Studii de lingvistică 14, nr.2, 2024, 63 - 94

Towards a new Contextualized Annotation Schema for Unacceptable and Extreme Speech (CUES) to Unleash Generalization Capability of ML models

Dimitra Niaouri[1]
Michele Linardi[2]
Julien Longhi[3]

Abstract: In an era marked by global crises and social challenges, including inequality, unrest, and the proliferation of extreme online content, the need for effective Machine Learning (ML) solutions to detect Socially Unacceptable Discourse (SUD) is paramount. However, existing ML models face significant challenges in accurately classifying such content due to issues such as biased annotations, limited contextual understanding, and the neglect of multimodal elements. Additionally, binary classes in annotated datasets limit SUD representation, affecting real-world discriminative capacity, while multiclass frameworks expose generalization gaps and label semantics inconsistencies, hindering multi-source learning. This paper presents a novel approach aimed at enhancing the capabilities of state-of-the-art (SOTA) ML models by providing a set of guidelines that will allow to semantically enrich existing ad-hoc annotation schemas and better leverage state-of-the-art machine learning classifiers. Our methodology focuses on refining labels and improving model generalization by incorporating diverse contextual factors underlying the spread of unacceptable speech. We address the limitations of existing annotated datasets, including class imbalances and overlapping classes, and propose a systematic evaluation of our annotation schema across various ML models. By investigating user information and multimodal elements from online platforms, we aim to better understand the socio-cultural environment in which SUD arises. Through our approach, we highlight the significance of context in enhancing the effectiveness of ML algorithms for detecting extreme online content.

Keywords: Socially Unacceptable Discourse Analysis, Hate Speech Analysis, Machine Learning, Annotation Schema, Unacceptable Discourse Context Modeling

[1] ETIS UMR-8051, AGORA, Cergy Paris Université ; dimitra.niaouri@cyu.fr.
[2] ETIS UMR-8051, Cergy Paris Université ; michele.linardi@cyu.fr.
[3] AGORA, Cergy Paris Université ; julien.longhi@cyu.fr.

1. Introduction[4]

Over the last decades, the widespread adoption of social media has profoundly altered the landscape of human communication and global information sharing (Calderón *et al.*, 2020; Carneiro *et al.*, 2023). While the rapid spread of information and the ability to connect with a broad audience are clear benefits of these platforms, they also present challenges. The potential for anonymity and lack of accountability can foster the spread of socially unacceptable ideas and contribute to radicalized public discourse, often advocating to forms of discrimination such as racism, sexism, homophobia and many others (*ibid.*).

1.1. Positioning with respect to the Corpus Analysis

One manifestation of harmful communication is Socially Unacceptable Discourse (SUD), which refers to a range of offensive and aggressive communication behaviors. SUD includes direct and indirect threats, offensive language, incitement to violence, negative stereotypes, generalizations, and provocative or obscene statements (Vehovar *et al.*, 2020; de Maiti & Fišer, 2021). According to Okulska and Kołos (2023), some of the linguistic characteristics of hate speech (and consequently SUD) may include a high frequency of nouns that objectify target groups, the use of third-person plural pronouns to create an "us *vs* them" dynamic, and reliance on the present tense for immediacy and authority. Additionally, imperative forms are being used to call for harmful actions along with complex nominal phrases to negatively characterize targets.

The concept of SUD can be positioned within a broader framework of related terms, particularly in relation to uncivil discourse and extremist narratives. Uncivil discourse is characterized by communication that conveys a disrespectful tone, which aligns closely with SUD's inclusion of aggressive behaviors like threats, offenses, and provocations (Coe *et al.* 2014; Rossini 2020). It often includes behaviors such as name-calling, vulgarity, and hostility, among others (Coe *et al.* 2014; Kenski *et al.*, 2020), which are frequently identified in both types of discourse. Closely related to SUD are extremist narratives, which, unlike the often-unstructured nature of SUD, are deliberate, ideologically driven, and use toxic communication to mobilize support for radical ideologies. Extremist narratives often leverage hostile language to frame "us *vs* them" dichotomies, amplifying polarization and spreading structured,

[4] Acknowledgments: the work presented in this paper is part of the ARENAS project. This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101094731. This work was granted access to the HPC resources of IDRIS under the allocation 2024-AD010615085R1 made by GENCI.

divisive worldviews (Postigo-Fuentes *et al.* 2024).

Given the pervasive nature of toxic communication online, detecting and characterizing[5] such forms of harmful discourse is an essential requirement for effective social media moderation and analysis. This process involves the ability to automatically differentiate between various types of discourses, such as insults, cyberbullying, and homophobia, among others. This capability is crucial for several reasons:

(I) **Tailored Interventions:** Understanding the specific type of toxicity enables targeted interventions, enhancing effectiveness in addressing the issue;

(II) **Understanding Root Causes:** Identifying toxic discourse provides insights into underlying causes like prejudice and a lack of online etiquette, informing strategies for promoting positive interactions;

(III) **Legal Frameworks:** Characterizing toxic discourse aids in establishing appropriate legal frameworks for moderation, ensuring regulatory measures align with the nature of toxic behaviors.

1.2. Positioning with respect to the Machine Learning State-Of-The-Art

While machine learning (ML) holds promise for automating content detection, significant obstacles challenge its effectiveness. Analysts face several challenges when employing current ML solutions for detecting SUD. Textual features often overlap, demanding careful analysis and clear criteria for differentiation due to the absence of standard SUD annotation guidelines (Fišer *et al.*, 2017). Furthermore, prior studies have shown that annotator bias can substantially impact outcomes, which complicates ML models' generalization capabilities (Badjatiya *et al.*, 2019; Yuan *et al.*, 2022; Davidson *et al.*, 2019; Yuan and Rizoiu, 2022). Davani *et al.* (2023) specifically highlighted that hate speech classifiers often reflect societal stereotypes against marginalized groups, leading to systematic biases and potentially perpetuating social

[5] The ARENAS project, within which this study falls under, aims to analyze extremist narratives as discursive phenomena with significant linguistic and semiotic dimensions. By examining the language and underlying narratives of extremism, the project aims to provide concrete recommendations for the detection, characterization, and prevention of extremist narratives. In this way, the ARENAS project represents a substantial advancement in the scholarly investigation of toxic discourse detection, moving beyond traditional boundaries to examine the intricate interplay of language, ideology, and conviction within extremist narratives.

inequalities. Contextual and topic dependencies also play a crucial role in discourse characterization, as language intent is highly context-sensitive (Sheth *et al.*, 2022). Moreover, the prevalence of binary schemas in publicly available annotated datasets limits the representation of SUD (Sulc and de Maiti, 2020), reducing models' discriminative capacity in real-world settings, while an exclusive emphasis on textual context at the expense of multimodal cues impedes the model's ability to interpret a wide range of signals (e.g., visual, auditory).

In response to these challenges, various studies have proposed different approaches. Carneiro *et al.* (2023) constructed a diverse corpus with 12 classes from 13 public datasets to address multiclass detection but found generalization issues, stressing the need for better annotation schemas. Perifanos and Goutsos (2021) combined textual and visual modalities for abusive content detection, yielding promising results, but their binary classification setup limited its scope. Other studies explored contextual cues, with textual context research emphasizing the importance of parent comments and article titles. However, findings were uneven due to limitations such as small sample sizes (Gao & Huang, 2017) and insufficient classifier accuracy (Pavlopoulos *et al.*, 2020), as well as the omission of critical contextual elements (Mubarak *et al.*, 2017) and the lack of adequate information for replicating the study (Pavlopoulos *et al.*, 2017). However, context-aware models have outperformed context-agnostic ones (Xenos *et al.*, 2021) despite the possibility of dataset biases (Zhou *et al.*, 2023). In multimodal contexts, integrating images and videos significantly enhanced hate speech and cyberbullying detection (Yang *et al.*, 2019; Hosseinmardi *et al.*, 2015; Sheth *et al.*, 2022). Research on community-based context underscored the importance of social dynamics and community graphs (Unsvåg & Gambäck, 2018; Mishra *et al.*, 2019; Ziems *et al.*, 2020; Sheth *et al.*, 2022; Kurrek *et al.*, 2022; Nagar *et al.*, 2023), while sociopolitical and pragmatic contexts, including nuances like irony and sarcasm, were critical for interpreting discourse (de Fina & Georgakopoulou, 2020; ElSherief *et al.*, 2021; Paveau, 2013b).

In line with previous works, this paper aims to elevate the proficiency of state-of-the-art (SOTA) ML models through specialized annotation schema guidelines. Our methodology revolves around augmenting model generalizability by incorporating diverse contextual factors that influence SUD detection. We tackle prevalent issues in existing annotated datasets, such as class imbalances and overlapping classes, proposing a methodological evaluation across ML model families, including Shallow Learning Models, Masked Language Models, and Causal Language Models.

We aim to make a novel contribution to the applied linguistic field by harnessing ML solutions and delving into the intricate relationship between the two fields (Lin, 2021; Linzen, 2019). We additionally strive to enhance the interpretability of our models, recognizing the immense potential within the domain of Large Language Model (LLM) explainability (Zhao *et al.*, 2024; Slack *et al.*, 2023). Consequently, we seek to establish a connection between corpus annotation and ML analysis by shedding a light on how annotations impact model performance and how ML can streamline the annotation process.

