Antisocial Behavior Prediction: A Survey and Practical Guide

Anaïs Ollagnier
Université Côte d'Azur, Inria, CNRS, I3S / 930 route des Colles, BP 145,
Sophia Antipolis Cedex, 06903, France
ollagnier@i3s.unice.fr

Abstract

Antisocial behavior (ASB) on social media encompasses online behaviors that harm individuals, groups, or platform ecosystems, including hate speech, harassment, cyberbullying, trolling, and coordinated abuse. While most prior work has focused on detecting harm after it occurs, a growing body of research on ASB prediction seeks to forecast future harmful outcomes before they materialize, including—but not limited to—hate-speech diffusion, conversational derailment, and user recidivism. However, this emerging field remains fragmented, with limited conceptual grounding and few integrative frameworks. This paper establishes a foundation for ASB prediction by introducing a structured taxonomy spanning temporal, structural, and behavioral dimensions. Drawing on 49 machine learning studies identified through a literature review, we map predictive goals to datasets, modeling choices, and evaluation practices, and identify key challenges, including the lack of standardized benchmarks, the dominance of text-centric representations, and trade-offs between accuracy and interpretability. We conclude by outlining actionable directions toward more robust, generalizable, and responsible ASB prediction systems.

1 Introduction

Social media platforms have reshaped how people consume information and engage in public discourse (Kahn and Kellner, 2004; Brown et al., 2007; Quattrociocchi et al., 2014). While these systems enable large-scale participation (Criado et al., 2013; Etta et al., 2022), they also facilitate harmful dynamics such as misinformation (Vicario et al., 2016), echo chambers (Cinelli et al., 2021), and hostility (Cheng et al., 2017; Saveski et al., 2021). These conditions often lead to antisocial behavior (ASB)—understood here as online behaviors that disregard or violate the rights and well-being of others, including hate speech, harassment, trolling, and related forms of harmful interaction (Ollagnier et al., 2023a,b; Chowdhury et al., 2019)—with serious individual and societal consequences (Parent et al., 2019; Saha et al., 2019).

While harmful content detection has been extensively studied, ASB prediction—defined as the computational modeling of future harmful behaviors—remains underexplored, as evidenced by our systematic review identifying only 49 machine-learning studies since 2014 (Appendix A). The field has nonetheless accelerated sharply: from isolated contributions in 2014–2017 (one paper per year), to modest growth in 2018–2019, and a marked surge from 2020 onward. This expansion reflects heightened awareness of online harms, broader access to social media data, rapid advances in NLP and machine learning (ML), and growing regulatory and societal pressures—including the EU Digital Services Act[1] (2025), the UN's International Day for Countering Hate Speech[2], and grassroots counter-speech movements such as #IamHere[3]. Importantly, this evolution is not only quantitative but also qualitative: since 2020, ASB prediction has gained visibility in top-tier venues such as EMNLP, NAACL, ICWSM, KDD, and IEEE Transactions on Network Science and Engineering, reflecting growing methodological sophistication and interdisciplinary engagement. Yet the field remains fragmented, with limited conceptual grounding and few integrative frameworks.

To address this gap, we draw on our systematic literature review of machine-learning–oriented studies to propose a structured synthesis of progress to date, identify best practices, and highlight current limitations alongside open challenges. After introducing background on ASB (§2), the survey is structured around three dimensions: task design (§3), data selection (§4), and evaluation

[1] EU Digital Services Act
[2] UN Strategy and Plan of Action on Hate Speech
[3] #IamHere International

The Proceedings for the 15th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA 2026), pages 235–251
March 29, 2026 ©2026 Association for Computational Linguistics

(§5). We then discuss key challenges and promising research directions (§6), with full details of the review methodology provided in Appendix A.

2 Background

2.1 Definitions

Hate speech has traditionally occupied a central place in academic and legal debates, yet it represents only one subset of abusive language, typically defined by its *function* (e.g., incitement) and *target* (identity-based groups) (Poletto et al., 2021). More broadly, ASB originates in social and clinical psychology, where it denotes a persistent pattern of disregard for or violation of the rights and well-being of others, as formalized in the DSM-5 (American Psychiatric Association, 2013). In online contexts, we use ASB not as a clinical diagnosis but as a behavioral construct capturing observable actions that harm individuals, groups, or platform ecosystems. Adopting ASB as an umbrella term reduces the conceptual fragmentation introduced by overlapping notions such as *online abuse*, *toxic language*, and *cyberaggression* (Poletto et al., 2021; Alkomah and Ma, 2022; Jahan and Oussalah, 2023). Drawing on social psychology, platform governance frameworks (Meta, 2022), and computational social science (Gruzd et al., 2020; Haythornthwaite, 2023), we define online ASB along three dimensions: **personal harms** (e.g., harassment, cyberbullying), **group-directed harms** (e.g., identity-based hate, stereotyping), and **environmental disruptions** (e.g., trolling, misinformation, coordinated harassment). This definition integrates both linguistic content and behavioral patterns, providing a clear foundation for ASB prediction.

2.2 Taxonomy of ASB computational tasks

To address the absence of a structured framework for ASB prediction, we propose a taxonomy grounded in the collected literature. Adapting the thematic analysis framework of Zhou et al. (2022) to our goals, we organize predictive tasks along two complementary dimensions: their *temporal orientation*, which specifies when harm is anticipated (e.g., emergence, escalation, or diffusion), and their *operational purpose*, which clarifies why it is modeled (e.g., moderation, risk assessment, or intervention). This two-dimensional taxonomy yields five task categories, illustrated below with scenarios (see Appendix B for a detailed overview).

Harm Emergence. This task forecasts whether an initially civil interaction—such as a thread, comment exchange, or reply chain—will develop into harmful behavior. Models look for subtle precursors like escalating rhetoric, sentiment shifts, or power dynamics to anticipate toxicity before explicit harmful content appears.

> **User A:** I don't think the article should call the policy "controversial."
> **User B:** Reliable sources use that term, so it seems justified.
> **User A: You're just pushing your agenda.**
> **User B:** That's unfair—please stick to the evidence.
> **User A:** People like you ruin every discussion here!
> **Observed pattern:** The bolded remark introduces an adversarial tone in an otherwise civil exchange, signaling a possible shift toward personal attacks.

Harm Propagation. Rather than predicting if harm will occur, this task models how existing harmful content will spread across networks. It estimates the reach, speed, and diffusion trajectory of toxic material, identifying which hateful posts are likely to go viral or which communities will amplify them.

> **User A:** Those people are ruining our country. Share this so everyone knows the truth!
> **User B:** Exactly! Everyone in my group needs to see this. (shares to a community group)
> **User C:** This is going viral—just shared it with 10 friends.
> **User D:** Spreading fast—check out this hateful meme. (reshare with image)
> **Observed pattern:** Rapid resharing across groups indicates accelerating virality, pointing to high exposure risk if left unchecked.

Early Harm Detection. This task emphasizes rapid intervention, predicting harmful behavior from minimal early signals—often just the first few messages in a thread. The aim is to enable triage before harm escalates.

> **User A:** "Does anyone have tips for first-time job seekers?"
> **User B:** "Maybe start by not being so lazy."
> **User C:** "Yeah, people like you just want handouts."
> **Observed pattern:** Even at the outset, the exchange shows dismissive and stereotyping remarks, making it a likely candidate for escalation into bullying.

Behavioral Risk. Here, the focus shifts from content to individuals. Models assess whether a user is likely to engage in, or become a target of, ASB. They build risk profiles from prior behavior, lin-

guistic cues, and engagement patterns.

Perpetration risk (user likely to engage in ASB).
User X (Mon): "Debate's heated, but let's keep it civil."
User X (Wed): "Some people just don't belong here..."
User X (Thu): joins *@HardLineForum*; likes several posts with coded slurs.
User X (Fri): "They're parasites—open your eyes."
Observed pattern: A progression from neutral to exclusionary to dehumanizing language, coupled with extremist ties, reveals growing perpetration risk.

Proactive Moderation. Models in this category support interventions before harm materializes. They predict harmfulness at the moment of posting, suggest safer edits, or prioritize items for review.

User A (drafting a comment): "This journalist is a lying idiot—don't trust anything she writes."
Observed pattern: At the drafting stage, the system flags personal insults and polarizing phrasing, offering a constructive rewording before publication.

3 Step 1: Design your task

The first step in building an ASB prediction system is to determine which task(s) you aim to address. Tasks vary depending on whether you predict categories or continuous outcomes, and at which level—actors or content—they occur.

3.1 Classifying ASB

Classification tasks assign a discrete label to an observation to characterize the phenomenon. Schemes range from binary (e.g., toxic vs. non-toxic content (Kennedy et al., 2020a; Saveski et al., 2021) or banned vs. not banned users (Cheng et al., 2015)) to multi-class setups, such as high/medium/low incivility (Yu et al., 2024).

