario, we consider MeToo stories from Reddit, expressing experiences of sexual harassment and assault. We surveyed 29 helpers recruited from MeToo-related subreddits and found that reading and responding to long MeToo stories is demanding. Hence, we built a computational tool that identifies three important components of a MeToo story. Due to multiple variations in the story structure, identifying and highlighting these components was challenging. Our survey concluded that the tool was successful in reducing the demandingness of reading and responding to such stories.

Using MeToo stories, we also studied how violations stories are linguistically expressed. We found freely given as the most prominent violation and stories describing it frequently mention time-sensitive attributes, as compared to other violations. Some violations (explicit and enthusiastic; informed and specific) have positive correlations and hence co-occur. The LIWC-22 (Boyd et al. 2022) analysis found that stories indicating consent violations illustrate low confidence and leadership but high authenticity.

In the mixed scenario, we study inciting speech used in the religious context. Inciting speech is both structurally and linguistically different from hate speech. Hence, it may be difficult to be identified by traditional hate speech detectors. We found three rhetorical strategies in incitement and identified them using our computational tool. Our analysis showed that these strategies differ in their textual signatures. One post can contain multiple strategies but a sentence typically has only one. We found the prevalence of such inciting speech on other platforms such as Gab and Twitter, targeting groups based on non-religious attributes such as sexual orientation and gender.

6.2 Social Impact

We discuss the possible social impact of our projects below. First, we describe how the problems we solved can potentially prevent abusers from causing harassment. Second, we describe how our computational methods can positively impact the victims.

6.2.1 Preventing Abusers

Abusers can directly or indirectly harm the victims (Figure 1.1). For instance, abusers directly harass the victim in the real-world scenario, however, exploit the online platform to harass in the other two scenarios. We discuss the notion of preventing such harassment in each of the scenarios below.

Abusers exploit rogue apps for spying and stalking purposes. The more the rogue apps, the more options they have for such misuse. Moreover, some legitimate apps can be easily found to have rogue functionalities and hence misused (Chatterjee et al. 2018). In order to prevent such misuse, these apps should be brought into the attention of app distribution platforms and developers. We did so as described in Section 2.7.2.

In the real-world scenario, abusers sexually harass or assault the victims. Since the harm is directly done, we are incapable to prevent it. However, we mitigate its consequences by assisting in providing support, as discussed in Section 6.2.2.

The case of mixed harassment is peculiar. Abusers exploit online platform to cause both online and real-world harassment. As shown by our study, inciting speech is not only prominent against religious communities but also against other targets such as people of an ethnicity, sexual orientation, and so on. We found three types of rhetorical strategies that abusers leverage to express incitement. Delving into such linguistic patterns are useful in identifying such speech and preventing exploitation of online platforms.

6.2.2 Helping Victims

This thesis emphasizes the social good of victims by providing responsible NLP solutions. In the online scenario, awareness of identified rogue apps may prevent future users (potential victims) from installing them, curbing future misuse. Moreover, timely identification of such apps and reporting them will be the first step toward their rectification. Our proposed approach, iROGUE, can do so based on reviews received in real time. This step will make such apps safe and privacy-sensitive from the victims' perspective.

In the real-world scenario, providing support to the victims of sexual harassment and assault is crucial to help them cope with trauma and depression (Andalibi et al. 2016). We built a tool that can highlight important parts of a MeToo story. Through the survey of helpers, we found that the generated highlights are not only important but significantly reduce the effort of reading and responding to MeToo stories. As a result, our tool has proven effective at assisting helpers in providing support.

In the mixed scenario, we built a computational model to identify inciting speech. Such online speech may be disturbing for the target to read. Moreover, it may induce real-world animosity and violence against the target. Identification models can be used for online moderation, potentially preventing the reach of such speech to the instrument and target.

6.3 Future Work

We contribute three datasets: METHREE, MYCONSENT, and INCITE. In conjunction with our findings, these datasets open up interesting directions for future research. We describe them in detail below.

Cross-domain models can be trained to leverage knowledge in these datasets and identify similar problems on other platforms. For example, INCITE can be used to learn inciting patterns from WhatsApp and uncover similar patterns on other social media. Such models will not only uncover interesting findings but also identify antisocial speech on multiple platforms. This research may unveil a new category of antisocial speech which is under-studied and requires attention. These efforts would be beneficial in real-time content moderation addressing new kinds of problematic online text.

A comparative study between multiple platforms will provide an in-depth understanding of the whole landscape of harassment. For example, investigating violation stories on the Safecity (Safecity 2023) website and comparing them with MYCONSENT will provide a broad view of real-world harassment. Stories on such websites may not be as long as Reddit stories, hence language expressing violations may vary. Identifying differences in language from MYCONSENT is an interesting direction. Moreover, many corporate portals allow employees to report their harassment or assault experiences. Understanding incidents, effects, and consent through stories from these various sources can help the authorities in providing effective support to the survivors.

While tackling harassment, there is a scope for building explainable models that provide transparency to users on how the prediction was made. For example, answering why a user's post is flagged as inciting and how they can improve its language. This may include suggesting actions to the user to correct specific parts of their post. In the online harassment scenario, a potential user can be warned about the harms of the app before installing it. This can be done by mining previous reviews of that app and generating alarms based on the reported rogue behavior.

We focus on specific problems in each of the scenarios. However, many targeted problems in such scenarios are still unaddressed. In future, there is a scope of understanding cases such as racial harassment, domestic violence, and so on. Social media studies focusing on such sensitive topics may explore a different spectrum of societal problems.

All these directions accelerate research in multiple areas such as computational social science and responsible computing. By acknowledging diverse experiences and contexts of harassment, effective and sustainable AI solutions should be built to have a high societal impact. At a large scale, such solutions would promote online and real-world interactions that are safe and inclusive.

