

RSD framework for the detection of covert online harms

Henna Paakki

University of Helsinki

`henna.paakki@helsinki.fi`

Abstract

In this paper, I propose the RSD framework for computational detection of covert harmful online behaviors. The framework’s three dimensions of Repetition, Severity and Disregard are used together to assess the harmfulness of norm-violating behaviors. These are observable and operationalizable dimensions that aim to support computational detection, enabling evidence-based identification of harmful behavior in conjunction with community norms. While the computational detection of overt harmful behaviors like aggression and hate speech has received a lot of attention, covert harmful behaviors like trolling have been much more challenging to identify. I argue that covert behaviors can be traced by models that examine norm violations crossing the thresholds of the dimensions of RSD.

1 Introduction

Computational detection of online harms has been an important growing area of research over the past decade. However, it has become clear that online harms often assume covert, hidden forms of abuse that elude detection (Herring et al., 2002; Hardaker, 2010). Actors intending to implicate harm also constantly come up with novel ways to evade existing detection methods, and new strategies to sow discord online (Gröndahl et al., 2018). Although there has been steady advancement in methods over the years, elusive and covert strategies for implicating harm pose a significant challenge within the realm of computational detection.

By covert harms, I refer to behaviors that exploit indirect or hidden strategies to disrupt conversations or spaces, to sow discord, harass, exploit or deceive, harm or manipulate an individual or a group. These may encompass – but are not limited to – trolling, disinformation dissemination, covert harassment, manipulation, gas-lighting and online deception. By harm, more specifically, here

I refer to emotional, psychological or relational harms (Scheuerman et al., 2021): violations that threaten the autonomy or consent, privacy or well-being of an individual or group (Banko et al., 2020).

The disinformation campaigns and trolling that have flooded online spaces after 2014 have leveraged covert strategies to manipulate and disrupt online conversations and public opinion (Bennett and Livingston, 2018; Francois and Lin, 2021; Badawy et al., 2018). Since then, online spaces have increasingly been increasingly targeted with attempts to manipulate public opinion with regard to political and societal issues. Current disinformation research highlights the ‘liquid’ character of constantly evolving disinformation tactics which is why unmasking them is extremely difficult (Ferreira, 2022). This complexity is exacerbated by the deceptive use of machine learning and generative AI for the creation and amplification of dis- and misinformation (Zagni and Canetta, 2023). While these phenomena have drawn attention to how crucial it is to develop both regulation and detection methods to mitigate such coordinated manipulation, they also often exploit covert strategies that require a novel set of methods to detect.

While coordinated online mass manipulation is a relatively new phenomenon, the strategies of influence are not. The campaigns often exploit similar strategies of communication and interaction as other fore-mentioned covert harmful behaviors, for example strategic digression of conversation (Paakki et al., 2021, 2024; Hardaker, 2013). I highlight that for the detection of covert online harms, there is a need to understand and explicate the ways in which covert harmful behaviors violate manners of constructive communication or interaction, i.e. how they violate social norms. In essence, this requires context-driven definition of the social norms central to a community or space to break down the ways in which these norms might be violated, and thus what constitutes harmful behavior in said

context.

Traditionally the focus of computational detection approaches has been more on the overt end of the spectrum of online harms, such as verbal aggression and hatespeech (Albladi et al., 2025). The challenge we face with especially the more covert forms of online harms is that they employ a constantly changing set of strategies, which elude computational models and definitions (Ferreira, 2022).

I argue that the field could benefit from frameworks that address this dynamic nature of online harms. I discuss key issues related to 1.) definition of covert harmful online behaviors, 2.) norm violation based identification of their features, and 3.) contextual definition and detection of ‘harm’.

The contribution of this paper is to outline the RSD framework for computational detection of covert online harms, consisting of three dimensions: Repetition, Severity and Disregard. I theorize that a common feature of covert harmful behaviors is that they violate social norms in covert ways, utilizing deception and manipulation. I argue that examining social norm violations in relation to the three dimensions of RSD can help improve their computational detection. Assessing norm violations from the perspective of Repetition, Severity and Disregard can help define thresholds for what is considered harmful behavior. The dimensions can also help develop more refined models that can discern between non-harmful and harmful covert norm violations.

2 Literature review

In this section, I will first describe earlier approaches to the computational detection of harmful online behaviors, and then I will provide a summary of some of the key theoretical approaches to the related behaviors.

2.1 Computational detection of online harms

Online harms may encompass a range of different phenomena. Here, I focus on harmful behaviors in particular: these may encompass, but are not limited to, aggression, cyberbullying, hate speech, trolling, disinformation dissemination and deception.

Computational detection of harmful online behaviors has traditionally concentrated on more overt behaviors like (verbal) aggression, which is characterized by aggressive or offensive, vulgar or rude language (Hamilton, 2012). Online aggres-

sion, in fact, has been widely studied, and effectively detected using computational models (see e.g., Chatzakou et al., 2017; Nobata et al., 2016; Wang and Cardie, 2014; Razavi et al., 2010). Arguably, when compared to covert behaviors, this is an easier task to automate as models can be based on the identification of explicit word use or other linguistic features that are overtly and readily identifiable in communication. Online aggression has been identified in earlier work by investigating sentiment or offensive or taboo words present in messages (Razavi et al., 2010; Sood et al., 2012), other linguistic features (Sahana et al., 2020), impoliteness (Zhang et al., 2018), or ad hominem attacks (Zhang et al., 2018). Others have used deep learning, neural networks and LMs/LLMs to detect aggression based on its textual features (Aroyehun and Gelbukh, 2018; Kumar et al., 2018; Risch and Krestel, 2020; Zampieri et al., 2019).