2. SUD annotation schemas

Numerous works have contributed to the development of annotated datasets for hate speech analysis, including efforts by Davidson *et al.* (2017), Founta *et al.* (2018), Qian *et al.* (2019), and Grimminger and Klinger (2021). One of the most well-known resources is "hatespeechdata[6]", which gathers various datasets and their corresponding links. Upon these foundational works, our study analyzes state-of-the-art annotations from various corpora to evaluate the performance of machine learning (ML) models in detecting SUD. Since SUD covers numerous textual characteristics, we identify an extensive perimeter covering multiple scenarios. We note that a specific discourse analysis naturally requires identifying different features and entities. In this sense, racist content identification is one clear example (Potter and Wetherell, 1988), where we are not only required to identify multiple linguistic features (abusive, aggressive, hate speech) but also the involved entities (e.g., superior groups attacking a specific identity or race) and the context in which subtle dynamics can hide prejudice or other kinds of discrimination.

Following Carneiro *et al.* (2023), we use data sources from multiple platforms to assess the effectiveness of the latest ML solutions for detecting socially unacceptable content. We selected 13 public datasets originating from different annotation schemas, totaling 470,768 samples across 12 classes. By integrating these datasets, we move beyond earlier studies, which often relied on binary classifications or limited class scopes. The selected datasets broadly cover multiple SUD categories, such as racism, homophobia, sexism, abuse, harassment, and offensive and toxic speech.

[6] https://hatespeechdata.com/

Dataset	Source	Sample type	Samples	Topic
Davidson	Davidson *et al.*, 2017	Tweets	25,000	Generic
Founta	Founta *et al.*, 2018	Tweets	100,000	Generic
Fox	Yuan and Rizoiu, 2022	Threads	1,528	Fox News Posts
Gab	Qian *et al.*, 2019	Posts	34,000	Generic
Grimminger	Grimminger and Klinger, 2021	Tweets	3,000	US Presidential Election
HASOC2019	Wang *et al.*, 2019	Facebook Twitter posts	12,000	Generic
HASOC2020	Ghosh Roy *et al.*, 2021	Facebook posts	12,000	Generic
Hateval	MacAvaney *et al.*, 2019	Tweets	13,000	Misogynist and Racist content
Jigsaw	Van Aken *et al.*, 2018	Wikipedia talk pages	220,000	Generic
Olid	Zampieri *et al.*, 2019	Tweets	14,000	Generic
Reddit	Yuan and Rizoiu, 2022	Posts	22000	Toxic subjects
Stormfront	MacAvaney *et al.*, 2019	Threads	10,500	White Supremacy Forum
Trac	Aroyehun and Gelbukh, 2018	Facebook posts	15,000	Generic

Table 1: Summary of datasets (Carneiro *et al.*, 2023)

Table 1 summarizes the datasets we consider. Carneiro *et al.*, (2023) have also unified these corpora in a single dataset, namely the G^{SUD} corpus, which aims to reproduce a large scenario to study overlapping in SUD labels and the existence of bias and ambiguities in the annotation. In the next section, we present the annotation labels adopted in these datasets, with the distribution of the relative class in each dataset.

3. Annotation schema challenges and suggestions

In this section, we delve into the challenges confronting ML/DL methodologies in the automatic detection of SUD. We suggest solutions to enhance future annotation schemas with machine-interpretable data, boosting ML model performance. Challenges are identified in two main areas: uneven class distribution and lack of contextual cues.

3.1. Class distribution

In this section, we examine the distribution of annotated content across datasets selected for our study. We explore how this distribution impacts the efficacy of our models in distinguishing different classes. Figure 1 shows the class distribution across different datasets.

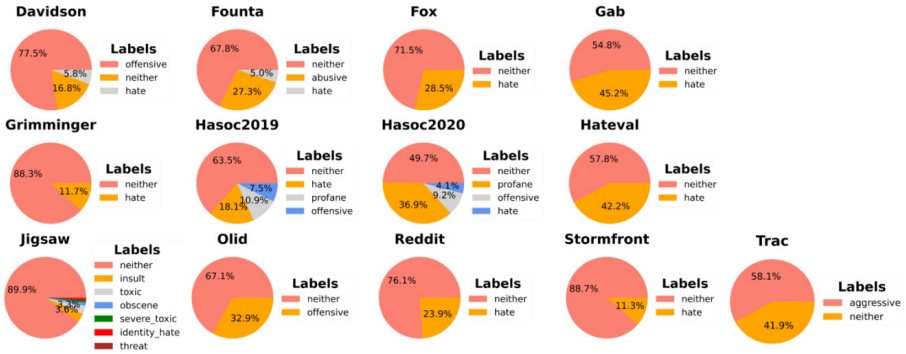

Figure 1: Class distribution of datasets

We observe a common pattern in almost all datasets: a class imbalance where the *neither* class overwhelmingly dominates the other classes. For instance, in Jigsaw, about 89.9% of the entries are labeled as *neither*, with the rest distributed among six other classes. Though the extent of this imbalance differs among datasets, the prevalence of the *neither* class remains consistent. As this imbalance presents challenges for ML models, especially in addressing underrepresented classes, during the annotation stage, it is relevant to report ambiguous cases in which annotator agreement is low and all the information that can guide ML practitioners to exclude biased samples and obtain more balanced class distributions.

3.2. Contextual considerations

This section emphasizes the need to integrate context into annotation schemas. Communication mediums like tweets contain a mix of elements closely tied to their specific context within a technology-driven environment (Paveau, 2013a). To accurately detect SUD, it is crucial to understand the broader context that extends beyond mere situational and domain-specific content analysis (Purohit *et al.*, 2020; Sheth *et al.*, 2022). This understanding should consider applicable human values, social norms, and cultural influences at the individual, group, and community levels (Purohit *et al.*, 2020; Sheth *et al.*, 2022). Acknowledging the importance of contextual information in building

effective SUD detection systems (de Gibert *et al.*, 2018; Pavlopoulos *et al.*, 2020; Vidgen *et al.*, 2021; Zhou *et al.*, 2023), our aim is to deepen understanding of how such factors influence the detection process and outline key considerations for annotation practices. We address five types of contexts: textual, multimodal, community-based, sociopolitical, and pragmatic.

3.2.1. Textual context

Textual context, comprising thematically linked vocabulary terms, provides a framework for understanding subsequent elements (Stairmand, 1997). As noted by Dey (2001), the ability to convey implicit situational information or context in conversations is crucial for increasing conversational bandwidth.

A key challenge in dataset development is the frequent neglect of contextual information during the annotation process (Nobata *et al.*, 2016; Wulczyn *et al.*, 2017; Waseem and Hovy, 2016). This limitation has been noted in previous research, which underscores the crucial role context plays in the effectiveness of automatic hate speech detection systems. For instance, Gao and Huang (2017) constructed an annotated dataset of hateful comments from Fox News articles by incorporating article titles and preceding comments. However, as noticed by Pavlopoulos *et al.* (2020), this innovative approach faced limitations, including a small sample size, lack of reproducibility, and reliance on a single annotator, which may compromise the reliability of the findings.

Similarly, Mubarak *et al.* (2017) provided annotators with the titles of relevant news articles but omitted parent comments, thereby missing a significant contextual component that could inform the interpretation of toxicity. Building on this emphasis on context, Pavlopoulos *et al.* (2017) employed professional moderators to evaluate entire comment threads, allowing for a more comprehensive assessment of toxicity. Nonetheless, their dataset lacked the precise text of the comments, complicating further analysis and replication of the findings.

In a subsequent study, Pavlopoulos *et al.* (2020) explored the influence of context on human judgments of toxicity and classifier performance using a dataset of Wikipedia Talk Pages. Their findings indicated that while context could amplify or mitigate perceived toxicity, it did not enhance classifier accuracy, suggesting a need for larger, context-aware datasets. Contrary to that, Xenos *et al.* (2021) categorized each entry in their dataset based on "context sensitiveness", determined by comparing annotators who had access to the previous (parent) post with those who did not. Their findings showed that classifiers performed better on posts that were more context-sensitive. Similarly, Zhou *et al.* (2023) demonstrated that context-aware models surpass context-agnostic models in hate speech assessment using a

dataset of 33k offensive statements with machine-generated contexts. Nonetheless, the reliance on GPT-3 generated data raises concerns about biases and stereotype reinforcement, suggesting a need for more nuanced methodologies in future research.

Our analysis further emphasizes the significance of contextual understanding in this domain, as illustrated in Table 2.

Text	**Label**
"so you admit being a woman"	hate
"chris jones is gay"	toxic
"learning about with an earthquake example"	abusive
"that's because you are an old man"	offensive
"how is this racist"	toxic

Table 2: Examples of annotated data as SUD lacking context

Instances classified as hate speech or toxic, like "so you admit being a woman" or "Chris Jones is gay", highlight this importance. Without context, assessing their meaning and potential harm is challenging, even for human interpreters, let alone ML algorithms. Moreover, seemingly harmless phrases like "learning about with an earthquake example" are annotated as abusive. Similarly, instances like "that's because you are an old man" and "how is this racist" are labeled offensive and toxic, respectively. In these cases, the lack of context impedes the ability to grasp the implications of the statements making it necessary to incorporate context to the annotation process.