REFERENCES

Larry Alexander. Incitement and Freedom of Speech. In *Freedom of Speech and Incitement Against Democracy*, pages 101–118. Brill Nijhoff, Online, 2001.

Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. A Deep Dive into Multilingual Hate Speech Classification. In *Proceedings of the Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track: European Conference, ECML PKDD*, pages 423–439, Ghent, Belgium, 2021. Springer International Publishing.

Nazanin Andalibi, Oliver Haimson, Munmun Choudhury, and Andrea Forte. Understanding Social Media Disclosures of Sexual Abuse Through the Lenses of Support Seeking and Anonymity. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 3906–3918, San Jose, California, May 2016. Association for Computing Machinery.

Debasish Basak, Srimanta Pal, and Dipak Patranabis. Support Vector Regression. *Neural Information Processing, Letters and Reviews*, 11, November 2007.

Tobias Bauer, Emre Devrim, Misha Glazunov, William Lopez Jaramillo, Balaganesh Mohan, and Gerasimos Spanakis. #MeTooMaastricht: Building a Chatbot to Assist Survivors of Sexual Harassment. In *Machine Learning and Knowledge Discovery in Databases*, pages 503–521, Cham, September 2020. Springer International Publishing.

Andrew R. Besmer, Jason Watson, and M. Shane Banks. Investigating User Perceptions of Mobile App Privacy: An Analysis of User-Submitted App Reviews. *International Journal of Information Security and Privacy (IJISP)*, 14(4):74–91, October 2020.

Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. A Dataset of Hindi-English Code-Mixed Social Media Text for Hate Speech Detection. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 36–41, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

Douglas G Bonett. Statistical Inference for G-indices of Agreement. *Journal of Educational and Behavioral Statistics*, 47(4):438–458, 2022.

Hanen Borchani, Gherardo Varando, Concha Bielza, and Pedro Larrañaga. A Survey on Multi-Output Regression. *Wiley Interdisciplinary Reviews Data Mining and Knowledge Discovery*, 5 (5):216–233, September 2015.

Sandrine Boudana and Elad Segev. Theorizing Provocation Narratives as Communication Strategies: Provocation Narratives. *Communication Theory*, 27:329–346, 06 2017.

Ryan L. Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W. Pennebaker. The Development and Psychometric Properties of LIWC-22. 2022.

Denis Eka Cahyani and Irene Patasik. Performance Comparison of TF-IDF and Word2Vec Models for Emotion Text Classification. *Bulletin of Electrical Engineering and Informatics*, 10(5):2780–2788, 2021.

Clay Calvert. Hate Speech and Its Harms: A Communication Theory Perspective. *Journal of Communication*, 47(1):4–19, 1997.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Universal Sentence Encoder. *CoRR*, abs/1803.11175:1–7, 2018.

Jair Cervantes, Farid Garcia-Lamont, Lisbeth Rodriguez-Mazahua, and Asdrubal Lopez. A Comprehensive Survey on Support Vector Machine Classification: Applications, Challenges and Trends. *Neurocomputing*, 408:189–215, 2020.

Kelly M. Champion and Daniel L. Clay. Individual Differences in Responses to Provocation and Frequent Victimization by Peers. *Child Psychiatry and Human Development*, 37:205–220, 2007.

Mohit Chandra, Manvith Reddy, Shradha Sehgal, Saurabh Gupta, Arun Balaji Buduru, and Ponnurangam Kumaraguru. "A Virus Has No Religion": Analyzing Islamophobia on Twitter During the COVID-19 Outbreak. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, pages 67–77, New York, NY, USA, 2021. Association for Computing Machinery. doi: 10.1145/3465336.3475111.

Rahul Chatterjee, Periwinkle Doerfler, Hadas Orgad, Sam Havron, Jackeline Palmer, Diana Freed, Karen Levy, Nicola Dell, Damon McCoy, and Thomas Ristenpart. The Spyware Used in Intimate Partner Violence. In *Proceedings of the 39th IEEE Symposium on Security and Privacy (SP)*, pages 441–458, San Francisco, CA, USA, May 2018. IEEE Press.

Ning Chen, Jialiu Lin, Steven C. H. Hoi, Xiaokui Xiao, and Boshen Zhang. AR-Miner: Mining Informative Reviews for Developers from Mobile App Marketplace. In *Proceedings of the 36th International Conference on Software Engineering*, pages 767–778, New York, May 2014. Association for Computing Machinery.

Jianpeng Cheng and Mirella Lapata. Neural Summarization by Extracting Sentences and Words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 484–494, Berlin, Germany, August 2016. Association for Computational Linguistics.

Patricia Chiril, Endang Wahyu Pamungkas, Farah Benamara, Véronique Moriceau, and Viviana Patti. Emotionally Informed Hate Speech Detection: A Multi-target Perspective. *Cognitive Computation*, pages 1–31, 2022.

Jacob Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences.* Lawrence Erlbaum Associates, Hillsdale, New Jersey, 2nd edition, 1988.

Aron Culotta and Andrew McCallum. Reducing Labeling Effort for Structured Prediction Tasks. In *American Association for Artificial Intelligence*, volume 5, pages 746–751, 2005.

Ido Dagan and Sean P. Engelson. *Committee-based Sampling for Training Probabilistic Classifiers*, pages 150–157. Elsevier, 1995.

Mithun Das, Somnath Banerjee, and Animesh Mukherjee. Data Bootstrapping Approaches to Improve Low Resource Abusive Language Detection for Indic Languages. In *Proceedings of the 33rd ACM Conference on Hypertext and Social Media*, pages 32–42, New York, NY, USA, 2022. Association for Computing Machinery.