Actor classification. Actor-level tasks predict user-level risks and behaviors. Models estimate whether a user will be banned (Cheng et al., 2015; Hickey et al., 2025), join or escalate activity in harmful communities, spread toxic narratives (Irani et al., 2021), or generate abusive replies (Tsantarliotis et al., 2017). Some studies also forecast longitudinal trends, such as users' contributions to future harm or their recidivism risk (Song et al., 2025).

Content classification. Content-level tasks focus on posts, threads, or conversations, predicting whether an interaction will lead to or amplify harm. Examples include forecasting whether a comment thread will derail into personal attacks (Nonaka and Yoshida, 2025; Altarawneh et al., 2023), whether a post will trigger cyberbullying (Hosseinmardi et al., 2015), or whether a tweet will spark widespread toxic replies (Al-Merekhi et al., 2020).

3.2 Regressing ASB

Regression tasks predict continuous measures of harm offering finer-grained signals that reflect intensity or likelihood.

Actor regression. At the actor level, models forecast continuous user outcomes, such as how a user's harmful behavior may evolve, their probability of returning after moderation, or their trajectory of engagement with toxic communities (Chelmis and Yao, 2019; Levy et al., 2022).

Content regression. At the content level, models estimate harm scores for posts or conversations, such as the predicted intensity of harassment in a thread (Dahiya et al., 2021; Meng et al., 2023), expected hate triggered by a new post, or diffusion strength of harmful narratives (Han et al., 2021).

Classification often over-simplifies behaviors, producing false positives on strong but benign language or false negatives on subtle harms. Mitigation: adopt multi-label or hierarchical schemes, add contextual metadata, and treat annotator disagreement as informative (Lambert et al., 2022; Kim et al., 2025).
Regression suffers from label subjectivity and drift: harm scores vary with annotators, time, and platform norms, leading to uncalibrated outputs. Mitigation: apply calibration techniques, use multi-rater judgments, and train on longitudinal data to capture evolving behaviors (Meng et al., 2023; Gajo et al., 2023; Alharthi et al., 2025).

3.3 Timing of Prediction: Ex-ante vs. Peeking

Beyond deciding what to predict, you must also decide **when** to predict. A detailed overview of feature-based model strategies and formulations is provided in Appendix C.

Ex-ante prediction Ex-ante models predict harm *at* or *before* content is posted, using only static features such as text, images, user metadata, or network structure. For example, some models predict whether an Instagram post will trigger cyberbullying using only the image, caption, and poster's social graph (Hosseinmardi et al., 2015, 2016). Context-aware variants add structural or personality features while remaining anticipatory (Han et al., 2021; Liu et al., 2024a; Solovev and Pröllochs, 2023). Ex-ante strategies are especially com-

mon in classification tasks because they enable immediate moderation.

Peeking prediction. Peeking strategies allow models to observe early interactions—such as initial replies—before predicting. This often improves accuracy but introduces delay, making them ideal for early warning systems. Examples include forecasting hate intensity in threads or detecting cyberbullying after initial comment patterns (López-Vizcaíno et al., 2023a; Solovev and Pröllochs, 2023). Progressive approaches update predictions as data arrives and can stop once confidence is sufficient (Al-Merekhi et al., 2020; Meng et al., 2023).

> Ex-ante and peeking strategies present distinct trade-offs for ASB prediction. Ex-ante approaches enable instant action but rely only on static features, often reducing accuracy and increasing false positives. In contrast, peeking strategies achieve higher accuracy by incorporating early conversational signals but delay intervention, risking harmful escalation. A practical solution, inspired by principles of information diffusion, is an a priori strategy: treat ex-ante predictions as low-confidence hypotheses (e.g., soft warnings or reduced reach) that are progressively refined as interaction evidence accumulates (Zhou et al., 2022). This hybrid approach ensures moderation that is both timely and adaptive.

4 Step 2: Select the data

After defining the task, the next step is to decide whether to build a new dataset or rely on existing resources. This section synthesizes, based on our systematic review of ASB prediction studies, the main data collection procedures reported in the literature and the characteristics of publicly available datasets employed.

4.1 Collecting your own data

The following analysis framework builds on the collection strategies outlined by Bonaldi et al. (2024).

Crawling. Crawling refers to the automated collection of real content from online platforms such as social networking sites (e.g., YouTube (Kennedy et al., 2020a), Twitter/X (Saveski et al., 2021; Alharthi et al., 2025)), discussion forums (e.g., Wikipedia Talk Pages (Altarawneh et al., 2023; Nonaka and Yoshida, 2025), Reddit (Yu et al., 2024; Hickey et al., 2025)), or news platforms (e.g., CNN, NYT (Han et al., 2021; Liu et al., 2024a)). This approach is particularly attractive because it can yield large-scale and diverse data samples, capturing a wide variety of contexts and interaction styles. However, the way content is retrieved imposes important constraints. Most crawling procedures either (i) rely on selected "representative" keywords to filter and retrieve relevant posts, or (ii) prioritize content already highlighted by platform-specific popularity metrics (e.g., top-level comments on Reddit or trending videos on YouTube). While the first method helps to collect data that directly reflects known forms of ASB, it tends to suffer from limited lexical coverage and reduced sensitivity to alternative rhetorical strategies. Moreover, in the context of ASB prediction, even non-aggressive or neutral behaviors can evolve into toxic interactions (Al-Merekhi et al., 2020; Solovev and Pröllochs, 2023), making keyword-based sampling insufficient. Conversely, using more generic content without pre-filtering can expand lexical diversity and contextual coverage, but it introduces a new layer of complexity. Such content must undergo a meticulous curation process that depends heavily on each platform's structure and interaction dynamics—for instance, retweet cascades on Twitter or multi-turn, threaded exchanges on Reddit. This complexity introduces additional limitations: platform-specific interaction patterns can bias the data, the resulting dataset may present imbalanced or sparse evidence of the targeted phenomena, and meaningful cross-platform comparisons become substantially more challenging. Finally, crawled data is often ephemeral—posts and accounts can be deleted or modified—hindering long-term reproducibility and making it difficult for future researchers to replicate or validate findings (Klubicka and Fernández, 2018; Florio et al., 2020).

Hybrid and fully automated collection. Hybrid approaches combine automated collection methods (such as web crawling, platform APIs, or bulk dataset ingestion) with manual interventions in the data collection pipeline (such as language adaptation or dataset post-editing). This allows researchers to scale up data acquisition while still ensuring that the collected material matches the domain and task requirements. In Kim et al. (2025), for example, an existing dataset was automatically ingested and supplemented with additional Korean SNS conversations. To increase coverage, corpora from AI Hub were integrated, with only the general and sexual conversation subsets retained.

Crowdsourcing and **nichesourcing** are established data collection strategies in related domains dealing with abusive language (Bonaldi et al., 2024;

Ollagnier et al., 2024), but, to the best of our knowledge, they have not yet been applied to ASB prediction. Crowdsourcing relies on non-expert annotators to generate, rephrase, or simulate ASB content and practices. This approach enables the collection of large quantities of data at relatively low cost. However, the resulting material often reflects simpler argument patterns and limited rhetorical diversity (Qian et al., 2019), which restricts its usefulness for modeling nuanced precursors of ASB. Nichesourcing, by contrast, relies on expert contributors—such as specialists, moderators, or trained practitioners—to create or simulate ASB content. This strategy yields data of higher quality, with richer syntactic variety and greater semantic complexity, particularly valuable for capturing subtle cues or early warning signs of harmful behavior. Its main drawback is high cost and limited scalability, which make it difficult to assemble large datasets (Chung et al., 2019).

> Crawling remains the most common procedure for collecting ASB data in the wild. However, crowdsourcing and nichesourcing, while not widely explored for this purpose, offer promising alternatives. Well-designed crowdsourcing tasks, such as scenario-based role-playing exercises (Ollagnier, 2024), could elicit a wide range of interpersonal behaviors that may escalate into toxicity, thereby enriching the triggers considered in ASB prediction models. Likewise, nichesourcing—drawing on domain experts—could capture rare or highly specific interaction patterns that precede harmful outcomes, providing high-quality examples to strengthen computational modeling.

4.2 Choosing from existing datasets

Rather than collecting new data, a practical alternative is to leverage existing datasets. We review the publicly available resources identified in our survey of ASB prediction, summarized in Table 1.