Our proposal, therefore, emphasizes incorporating in the corpus each instance's textual context, originating from parent posts, hashtags, replies and mentions, among others. Systematically considering such cues ensures a comprehensive dataset while excluding them may weaken automatic detection models' effectiveness, leading to unreliable outcomes.

3.2.2. Multimodal context

Research on multimodal hate speech, which encompasses the integration of images and text, remains sparse. However, studies have stressed and demonstrated that incorporating image features can significantly enhance the detection of hate speech and cyberbullying (Yang *et al.*, 2019; Hosseinmardi *et al.*, 2015; Sheth *et al.*, 2022). These findings indicate that the integration of visual content improves classification accuracy and facilitates the identification of cyberbullying instances more effectively. Mostafazadeh *et al.* (2017)

further emphasized the impact of sharing images on social media, highlighting the role of visual context in communication. This stresses the significance of multimodality, integrating text, images, and videos for effective meaning conveyance. Jackiewicz's (2018) research further explores the dynamics between text and image, revealing nuances of complementarity, redundancy, and opposition. Understanding and managing these elements is essential in navigating communication channels. Building on this, Kiela *et al.* (2020) developed a challenging dataset for detecting hate speech in multimodal memes, where unimodal models are insufficient, and only those that integrate both text and image can succeed. By including complex examples that require a deeper level of analysis, they showed that current state-of-the-art models significantly underperform compared to humans. The imperative integration of multimodal context within ML algorithms is exemplified in Figure 2.

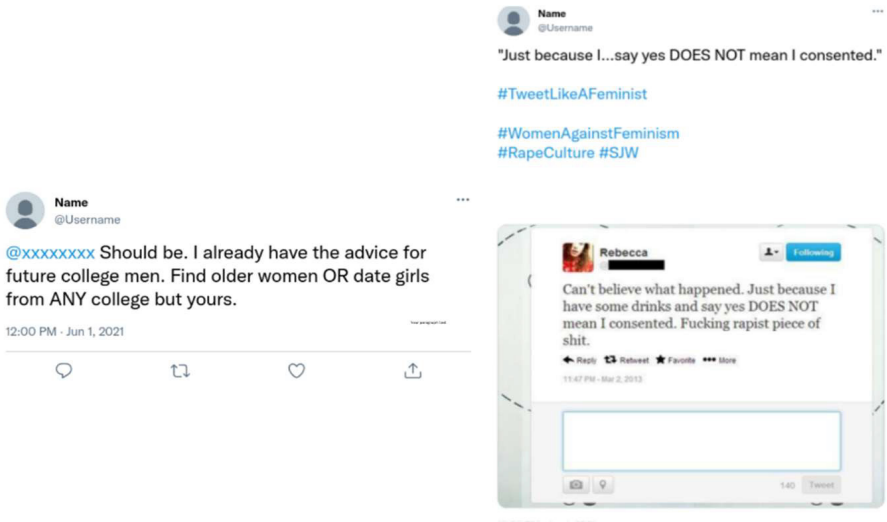

Figure 2: Example of an out-of-context tweet: (a) initial tweet; (b) parent tweet

Consider the tweet of Figure 2a, which is annotated as sexism. Isolating this tweet for classification by a model cannot guarantee its accurate categorization due to the absence of context. However, when analyzed in conjunction with its parent tweet (Figure 2b), the importance of contextual information becomes apparent as the tweet in question is understood as an attempt to invalidate the experience of a potential victim of sexual abuse. The parent tweet, being multimodal, combines textual and visual elements, necessitating specific annotations. Without this context, algorithms might misclassify tweets, missing

their intended meaning. According to Vidak and Jackiewicz (2016), multimodal textual tools, such as hashtags, hyperlinks, and multimedia integrations, play a vital role in expanding the constrained tweet format while also conveying emotions, biases, and sparking controversy. These tools not only serve technical purposes on Twitter but frequently take on syntactic and discursive roles, much like words, phrases, or even complete sentences. For instance, spontaneous hashtags can express viewpoints, biases, or emotions without necessarily forming a thematic thread. Additionally, Stowe *et al.* (2018) advocate for analyzing a user's entire stream and behavior within event contexts, emphasizing the importance of considering both preceding and subsequent tweets for context, while the study of Yu *et al.* (2023) provides insights into the characteristics of replies to real hate speech content in online discourses. Expanding on this, we propose that annotation schemas should actively incorporate multimodal cues, such as annotated images or videos along with multimodal textual tools already addressed in previous studies, to enhance contextual comprehension. By embracing these cues, ML models can better capture details often overlooked in unimodal analyses.

3.2.3. Community based context

In SUD detection, understanding community-based discourse dynamics is crucial for attributing speech acts to specific social groups. This understanding can be enriched through the concept of homophily, a phenomenon observed in real-world and online interactions, wherein individuals tend to associate with those who exhibit perceived similarities (Mishra *et al.*, 2019). These similarities encompass various dimensions, including geographical location, age, language, among others. Mishra *et al.* (2019) proposed that techniques aimed at constructing community graphs, using information about a defined community, facilitate inference about individuals within that community and enhance the performance of automatic detection models. In their study, they conceptualized user communities as graphs and conducted experiments employing classification methods for tweets. Their findings suggested improvements in classification when author profiling through community graphs was employed. Similarly, Sheth *et al.* (2022) proposed a Knowledge-infused Learning framework (K-iL), which enables the model to identify varying interpretations of toxic concepts from diverse perspectives, thereby minimizing ambiguity. This framework is designed to tackle multiple analytical levels, including content, individual, and community dimensions, ensuring that individual-level details evolve in response to interactions within their network.

Unsvåg & Gambäck (2018) investigated how incorporating user-related features (like followers, friends, activity, and profile details) could

improve the accuracy of hate speech detection on Twitter. The authors analyzed datasets in English, Portuguese, and German, and found that while no strong correlation emerged between user characteristics and hate speech, combining user features with text-based features resulted in slight improvements in detection performance. Network-related features, such as the number of followers and friends, consistently contributed the most to enhancing classifier performance.

Ziems *et al.* (2020) tackled the challenge of detecting cyberbullying in online communities by creating an original annotation framework and a comprehensive dataset of Twitter messages. They defined cyberbullying using five key criteria that account for its social and linguistic complexities, allowing for a nuanced representation of this behavior. The researchers employed a combination of text-based features, such as unigrams and sentiment scores, and innovative social network features to improve classification performance. Their findings emphasize that social dynamics and contextual information are crucial for effectively identifying cyberbullying, demonstrating that traditional text-only models often fall short in capturing the subtleties of harmful intent.

In their study, Nagar *et al.* (2023) propose a novel approach for detecting hate speech on Twitter by combining textual content with social context and user profile information. They use a Variational Graph Auto-encoder to jointly learn unified features based on social networks, language, and user data. This method allows for a more nuanced understanding of hate speech by recognizing the influence of an individual's social circle on their content and showing that social context significantly improves detection accuracy.

In a similar direction, Kurrek *et al.* (2022) explored how incorporating community context can improve the detection of abusive language online. By analyzing Reddit comments containing slurs and using subreddit embeddings to capture community behavior, the authors demonstrated that adding context from the community environment reduces false positives and improves the accuracy of abuse detection models.

An example showcasing the utility of constructing community graphs, found in the research of Dias Oliva *et al.* (2021), addresses the issue of the misclassification of LGBTQ-related terms as toxic by ML algorithms. The tweet of Figure 3 was inaccurately assigned a toxicity level of 90.85%.

Figure 3: Example of a non-hateful tweet classified as toxic

This mislabeling arises from the biases embedded in dataset annotations, where terms associated with LGBTQ identities inadvertently become associated with negativity making the incorporation of community-based information imperative for improved automatic detection results.

As a result, incorporating user information, such as anonymized data regarding followers, following and lists, from platforms such as "X", into annotation schemas could enable the construction of community graphs for SUD detection. By integrating these dimensions of user interactions, annotation schemas can better understand online community dynamics and reduce misclassifications.

3.2.4. Sociopolitical context

In sociolinguistics and linguistic anthropology, discourse analysis necessitates a nuanced understanding of how language is shaped by and situated within various contexts: interactional, local, national, and global (de Fina & Georgakopoulou, 2020). Speakers continuously interpret language through contextualization, influenced by their cultural, political, and historical surroundings. As suggested by de Fina and Georgakopoulou (2020), language is intricately connected to culture and plays a crucial role in sustaining social structures. Thus, to gain a comprehensive understanding of discourse, researchers must investigate the backgrounds of speakers, their everyday experiences, and the broader societal forces that impact them.

In this section, we further delve into the role of sociopolitical context in the SUD detection. We examine a tweet identified by Terkourafi *et al.* (2018) who investigated tweets of Steven Salaita regarding the Israeli-Palestinian conflict. This tweet along with a series of other tweets of him contributed to the revoking of his academic job offer in September 2014. The tweet in question (Figure 4), was posted amid heightened tensions between Israel and Palestine following the abduction of three Israeli teenagers and subsequent military operations during the Palestinian-Israeli conflict.

