



Research Article

The unappreciated role of intent in algorithmic moderation of abusive content on social media

A significant body of research is dedicated to developing language models that can detect various types of online abuse, for example, hate speech, cyberbullying. However, there is a disconnect between platform policies, which often consider the author's intention as a criterion for content moderation, and the current capabilities of detection models, which typically lack efforts to capture intent. This paper examines the role of intent in the moderation of abusive content. Specifically, we review state-of-the-art detection models and benchmark training datasets to assess their ability to capture intent. We propose changes to the design and development of automated detection and moderation systems to improve alignment with ethical and policy conceptualizations of these abuses.

Authors: Xinyu Wang (1), Sai Koneru (1), Pranav Narayanan Venkit (1), Brett Frischmann (2), Sarah Rajtmajer (1)

Affiliations: (1) College of Information Sciences and Technology, Pennsylvania State University, USA, (2) Charles Widger School of Law, Villanova University, USA

How to cite: Wang, X., Koneru, S., Venkit, P. N., Frischmann, B., & Rajtmajer, S. (2025). The unappreciated role of intent in algorithmic moderation of abusive content on social media. *Harvard Kennedy School (HKS) Misinformation Review*, 6(3).

Received: March 25th, 2025. Accepted: July 2nd, 2025. Published: July 29th, 2025.

Research questions

- What role does intent play in existing social media policies for abuse moderation?
- What is the current state of annotating and detecting common forms of online abuse, focusing on hate speech and cyberbullying?
- How can intent be incorporated into existing annotation, detection, and moderation pipelines to align with content moderation policies?

Essay summary

- As social media platforms work to balance free expression with the prevention of harm and abuse, user intent is often cited in platform policies as a determinant of appropriate action. However, we surveyed recent scholarly research and found that the role of intent is underappreciated or, often, wholly ignored during the annotation, detection, and in-practice moderation of online abuse.
- Capturing users' intent from their text is an exceptionally hard problem. It is hard for a human reader to understand the intent of an author; it is even more challenging for an algorithm to do the same. Despite advancements in NLP, today's state-of-the-art approaches cannot reliably infer

¹ A publication of the Shorenstein Center on Media, Politics and Public Policy at Harvard University, John F. Kennedy School of Government.

user intent from short text, particularly if not provided with sufficient context. We highlight the features and metadata considered by existing algorithms, their prevalence, and their plausible impacts on understanding intent. Based on our findings, we put forth recommendations for the design of abuse detection and mitigation frameworks that include 1) robust training datasets annotated with context that reflect the complexities of intent, 2) state-of-the-art detection models that use contextual information as input and provide explanations as output, 3) moderation systems that combine automated detection with wisdom of the crowd to accommodate evolving social norms, and 4) friction-focused platform designs that both offer users opportunities to reflect on their intent before sharing and generate useful data regarding user intent.

Implications

Substantial interdisciplinary literature seeks to define, detect, measure, and model different types of abusive content online. Industry efforts to moderate abusive content at scale rely heavily on automated, algorithmic systems. These systems routinely fail (in terms of both false positives and false negatives). This paper explains why and proposes a series of potential improvements.

The basic argument is straightforward. Whether or not content is abusive, according to most platform policies and legal definitions, depends on the state of mind of a human being, usually the speaker. Yet algorithmic systems that process content (e.g., the speaker's words) are unable to reliably determine a person's state of mind. More and different kinds of information are needed.

The term *abuse* spans a spectrum of harmful language, including generalized hate speech and specific offenses such as sexism and racism. Definitions of different types of abuse often intersect and lack precise boundaries. Later, we consider various definitions and taxonomies proposed to describe abusive content.

Common to most definitions of digital abuse is some notion of *intent* (e.g., French et al., 2023; Hashemi, 2021; Molina et al., 2021; Vidgen & Derczynski, 2020). Intent is a subjective state of mind attributable to an actual person, typically the speaker, poster, or sharer. In cognitive science, ethics, law, and philosophy, intent is a contested and complicated concept. As Frischmann & Selinger (2018) explain:

Intention is a mental state that is part belief, part desire, and part value. My intention to do something—say to write th[is] explanatory text... or to eat an apple—entails (1) beliefs about the action, (2) desire to act, and (3) some sense of value attributable to the act (p. 364).

In pragmatic ethical and legal contexts, the focus often turns to evidence of intent. For example, a written signature is considered an objective manifestation that a person intends to enter into a contract (Frischmann & Vardi, 2024). Thus, by including intent in definitions of abusive content, the implicit challenge is divining subjective state of mind from evidence manifested in the content itself and the surrounding context.

While we focus on intent as a central element in how platforms define and moderate abuse, we do not claim that interpreting intent is the only or universally preferred approach. Some frameworks prioritize observable harms or outcomes, especially in contexts like mis- and disinformation (Mirza et al., 2023; Scheuerman et al., 2021). Our aim is not to displace these frameworks but to critically assess the feasibility and implications of relying on intent when it is embedded in platform policy and annotation practices.

Unsurprisingly, intent is exceedingly difficult to capture algorithmically through analysis of short text (Gao & Huang, 2017; MacAvaney et al., 2019; Wang et al., 2020). ToxicBERT (Hanu & Unitary team, 2020), for example, can label sentences as hateful or toxic but lacks the ability to interpret context in the input

(MacAvaney et al., 2019; Wang et al., 2020). ToxicBERT flags the sentence “I’m going to kill you if you leave the dishes for me again” as toxic and threatening; it fails to differentiate between literal and figurative language. Supervised algorithms like ToxicBERT rely on curated, typically human-annotated training datasets. However, the complexities of intent are eventually reduced to simple class labels—such as hate speech or not hate speech. The underlying assumption of this approach is that human annotators are furnished with adequate contextual information to make these judgements.

In this paper, we survey the landscape of the online abuse moderation policies of major social media platforms. We examine existing taxonomies of online abuse and survey existing algorithms for abuse detection. We focus on hate speech and bullying due to their prevalence, but our work can and should be extended to other content categories for which intent is relevant, including mis-/disinformation (Kruger et al., 2024). We examine the training datasets underlying these algorithms and the role (if any) of intent during dataset annotation. Finally, we survey the set of features extracted from training datasets and used for algorithm development. These features reflect the varied context available to algorithmic content moderation systems. Our findings motivate a set of recommendations to better align abuse detection algorithms with platform policies (see Figure 1).

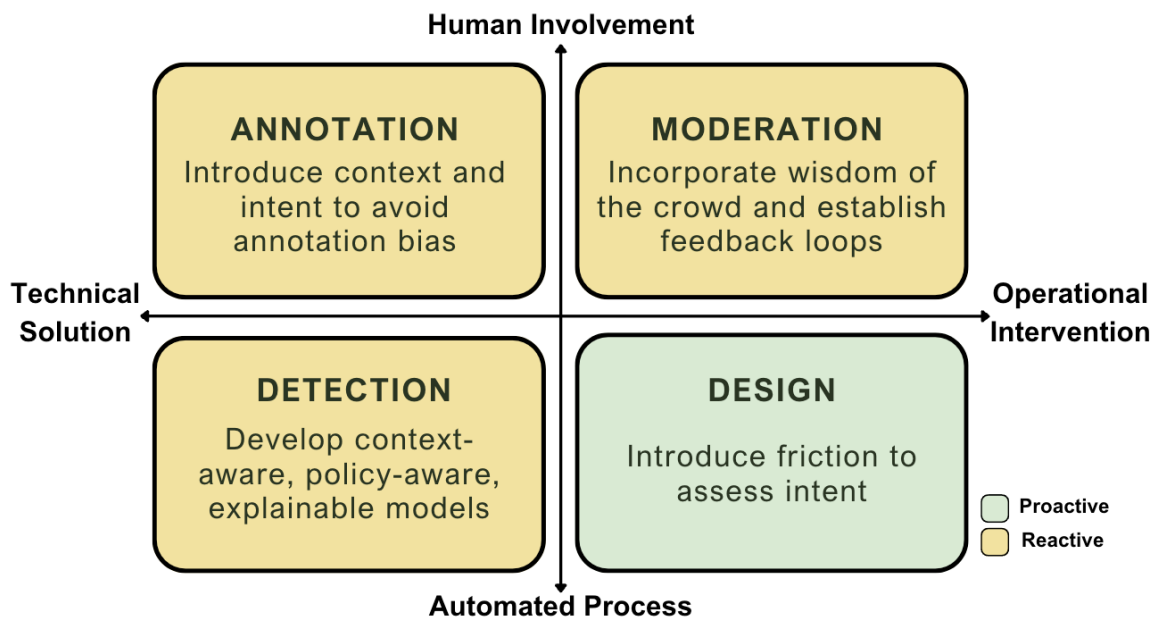


Figure 1. Recommendations to better capture intent.

Recommendation: Annotation—Introduce context and intent during dataset annotation

Dataset curators must recognize social, cultural, and other contextual variations present in natural language, and design annotation tasks sensitive to and mindful of these differences. Moreover, dataset curators must provide sufficient context to annotators to permit accurate assessment of intent, in particular (Anuchitanukul & Ive, 2022). Examples of such context include conversation threads between initiators and targets, user history and metadata, or norms defined by the specific platform. We propose a combined codebook-datasheet in the form of a structured set of questions for dataset curators and annotators to consider prior to annotation—particularly for online abuse datasets. This framework integrates annotation-specific guidance (codebook) with dataset-level context (datasheet). The full set of questions is in Appendix A. Responses to these questions should be made transparent and explicit to all stakeholders prior to annotation. The structure of this framework is informed by qualitative approaches

that treat annotation as an interpretive task shaped by context and ambiguity, rather than a purely objective labeling (Charmaz, 2006). By encouraging dataset curators and annotators to reflect on definitions, contextual cues, and the relationships between initiators and targets, the Codebook-Datasheet promotes more consistent and intent-sensitive annotation practices.

Challenge: Trade-offs between capturing intent and achieving high annotation agreement

A primary challenge in enhancing the annotation phase with contextual details is the trade-off between accurately inferring an individual's intent reflected through content and maintaining high agreement among annotators (Ross et al., 2017). Context can be subjective, and different annotators might interpret the same information differently based on their backgrounds, experiences, and biases (Joseph et al., 2017; Lynn et al., 2019). Providing more context also raises concerns about privacy, annotator fatigue, and task scalability. On the other hand, annotator disagreements are sometimes due to lack of sufficient context (Zhang et al., 2023). In cases where intent is ambiguous, omitting context may lead to systematic mislabeling or overly reductive interpretations of the content. Consequently, establishing standardized guidelines that incorporate diverse perspectives is essential to mitigate this concern and improve annotation consistency.

Recommendation: Detection—Develop context-aware, policy-aware, explainable models

Incorporation of contextual features into detection algorithms can substantially improve the accuracy of intent-based abuse detection (Markov & Daelemans, 2022; Menini et al., 2021). Recent progress in NLP, such as the introduction of retrieval-augmented generation (RAG) for large language models (Li et al., 2024), supports access to relevant contextual information from knowledge bases—for example, a user's past interactions—in real-time (Shi et al., 2024). Unlike earlier methods that rely on fixed input windows or static embeddings of prior conversations, RAG retrieves the most semantically relevant context dynamically, enabling more targeted and interpretable use of prior information. This may be especially beneficial for platforms aiming to implement context-aware moderation policies that adapt to dynamic social norms. Likewise, state-of-the-art large language models (LLMs) are more effective at executing rule-based moderation (Kumar et al., 2023); platform policies can be retrieved as context and used directly during the algorithmic decision process. Importantly, advances in explainable AI (XAI) should help users and developers understand model outputs (Islam et al., 2023; Kohli & Devi, 2023). Examples of XAI techniques include LIME, SHAP, and attention heatmaps, which help explain why a model flagged a post as abusive (Muhammadiyah et al., 2025). These tools can help moderators identify whether misclassifications stem from sarcasm, missing context, or ambiguous language—factors especially relevant when intent is at the core. XAI can be used, for example, to pinpoint why a model might misunderstand particular intentions or contexts and ensure compliance with social media moderation policies that require explanations of AI-driven decisions. We advocate for the implementation of policy-aware and dynamic abuse detection, which can be facilitated by XAI and retrieval-augmented models.

Challenge: Balancing performance and applicability

Incorporating contextual features presents a significant challenge, particularly in terms of performance, as standard detection models often prove to be overly optimistic (Menini et al., 2021). Ensuring that the system achieves high performance without compromising its applicability can be difficult; thus, advanced machine learning models and continuous algorithm training with updated datasets are required to address this balance effectively (Scheuerman et al., 2021). Additionally, integrating XAI methods improves transparency, but they may also reduce model efficiency when requiring simplification of underlying

architectures (Crook et al., 2023). Thus, designing explainability into systems without sacrificing model performance remains an ongoing area of research.

Recommendation: Moderation—Incorporate wisdom of the crowd and establish feedback loops

While algorithmic systems play an essential role in content moderation at scale, they also have limitations. Algorithms struggle to process and make sense of the nuance and subtlety of human communication (Bender & Koller, 2020). Humans, although generally better at understanding context and intent, can exhibit inconsistency and bias (Basile et al., 2022).

We suggest that moderation can be improved by hybrid approaches, leveraging “wisdom of the crowd”—utilizing user reports and community feedback—alongside, and even integrated with, algorithmic systems to identify and moderate abusive content (Allen et al., 2021; Pröllochs, 2022). The potential effectiveness of crowd-based moderation is evident in several real-world implementations, such as Twitter’s Birdwatch (now X’s Community Notes), where users collaboratively add contextual notes to potentially harmful or misleading content;² Reddit’s voting and reporting mechanisms that shape content visibility;³ and Wikipedia’s multi-layered editorial and review workflows.⁴ Additionally, we recommend establishing clear pathways for feedback so that moderators can provide insights back to the automated detection systems to support iterative improvements, for example, as new labeled training data. Throughout these processes, platforms must design moderation systems to be sensitive to evolving definitions of inappropriate content and dynamic societal norms across regions and cultures.

Challenge: Managing the scale and the biases inherent in crowd-sourced moderation

The challenge in this phase lies in managing the scale of data and potential biases that can arise from crowd-sourced inputs. User reports can be influenced by personal biases or coordinated attacks, leading to false positives or negatives (Jhaver et al., 2019). Uneven participation rates mean that moderation decisions may disproportionately reflect the views of a narrow group of users. Moreover, the quality of crowd input may vary widely depending on task design, interface clarity, and community norms (Gadiraju et al., 2015). Implementing robust filtering algorithms to verify and validate user-generated reports before they influence the moderation process is crucial for maintaining the integrity of the system.

Recommendation: Design—Introduce friction to generate evidence of intent and enable more intentional actions

For the most part, we focus on the detection and evaluation of content flowing through social media systems as if the moderation pipeline is designed to operate independently, only triggering governance procedures upon detection of content that violates a policy. Of course, the actions of users depend to some degree on the affordances of the social media platform itself.

If we relax the assumption of independence between platform design and content moderation pipeline design, a range of friction-in-design (Frischmann & Selinger, 2018; Frischmann & Benesch, 2023) measures might generate reliable evidence of intent to better guide content moderation systems. For example, platforms might introduce prompts that query users about their intentions when users post or

² See: https://blog.x.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation; <https://help.x.com/en/using-x/community-notes>

³ See: <https://support.reddithelp.com/hc/en-us/articles/7419626610708-What-are-upvotes-and-downvotes>; <https://support.reddithelp.com/hc/en-us/articles/360058309512-How-do-I-report-a-post-or-comment>

⁴ See: https://en.wikipedia.org/wiki/Wikipedia:Editing_policy

share (certain types of) content. Such prompts might be triggered based on different criteria, such as accuracy, authenticity, or intended audience. The prompt could provide a simple means for the user to express their intentions. The prompt might be framed in terms of *purpose*. Not only would a response provide potentially reliable evidence of intent that would be useful for moderation, but it also would provide the user with an opportunity to think about their own intentions. Other friction-in-design measures could provide users with knowledge about the potential consequences of their actions. Such measures would generate another source of reliable evidence about intent.

Prior research has shown that friction-based interventions—such as prompts during content creation, interstitial warnings, limits on message forwarding, and credibility nudges—can reduce misinformation sharing, offensive language, and improve user reflection (Clayton et al., 2020; Egelman et al., 2008; Fazio, 2020; Jahn et al., 2023; Kaiser et al., 2021). These examples underscore that friction can support both behavioral change and the elicitation of user intent for moderation purposes.

Challenge: Aligning business models and regulatory frameworks with slow governance for a safe, healthy digital environment

The basic idea of the friction-in-design strategy is to focus on the dynamic interactions between the social media platform and the content moderation pipeline from a design perspective. The associated challenges here are substantively different from the technical challenges we have outlined in prior sections. The success of friction-in-design approaches center around platform business models and regulatory frameworks that prioritize and reward healthy information ecosystems (Frischmann & Benesch, 2023). Frischmann & Sanfilippo (2023) emphasize the need to replace the dominant platform design logic that prioritizes frictionless, seamless interactions with one that embraces *slow governance*. Here, what that means is utilizing digital speedbumps not only to slow traffic but also, more importantly, for the instrumental functions of generating reliable evidence of intent and affording users the opportunity to be more intentional in their online behavior. Yet, as is probably all too obvious, this presents a fundamental challenge to existing business models.

Findings

Finding 1: Existing taxonomies of digital abuse are diverse and often ambiguous, making consistent categorization difficult.

Extensive literature details the diverse forms of online harm across various digital platforms (e.g., Arora et al., 2023; Im et al., 2022; Keipi et al., 2016; Scheuerman et al., 2021). These studies categorize a range of abusive speech, with prominent themes including hate speech, cyberbullying, and discrimination (Arora et al., 2023; Ghosh et al., 2024; Scheuerman et al., 2021). Each of these harms is consequent to nuanced user interactions, which are often platform-specific, culture-sensitive, and context-dependent. Prior work has highlighted the differing definitions of online harms in the literature (Fortuna & Nunes, 2018); this ambiguity poses challenges for the development of consistent and effective moderation pipelines.