Shape of the interactions. A crucial dimension to consider is the structure of interactions captured in a dataset, as this directly determines the types of ASB tasks it can support. Some datasets focus on *isolated content*, such as individual tweets or toxic news comments, which are treated as independent units without any interaction history; these are commonly used to address tasks such as harm emergence, behavioral risk, and proactive moderation (Kennedy et al., 2020a; Han et al., 2021; Irani et al., 2021). Others capture *local interaction* patterns, for instance short-lived exchanges like immediate replies or audience responses, which are often leveraged in behavioral risk or early harm detection tasks within platforms such as Reddit (Yu et al., 2024; Song et al., 2025; Hickey et al., 2025). A third group encompasses *conversational flow* datasets that preserve multi-turn, threaded discussions, as seen in Reddit threads (Bao et al., 2021) or Wikipedia Talk Pages (Altarawneh et al., 2023; Nonaka and Yoshida, 2025), making them particularly suitable for modeling discourse evolution and for tasks such as early harm detection or proactive moderation. Finally, some datasets reflect *networked spread*, modeling how harmful content propagates through user networks—for example, via retweet cascades on Twitter—thereby enabling studies of diffusion processes and influence dynamics relevant to harm emergence or harm propagation (Saveski et al., 2021; Dahiya et al., 2021).

Targets of hate. As in related domains dealing with abusive language, ASB prediction datasets often focus on phenomena directed toward specific, and often multiple, targeted minorities. For example, Kennedy et al. (2020a) propose a rich annotation scheme with eight broad target identity groups (race/ethnicity, religion, national origin or citizenship, gender, sexual orientation, age, disability, and political ideology) and further refines these into 42 specific subgroups. In a complementary direction, Lambert et al. (2022) classify deleted comments based on normative violations, including misogynistic content, hate speech targeting racial or sexual minorities, attacks over opposing political views (e.g., pro- or anti-Trump), as well as abusive comments directed at moderators or individuals, such as name-calling or accusations that others are "too sensitive". Other datasets capture more niche contexts: Meng et al. (2023) focus on anti-Asian hate, reflecting culture- and ethnicity-related hostility, while Gajo et al. (2023) examine incel forums, a setting where hate speech manifests through misogynistic and gendered discourse.

Types of hate addressed. Beyond specifying targets, many datasets classify hate speech by the *form* or *type* of negative reference expressed toward those groups. Building on taxonomies such as Chung et al. (2019), these categories capture whether the utterance manifests as insults, stereotyping, dehumanization, threats, or calls to violence. For instance, Kennedy et al. (2020a) annotate ten fine-grained dimensions, including (dis)respect, humiliation, inferior status, violence, genocide, and attack/defense. In contrast, broader schemes like Saveski et al. (2021) emphasize

Dataset	Size	Source(s)	Lang.	Target	Add. Info
Isolated Content					
Kennedy et al. (2020a)	135,556	YouTube, Twitter, Reddit	EN	✓	✓
Han et al. (2021)	5,571	Patch website	EN	×	✓
Irani et al. (2021)	40,000	Twitter	EN/ES	×	×
Lambert et al. (2022)	5,059	Reddit	EN	✓	✓
Solovev and Pröllochs (2023)	691,237	Twitter	EN	×	✓
Liu et al. (2024a)	5,571	Patch website	EN	×	✓
Conversational Flow					
Bao et al. (2021)	2,388	Reddit	EN	×	×
Meng et al. (2023)	890,372	Twitter	EN	✓	✓
Altarawneh et al. (2023)	11,030	WTP/CMV	EN	×	✓
Gajo et al. (2023)	5,503	Incels.is / Il forum dei brutti	EN/IT	✓	✓
Alharthi et al. (2025)	2,369	Twitter	EN	×	✓
Kim et al. (2025)	15,000	Chat platforms	KO	✓	×
Nonaka and Yoshida (2025)	11,030	WTP/CMV	EN	×	✓
Local Interaction					
Yu et al. (2024)	34,115	Reddit	EN	×	✓
Song et al. (2025)	5,724	Reddit	EN	×	✓
Hickey et al. (2025)	50,003	Reddit	EN	×	✓
Networked Spread					
Saveski et al. (2021)	1.18M	Twitter	EN	×	✓
Dahiya et al. (2021)	4,533	Twitter	EN	×	×

Table 1: Overview of publicly available datasets organized by interaction type.

generic toxicity markers such as rudeness or disrespect. Domain-specific datasets also highlight particular forms: Liu et al. (2024a) and Han et al. (2021) classify hate-related news comments into subtypes such as homicide or kidnapping, while Alharthi et al. (2025) provide aggregate indicators including counts of insults, profanity, threats, and identity attacks. Other resources focus on structural consequences of hate, for example examining how misogyny reinforces systemic oppression (Gajo et al., 2023; Hickey et al., 2025).

Languages. Most reported datasets are based on English-language corpora, with only a few extending to other languages, such as Korean (Kim et al., 2025). Others take a bilingual perspective—e.g., English–Spanish (Irani et al., 2021) and English–Italian (Gajo et al., 2023)—enabling cross-lingual and comparative analyses.

Additional information. Several datasets incorporate rich auxiliary signals that characterize conversations and users in greater detail. For example, Nonaka and Yoshida (2025) provide pairs of hate speech and counterspeech, further annotated according to the hater's subsequent behavior (no reentry, hateful reentry, or non-hateful reentry). Other resources, such as Solovev and Pröllochs (2023), enrich each instance with a broad set of metadata, including text complexity, temporal information, engagement indicators, and lexical features (e.g., moral and emotional content). Likewise, Yu et al. (2024) supply conversation-level attributes such as hate scores, subreddit tags, and reply-level labels. Some datasets also document annotator characteristics and target groups: for instance, Kennedy et al. (2020a) report annotator demographics and specify 40 fine-grained target subgroups.

> Dataset selection is primarily guided by the target task, as different tasks require specific features (e.g., conversational flow for harm propagation). However, selection should also consider methodological factors such as construct validity, biases, and evidential support. To aid in this process, Liu et al. (2024b) offer guiding questions that clarify what capabilities a dataset measures. Practical aspects like dataset size and coverage also matter: few-shot prompting may suffice for small datasets, while fine-tuning typically demands larger corpora. Rich metadata—such as multiple annotator perspectives—can be especially valuable for subjective tasks, as explored in Frenda et al. (2025). Finally, transparent annotation procedures are critical, as they directly influence reliability and validity (see Appendix D). Beyond task-specific corpora, datasets from related domains (e.g., toxic language detection, cyberbullying, misinformation) should be systematically scrutinized, as their reuse can expand coverage and enhance comparability across ASB prediction research.

5 Step 3: Evaluate

In this step, we review how the literature evaluates ASB prediction models, focusing on the task

types outlined in 2.2. Since these tasks are most often formulated as either classification or regression problems, the choice of evaluation strategies naturally varies with the formulation.

5.1 Evaluating classification

Most ASB prediction studies formulate tasks such as early harm detection, harm emergence prediction, or proactive moderation as classification problems (Saveski et al., 2021; Lambert et al., 2022; Hickey et al., 2025). When ground-truth labels are available, models are commonly assessed using *accuracy*, *precision*, *recall*, and *F1-score*, often complemented by *ROC curves* and the *area under the ROC curve* to capture performance across decision thresholds. Several works report *macro-averaged metrics* (e.g., macro F1 or macro precision) to handle the class imbalance typical of ASB datasets, while others employ *latency-aware measures*, such as latency-weighted F1 (López-Vizcaíno et al., 2021) or time-aware precision (López-Vizcaíno et al., 2023a), to better capture the timing sensitivity required in early detection tasks. In ranking-based evaluation settings, measures such as *MAP@20* or *HITS@20* are used to assess whether models correctly prioritize high-risk content or users (Makkar and Chakraborty, 2020; Masud et al., 2021). *Complementary cumulative distribution functions* have also been applied to visualize performance across thresholds (Solovev and Pröllochs, 2023). Finally, some studies benchmark models against human judgment to verify performance; for example, Kim et al. (2025) introduce the *Human-to-Model Ratio*, which evaluates how closely a model's detection speed aligns with that of human annotators in early detection scenarios. Complementing these quantitative measures, qualitative error analysis is often employed to reveal systematic weaknesses or biases in model predictions (Yu et al., 2024; Song et al., 2025).

5.2 Evaluating Regression

When ASB prediction tasks are framed to produce continuous outputs evaluation typically relies on metrics that capture both correlation and error. Correlation-based measures, such as the Pearson or Spearman correlation coefficients (Dahiya et al., 2021; Meng et al., 2023), assess how well the predicted scores align with the ground truth trends, providing insight into the model's ability to rank or follow underlying dynamics. Complementarily, error-based measures such as Root Mean Square Error, Mean Absolute Percentage Error, Symmetric Mean Absolute Percentage Error, or L2 loss quantify the magnitude of prediction errors (Wu et al., 2022; Hebert et al., 2023), thus revealing how close the predictions are in absolute terms. Some studies additionally report coefficients of determination (R-squared) to capture the proportion of variance explained by the model (Han et al., 2021; Liu et al., 2024a), or information-theoretic criteria such as log-likelihood scores and the Bayesian Information Criterion to balance goodness of fit with model complexity (Levy et al., 2022).