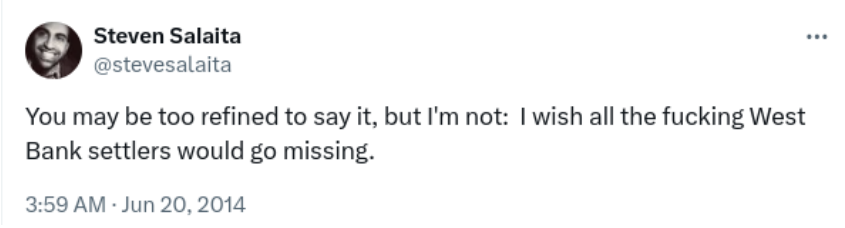

Figure 4: Example of a tweet placed inside a sociopolitical context

This example highlights the challenge for ML models in grasping sociopolitical context and making informed classification decisions.

Therefore, it is advantageous to incorporate in the corpus parent posts, replies to posts, links related to the tweet/instance in question or any other available cues to facilitate the optimal contextual understanding of the model.

3.2.5. Pragmatic context

The final category of context we explore is the pragmatic context. This term serves as an overarching concept encompassing various pragmatic cues that may be embedded within a given context. Examples may include irony, sarcasm, and hyperbole, among others. Studies have highlighted the presence of these cues within platforms like Twitter (e.g., Barbieri & Saggion, 2014; Karoui *et al.*, 2015; Farías *et al.*, 2016). Notably, ElSherief *et al.* (2021) introduced a large-scale benchmark corpus for implicit hate speech (e.g. incitement to violence, inferiority language, irony, stereotypes and misinformation), providing fine-grained labels and enabling better modeling of these nuanced forms of hate. Similarly, Karoui *et al.* (2017) developed a comprehensive annotation schema for irony, which incorporates multiple layers, and applied it to a dataset comprising tweets in French, English, and Italian. In Table 3, we offer an illustration of such instances.

Text	**Label**
"great parliamentary quotes of our times"	hate
"great work at twitter taking black kids ideas and not crediting him or giving h"	abusive
"wow so refreshing and informative it was fun both of you are awesome keep fighting keep anchoring"	aggressive

Table 3: Examples of ironic and sarcastic tweets

Examining the examples in the table reveals that ironic or sarcastic tweets frequently feature terms positively correlated with expressions like 'great', 'refreshing', and 'awesome'. These terms are typically linked with positive sentiments, making it even harder to grasp the intended meaning of the utterance. In the same direction, Paveau (2013b) argued that hashtags on platforms like Twitter often serve not only as indicators of topic or content but also as pragmatic tools to convey emotions and modulate meaning (e.g. #sarcasm, #irony, #humor). Hashtags can add emotional or interpretive layers to statements, functioning as complementary cues that blur the line between expressing subjective states and offering interpretative

guidance. This includes lexical expressions of emotions, with hashtags like #anger, #joy, and #scandalized providing additional layers of meaning beyond the content of the tweet itself (Paveau, 2013b). As a result, we argue that there exists a necessity for broader inclusion of cues capable of indicating pragmatics in annotations, such as emoticons, or hashtags, for an ML model to be able to address these phenomena.

4. Models

To better understand the well acknowledged annotation issues found in datasets we designed a framework of SOTA models differentiating between 3 model families: Shallow Learning Models (SLMs), Masked Language Models (MLMs), and Causal Language Models (CLMs). In the following sections we present the specific characteristics of each category, while details about the hyperparameters of the models can be found in our online repository (Niaouri *et al.*, 2024).

4.1. Shallow Learning Models

Shallow learning models, defined as a category encompassing traditional ML algorithms proposed before 2006, are characterized by their simplicity, typically featuring few layers or processing units (Xu *et al.*, 2021). These models are well-suited for tasks with straightforward data patterns. However, their basic architecture may limit their capacity to capture complex relationships and adapt to new data. Consequently, the performance of such models heavily relies on the efficacy of the feature extraction process (Janiesch *et al.*, 2021). Within this overarching classification, we specifically explore Gradient Boosting (GB) (Friedman, 2001), Logistic Regression (LR) (Wright, 1995), Multinomial Naive Bayes (MNB) (Kibriya *et al.*, 2004), Random Forest (RF) (Breiman, 2001), and Support Vector Machines (SVM) (Hearst *et al.*, 1998).

4.2. Masked Language models

Masked language models (MLMs), as described in Devlin *et al.* (2019), are deep learning models that have been trained to fill in the blanks for masked tokens in a given input sequence. Specifically, MLMs aim to predict the original vocabulary identity of a masked word, relying solely on the context provided by surrounding words. The key advantage of these models is their ability to consider both preceding and subsequent tokens in the input sequence, enabling a bidirectional understanding during the prediction process. Masked Language models are acclaimed for their high performances in classification tasks. Within this category, we finetune and assess the performance of $\mathbf{BERT}_{\mathbf{BASE}}$ (Devlin *et al.*,

2019; Yuan and Rizoiu, 2022) and some of its architectural variants introduced to enhance overall performance and reduce computational complexity, namely **ALBERT$_{\text{BASE}}$** (Lan *et al.*, 2019), **RoBERTa$_{\text{BASE}}$** (Liu *et al.*, 2019) and **ELECTRA$_{\text{BASE}}$** (Clark *et al.*, 2020).

4.3. Causal Language models

As explained in the previous section, MLMs are bidirectional models trained to comprehend context from both directions. In contrast, CLMs are unidirectional models that only consider the preceding context for predictions. CLMs are trained to anticipate the next token in a sequence solely based on prior tokens, making them particularly adept at text generation tasks. The CLM models fine-tuned and evaluated in this study are **Llama 2 (Llama-2-7b-hf)** (Touvron *et al.* 2023), **Mistral (Mistral-7B-v0.1)** (Jiang *et al.*, 2023) and **MPT (mpt-7b)** (MosaicML NLP team, 2023).

5. Results

We present the main results of our empirical assessment across the three model categories under investigation. Across all our trials, we divided the datasets into three subsets: 80% designated for training, 10% for validation, and 10% for testing. Ensuring methodological transparency and reproducibility, we have furnished the requisite code, datasets, and procedural guidelines in an online repository (Niaouri *et al.*, 2024). Table 4 contains the optimal performing model per class and dataset.

	Macro F1 Score												**Best model**
	Abusive	**Aggressive**	**Hate**	**Identity Hate**	**Insult**	**Neither**	**Obscene**	**Offensive**	**Profane**	**Severe Toxic**	**Threat**	**Toxic**	
G^{SUD}	0.79 0.8 0.8	0.64 0.64 0.67	0.6 0.6 0.68	0.36 0.38 0.42	0.5 0.51 0.5	0.94 0.94 0.94	0.25 0.34 0.25	0.75 0.75 0.75	0.31 0.33 0.37	0.40 0.42 0.42	0.43 0.46 0.46	0.18 0.2 0.17	**BERT** **ELECTRA** **RoBERTa**
Davidson	-	-	0.46	-	-	0.9	-	0.94	-	-	-	-	ELECTRA
Founta	0.89	-	0.42	-	-	0.91	-	-	-	-	-	-	MISTRAL
Fox	-	-	0.67	-	-	0.82	-	-	-	-	-	-	MISTRAL
Gab	- -	- -	0.89 0.88 0.89	- - -	- - -	0.91 0.91 0.91	- - -	- - -	- - -	- - -	- - -	- - -	GB ALBERT RoBERTa
Grimminger	-	-	0.58	-	-	0.95	-	-	-	-	-	-	ELECTRA
HASOC2019	-	-	0.29	-	-	0.8	-	0.36	0.57	-	-	-	ELECTRA
HASOC2020	-	-	0.22	-	-	0.91	-	0.3	0.83	-	-	-	ELECTRA

Hateval	- - -	- - -	0.75 0.75 0.76	- - -	- - -	0.79 0.8 0.78	- - -	- - -	- - -	- - -	- - -	- - -	ELECTRA RoBERTa MISTRAL
Jigsaw	-	-	-	0.46	0.57	0.98	0.38	-	-	0.4	0.56	0.3	ELECTRA
Olid	- -	- -	- -	- -	- -	0.85 0.84	- -	0.67 0.68	- -	- -	- -	- -	BERT ELECTRA
Reddit	- -	- -	0.77 0.78	- -	- -	0.92 0.93	- -	- -	- -	- -	- -	-	LLAMA 2 MISTRAL
Storm-front	-	-	0.6	-	-	0.96	-	-	-	-	-	-	RoBERTa
Trac	-	0.84	-	-	-	0.71	-	-	-	-	-	-	MISTRAL

Table 4: Optimal performing model per class and dataset

ELECTRA consistently emerges as the leading performer, as evidenced by its attainment of the highest Macro F1 score across eight datasets, notably including the G^{SUD} corpus. Comparative evaluation of BERT variants alongside the original BERT model reveals a superiority of ELECTRA and RoBERTa in differentiating SUD classes. While Mistral demonstrates competitive efficacy, its performance in the G^{SUD} dataset fails to meet anticipated levels of generalization capability. Our findings also reveal unstable classification outcomes within the *hate* and *offensive* classes (majority classes) and low performances for the underrepresented classes (i.e., *severe toxic*, *threat*, and *toxic*). Among all models assessed, the shallow learning models exhibited the most inferior performance across datasets.