Bram Orobio de Castro, Esmée E. Verhulp, and Kevin Runions. Rage and Revenge: Highly Aggressive Boys' Explanations for their Responses to Ambiguous Provocation. *European Journal of Developmental Psychology*, 9(3):331–350, 2012.

Bonnie-Elene Deal, Lourdes S. Martinez, Brian H. Spitzberg, and Ming-Hsiang Tsou. "I Definitely Did Not Report It When I Was Raped…# WeBelieveChristine# MeToos": A Content Analysis of Disclosures of Sexual Assault on Twitter. *Social Media+ Society*, 6(4): 2056305120974610, 2020.

Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. Hate Me, Hate Me Not: Hate Speech Detection on Facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, pages 86–95, 2017.

Peter Devine, Yun Sing Koh, and Kelly Blincoe. Evaluating Unsupervised Text Embeddings on Software User Feedback. In *Proceedings of the IEEE 29th International Requirements Engineering Conference Workshops (REW)*, pages 87–95, 2021. doi: 10.1109/REW53955.2021.00020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

Venkatesh T. Dhinakaran, Raseshwari Pulle, Nirav Ajmeri, and Pradeep K. Murukannaiah. App Review Analysis via Active Learning: Reducing Supervision Effort without Compromising Classification Accuracy. In *Proceedings of 26th IEEE International Requirements Engineering Conference (RE)*, pages 170–181, Banff, Canada, 2018. IEEE.

Stephan Dreiseitl and Lucila Ohno-Machado. Logistic Regression and Artificial Neural Network Classification Models: a Methodology Review. *Journal of Biomedical Informatics*, 35(5): 352–359, 2002.

EEOC. Harassment, 2023. URL `https://www.eeoc.gov/sexual-harassment`. United States Equal Employment Opportunity Commission. Accessed 2023-07-19.

Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. Hate Lingo: A Target-based Linguistic Analysis of Hate Speech in Social Media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12, 2018.

Anjalie Field, Gayatri Bhat, and Yulia Tsvetkov. Contextual Affective Analysis: A Case Study of People Portrayals in Online #Metoo Stories. In *Proceedings of the 13th International AAAI Conference on Web and Social Media*, volume 13, pages 158–169, 2019.

Kimberly Field-Springer, Hannah Draut, Fran Babrow, and Maddie Sandman. (Re)claiming Stories in the #MeToo Movement: Righting Epistemic Wrongs of Physical, Mental, and Emotional Harms of Sexual Violence. *Health Communication*, 37(8):982–991, 2022. doi: 10.1080/10410236.2021.1880052.

Tensor Flow. Tensorflow Hub, 2023. URL `https://www.tensorflow.org/hub`. Accessed 2023-08-05.

Michelle A. Fortier, David Dilillo, Terri L. Messman-Moore, James Peugh, Kathleen A. DeNardi, and Kathryn J. Gaffey. Severity of Child Sexual Abuse and Revictimization: The Mediating Role of Coping and Trauma Symptoms. *Psychology of Women Quarterly*, 33(3):308–320, 2009.

Diana Freed, Sam Havron, Emily Tseng, Andrea Gallardo, Rahul Chatterjee, Thomas Ristenpart, and Nicola Dell. Is My Phone Hacked? Analyzing Clinical Computer Security Interventions with Survivors of Intimate Partner Violence. *Proceedings of the 17th ACM Conference on Human-Computer Interaction*, 3:1–24, 2019.

Björn Gambäck and Utpal Kumar Sikdar. Using Convolutional Neural Networks to Classify Hate-Speech. In *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90, Vancouver, BC, Canada, August 2017. Association for Computational Linguistics.

César García, Alejandro Guerrero, Joshua Zeitsoff, Srujay Korlakunta, Pablo Fernandez, Armando Fox, and Antonio Ruiz-Cortés. Bluejay: a Cross-Tooling Audit Framework for Agile Software Teams. In *Proceedings of 43rd IEEE International Conference on Software Engineering: Software Engineering Education and Training (ICSE-SEET)*, pages 283–288. IEEE, 2021.

Arijit Ghosh Chowdhury, Ramit Sawhney, Puneet Mathur, Debanjan Mahata, and Rajiv Ratn Shah. Speak up, Fight Back! Detection of Social Media Disclosures of Sexual Harassment. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 136–146, Minneapolis, Minnesota, June 2019a. Association for Computational Linguistics.

Arijit Ghosh Chowdhury, Ramit Sawhney, Rajiv Ratn Shah, and Debanjan Mahata. #YouToo? Detection of Personal Recollections of Sexual Harassment on Social Media. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2527–2537, Florence, Italy, July 2019b. Association for Computational Linguistics.

Aristides Gionis, Piotr Indyk, and Rajeev Motwani. Similarity Search in High Dimensions via Hashing. In *Proceedings of the Very Large Data Bases*, volume 99, pages 518–529, 1999.

Steven N. Gold, Stacey A. Ketchman, Irene Zucker, Melissa A. Cott, and Alfred H. Sellers. Relationship Between Dissociative and Medically Unexplained Symptoms in Men and Women Reporting Childhood Sexual Abuse. *Journal of Family Violence*, 23:569–575, 2008.

Hui Guo and Munindar P. Singh. Caspar: Extracting and Synthesizing User Stories of Problems from App Reviews. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*, pages 628–640, New York, NY, USA, July 2020. Association for Computing Machinery. ISBN 9781450371216. doi: 10.1145/3377811.3380924.