Various approaches to classifying online abuse emphasize different parameters. Some taxonomies highlight the *target* of abuse—whether directed at individuals, groups, or concepts (Al Mazari, 2013; Vidgen et al., 2019; Waseem et al., 2017). Others focus on *characteristics* of the abuse—whether it is explicit or implicit (Mladenović et al., 2021), and still others explore subcategories of abuse or harm (Gashroo & Mehrotra, 2022; Lewandowska-Tomaszczyk et al., 2023). An overview of existing taxonomies is provided in Appendix B. Our work focuses on intent, which features prominently in platform policies

but is often implicit in current taxonomies. For example, intent may be operationalized through the notion of targeting. We emphasize two common types of explicit online abuse: hate speech and cyberbullying (Wiegand et al., 2019). These forms of abuse have not only received significant attention in natural language processing (NLP) but reflect intent—deliberate aim to harm or intimidate specific individuals or groups. To clarify how these categories are typically defined in the literature, we provide the following commonly used definitions:

- *Hate speech*: speech that attacks or discriminates against a person or group on the basis of attributes such as race, religion, ethnic origin, national origin, sex, disability, sexual orientation, or gender identity (Brown, 2017; Lepoutre et al., 2023)
- *Cyberbullying*: the use of electronic communication technologies like the internet, social media, and mobile phones to intentionally harass, threaten, humiliate, or target another person or group (Campbell & Bauman, 2018; Wright, 2021)

Finding 2: Platform moderation policies invoke user intent, yet intent is difficult to observe directly.

Intentions are a state of mind. People exercise their autonomy by acting upon their intentions. Most often, we rely on external manifestations, such as what people say or do (Frischmann & Selinger, 2018) to learn another person’s intent. In content moderation, like other contexts, intent is thus inexorably intertwined with actions. When attempting to determine the intent associated with abusive content, one must identify the relevant actor(s) and action(s). Notably, content is a sociotechnical artifact with history and context, and multiple relevant actors and actions may be involved with its arrival in the platform. Major platforms like Twitter (now “X”) and Facebook (now “Meta”) emphasize the importance of intent in content moderation. For example, Twitter’s guidelines state (emphasis added):

Violent entities are those that **deliberately** target humans or essential infrastructure with physical violence and/or violent rhetoric as a means to further their cause.

Hateful entities are those that have **systematically and intentionally** promoted, supported and/or advocated for hateful conduct, which includes promoting violence or engaging in targeted harassment towards a protected category.⁵

Instagram also addresses the complexity of moderating hate speech by considering the intent associated with the act of sharing. The platform allows content that might be deemed hateful if it is shared to challenge or raise awareness about the issues discussed, provided this intent is clearly communicated (emphasis added):

It’s never OK to encourage violence or attack anyone based on their race, ethnicity, national origin, sex, gender, gender identity, sexual orientation, religious affiliation, disabilities, or diseases. When hate speech is being shared to challenge it or to raise awareness, we may allow it. In those instances, we ask that you express your **intent** clearly.⁶

⁵ See: <https://help.Twitter.com/en/rules-and-policies/violent-entities> (accessed May 2024).

⁶ See: <https://help.instagram.com/477434105621119> (accessed May 2024).

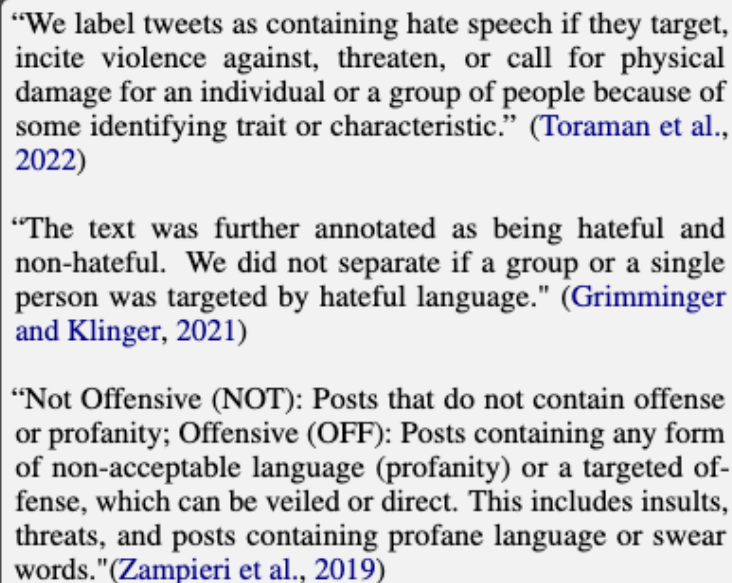
TikTok describes hate speech as intentional as well (emphasis added):

(...) content that **intends to** or does attack, threaten, incite violence against, or dehumanize an individual or group of individuals on the basis of protected attributes like race, religion, gender, gender identity, national origin, and more.⁷

Finding 3: Many publicly available datasets of abusive content lack detailed annotation procedures and do not systematically incorporate intent or platform-specific context.

Our review highlights several key challenges to the alignment between platforms' policies around abuse and the datasets used to train abuse detection algorithms:

- *Ambiguity in definitions of digital abuse:* Compounding the diversity of definitions and taxonomies of abuse proposed in academic work and discussed above, instructions to annotators are often vague. This can result in lack of reusable training data and benchmarks. In Figure 2, we provide excerpts describing annotation processes from surveyed papers. While we cannot determine whether more comprehensive information was provided to annotators due to limited reporting, excerpts alone reveal substantial variability in definitions and instructions.



“We label tweets as containing hate speech if they target, incite violence against, threaten, or call for physical damage for an individual or a group of people because of some identifying trait or characteristic.” (Toraman et al., 2022)

“The text was further annotated as being hateful and non-hateful. We did not separate if a group or a single person was targeted by hateful language.” (Grimminger and Klinger, 2021)

“Not Offensive (NOT): Posts that do not contain offense or profanity; Offensive (OFF): Posts containing any form of non-acceptable language (profanity) or a targeted offense, which can be veiled or direct. This includes insults, threats, and posts containing profane language or swear words.” (Zampieri et al., 2019)

Figure 2. Example of annotation procedure provided in the surveyed papers.

- *Inconsistent consideration of intent during annotation:* Table 1 shows an excerpt from the dataset paper summary (see Appendix D for the full table). Of the dataset papers surveyed, 47.6% (20 of 42) explicitly mentioned intent during annotation, while 35.7% (15 of 42) provided context to annotators to help them better infer intent. In addition, 33.3% (14 of 42) of the papers required annotators to identify the target of abuse. Just three papers (Van Hee et al., 2018; Vidgen, Nguyen, et al., 2021; Ziems et al., 2020) provided contextual information to annotators, explicitly

⁷ See: <https://newsroom.tiktok.com/en-us/countering-hate-on-tiktok> (accessed May 2024).

acknowledged the role of intent in annotation guidelines, and requested annotators to identify the target of the abuse.

- *Cross-platform differences*: Text and annotation instructions are typically provided to annotators outside of the platforms from which they were collected. Labels are considered universal, rather than tailored to the environments and operational settings of specific social media platforms.

Table 1. Example datasets for social media abuse.

Reference	Source	Author-defined scope	Content type	Context provided	Target annotation	Intent mentioned
(Waseem & Hovy, 2016)	Twitter	Hate Speech	Text	No	No	No
(Waseem, 2016)	Twitter	Hate Speech	Text	No	No	No
(Golbeck et al., 2017)	Twitter	Online Harassment	Text	No	No	Yes
(Chatzakou et al., 2017)	Twitter	Cyberbullying	Multimodal	Metadata	No	Yes
(Gao & Huang, 2017)	Fox News	Hate Speech	Text	Conversation	No	No
(Davidson et al., 2017)	Twitter	Hate Speech	Text	No	No	Yes
(Gao et al., 2017)	Twitter	Hate Speech	Text	No	No	No
(Jha & Mamidi, 2017)	Twitter	Sexism	Text	No	No	No
(Van Hee et al., 2018)	ASKfm	Cyberbullying	Text	Conversation	Yes	Yes
(Fersini et al., 2018)	Twitter	Misogyny	Text	No	Yes	Yes

Finding 4: Current detection algorithms often fail to account for contextual and dynamic aspects of hate speech and cyberbullying, limiting their ability to infer intent and generalize across settings.

We prompted GPT-3.5 to classify several statements as abusive or not abusive, both with and without contextual cues (see Figure 3). We used GPT-3.5 because it is a widely available, general-purpose language model that can respond to contextual prompts, allowing us to observe how variations in context and intent may affect classification outcomes. We observe the impact of context and intention on model interpretations of each statement. However, many existing models, like those based on BERT, fall short in capturing context to infer intent. We evaluated four different models (Barbieri et al., 2020; Das et al., 2022; Mathew et al., 2021; Vidgen, Thrush, et al., 2021) using the same examples as those tested with GPT-3.5. The results (see Figure 3) show inconsistent outputs and a significant reliance on word-level cues, failing to account for the underlying user intent that contextualizes meaning.

 1	'Clearly, he is a smart and functioning kid.'	Not Abusive 
	'Clearly, he is a smart and functioning kid.' Explicit Intent: This sentence was sarcastic.	It's Abusive 
 2	'I prefer to live in a neighborhood with people like us.'	Not Abusive 
	'I prefer to live in a neighborhood with people like us.' Topic Context: This sentence talks about segregation.	It's Abusive 
 3	'You are a witch.'	It's Abusive 
	'You are a witch.' Target Identification: This is said to a girl with a witch costume.	Not Abusive 
 4	'Damn! I am so going to kill my roommate!'	It's Abusive 
	'Damn! I am so going to kill my roommate!' Conversational Context: 'Your roommate didn't do the dishes today, hahaha'	Not Abusive 

	1	2	3	4
Mathew et al. 2021	Normal (0.522)	Abusive (0.513)	Abusive (0.534)	Abusive (0.510)
Barbieri et al. 2020	Normal (0.503)	Normal (0.885)	Offensive (0.893)	Offensive (0.763)
Vidgen et al. 2021b	Normal (1.000)	Normal (1.000)	Hate (0.961)	Normal (1.000)
Das et al. 2022	Normal (0.990)	Normal (0.990)	Normal (0.627)	Abusive (0.726)

Figure 3. Example of prompts and detection model performance to showcase the importance of context and intent in understanding online abuse.

Our context analysis of the papers on online abuse detection algorithms confirmed that while detection models have made notable advances in identifying abusive content through analysis of textual data, they often fail to consider the complex nature of social media interactions, which include aspects like user status, social network structures, and offline context. Figure 4(A) shows the co-occurrence network of ten features extracted and used to capture context and infer intent. The size of each node reflects how many papers included the corresponding feature, while the width of each edge represents how often two features appeared together in the same paper. We observe that user metadata are the most frequently considered features, with 16 (9.5%) algorithms using them for detection. Similarly, linguistic cues are used in 14 (8.3%) studies and post metadata in 10 (5.9%). In contrast, psychological and cognitive dimensions, and policy or rule-aware models are less frequently explored, with just four (2.2%) and two (1.1%) studies, respectively. A majority of reviewed papers, 58.9% (99 of 168), focus on comparing model performance and benchmarking against various metrics. Figure 4(B) shows a histogram of papers according to the number of intent-related features considered in a paper. Among 168 surveyed papers, three of them present models incorporating six out of the ten features listed above, meaningfully including context and potentially addressing the issue of intent (Dhingra et al., 2023; López-Vizcaíno et al., 2021; Ziems et al., 2020). Building upon Ziems et al. (2020), a subsequent paper utilized the dataset from the earlier study and applied methodological improvements (Dhingra et al., 2023).

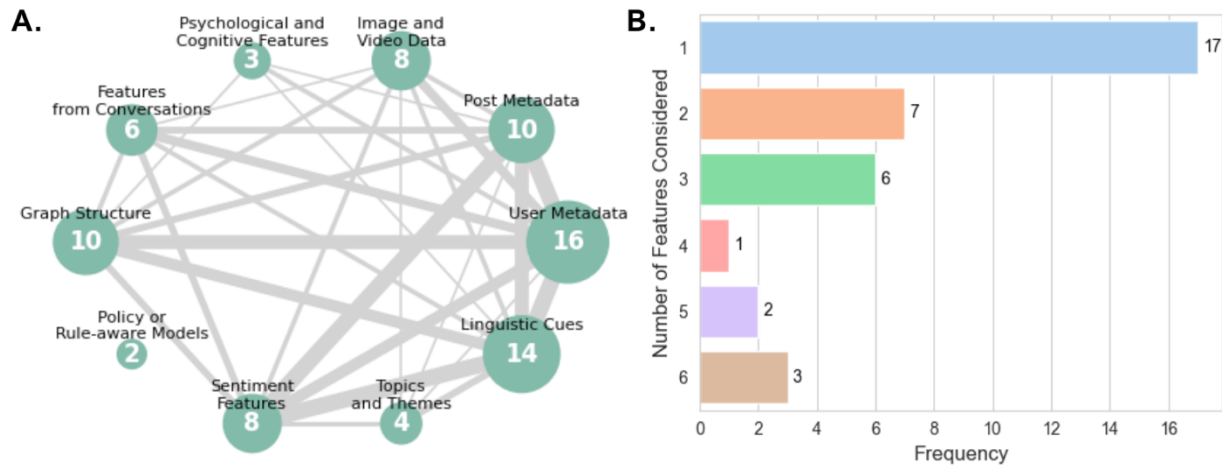


Figure 4. (A) Co-occurrence network of features considered by detection models, (B) Distribution of models/papers considering contextual features, based on the number of features considered (from 1[min] to 6[max]).

Based on these observations and our review of the literature, we summarized the current limitations of detection algorithms for online abuse.

- *Insufficient context for inferring intent:* Because intent is rarely explicit in language, detecting it often requires additional context beyond the text itself. Traditional text-based models, though robust in their linguistic analyses, fall short in understanding the complexities of group dynamics in the spread of online abuse (Salminen et al., 2018). For instance, an abusive narrative might emerge and propagate not merely because of its textual content but due to the influence and endorsement of a closely-knit group within the network (Marwick & Lewis, 2017). These group-based interactions are influenced by, for example, shared ideologies, mutual affiliations, or even orchestrated campaigns, which sometimes employ subtle linguistic cues not easily detectable by conventional text-based models. The complexity is further compounded when they employ tactics like code-switching, euphemisms, or meme-based communication, thereby effectively circumventing text-based detection mechanisms.
- *Static perspectives:* Social norms are dynamic and influence users' interpretation of what constitutes abusive content (Crandall et al., 2002). Traditional detection models often fail to account for these longitudinal shifts. For example, a phrase may take on a new meaning in the digital realm, and the rapid evolution of internet slang also necessitates a more flexible approach for continuous training and updating.
- *Poor generalizability:* Model performance can fluctuate significantly even when utilizing the same dataset, with optimal outcomes frequently exclusive to that specific dataset (Leo et al., 2023). These inconsistencies highlight the need for robust testing to understand the generalizability of existing methods in new contexts.

Methods

Online abuse datasets

We reviewed the documentation provided in the associated papers of existing online abuse detection datasets, focusing on any available descriptions of data sources, categories, modality, and annotation

procedures. We categorize each based on the extent to which it considers the context and intent of abusive expressions during annotation. We reviewed papers available between 2016 and 2024 using Scopus search along with datasets referenced in Vidgen & Derczynski (2020)⁸ and citation search. The dataset selection and review process involved two researchers. One researcher drafted the initial table, and the second independently reviewed the associated papers in detail to identify any discrepancies. No inconsistencies were observed during this process. Further details, search terms, inclusion criteria, and a PRISMA diagram for the screening pipeline are provided in Appendix C.

Appendix D outlines papers reviewed, noting whether any contextual information was provided to annotators, whether intent was explicitly mentioned in annotation instructions, and whether the annotation task included identification of a target person or group.

- Intent mentioned: We capture whether annotation guidelines mention intent, intention, etc., anywhere in instructions or definitions.
- Context provided: When provided, context takes one of two forms: 1) conversations (annotators are provided with conversation/text surrounding the text being annotated, enabling them to infer intent by understanding the broader exchange) and 2) metadata (annotators are provided with user-level metadata, for example, profile content, geographical location, or post-level metadata, for example, images or text extracted from images, which may offer additional clues about intent).
- Target annotation: Annotation guidelines request that the person or group targeted by a comment be identified during annotation.

Online abuse detection algorithms

We reviewed the papers that described and examined abuse detection models, categorized features utilized by these models, and identified the gaps that exist in effectively detecting abuse. We queried SCOPUS for papers presenting abuse detection algorithms published between 2013 and early 2024. Similar to our survey of datasets, specific search terms, inclusion criteria, and a PRISMA diagram for the screening pipeline are provided in Appendix C. The review and categorization were also conducted by two researchers who cross-checked decisions and engaged in discussion to ensure consistency. No inconsistencies were reported during this process.

We categorized the features used by detection models in our survey that go beyond traditional text-based methods such as bag-of-words, TF-IDF, and embeddings. These features fall into the following ten categories: user metadata, post metadata, image and video content, psychological and cognitive signals, conversational context, graph structure, policy- or rule-aware signals, sentiment, topical or thematic information, and linguistic cues. Appendix E provides a detailed breakdown and examples for each category, and Appendix F includes a full mapping of these features to the corresponding papers in our survey.

Bibliography

Al Mazari, A. (2013). Cyber-bullying taxonomies: Definition, forms, consequences and mitigation strategies. In *2013 5th International Conference on Computer Science and Information Technology* (pp. 126–133). IEEE. <https://doi.org/10.1109/CSIT.2013.6588770>

⁸ See: <https://hatespeechdata.com/> (accessed May 2024).