6. Empirical Analysis of Machine Learning Models Performance

In this section, we present a detailed qualitative analysis of the best performing model among eleven different SOTA ML techniques falling under two categories, namely shallow learning models and large language models based on deep learning.

We note that the large language model ELECTRABASE (Clark *et al.*, 2020) consistently emerges as the dominant performer compared to other eleven SOTA ML models that we evaluated, as it obtains the highest Macro F1 score (0.54% in the G^{SUD} corpus) across seven datasets (over thirteen, see Table 1). Further details of our empirical analysis and comparison can be found in Niaouri *et al.* (in press).

6.1. Error Analysis and Discriminatory Ability Assessment of ELECTRA

Here we present the results of our error analysis conducted over SUD detection using the ELECTRA model. In Figure 5, we provide a confusion matrix computed on our test set to enhance comprehension of which classes were frequently misclassified. The y-axis denotes the actual labels, while the x-axis represents the predicted ones.

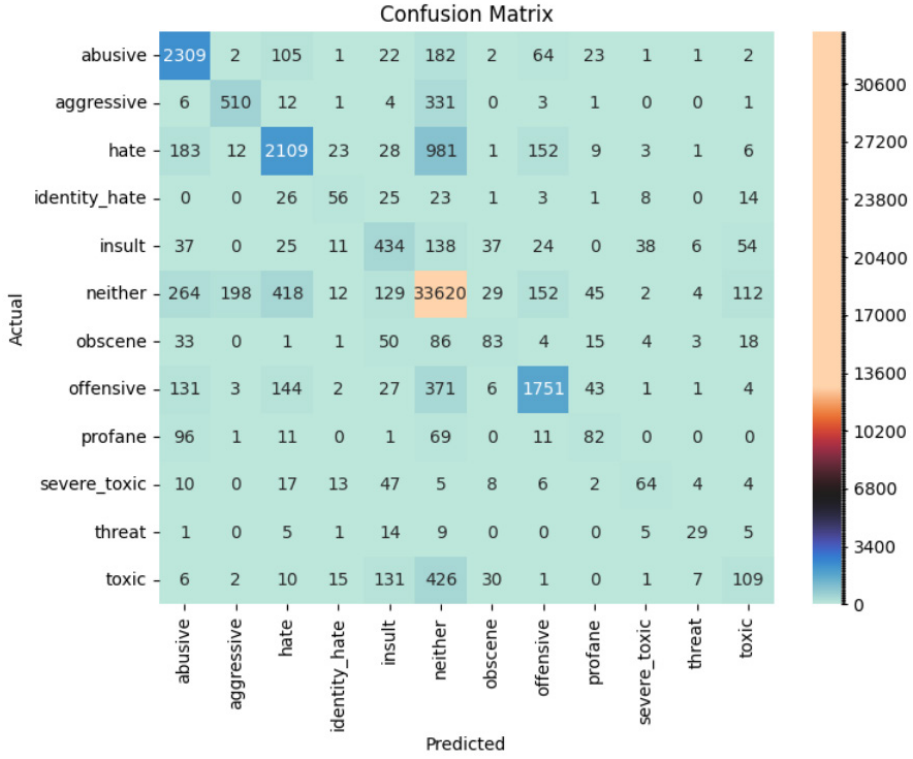

Figure 5: Confusing matrix of multi-class SUD classification

We note that the *abusive* class is predominantly predicted accurately alongside the *neither* and *offensive* classes. Also, in most instances, the SUD classes are primarily misclassified as the *neither* class, indicating a probable lack of contextual understanding. Instances within the *obscene* category exhibit a higher inclination to be classified as *neither* rather than their correct class. Similarly, the *toxic* class tends to be misclassified as *neither*. Moreover, instances arise wherein SUD classes are inaccurately categorized as other SUD classes. For instance, the term *profane* is more frequently misclassified as *abusive* rather than its appropriate class.

To better understand the discriminatory ability of ELECTRA we visualized the textual representation produced by the model focusing specifically on the output of its output pooled layer. We employed t-distributed Stochastic Neighbor Embedding (t-SNE) to reduce the dimensionality of the output to two dimensions. The resulting plot (Figure 6) depicts the outcomes of the testing set under the G^{SUD} training scenario.

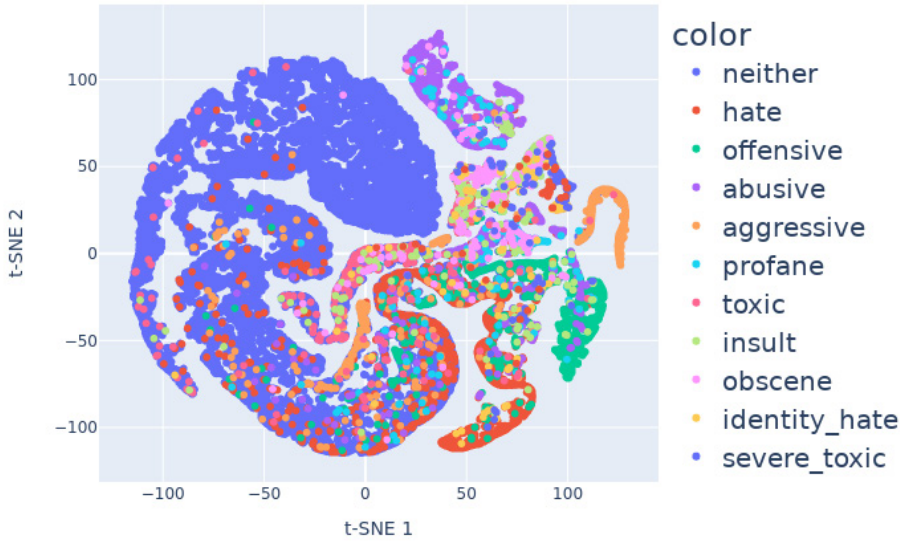

Figure 6: Two components t-SNE visualization of samples embedding produced by ELECTRA output pooled layer

Observations suggest that classes, such as *abusive* and *aggressive*, exhibit clear clusters, indicating a pronounced separation within the dataset. This pattern highlights the exclusive occurrence of these categories within a single dataset. Conversely, classes like *profane*, *obscene*, *threat*, *severe toxic*, and *toxic* display more scattered data points in the plot.

6.2. Error analysis and Confidence Levels in ELECTRA's Predictions

We further examined the specific errors generated by the model alongside the corresponding levels of confidence associated with these predictions. This involved inspecting our dataset and constructing graphical representations that depicted the confidence levels attributed to erroneous predictions. We aimed to trace the instances wherein the model exhibited higher confidence in its predictions. Scores closer to 1 indicate a higher degree of confidence in the model's predictions, while scores closer to 0 suggest lower levels of confidence.

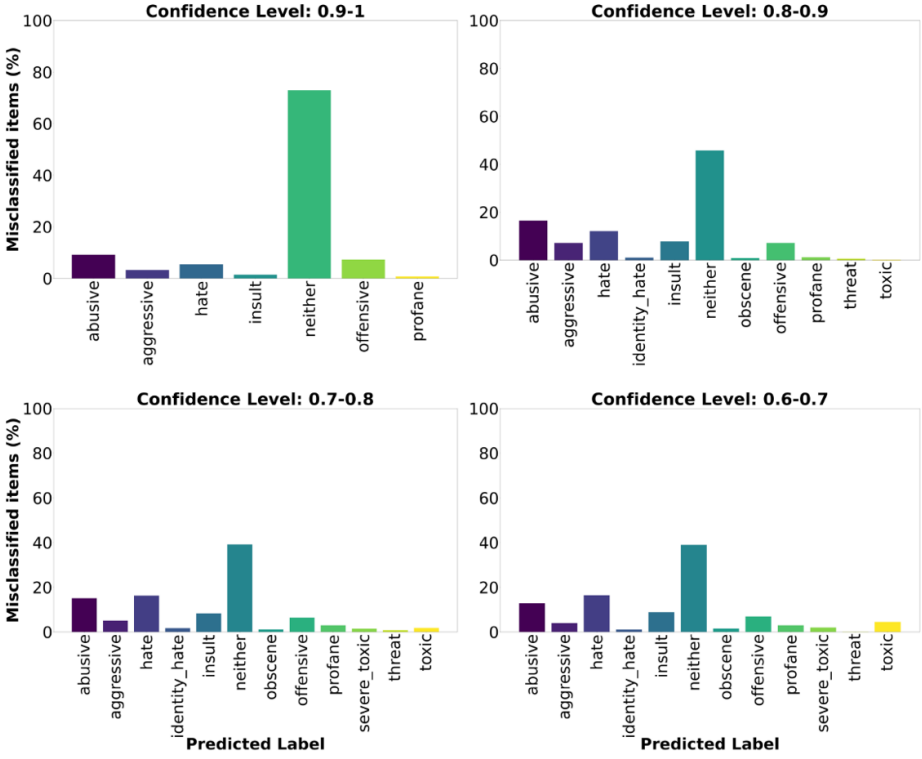

Figure 7: Misclassified instances per class and level of model confidence (0.6-1)

A notable trend emerged: highly confident predictions predominantly corresponded to misclassifications as the *neither* class, despite the ground truth being an SUD class. In Figure 7, within the confidence range of 0.9-1, a significant prevalence of misclassified texts as *neither* was observed, despite the fact that their actual label is an SUD category. Notably, as prediction confidence decreased, a more diverse spectrum of classes was predicted, with a persistent gap between *neither* and SUD classes. Although this gap exhibited a slight reduction as confidence levels decreased, it remains pronounced throughout the examined ranges.