> ASB prediction research remains constrained by an overreliance on standard classification and regression metrics and by the lack of context-sensitive evaluation protocols. While such metrics are useful, they are insufficient in high-stakes settings where error type, fairness, and user trust matter. Building on principles for systematic benchmark design (Liu et al., 2024b), best practices for evaluating ASB prediction should include cost-sensitive and asymmetric error measures, calibration and subgroup analyses, and human-centered criteria such as perceived harm and legitimacy (Olteanu et al., 2017). Metrics drawn from adjacent domains—such as diffusion or reach-based measures—can further capture the temporal and social impact of predicted harms (Zhou et al., 2022).

6 Open Challenges in ASB Prediction

Standardized Benchmarking. ASB prediction lacks standardized benchmarks and evaluation protocols, both globally and within predictive settings. Current tasks are underspecified, rely on heterogeneous datasets, and neglect cultural and pragmatic variation, which undermines reproducibility and cross-platform generalization. To address this, the field must move beyond ad hoc evaluations toward linguistically and culturally grounded frameworks that account for the temporal, structural, and ethical dimensions of ASB. Promising directions include meta-collection of existing resources (Piot et al., 2024) and the adaptation of Evidence-Centered Benchmark Design (Liu et al., 2024b), enabling systematic evidence collection about model capabilities. Incorporating insights from linguistics (e.g., speech act theory, relevance theory, dynamic and epistemic logic) and cultural psychology (e.g., the GLOBE framework (Karinshak et al., 2024)) would enhance both the modeling and the evaluation of a model's ability to capture communicative intentions, conversational dynamics, and cultural variation. Developing such benchmarks is critical for protocols that are rigorous, comparable, and

sensitive to the diversity of online harms. Lessons from related shared tasks—such as *eRisk Lab*[4] and *PAN Author Profiling*[5]—show how community-driven evaluations can foster standardization, comparability, and faster progress toward deployable solutions in ASB prediction.

Multimodal and Pragmatic Representation Learning. Although our review considered ML studies broadly, ASB prediction remains dominated by text-centric features such as lexical cues or sentiment markers, often overlooking the multimodal and pragmatic dimensions of online interactions. As noted in Section 4, even neutral or non-aggressive behaviors can escalate into toxicity, underscoring the need to model evolving interaction dynamics (Ollagnier, 2024). Harmful behavior also frequently manifests through images, videos, memes, or subtle pragmatic cues such as politeness strategies, idioms, and conversational flow (Fried et al., 2023). Addressing these phenomena requires advances beyond current text-only pipelines, including the encoding of pragmatic knowledge, the integration of heterogeneous information networks linking users, posts, and contexts, and the development of graph-based architectures capable of combining multimodal and discourse-level signals (Wang et al., 2024). In parallel, recent progress in LLMs offers an additional research avenue, as such models provide richer contextual representations that may support predictive modeling when appropriately grounded (Albladi et al., 2025). Progress across these complementary directions is essential for developing ASB prediction models that are not only accurate but also sensitive to the complex and gradual ways in which harmful behavior develops and escalates.

Interpretability and Human Oversight. Modern ASB prediction models—such as transformers, GNNs, or multimodal systems—achieve strong performance but remain opaque (Mathew et al., 2021). In predictive settings, where errors directly impact user rights and content visibility, such opacity poses serious challenges to accountability. Regulations like the GDPR stress a "right to explanation" (Council, 2016), making transparency essential to avoid unfair sanctions or missed harms. Existing explainability methods provide limited value: they often reduce to feature importance, ignore cultural and linguistic nuance, and lack consistent evaluation (Nguyen et al., 2021). A key challenge is to design explanation frameworks that incorporate pragmatic reasoning, cultural grounding, and user feedback. Human-in-the-loop oversight can improve transparency and reduce bias. Progress requires interactive workflows, metrics for human–AI alignment, and governance models that ensure fairness and trust in high-stakes moderation (Kotarcic et al., 2022).

Global Challenges in ASB Prediction. Across the literature, several recurring challenges echo broader concerns in tackling ASB. First, **language and cultural diversity** remains limited, with most datasets English-centric and poorly generalizing across regions and communities (Alkomah and Ma, 2022). Second, the **granularity of labeling** is often too coarse, obscuring subtle forms of abuse such as implicit hate or trolling; richer annotation schemes (e.g., multi-label taxonomies, continuous ratings, or annotator demographics) would enable more precise modeling (Vidgen et al., 2019; Kennedy et al., 2020a; Kirk et al., 2022). Third, **data scope and representativeness** are constrained by small, biased datasets that underrepresent marginalized voices and multimodal contexts, as supported in Table 1. Finally, the problem of **evolving behavior and concept drift** means that models trained on static corpora struggle against rapidly changing discourse, slang, and evasion tactics (Florio et al., 2020). Together, these challenges highlight the need for multilingual, fine-grained, cross-platform, representative, and adaptive approaches if ASB prediction is to be both robust and socially responsible.

7 Conclusion

We presented an in-depth review of the emerging field of ASB prediction, introducing a taxonomy that structures existing ML research into five core task types, together with guidance on task design, dataset curation, and evaluation. We further identified key open challenges, including the creation of standardized benchmarks, advances in multimodal and pragmatic representation learning, and the development of explainable, human-centered systems. Our aim is to provide both a consolidated overview of current progress and a foundation for future work, offering a structured entry point for researchers, developers, and practitioners seeking to advance the predictive understanding and mitigation of harmful online behavior.

[4]https://erisk.irlab.org/

[5]https://pan.webis.de/clef21/pan21-web/author-profiling.html

8 Limitations

To an external observer, the number of papers reviewed in this study may seem limited. However, this reflects both the emerging status of ASB prediction as a distinct research subfield and our focused inclusion criteria. We considered only machine-learning–oriented studies that define a clear prediction task—such as classification or regression—supported by a dataset, modeling framework, or empirical evaluation. This focus ensures that the review highlights work advancing predictive methodologies for ASB. Our search covered peer-reviewed conferences, journals, and preprints indexed in major databases (Google Scholar, DBLP, Scopus, IEEE Xplore, SpringerLink, and ScienceDirect), following established practices in abusive language research (e.g., (Tontodimamma et al., 2021; Jahan and Oussalah, 2023)).

A further limitation arises from restricted access to closed-access venues, leading to possible underrepresentation of research outside open-access repositories or preprint archives. Despite extensive keyword-driven searches, some relevant studies using alternative terms (e.g., "harmful content forecasting", "online risk modeling") may have been missed. To reduce omissions, the author supplemented automated searches with a curated archive of domain-relevant publications. As the field grows, future surveys would benefit from broader interdisciplinary coverage and more inclusive search strategies to fully capture the diversity of predictive work on ASB online.

9 Ethical Considerations

Research on ASB prediction inevitably involves the collection, annotation, and modeling of harmful or offensive content, which poses ethical risks for both researchers and affected users. Annotation tasks, in particular, can expose annotators and development teams to distressing material, potentially impacting mental health. In line with prior recommendations (Vidgen et al., 2019; Kirk et al., 2022), projects involving ASB data should incorporate protective measures, including rotating exposure schedules, mental health support resources, and the use of clear content warnings.

Privacy and data ethics are equally critical in this domain, especially when models are trained on user-generated content from social media platforms. Even with anonymization, re-identification risks remain—particularly for marginalized or vulnerable communities whose linguistic patterns or behavioral markers may be unique, underrepresented, or stigmatized. These concerns are amplified in predictive tasks such as user risk profiling, where models infer the likelihood of individuals engaging in or becoming targets of ASB based on prior behavior, linguistic cues, and engagement patterns. While such approaches may support early intervention or moderation, they also introduce significant ethical risks, including surveillance-like monitoring, reputational bias, and opaque classification processes. In proactive moderation and early-warning systems, false positives or disproportionate targeting may further marginalize at-risk populations, undermining the very goals of safety and fairness. To mitigate these risks, ASB prediction systems must adhere to core principles of fairness, transparency, and accountability, as emphasized in regulatory frameworks such as the GDPR (Council, 2016), and should align with ongoing efforts in explainable AI (Mehta and Passi, 2022) to ensure that model decisions are interpretable and contestable.