HAF. Deadly Attacks on Hindu Americans, 2023. URL `https://www.hinduamerican.org/press/anti-hindu-rate-crimes-united-states-rise-fbi-report`. Hindu American Foundation on Religious Harassment. Accessed 2023-11-25.

Kevin A. Hallgren. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in Quantitative Methods for Psychology*, 8(1):23–34, 2012.

Amanul Haque, Vaibhav Garg, Hui Guo, and Munindar P Singh. Pixie: Preference in Implicit and Explicit Comparisons. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 106–112, Dublin, Ireland, 2022. Association for Computational Linguistics.

Naeemul Hassan, Amrit Poudel, Jason Hale, Claire Hubacek, Khandaker Tasnim Huq, Shubhra Kanti Karmaker Santu, and Syed Ishtiaque Ahmed. Towards Automated Sexual Violence Report Tracking. In *Proceedings of the 14th International AAAI Conference on Web and Social Media*, pages 250–259, Georgia, USA, June 2020.

Sam Havron, Diana Freed, Rahul Chatterjee, Damon McCoy, Nicola Dell, and Thomas Ristenpart. Clinical Computer Security for Victims of Intimate Partner Violence. In *Proceedings of the 28th USENIX Security Symposium*, pages 105–122, Santa Clara, January 2019. USENIX Association.

Wesley Newcomb Hohfeld. *Fundamental Legal Conceptions as Applied in Judicial Reasoning and other Legal Essays*. Yale University Press, New Haven, Connecticut, 1919. A 1919 printing of articles from 1913.

J.W. Holley and G.A. Lienert. The G Index of Agreement in Multiple Ratings. *Educational and Psychological Measurement*, 34(4):817–822, 1974.

Qianjia Huang, Vivek K. Singh, and Pradeep K. Atrey. On Cyberbullying Incidents and Underlying Online Social Relationships. *Journal of Computational Social Science*, 1:241–260, 2018.

ICCPR. Incitement, 2023. URL `https://www.ohchr.org/sites/default/files/Documents/Issues/Opinion/SeminarRabat/Rabat_threshold_test.pdf`. International Covenant on Civil and Political Rights on Incitement. Accessed 2023-02-09.

Jane Im, Jill Dimond, Melody Berton, Una Lee, Katherine Mustelier, Mark S. Ackerman, and Eric Gilbert. Yes: Affirmative Consent as a Theoretical Framework for Understanding and

Imagining Social Platforms. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–18. Association for Computing Machinery, 2021.

Aishwarya Jadhav and Vaibhav Rajan. Extractive Summarization with SWAP-NET: Sentences and Words from Alternating Pointer Networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 142–151, Melbourne, Australia, July 2018. Association for Computational Linguistics.

Hannah Javidi, Laura Widman, Anne J. Maheux, Jordyn McCrimmon, Reina Evans-Paulson, and Whitney Becker. Pact: Developing and Evaluating a Digital Sexual Consent Program for Youth. *The Journal of Sex Research*, pages 1–15, 2023. doi: 10.1080/00224499.2023.2208560.

Kristen N. Jozkowski. "Yes Means Yes"? Sexual Consent Policy and College Students. *Change: The Magazine of Higher Learning*, 47(2):16–23, 2015.

Sweta Karlekar and Mohit Bansal. SafeCity: Understanding Diverse Forms of Sexual Harassment Personal Stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2805–2811, Brussels, Belgium, October 2018. Association for Computational Linguistics.

Aparup Khatua, Erik Cambria, and Apalak Khatua. Sounds of Silence Breakers: Exploring Sexual Violence on Twitter. In *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 397–400, Barcelona, 2018. IEEE Press.

Felix Koch. Consent as a Normative Power. In *The Routledge Handbook of the Ethics of Consent*, pages 32–43. Routledge, 2018.

Zijad Kurtanović and Walid Maalej. Mining User Rationale from Software Reviews. In *Proceedings of the 25th IEEE International Requirements Engineering Conference (RE)*, pages 61–70, Lisbon, Portugal, September 2017. IEEE Press.

Peng Li, Yinglin Wang, Wei Gao, and Jing Jiang. Generating Aspect-Oriented Multi-Document Summarization with Event-Aspect Model. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1137–1146, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics.

Yingchi Liu, Quanzhi Li, Xiaozhong Liu, Qiong Zhang, and Luo Si. Sexual Harassment Story Classification and Key Information Identification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2385–2388, New York, NY, USA, November 2019a. Association for Computing Machinery.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized Bert Pretraining Approach. *arXiv preprint arXiv:1907.11692*, 2019b.

Piotr Łubinski. Social Media Incitement to Genocide: ECHR Countries Perspective. In *The Concept of Genocide in International Criminal Law*. Taylor & Francis, 2017.

Lydia Manikonda, Ghazaleh Beigi, Subbarao Kambhampati, and Huan Liu. #Metoo through the Lens of Social Media. In *Proceedings of the 11th International Conference on Social, Cultural, and Behavioral Modeling*, pages 104–110, Washington, USA, November 2018. Springer Verlag.

Christopher D. Manning. Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics? In *Proceedings of the Computational Linguistics and Intelligent Text Processing*, pages 171–189, Berlin, Heidelberg, February 2011. Springer.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875, 2021.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS, pages 3111–3119, Lake Tahoe, Nevada, December 2013. Neural Information Processing Systems Foundation.

George A. Miller. WordNet: A Lexical Database for English. *Communications of the ACM*, 38 (11):39–41, November 1995. doi: 10.1145/219717.219748.

Saif Mohammad and Peter Turney. Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34, Los Angeles, California, June 2010. Association for Computational Linguistics.

Saif M. Mohammad. Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words. In *Proceedings of The Annual Conference of the Association for Computational Linguistics (ACL)*, Melbourne, Australia, 2018.