Hate speech detection is similarly a widely studied topic: early studies approached it by using (linguistic) feature-based classifiers and classic machine learning (ML), like SVM (Albladi et al., 2025; Malik et al., 2022), later attempts using deep learning and transformers (Ramos et al., 2024; MacAvaney et al., 2019), and hybrid methods (Zhou et al., 2020). Hate speech detection, like aggression detection, is usually framed as a text classification problem (Fortuna and Nunes, 2018; Mullah and Zainon, 2021; Rawat et al., 2024). Rule and keyword-based methods have been widely used, but they tend to miss subtle, indirect, or context-dependent hate speech (Albladi et al., 2025).

Early cyberbullying detection has also traditionally relied on rule-based systems, linguistic text features and classical machine learning (Hee et al., 2018; Salawu et al., 2020). Later approaches have employed deep learning, transformers or hybrid models (Dadvar and Eckert, 2020; Chatzakou et al., 2017). Across studies, supervised text classification is a dominant approach, but more recent stronger systems examine user, network, and social media features (Chatzakou et al., 2017, 2019; Bozyigit et al., 2021), and social behavior (Ziems et al., 2020).

Covert harmful behaviors have received less attention in computational detections, likely because they are challenging to detect, even for humans, due to their concealed and ambiguous nature.

Trolling is a case in point: it often uses covert means to harass, digress or disrupt online spaces (Hardaker, 2010; Herring et al., 2002; Paakki et al.,

2021; Phillips, 2011). Many studies have approached trolling as a text classification problem as well, treating it similarly to aggression (Cambria et al., 2010; Capistrano et al., 2019; Seah et al., 2015; Shekhar et al., 2023). However, others have found that trolling is often covertly enacted (Hardaker, 2013; Paakki et al., 2021; Herring et al., 2002; Paakki et al., 2024), and efficient trolling detection requires features beyond explicit language used, features like how actors behave in interaction, for more accurate results (Paakki et al., 2024; Mojica de la Vega and Ng, 2018; Guy and Shapira, 2018; Kumar et al., 2014).

Another example of covert harmful behavior is disinformation dissemination, which usually utilizes manipulative and indirect means of influence (Bennett and Livingston, 2018; Wardle and Derakhshan, 2017). Trolling has also been used as a conversational strategy for disinformation dissemination (Mihaylov et al., 2018; Zelenkauskaitė and Niezgodą, 2017; Zannettou et al., 2019). Approaches to detecting such behaviors have often looked at stylistic and linguistic features and similarity of shared content in contrast to previously revealed disinformation (e.g., articles, messages, accounts) using machine learning, neural networks, transformers, or Large Language Models (LLMs) (Potthast et al., 2018; Rashkin et al., 2017; Shu et al., 2017; Zhou and Zafarani, 2020). Some have sought to identify disinformation dissemination by investigating networks that amplify misleading content online, often utilizing network analysis methods (Giglietto et al., 2020; Mihaylov et al., 2018; Tangherlini et al., 2020). Others have analyzed user accounts (Guo et al., 2022), cross-checking of media sources (Fung et al., 2021), and source reliability to identify disinformation dissemination (Guo et al., 2022).

2.2 Theories of harmful behavior

Online harms have been researched in many fields of research, including NLP, computer science and social cybersecurity, Human-Computer Interaction (HCI), psychology, social sciences and humanities, and philosophy. There are many competing definitions of what constitutes harm, and there is no single standardized definition. Established frameworks, however, distinguish between different types of harm, including physical, emotional or psychological, relational and financial harm (Scheuerman et al., 2021; Schoenebeck et al., 2023; Agrafiotis et al., 2018). In a typology of harmful

content by Banko et al. 2020, harm is constituted by acts that cross a threshold of severeness by violating laws, an individual's privacy of consent, causing immediate threat or physical or psychological trauma, or directly attacking an individual or a vulnerable target (Banko et al., 2020).

Here, I approach harm as relational: behavior both experienced by an individual or group as distressing or harmful, and behavior that violates the consent or autonomy, privacy or well-being of an individual or group. I concentrate on harms implicated in online interaction, thus consisting mainly of emotional, psychological or relational harms.

Harmful online behaviors have often been defined through intentionality – the concept that some behavior is harmful if an actor has the intention to cause harm through their actions (Coles and West, 2016; Golf-Papez and Veer, 2017; Keller et al., 2020; Zannettou et al., 2019). For instance, cyberbullying has been defined as intentional and repeated hostile demeanor, which is specifically directed at one vulnerable individual who is already known to the bully; trolls, on the other hand, are seen to intentionally target groups or communities (Coles and West, 2016; Golf-Papez and Veer, 2017).

Research on what constitutes these different covert harmful behaviors is also not uniform in their definitions. Some have defined it as boundary-maintenance between novice and experienced Internet users (Graham, 2019), political manipulation (Bennett and Livingston, 2018; Akhtar and Morrison, 2019), or manipulation or disruption of online conversations (Hardaker, 2010; Paakki et al., 2021). Meanwhile, others have stressed that trolling is not always harmful or malicious (Cruz et al., 2018; Sanfilippo et al., 2017), but may be used for grassroots activism (Spallaccia et al., 2019) or humor (Bishop, 2014). Thus, it needs careful consideration as to what is considered harmful behavior.

The listed behaviors are also not uniformly overt or covert, but may take advantage of a range of strategies. For example, trolling might sometimes utilize overt provocation, although it often relies on covert influence (Hardaker, 2013; Paakki et al., 2021). Although aggression and hate speech are often construed as more overt behaviors, they also might take more covert or indirect forms (ElSherief et al., 2021; Garandeau and Cillessen, 2006). These considerations underline the need for more robust novel methods for detecting covert online harms.

3 Limitations of the detection of online harms

Within the field of detecting online harms, some persistent caveats remain. For the detection of many of the covert harmful online behaviors described in the previous section, no single detection approach works across different datasets, contexts, languages, and definitions of the phenomenon (Yin and Zubiaga, 2021; Ramos et al., 2024; Salawu et al., 2020). Notable challenges are related to reliance on surface features, definition of the to-be-detected phenomenon, and requirement of intentionality.