- Albanyan, A., & Blanco, E. (2022). Pinpointing fine-grained relationships between hateful tweets and replies. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10), 10418–10426. <https://doi.org/10.1609/aaai.v36i10.21284>
- Albanyan, A., Hassan, A., & Blanco, E. (2023). Not all counterhate tweets elicit the same replies: A fine-grained analysis. In A. Palmer & J. Camacho-Collados (Eds.), *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)* (pp. 71–88). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.starsem-1.8>
- Ali, S., Blackburn, J., & Stringhini, G. (2025). *Evolving hate speech online: An adaptive framework for detection and mitigation*. arXiv. <https://doi.org/10.48550/arXiv.2502.10921>
- Allen, J., Arechar, A. A., Pennycook, G., & Rand, D. G. (2021). Scaling up fact-checking using the wisdom of crowds. *Science Advances*, 7(36), eabf4393. <https://doi.org/10.1126/sciadv.abf4393>
- Anuchitanukul, A., & Ive, J. (2022). SURF: Semantic-level unsupervised reward function for machine translation. In M. Carpuat, M.-C. de Marneffe, & I. V. Meza Ruiz (Eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4508–4522). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.naacl-main.334>
- Arora, A., Nakov, P., Hardalov, M., Sarwar, S. M., Nayak, V., Dinkov, Y., Zlatkova, D., Dent, K., Bhatawdekar, A., Bouchard, G., & Augenstein, I. (2023). Detecting harmful content on online platforms: What platforms need vs. Where research efforts go. *ACM Computing Surveys*, 56(3), 1–17. <https://doi.org/10.1145/3603399>
- Barbieri, F., Camacho-Collados, J., Espinosa Anke, L., & Neves, L. (2020). TweetEval: Unified benchmark and comparative evaluation for tweet classification. In T. Cohn, Y. He, & Y. Liu (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 1644–1650). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.findings-emnlp.148>
- Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F. M., Rosso, P., & Sanguinetti, M. (2019). SemEval-2019 Task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki, & S. M. Mohammad (Eds.), *Proceedings of the 13th International Workshop on Semantic Evaluation* (pp. 54–63). Association for Computational Linguistics. <https://doi.org/10.18653/v1/S19-2007>
- Basile, V., Caselli, T., Balahur, A., & Ku, L.-W. (2022). Bias, subjectivity and perspectives in natural language processing. *Frontiers in Artificial Intelligence*, 5, 926435. <https://doi.org/10.3389/frai.2022.926435>
- Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In D. Jurafsky, J. Chai, N. Schuster, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.463>
- Brown, A. (2017). What is hate speech? Part 1: The myth of hate. *Law and Philosophy*, 36, 419–468. <https://doi.org/10.1007/s10982-017-9297-1>
- Campbell, M., & Bauman, S. (2018). Cyberbullying: Definition, consequences, prevalence. In *Reducing cyberbullying in schools* (pp. 3–16). Elsevier. <https://doi.org/10.1016/B978-0-12-811423-0.00001-8>
- Caselli, T., Basile, V., Mitrović, J., Kartoziya, I., & Granitzer, M. (2020). I feel offended, don't be abusive! Implicit/explicit messages in offensive and abusive language. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 6193–6202). European Languages Resources Association. <https://aclanthology.org/2020.lrec-1.760/>

- Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. SAGE.
- Clayton, K., Blair, S., Busam, J. A., Forstner, S., Glance, J., Green, G., Kawata, A., Kovvuri, A., Martin, J., Morgan, E., Sandhu, M., Sang, R., Scholz-Bright, R., Welch, A. T., Wolff, A. G., Zhou, A., & Nyhan, B. (2020). Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior*, 42(4), 1073–1095. <https://doi.org/10.1007/s11109-019-09533-0>
- Crandall, C. S., Eshleman, A., & O'Brien, L. (2002). Social norms and the expression and suppression of prejudice: The struggle for internalization. *Journal of Personality and Social Psychology*, 82(3), 359–78. <https://pubmed.ncbi.nlm.nih.gov/11902622/>
- Crook, B., Schlüter, M., & Speith, T. (2023). *Revisiting the performance-explainability trade-off in explainable artificial intelligence (XAI)*. arXiv. <https://doi.org/10.48550/arXiv.2307.14239>
- Das, M., Banerjee, S., & Mukherjee, A. (2022). *Data bootstrapping approaches to improve low resource abusive language detection for Indic languages*. arXiv. <https://doi.org/10.48550/arXiv.2204.12543>
- Dhingra, N., Chawla, S., Saini, O., & Kaushal, R. (2023). An improved detection of cyberbullying on social media using randomized sampling. *International Journal of Bullying Prevention*, 1–13. <https://doi.org/10.1007/s42380-023-00188-4>
- Egelman, S., Cranor, L. F., & Hong, J. (2008). You've been warned: An empirical study of the effectiveness of web browser phishing warnings. In *CHI '08: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1065–1074). Association for Computing Machinery. <https://doi.org/10.1145/1357054.1357219>
- Fazio, L. (2020). Pausing to consider why a headline is true or false can help reduce the sharing of false news. *Harvard Kennedy School (HKS) Misinformation Review*, 1(2). <https://doi.org/10.37016/mr-2020-009>
- Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4), 1–30. <https://doi.org/10.1145/3232676>
- French, A., Storey, V. C., & Wallace, L. (2023). A typology of disinformation intentionality and impact. *Information Systems Journal*, 34(4), 1324–1354. <https://doi.org/10.1111/isj.12495>
- Frischmann, B., & Benesch, S. (2023). Friction-in-design regulation as 21st century time, place, and manner restriction. *Yale Journal of Law and Technology*, 25(1), 376–447. <https://yjolt.org/friction-design-regulation-21st-century-time-place-and-manner-restriction>
- Frischmann, B. M., & Vardi, M. Y. (2024). *Better digital contracts with prosocial friction-in-design*. SSRN. <http://dx.doi.org/10.2139/ssrn.4918003>
- Frischmann, B., & Sanfilippo, M. (2023). Slow-governance in smart cities: An empirical study of smart intersection implementation in four U. S. college towns. *Internet Policy Review*, 12(1), 1–31. <https://doi.org/10.14763/2023.1.1703>
- Frischmann, B., & Selinger, E. (2018). *Re-engineering humanity*. Cambridge University Press. <https://doi.org/10.1017/9781316544846>
- Gadiraju, U., Kawase, R., Dietze, S., & Demartini, G. (2015). Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In *CHI '15: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 1631–1640). Association for Computing Machinery. <https://doi.org/10.1145/2702123.2702443>
- Gao, L., & Huang, R. (2017). Detecting online hate speech using context aware models. In R. Mitkov & G. Angelova (Eds.), *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017* (pp. 260–266). INCOMA Ltd. https://doi.org/10.26615/978-954-452-049-6_036

- Gashroo, O. B., & Mehrotra, M. (2022). Analysis and classification of abusive textual content detection in online social media. In G. Rajakumar, K. Du, C. Vuppalapati, & G. N. Beligiannis (Eds.), *Intelligent Communication Technologies and Virtual Mobile Networks: Proceedings of ICICV 2022* (pp. 173–190). Springer. https://doi.org/10.1007/978-981-19-1844-5_15
- Ghosh, S., Venkit, P. N., Gautam, S., Wilson, S., & Caliskan, A. (2024). Do generative ai models output harm while representing non-western cultures: Evidence from a community-centered approach. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 476–489). AAAI Press. <https://dl.acm.org/doi/10.5555/3716662.3716702>
- Grimminger, L., & Klinger, R. (2021). Hate towards the political opponent: A Twitter corpus study of the 2020 us elections on the basis of offensive speech and stance detection. In O. De Clercq, A. Balahur, J. Sedoc, V. Barriere, S. Tafreshi, S. Buechel, & V. Hoste (Eds.), *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (pp. 171–180). Association for Computational Linguistics. <https://aclanthology.org/2021.wassa-1.18>
- Hanu, L. & Unitary team. (2020). *Detoxify*. Github. <https://github.com/unitaryai/detoxify>
- Hashemi, M. (2021). A data-driven framework for coding the intent and extent of political tweeting, disinformation, and extremism. *Information*, 12(4), 148. <https://doi.org/10.3390/info12040148>
- Im, J., Schoenebeck, S., Iriarte, M., Grill, G., Wilkinson, D., Batool, A., Alharbi, R., Funwie, A., Gankhuu, T., Gilbert, E., & Naseem, M. (2022). Women’s perspectives on harm and justice after online harassment. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2), 1–23. <https://doi.org/10.1145/3555775>
- Islam, M. R., Bataineh, A. S., & Zulkernine, M. (2023). Detection of cyberbullying in social media texts using explainable artificial intelligence. In G. Wang, H. Wang, G. Min, N. Georgalas, & W. Meng (Eds.), *Ubiquitous Security: UbiSec 2023* (pp. 319–334). Springer. https://doi.org/10.1007/978-981-97-1274-8_21
- Jha, A., & Mamidi, R. (2017). When does a compliment become sexist? Analysis and classification of ambivalent sexism using Twitter data. In D. Hovy, S. Volkova, D. Bamman, D. Jurgens, B. O’Connor, O. Tsur, & A. S. Doğruöz (Eds.), *Proceedings of the Second Workshop on NLP and Computational Social Science* (pp. 7–16). Association for Computational Linguistics. <https://doi.org/10.18653/v1/W17-2902>
- Jhaver, S., Birman, I., Gilbert, E., & Bruckman, A. (2019). Human-machine collaboration for content regulation: The case of reddit automoderator. *ACM Transactions on Computer-Human Interaction*, 26(5), 1–35. <https://doi.org/10.1145/3338243>
- Joseph, K., Friedland, L., Hobbs, W., Lazer, D., & Tsur, O. (2017). ConStance: Modeling annotation contexts to improve stance classification. In M. Palmer, R. Hwa, & S. Riedel (Eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 1115–1124). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D17-1116>
- Kaiser, B., Wei, J., Lucherini, E., Lee, K., Matias, J. N., & Mayer, J. (2021). Adapting security warnings to counter online disinformation. In *30th USENIX Security Symposium (USENIX Security 21)* (pp. 1163–1180). <https://www.usenix.org/conference/usenixsecurity21/presentation/kaiser>
- Keipi, T., Näsi, M., Oksanen, A., & Räsänen, P. (2016). *Online hate and harmful content: Cross-national perspectives*. Taylor & Francis.
- Kohli, A., & Devi, V. S. (2023). Explainable offensive language classifier. In B. Luo, L. Cheng, Z.-G. Wu, H. Li, & C. Li (Eds.), *International Conference on Neural Information Processing* (pp. 299–313). Springer. http://dx.doi.org/10.1007/978-981-99-8132-8_23
- Kruger, A., Saletta, M., Ahmad, A., & Howe, P. (2024). Structured expert elicitation on disinformation, misinformation, and malign influence: Barriers, strategies, and opportunities. *Harvard Kennedy School (HKS) Misinformation Review*, 5(7). <https://doi.org/10.37016/mr-2020-169>

- Kumar, D., AbuHashem, Y., & Durumeric, Z. (2023). *Watch your language: Investigating content moderation with large language models*. arXiv. <https://doi.org/10.48550/arXiv.2309.14517>
- Leo, C. O., Santoso, B. J., & Pratomo, B. A. (2023). Enhancing hate speech detection for social media moderation: A comparative analysis of machine learning algorithms. In *2023 International Conference on Advanced Mechatronics, Intelligent Manufacture and Industrial Automation (ICAMIMIA)* (pp. 960–964). IEEE. <https://doi.org/10.1109/ICAMIMIA60881.2023.10427779>
- Lepoutre, M., Vilar-Lluch, S., Borg, E., & Hansen, N. (2023). What is hate speech? The case for a corpus approach. *Criminal Law and Philosophy*, 1–34. <https://doi.org/10.1007/s11572-023-09675-7>
- Lewandowska-Tomaszczyk, B., Bączkowska, A., Liebeskind, C., Valunaite Oleskeviciene, G., & Žitnik, S. (2023). An integrated explicit and implicit offensive language taxonomy. *Lodz Papers in Pragmatics*, 19(1), 7–48. <https://doi.org/10.1515/lpp-2023-0002>
- Li, G., Lu, W., Zhang, W., Lian, D., Lu, K., Mao, R., Shu, K., & Liao, H. (2024). *Re-search for the truth: Multi-round retrieval-augmented large language models are strong fake news detectors*. arXiv. <https://doi.org/10.48550/arXiv.2403.09747>
- López-Vizcaíno, M. F., Nóvoa, F. J., Carneiro, V., & CACHED, F. (2021). Early detection of cyberbullying on social media networks. *Future Generation Computer Systems*, 118, 219–229. <https://doi.org/10.1016/j.future.2021.01.00>
- Lynn, V., Giorgi, S., Balasubramanian, N., & Schwartz, H. A. (2019). Tweet classification without the tweet: An empirical examination of user versus document attributes. In S. Volkova, D. Jurgens, D. Hovy, D. Bamman, & O. Tsur (Eds.), *Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science* (pp. 18–28). Association for Computational Linguistics. <https://doi.org/10.18653/v1/W19-2103>
- MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PLoS ONE*, 14(8). <https://doi.org/10.1371/journal.pone.0221152>
- Mahadevan, A., & Mathioudakis, M. (2024). Cost-aware retraining for machine learning. *Knowledge-Based Systems*, 293, 111610. <https://doi.org/10.1016/j.knosys.2024.111610>
- Markov, I., & Daelemans, W. (2022). The role of context in detecting the target of hate speech. In R. Kumar, A. Kr. Ojha, M. Zampieri, S. Malmasi, & D. Kadar (Eds.), *Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022)* (pp. 37–42). Association for Computational Linguistics. <https://aclanthology.org/2022.trac-1.5>
- Marwick, A., & Lewis, R. (2017). *Media manipulation and disinformation online*. Data & Society. <https://datasociety.net/library/media-manipulation-and-disinfo-online/>
- Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P., & Mukherjee, A. (2021). Hatexplain: A benchmark dataset for explainable hate speech detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17), 14867–14875. <https://cdn.aaai.org/ojs/17745/17745-13-21239-1-2-20210518.pdf>
- Menini, S., Aproso, A. P., & Tonelli, S. (2021). *Abuse is contextual, what about nlp? The role of context in abusive language annotation and detection*. arXiv. <https://doi.org/10.48550/arXiv.2103.14916>
- Mirza, S., Begum, L., Niu, L., Pardo, S., Abouzied, A., Papotti, P., & Pöpper, C. (2023). *Tactics, threats & targets: Modeling disinformation and its mitigation* [Paper presentation]. 2023 Network and Distributed System Security Symposium. San Diego, CA, USA. <https://doi.org/10.14722/ndss.2023.23657>
- Mladenović, M., Ošmjanski, V., & Stanković, S. V. (2021). Cyber-aggression, cyberbullying, and cyber-grooming: A survey and research challenges. *ACM Computing Surveys (CSUR)*, 54(1), 1–42. <https://doi.org/10.1145/3424246>

- Molina, M. D., Sundar, S. S., Le, T., & Lee, D. (2021). "Fake news" is not simply false information: A concept explication and taxonomy of online content. *American Behavioral Scientist*, 65(2), 180–212. <https://doi.org/10.1177/0002764219878224>
- Muhammadiyah, M., Rahman, R., & Wei, S. (2025). Interpretation of deep learning models in natural language processing for misinformation detection with the explainable AI (XAI) approach. *Journal of Computer Science Advancements*, 3(2), 2. <https://doi.org/10.70177/jsca.v3i2.2104>
- Pröllochs, N. (2022). Community-based fact-checking on Twitter's birdwatch platform. *Proceedings of the International AAAI Conference on Web and Social Media*, 16, 794–805. <https://doi.org/10.1609/icwsm.v16i1.19335>
- Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2017). *Measuring the reliability of hate speech annotations: The case of the European refugee crisis*. arXiv. <https://doi.org/10.17185/dupublico/42132>
- Salminen, J., Almerexhi, H., Milenković, M., Jung, S., An, J., Kwak, H., & Jansen, B. (2018). Anatomy of online hate: Developing a taxonomy and machine learning models for identifying and classifying hate in online news media. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1). <https://doi.org/10.1609/icwsm.v12i1.15028>
- Scheuerman, M. K., Jiang, J. A., Fiesler, C., & Brubaker, J. R. (2021). A framework of severity for harmful content online. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2), 1–33. <https://doi.org/10.1145/3479512>
- Shi, K., Sun, X., Li, Q., & Xu, G. (2024). *Compressing long context for enhancing RAG with AMR-based concept distillation*. arXiv. <https://doi.org/10.48550/arXiv.2405.03085>
- Toraman, C., Şahinuç, F., & Yilmaz, E. (2022). Large-scale hate speech detection with cross-domain transfer. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 2215–2225). European Language Resources Association. <https://aclanthology.org/2022.lrec-1.238>
- Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., De Pauw, G., Daelemans, W., & Hoste, V. (2018). Automatic detection of cyberbullying in social media text. *PloS One*, 13(10), e0203794. <https://doi.org/10.1371/journal.pone.0203794>
- Vidgen, B., & Derczynski, L. (2020). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PloS One*, 15(12), e0243300. <https://doi.org/10.1371/journal.pone.0243300>
- Vidgen, B., Harris, A., Nguyen, D., Tromble, R., Hale, S., & Margetts, H. (2019). Challenges and frontiers in abusive content detection. In S. T. Roberts, J. Tetreault, V. Prabhakaran, & Z. Waseem (Eds.), *Proceedings of the Third Workshop on Abusive Language Online* (pp. 80–93). Association for Computational Linguistics. <https://doi.org/10.18653/v1/W19-3509>
- Vidgen, B., Nguyen, D., Margetts, H., Rossini, P., & Tromble, R. (2021). Introducing CAD: The contextual abuse dataset. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, & Y. Zhou (Eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 2289–2303). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.naacl-main.182>
- Vidgen, B., Thrush, T., Waseem, Z., & Kiela, D. (2021). Learning from the worst: Dynamically generated datasets to improve online hate detection. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (Vol. 1, pp. 1667–1682). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.acl-long.132>

- Wang, K., Lu, D., Han, C., Long, S., & Poon, J. (2020). Detect all abuse! Toward universal abusive language detection models. In D. Scott, N. Bel, & C. Zong (Eds.), *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 6366–6376). International Committee on Computational Linguistics. <https://doi.org/10.18653/v1/2020.coling-main.560>
- Waseem, Z., Davidson, T., Warmusley, D., & Weber, I. (2017). Understanding abuse: A typology of abusive language detection subtasks. In Z. Waseem, W. H. K. Chung, D. Hovy, & J. Tetreault (Eds.), *Proceedings of the First Workshop on Abusive Language Online* (pp. 78–84). Association for Computational Linguistics. <https://doi.org/10.18653/v1/W17-3012>
- Wiegand, M., Ruppenhofer, J., & Kleinbauer, T. (2019). Detection of abusive language: The problem of biased datasets. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol 1, pp. 602–608). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1060>
- Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). Predicting the type and target of offensive posts in social media. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol 1, pp. 1415–1420). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1144>
- Zhang, W., Guo, H., Kivlichan, I. D., Prabhakaran, V., Yadav, D., & Yadav, A. (2023). *A taxonomy of rater disagreements: Surveying challenges & opportunities from the perspective of annotating online toxicity*. arXiv. <https://doi.org/10.48550/arXiv.2311.04345>
- Ziems, C., Vigfusson, Y., & Morstatter, F. (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. <https://doi.org/10.1609/icwsm.v14i1.7345>

Authorship

Sai Koneru and Pranav Narayanan Venkit contributed equally to this work.