Finally, as generative and predictive models become increasingly integrated into real-time moderation workflows, issues of responsibility and accountability become paramount. These systems should not operate autonomously in high-stakes settings. Instead, human-in-the-loop oversight, transparent decision-making processes, and strong governance mechanisms are essential to prevent overreach, hallucinated inferences, or unintended harm amplification. Aligning technical performance with ethical and social responsibility must remain a central goal in the development and deployment of ASB prediction systems.

References

Hind A. Al-Merekhi, Haewoon Kwak, Joni Salminen, and Bernard J. Jansen. 2020. Are these comments triggering? predicting triggers of toxicity in online discussions. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 3033–3040. ACM / IW3C2.

Aish Albladi, Minarul Islam, Amit Das, Maryam Bigonah, Zheng Zhang, Fatemeh Jamshidi, Mostafa Rahgouy, Nilanjana Raychawdhary, Daniela Marghitu, and Cheryl D. Seals. 2025. Hate speech detection using large language models: A comprehensive review. *IEEE Access*, 13:20871–20892.

Raneem Alharthi, Rajwa Alharthi, Ravi Shekhar, Aiqi Jiang, and Arkaitz Zubiaga. 2025. Will I get hate

speech predicting the volume of abusive replies before posting in social media. *CoRR*, abs/2503.03005.

Fatimah Alkomah and Xiaogang Ma. 2022. A literature review of textual hate speech detection methods and datasets. *Inf.*, 13(6):273.

Enas Altarawneh, Ameeta Agrawal, Michael Jenkin, and Manos Papagelis. 2023. Conversation derailment forecasting with graph convolutional networks. *CoRR*, abs/2306.12982.

American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*, 5 edition. American Psychiatric Publishing.

Jisun An, Haewoon Kwak, Claire Seungeun Lee, Bogang Jun, and Yong-Yeol Ahn. 2021. Predicting anti-asian hateful users on twitter during COVID-19. pages 4655–4666.

Jiajun Bao, Junjie Wu, Yiming Zhang, Eshwar Chandrasekharan, and David Jurgens. 2021. Conversations gone alright: Quantifying and predicting prosocial outcomes in online conversations. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 1134–1145. ACM / IW3C2.

Helena Bonaldi, Yi-Ling Chung, Gavin Abercrombie, and Marco Guerini. 2024. NLP for counterspeech against hate: A survey and how-to guide. In *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 3480–3499. Association for Computational Linguistics.

Jo Brown, Amanda J Broderick, and Nick Lee. 2007. Word of mouth communication within online communities: Conceptualizing the online social network. *Journal of interactive marketing*, 21(3):2–20.

Charalampos Chelmis and Mengfan Yao. 2019. Minority report: Cyberbullying prediction on instagram. In *Proceedings of the 11th ACM Conference on Web Science, WebSci 2019, Boston, MA, USA, June 30 - July 03, 2019*, pages 37–45. ACM.

Justin Cheng, Michael S. Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW 2017, Portland, OR, USA, February 25 - March 1, 2017*, pages 1217–1230. ACM.

Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2015. Antisocial behavior in online discussion communities. In *Proceedings of the Ninth International Conference on Web and Social Media, ICWSM 2015, University of Oxford, Oxford, UK, May 26-29, 2015*, pages 61–70. AAAI Press.

Arijit Ghosh Chowdhury, Ramit Sawhney, Rajiv Ratn Shah, and Debanjan Mahata. 2019. #youtoo? detection of personal recollections of sexual harassment on social media. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2527–2537. Association for Computational Linguistics.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.

Yi-Ling Chung, Serra Sinem Tekiroglu, and Marco Guerini. 2020. Italian counter narrative generation to fight online hate speech. In *Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020, Bologna, Italy, March 1-3, 2021*, volume 2769 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proc. Natl. Acad. Sci. USA*, 118(9):e2023301118.

European Council. 2016. EU regulation 2016/679 general data protection regulation (gdpr). *Official Journal of the European Union*, 59(6):1–88.

J. Ignacio Criado, Rodrigo Sandoval-Almazán, and J. Ramón Gil-García. 2013. Government innovation through social media. *Gov. Inf. Q.*, 30(4):319–326.

Snehil Dahiya, Shalini Sharma, Dhruv Sahnan, Vasu Goel, Emilie Chouzenoux, Víctor Elvira, Angshul Majumdar, Anil Bandhakavi, and Tanmoy Chakraborty. 2021. Would your tweet invoke hate on the fly? forecasting hate intensity of reply threads on twitter. In *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, pages 2732–2742. ACM.

Gabriele Etta, Matteo Cinelli, Niccolò Di Marco, Michele Avalle, Alessandro Panconesi, and Walter Quattrociocchi. 2024. A topology-based approach for predicting toxic outcomes on twitter and youtube. *IEEE Trans. Netw. Sci. Eng.*, 11(5):4875–4885.

Gabriele Etta, Emanuele Sangiorgio, Niccolò Di Marco, Michele Avalle, Antonio Scala, Matteo Cinelli, and Walter Quattrociocchi. 2022. Characterizing engagement dynamics across topics on facebook. *CoRR*, abs/2211.15988.

Tope Christopher Christopher Falade, Niloofar Yousefi, and Nitin Agarwal. 2024. Toxicity prediction in reddit. In *30th Americas Conference on Information Systems: Elevating Life through Digital Social*

Entrepreneurship, AMCIS 2024, Salt Lake City, UT, USA, August 15-17, 2024. Association for Information Systems.

Komal Florio, Valerio Basile, Marco Polignano, Pierpaolo Basile, and Viviana Patti. 2020. Time of your hate: The challenge of time in hate speech detection on social media. *Applied Sciences*, 10(12).

Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2025. Perspectivist approaches to natural language processing: a survey. *Lang. Resour. Evaluation*, 59(2):1719–1746.

Daniel Fried, Nicholas Tomlin, Jennifer Hu, Roma Patel, and Aida Nematzadeh. 2023. Pragmatics in language grounding: Phenomena, tasks, and modeling approaches. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 12619–12640. Association for Computational Linguistics.

Paolo Gajo, Arianna Muti, Katerina Korre, Silvia Bernardini, and Alberto Barrón-Cedeño. 2023. On the identification and forecasting of hate speech in inceldom. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing, RANLP 2023, Varna, Bulgaria, 4-6 September 2023*, pages 373–384. INCOMA Ltd., Shoumen, Bulgaria.

Anatoliy Gruzd, Philip Mai, and Zahra Vahedi. 2020. Studying anti-social behaviour on reddit with communalytic. *The SAGE Handbook of Social Media Research Methods*, pages 503–520.

Songqiao Han, Hailiang Huang, Jiangwei Liu, and Shengsheng Xiao. 2021. American hate crime trends prediction with event extraction. *arXiv preprint arXiv:2111.04951*.

Caroline Haythornthwaite. 2023. Moderation, networks, and anti-social behavior online. *Social Media + Society*, 9(3):20563051231196874.

Liam Hebert, Lukasz Golab, and Robin Cohen. 2023. Predicting hateful discussions on reddit using graph transformer networks and communal context. volume abs/2301.04248.

Daniel Hickey, Daniel M. T. Fessler, Matheus Schmitz, Kristina Lerman, and Keith Burghardt. 2025. The peripatetic hater: Predicting movement among hate subreddits. In *Proceedings of the Nineteenth International AAAI Conference on Web and Social Media, June 23-26, 2025, Copenhagen, Denmark*, pages 786–803. AAAI Press.

Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2015. Prediction of cyberbullying incidents on the instagram social network. *CoRR*, abs/1508.06257.

Homa Hosseinmardi, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2016. Prediction of cyberbullying incidents in a media-based social network. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016, San Francisco, CA, USA, August 18-21, 2016*, pages 186–192. IEEE Computer Society.

Mia Mohammad Imran, Robert Zita, Rebekah Copeland, Preetha Chatterjee, Rahat Rizvi Rahman, and Kostadin Damevski. 2025. Understanding and predicting derailment in toxic conversations on github. *CoRR*, abs/2503.02191.

Darius Irani, Avyakta Wrat, and Silvio Amir. 2021. Early detection of online hate speech spreaders with learned user representations. In *Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021*, volume 2936 of *CEUR Workshop Proceedings*, pages 2004–2010. CEUR-WS.org.

Md Saroar Jahan and Mourad Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546:126232.

Richard Kahn and Douglas Kellner. 2004. New media and internet activism: From the 'battle of seattle' to blogging. *New Media Soc.*, 6(1):87–95.

Elise Karinshak, Amanda Hu, Kewen Kong, Vishwanatha Rao, Jingren Wang, Jindong Wang, and Yi Zeng. 2024. LLM-GLOBE: A benchmark evaluating the cultural values embedded in LLM output. *CoRR*, abs/2411.06032.