First, models relying on overt text features are easily deceived (Gröndahl et al., 2018). Many approaches to detecting harmful online behaviors, as described in the previous section, have framed the problem as a text classification task, but such approaches often miss more subtle, indirect, covert behaviors. Even phenomena that are often more overt, like hate speech, may be enacted through more indirect ways, e.g. via the use of euphemisms. In addition, one of the greatest challenges has been the fact that many detection approaches relying on linguistic or textual features operate post hoc. They rely on training data gathered after some harmful content has been identified and aim to detect similar future content. Such methods are unavoidably reactive, and the past years have shown that more forward-looking methods are needed, ones that could better respond to the problem of ever-changing elusive strategies used to enact online harms.

To give an example, disinformation dissemination has often been identified based on textual content – linguistic features of individual messages or texts (Guo et al., 2022). There exists an underlying assumption that the veracity or even intentional misleadingness of information can be detected from these features of individual texts. Studies have found that the generalizability of text classification models beyond the material used in their training (and possibly the topics they represent) is poor (Guo et al., 2022).

It seems thus that efficient detection of especially covert harmful behaviors requires a different set of methods that look beyond explicit language used, features like how actors behave in interaction (Bozyigit et al., 2021; Chatzakou et al., 2019; Guy and Shapira, 2018; Paakki et al., 2024; Ziems et al., 2020).

Second, the field of automatic detection of online harms has often faced theoretical and definitional challenges. Researchers have noted that for the detection of many of the listed online harms, there is no consensus on the definition of the specific type of harm, like cyberbullying or aggression (Nobata et al., 2016; Paakki et al., 2024; Salawu et al., 2020). Clear definitions and distinctions between different harmful behaviors would arguably improve comparability of approaches, and aid researchers in developing context-driven detection methods.

Third, intentionality has been a central defining feature of many harmful behaviors that have been the focus of computational detection. This has been part of definitions of cyberbullying (Ziems et al., 2020), trolling (Hardaker, 2010), and disinformation dissemination (Wardle and Derakhshan, 2017). However, this is problematic for computational detection, because verifying true intentionality is challenging if not impossible to detect the intent behind someone's actions taken online. Actors are also happy to exploit these ambiguities regarding intentionality. True intentionality cannot be derived from text-based analysis, because if we understand it as purposefully and knowingly committing some action, it is mind-internal. Even self-proclaimed intentionality does not necessarily mean we get objective information about true intentionality: individuals may proclaim some intent, but they may lie or change their interpretation of their own behavior. This applies especially to deceptive behaviors.

4 Toward a norm-violation focused approach to online harms

Researchers have emphasized community norms and their violation as an important foundation for detecting harmful behaviors (Bruckman et al., 2006; Paakki et al., 2024). Harmful behavior like malicious, manipulative trolling can be efficiently detected by investigating actions taken that repeatedly violate community norms (Paakki et al., 2024). Overall, cross-disciplinary studies utilizing theories from SSH research and psychology have illustrated promising novel approaches to computational detection of harmful online behaviors (Ziems et al., 2020; Paakki et al., 2024).

Here, I will propose a framework for the detection of covert harmful online behaviors that relies on the theoretical understanding of social norms,

and how they might be violated through manipulation and deception.

I argue that a unifying feature to a range of covert harmful online behaviors is that they violate social norms in covert ways, utilizing deception and manipulation. There is a range of different social norms we live by, some of them more unconscious than others. Legal regulations, of course, set the most imperative boundaries for human behavior, often with dire consequences and a lower threshold for what constitutes as crossing the threshold for harmful reprehensible behavior. Covert strategies for implicating harm thus often avoid such clear boundary-crossing, and tend to utilize means that are not as clear.

Social norms shape how people should behave in social situations, providing predictability in interaction (Garfinkel, 1967; Stivers et al., 2024). Norms can be 'mores': norms with moral underpinnings, or they can 'folkways', norms, which are less restrictive, commonly accepted ways to interact with others (Garfinkel, 1967) (tracing back to Sumner, 1906). Mores can be formal written norms, or unwritten, and reflect the moral principles of a social group or community. Their violations are usually regarded more seriously with consequences. Less restrictive norms, in turn, reflect practical unwritten rules about appropriate social behavior without moral considerations, like how to greet someone new, and usually with less serious or no consequences for their violation (Garfinkel, 1967). However, these unwritten or unconscious norms may be strategically exploited to influence or harass others (Paakki et al., 2024, 2021).

While more binding norms may be reflected in legal regulations to govern lawful and moral behavior, unwritten norms are more easily exploited without consequence if a community does not clearly define thresholds for violation of these norms (see e.g. (Herring et al., 2002)). For example, although sending a specific emoji into a conversation on an online forum is usually not violating any norm, posting it there a thousand times may often be considered distressing.

Covert harmful behaviors often utilize deception and manipulation to influence or harass others (Donath, 2002; Paakki et al., 2021; Herring et al., 2002). Deception has been defined in communication research as "knowingly transmitting messages to a receiver with the intent to foster a false belief or conclusion" (Buller and Burgoon, 1996). Some more specific forms include identity

deception, which is to mislead others with regard to an actor's identity online (Donath, 2002). In many communities, deception is regarded as a negative, undesired phenomenon, for example because it is seen to violate rules of constructive interaction (Grice, 1989).

Manipulation, on the other hand, is defined as interpersonal or social influence that threatens its target's autonomy – often through covert means and without revealing the manipulator's goals (Coons and Weber, 2014; Susser et al., 2019). Some have defined manipulation through outcomes – the target losing their autonomy or ending up behaving in ways that are not beneficial to themselves (Jongepier and Klenk, 2022). Another approach is to define manipulation as hidden influence where the manipulator shows negligence with regard to the means they use – and the norms they may violate – to influence others or to get them to act in some desired way (Klenk, 2021). The norm violations in question may involve the manipulator violating social norms, or getting the victim's violation of self-interest (Jongepier and Klenk, 2022).