We show an example of misclassified instances in Table 5. Here, the instances are classified as *neither* with a confidence level exceeding 0.9.

Text	Predicted Label	Ground Truth
"first time experience for my son and daughter all the way from Scotland"	neither	abusive
"wow so refreshing and informative it was fun both of you are awesome keep fighting keep anchoring"	neither	aggressive
"its a nice thought"	neither	hate
"lol"	neither	hate
"i never talked about gun control in any form"	neither	offensive
"white house washigton dc"	neither	profane

Table 5: Misclassified instances as *neither* with high level of confidence (over 0.9)

Analysis of these examples reveals a main issue across the sampled instances: the absence of context. This absence poses challenges to accurately interpret the intended meaning within these texts. Challenges may include sociopolitical references, emotional nuances, and situational descriptions, among others. Sociopolitical references, such as the mention of "White House, Washington DC", inherently carry layers of meaning dependent on broader sociopolitical context apart from conversational context. Similarly, statements like "I never talked about gun control in any form" necessitate an understanding of the discourse surrounding gun control, the speaker's position, and the underlying conversation context. Moreover, contextual cues can manifest in subtler forms, such as irony (e.g., "wow so refreshing … keep anchoring"), requiring comprehension of rhetorical devices and pragmatic conventions to understand the intended meaning.

We further examined misclassifications made by the model that confidently assigned texts with a wrong SUD label. The high confidence of these predictions suggests that the model may have developed clear patterns leading to such classifications. To analyze these misclassifications, we selected the 10 most frequently attested tokens per class, considering instances where the model's confidence exceeded 0.9 after preprocessing our data by removing stopwords. Figure 8 presents the most frequently attested tokens per class in utterances misclassified as the wrong SUD class.

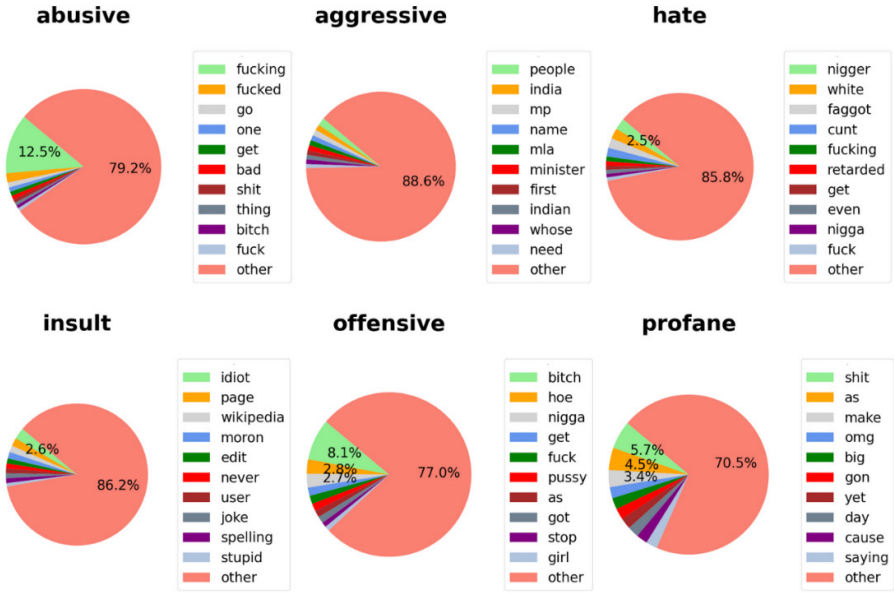

Figure 8: Top 10 most frequently attested words per class misclassified as the wrong SUD class

Misclassified instances predominantly classified as ***abusive*** were found to contain a significant proportion of tokens featuring the lemma "f*ck", comprising ~15% of the tokens of this class. Examples such as "squidwards a f*cking philosopher" illustrate this pattern (see Table 6). Additionally, terms such as "b*tch" and "s*ht" were also notable in these texts as the examples "b*tches be crazy" and "this sh*t would be insane" suggest. In the ***aggressive*** class, the lemma "India" was also among the most frequently occurring tokens, as exemplified by phrases like "they are ... in India ... burhan". It's worth noting that the aggressive class stemmed from a single dataset, potentially indicating a specialized content. The ***hate*** class encompassed various instances of racist, homophobic, and sexist content, with words like "n*gger", "n*gga", "f*ggot", and "c*nt" being prevalent as exemplified by instances like: "n*gger shut up sit down", "if you still hate this n*gga", "well do it then b*tch ass f*ggot" and "they sound like soulless c*nts". Instances categorized under the ***insult*** class included terms such as "idiot" (e.g., "you are an idiot ... yourself"), "moron" (e.g., "you dont want ... moron"), and "joke" (e.g., "ur a ... a joke ..."), representing less severe forms of derogatory language. The ***offensive*** class predominantly featured sexist content, including terms like "b*tch" (e.g., "b*tch text me"), "h*e" (e.g., "i cant ... h*e"), "p*ssy" (e.g., "that p*ssy ... a pass"),

and "girl" (e.g., "the same … like girl bye"), along with racially charged terms like "n*gga" (e.g., "dont my n*gga …"). Finally, the ***profane*** class was primarily associated with the term "sh*t", as exemplified by phrases such as "you just can't make this sh*t up".

In Table 6 we provide the previously analysed examples along with information regarding their ground truth label.

Text	Predicted Label	Ground Truth
"squidwards a **f*cking** philosopher"	abusive	hate
"**b*tches** be crazy"	abusive	neither
"this **sh*t** would be insane"	abusive	neither
"they are the worst news channel in **india** they should be banned they call burhan"	aggressive	neither
"**n*gger** shut up sit down"	hate	identity hate
"if you still hate this **n*gga"**	hate	abusive
"well do it then b*tch ass **f*ggot**"	hate	identity hate
"they sound like soulless **c*nts**"	hate	offensive
"you are an **idiot** if you actually think that updating information about his son constitutes vandalism go f*ck yourself"	insult	severe toxic
"you dont want the truth she is and was a tramp stop editing the truth **moron**"	insult	toxic
"ur a f*k head stop deleting it its a **joke** for my friend"	insult	toxic
"**b*tch** text me"	offensive	hate
"i cant f*ck a scary **h*e**"	offensive	hate
"that **p*ssy** is way too old to try for a pass"	offensive	hate
"the same b*tch is all on my boo s shit like **girl** bye"	offensive	neither
"dont my **n*gga** me honkie"	offensive	hate
"you just can t make this **sh*t** up"	profane	abusive

Table 6: Misclassified instances as SUD with high level of confidence (over 0.9)

To see how those erroneous yet certain predictions of the model were based on the patterns learned from the annotated data we conducted the same analysis on the correctly predicted data with prediction confidence over 0.9. In Figure 9 we present the most frequently attested tokens per class classified as the correct SUD class. Under this condition two more classes arose: *identity hate* and *threat*.

Figure 9: Top 10 most frequently attested words per class classified as the correct SUD class

The same patterns as in the previous experimental condition were attested for the classified instances as the correct SUD class. Presented below (Table 7) are examples of accurately predicted labels. Similarly to the previous condition, the most common lemma found in instances under the ***abusive*** class is "f*ck" as exemplified in "... f*cking police". The derogatory term "b*tch" featured prominently among the top 10 frequently occurring terms in this class, evident in examples such as "if i gotta be a b*tch ...". Moving on to the ***aggressive*** class, consistent with earlier findings, the term "India" surfaced frequently, indicating a specific topic within the corpus, as seen in instances like "... indian politician". Similarly, the ***hate*** class was characterized by racist, homophobic, and sexist content, with words like "n*gger" and "f*ggot" commonly appearing, exemplified by phrases such as "faggot niggers". Other frequently used terms are "white" (e.g., "this is the whitest n*gger ive everseen"), "jew" (e.g., "its all pig jew to me"), "retarded" (e.g., "f*cking retarded apes") and "c*nt" (e.g., "the c*nt lost"). Prevalent is the existence of the lemma "f*ck" (e.g. "f*ck off..."). Likewise, the ***identity hate*** class showed a tendency towards racism. Among the 10 most frequently used terms in this class are the terms "black" (e.g., "ut u racist ... cause im black"), "n*gger" and "n*gga" (e.g., "rich n*ggaaaazzz your a n*gger"). In contrast (and similarly to our previous analysis), the ***insult*** class contained less severe derogatory terms like "idiot" (e.g., "get ... page idiot") and "stupid" (e.g., "you ... are stupid"). The class ***offensive***, as before, has been mostly populated by sexist terms such as "b*tch", "h*e", and "p*ssy" as attested in the examples "love a b*tch fight", "you are a h*e h*e amp a h*e", and "the karate kid is such a p*ssy". Other terms we find are "n*gga" (e.g.,

"… i got n*ggas"), and "sh*t" (e.g., "dat … sh*t lls"). Under the class ***profane***, as expected, we find again the term "sh*t" (e.g., "miss … sh*t") and "f*ck" (e.g., "i m this … as f*ck"). Finally, for the last class ***threat***, among the most prominent terms attested is that of "kill" as evidenced in "i am going to kill … you".