Chris J. Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. 2020a. Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application. *CoRR*, abs/2009.10277.

Chris J. Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. 2020b. Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application. *CoRR*, abs/2009.10277.

Sahrish Khan, Rabeeh Ayaz Abbasi, Muddassar Azam Sindhu, Sachi Arafat, Akmal Saeed Khattak, Ali Daud, and Mubashar Mushtaq. 2024. Predicting the victims of hate speech on microblogging platforms. *Heliyon*, 10(23).

Dohyeon Kim, Taehoon Kim, and Jihoon Yang. 2025. Early detection of online grooming with language models. In *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing, SAC 2025, Catania International Airport, Catania, Italy, 31 March 2025 - 4 April 2025*, pages 963–970. ACM.

Hannah Kirk, Abeba Birhane, Bertie Vidgen, and Leon Derczynski. 2022. Handling and presenting harmful

text in NLP research. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 497–510. Association for Computational Linguistics.

Barbara Kitchenham. 2004. Procedures for performing systematic reviews. Technical report, Keele University, Department of Computer Science.

Filip Klubicka and Raquel Fernández. 2018. Examining a hate speech corpus for hate speech detection and popularity prediction. *CoRR*, abs/1805.04661.

Ana Kotarcic, Dominik Hangartner, Fabrizio Gilardi, Selina Kurer, and Karsten Donnay. 2022. Human-in-the-loop hate speech classification in a multilingual context. *CoRR*, abs/2212.02108.

Charlotte Lambert, Ananya Rajagopal, and Eshwar Chandrasekharan. 2022. Conversational resilience: Quantifying and predicting conversational outcomes following adverse events. In *Proceedings of the Sixteenth International AAAI Conference on Web and Social Media, ICWSM 2022, Atlanta, Georgia, USA, June 6-9, 2022*, pages 548–559. AAAI Press.

Sharon Levy, Robert E. Kraut, Jane A. Yu, Kristen M. Altenburger, and Yi-Chia Wang. 2022. Understanding conflicts in online conversations. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, pages 2592–2602. ACM.

Ken-Yu Lin, Roy Ka-Wei Lee, Wei Gao, and Wen-Chih Peng. 2021. Early prediction of hate speech propagation. In *2021 International Conference on Data Mining, ICDM 2021 - Workshops, Auckland, New Zealand, December 7-10, 2021*, pages 967–974. IEEE.

Jiangwei Liu, Xiangzhen Jia, You Wu, Jingshu Zhang, and Xiaohong Huang. 2024a. From news to knowledge: Predicting hate crime trends through event extraction from media content. *IAENG International Journal of Applied Mathematics*, 54(4).

Yu Lu Liu, Su Lin Blodgett, Jackie C. K. Cheung, Vera Liao, Alexandra Olteanu, and Ziang Xiao. 2024b. ECBD: evidence-centered benchmark design for NLP. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 16349–16365. Association for Computational Linguistics.

Manuel F. López-Vizcaíno, Francisco Javier Nóvoa, Thierry Artières, and Fidel Cacheda. 2023a. Site agnostic approach to early detection of cyberbullying on social media networks. *Sensors*, 23(10):4788.

Manuel F. López-Vizcaíno, Francisco Javier Nóvoa, Thierry Artières, and Fidel Cacheda. 2023b. Site agnostic approach to early detection of cyberbullying on social media networks. *Sensors*, 23(10):4788.

Manuel F. López-Vizcaíno, Francisco Javier Nóvoa, Victor Carneiro, and Fidel Cacheda. 2021. Early detection of cyberbullying on social media networks. *Future Gener. Comput. Syst.*, 118:219–229.

Sakshi Makkar and Tanmoy Chakraborty. 2020. *Hate speech diffusion in twitter social media*. Ph.D. thesis, IIIT-Delhi.

Sarah Masud, Subhabrata Dutta, Sakshi Makkar, Chhavi Jain, Vikram Goyal, Amitava Das, and Tanmoy Chakraborty. 2021. Hate is the new infodemic: A topic-aware modeling of hate speech diffusion on twitter. In *37th IEEE International Conference on Data Engineering, ICDE 2021, Chania, Greece, April 19-22, 2021*, pages 504–515. IEEE.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14867–14875. AAAI Press.

Harshkumar Mehta and Kalpdrum Passi. 2022. Social media hate speech detection using explainable artificial intelligence (xai). *Algorithms*, 15(8).

Qing Meng, Tharun Suresh, Roy Ka-Wei Lee, and Tanmoy Chakraborty. 2023. Predicting hate intensity of twitter conversation threads. *Knowl. Based Syst.*, 275:110644.

Meta. 2022. Facebook community standards. `https://transparency.fb.com/policies/community-standards/`. (2022c).

David Moher, Alessandro Liberati, Jennifer Tetzlaff, and Douglas G Altman. 2009. Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *Bmj*, 339.

Djedjiga Mouheb, Masa Hilal Abushamleh, Maya Hilal Abushamleh, Zaher Al Aghbari, and Ibrahim Kamel. 2019. Real-time detection of cyberbullying in arabic twitter streams. In *10th IFIP International Conference on New Technologies, Mobility and Security, NTMS 2019, Canary Islands, Spain, June 24-26, 2019*, pages 1–5. IEEE.

Hamdy Mubarak, Samir Abdaljalil, Azza Nassar, and Firoj Alam. 2023. Detecting and reasoning of deleted tweets before they are posted. *CoRR*, abs/2305.04927.

Hung Truong Thanh Nguyen, Hung Quoc Cao, Khang Vo Thanh Nguyen, and Nguyen Dinh Khoi Pham. 2021. Evaluation of explainable artificial intelligence: Shap, lime, and cam. In *Proceedings of the FPT AI Conference*, pages 1–6.

Kenya Nonaka and Mitsuo Yoshida. 2025. Zero-shot prediction of conversational derailment with large language models. *IEEE Access*, 13:55081–55093.

Anaïs Ollagnier. 2024. Cyberagressionado-v2: Leveraging pragmatic-level information to decipher online hate in french multiparty chats. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 4287–4298. ELRA and ICCL.

Anaïs Ollagnier, Elena Cabrio, and Serena Villata. 2023a. Harnessing bullying traces to enhance bullying participant role identification in multi-party chats. In *Proceedings of the Thirty-Sixth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2023, Clearwater Beach, FL, USA, May 14-17, 2023*. Florida Online Journals.

Anaïs Ollagnier, Elena Cabrio, and Serena Villata. 2023b. Unsupervised fine-grained hate speech target community detection and characterisation on social media. *Soc. Netw. Anal. Min.*, 13(1):58.

Anaïs Ollagnier, Elena Cabrio, Serena Villata, and Valerio Basile. 2024. Cyberagressionado-large: French multiparty chat dataset to address online hate. *Revue TAL : traitement automatique des langues*, 65(3):21–44.

Alexandra Olteanu, Kartik Talamadupula, and Kush R. Varshney. 2017. The limits of abstract evaluation metrics: The case of hate speech detection. In *Proceedings of the 2017 ACM on Web Science Conference*, WebSci '17, page 405–406, New York, NY, USA. Association for Computing Machinery.

Mark C. Parent, Taylor D. Gobble, and Aaron Rochlen. 2019. Social media behavior, toxic masculinity, and depression. *Psychology of Men & Masculinity*, 20(3):277–287. Epub 2018 Apr 23.

Paloma Piot, Patricia Martín-Rodilla, and Javier Parapar. 2024. Metahate: A dataset for unifying efforts on hate speech detection. In *Proceedings of the Eighteenth International AAAI Conference on Web and Social Media, ICWSM 2024, Buffalo, New York, USA, June 3-6, 2024*, pages 2025–2039. AAAI Press.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Lang. Resour. Evaluation*, 55(2):477–523.

Nektaria Potha and Manolis Maragoudakis. 2014. Cyberbullying detection using time series modeling. In *2014 IEEE International Conference on Data Mining Workshops, ICDM Workshops 2014, Shenzhen, China, December 14, 2014*, pages 373–382. IEEE Computer Society.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth M. Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4754–4763. Association for Computational Linguistics.

Walter Quattrociocchi, Guido Caldarelli, and Antonio Scala. 2014. Opinion dynamics on interacting networks: media competition and social influence. *Scientific Reports*, 4:4938.

Koustuv Saha, Eshwar Chandrasekharan, and Munmun De Choudhury. 2019. Prevalence and psychological effects of hateful speech in online college communities. In *Proceedings of the 11th ACM Conference on Web Science, WebSci 2019, Boston, MA, USA, June 30 - July 03, 2019*, pages 255–264. ACM.

Martin Saveski, Brandon Roy, and Deb Roy. 2021. The structure of toxic conversations on twitter. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 1086–1097. ACM / IW3C2.