Funding

This work was partially supported by NSF award #2318460.

Competing interests

The authors declare no competing interests.

Ethics

This study does not involve human or animal subjects.

Copyright

This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided that the original author and source are properly credited.

Data availability

All materials needed to replicate this study are available via the Harvard Dataverse:

<https://doi.org/10.7910/DVN/PQ1EH3>

Appendix A: Datasheet codebook for dataset annotation

Regarding information provided to annotators:

- What are the definitions and scope of online abuse presented to the annotator?
- What underlying taxonomy is provided, and how should it be applied (e.g., modified, integrated) during annotation?
- What contextual information is provided to annotators to assist in the annotation process?
- What is the platform's content moderation policy, and does the annotation rubric align with this policy?
- What is the platform's moderation policy for abusive content and does the annotation rubric adhere to the policy?

Regarding information solicited from annotators:

- Is abusive or offensive language present?
- Is there identifiable intent behind the dissemination of the content, if there is abusive language present?
- Who are the initiators and the targets?

Regarding assessment and reporting:

- When was the data collected and when was it labeled?
- Who are the annotators (demographics, etc.)?
- What is the agreement score amongst annotators?
- What are the data points with low agreement? What are potential reasons for disagreement?

Appendix B: Summary of existing taxonomies

Table B2. Summary of existing taxonomies of digital abuse on social media.

Reference	Author-defined scope	Categories
(Nocentini et al., 2010)	Cyberbullying	written-verbal behavior, visual behavior, exclusion, impersonation
(Al Mazari, 2013)	Cyberbullying	child cyberbullying, cybergrooming, adults cyberstalking, workplace cyber-bullying
(Agrafiotis et al., 2016)	Cyber harm	physical, psychological, economic, reputational, cultural, political
(Miró-Llinares & Rodríguez-Sala, 2016)	Hate speech & violent communication	violent incitement, personal offence, discrimination incitement, collective offence
(Waseem et al., 2017)	Online abuse	two-fold typology: directed towards a specific individual or entity & used towards a generalized other; explicit & implicit
(Salminen et al., 2018)	Online hate	targets: financial power (corporation, wealthy), political issues (terrorism, politics, ideology), racism & xenophobia (anti-white, anti-black, xenophobia), religion (anti-Islam, antisemitic), specific nation(s), specific person, media (towards media company, other), armed forces (police, military), behavior (humanity, other); language: accusations, humiliation, swearing, promoting violence
(Anzovino et al., 2018)	Misogyny	discredit, stereotype and objectification, sexual harassment and threats of violence, dominance, derailing
(Vidgen et al., 2019)	Online abuse	individual-directed abuse, identity-directed abuse, concept-directed abuse
(Vidgen & Derczynski, 2020)	Online abuse	person-directed abuse, group-directed abuse, flagged content, uncivil content, mixed
(Banko et al., 2020)	Online harm	hate and harassment (doxxing, identity attack, identity misrepresentation, insult, sexual aggression, threat of violence), self-inflicted harm (eating disorder promotion, self-harm), ideological harm (extremism, terrorism & organized crime, misinformation), exploitation (adult sexual services, scams, child sexual abuse material)
(Vidgen, Nguyen, et al., 2021)	Online abuse	identity-directed abuse (derogation, animosity, threatening, glorification, dehumanization), affiliation-directed abuse (derogation, animosity, threatening, glorification, dehumanization), person-directed abuse (abuse to them, abuse about them)

Reference	Author-defined scope	Categories
(Sajadi Ansari et al., 2021)	Cyberbullying	flaming, harassment, sexual, threat, trickery
(Mladenović et al., 2021)	Objectionable content (cyberbullying)	fourfold typology: expression (explicit, implicit), targeting (targeted, untargeted), orientation (directed, generalized), frequency (repeated, unrepeated)
(Alrashidi et al., 2022)	Abusive content	abusive and offensive language, hate speech, cyberbullying, targeted groups (religious and racism, gender and misogyny)
(Gashroo & Mehrotra, 2022)	Online abuse	abusive language, aggression, cyberbullying, insults, personal attacks, provocation, racism, sexism, toxicity
(Lewandowska-Tomaszczyk et al., 2023)	Offensive language	taboo (obscene, profane), insulting (abusive), hate speech (slur), harassment (cyberbullying), toxic
(Kogilavani et al., 2023)	Offensive language	aggression, cyberbullying, hate speech, offensive language, toxic comments

Appendix C: Inclusion criteria and PRISMA diagram

Query for dataset papers

KEY ("social media" AND "dataset" AND ("NLP" OR "Natural Language Processing") AND ("hate speech" OR "abus*" OR "offens*" OR "cyberbully*")) OR TITLE ("social media" AND "dataset" AND ("hate speech" OR "abus*" OR "offens*" OR "cyberbully*")) AND (LIMIT-TO (LANGUAGE, "English"))

Inclusion criteria for dataset papers

We applied the following inclusion criteria:

- 1) The paper presents a novel dataset for which annotation procedures are described.
- 2) The dataset is intended for training and testing algorithm(s) aimed at abuse detection.
- 3) The dataset is curated from one or more widely used social media platforms.
- 4) The dataset is in English.⁹
- 5) The dataset includes textual content.

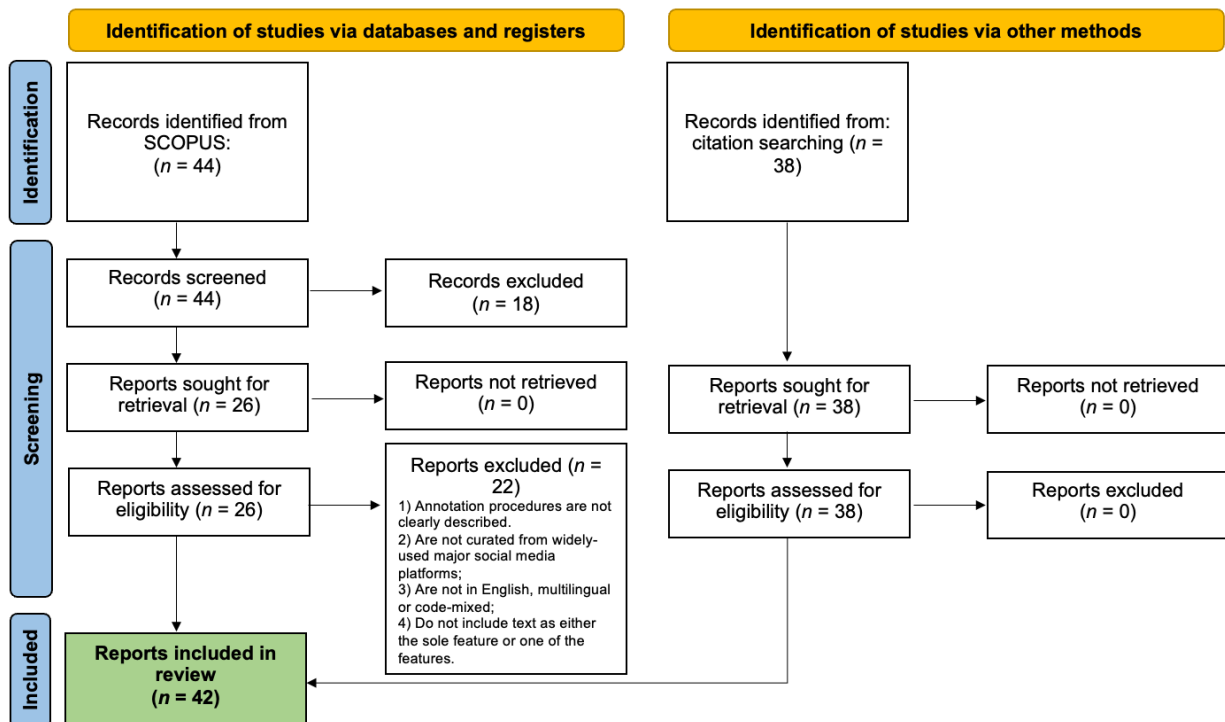


Figure C1. PRISMA diagram for the selection of papers presenting labeled datasets for online abuse.

⁹ Focusing on monolingual settings allows us to address these issues directly before extending analyses to multiple languages, where cultural variations and linguistic nuances further complicate intent inference.

Query for algorithm papers

KEY (“social media” AND (“NLP” OR “Natural Language Processing”) AND “de-taction” AND (“hate speech” OR “abus*” OR “offens*” OR “cyberbully*”)) OR TITLE (“social media” AND “detection” AND (“hate speech” OR “abus*” OR “offens*” OR “cyberbully*”)) AND (LIMIT-TO (LANGUAGE, “English”))

Inclusion criteria for algorithm papers

Similar to our survey of datasets, we removed papers that were not accessible, not written in English, or which did not describe an algorithm for detection of online abuse. We removed models designed for multilingual or non-English tasks.

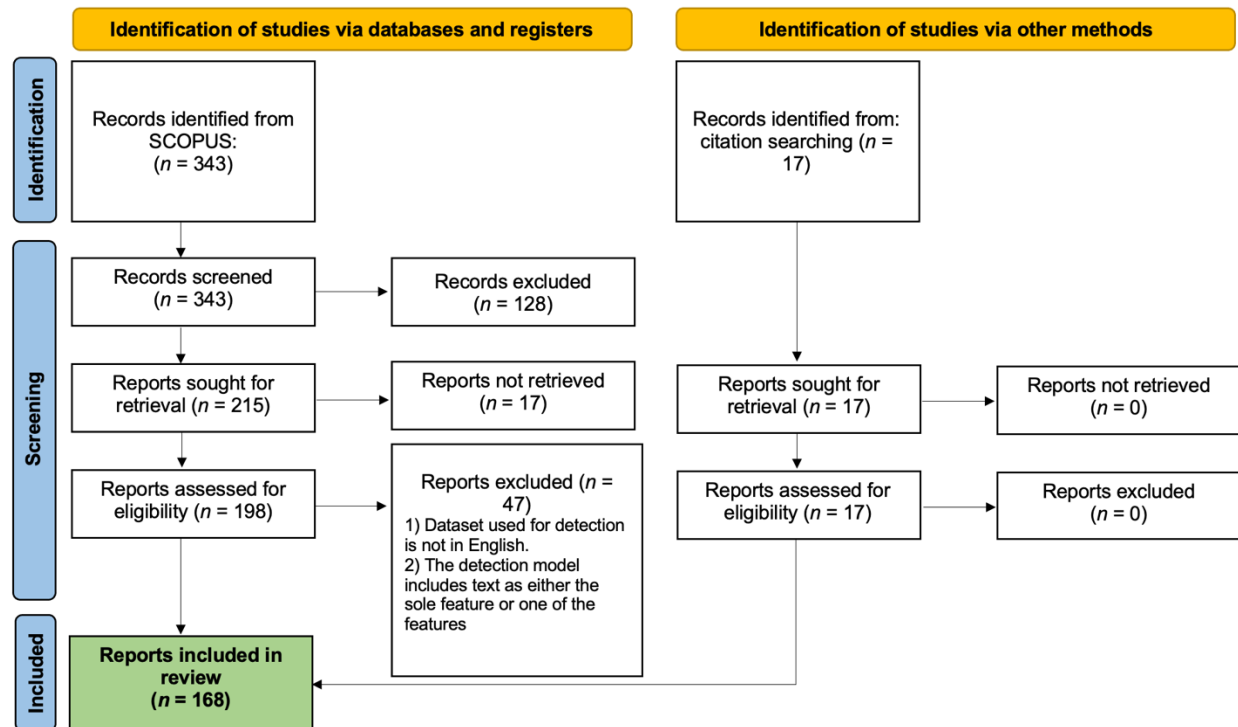


Figure C2. PRISMA diagram for the selection of papers presenting detection algorithms for online abuse.

Appendix D: Summary of datasets

Table D3. Summary of datasets for social media abuse.

Reference	Source	Author-defined scope	Content type	Context provided	Target annotation	Intent mentioned
(Waseem & Hovy, 2016)	Twitter	Hate speech	Text	No	No	No
(Waseem, 2016)	Twitter	Hate speech	Text	No	No	No
(Golbeck et al., 2017)	Twitter	Online harassment	Text	No	No	Yes
(Chatzakou et al., 2017)	Twitter	Cyberbullying	Multimodal	Metadata	No	Yes
(Gao & Huang, 2017)	Fox News	Hate speech	Text	Conversation	No	No
(Davidson et al., 2017)	Twitter	Hate speech	Text	No	No	Yes
(Gao et al., 2017)	Twitter	Hate speech	Text	No	No	No
(Jha & Mamidi, 2017)	Twitter	Sexism	Text	No	No	No
(Van Hee et al., 2018)	ASKfm	Cyberbullying	Text	Conversation	Yes	Yes
(Fersini et al., 2018)	Twitter	Misogyny	Text	No	Yes	Yes
(Ribeiro et al., 2018)	Twitter	Hate speech	Multimodal	Metadata	No	No
(ElSherief et al., 2018)	Twitter	Hate speech	Text	No	Yes	No
(Founta et al., 2018)	Twitter	Abusive behavior	Multimodal	No	No	Yes
(Rezvan et al., 2018)	Twitter	Online harassment	Text	No	No	Yes
(Salminen et al., 2018)	YouTube, Facebook	Online Hate	Text	No	Yes	Yes
(M. Zampieri et al., 2019)	Twitter	Offensive language	Text	No	Yes	No
(J. Qian et al., 2019)	Reddit, Gab	Hate speech	Text	Conversation	No	No
(Ousidhoum et al., 2019)	Twitter	Hate speech	Text	No	Yes	No
(Basile et al., 2019)	Twitter	Hate speech	Text	No	Yes	No
(Mandl et al., 2019)	Twitter	Hate Speech, Offensive Content	Text	No	Yes	No

Reference	Source	Author-defined scope	Content type	Context provided	Target annotation	Intent mentioned
(Wijesiriwardene et al., 2020)	Twitter	Toxicity	Multimodal	Conversation	No	Yes
(Gomez et al., 2020)	Twitter	Hate speech	Multimodal	Metadata	No	No
(Vidgen et al., 2020)	Twitter	Hate speech	Text	No	No	Yes
(Caselli et al., 2020)	Twitter	Abusive language	Text	No	No	Yes
(Ziems et al., 2020)	Twitter	Cyberbullying	Text	Conversation	Yes	Yes
(C. J. Kennedy et al., 2020)	Twitter, Reddit, YouTube	Hate speech	Text	No	Yes	Yes
(Aggarwal et al., 2020)	Twitter, Reddit, Formspring	Cyberbullying	Text	No	No	No
(Suryawanshi et al., 2020)	Kaggle, Reddit, Facebook, Twitter, Instagram	Offensive language	Multimodal	Metadata	No	Yes
(Van Bruwaene et al., 2020)	Facebook, Instagram, Twitter, Pinterest, Tumblr, YouTube	Cyberbullying	Text	Metadata	No	Yes
(Grimminger & Klinger, 2021)	Twitter	Hate speech	Text	No	No	No
(Qureshi & Sabih, 2021)	Twitter	Hate speech	Text	No	No	No
(Salawu et al., 2021)	Twitter	Cyberbullying	Text	No	No	Yes
(Samory et al., 2021)	Twitter	Sexism	Text	No	No	No
(He et al., 2021)	Twitter	Hate speech	Multimodal	No	No	Yes
(Vidgen, Nguyen, et al., 2021)	Reddit	Abusive language	Text	Conversation	Yes	Yes
(Ashraf et al., 2021)	YouTube	Abusive language	Text	Conversation	No	Yes
(Mathew et al., 2021)	Twitter, Gab	Hate speech	Text	No	Yes	No
(Albanyan & Blanco, 2022)	Twitter	Hate speech	Text	Conversation	No	No

Reference	Source	Author-defined scope	Content type	Context provided	Target annotation	Intent mentioned
(Toraman et al., 2022)	Twitter	Hate speech	Text	No	No	No
(Thapa et al., 2022)	Twitter	Hate speech	Multimodal	Metadata	No	No
(B. Kennedy et al., 2022)	Gab	Hate speech	Text	No	Yes	Yes
(Albanyan et al., 2023)	Twitter	Hate speech	Text	Conversation	Yes	No

Appendix E: Categorization of features considered by the detection models

User metadata

Information about a user or an account, whether the speaker or the target of a potentially abusive comment, may help to infer intentionality or harm. For example, certain words might be acceptable among some users, whereas the same words could be considered abusive when used by others. Likewise, patterns of behavior can be indicative of intent, for example, users who repeatedly engage in abusive behavior may be more intentional. This can be operationalized through characterization of the history of a user's activities (Dadvar et al., 2013). More standardized user-level metadata, such as the geographical location of the user and the follower-following statistics of the message sender, have been shown to correlate with the occurrence of abusive content and are integrated as features in detection models (Bozyiğit et al., 2021).