Deception and manipulation are central strategies in covert online harms like trolling. What unifies them is that both violate social norms that are often held in high regard in many communities, such as autonomy and truthfulness. However, at the micro level the violated norms are less restrictive unwritten norms, which enable the violator to not be too apparent and to escape consequences. Thus, I argue that research on the computational detection of covert online harms could benefit from closer examination of how models could identify norm violations.

4.1 Unwritten social norms

Unwritten social norms are, of course, challenging to pin-point, because they are often unconscious. Earlier research on human interaction has a long tradition in theorizing about these unwritten social rules. Theories regarding general accepted or constructive manners of interaction and communication can help to define more explicit community norms and criteria for identifying their violations. Next, I will describe Paul Grice's (1989) theory regarding unwritten norms of constructive interaction, which, at least in a Western context, can be helpful for identifying unwritten norms of interaction and communication. These include Paul Grice's theory of the Cooperative Principle (1975, 1989) and the conversational Maxims.

According to Grice, we humans tend to strive for cooperativeness, producing turns in interaction that advance a conversation and help to bring it forth toward mutual understanding – the Cooperative Principle (Grice, 1975, 1989). These rules are often unconsciously followed and expected. The conversational maxims, on the other hand, delineate unwritten rules according to which people tend to behave and expect others to behave, in order to contribute to the Cooperative Principle. These are the maxim of Quantity, maxim of Quality, maxim of Relevance and maxim of Manner.

Maxim of Quantity sets the expectation that participants will provide the required amount of information in their contributions to conversation so that it is understandable but does not include too much extra information (Grice, 1989). **The maxim of Quality** sets expectations with regard to the veracity of a contribution or information which is provided. The core expectation is that people will not present information they know to be false or unsure, and that they should not state a claim in case they do not have adequate evidence to back the claim (Grice, 1989). **The maxim of Relevance** refers to the expectation that participants will construct their contributions in a way that is relevant to the ongoing conversation (Grice, 1989). **The maxim of Manner**, in turn, is about how something is said: we expect others to avoid obscurity and ambiguity, and instead seek to communicate in a brief and orderly manner (Grice, 1989). Other norms, which should be considered in conjunction with the maxims to understand human behaviors in interaction, include politeness (Grice, 1989).

These maxims and their violations have also been identified as fruitful for understanding deceptive behavior (McCornack, 1992) and disinformation (Søe, 2021). In particular, violations of the maxim of Quality are covertly used to generate misleading epistemic implicatures, but also by violating other maxims in subtle ways (McCornack, 1992). Arguably, our tendency to expect that a deviation somehow adheres to the Cooperative Principle (Grice, 1989) opens possibilities for the norm-violator to continue their behavior.

5 RSD framework for covert harm detection

For the purposes of computational detection of covert online harms, I suggest that violations of norms can be identified through the lens of the

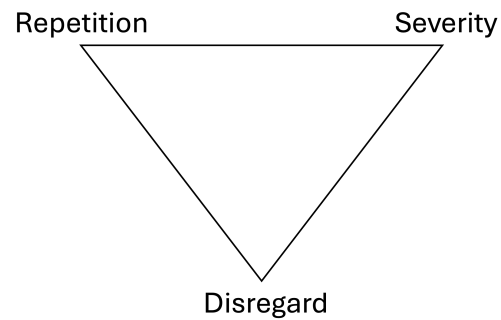


Figure 1: the RSD framework: dimensions for assessing harmfulness of norm violation.

‘RSD framework’, Figure 1.

Since communication, in essence, is to influence and to be influenced by others, as noted by Paul Grice (Grice, 1989), the question remains what then constitutes ‘harmful’ influence, or harmful online behavior. I suggest the RSD framework for assessing the harmfulness of norm-violations through the dimensions of **Repetition**, **Severity** and **Disregard**.

Repetition refers to whether a norm violation is repeated several times, or whether it is a one-time act. With regard to covert behaviors, repetition is perhaps one of the few objective, observable features that can be used as a kind of proxy for intentionality. It assumes that if a type of norm violation is repeatedly committed by the same actor in a social space, this signals some level of intent to commit the violation. Exceptions might include unwittingly committing said violation due to difficulty of understanding some unwritten social rule. Here explicit community norms can be helpful. For instance, covert behaviors like trolling may be extremely difficult to identify based on few messages in interaction (Paakki et al., 2024), and the requirement for repetition might protect both the individual as well as the group in unclear cases.

Severity relates to the seriousness of the norm violation. As mentioned, social and societal norms differ in levels of gravity: some are reflected in formal (written) legal regulations while others are informal (unwritten) norms with possibly no consequences for their violation.

Disregard, finally, refers to whether someone who violates a norm acknowledges a notification from the community about their unacceptable behavior. This is an important dimension to consider, because, especially when it comes to unwritten norms, there is a chance it was unintended. This

#	User	Post
1	T	Three Fatal Flaws in Greenhouse Gas Climate Science - Principia Scientific International [link] [shares content disproving climate change, rest not shown]
2	A	@T The flaws are yours. I suggest you go back and look at the science again.
3	T	@A, I think you need to resit Physics. Even NASA now admit it. [academia.edu link] " THE EARTH IS ACTUALLY COOLING," [rest not shown]
4	A	@T You are quoting propoganda websites saying what NASA says, not what NASA actually says. [rest not shown]

Table 1: Example of disinformation dissemination from Paakki (2024).

offers an opportunity for repair, which is an important part of how participants build mutual understanding in interaction (Grice, 1989; Meredith and Stokoe, 2014). Disregard also relates to the concept of accountability: when deviating from a norm, in well-intending interaction participants are expected to account for their actions – to explain deviant behavior. For example, this might be important when confronting someone about disinformation they have shared to an online community. Disregarding requests from others to account for a norm violation would count as a negative signal within the RSD framework.

The three dimensions of RSD operate together when assessing covert harmful behavior. While some norms might be so severe that even a one-time violation might result in banning the violator, many norm violations that are used strategically in covert behaviors are less severe. Thus, to detect harmful behavior, all dimensions need to be considered together.