Saif M. Mohammad and Peter D. Turney. Crowdsourcing a Word-Emotion Association Lexicon. *Computational Intelligence*, 29(3):436–465, August 2013.

Duc Cuong Nguyen, Erik Derr, Michael Backes, and Sven Bugiel. Short Text, Large Effect: Measuring the Impact of User Reviews on Android App Security & Privacy. In *Proceedings of the 40th IEEE Symposium on Security and Privacy (SP)*, pages 555–569, San Francisco, May 2019. IEEE Computer Society.

NLTK. Natural language toolkit, 2023. URL `https://www.nltk.org/api/nltk.tokenize.html`. NLTK Library. Accessed 2023-07-14.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive Language Detection in Online User Content. In *Proceedings of the 25th International Conference on World Wide Web*, pages 145–153, Republic and Canton of Geneva, CHE, 2016. International World Wide Web Conferences Steering Committee.

Lorelli S. Nowell, Jill M. Norris, Deborah E. White, and Nancy J. Moules. Thematic Analysis: Striving to Meet the Trustworthiness Criteria. *International Journal of Qualitative Methods*, 16(1):1609406917733847, 2017.

NSVR. Harassment statistics, 2023. National Sexual Violence Resource Centre. Accessed 2023-07-19.

Tully O'Neill. 'Today I Speak': Exploring How Victim-Survivors Use Reddit. *International Journal for Crime, Justice and Social Democracy*, 7(1):44–59, 2018.

Planned Parenthood. Sexual Consent, 2021. URL `https://www.plannedparenthood.org/learn/relationships/sexual-consent`. Sexual Consent and Its Attributes. Accessed 2024-02-03.

Ji Ho Park and Pascale Fung. One-step and Two-step Classification for Abusive Language Detection on Twitter. In *Proceedings of the First Workshop on Abusive Language Online*, pages 41–45, Vancouver, BC, Canada, August 2017. Association for Computational Linguistics.

Ivan Pashchenko, Henrik Plate, Serena Elisa Ponta, Antonino Sabetta, and Fabio Massacci. Vuln4real: A Methodology for Counting Actually Vulnerable Dependencies. *IEEE Transactions on Software Engineering*, 48(5):1592–1609, 2020.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global Vectors for Word Representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

Juan Carlos Pereira-Kohatsu, Lara Quijano-Sánchez, Federico Liberatore, and Miguel Camacho-Collados. Detecting and Monitoring Hate Speech in Twitter. *Sensors*, 19(21):4654, 2019.

Henning Perl, Sergej Dechand, Matthew Smith, Daniel Arp, Fabian Yamaguchi, Konrad Rieck, Sascha Fahl, and Yasemin Acar. Vccfinder: Finding Potential Vulnerabilities in Open-Source Projects to Assist Code Audits. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 426–437, 2015.

Pew. Statistics Indicating Harassment of Religious Groups, 2023. URL `https://www.pewresearch.org/religion/2021/09/30/harassment-of-religious-groups-reaches-new-peak-in-2019/`. Pew Research Center on Religious Harassment. Accessed 2023-11-25.

Pushshift. Pushshift API, 2023. URL `https://psaw.readthedocs.io/en/latest/`. Reddit's Pushshift API. Accessed 2023-02-14.

PyDictionary. Dictionary in Python, 2023. URL https://pypi.org/project/PyDictionary/. Accessed 2023-08-05.

Pythesaurus. Thesaurus in Python, 2023. URL https://pypi.org/project/py-thesaurus/. Accessed 2023-08-05.

RAINN. Assault, 2023. URL https://www.rainn.org/articles/sexual-assault. Rape, Abuse & Incest National Network. Accessed 2023-07-19.

Reddit. Subreddit MeToo, 2023a. URL https://www.reddit.com/r/meToo/. MeToo Subreddit. Accessed 2023-04-20.

Reddit. Subreddit Sexual Assault, 2023b. URL https://www.reddit.com/r/sexualassault/. Sexual Assault Subreddit. Accessed 2023-04-20.

Reddit. Subreddit Sexual Harassment, 2023c. URL https://www.reddit.com/r/SexualHarassment/. Sexual Harassment Subreddit. Accessed 2023-04-20.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.

Ana Reyes-Menendez, Jose Ramon Saura, and Stephen B. Thomas. Exploring Key Indicators of Social Identity in the #MeToo Era: Using Discourse Analysis in UGC. *International Journal of Information Management*, 54:102129, 2020.

Debra Rickwood, Frank P. Deane, Coralie J. Wilson, and Joseph Ciarrochi. Young People's Help-seeking for Mental Health Problems. *Australian e-Journal for the Advancement of Mental Health*, 4(3):218–251, 2005. doi: 10.5172/jamh.4.3.218.

Privacy Rights. Online harassment, 2023. URL https://privacyrights.org/consumer-guides/online-harassment-cyberstalking. Online Harassment Definition. Accessed 2023-02-09.

Kevin A. Roundy, Paula Barmaimon Mendelberg, Nicola Dell, Damon McCoy, Daniel Nissani, Thomas Ristenpart, and Acar Tamersoy. The Many Kinds of Creepware Used for Interpersonal Attacks. In *Proceedings of the 41th IEEE Symposium on Security and Privacy (SP)*, pages 753–770, Los Alamitos, May 2020. IEEE Computer Society.

Safecity. Safecity Website, 2023. URL https://www.safecity.in/. Safecity Website. Accessed 2023-02-14.

Punyajoy Saha, Binny Mathew, Kiran Garimella, and Animesh Mukherjee. "Short is the Road That Leads from Fear to Hate": Fear Speech in Indian Whatsapp Groups. In *Proceedings of the Web Conference*, pages 1110–1121, New York, 2021. Association for Computing Machinery.

