Title: Categorization of features considered by the detection models appendix for "The unappreciated role of intent in algorithmic moderation of abusive content on social media"
Authors: Xinyu Wang (1), Sai Koneru (1), Pranav Narayanan Venkit (1), Brett Frischmann (2), Sarah Rajtmajer (1)
Date: July 29th, 2025
Note: The material contained herein is supplementary to the article named in the title and published in the Harvard Kennedy School (HKS) Misinformation Review.

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.