Post metadata

Most social media platforms attach metadata to each post, for example, engagement metrics, mentions, and hashtags. These can reflect the broader context of a message. For instance, high engagement levels (likes, shares, comments) might indicate the popularity of (or controversy around) a post, while particular mentions and hashtags can indicate relevance to specific communities or ongoing discussions. Suhas Bharadwaj et al. (2022) incorporate hashtags and emojis as distinct features separate from the main text content. Bozyiğit et al. (2021) integrate post-level metadata, such as the number of retweets or mentions, to improve the performance of these models for detection of cyberbullying.

Image and video data

Many platforms have evolved to include a variety of media formats. Recognizing this, some researchers have extended their focus beyond text to include images and videos (Nisha & Jebathangam, 2022; Qiu et al., 2022). The additional context that visual and audio elements can provide may improve the detection of abusive content.

Psychological and cognitive features

Patterns of language may reflect personality, emotional states, and psychological traits (C. Alonso & Romero, 2017). Understanding the psychological and cognitive dimensions of users' behavior is particularly critical for understanding intent. Balakrishnan et al. (2020) incorporate multidimensional personality traits as features for cyber-aggression detection models.

Conversations

The conversation thread and previous interactions can offer useful context around potentially abusive language and provide evidence of intent. Ziems et al. (2020) incorporated features such as timeline similarity and mentions overlap based on shared conversations between the author and the target.

Graph structure

The relationships and interactions within social networks—such as who users connect with, how they interact with these connections, and the nature of the communities they are part of—can offer clues about users' intent. For instance, users embedded in tight networks may adopt similar communication patterns, which could be innocuous or abusive depending on norms of that group. Authors have incorporated network centrality measures for detection of cyberbullying (V. K. Singh et al., 2016).

Policy or rule-aware models

Norms within various online communities can shape what is viewed as inappropriate (Chandrasekharan et al., 2018). Policy or rule-aware models aim to ensure that automated systems adhere to guidelines and standards. The approach is particularly effective in environments where regulations may vary significantly, for example, across cultural contexts. D. Kumar et al. (2023) conducted prompt engineering to incorporate large language models into content moderation by including rules within the prompts. Calabrese et al. (2022) proposed a representation of moderation policies tailored for machine interpretation and illustrated how techniques from intent classification and slot filling can be applied to detect abusive content.

Sentiment

Sentiment analysis is a valuable component of many detection models. Sentiment features provide insights into the emotional tone of language which might not be apparent through baseline text analysis (Geetha et al., 2021).

Topics and themes

Topic modeling techniques, such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003) or theme categorization (Perera & Fernando, 2021), allow detection models to understand the subject matter of discussions. Models can learn whether certain topics are more likely to involve harmful language or cyberbullying. Murshed et al. (2023) employed a clustering-based topic modeling technique to improve the accuracy of cyberbullying detection. Perera & Fernando (2021) measured frequency of themes/categories associated with cyberbullying, for example, racist, sexual, and physical, to improve detection.

Linguistic cues

Words and phrases that are associated with offensive or abusive language are commonly used for abuse detection. This includes explicit language, slurs, and aggressive or threatening terms. Common approaches include constructing personalized dictionaries and using Linguistic Inquiry and Word Count (LIWC) for feature extraction (Geetha et al., 2021). Since TF-IDF and bag-of-words approaches are standard practices in NLP, we do not categorize them as nuanced uses of linguistic cues.

Appendix F: Detection paper categorization

User metadata

Al-Garadi et al., 2016; Balakrishnan et al., 2020; Bozyiğit et al., 2021; Cheng et al., 2019b; Dadvar et al., 2013, 2014; Dhingra et al., 2023; Escalante et al., 2017; Y. Liu et al., 2019; López-Vizcaíno et al., 2021; Nagar et al., 2022; Nisha & Jebathangam, 2022; W. Qian et al., 2023; Qiu et al., 2022; Sajadi Ansari et al., 2021; Ziems et al., 2020

Post metadata

Babaeianjelodar et al., 2022; Balakrishnan et al., 2020; Bozyiğit et al., 2021; Dhingra et al., 2023; Geetha et al., 2021; Y. Liu et al., 2019; López-Vizcaíno et al., 2021; Nisha & Jebathangam, 2022; Suhas Bharadwaj et al., 2022; Ziems et al., 2020

Image and video data

Cheng et al., 2019b; López-Vizcaíno et al., 2021; Nisha & Jebathangam, 2022; Qiu et al., 2022; N. M. Singh & Sharma, 2024; V. K. Singh et al., 2017; Thapa et al., 2022; Wang, Xiong, et al., 2020

Psychological and cognitive features

Al-Garadi et al., 2016; Balakrishnan et al., 2020; Cheng et al., 2019a

Features from conversations

Ashraf et al., 2021; H.-Y. Chen & Li, 2020; Dhingra et al., 2023; Nisha & Jebathangam, 2022; W. Qian et al., 2023; Ziems et al., 2020

Graph structure

Cécillon et al., 2021; Cheng et al., 2019a, 2019b; Dhingra et al., 2023; Q. Huang et al., 2014; Y. Liu et al., 2019; Nagar et al., 2022; Qiu et al., 2022; V. K. Singh et al., 2016; Ziems et al., 2020

Policy or rule-aware models

Calabrese et al., 2022; D. Kumar et al., 2023

Sentiment features

Babaeianjelodar et al., 2022; Dhingra et al., 2023; Geetha et al., 2021; Y. Liu et al., 2019; López-Vizcaíno et al., 2021; Nisha & Jebathangam, 2022; Perera & Fernando, 2021; Ziems et al., 2020

Topics and themes

López-Vizcaíno et al., 2021; Murshed et al., 2023; Perera & Fernando, 2021; Van Hee et al., 2018

Linguistic cues

Babaeianjelodar et al., 2022; Cheng et al., 2019a, 2019b; Dadvar et al., 2013; Dhingra et al., 2023; Geetha et al., 2021; Z. Li & Shimada, 2022; Y. Liu et al., 2019; López-Vizcaíno et al., 2021; Perera & Fernando, 2021; Sajadi Ansari et al., 2021; Van Hee et al., 2018; N. Zampieri et al., 2021; Ziems et al., 2020

Model evaluation, improvement and application

Abro et al., 2020; Aggarwal et al., 2020; Agnes et al., 2023; Ahmed et al., 2021, 2022; Aind et al., 2020; Akinyemi et al., 2023; Alksasbeh et al., 2021; P. Alonso et al., 2019, 2020; Alotaibi et al., 2021; Anjum & Katarya, 2022; Antypas & Camacho-Collados, 2023; Awal et al., 2021; Baydogan & Alatas, 2021, 2022; Beddiar et al., 2021, p. 24; Bhagya & Deepthi, 2021; Bokolo & Liu, 2023; Buan & Ramachandra, 2020; Bunde, 2021; Chandrasekaran et al., 2022; Chelmis & Zois, 2021; H. Chen et al., 2017, 2018; Daniel et al., 2023; De Souza & Da Costa-Abreu, 2020; Fale et al., 2023; Gopalan et al., 2023; Haider et al., 2023; Hamdy et al., 2020; Hani et al., 2019; Harish et al., 2023; Herath et al., 2020; Y. Huang et al., 2022; Ibrahim et al., 2020; Islam et al., 2023; Iwendi et al., 2023; Jahan et al., 2022; Jain et al., 2021; Karatsalos & Panagiotakis, 2020; Kavatagi & Rachh, 2021; Kavitha et al., 2023; Kazbekova et al., 2023; Kohli & Devi, 2023; Kovács et al., 2021, 2022; A. Kumar et al., 2021; A. Kumar & Kumar, 2023; A. Kumar & Sachdeva, 2022; Leo et al., 2023; P. Liu et al., 2019; M. López-Vizcaíno et al., 2023; Lu et al., 2020; Malik et al., 2021; Mathur et al., 2023; Mehta & Passi, 2022; Mercan et al., 2021; Mohtaj & Möller, 2022; Mozafari et al., 2020a, 2020b; Muneer et al., 2023; Murshed et al., 2022; Muzakir et al., 2023; Nascimento et al., 2022; Nath et al., 2022; Nitya Harshitha et al., 2024; Omran et al., 2023; Pahuja et al., 2023; Pariyani et al., 2021; Pavlopoulos et al., 2019; Phung et al., 2020; Pradhan et al., 2020; Preetham & Anitha, 2023; Qureshi & Sabih, 2021; Ramiandrisoa, 2022; Reddy et al., 2023; Sachdeva et al., 2021; Saha et al., 2019; Sahana et al., 2023; Salehghohari et al., 2022; Sathishkumar et al., 2023; Shankar et al., 2022; Sharif et al., 2021; P. Sharma & Tiwari, 2023; N. K. Singh et al., 2022; R. K. Singh et al., 2023; Sookarah & Ramwodin, 2022; Sultan et al., 2023; Tanase et al., 2020; Themeli et al., 2019; Thenmozhi et al., 2019, 2020; Xiang et al., 2021; Xingyi & Adnan, 2024; Yao, 2019; Yao et al., 2019; Yi & Zubiaga, 2023a; Yuan et al., 2023

Surveys

Ali et al., 2018; Aljohani et al., 2023; Alkomah & Ma, 2022; Alrashidi et al., 2022; Alrehili, 2019; Ambareen & Meenakshi Sundaram, 2023; Anjum & Katarya, 2024; Bhatt et al., 2023; Bilen, 2023; Elsafoury et al., 2021; Fersini et al., 2018; Gandhi et al., 2024; Gangurde et al., 2022; Gashroo & Mehrotra, 2022; Gongane et al., 2022, 2024; Gudumotu et al., 2023; Hussein & Aleqabie, 2023; Istaiteh et al., 2020; Jahan & Oussalah, 2023; Kaur et al., 2021; R. Kumar & Bhat, 2022; Mansur et al., 2023; Miran & Yahia, 2023; Modi, 2018; Mullah & Zainon, 2021; Rawat et al., 2024; Shakeel & Dwivedi, 2022; A. Sharma & Bhalla, 2022; G. Sharma et al., 2022; Vora et al., 2023; Yi & Zubiaga, 2023b; Yin & Zubiaga, 2021

Appendix G: Bibliography for appendices A–F

- Abro, S., Shaikh, S., Ali, Z., Khan, S., Mujtaba, G., & Khand, Z. H. (2020). Automatic hate speech detection using machine learning: A comparative study. *International Journal of Advanced Computer Science and Applications*, 11(8), 484–491. <https://doi.org/10.14569/IJACSA.2020.0110861>
- Aggarwal, A., Maurya, K., & Chaudhary, A. (2020). Comparative study for predicting the severity of cyberbullying across multiple social media platforms. In *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 871–877). IEEE. <https://doi.org/10.1109/ICICCS48265.2020.9121046>
- Agnes, S. A., Solomon, A. A., & Tamilmaran, D. J. C. (2023). Abusive comment detection in social media with bidirectional LSTM model. In *2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT)* (pp. 1368–1373). IEEE. <https://doi.org/10.1109/ICSSIT55814.2023.10060887>
- Agrafiotis, I., Bada, M., Cornish, P., Creese, S., Goldsmith, M., Ignatuschtschenko, E., Roberts, T., & Upton, D. M. (2016). *Cyber harm: Concepts, taxonomy and measurement*. SSRN. <https://dx.doi.org/10.2139/ssrn.2828646>
- Ahmed, T., Ivan, S., Kabir, M., Mahmud, H., & Hasan, K. (2022). Performance analysis of transformer-based architectures and their ensembles to detect trait-based cyberbullying. *Social Network Analysis and Mining*, 12(1), 99. <https://doi.org/10.1007/s13278-022-00934-4>
- Ahmed, T., Kabir, M., Ivan, S., Mahmud, H., & Hasan, K. (2021). Am I being bullied on social media? An ensemble approach to categorize cyberbullying. In *2021 IEEE International Conference on Big Data (Big Data)* (pp. 2442–2453). IEEE. <https://doi.org/10.1109/BigData52589.2021.9671594>
- Aind, A. T., Ramnaney, A., & Sethia, D. (2020). Q-bully: A reinforcement learning based cyberbullying detection framework. In *2020 International Conference for Emerging Technology (INCET)* (pp. 1–6). IEEE. <https://doi.org/10.1109/INCET49848.2020.9154092>
- Akinyemi, J. D., Ibitoye, A. O., Oyewale, C. T., & Onifade, O. F. (2023). Cyberbullying detection and classification in social media texts using machine learning techniques. In Z. Hu, I. Dychka, & M. He (Eds.), *International Conference on Computer Science, Engineering and Education Applications* (pp. 440–449). Springer. https://doi.org/10.1007/978-3-031-36118-0_40
- Al-Garadi, M. A., Varathan, K. D., & Ravana, S. D. (2016). Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. *Computers in Human Behavior*, 63, 433–443. <https://doi.org/10.1016/j.chb.2016.05.051>
- Ali, W. N. H. W., Mohd, M., & Fauzi, F. (2018). Cyberbullying detection: An overview. In *2018 Cyber Resilience Conference (CRC)* (pp. 1–3). IEEE. <https://doi.org/10.1109/CR.2018.8626869>
- Aljohani, E. J., Yafooz, W. M., & Alsaeedi, A. (2023). Cyberbullying detection approaches: A review. In *2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA)* (pp. 1310–1316). IEEE. <https://doi.org/10.1109/ICIRCA57980.2023.10220688>
- Alkomah, F., & Ma, X. (2022). A literature review of textual hate speech detection methods and datasets. *Information*, 13(6), 273. <https://doi.org/10.3390/info13060273>
- Alksasbeh, M. Z., Alqaralleh, B. A., Abukhalil, T., Abukaraki, A., Al Rawashdeh, T., & Al-Jaafreh, M. (2021). Smart detection of offensive words in social media using the soundex algorithm and permuterm index. *International Journal of Electrical and Computer Engineering (IJECE)*, 11(5), 4431–4438. <http://doi.org/10.11591/ijece.v11i5.pp4431-4438>
- Alonso, C., & Romero, E. (2017). Aggressors and victims in bullying and cyberbullying: A study of personality profiles using the five-factor model. *The Spanish Journal of Psychology*, 20, E76. <https://doi.org/10.1017/sjp.2017.73>

- Alonso, P., Saini, R., & Kovács, G. (2019). The North at HASOC 2019: Hate speech detection in social media data. In P. Mehta, P. Rosso, P. Majumder, & M. Mitra (Eds.), *Working notes of FIRE 2019: Forum for Information Retrieval Evaluation, Kolkata, India, December 12–15, 2019* (pp. 293–299). CEUR-WS.org. <https://ceur-ws.org/Vol-2517/T3-15.pdf>
- Alonso, P., Saini, R., & Kovács, G. (2020). Hate speech detection using transformer ensembles on the HASOC dataset. In A. Karpov & R. Potapova (Eds.), *Speech and Computer: 22nd International Conference, SPECOM 2020, St. Petersburg, Russia, October 7–9, 2020, Proceedings* (pp. 13–21). Springer. https://doi.org/10.1007/978-3-030-60276-5_2
- Alotaibi, M., Alotaibi, B., & Razaque, A. (2021). A multichannel deep learning framework for cyberbullying detection on social media. *Electronics*, 10(21), 2664. <https://doi.org/10.3390/electronics10212664>
- Alrashidi, B., Jamal, A., Khan, I., & Alkhathlan, A. (2022). A review on abusive content automatic detection: Approaches, challenges and opportunities. *PeerJ Computer Science*, 8. <https://doi.org/10.7717/PEERJ-CS.1142>
- Alrehili, A. (2019). Automatic hate speech detection on social media: A brief survey. *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)* (pp. 1–6). IEEE. <https://doi.org/10.1109/AICCSA47632.2019.9035228>
- Ambareen, K., & Meenakshi Sundaram, S. (2023). A survey of cyberbullying detection and performance: Its impact in social media using artificial intelligence. *SN Computer Science*, 4(6), 859. <https://doi.org/10.1007/s42979-023-02301-2>
- Anjum, & Katarya, R. (2022). Analysis of online toxicity detection using machine learning approaches. In G. Sanyal, C. M. Travieso-González, S. Awasthi, C. M. A. Pinto, & B. R. Purushothama (Eds.), *International Conference on Artificial Intelligence and Sustainable Engineering: Select Proceedings of AISE 2020* (Vol. 836, pp. 381–392). Springer. https://doi.org/10.1007/978-981-16-8542-2_29
- Anjum, & Katarya, R. (2024). Hate speech, toxicity detection in online social media: A recent survey of state of the art and opportunities. *International Journal of Information Security*, 23, 577–608. <https://doi.org/10.1007/s10207-023-00755-2>
- Antypas, D., & Camacho-Collados, J. (2023). Robust hate speech detection in social media: A cross-dataset empirical evaluation. In Y. Chung, P. Röttger, D. Nozza, Z. Talat, & A. Mostafazadeh Davani (Eds.), *The 7th Workshop on Online Abuse and Harms (WOAH)* (pp. 231–242). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.woah-1.25>
- Anzovino, M., Fersini, E., & Rosso, P. (2018). Automatic identification and classification of misogynistic language on Twitter. In M. Silberstein, F. Atigui, E. Kornysheva, E. Métais, & F. Mezziane (Eds.), *Natural Language Processing and Information Systems 23rd International Conference on Applications of Natural Language to Information Systems, NLDB 2018, Paris, France, June 13-15, 2018, Proceedings* (pp. 57–64). Springer. https://doi.org/10.1007/978-3-319-91947-8_6
- Ashraf, N., Zubiaga, A., & Gelbukh, A. (2021). Abusive language detection in YouTube comments leveraging replies as conversational context. *PeerJ Computer Science*, 7. <https://doi.org/10.7717/peerj-cs.742>
- Awal, M. R., Cao, R., Lee, R. K. W., & Mitrović, S. (2021, May). Angrybert: Joint learning target and emotion for hate speech detection. In K. Karlapalem, H. Cheng, N. Ramakrishnan, R. K. Agrawal, P. K. Reddy, J. Srivastava, & T. Chakraborty (Eds.), *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 701–713). Springer. https://doi.org/10.1007/978-3-030-75762-5_55