Punyajoy Saha, Kiran Garimella, Narla Komal Kalyan, Saurabh Kumar Pandey, Pauras Mangesh Meher, Binny Mathew, and Animesh Mukherjee. On the Rise of Fear Speech in Online Social Media. *Proceedings of the National Academy of Sciences*, 120(11):e2212270120, 2023.

Joni Salminen, Maximilian Hopf, Shammur A. Chowdhury, Soongyo Jung, Hind Almerekhi, and Bernard J. Jansen. Developing an Online Hate Classifier for Multiple Social Media Platforms. *Human-centric Computing and Information Sciences*, 10:1–34, 2020.

Hubert Schnüriger. What is Consent? In Peter Schaber and Andreas Müller, editors, *The Routledge Handbook of the Ethics of Consent*, chapter 2, pages 21–31. Routledge, London, 2018. doi: 10.4324/9781351028264-3.

Abigail See, Peter J. Liu, and Christopher D. Manning. Get To The Point: Summarization with Pointer-Generator Networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada, July 2017. Association for Computational Linguistics.

Andrew Sellars. Defining Hate Speech. *Berkman Klein Center Research Publication*, pages 16–48, 2016.

Burr Settles. *Active Learning*. Number 6 in Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool, Cham, Switzerland, 2012. doi: 10.1007/978-3-031-01560-1.

Maria Grazia Sindoni. Direct Hate Speech Vs. Indirect Fear Speech. A Multimodal Critical Discourse Analysis of the Sun's Editorial "1 in 5 Brit Muslims' Sympathy for Jihadis". *Lingue e Linguaggi*, 28:267–292, 2018.

Munindar P Singh. Consent as a Foundation for Responsible Autonomy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12301–12306, 2022.

Vivek K. Singh and Connor Hofenbitzer. Fairness Across Network Positions in Cyberbullying Detection Algorithms. In Francesca Spezzano, Wei Chen, and Xiaokui Xiao, editors, *ASONAM: International Conference on Advances in Social Networks Analysis and Mining*, pages 557–559, British Columbia, Canada, aug 2019. Association for Computing Machinery.

Sklearn. One versus One Approach, 2023. URL `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html`. Sklearn Metrics for AUC. Accessed 2023-07-14.

Paul F. Smith, Siva Ganesh, and Ping Liu. A Comparison of Random Forest Regression and Multiple Linear Regression for Prediction in Neuroscience. *Journal of Neuroscience Methods*, 220:85–91, October 2013.

Dangerous Speech. Definition of Dangerous Speech, 2023. URL `https://dangerousspeech.org/faq/?faq=200`. Resources Related to Dangerous Speech. Accessed 2023-07-09.

Apple App Store. AirBeam Video Surveillance App, 2023a. URL `https://apps.apple.com/us/app/airbeam-video-surveillance/id428767956`. Accessed 2023-10-08.

Apple App Store. Apple App Store, 2023b. URL https://www.apple.com/app-store/. Accessed 2023-11-11.

Apple App Store. Find My Kids: Parental Control, 2023c. URL https://apps.apple.com/us/app/id994098803. Accessed 2023-10-08.

Apple App Store. Find My iPhone App, 2023d. URL https://apps.apple.com/us/app/find-my-iphone/id376101648. Accessed 2023-10-08.

Apple App Store. Kik App, 2023e. URL https://apps.apple.com/us/app/kik/id357218860. Accessed 2023-10-08.

Apple App Store. Life360 App, 2023f. URL https://apps.apple.com/us/app/life360-find-family-friends/id384830320. Accessed 2023-10-08.

Apple App Store. Nortan Mobile Security App, 2023g. URL https://buy-static.norton.com/norton/ps/bb/ushard/4up_mnav05w_us_en_fl_tw_branded_mix-n360.html. Accessed 2023-10-08.

Apple App Store. Ourpact Jr. Child App, 2023h. URL https://apps.apple.com/us/app/id1127917970. Accessed 2023-10-08.

Apple App Store. Saferkid Text Monitoring App, 2023i. URL https://apps.apple.com/us/app/saferkid-text-monitoring-app/id1143802529. Accessed 2023-10-08.

Apple App Store. Smart Family Companion App, 2023j. URL https://apps.apple.com/us/app/smart-family-companion/id1352914754. Accessed 2023-10-08.

Apple App Store. 3Fun: Threesome & Swingers App, 2023k. URL https://apps.apple.com/app/id1164067996. Accessed 2023-10-08.

Apple App Store. Popular Utilities Apps, 2023l. URL https://apps.apple.com/us/genre/ios-utilities/id6002. Accessed 2023-11-11.

Google Play Store. Google Play Store, 2023m. URL https://play.google.com/store/apps. Accessed 2023-02-15.

Susan Benesch. Defining and Diminishing Hate Speech. *State of the World's Minorities and Indigenous Peoples*, 2014:18–25, 2014.

Emily Tseng, Rosanna Bellini, Nora McDonald, Matan Danos, Rachel Greenstadt, Damon McCoy, Nicola Dell, and Thomas Ristenpart. The Tools and Tactics Used in Intimate Partner Surveillance: An Analysis of Online Infidelity Forums. In *Proceedings of the 29th USENIX Security Symposium*, pages 1893–1909, Virtual, August 2020. USENIX Association.

Emily Tseng, Diana Freed, Kristen Engel, Thomas Ristenpart, and Nicola Dell. A Digital Safety Dilemma: Analysis of Computer-Mediated Computer Security Interventions for Intimate Partner Violence During COVID-19. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–17, Virtual, 2021. Association for Computing Machinery.