To illustrate how RSD could be applied, consider Table 1. The conversation is an extract from a longer interaction in a news comment discussion (Paakki, 2024). T (the Troll) continuously shares content in the conversation that is considered by others as disinformation, and continues to do so despite being told off by others, like A in post 4. They deflect accountability, and in terms of RSD can be seen to score quite high on Repetition, Severity and Disregard. While R may be measured through number of violations in a conversation, D may be operationalized for instance through examination of accusations or accountability requests, following Paakki et al. 2024. Severity thresholds ultimately need to be contextually defined.

In other words, the boundary for harmful behavior could be crossed by repeatedly disregarding requests to explain oneself, while continuing to digress or harass the conversation, though not in

too extravagant ways. By paying attention to the three dimensions, the framework could also help set the boundary for behaviors that are not considered harmful in the community. For example, consider someone repeatedly posting to a group without using any commas. Although the level of repetition might be high, and they might completely disregard others asking them to use commas, this may not cross the threshold of harmful behavior, since the level of severity is too low.

Finally, norm violations need to be defined from a community and context-driven perspective: what are the central important norms in that specific community, how they might be violated, and what are the significant thresholds for considering a norm violation as harmful. Different communities may have different criteria.

6 Discussion

I have described the challenges that years of research into the computational detection of online harms have faced with covert forms of harmful online behaviors. While many models and approaches have been developed for detecting different harmful behaviors, covert harmful behaviors have remained challenging to detect due to their elusive and indirect nature. I have argued in this paper that approaches to detecting covert harmful behaviors could benefit from operationalizations of if and how community norms are violated.

For the detection of harmful behavior, I proposed the RSD framework, which seeks to assess the harmfulness of norm-violating behavior within three dimensions: Repetition, Severity and Disregard. I argue that the framework offers means for preparedness and readiness for identifying online behaviors that covertly violate community norms, and for identifying when the said behaviors cross the threshold for becoming harmful. This is important for communities in order to protect safe

spaces and to define thresholds for unacceptable or harmful behavior, and also criteria that help them justify why some penalty – like banning a user – was deemed necessary. Such criteria are especially important when it comes to unwritten norms that usually do not have any severe consequences or legal ramifications.

Although it is challenging to explicitly articulate social rules that are more unspoken rather than written, like 'generally accepted' rules of social behavior, from the perspective of online harm detection and prevention, it is also very important to do so. Like Susan Herring noted already in 2002, online communities may make themselves vulnerable to different types of harassment and manipulation by being too lax about the acceptable norms of behavior in their community.

Usually humans tend to expect others to act according to the unwritten generally accepted norms of constructive conversation. This means that we tend to interpret deviation from a norm first in a positive light, e.g., as a mistake, and the deviator is thus given a chance for repair or to account for their actions. It is thus necessary to consider some threshold for repetition and severity of said deviation, and also whether the said person explained their actions when notified or stopped acting in a certain way when asked.

Theories regarding conversational and interpersonal interaction may help to identify the sets of unwritten social norms that are important to a community, and to define in advance what would be considered a harmful violation. I delineated Gricean theory that can help identify unwritten social norms that may be used in covert harassment or manipulation against community members to disrupt the community. Though it is a somewhat idealized view of interaction, it offers insight into the unwritten social rules that may be broken in interaction.

Finally, harm should be defined in juxtaposition to non-harm. This means that having a clear set of community norms helps the community to justify why some behavior is not acceptable there. This also assumes that community norms are explicitly defined, and can guide computational detection. This is important, because what constitutes hate speech, for instance, is dependent on cultural and social context as well as its prevalent definitions (Nguyen et al., 2020).

It has been challenging to deal with covert behaviors like trolling online, because they operate in indirect and hidden ways. Also, many of the covert

behaviors described here are often defined through intentionality, which is not observable or an objective quality in online interaction. Explicit definitions of community norms can be used together with the RSD framework to investigate the observable characteristics of behaviors, and to arrive at an evidence-based assessment of norm violations and their harmfulness. In this way, the approach does not rely on intentionality.

Harm may also be a subjective experience: that is to say that individuals may experience different things as harassment. This is a potential limitation for the framework and for computational detection, and thus human moderation is key. The Disregard dimension of the RSD framework attempts to offer means to operationalize experiences of harm.

Finally, a significant question in the field is how to respond to the problems observed: *ex ante* (in advance) or *ex post* (afterwards). Online moderation can operate both ways. Many models aimed at recognition of harmful behaviors function after-the-fact, based on the content of already identified harms. However, the strategies and forms of covert harmful behaviors online evolve rapidly (Ferreira, 2022), so *ex post* approaches like text-based methods looking for similar content are easily left lagging behind. This underlines the need for novel frameworks, like the RSD framework, for the detection of online harms that can anticipate norm violations, even ones that operate in covert and hidden ways. The RSD framework can guide the operationalization of computational detection of especially covert harms through evidence-based assessment of norm violations.

Limitations

There have been also other persistent limitations to the detection of covert online harms, some of which I have not discussed here. Some have pointed out the dangers of potentially biased or politically charged detection or moderation (Garg et al., 2023; Paakki et al., 2024), ethical questions like explainability (Kiritchenko et al., 2021), and scalability issues for reliable detection (Paakki et al., 2024). In practice, methods are often limited by the requirement for scalability: the margin of error must be very small when processing huge data sets. These were out of the scope for this paper to consider in more depth. However, operationalization of detection based on community norms may be a helpful starting point for mitigating bias and limiting model space to the community in question.