- Babaeianjelodar, M., Poorna Prudhvi, G., Lorenz, S., Chen, K., Mondal, S., Dey, S., & Kumar, N. (2022). Interpretable and high-performance hate and offensive speech detection. In J. Y. C. Chen, G. Fragomeni, H. Degen, & S. Ntoa (Eds.), *International Conference on Human-Computer Interaction* (pp. 233–244). Springer. https://doi.org/10.1007/978-3-031-21707-4_18
- Balakrishnan, V., Khan, S., & Arabnia, H. R. (2020). Improving cyberbullying detection using Twitter users' psychological features and machine learning. *Computers & Security*, 90, 101710. <https://doi.org/10.1016/j.cose.2019.101710>
- Banko, M., MacKeen, B., & Ray, L. (2020). A unified taxonomy of harmful content. In S. Akiwowo, B. Vidgen, V. Prabhakaran, & Z. Waseem (Eds.), *Proceedings of the Fourth Workshop on Online Abuse and Harms* (pp. 125–137). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.alw-1.16>
- Baydogan, C., & Alatas, B. (2021). Metaheuristic ant lion and moth flame optimization-based novel approach for automatic detection of hate speech in online social networks. *IEEE Access*, 9, 110047–110062. <https://doi.org/10.1109/ACCESS.2021.3102277>
- Baydogan, C., & Alatas, B. (2022). Deep-Cov19-Hate: A textual-based novel approach for automatic detection of hate speech in online social networks throughout COVID-19 with shallow and deep learning models. *Tehnicki Vjesnik*, 29(1), 149–156. <https://doi.org/10.17559/TV-20210708143535>
- Beddiar, D., Jahan, M., & Oussalah, M. (2021). *Data expansion using back translation and paraphrasing for hate speech detection. Online Social Networks and Media*, 24, 100153. <https://doi.org/10.1016/j.osnem.2021.100153>
- Bhagya, J., & Deepthi, P. (2021). Cyberbullying detection on social media using SVM. In V. Suma, J. I. Chen, Z. Baig, & H. Wang (Eds.), *Inventive Systems and Control: Proceedings of ICISC 2021* (pp. 17–27). Springer. https://doi.org/10.1007/978-981-16-1395-1_2
- Bhatt, C., Saini, N., Chauhan, R., & Sahoo, A. K. (2023). Machine learning techniques for hate speech detection on social media. In *2023 3rd International Conference on Innovative Sustainable Computational Technologies (CISCT)* (pp. 1–5). IEEE. <https://doi.org/10.1109/CISCT57197.2023.10351228>
- Bilen, A. (2023). A review: Detection of discrimination and hate speech shared on social media platforms using artificial intelligence methods. In M. Kılıç & S. Bozkuş Kahyaoğlu (Eds.), *Algorithmic discrimination and ethical perspective of artificial intelligence* (pp. 171–181). Springer. https://doi.org/10.1007/978-981-99-6327-0_12
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(2003), 993–1022. <https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf>
- Bokolo, B. G., & Liu, Q. (2023). Cyberbullying detection on social media using machine learning. In *IEEE INFOCOM 2023-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)* (pp. 1–6). IEEE. <https://doi.org/10.1109/INFOCOMWKSHPS57453.2023.10226114>
- Bozyiğit, A., Utku, S., & Nasibov, E. (2021). Cyberbullying detection: Utilizing social media features. *Expert Systems with Applications*, 179, 115001. <https://doi.org/10.1016/j.eswa.2021.115001>
- Buan, T. A., & Ramachandra, R. (2020). Automated cyberbullying detection in social media using an SVM activated stacked convolution LSTM network. In *Proceedings of the 2020 4th International Conference on Compute and Data Analysis* (pp. 170–174). Association for Computing Machinery. <https://doi.org/10.1145/3388142.3388147>
- Bunde, E. (2021). AI-assisted and explainable hate speech detection for social media moderators—A design science approach. In *Proceedings of the 54th Hawaii International Conference on System Sciences (HICSS)*. (pp. 1264–1273). <https://doi.org/10.24251/HICSS.2021.154>

- Calabrese, A., Ross, B., & Lapata, M. (2022). Explainable abuse detection as intent classification and slot filling. *Transactions of the Association for Computational Linguistics*, 10, 1440–1454.
https://doi.org/10.1162/tacl_a_00527
- Cécillon, N., Labatut, V., Dufour, R., & Linares, G. (2021). Graph embeddings for abusive language detection. *SN Computer Science*, 2, 37. <https://doi.org/10.1007/s42979-020-00413-7>
- Neelakandan, S., Sridevi, M., Chandrasekaran, S., Murugeswari, K., Singh Pundir, A. K., Sridevi, R., & Lingaiah, T. B. (2022). Deep learning approaches for cyberbullying detection and classification on social media. *Computational Intelligence and Neuroscience*, 2022(1), 163458.
<https://doi.org/10.1155/2022/2163458>
- Chandrasekharan, E., Samory, M., Jhaver, S., Charvat, H., Bruckman, A., Lampe, C., Eisenstein, J., & Gilbert, E. (2018). The Internet's hidden rules: An empirical study of Reddit norm violations at micro, meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW), 1–25. <https://doi.org/10.1145/3274301>
- Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., & Vakali, A. (2017). Mean birds: Detecting aggression and bullying on Twitter. In *Proceedings of the 2017 ACM on Web Science Conference* (pp. 13–22). Association for Computing Machinery.
<https://doi.org/10.1145/3091478.3091487>
- Chelmiss, C., & Zois, D.-S. (2021). Dynamic, incremental, and continuous detection of cyberbullying in online social media. *ACM Transactions on the Web (TWEB)*, 15(3), 1–33.
<https://doi.org/10.1145/3448014>
- Chen, H., McKeever, S., & Delany, S. J. (2017). Harnessing the Power of Text Mining for the Detection of Abusive Content in Social Media. In: P. Angelov, A. Gegov, C. Jayne, & Q. Shen (Eds.), *Advances in Computational Intelligence Systems. Contributions presented at the 16th UK Workshop on Computational Intelligence, September 7–9, 2016, Lancaster, UK* (pp. 187–205). Springer.
https://doi.org/10.1007/978-3-319-46562-3_12
- Chen, H., McKeever, S., & Delany, S. J. (2018). A comparison of classical versus deep learning techniques for abusive content detection on social media sites. In S. Staab, O. Koltsova, & D. I. Ignatov (Eds.), *Social Informatics: 10th International Conference, SocInfo 2018, St. Petersburg, Russia, September 25-28, 2018, Proceedings, Part I* (pp. 117–133). Springer.
https://doi.org/10.1007/978-3-030-01129-1_8
- Chen, H.-Y., & Li, C.-T. (2020). HENIN: Learning heterogeneous neural interaction networks for explainable cyberbullying detection on social media. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 2543–2552). Association for Computational Linguistics.
<https://doi.org/10.18653/v1/2020.emnlp-main.200>
- Cheng, L., Li, J., Silva, Y., Hall, D., & Liu, H. (2019a). PI-bully: Personalized cyberbullying detection with peer influence. In S. Kraus (Ed.), *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 5829–5835). International Joint Conferences on Artificial Intelligence. <https://doi.org/10.24963/ijcai.2019/808>
- Cheng, L., Li, J., Silva, Y. N., Hall, D. L., & Liu, H. (2019b). Xbully: Cyberbullying detection within a multi-modal context. In *WSDM '19: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining* (pp. 339–347). Association for Computing Machinery.
<https://doi.org/10.1145/3289600.3291037>
- Dadvar, M., Trieschnigg, D., & De Jong, F. (2014). Experts and machines against bullies: A hybrid approach to detect cyberbullies. In M. Sokolova & P. Beek (Eds.), *Advances in Artificial Intelligence: Proceedings of the 27th Canadian Conference on Artificial Intelligence, Canadian AI 2014, Montréal, QC, Canada, May 6-9, 2014* (pp. 275–281). Springer.
https://doi.org/10.1007/978-3-319-06483-3_25

- Dadvar, M., Trieschnigg, D., Ordelman, R., & De Jong, F. (2013). Improving cyberbullying detection with user context. In P. Serdyukov, P. Braslavski, S. O. Kuznetsov, J. Kamps, S. Rüger, E. Agichtein, I. Segalovich, & E. Yilmaz (Eds.), *Advances in Information Retrieval: 35th European Conference on IR Research, ECIR 2013, Moscow, Russia, March 24-27, 2013. Proceedings* (pp. 693–696). Springer. https://doi.org/10.1007/978-3-642-36973-5_62
- Daniel, R., Murthy, T. S., Kumari, C. D. V. P., Lydia, E. L., Ishak, M. K., Hadjouni, M., & Mostafa, S. M. (2023). Ensemble learning with tournament selected glowworm swarm optimization algorithm for cyberbullying detection on social media. *IEEE Access*, 11, 123392–123400. <https://doi.org/10.1109/ACCESS.2023.3326948>
- Davidson, T., Warmesley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1), 512–515. <https://doi.org/10.1609/icwsm.v11i1.14955>
- De Souza, G. A., & Da Costa-Abreu, M. (2020). Automatic offensive language detection from Twitter data using machine learning and feature selection of metadata. In *2020 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–6). IEEE. <https://doi.org/10.1109/IJCNN48605.2020.9207652>
- Elsafoury, F., Katsigiannis, S., Pervez, Z., & Ramzan, N. (2021). When the timeline meets the pipeline: A survey on automated cyberbullying detection. *IEEE Access*, 9, 103541–103563. <https://doi.org/10.1109/ACCESS.2021.3098979>
- ElSherief, M., Nilizadeh, S., Nguyen, D., Vigna, G., & Belding, E. (2018). Peer to peer hate: Hate speech instigators and their targets. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1), 52–61. <https://doi.org/10.1609/icwsm.v12i1.15038>
- Escalante, H. J., Villatoro-Tello, E., Garza, S. E., López-Monroy, A. P., Montes-y-Gómez, M., & Villaseñor-Pineda, L. (2017). Early detection of deception and aggressiveness using profile-based representations. *Expert Systems with Applications*, 89, 99–111. <https://doi.org/10.1016/j.eswa.2017.07.040>
- Fale, P. N., Goyal, K. K., & Shivani, S. (2023). A hybrid deep learning approach for abusive text detection. *AIP Conference Proceedings*, 2753(1). <https://doi.org/10.1063/5.0128071>
- Fersini, E., Rosso, P., & Anzovino, M. (2018). Overview of the task on automatic misogyny identification at IberEval 2018. In P. Rosso, J. Gonzalo, R. Martínez, S. Montalvo, & J. Carrillo de Albornoz (Eds.), *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018, CEUR Workshop Proceedings 2150* (pp. 214–228). <https://boa.unimib.it/handle/10281/219328>
- Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., & Kourtellis, N. (2018). Large scale crowdsourcing and characterization of Twitter abusive behavior. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1). <https://doi.org/10.1609/icwsm.v12i1.14991>
- Gandhi, A., Ahir, P., Adharyu, K., Shah, P., Lohiya, R., Cambria, E., Poria, S., & Hussain, A. (2024). Hate speech detection: A comprehensive review of recent works. *Expert Systems*, 41(8), e13562. <https://doi.org/10.1111/exsy.13562>
- Gangurde, A., Mankar, P., Chaudhari, D., & Pawar, A. (2022). A systematic bibliometric analysis of hate speech detection on social media sites. *Journal of Scientometric Research*, 11(1), 100–111. <https://dx.doi.org/10.5530/jscires.11.1.10>

- Gao, L., Kuppersmith, A., & Huang, R. (2017). Recognizing explicit and implicit hate speech using a weakly supervised two-path bootstrapping approach. In G. Kondrak & T. Watanabe (Eds.), *Proceedings of the Eighth International Joint Conference on Natural Language Processing* (Vol. 1, pp. 774–782). Asian Federation of Natural Language Processing. <https://aclanthology.org/I17-1078>
- Geetha, R., Karthika, S., Sowmika, C. J., & Janani, B. M. (2021). Auto-Off ID: Automatic detection of offensive language in social media. *Journal of Physics: Conference Series*, 1911(1), 012012. <https://doi.org/10.1088/1742-6596/1911/1/012012>
- Golbeck, J., Ashktorab, Z., Banjo, R. O., Berlinger, A., Bhagwan, S., Buntain, C., Cheakalos, P., Geller, A. A., Gnanasekaran, R. K., Gunasekaran, R. R., Hoffman, K. M., Hottle, J., Jienjitlert, V., Khare, S., Lau, R., Martindale, M. J., Naik, S., Nixon, H. L., Ramachandran, P., ... D. M. Wu. (2017). A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on Web Science Conference* (pp. 229–233). Association for Computing Machinery. <https://doi.org/10.1145/3091478.3091509>
- Gomez, R., Gibert, J., Gomez, L., & Karatzas, D. (2020). Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 1470–1478). IEEE. <https://doi.ieeecomputersociety.org/10.1109/WACV45572.2020.9093414>
- Gongane, V. U., Munot, M. V., & Anuse, A. (2022). Feature representation techniques for hate speech detection on social media: A comparative study. In *2022 International Conference on Signal and Information Processing (IconSIP)* (pp. 1–6). IEEE. <https://doi.org/10.1109/ICoNSIP49665.2022.10007458>
- Gongane, V. U., Munot, M. V., & Anuse, A. D. (2024). A survey of explainable AI techniques for detection of fake news and hate speech on social media platforms. *Journal of Computational Social Science*, 7, 587–623. <https://doi.org/10.1007/s42001-024-00248-9>
- Gopalan, A., Mohanavel, V., Geo, A. A., Rajkumar, G. V., Kavitha, T., & Pooja, P. (2023). Experimental evaluation of robust cyberbullying detection over social media using intelligent learning scheme. In *2023 International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE)* (pp. 1–7). IEEE. <https://doi.org/10.1109/RMKMATE59243.2023.10368747>
- Gudumotu, C. E., Nukala, S. R., Reddy, K., Konduri, A., & Gireesh, C. (2023). A survey on deep learning models to detect hate speech and bullying in social media. In A. Biswas, V. B. Semwal, & D. Singh (Eds.), *Artificial Intelligence for Societal Issues* (pp. 27–44). Springer. https://doi.org/10.1007/978-3-031-12419-8_2
- Haider, F., Dipty, I., Rahman, F., Assaduzzaman, M., & Sohel, A. (2023). Social media hate speech detection using machine learning approach. In S. C. K. R., N. Sujaudeen, A. Beulah, & H. S. Hamead (Eds.), *Computational Intelligence in Data Science: 6th IFIP TC 12 International Conference, ICCIDS 2023, Chennai, India, February 23–25, 2023, revised selected papers* (pp. 218–229). Springer. https://doi.org/10.1007/978-3-031-38296-3_17
- Hamdy, E., Mitrović, J., & Granitzer, M. (2020). nlpUP at SemEval-2020 Task 12: A blazing fast system for offensive language detection. In A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, & E. Shutova (Eds.), *Proceedings of the Fourteenth Workshop on Semantic Evaluation* (pp. 2098–2104). International Committee for Computational Linguistics. <https://doi.org/10.18653/v1/2020.semeval-1.278>
- Hani, J., Nashaat, M., Ahmed, M., Emad, Z., Amer, E., & Mohammed, A. (2019). Social media cyberbullying detection using machine learning. *International Journal of Advanced Computer Science and Applications*, 10(5), 703–707. <https://dx.doi.org/10.14569/IJACSA.2019.0100587>

- Harish, D., Alamelu, M., Manimaran, M., & Jayashakthi, V. P. (2023). Automatic detection of cyberbullying on social media using machine learning. In *2023 2nd International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)* (pp. 1–6). IEEE. <https://doi.org/10.1109/ICAECA56562.2023.10201149>
- He, B., Ziems, C., Soni, S., Ramakrishnan, N., Yang, D., & Kumar, S. (2021). Racism is a virus: Anti-Asian hate and counterspeech in social media during the COVID-19 crisis. In M. Coscia, A. Cuzzocrea, & K. Shu (Eds.), *ASONAM '21: Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 90–94). Association for Computing Machinery. <https://doi.org/10.1145/3487351.3488324>
- Herath, M., Atapattu, T., Dung, H. A., Treude, C., & Falkner, K. (2020). AdelaideCyC at SemEval-2020 task 12: Ensemble of classifiers for offensive language detection in social media. In A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, & E. Shutova (Eds.), *Proceedings of the Fourteenth Workshop on Semantic Evaluation* (pp. 1516–1523). International Committee for Computational Linguistics. <https://doi.org/10.18653/v1/2020.semeval-1.198>
- Huang, Q., Singh, V. K., & Atrey, P. K. (2014). Cyber bullying detection using social and textual analysis. In *SAM '14: Proceedings of the 3rd International Workshop on Socially-Aware Multimedia* (pp. 3–6). Association for Computing Machinery. <https://doi.org/10.1145/2661126.2661133>
- Huang, Y., Song, R., Giunchiglia, F., & Xu, H. (2022). A multitask learning framework for abuse detection and emotion classification. *Algorithms*, 15(4). <https://doi.org/10.3390/a15040116>
- Hussein, F. N. A., & Aleqabie, H. J. (2023). Cyberbullying detection on social media: A brief survey. In *2023 Second International Conference on Advanced Computer Applications (ACA)* (pp. 1–6). IEEE. <https://doi.org/10.1109/ACA57612.2023.10346758>
- Ibrahim, M., Torki, M., & El-Makky, N. (2020). AlexU-BackTranslation-TL at SemEval-2020 Task 12: Improving offensive language detection using data augmentation and transfer learning. In A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, & E. Shutova (Eds.), *Proceedings of the Fourteenth Workshop on Semantic Evaluation* (pp. 1881–1890). International Committee for Computational Linguistics. <https://doi.org/10.18653/v1/2020.semeval-1.248>
- Istaiteh, O., Al-Omouh, R., & Tedmori, S. (2020). Racist and sexist hate speech detection: Literature review. In *2020 International Conference on Intelligent Data Science Technologies and Applications (IDSTA)* (pp. 95–99). <https://doi.org/10.1109/IDSTA50958.2020.9264052>
- Iwendi, C., Srivastava, G., Khan, S., & Maddikunta, P. K. R. (2023). Cyberbullying detection solutions based on deep learning architectures. 29(3), 1839–1852. <https://doi.org/10.1007/s00530-020-00701-5>
- Jahan, M. S., Beddiar, D. R., Oussalah, M., & Mohamed, M. (2022). Data expansion using wordnet-based semantic expansion and word disambiguation for cyberbullying detection. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 1761–1770). European Language Resources Association. <https://aclanthology.org/2022.lrec-1.187>
- Jahan, M. S., & Oussalah, M. (2023). A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546, 126232. <https://doi.org/10.1016/j.neucom.2023.126232>
- Jahn, L., Rendsvig, R. K., Flammini, A., Menczer, F., & Hendricks, V. F. (2023). Friction interventions to curb the spread of misinformation on social media. arXiv. <https://doi.org/10.48550/arXiv.2307.11498>