UN. Hate speech, 2023. URL `https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech`. United Nations on Hate Speech. Accessed 2023-07-09.

Alper Kursat Uysal and Serkan Gunal. The Impact of Preprocessing on Text Classification. *Information Processing & Management*, 50(1):104–112, 2014.

Bertie Vidgen and Taha Yasseri. Detecting Weak and Strong Islamophobic Hate Speech on Social Media. *Journal of Information Technology & Politics*, 17(1):66–78, 2020.

Fedor Vitiugin, Yasas Senarath, and Hemant Purohit. Efficient Detection of Multilingual Hate Speech by Using Interactive Attention Network with Minimal Human Feedback. In *Proceedings of the 13th ACM Web Science Conference*, pages 130–138, 2021.

Ting-Yu Wang, Haiming Jin, and Klara Nahrstedt. mAuditor: Mobile Auditing Framework for mHealth Applications. In *Proceedings of the 2015 Workshop on Pervasive Wireless Healthcare*, pages 7–12, 2015.

Merriam Webster. Islamophobia Meaning, 2023. URL `https://www.merriam-webster.com/dictionary/Islamophobia`. Merriam Webster on Islamophobia. Accessed 2023-07-09.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, MercÃ¨ Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific data*, 3 (1):1–9, 2016.

Mingyuan Xia, Lu Gong, Yuanhao Lyu, Zhengwei Qi, and Xue Liu. Effective Real-Time Android Application Auditing. In *Proceedings of the 36th IEEE Symposium on Security and Privacy*, pages 899–914, San Jose, May 2015. IEEE.

Min Xu, Pakorn Watanachaturaporn, Pramod K. Varshney, and Manoj K. Arora. Decision Tree Regression for Soft Classification of Remote Sensing Data. *Remote Sensing of Environment*, 97(3):322–336, 2005. doi: https://doi.org/10.1016/j.rse.2005.05.008.

Peng Yan, Linjing Li, Weiyun Chen, and Daniel Zeng. Quantum-inspired Density Matrix Encoder for Sexual Harassment Personal Stories Classification. In *2019 IEEE International Conference on Intelligence and Security Informatics, ISI*, pages 218–220, Shenzhen, China, July 2019. Institute of Electrical and Electronics Engineers Inc.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R. Salakhutdinov, and Quoc V. Le. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *Advances in Neural Information Processing Systems*, 32, 2019.

Wenjie Yin, Vibhor Agarwal, Aiqi Jiang, Arkaitz Zubiaga, and Nishanth Sastry. AnnoBERT: Effectively Representing Multiple Annotators' Label Choices to Improve Hate Speech Detection. In *Proceedings of the 17th International AAAI Conference on Web and Social Media*, pages 902–913, Limassol, Cyprus, June 2023.

William J. Youden. Index for Rating Diagnostic Tests. *Cancer*, 3(1):32–35, 1950.

Mohammed Zakariah. Classification of Large Datasets Using Random Forest Algorithm in Various Applications: Survey. *International Journal of Engineering and Innovative Technology (IJEIT)*, 4:189–198, September 2014.

Renxian Zhang, Wenjie Li, and Dehong Gao. Generating Coherent Summaries with Textual Aspects. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, pages 1727–1733, Toronto, Ontario, Canada, July 2012. AAAI Press.

Ziqi Zhang, David Robinson, and Jonathan Tepper. Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network. In *European Semantic Web Conference*, pages 745–760. Springer, 2018.

Yixin Zou, Allison McDonald, Julia Narakornpichit, Nicola Dell, Thomas Ristenpart, Kevin Roundy, Florian Schaub, and Acar Tamersoy. The Role of Computer Security Customer Support in Helping Survivors of Intimate Partner Violence. In *Proceedings of the 30th USENIX Security Symposium*, pages 429–446, Virtual, 2021. USENIX Association.

APPENDIX

APPENDIX

A

LABELING CODEBOOK

A.1 Labeling MeThree Dataset

We follow the following labeling instructions to annotate sentences in MeThree:

1. Incident: Any unwelcome sexual advances, sexual behavior, requests for sexual favors, verbal or physical acts of a sexual nature, offensive jokes, or remarks that were either sexual or based on someone's gender were labeled as incidents (EEOC 2023; RAINN 2023).

2. Effects on the survivor: We considered survivors' feelings and emotions that arose during or after the incident. Examples range from feeling uncomfortable (due to the abuser's actions) to being afraid (emotion: fear) of reporting the incident.

3. Advice requested: We considered sentences in which survivors asked for suggestions on topics related to the incident, e.g., whether to report the incident, where to get therapy from, and how to face the abuser again.

Table A.1 shows examples of each category. For confidentiality, we have masked details that might be used to identify survivors or abusers. The first example describes inappropriate physical behavior and is considered an incident. The second example describes that the survivor is sexually exploited (incident) and suffers from depression and anxiety (effects). In the third

Table A.1: Relevant examples according to the labeling instructions. Columns I, E, and A reflect the incident, effects, and advice requested, respectively.

Sentence	I	E	A	Rationale
... he slid his hand up my leg and into my shorts.	✓			Inappropriate physical behavior
... I was sexually used by <abuser> on many occasions ... I am in a constant battle with major depression, crippling real event OCD (I ruminate for 16 hours/day) & debilitating anxiety.	✓	✓		Sexual abuse leading to depression and anxiety
... I'm freaking out and have no one to talk to because no one knows about him or what happened ... What do I do?		✓	✓	Freaking out and asking about future action
Does anyone know how a legal advocate works and what you experienced with them?			✓	Asking for legal advice

example, the survivor expresses fear (by mentioning "freak out") and seeks advice about dealing with it. In the last example, the survivor seeks advice relating to the legal process. Table A.2 reports the Cohen's kappa score among annotators.