Kirill Solovev and Nicolas Pröllochs. 2023. Moralized language predicts hate speech on social media. *PNAS nexus*, 2(1):pgac281.

Xiaoying Song, Sharon Lisseth Perez, Xinchen Yu, Eduardo Blanco, and Lingzi Hong. 2025. Echoes of discord: Forecasting hater reactions to counterspeech. In *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 4892–4905. Association for Computational Linguistics.

Billy Spann and Nitin Agarwal. 2023. Predicting toxicity in reddit discussion threads. In *Proceedings of the 16th International Conference on Social Computing, Behavioral-Cultural Modeling & Prediction and Behavior Representation in Modeling and Simulation (SBP-BRiMS 2023)*.

Sajedul Rahim Talukder and Bogdan Carbunar. 2018. Abusniff: Automatic detection and defenses against abusive facebook friends. In *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018*, pages 385–394. AAAI Press.

Benedetta Tessa, Lorenzo Cima, Amaury Trujillo, Marco Avvenuti, and Stefano Cresci. 2024. Beyond trial-and-error: Predicting user abandonment after a moderation intervention. *CoRR*, abs/2404.14846.

Alice Tontodimamma, Eugenia Nissi, Annalina Sarra, and Lara Fontanella. 2021. Thirty years of research into hate speech: topics of interest and their evolution. *Scientometrics*, 126(1):157–179.

Paraskevas Tsantarliotis, Evaggelia Pitoura, and Panayiotis Tsaparas. 2017. Defining and predicting troll vulnerability in online social media. *Soc. Netw. Anal. Min.*, 7(1):26:1–26:15.

Michela Del Vicario, Gianna Vivaldo, Alessandro Bessi, Fabiana Zollo, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. 2016. Echo chambers: Emotional contagion and group polarization on facebook. *CoRR*, abs/1607.01032.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.

Ying Wang, Yingji Li, Yue Wu, and Xin Wang. 2024. Exploring multiple hypergraphs for heterogeneous graph neural networks. *Expert Syst. Appl.*, 236:121230.

Xiao-Kun Wu, Tian-Fang Zhao, Lu Lu, and Wei-Neng Chen. 2022. Predicting the hate: A gstm model based on covid-19 hate speech datasets. *Information Processing & Management*, 59(4):102998.

Xinchen Yu, Eduardo Blanco, and Lingzi Hong. 2024. Hate cannot drive out hate: Forecasting conversation incivility following replies to hate speech. In *Proceedings of the Eighteenth International AAAI Conference on Web and Social Media, ICWSM 2024, Buffalo, New York, USA, June 3-6, 2024*, pages 1740–1752. AAAI Press.

Jiaqing Yuan and Munindar P. Singh. 2023. Conversation modeling to predict derailment. In *Proceedings of the Seventeenth International AAAI Conference on Web and Social Media, ICWSM 2023, Limassol, Cyprus, June 5-8, 2023*, pages 926–935. AAAI Press.

Yunfan Zhang, Kathleen McKeown, and Smaranda Muresan. 2025. Forecasting communication derailments through conversation generation. *CoRR*, abs/2504.08905.

Fan Zhou, Xovee Xu, Goce Trajcevski, and Kunpeng Zhang. 2022. A survey of information cascade analysis: Models, predictions, and recent advances. *ACM Comput. Surv.*, 54(2):27:1–27:36.

A Methodology of the Review

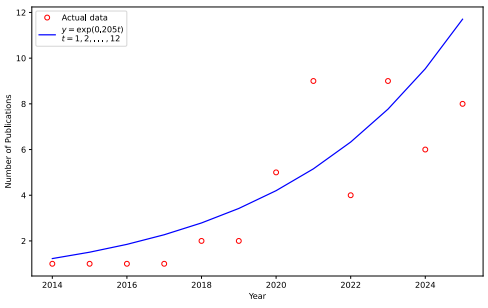

Figure 1: Number of publications on ASB prediction per year: observed and expected distribution.

Figure 1 illustrates the steady growth of research on ASB prediction, reflecting the increasing interest in proactive and forward-looking approaches to online harm. To ensure methodological rigor, transparency, and reproducibility, we conducted this survey following established guidelines for systematic literature reviews in computer science and engineering (Kitchenham, 2004), in conjunction with the PRISMA framework (Moher et al., 2009). The study selection process is summarized in Fig. 2.

The objective of this review is to provide a structured synthesis of progress to date in ASB prediction tasks within the ML literature, clearly distinguishing prediction-oriented approaches from the more mature body of work on detection and classification. We operationalize prediction as tasks that forecast future outcomes or risks using supervised ML techniques, rather than merely identifying existing harmful content. Accordingly, we include only peer-reviewed, openly accessible studies that adhere to this definition, while systematically excluding work limited to detection, identification, profiling, or post-hoc analysis without a forward-looking objective.

Following prior surveys of harmful language and online abuse (Poletto et al., 2021; Jahan and Oussalah, 2023), we retrieved candidate papers from major scholarly databases, including Google Scholar, DBLP, Scopus, IEEE Xplore, SpringerLink, and ScienceDirect. To streamline and centralize retrieval, we used the `SciLEx` aggregator[6]. Search queries combined prediction-oriented keywords (e.g., prediction, forecasting, early detection, propagation, temporal modeling, real-time prediction) with ASB-related terms (e.g., antisocial behaviour, cyberaggression, hate speech, cyberbullying, online abuse, toxicity, online grooming, hate crime). Searches were applied to titles and abstracts, restricted to English-language publications between 2010 and 2025, with the final search conducted on June 17, 2025.

The screening process proceeded in multiple stages. First, titles and abstracts were reviewed to remove irrelevant or tangential works. Second, full-text screening was conducted to verify that remaining studies met all inclusion criteria, particularly with respect to their predictive framing and ML-based methodology. Duplicate records retrieved across databases were removed. Each included paper was then annotated using a classifi-

[6] https://github.com/Wimmics/SciLEx/

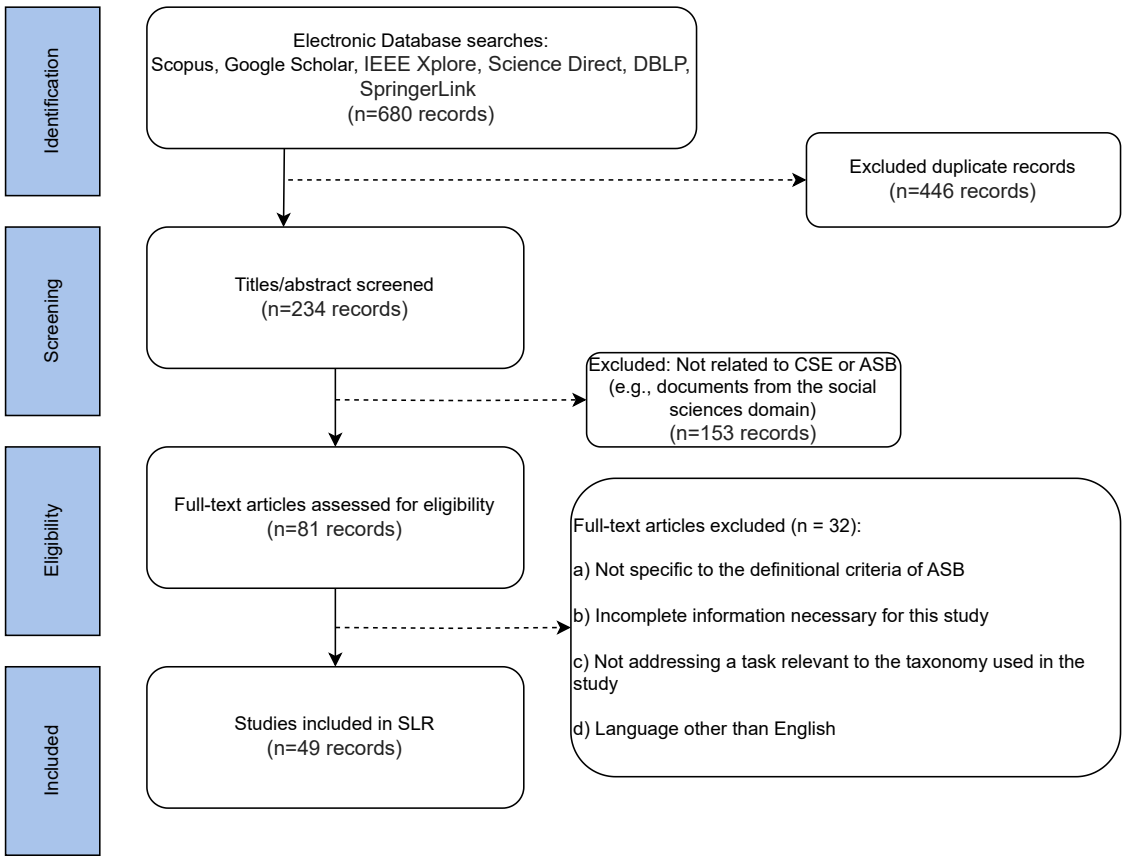

Figure 2: PRISMA flowchart for selection of antisocial behaviour prediction studies.

cation scheme capturing its disciplinary grounding (e.g., NLP, computational social science), primary research contribution (e.g., predictive modeling, resource creation, intervention design), methodological approach (e.g., neural, traditional ML, or hybrid), and the specific ASB prediction task addressed, following the taxonomy in Section 2.2.