- Jain, V., Kumar, V., Pal, V., & Vishwakarma, D. K. (2021). Detection of cyberbullying on social media using machine learning. In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 1091–1096). IEEE.
<https://doi.org/10.1109/ICCMC51019.2021.9418254>
- Karatsalos, C., & Panagiotakis, Y. (2020). Attention-based method for categorizing different types of online harassment language. In P. Cellier & K. Driessens (Eds.), *Machine learning and knowledge discovery in databases: International workshops of ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part II* (pp. 321–330). Springer.
https://doi.org/10.1007/978-3-030-43887-6_26
- Kaur, S., Singh, S., & Kaushal, S. (2021). Abusive content detection in online user-generated data: A survey. *Procedia Computer Science*, 189, 274–281. <https://doi.org/10.1016/j.procs.2021.05.098>
- Kavatagi, S., & Rachh, R. (2021). A context aware embedding for the detection of hate speech in social media networks. In *2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)* (pp. 1–4). IEEE.
<https://doi.org/10.1109/SMARTGENCON51891.2021.9645877>
- Kavitha, S., Anchitalagammai, J., Murali, S., Deepalakshmi, R., Himall, L., & Suryakanth, M. (2023). Smart language checker: A machine learning solution for offensive language detection in social media. In *2023 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI)* (pp. 1–6). IEEE. <https://doi.org/10.1109/ICDSAAI59313.2023.10452454>
- Kazbekova, G., Ismagulova, Z., Kemelbekova, Z., Tileubay, S., Baimurzaev, B., & Bazarbayeva, A. (2023). Offensive language detection on online social networks using hybrid deep learning architecture. *International Journal of Advanced Computer Science and Applications*, 14(11), 793–805.
<https://doi.org/10.14569/IJACSA.2023.0141180>
- Kennedy, B., Atari, M., Mostafazadeh Davani, A., Yeh, L., Omrani, A., Kim, Y., Coombs, K. Jr., Havaladar, S., Portillo-Wightman, G., Gonzalez, E., Hoover, J., Azatian, A., Hussain, A., Lara, A., Cardenas, G., Omary, A., Park, C., Wang, X., Wijaya, C., ... M. Dehghani. (2022). Introducing the Gab hate corpus: Defining and applying hate-based rhetoric to social media posts at scale. *Language Resources and Evaluation*, 56, 79–108. <https://doi.org/10.1007/s10579-021-09569-x>
- Kennedy, C. J., Bacon, G., Sahn, A., & von Vacano, C. (2020). *Constructing interval variables via faceted Rasch measurement and multitask deep learning: A hate speech application*. arXiv.
<https://doi.org/10.48550/arXiv.2009.10277>
- Kogilavani, S. V., Malliga, S., Jaibinaya, K., Malini, M., & Kokila, M. M. (2023). Characterization and mechanical properties of offensive language taxonomy and detection techniques. *Materials Today: Proceedings*, 81.2., 630–633. <https://doi.org/10.1016/j.matpr.2021.04.102>
- Kovács, G., Alonso, P., & Saini, R. (2021). Challenges of hate speech detection in social media: Data scarcity, and leveraging external resources. *SN Computer Science*, 2(2), 95.
<https://doi.org/10.1007/s42979-021-00457-3>
- Kovács, G., Alonso, P., Saini, R., & Liwicki, M. (2022). Leveraging external resources for offensive content detection in social media. *AI Communications*, 35(2), 87–109. <https://doi.org/10.3233/AIC-210138>
- Kumar, A., & Kumar, S. (2023). Hate speech detection in multi-social media using deep learning. In R. N. Shaw, M. Paprzycki, & A. Ghosh (Eds.), *Advanced Communication and Intelligent Systems Second International Conference, ICACIS 2023, Warsaw, Poland, June 16–17, 2023, Revised Selected Papers, Part I* (pp. 59–70). Springer. https://doi.org/10.1007/978-3-031-45121-8_6
- Kumar, A., & Sachdeva, N. (2022). A Bi-GRU with attention and CapsNet hybrid model for cyberbullying detection on social media. *World Wide Web*, 25(4), 1537–1550.
<https://doi.org/10.1007/s11280-021-00920-4>

- Kumar, A., Tyagi, V., & Das, S. (2021). Deep learning for hate speech detection in social media. In *2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON)* (pp. 1–4). IEEE. <https://doi.org/10.1109/GUCON50781.2021.9573687>
- Kumar, R., & Bhat, A. (2022). A study of machine learning-based models for detection, control, and mitigation of cyberbullying in online social media. *International Journal of Information Security*, 21(6), 1409–1431. <https://doi.org/10.1007/s10207-022-00600-y>
- Li, Z., & Shimada, K. (2022). Combining pre-trained language models and features for offensive language detection. In *2022 13th International Congress on Advanced Applied Informatics Winter (IIAI-AAI-Winter)* (pp. 5–10). IEEE. <https://doi.org/10.1109/IIAI-AAI-Winter58034.2022.00012>
- Liu, P., Li, W., & Zou, L. (2019). NULI at SemEval-2019 Task 6: Transfer learning for offensive language detection using bidirectional transformers. In J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki, & S. M. Mohammad (Eds.), *Proceedings of the 13th International Workshop on Semantic Evaluation* (pp. 87–91). Association for Computational Linguistics. <https://doi.org/10.18653/v1/S19-2011>
- Liu, Y., Zavarsky, P., & Malik, Y. (2019). Non-linguistic features for cyberbullying detection on a social media platform using machine learning. In J. Vaidya, X. Zhang, & J. Li (Eds.), *Cyberspace Safety and Security: 11th International Symposium, CSS 2019, Guangzhou, China, December 1–3, 2019, Proceedings, Part I* (pp. 391–406). Springer. https://doi.org/10.1007/978-3-030-37337-5_31
- López-Vizcaíno, M., Nóvoa, F. J., Artieres, T., & Cacheda, F. (2023). Site agnostic approach to early detection of cyberbullying on social media networks. *Sensors*, 23(10), 4788. <https://doi.org/10.3390/s23104788>
- Lu, N., Wu, G., Zhang, Z., Zheng, Y., Ren, Y., & Choo, K.-K. R. (2020). Cyberbullying detection in social media text based on character-level convolutional neural network with shortcuts. *Concurrency and Computation: Practice and Experience*, 32(23), e5627. <https://doi.org/10.1002/cpe.5627>
- Malik, P., Aggrawal, A., & Vishwakarma, D. K. (2021). Toxic speech detection using traditional machine learning models and bert and fasttext embedding with deep neural networks. In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 1254–1259). IEEE. <https://doi.org/10.1109/ICCMC51019.2021.9418395>
- Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandlia, C., & Patel, A. (2019). Overview of the HASOC track at FIRE 2019: Hate speech and offensive content identification in Indo-European languages. In P. Majumder, M. Mitra, S. Gangopadhyay, & P. Mehta (Eds.), *Proceedings of the 11th Annual Meeting of the Forum for Information Retrieval Evaluation* (pp. 14–17). Association for Computing Machinery. <https://doi.org/10.1145/3368567.3368584>
- Mansur, Z., Omar, N., & Tiun, S. (2023). Twitter hate speech detection: A systematic review of methods, taxonomy analysis, challenges, and opportunities. *IEEE Access*, 11, 16226–16249. <https://doi.org/10.1109/ACCESS.2023.3239375>
- Mathur, S. A., Isarka, S., Dharmasivam, B., & Jaidhar, C. (2023). Analysis of tweets for cyberbullying detection. In *2023 Third International Conference on Secure Cyber Computing and Communication (ICSCCC)* (pp. 269–274). IEEE. <https://doi.org/10.1109/ICSCCC58608.2023.10176416>
- Mehta, H., & Passi, K. (2022). Social media hate speech detection using explainable artificial intelligence (XAI). *Algorithms*, 15(8), 291. <https://doi.org/10.3390/a15080291>
- Mercan, V., Jamil, A., Hameed, A. A., Magsi, I. A., Bazai, S., & Shah, S. A. (2021). Hate speech and offensive language detection from social media. In *2021 International Conference on Computing, Electronic and Electrical Engineering (ICE Cube)* (pp. 1–5). IEEE. <https://doi.org/10.1109/ICECube53880.2021.9628255>
- Miran, A. Z., & Yahia, H. S. (2023). Hate speech detection in social media (Twitter) using neural network. *Journal of Mobile Multimedia*, 19(3), 765–798. <http://dx.doi.org/10.13052/jmm1550-4646.1936>

- Miró-Llinares, F., & Rodríguez-Sala, J. J. (2016). Cyber hate speech on Twitter: Analyzing disruptive events from social media to build a violent communication and hate speech taxonomy. *International Journal of Design & Nature and Ecodynamics*, 11(3), 406–415.
<https://doi.org/10.2495/DNE-V11-N3-406-415>
- Modi, S. (2018). AHTDT- Automatic hate text detection techniques in social media. In *2018 International Conference on Circuits and Systems in Digital Enterprise Technology (ICCSDET)* (pp. 1–3). IEEE.
<https://doi.org/10.1109/ICCSDET.2018.8821128>
- Mohtaj, S., & Möller, S. (2022). On the importance of word embedding in automated harmful information detection. In P. Sojka, A. Horák, I. Kopeček, & K. Pala (Eds.), *Text, Speech, and Dialogue: 25th International Conference, TSD 2022, Brno, Czech Republic, September 6–9, 2022, Proceedings* (pp. 251–262). Springer. https://doi.org/10.1007/978-3-031-16270-1_21
- Mozafari, M., Farahbakhsh, R., & Crespi, N. (2020a). A BERT-based transfer learning approach for hate speech detection in online social media. In H. Cherifi, S. Gaito, J. F. Mendes, E. Moro, & L. M. Rocha (Eds.), *Complex networks and their applications VIII: Volume 1, proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019* (pp. 928–940). Springer. https://doi.org/10.1007/978-3-030-36687-2_77
- Mozafari, M., Farahbakhsh, R., & Crespi, N. (2020b). Hate speech detection and racial bias mitigation in social media based on BERT model. *PloS One*, 15(8), e0237861.
<https://doi.org/10.1371/journal.pone.0237861>
- Mullah, N. S., & Zainon, W. M. N. W. (2021). Advances in machine learning algorithms for hate speech detection in social media: A review. *IEEE Access*, 9, 88364–88376.
<https://doi.org/10.1109/ACCESS.2021.3089515>
- Muneer, A., Alwadain, A., Ragab, M. G., & Alqushaibi, A. (2023). Cyberbullying detection on social media using stacking ensemble learning and enhanced BERT. *Information (Switzerland)*, 14(8), 467.
<https://doi.org/10.3390/info14080467>
- Murshed, B. A. H., Abawajy, J., Mallappa, S., Saif, M. A. N., & Al-Ariki, H. D. E. (2022). DEA-RNN: A hybrid deep learning approach for cyberbullying detection in Twitter social media platform. *IEEE Access*, 10, 25857–25871. <https://doi.org/10.1109/ACCESS.2022.3153675>
- Murshed, B. A. H., Suresha, Abawajy, J., Saif, M. A. N., Abdulwahab, H. M., & Ghanem, F. A. (2023). FAEO-ECNN: Cyberbullying detection in social media platforms using topic modelling and deep learning. *Multimedia Tools and Applications*, 82(30), 46611–46650.
<https://doi.org/10.1007/s11042-023-15372-3>
- Muzakir, A., Adi, K., & Kusumaningrum, R. (2023). Classification of hate speech language detection on social media: Preliminary study for improvement. In M. B. Ahmed, B. A. Abdelhakim, B. K. Ane, & D. Rosiyadi (Eds.), *Emerging trends in intelligent systems & network security* (pp. 146–156). Springer. https://doi.org/10.1007/978-3-031-15191-0_14
- Nagar, S., Gupta, S., Bahushruth, C., Barbhuiya, F. A., & Dey, K. (2022). Hate speech detection on social media using graph convolutional networks. In R. M. Benito, C. Cherifi, H. Cherifi, E. Moro, L. M. Rocha, & M. Sales-Pardo (Eds.), *Complex networks & their applications X: Volume 2, proceedings of the Tenth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2021 10* (pp. 3–14). Springer. https://doi.org/10.1007/978-3-030-93413-2_1
- Nascimento, F. R. S., Cavalcanti, G. D. C., & Da Costa-Abreu, M. (2022). Unintended bias evaluation: An analysis of hate speech detection and gender bias mitigation on social media using ensemble learning. *Expert Systems with Applications*, 201, 117032.
<https://doi.org/10.1016/j.eswa.2022.117032>

- Nath, N., George, J. P., Kesan, A., & Rodrigues, A. (2022). An efficient deep learning-based hybrid architecture for hate speech detection in social media. In S. Shukla, X. Gao, J. V. Kureethara, & D. Mishra (Eds.), *Data science and security, proceedings of IDSCS 2022* (pp. 347–355). Springer. https://doi.org/10.1007/978-981-19-2211-4_30
- Nisha, M., & Jebathangam, J. (2022). Detection and classification of cyberbullying in social media using text mining. In *2022 6th International Conference on Electronics, Communication and Aerospace Technology* (pp. 856–861). IEEE. <https://doi.org/10.1109/ICECA55336.2022.10009445>
- Nitya Harshitha, T., Prabu, M., Suganya, E., Sountharajan, S., Bavirisetti, D. P., Gadde, N., & Uppu, L. S. (2024). ProTect: A hybrid deep learning model for proactive detection of cyberbullying on social media. *Frontiers in Artificial Intelligence*, 7, 1269366. <https://doi.org/10.3389/frai.2024.1269366>
- Nocentini, A., Calmaestra, J., Schultze-Krumbholz, A., Scheithauer, H., Ortega, R., & Menesini, E. (2010). Cyberbullying: Labels, behaviours and definition in three European countries. *Australian Journal of Guidance and Counselling*, 20(2), 129–142. <https://doi.org/10.1375/ajgc.20.2.129>
- Omran, E., Al Tatarwah, E., & Al Qundus, J. (2023). A comparative analysis of machine learning algorithms for hate speech detection in social media. *Online Journal of Communication and Media Technologies*, 13(4), e202348. <https://doi.org/10.30935/ojcm/13603>
- Ousidhoum, N., Lin, Z., Zhang, H., Song, Y., & Yeung, D.-Y. (2019). Multilingual and multi-aspect hate speech analysis. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 4675–4684). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D19-1474>
- Pahuja, V., Neema, S., & Dubey, R. (2023). Securing social spaces: Cyberbullying detection with ML and DL on social media platforms. In *2023 International Conference on Sustainable Communication Networks and Application (ICSCNA)* (pp. 1471–1476). IEEE. <https://doi.org/10.1109/icscna58489.2023.10370114>
- Pariyani, B., Shah, K., Shah, M., Vyas, T., & Degadwala, S. (2021). Hate speech detection in Twitter using natural language processing. In *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)* (pp. 1146–1152). IEEE. <https://doi.org/10.1109/ICICV50876.2021.9388496>
- Pavlopoulos, J., Thain, N., Dixon, L., & Androutsopoulos, I. (2019). ConvAI at SemEval-2019 Task 6: Offensive language identification and categorization with perspective and BERT. In J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki, & S. M. Mohammad (Eds.), *Proceedings of the 13th International Workshop on Semantic Evaluation* (pp. 571–576). Association for Computational Linguistics. <https://doi.org/10.18653/v1/S19-2102>
- Perera, A., & Fernando, P. (2021). Accurate cyberbullying detection and prevention on social media. *Procedia Computer Science*, 181, 605–611. <https://doi.org/10.1016/j.procs.2021.01.207>
- Phung, H. T., Dang, H. K. L., & Pham, M. T. (2020). Cyberbullying detection based on word curve representations using B-spline interpolation. In *Proceedings of the 4th International Conference on Future Networks and Distributed Systems*, 50, 1–7. <https://doi.org/10.1145/3440749.3442657>
- Pradhan, A., Yatam, V. M., & Bera, P. (2020). Self-attention for cyberbullying detection. In *2020 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA)* (pp. 1–6). IEEE. <https://doi.org/10.1109/CyberSA49311.2020.9139711>
- Preetham, J., & Anitha, J. (2023). Offensive language detection in social media using ensemble techniques. In *2023 International Conference on Circuit Power and Computing Technologies (ICCPCT)* (pp. 805–808). IEEE. <https://doi.org/10.1109/ICCPCT58313.2023.10245673>