References

- Ioannis Agraftiotis, Jason R C Nurse, Michael Goldsmith, Sadie Creese, and David Upton. 2018. [A taxonomy of cyber-harms: Defining the impacts of cyber-attacks and understanding how they propagate](#). *Journal of Cybersecurity*, 4(1):tyy006.
- Shazia Akhtar and Catriona M. Morrison. 2019. [The prevalence and impact of online trolling of UK members of parliament](#). *Computers in Human Behavior*, 99:322–327.
- Aish Albladi, Minarul Islam, Amit Das, Maryam Bigonah, Zheng Zhang, Fatemeh Jamshidi, Mostafa Rahgouy, Nilanjana Raychawdhary, Daniela Marghitu, and Cheryl Seals. 2025. [Hate speech detection using large language models: A comprehensive review](#). *IEEE Access*, 13:20871–20892.
- Segun Taofeek Aroyehun and Alexander Gelbukh. 2018. Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying*, TRAC '18, pages 90–97, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Adam Badawy, Emilio Ferrara, and Kristina Lerman. 2018. [Analyzing the digital traces of political manipulation: The 2016 Russian interference Twitter campaign](#). In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 258–265, Barcelona. IEEE.
- Michele Banko, Brendon MacKeen, and Laurie Ray. 2020. [A unified taxonomy of harmful content](#). In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 125–137, Online. Association for Computational Linguistics.
- W. Lance Bennett and Steven Livingston. 2018. [The disinformation order: Disruptive communication and the decline of democratic institutions](#). *European Journal of Communication*, 33(2):122–139.
- Jonathan Bishop. 2014. Representations of ‘trolls’ in mass media communication: a review of media-texts and moral panics relating to ‘internet trolling’. *International Journal of Web Based Communities*, 10(1):7–24.
- Alican Bozyigit, S. Utku, and E. Nasibov. 2021. [Cyberbullying detection: Utilizing social media features](#). *Expert Syst. Appl.*, 179:115001.
- Amy Bruckman, Catalina Danis, Cliff Lampe, Janet Sternberg, and Chris Waldron. 2006. [Managing deviant behavior in online communities](#). In *CHI '06 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '06, page 21–24, New York, NY, USA. Association for Computing Machinery.
- David B Buller and Judee K Burgoon. 1996. Interpersonal deception theory. *Communication theory*, 6(3):203–242.
- Erik Cambria, Praphul Chandra, Avinash Sharma, and Amir Hussain. 2010. [Do not feel the trolls](#). In *Proceedings of ISWC, ISWC '10*, pages 1–12, Aachen, Germany. CEUR Workshop Proceedings.
- Jose Lorenzo C. Capistrano, Jessie James P. Suarez, and Prospero C. Naval. 2019. [Salsa: Detection of cyber trolls using sentiment, aggression, lexical and syntactic analysis of tweets](#). In *Proceedings of the 9th International Conference on Web Intelligence, Mining and Semantics, WIMS '19*, pages 1–6, New York, NY, United States. ACM.
- Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. [Mean birds: Detecting aggression and bullying on twitter](#). In *Proceedings of the 2017 ACM on Web Science Conference*, WebSci '17, page 13–22, New York, NY, USA. Association for Computing Machinery.
- Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Emiliano De Cristofaro, G. Stringhini, A. Vakali, and N. Kourtellis. 2019. [Detecting cyberbullying and cyberaggression in social media](#). *ACM Transactions on the Web (TWEB)*, 13:1 – 51.
- Bryn Alexander Coles and Melanie West. 2016. Trolling the trolls: Online forum users constructions of the nature and properties of trolling. *Computers in Human Behavior*, 60:233–244.
- Christian Coons and Michael Weber. 2014. *Manipulation: theory and practice*. Oxford University Press.
- Angela Gracia B. Cruz, Yuri Seo, and Mathew Rex. 2018. [Trolling in online communities: A practice-based theoretical perspective](#). *The Information Society*, 34(1):15–26.
- Maral Dadvar and Kai Eckert. 2020. Cyberbullying detection in social networks using deep learning based models. In *International conference on big data analytics and knowledge discovery*, pages 245–255. Springer.
- Judith S. Donath. 2002. *Identity and deception in the virtual community*. Routledge, London.
- Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 345–363.
- Ricardo Ribeiro Ferreira. 2022. Liquid disinformation tactics: Overcoming social media countermeasures through misleading content. *Journalism Practice*, 16(8):1537–1558.
- Paula Fortuna and Sérgio Nunes. 2018. [A survey on automatic detection of hate speech in text](#). *ACM Computing Surveys (CSUR)*, 51:1 – 30.