Text	Predicted Label	Ground Truth
"im calling the **f*cking** police"	abusive	abusive
"if i gotta be a **b*tch** imma be a bad one"	abusive	abusive
"kill all **indian** politician"	aggressive	aggressive
"**f*ggot n*ggers**"	hate	hate
"this is the **whitest** n*gger ive everseen"	hate	hate
"its all pig **jew** to me"	hate	hate
"f*cking **retarded** apes"	hate	hate
"the **c*nt** lost"	hate	hate
"f*ck off **retard** cuck"	hate	hate
"ut u racist **n*gga** u blocked me cause im black"	identity hate	identity hate
"rich **n*ggaaaazzz** your a **n*gger**"	identity hate	identity hate
"get the hell off my page **idiot**"	insult	insult
"you suck get a life get off wikipedia you are **stupid**"	insult	insult
"love a **b*tch** fight"	offensive	offensive
"you are a **h*e hoe** amp a **h*e**"	offensive	offensive
"the karate kid is such a **p*ssy**"	offensive	offensive
"you got hoes i got **n*ggas**"	offensive	offensive
"dat bitch wild as **sh*t** lls"	offensive	offensive
"miss me with that **sh*t**"	profane	profane
"i m this old ok we **f*cking** get it yall old as **f*ck**"	profane	profane
"i am going to **kill** you i am going to murder you"	threat	threat

Table 7: Correctly classified instances as SUD with high level of confidence (over 0.9)

The analysis reveals how annotated data influences the model's prediction patterns. The similarity between correctly classified SUD instances and their misclassifications indicates the influence of patterns in the annotated data on model errors. Specific terms within the data influenced the model's decisions, but this wasn't always

sufficient for accurate predictions. A critical factor for these errors is likely the lack of context in the annotated data. Context is crucial for disentangling different types of discourse that may share common linguistic forms and biases but belong to distinct classes. As a result, the interplay between shared linguistic forms across different classes, combined with the complexities of different annotation schemas, highlights the need for enhanced contextual understanding.

7. Conclusions and suggestions for future work

The findings presented in this paper underscore the critical role of contextual understanding in the detection of SUD in online platforms using ML algorithms. By emphasizing the influence of textual, multimodal, community-based, sociopolitical, and pragmatic contexts on the detection process, our study highlights the necessity for balanced datasets and clear annotation practices to mitigate unintended biases. We propose integrating contextual indicators such as hashtags, emoticons, mentions, parent posts and links, among others, and user information, such as anonymized data regarding followers and following, into annotation schemas to enhance the effectiveness of ML algorithms. Moreover, our error analysis reveals challenges in model generalization due to overlapping features within discourse, prompting the need for well-defined multi-class configurations that accurately reflect real-world scenarios.

As future research, we acknowledge the importance of incorporating contextual information into our analysis. While our study considers a large-scale dataset, we require the annotation of contextual elements to study their benefit to SUD detection. As noted, prior research, such as the work by Pavlopoulos *et al.* (2020), has examined the role of context in a specific and low-scale scenario. In future studies, we aim to validate the critical role of context by utilizing larger and context-rich datasets.

References

Aroyehun, S. T., Gelbukh, A. (2018), “Aggression Detection in Social Media: Using Deep Neural Networks, Data Augmentation, and Pseudo Labeling”, *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, Santa Fe, New Mexico, p. 90-97, https://aclanthology.org/W18-4411.

Badjatiya, P., Gupta, M., Varma, V. (2019), « Stereotypical Bias Removal for Hate Speech Detection Task using Knowledge-based Generalizations”, *The World Wide Web Conference*, p. 49-59, https://dl.acm.org/doi/pdf/10.1145/3308558.3313504.

Barbieri, F., Saggion, H. (2014), “Modelling Irony in Twitter”, *Conference of the European Chapter of the Association for Computational Linguistics*, https://doi.org/10.3115/v1/E14-3007.

Breiman, L. (2001), "Random Forests", *Machine Learning*, 45, p. 5-32, https://doi.org/10.1023/A:1010933404324.

Calderón, C. A., de la Vega, G., Herrero, D. B. (2020), "Topic modeling and characterization of hate speech against immigrants on Twitter around the emergence of a far-right party in Spain", *Social Sciences*, 9/11, 188. https://doi.org/10.3390/socsci9110188.

Carneiro, B. M., Linardi, M., Longhi, J. (2023), "Studying Socially Unacceptable Discourse Classification (SUD) through different eyes: 'Are we on the same page?' ", *Proceedings of the 10th International Conference on CMC and Social Media Corpora for the Humanities (CMC-Corpora 2023),* 14-15 September 2023, University of Mannheim, Germany, https://ids-pub.bsz-bw.de/frontdoor/index/index/docId/12095.

Clark, K., Luong, M. T., Le, Q. V., Manning, C. D. (2020), "Electra: Pre-training text encoders as discriminators rather than generators", arXiv:2003.10555, https://doi.org/10.48550/arXiv.2003.10555.

Coe, K., Kenski, K., Rains, S. A. (2014), "Online and uncivil? Patterns and determinants of incivility in newspaper website comments", *Journal of communication*, 64/4, p. 658-679, https://doi.org/10.1111/jcom.12104.

Davani, A. M., Atari, M., Kennedy, B., Dehghani, M. (2023), "Hate speech classifiers learn normative social stereotypes", *Transactions of the Association for Computational Linguistics,* 11, p. 300-319, https://doi.org/10.1162/tacl_a_00550.

Davidson, T., Bhattacharya, D., Weber, I. (2019), "Racial bias in hate speech and abusive language detection datasets", *Proceedings of the Third Workshop on Abusive Language Online*, https://doi.org/10.18653/v1/W19-3504.

Davidson, T., Warmsley, D., Macy, M.W., 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), p. 512-515, https://doi.org/10.1609/icwsm.v11i1.14955.

de Gibert, O., Pérez, N., García-Pablos, A., Cuadros, M. (2018), "Hate Speech Dataset from a White Supremacy Forum", *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, Brussels, p.11-20, https://doi.org/10.18653/v1/W18-5102.

de Fina, A., Georgakopoulou, A. (eds) (2020), *The Cambridge handbook of discourse studies*, Cambridge University Press.

de Maiti, K., Fišer, D. (2021), "Working with socially unacceptable discourse online: Researchers' perspective on distressing data", *Proceedings of the 8th Conference on CMC and Social Media Corpora for the Humanities (CMC-Corpora 2021)*, p. 78-82.

Devlin, J., Chang, M., Lee, K., Toutanova, K. (2019), "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", *Proceedings of the 2019 Conference of North American Chapter of the Association for Computational Linguistics*, p. 4171-4186, https://doi.org/10.18653/v1/N19-1423.

Dey, A. K. (2001), "Understanding and Using Context", *Personal and Ubiquitous Computing*, 5/4-7, https://doi.org/10.1007/s007790170019.

Dias Oliva, T., Antonialli, D.M., Gomes, A. (2021), "Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online", *Sexuality & Culture*, 25, p. 700-732, https://doi.org/10.1007/s12119-020-09790-w.

Elsherief, M., Ziems, C., Muchlinski, D. A., Anupindi, V., Seybolt, J., Choudhury, M. D., Yang, D. (2021), "Latent Hatred: A Benchmark for Understanding Implicit Hate Speech", *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 345-363, https://doi.org/10.18653/v1/2021.emnlp-main.29.

Farías, D. I., Patti, V., Rosso, P. (2016), "Irony Detection in Twitter", *ACM Transactions on Internet Technology (TOIT)*, 16, p. 1-24, https://doi.org/10.1145/2930663.

Fišer, D., Erjavec, T., Ljubesic, N. (2017), "Legal Framework, Dataset and Annotation Schema for Socially Unacceptable Online Discourse Practices in Slovene", *Proceedings of the First Workshop on Abusive Language Online*, Vancouver, p. 46-51, https://doi.org/10.18653/v1/W17-3007.

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.

Friedman, J. H. (2001), "Greedy function approximation: a gradient boosting machine", *Annals of statistics*, 29/5, p. 1189-1232.

Gao, L., Huang, R. (2017), "Detecting online hate speech using context aware models", *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, Varna, p. 260-266, https://doi.org/10.26615/978-954-452-049-6_036.