Table A.2: Cohen's kappa scores while labeling METHREE.

Annotators	Incident	Effects	Advice requested
A_1, A_2	0.798	0.793	0.891
A_2, A_3	0.720	0.725	0.843
A_3, A_1	0.795	0.801	0.861
Total	0.772	0.774	0.865

A.2 Labeling INCITE Dataset

Two researchers in this project were the annotators. They were given a sentence and asked to label one of the following categories: (i) Identity, (ii) Imputed misdeeds, (iii) Exhortation, and (iv) None. Along with the sentence to label, they read preceding and succeeding sentences to get enough context while labeling.

The labeling was conducted in three phases. In the first phase, the annotators were provided with initial labeling instructions, which they used to label a total of 400 sentences, in four rounds of 100 sentences each. In this phase, Cohen's kappa score was 0.48 (moderate agreement) (Cohen 1960). After each round, the annotators discussed their disagreements, which helped in finalizing the labeling instructions. In the second phase, the two annotators labeled 600 sentences using the final instructions, leading to Cohen's kappa score of 0.73 (substantial agreement). In the third phase, the remaining sampled sentences were split among the two annotators, such that only one annotator labeled a sentence.

The final labeling instructions (codebook), including definitions and examples of each rhetorical strategy, are shown below:

- **Identity:** Sentences that can stereotype or criticize the target, including its beliefs such as its ideologies, sacred books (in a religious context), and leaders, such as

  "In Kashmir, every person who speaks 'Murdabad' is a Muslim" (targeting Muslims)

  "…Muslim children are well taught Jihadi Quran in madrasa" (targeting Quran, the main Islamic scripture)

  The first example claims that Muslims say 'Murdabad' (the Hindi word for India's dismissal). Moreover, the second example ('madrasa' means Islamic school) mentions the Quran to be Jihadi, meaning it spreads terrorism. Indirectly, the latter example typecasts all Muslim children to become terrorists. Such sentences invoke hostility or anger against Muslims and hence are considered inciting.

- **Imputed misdeeds:** Sentences that express ongoing or past misdeeds (real or fake) of the target. These include oppressive incidents (such as violence) allegedly perpetrated by the target. A relevant sentence structure is "X did violence to Y," where X is the target and Y is an individual or some other group of people. Such an example is presented below.

  "<A person> - set fire - heavy damage at Sartala; 3 out of four rooms were destroyed"

- **Exhortation:** Sentences that urge readers to act against the target. For example, the following sentence asks readers to support Hindus in the fight against Muslims. This support-seeking is open to interpretation and can mean anything from boycotts to violence against Muslims.

  "Be violent for religion"

> *"Always support Hindus in the fight of Hindu Muslim, right wrong no matter, all of them later!"*

> We also considered subtle cases that indicate some action. For example, an indirect call to shut down a madrasa (Islamic school) is depicted in the following:

> *"If the madrasa is not closed, after 15 years, more than half of the Muslims of the country will be supportive of the ISIS ideology"*

The sentences not indicating any of the above types were labeled None. Combining the sentences annotated in the three labeling phases, we curated a dataset of 7,000 sentences.

A.3 Labeling MyConsent Dataset

The three researchers in this project were annotators. We defined our labeling instructions following the definitions given in the FRIES framework. The instructions contained language examples (shown below) that helped annotators to identify each consent violation. We divided the labeling process into two phases. First, 400 sentences were distributed among all annotators such that two of them labeled each sentence. For interrater agreement, we computed the G-index of agreement (Holley and Lienert 1974), which is preferred over Cohen's kappa (Cohen 1960) for datasets with multiple categories labeled "no" (Bonett 2022). We report G-index in Table A.3. According to Bonett (2022), the agreement is "excellent" (0.75–1.00 range) for reversible and specific categories, and "good" (0.51–0.75 range) for other categories. Second, after achieving reliable agreement, we equally distributed the remaining 600 sentences such that each is labeled by one annotator. After removing duplicates from a total of 1,000 stories, we were left with 989 labeled stories.

Violation of Freely given: *"I was sleeping", "My gynecologist/doctor/teacher/relative/massage therapist did something . . . and I felt pressured to be part of it" (uneven power dynamics), "I was drunk", "I was minor/underage", "I was tricked into it", "I was high on drugs and don't remember it"*

Violation of Reversible: *"I said no but he/she/they kept going", "I wanted to stop but he/she/they didn't let me"*

Violation of Informed: *"I went in for a hug and then he/she/they grabbed my br**sts/b*tt/XYZ", "I thought we were using a condom but then I found out the other person never put one on", "I thought we had both been tested but I found out he/she/they lied and had an STD", "I just wanted to kiss for a second but then he/she/they made out with me for a whole minute"*

Violation of Enthusiastic: *"I didn't really want to but I wanted to make him/her/them happy", "I felt pressured so I gave in", "I let it happen", "I wasn't that excited"*

Violation of Explicit: *"I didn't say anything that would make him/her/them think I wanted to", "He assumed", "I said maybe", "I said I'd be okay with it later but he/she/they said no we need to do it now"*

Violation of Specific: *"We were making out and I wanted to stop there but we did more", "I told him/her/them I wanted to just do XYZ but then he/she/they did something else..."*

Table A.3: G-index of agreement while labeling MyConsent.

Violation type	G-index of agreement (Holley and Lienert 1974)
Freely given	0.64
Reversible	0.76
Informed	0.81
Enthusiastic	0.58
Explicit	0.56
Specific	0.76
Unknown	0.67