This systematic and multi-stage process resulted in a final corpus of 49 papers, which form the empirical basis for our synthesis of the state of the art and the identification of open challenges and future directions in ASB prediction research.

B Task Category Distribution in ASB Prediction

The distribution of task categories in current ASB prediction research, illustrated in Figure 3, reveals a clear concentration around early harm detection and harm emergence prediction, which together represent the majority of studied scenarios. Early detection tasks are designed to anticipate escalation risks in conversations, flag toxic replies at their onset, or detect early signs of conflict based on minimal interaction cues (e.g., (Al-Merekhi et al., 2020; Mubarak et al., 2023; Kim et al., 2025)). Harm emergence tasks focus on forecasting whether harmful behavior will surface in the course of an interaction, such as predicting conversational derailment or the likelihood of toxic replies (e.g., (Nonaka and Yoshida, 2025; Altarawneh et al., 2023)). Harm Propagation Prediction forms a substantial secondary cluster, encompassing tasks like modeling the spread and virality of hate speech, estimating future hate intensity, or forecasting how toxic content diffuses across social networks (Makkar and Chakraborty, 2020; Meng et al., 2023; Alharthi et al., 2025). In parallel, behavioral risk prediction—though less frequent—covers user-centric forecasting, such as predicting recidivism, moderation outcomes, or the likelihood of becoming a hate speech amplifier (Potha and Maragoudakis, 2014; An et al., 2021; Khan et al., 2024). Lastly, proactive moderation support includes predictive tasks aimed at assisting moderation efforts in advance, such as estimating abuse likelihood at posting time, predicting

conversation resilience, or evaluating harassment severity through multimodal features (Talukder and Carbunar, 2018; Bao et al., 2021; Lambert et al., 2022). Overall, this landscape reflects a multidimensional field, with strong emphasis on early content-level interventions and growing attention toward user trajectories and longitudinal harm forecasting.

C Ex-ante Prediction versus Peeking Strategy

As summarized in Table 2, ex-ante strategies are more prevalent in the literature, particularly for classification tasks (Klubicka and Fernández, 2018; Masud et al., 2021), due to their appeal in proactive moderation. However, peeking and progressive strategies are gaining traction, especially in multi-turn conversation or cascade-based tasks where the temporal evolution of content is central.

D Annotators Training Procedure

Engaging with hate speech and counterspeech is known to have significant social and psychological consequences for those involved. This has led to the development of specific guidelines and best practices to ensure that such work is conducted responsibly (Chung et al., 2020; Ollagnier, 2024; Bonaldi et al., 2024). In the same way, research on ASB should adopt precautions similar to those recommended in other domains dealing with abusive language. Recognizing, labeling, and interpreting ASB in online content—whether for escalation forecasting, propagation modeling, or proactive moderation—requires not only clear technical guidelines but also domain-specific sensitivity and practical experience. Annotators are frequently exposed to distressing material and are tasked with making fine-grained judgments about nuanced patterns of harm. When they lack prior expertise in ASB labeling, targeted training becomes essential to help them develop the necessary skills and ensure consistent, high-quality annotations, as highlighted in prior work on abusive language annotation and harm labeling (Vidgen et al., 2019).

The most commonly reported procedure for training annotators includes the following steps:

a) **Reviewing guidelines and documentation:** annotators first study platform or research-specific annotation guidelines and public documentation on harmful behaviors (e.g., definitions of harassment, trolling, or hate speech).

b) **Learning from examples:** they are then exposed to curated examples of ASB labeling (including edge cases such as implicit hate or sarcasm) and, where relevant, examples of expert annotations for specific tasks (e.g., labeling early harm signals).

c) **Practice sessions:** annotators perform trial labeling on a subsample of posts or conversation threads, receiving iterative feedback.

d) **Discussion and calibration:** disagreements and difficult cases are discussed in regular meetings with an expert or lead annotator to achieve shared understanding and improve consistency.

Table 3 summarises the training steps reported in studies that explicitly describe how their annotators were trained, among those studies whose datasets are publicly available and documented. Importantly, because annotators are frequently exposed to harmful or distressing material, their well-being must be prioritized. Precautions include explaining the prosocial purpose of the work, limiting annotation sessions to a few hours per day, encouraging regular breaks, and providing structured opportunities to raise concerns or seek support (Vidgen et al., 2019; Kirk et al., 2022). These measures help ensure reliable annotations while safeguarding the mental health of those involved.

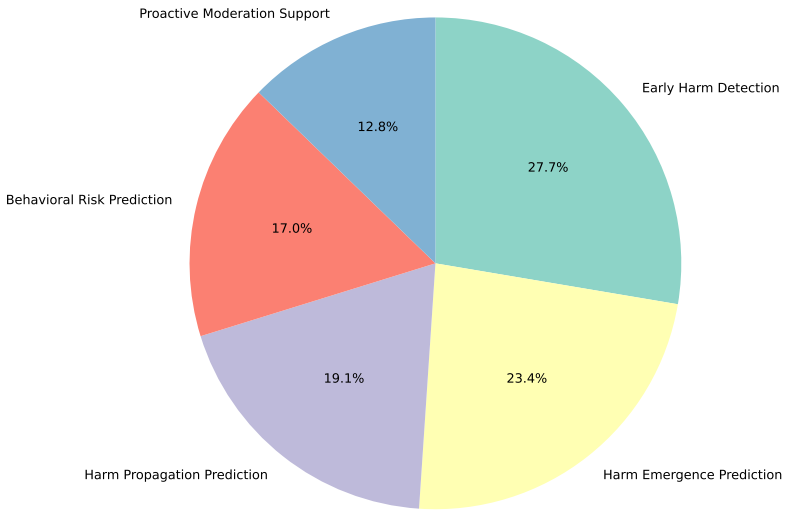

Figure 3: Distribution of ASB prediction task categories across reviewed papers.

Strategy	Formulation	Reference
Ex-ante	Classification	(Hosseinmardi et al., 2015, 2016; Klubicka and Fernández, 2018; Talukder and Carbunar, 2018; Mouheb et al., 2019; Makkar and Chakraborty, 2020; Masud et al., 2021; Bao et al., 2021; An et al., 2021; Irani et al., 2021; Mubarak et al., 2023; López-Vizcaíno et al., 2023b; Yuan and Singh, 2023; Spann and Agarwal, 2023; Altarawneh et al., 2023; Etta et al., 2024; Khan et al., 2024; Falade et al., 2024; Tessa et al., 2024; Kim et al., 2025; Imran et al., 2025; Nonaka and Yoshida, 2025; Zhang et al., 2025; Song et al., 2025; Hickey et al., 2025)
	Regression	(Kennedy et al., 2020b; Han et al., 2021; Wu et al., 2022; Levy et al., 2022; Hebert et al., 2023; Meng et al., 2023; Gajo et al., 2023; Yu et al., 2024; Liu et al., 2024a; Alharthi et al., 2025)
Peeking	Classification	(Potha and Maragoudakis, 2014; Cheng et al., 2015; Saveski et al., 2021; López-Vizcaíno et al., 2023a; Lin et al., 2021; Lambert et al., 2022; Solovev and Pröllochs, 2023)
	Regression	(Chelmis and Yao, 2019; Dahiya et al., 2021)

Table 2: Strategies and formulations of feature-based models.

Study	a	b	c	d
Kennedy et al. (2020a)	–	–	✓	–
Saveski et al. (2021)	✓	–	–	–
Bao et al. (2021)	✓	✓	✓	✓
Dahiya et al. (2021)	✓	–	–	–
Gajo et al. (2023)	✓	✓	✓	✓
Alharthi et al. (2025)	✓	✓	✓	✓
Kim et al. (2025)	✓	✓	–	–
Nonaka and Yoshida (2025)	✓	–	–	–
Song et al. (2025)	✓	✓	–	–
Hickey et al. (2025)	✓	✓	–	–

Table 3: Steps for annotators' training in studies that explicitly describe them, as detailed in D.