- Qian, J., Bethke, A., Liu, Y., Belding, E., & Wang, W. Y. (2019). A benchmark dataset for learning to intervene in online hate speech. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 4755–4764). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D19-1482>
- Qian, W., Yu, S., Nie, Z., Lu, X. S., Liu, H., & Huang, B. (2023). Improved hierarchical attention networks for cyberbullying detection via social media data. In *2023 IEEE International Conference on Networking, Sensing and Control (ICNSC)* (pp. 407–409). IEEE. <http://dx.doi.org/10.1109/ICNSC58704.2023.10319023>
- Qiu, J., Hegde, N., Moh, M., & Moh, T.-S. (2022). Investigating user information and social media features in cyberbullying detection. In *2022 IEEE International Conference on Big Data (Big Data)* (pp. 3063–3070). IEEE. <https://doi.org/10.1109/BigData55660.2022.10020305>
- Qureshi, K. A., & Sabih, M. (2021). Un-compromised credibility: Social media based multi-class hate speech classification for text. *IEEE Access*, 9, 109465–109477. <https://doi.org/10.1109/ACCESS.2021.3101977>
- Ramiandrisoa, F. (2022). Multi-task learning for hate speech and aggression detection. In L. Tamine, E. Amigó, & J. Mothe (Eds.), *Proceedings of the 2nd Joint Conference of the Information Retrieval Communities in Europe (CIRCLE 2022) Samatan, Gers, France, July 4-7, 2022*. https://ceur-ws.org/Vol-3178/CIRCLE_2022_paper_31.pdf
- Rawat, A., Kumar, S., & Samant, S. S. (2024). Hate speech detection in social media: Techniques, recent trends, and future challenges. *Wiley Interdisciplinary Reviews: Computational Statistics*, 16(2), e1648. <https://doi.org/10.1002/wics.1648>
- Reddy, B. A. C., Chandra, G. K., Sisodia, D. S., & Anuragi, A. (2023). Balancing techniques for improving automated detection of hate speech and offensive language on social media. In *2023 2nd International Conference for Innovation in Technology (INOCON)* (pp. 1–8). IEEE. <https://doi.org/10.1109/INOCON57975.2023.10101157>
- Rezvan, M., Shekarpour, S., Balasuriya, L., Thirunarayan, K., Shalin, V. L., & Sheth, A. (2018). A quality type-aware annotated corpus and lexicon for harassment research. In *Proceedings of the 10th ACM Conference on Web Science* (pp. 33–36). Association for Computing Machinery. <https://doi.org/10.1145/3201064.3201103>
- Ribeiro, M., Calais, P., Santos, Y., Almeida, V., & Meira Jr, W. (2018). Characterizing and detecting hateful users on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1). <https://doi.org/10.1609/icwsm.v12i1.15057>
- Sachdeva, J., Chaudhary, K. K., Madaan, H., & Meel, P. (2021). Text based hate-speech analysis. In *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)* (pp. 661–668). IEEE. <https://doi.org/10.1109/ICAIS50930.2021.9396013>
- Saha, P., Mathew, B., Goyal, P., & Mukherjee, A. (2019). *Hatemonitors: Language agnostic abuse detection in social media*. arXiv.
- Sahana, V., Anil Kumar, K. M., & Darem, A. A. (2023). A comparative analysis of machine learning techniques for cyberbullying detection on FormSpring in textual modality. *International Journal of Computer Network and Information Security*, 15(4), 36–47. <https://doi.org/10.5815/ijcnis.2023.04.04>
- Sajadi Ansari, F., Barhamgi, M., Khelifi, A., & Benslimane, D. (2021). An approach to detect cyberbullying on social media. In C. Attiogbé & S. Ben Yahia (Eds.), *Model and Data Engineering: 10th International Conference, MEDI 2021, Tallinn, Estonia, June 21–23, 2021, proceedings* (pp. 53–66). Springer. https://doi.org/10.1007/978-3-030-78428-7_5

- Salawu, S., Lumsden, J., & He, Y. (2021). A large-scale English multi-label Twitter dataset for cyberbullying and online abuse detection. In A. Mostafazadeh Davani, D. Kiela, M. Lambert, B. Vidgen, V. Prabhakaran, & Z. Waseem (Eds.), *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)* (pp. 146–156). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.woah-1.16>
- Salehgohari, A., Mirhosseini, M., Tabrizchi, H., & Koczy, A. V. (2022). Abusive language detection on social media using bidirectional long-short term memory. In *2022 IEEE 26th International Conference on Intelligent Engineering Systems (INES)* (pp. 000243–000248). IEEE. <https://doi.org/10.1109/INES56734.2022.9922628>
- Samory, M., Sen, I., Kohne, J., Flöck, F., & Wagner, C. (2021). “Call me sexist, but...”: Revisiting sexism detection using psychological scales and adversarial samples. *Proceedings of the International AAAI Conference on Web and Social Media*, 15, 573–584. <https://doi.org/10.1609/icwsm.v15i1.18085>
- Sathishkumar, R., Karthikeyan, T., Shamsundar, S., & Shamsundar, S. M. (2023). Ensemble text classification with TF-IDF vectorization for hate speech detection in social media. In *2023 International Conference on System, Computation, Automation and Networking (ICSCAN)* (pp. 1–7). IEEE. <http://dx.doi.org/10.1109/ICSCAN58655.2023.10395354>
- Shakeel, N., & Dwivedi, R. K. (2022). A survey on detection of cyberbullying in social media using machine learning techniques. In *Intelligent Communication Technologies and Virtual Mobile Networks, proceedings of ICICV 2022* (pp. 323–340). Springer. https://doi.org/10.1007/978-981-19-1844-5_25
- Shankar, K., Abirami, A., Indira, K., Angeline, C. N., & Shubhavya, K. (2022). Cyberbullying detection in social media using supervised ML and NLP techniques. In *Communication and Intelligent Systems, proceedings of ICCIS 2021* (pp. 817–828). Springer. https://doi.org/10.1007/978-981-19-2130-8_63
- Sharif, O., Hossain, E., & Hoque, M. M. (2021). NLP-CUET@DravidianLangTech-EACL2021: Offensive language detection from multilingual code-mixed text using Transformers. In B. R. Chakravarthi, R. Priyadharshini, A. Kumar M, P. Krishnamurthy, & E. Sherly (Eds.), *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages* (pp. 255–261). Association for Computational Linguistics. <https://aclanthology.org/2021.dravidianlangtech-1.35>
- Sharma, A., & Bhalla, R. (2022). Automatic and advance techniques for hate speech detection on social media: A review. In *2022 Algorithms, Computing and Mathematics Conference (ACM)* (pp. 54–61). IEEE. <https://doi.org/10.1109/ACM57404.2022.00017>
- Sharma, G., Brar, G. S., Singh, P., Gupta, N., Kalra, N., & Parashar, A. (2022). An exploration of machine learning and deep learning techniques for offensive text detection in social media—A systematic review. In D. Gupta, A. Khanna, A. E. Hassanien, S. Anand, & A. Jaiswal (Eds.), *International Conference on Innovative Computing and Communications, proceedings of ICCC 2022* (Vol. 3, pp. 541–559). Springer. http://dx.doi.org/10.1007/978-981-19-3679-1_45
- Sharma, P., & Tiwari, R. K. (2023). Deep learning approach for hate and non hate speech detection in online social media. In *2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS)* (pp. 492–496). IEEE. <https://doi.org/10.1109/ICTACS59847.2023.10390417>
- Singh, N. K., Singh, P., & Chand, S. (2022). Deep learning-based methods for cyberbullying detection on social media. In P. Nand, M. Singh, M. Kaur, & V. Jain (Eds.), *2022 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)* (pp. 521–525). IEEE. <https://doi.org/10.3390/fi15050179>

- Singh, N. M., & Sharma, S. K. (2024). An efficient automated multi-modal cyberbullying detection using decision fusion classifier on social media platforms. *Multimedia Tools and Applications*, 83(7), 20507–20535. <https://doi.org/10.1007/s11042-023-16402-w>
- Singh, R. K., Sanjay, H., SA, P. J., Rishi, H., & Bhardwaj, S. (2023). NLP based hate speech detection and moderation. In *2023 7th International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)* (pp. 1–5). IEEE. <https://doi.org/10.1109/CSITSS60515.2023.10333320>
- Singh, V. K., Ghosh, S., & Jose, C. (2017). Toward multimodal cyberbullying detection. In *CHI EA '17: Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 2090–2099). Association for Computing Machinery. <https://doi.org/10.1145/3027063.3053169>
- Singh, V. K., Huang, Q., & Atrey, P. K. (2016). Cyberbullying detection using probabilistic socio-textual information fusion. In R. Kumar, J. Caverlee, & H. Tong (Eds.), *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 884–887). IEEE. <https://doi.org/10.1109/ASONAM.2016.7752342>
- Sookarah, D., & Ramwodin, L. S. (2022). Combatting online harassment by using transformer language models for the detection of emotions, hate speech and offensive language on social media. In *2022 4th International Conference on Emerging Trends in Electrical, Electronic and Communications Engineering (ELECOM)* (pp. 1–6). IEEE. <https://doi.org/10.1109/ELECOM54934.2022.9965237>
- Suhas Bharadwaj, R., Kuzhalvaimozhi, S., & Vedavathi, N. (2022). A novel multimodal hybrid classifier based cyberbullying detection for social media platform. In R. Šilhavý, P. Šilhavý, & Z. Prokopová (Eds.), *Proceedings of the Computational Methods in Systems and Software* (pp. 689–699). Springer. http://dx.doi.org/10.1007/978-3-031-21438-7_57
- Sultan, D., Mendes, M., Kassenkhan, A., & Akylbekov, O. (2023). Hybrid CNN-LSTM network for cyberbullying detection on social networks using textual contents. *International Journal of Advanced Computer Science and Applications*, 14(9). <https://dx.doi.org/10.14569/IJACSA.2023.0140978>
- Suryawanshi, S., Chakravarthi, B. R., Arcan, M., & Buitelaar, P. (2020). Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In R. Kumar, A. Kr. Ojha, B. Lahiri, M. Zampieri, S. Malmasi, V. Murdock, & D. Kadar (Eds.), *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying* (pp. 32–41). European Language Resources Association (ELRA). <https://aclanthology.org/2020.trac-1.6>
- Tanase, M.-A., Cercel, D.-C., & Chiru, C. (2020). UPB at SemEval-2020 Task 12: Multilingual offensive language detection on social media by fine-tuning a variety of BERT-based models. In A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, & E. Shutova (Eds.), *Proceedings of the Fourteenth Workshop on Semantic Evaluation* (pp. 2222–2231). International Committee for Computational Linguistics. <https://doi.org/10.18653/v1/2020.semeval-1.296>
- Thapa, S., Shah, A., Jafri, F., Naseem, U., & Razzak, I. (2022). A multi-modal dataset for hate speech detection on social media: Case-study of Russia-Ukraine conflict. In A. Hürriyetoğlu, H. Tanev, V. Zavarella, & E. Yörük (Eds.), *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)* (pp. 1–6). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.case-1.1>
- Themeli, C., Giannakopoulos, G., & Pittaras, N. (2019). A study of text representations for hate speech detection. In A. Gelbukh (Ed.), *International Conference on Computational Linguistics and Intelligent Text Processing* (pp. 424–437). Springer. https://doi.org/10.1007/978-3-031-24340-0_32

- Thenmozhi, D., Pr, N., Arunima, S., & Sengupta, A. (2020). Ssn_nlp at SemEval 2020 Task 12: Offense target identification in social media using traditional and deep machine learning approaches. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation* (pp. 2155–2160). International Committee for Computational Linguistics. <https://doi.org/10.18653/v1/2020.semeval-1.286>
- Thenmozhi, D., Sharavanan, S., Chandrabose, A., & Chandrabose, A. (2019). SSN_NLP at SemEval-2019 Task 6: Offensive language identification in social media using traditional and deep machine learning approaches. In J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki, & S. M. Mohammad (Eds.), *Proceedings of the 13th International Workshop on Semantic Evaluation* (pp. 739–744). International Committee for Computational Linguistics. <https://doi.org/10.18653/v1/S19-2130>
- Van Bruwaene, D., Huang, Q., & Inkpen, D. (2020). A multi-platform dataset for detecting cyberbullying in social media. *Language Resources and Evaluation*, 54(4), 851–874. <https://doi.org/10.1007/s10579-020-09488-3>
- Vidgen, B., Hale, S., Guest, E., Margetts, H., Broniatowski, D., Waseem, Z., Botelho, A., Hall, M., & Tromble, R. (2020). Detecting East Asian prejudice on social media. In S. Akiwowo, B. Vidgen, V. Prabhakaran, & Z. Waseem (Eds.), *Proceedings of the Fourth Workshop on Online Abuse and Harms* (pp. 162–172). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.alw-1.19>
- Vora, D., Mukherjee, A., Repaka, S., Das, S., & Ingle, S. (2023). Multimodal cyberbullying detection on social media: Review and challenges. In *2023 International Conference on Integration of Computational Intelligent System (ICICIS)* (pp. 1–8). IEEE. <http://dx.doi.org/10.1109/ICICIS56802.2023.10430250>
- Wang, K., Xiong, Q., Wu, C., Gao, M., & Yu, Y. (2020). Multi-modal cyberbullying detection on social networks. In *2020 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8). IEEE. <https://doi.org/10.1109/IJCNN48605.2020.9206663>
- Waseem, Z. (2016). Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter. In D. Bamman, A. S. Doğruöz, J. Eisenstein, D. Hovy, D. Jurgens, B. O'Connor, A. Oh, O. Tsur, & S. Volkova (Eds.), *Proceedings of the First Workshop on NLP and Computational Social Science* (pp. 138–142). Association for Computational Linguistics. <https://doi.org/10.18653/v1/W16-5618>
- Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In J. Andreas, E. Choi, & A. Lazaridou (Eds.), *Proceedings of the NAACL Student Research Workshop* (pp. 88–93). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N16-2013>
- Wijesiriwardene, T., Inan, H., Kursuncu, U., Gaur, M., Shalin, V. L., Thirunarayan, K., Sheth, A., & Arpinar, I. B. (2020). Alone: A dataset for toxic behavior among adolescents on Twitter. In S. Aref, K. Bontcheva, M. Braghieri, F. Dignum, F. Giannotti, F. Grisolia, & D. Pedreschi (Eds.), *Social Informatics: 12th International Conference, SocInfo 2020, Pisa, Italy, October 6–9, 2020, proceedings* (pp. 427–439). Springer. https://doi.org/10.1007/978-3-030-60975-7_31
- Wright, M. F. (2021). Cyberbullying: Definition, behaviors, correlates, and adjustment problems. In *Encyclopedia of information science and technology* (5th ed., pp. 356–373). IGI Global. <http://dx.doi.org/10.4018/978-1-7998-3479-3.ch026>
- Xiang, T., MacAvaney, S., Yang, E., & Goharian, N. (2021). ToxCCLn: Toxic content classification with interpretability. In O. De Clercq, A. Balahur, J. Sedoc, V. Barriere, S. Tafreshi, S. Buechel, & V. Hoste (Eds.), *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (pp. 1–12). Association for Computational Linguistics. <https://aclanthology.org/2021.wassa-1.1>

- Xingyi, G., & Adnan, H. (2024). Potential cyberbullying detection in social media platforms based on a multi-task learning framework. *International Journal of Data and Network Science*, 8(1), 25–34. <https://doi.org/10.5267/j.ijdns.2023.10.021>
- Yao, M. (2019). Robust detection of cyberbullying in social media. In L. Liu & R. White (Eds.), *WWW '19: Companion proceedings of the 2019 World Wide Web Conference* (pp. 61–66). Association for Computing Machinery. <https://doi.org/10.1145/3308560.3314196>
- Yao, M., Chelmiss, C., & Zois, D.-S. (2019). Cyberbullying ends here: Towards robust detection of cyberbullying in social media. In L. Liu & R. White (Eds.), *WWW '19: The World Wide Web Conference* (pp. 3427–3433). Association for Computing Machinery. <https://doi.org/10.1145/3308558.3313462>
- Yi, P., & Zubiaga, A. (2023a). Learning like human annotators: Cyberbullying detection in lengthy social media sessions. In Y. Ding, J. Tang, J. Sequeda, L. Aroyo, C. Castillo, & G. Houben (Eds.), *WWW '23: Proceedings of the ACM Web Conference 2023* (pp. 4095–4103). Association for Computing Machinery. <https://doi.org/10.1145/3543507.3583873>
- Yi, P., & Zubiaga, A. (2023b). Session-based cyberbullying detection in social media: A survey. *Online Social Networks and Media*, 36, 100250. <https://doi.org/10.1016/j.osnem.2023.100250>
- Yin, W., & Zubiaga, A. (2021). Towards generalisable hate speech detection: A review on obstacles and solutions. *PeerJ Computer Science*, 7, 1–38. <https://doi.org/10.7717/PEERJ-CS.598>
- Yuan, L., Wang, T., Ferraro, G., Suominen, H., & Rizoïu, M.-A. (2023). Transfer learning for hate speech detection in social media. *Journal of Computational Social Science*, 6(2), 1081–1101. <https://doi.org/10.1007/s42001-023-00224-9>
- Zampieri, N., Illina, I., & Fohr, D. (2021). Multiword expression features for automatic hate speech detection. In E. Métais, F. Mezziane, H. Horacek, & E. Kapetanios (Eds.), *Natural Language Processing and Information Systems. 26th International Conference on Applications of Natural Language to Information Systems, NLDB 2021, Saarbrücken, Germany, June 23–25, 2021, proceedings* (pp. 156–164). Springer. https://doi.org/10.1007/978-3-030-80599-9_14