- Camille Francois and Herb Lin. 2021. [The strategic surprise of russian information operations on social media in 2016 in the united states: mapping a blind spot](#). *Journal of Cyber Policy*, 6(1):9–30.
- Yi Fung, Christopher Thomas, Revanth Gangi Reddy, Sandeep Polisetty, Heng Ji, Shih-Fu Chang, Kathleen McKeown, Mohit Bansal, and Avi Sil. 2021. [InfoSurgeon: Cross-media fine-grained information consistency checking for fake news detection](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1683–1698, Online. Association for Computational Linguistics.
- Claire F Garandeau and Antonius HN Cillessen. 2006. From indirect aggression to invisible aggression: A conceptual view on bullying and peer group manipulation. *Aggression and violent behavior*, 11(6):612–625.
- Harold Garfinkel. 1967. *Studies in ethnomethodology*. Englewood Cliffs, NJ: Prentice-Hall.
- Tanmay Garg, Sarah Masud, Tharun Suresh, and Tanmoy Chakraborty. 2023. Handling bias in toxic speech detection: A survey. *ACM Computing Surveys*, 55(13s):1–32.
- Fabio Giglietto, Nicola Righetti, Luca Rossi, and Gilda Marino. 2020. It takes a village to manipulate the media: coordinated link sharing behavior during 2018 and 2019 Italian elections. *Information, Communication & Society*, 23(6):867–891.
- Maja Golf-Papez and Ekant Veer. 2017. Don’t feed the trolling: rethinking how online trolling is being defined and combated. *Journal of Marketing Management*, 33(15-16):1336–1354.
- Elyse Graham. 2019. [Boundary maintenance and the origins of trolling](#). *New Media & Society*, 21(9):2029–2047.
- Paul Grice. 1975. Logic and conversation. In *Syntax and Semantics*, volume 3, Speech Acts, pages 41–58. New York, NY: Academic Press.
- Paul Grice. 1989. *Studies in the Way of Words*. Harvard University Press.
- Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N. Asokan. 2018. [All you need is "love": Evading hate speech detection](#). In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security, AISec ’18*, page 2–12, New York, NY, USA. Association for Computing Machinery.
- Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. [A survey on automated fact-checking](#). *Transactions of the Association for Computational Linguistics*, 10:178–206.
- Ido Guy and Bracha Shapira. 2018. [From royals to vegans: Characterizing question trolling on a community question answering website](#). In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR ’18*, page 835–844, New York, NY, USA. Association for Computing Machinery.
- Mark A. Hamilton. 2012. [Verbal aggression: Understanding the psychological antecedents and social consequences](#). *Journal of Language and Social Psychology*, 31(1):5–12.
- Claire Hardaker. 2010. [Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions](#). *Journal of Politeness Research. Language, Behaviour, Culture*, 6(2):215–242.
- Claire Hardaker. 2013. [“Uh. . . not to be nit-picky,,,,,but. . . the past tense of drag is dragged, not drug.”: An overview of trolling strategies](#). *Journal of Language Aggression and Conflict*, 1(1):58–86.
- Cynthia Van Hee, Gilles Jacobs, Chris Emmery, Bart Desmet, Els Lefever, B. Verhoeven, G. Pauw, Walter Daelemans, and Veronique Hoste. 2018. [Automatic detection of cyberbullying in social media text](#). *PLoS ONE*, 13.
- Susan Herring, Kirk Job-Sluder, Rebecca Scheckler, and Sasha Barab. 2002. [Searching for safety online: Managing "trolling" in a feminist forum](#). *The Information Society*, 18(5):371–384.
- Fleur Jongepier and Michael Klenk. 2022. Online manipulation: Charting the field. In *The Philosophy of Online Manipulation*, pages 15–48. Routledge.
- Franziska B Keller, David Schoch, Sebastian Stier, and JungHwan Yang. 2020. Political astroturfing on twitter: How to coordinate a disinformation campaign. *Political communication*, 37(2):256–280.
- Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen C Fraser. 2021. Confronting abusive language online: A survey from the ethical and human rights perspective. *Journal of Artificial Intelligence Research*, 71:431–478.
- Michael Klenk. 2021. [Interpersonal manipulation](#). *SSRN Electronic Journal*.
- Ritesh Kumar, Atul Kr. Ojha, Marcos Zampieri, and Shervin Malmasi, editors. 2018. [Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying \(TRAC-2018\)](#). Association for Computational Linguistics, Santa Fe, New Mexico, USA.
- Srijan Kumar, Francesca Spezzano, and VS Subrahmanian. 2014. Accurately detecting trolls in slashdot zoo via decluttering. In *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM ’14*, pages 188–195, Beijing, China. IEEE.

- Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and O. Frieder. 2019. [Hate speech detection: Challenges and solutions](#). *PLoS ONE*, 14.
- J. Malik, Guansong Pang, and Anton van den Hengel. 2022. [Deep learning for hate speech detection: a comparative study](#). *International Journal of Data Science and Analytics*, 20:3053 – 3068.
- Steven A McCornack. 1992. Information manipulation theory. *Communications Monographs*, 59(1):1–16.
- Joanne Meredith and Elizabeth Stokoe. 2014. [Repair: Comparing facebook ‘chat’ with spoken interaction](#). *Discourse & communication*, 8(2):181–207.
- Todor Mihaylov, Tsvetomila Mihaylova, Preslav Nakov, Lluís Màrquez, Georgi D. Georgiev, and Ivan Kolev Koychev. 2018. [The dark side of news community forums: Opinion manipulation trolls](#). *Internet Research*, 28(5):1292–1312.
- Luis Gerardo Mojica de la Vega and Vincent Ng. 2018. [Modeling trolling in social media conversations](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3701–3707, Miyazaki, Japan. European Language Resources Association (ELRA).
- Nanlir Sallau Mullah and W. Zainon. 2021. [Advances in machine learning algorithms for hate speech detection in social media: A review](#). *IEEE Access*, 9:88364–88376.
- Dong Nguyen, Maria Liakata, Simon DeDeo, Jacob Eisenstein, David Mimno, Rebekah Tromble, and Jane Winters. 2020. How we do things with words: Analyzing text as social and cultural data. *Frontiers in Artificial Intelligence*, 3:62.
- Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. [Abusive language detection in online user content](#). In *Proceedings of the 25th International Conference on World Wide Web, WWW ’16*, page 145–153, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.
- Henna Paakki. 2024. Asymmetric conversational strategies-methods for detecting manipulative online trolling.
- Henna Paakki, Heidi Vepsäläinen, and Antti Salovaara. 2021. [Disruptive online communication: How asymmetric trolling-like response strategies steer conversation off the track](#). *Computer Supported Cooperative Work (CSCW)*, pages 1–37.
- Henna Paakki, Heidi Vepsäläinen, Antti Salovaara, and Bushra Zafar. 2024. Detecting covert disruptive behavior in online interaction by analyzing conversational features and norm violations. *ACM Transactions on Computer-Human Interaction*, 31(2):1–43.
- Whitney Phillips. 2011. [LOLing at tragedy: Facebook trolls, memorial pages and resistance to grief online](#). *First Monday*, 16(12).
- Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. [A stylometric inquiry into hyperpartisan and fake news](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 231–240, Melbourne, Australia. Association for Computational Linguistics.
- Gil Ramos, Fernando Batista, Ricardo Ribeiro, Pedro Fialho, Sérgio Moro, António Fonseca, Rita Guerra, Paula Carvalho, Catarina Marques, and Cláudia Silva. 2024. [A comprehensive review on automatic hate speech detection in the age of the transformer](#). *Social Network Analysis and Mining*, 14.
- Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. [Truth of varying shades: Analyzing language in fake news and political fact-checking](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics.
- Anchal Rawat, Santosh Kumar, and Surender Singh Samant. 2024. [Hate speech detection in social media: Techniques, recent trends, and future challenges](#). *Wiley Interdisciplinary Reviews: Computational Statistics*, 16.
- Amir H. Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. 2010. [Offensive language detection using multi-level classification](#). In *Canadian Conference on Artificial Intelligence*, volume 6085 of *LNCS*, pages 16–27, Ottawa, Canada. Springer, Springer.
- Julian Risch and Ralf Krestel. 2020. [Bagging BERT models for robust aggression identification](#). In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 55–61, Marseille, France. European Language Resources Association (ELRA).
- B.S. Sahana, G. Sandhya, R.S. Tanuja, Sushma Ellur, and A. Ajina. 2020. [Towards a safer conversation space: Detection of toxic content in social media \(student consortium\)](#). In *2020 IEEE Sixth International Conference on Multimedia Big Data, BigMM*, pages 297–301, New Delhi, India. IEEE.
- Semiu Salawu, Yulan He, and Joan A. Lumsden. 2020. [Approaches to automated detection of cyberbullying: A survey](#). *IEEE Transactions on Affective Computing*, 11:3–24.
- Madelyn Sanfilippo, Shengnan Yang, and Pnina Fichman. 2017. Trolling here, there, and everywhere: Perceptions of trolling behaviors in context. *Journal of the Association for Information Science and Technology*, 68(10):2313–2327.

- Morgan Klaus Scheuerman, Jialun Aaron Jiang, Casey Fiesler, and Jed R. Brubaker. 2021. [A framework of severity for harmful content online](#). *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2).
- S. Schoenebeck, Cliff Lampe, and Penny Triu. 2023. [Online harassment: Assessing harms and remedies](#). *Social Media + Society*, 9.
- Chun Wei Seah, Hai Leong Chieu, Kian Ming A. Chai, Loo-Nin Teow, and Lee Wei Yeong. 2015. [Troll detection by domain-adapting sentiment analysis](#). In *18th International Conference on Information Fusion*, Fusion, pages 792–799, Sun City, South Africa. IEEE, IEEE.
- Shashi Shekhar, Hitendra Garg, Rohit Agrawal, Shivendra Shivani, and Bhisham Sharma. 2023. [Hatred and trolling detection transliteration framework using hierarchical lstm in code-mixed social media text](#). *Complex & Intelligent Systems*, 9(3):2813–2826.
- Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.
- Sille Obelitz S  . 2021. A unified account of information, misinformation, and disinformation. *Synthese*, 198(6):5929–5949.
- Sara Owsley Sood, Judd Antin, and Elizabeth Churchill. 2012. [Using crowdsourcing to improve profanity detection](#). In *AAAI Spring Symposium Series*, AAAI Publications, pages 69–74, Palo Alto, California. AAAI.
- Beatrice Spallaccia and 1 others. 2019. Trolling patriarchy: Online activism and feminist education against gendered cyberhate. In *Literature, Gender and Education for Children and Young Adults/Litt  rature, genre,   ducation pour l’enfance et la jeunesse*, pages 151–169. Bononia University Press (BUP).
- T. Stivers, Andrew Chalfoun, and Giovanni Rossi. 2024. [To err is human but to persist is diabolical: Toward a theory of interactional policing](#). *Frontiers in Sociology*, 9.
- William Graham Sumner. 1906. *Folkways: A study of the sociological importance of usages, manners, customs, mores, and morals*. New York: Ginn and Co.
- Daniel Susser, Beate Roessler, and Helen Nissenbaum. 2019. Technology, autonomy, and manipulation. *Internet policy review*, 8(2):1–22.
- Timothy R Tangherlini, Shadi Shahsavari, Behnam Shahbazi, Ehsan Ebrahimzadeh, and Vwani Roychowdhury. 2020. An automated pipeline for the discovery of conspiracy and conspiracy theory narrative frameworks: Bridgegate, pizzagate and storytelling on the web. *PloS one*, 15(6):e0233879.
- Lu Wang and Claire Cardie. 2014. [A piece of my mind: A sentiment analysis approach for online dispute detection](#). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 693–699, Baltimore, Maryland. Association for Computational Linguistics.
- Claire Wardle and Hossein Derakhshan. 2017. *Information disorder: Toward an interdisciplinary framework for research and policymaking*, volume 27. Council of Europe Strasbourg.
- Wenjie Yin and A. Zubiaga. 2021. [Towards generalisable hate speech detection: a review on obstacles and solutions](#). *PeerJ Computer Science*, 7.
- G Zagni and T Canetta. 2023. Generative ai marks the beginning of a new era for disinformation.
- Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. [SemEval-2019 task 6: Identifying and categorizing offensive language in social media \(OffensEval\)](#). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
- Savvas Zannettou, Michael Sirivianos, Jeremy Blackburn, and Nicolas Kourtellis. 2019. The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans. *Journal of Data and Information Quality (JDIQ)*, 11(3):1–37.
- Asta Zelenkauskait   and Brandon Niezgoda. 2017. [“stop kremlin trolls:” ideological trolling as calling out, rebuttal, and reactions on online news portal commenting](#). *First Monday*, 22(5).
- Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018. [Conversations gone awry: Detecting early signs of conversational failure](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1350–1361, Melbourne, Australia. Association for Computational Linguistics.
- Xinyi Zhou and Reza Zafarani. 2020. [A survey of fake news: Fundamental theories, detection methods, and opportunities](#). *ACM Comput. Surv.*, 53(5).
- Yanling Zhou, Yanyan Yang, Han Liu, Xiufeng Liu, and N. Savage. 2020. [Deep learning based fusion approach for hate speech detection](#). *IEEE Access*, 8:128923–128929.
- Caleb Ziems, Ymir Vigfusson, and Fred Morstatter. 2020. [Aggressive, repetitive, intentional, visible, and imbalanced: Refining representations for cyberbullying classification](#). *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):808–819.