Ghosh Roy, S., Narayan, U., Raha, T., Abid, Z., Varma, V. (2021), "Leveraging Multilingual Transformers for Hate Speech Detection", arXiv:2101.03207, https://doi.org/10.48550/arXiv.2101.03207.

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", *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 171-180, https://aclanthology.org/2021.wassa-1.18.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., Scholkopf, B. (1998), "Support vector machines", *IEEE Intelligent Systems and their applications*, 13/4, p. 18-28.

Hosseinmardi, H., Mattson, S. A., Rafiq, R. I., Han, R., Lv, Q., Mishra, S. (2015), "Detection of cyberbullying incidents on the instagram social network", *arXiv preprint arXiv:1503.03909*, https://doi.org/10.48550/arXiv.1503.03909.

Jackiewicz, A. (2018) « De l'hypertextualité dans des tweets polémiques », *Le discours hypertextualisé. Espaces énonciatifs mosaïques* (ch 4), https://hal.science/hal-01839465.

Janiesch, C., Zschech, P., Heinrich, K. (2021), "Machine learning and deep learning", *Electronic Markets*, 31/3, p. 685-695, https://doi.org/10.1007/s12525-021-00475-2.

Jiang, A. *et al.* (2023), Mistral 7B, https://doi.org/10.48550/arXiv.2310.06825.

Karoui, J., Benamara, F., Moriceau, V., Aussenac-Gilles, N., Belguith, L. H. (2015), "Towards a Contextual Pragmatic Model to Detect Irony in Tweets", *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (Volume 2: Short Papers), Beijing, p. 644-650, https://doi.org/10.3115/v1/P15-2106.

Karoui, J., Benamara, F., Moriceau, V., Patti, V., Bosco, C., Aussenac-Gilles, N. (2017), "Exploring the Impact of pragmatic phenomena on irony detection in tweets: a multilingual corpus study", *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistic*s, Volume 1, Long Papers, Valencia, p. 262-272, https://aclanthology.org/E17-1025.

Kenski, K., Coe, K., Rains, S. A. (2020), "Perceptions of uncivil discourse online: An examination of types and predictors", *Communication research*, 47/6, p. 795-814, https://doi.org/10.1177/0093650217699933.

Kibriya, A. M., Frank, E., Pfahringer, B., Holmes, G. (2004), "Multinomial Naive Bayes for Text Categorization Revisited", in Webb, G. I., Yu, X. (eds), *AI 2004: Advances in Artificial Intelligence. AI 2004. Lecture Notes in Computer Science,* vol 3339, Springer, Berlin, https://doi.org/10.1007/978-3-540-30549-1_43.

Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., Testuggine, D. (2020), "The hateful memes challenge: Detecting hate speech in multimodal memes", *Advances in neural information processing systems*, 33, p. 2611-2624, https://doi.org/10.48550/arXiv.2005.04790.

Kurrek, J., Saleem, H. M., Ruths, D. (2022), "Enriching abusive language detection with community context", arXiv preprint arXiv:2206.08445, https://doi.org/10.48550/arXiv.2206.08445.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R. (2019), "Albert: A lite bert for self-supervised learning of language representations", arXiv preprint arXiv:1909.11942, https://doi.org/10.48550/arXiv.1909.11942.

Lin, Z. (2021), "A Methodological Review of Machine Learning in Applied Linguistics", *English Language Teaching*, 14/1, p. 74-85. https://doi.org/10.5539/elt.v14n1p74.

Linzen, T. (2019), "What can linguistics and deep learning contribute to each other?", arXiv:1809.04179, https://doi.org/10.48550/arXiv.1809.04179.

MacAvaney, S., Yao, H., Yang, E., Russell, K., Goharian, N., Frieder, O. (2019), "Hate speech detection: Challenges and solutions", *PLoS ONE*, 14, https://doi.org/10.1371/journal.pone.0221152.

Mishra, P., Del Tredici, M., Yannakoudakis, H., Shutova, E. (2019), "Author profiling for hate speech detection", *arXiv preprint arXiv:1902.06734*, https://doi.org/10.48550/arXiv.1902.06734.

MosaicML NLP Team (2023), "Introducing MPT-7B: A New Standard for Open-Source, Commercially Usable LLMs", *Databricks*, https://www.databricks.com/blog/mpt-7b.

Mostafazadeh, N., Brockett, C., Dolan, W. B., Galley, M., Gao, J., Spithourakis, G.P., Vanderwende, L. (2017), "Image-Grounded Conversations: Multimodal Context for Natural Question and Response Generation", *Proceedings of the Eighth International Joint Conference on Natural Language Processing* (volume 1: Long Papers), Taipei p. 462-472, https://aclanthology.org/I17-1047.

Mubarak, H., Darwish, K., Magdy, W. (2017), "Abusive Language Detection on Arabic Social Media", *Proceedings of the First Workshop on Abusive Language Online,* Vancouver, p. 52-56, https://doi.org/10.18653/v1/W17-3008.

Nagar, S., Barbhuiya, F. A., Dey, K. (2023), "Towards more robust hate speech

detection: using social context and user data", *Social Network Analysis and Mining*, 13/1, https://doi.org/10.1007/s13278-023-01051-6.

Niaouri, D., Machado Carneiro, B., Linardi, M., Longhi, J. (2024), https://github.com/diniaouri/Machine_Learning_heading_to_SUD.

Niaouri, D., Machado Carneiro, B., Linardi, M., Longhi, J. (in press), "Machine Learning is heading to the SUD (Socially Unacceptable Discourse) analysis: from Shallow Learning to Large Language Models to the rescue, where do we stand?", *Digital linguistics*. De Gryuter.

Nobata, C., Tetreault, J.R., Thomas, A. O., Mehdad, Y., Chang, Y. (2016), "Abusive Language Detection in Online User Content", *Proceedings of the 25th International Conference on World Wide Web*, https://doi.org/10.1145/2872427.2883062.

Okulska, I., Kołos, A. (2023), "A Morpho-syntactic Analysis of Human-moderated Hate Speech Samples from Wykop. pl Web Service", *Półrocznik Językoznawczy Tertium*, 8/2, p. 54-71, https://doi.org/10.7592/Tertium.2023.8.2.245.

Paveau, M. A. (2013a), « Tweet [Dictionnaire] », *Technologies discursives*, https://doi.org/10.58079/uowx.

Paveau, M. A. (2013b), « Technodiscursivités natives sur Twitter. Une écologie du discours numérique », *Epistémè: revue internationale de sciences humaines et sociales appliquées*, 9, p. 139-176, https://doi.org/10.4000/pratiques.3533.

Pavlopoulos, J., Malakasiotis, P., Androutsopoulos, I. (2017), "Deep learning for user comment moderation", *arXiv preprint arXiv:1705.09993*, https://doi.org/10.48550/arXiv.1705.09993.

Pavlopoulos, J., Sorensen, J., Dixon, L., Thain, N., Androutsopoulos, I. (2020), "Toxicity detection: Does context really matter?", *arXiv preprint arXiv:2006.00998*, https://doi.org/10.48550/arXiv.2006.00998.

Perifanos, K., Goutsos, D. (2021), "Multimodal hate speech detection in Greek social media", *Multimodal Technologies and Interaction*, 5/7, 34, https://doi.org/10.3390/mti5070034.

Postigo-Fuentes, A. Y., Kailuweit R., Ziem A., Hartmann S. (2024), "Defining Extremist Narratives: A review of the current state of the art", in *HORIZON – CL2-2022-DEMOCRACY -01-05* [Report], Momentum Consulting, https://arenasproject.eu/download/1545/?tmstv=1721660114.

Potter, J., Wetherell, M. (1988), "Accomplishing attitudes: Fact and evaluation in racist discourse", *Text-Interdisciplinary Journal for the Study of Discourse*, 8/1-2, p. 51-68, https://doi.org/10.1515/text.1.1988.8.1-2.51.

Purohit, H., Shalin, V. L., Sheth, A. (2020), "Knowledge Graphs to Empower Humanity-Inspired AI Systems", *IEEE Internet Computing*, 24, p. 48-54, https://doi.org/10.1109/MIC.2020.3013683.

Qian, J., Bethke, A., Liu, Y., Belding-Royer, E. M., Wang, W.Y. (2019), "A Benchmark Dataset for Learning to Intervene in Online Hate Speech", *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP*), Hong Kong, p. 4755-4764, https://doi.org/10.18653/v1/D19-1482.

Rossini, P. G. (2020), "Beyond Incivility: Understanding Patterns of Uncivil and Intolerant Discourse in Online Political Talk", *Communication Research*, 49, p. 399-425, https://doi.org/10.1177/0093650220921314.

Sheth, A., Shalin, V. L., Kursuncu, U. (2022), "Defining and detecting toxicity on social media: context and knowledge are key", *Neurocomputing*, 490, p. 312-318, https://doi.org/10.48550/arXiv.2104.10788.

Slack, D., Krishna, S., Lakkaraju, H., Singh, S. (2023), "Explaining machine learning models with interactive natural language conversations using TalkToModel", *Nature Machine Intelligence*, 5/8, p. 873-883, https://doi.org/10.1038/s42256-023-00692-8.

Stairmand, M. A. (1997), "Textual context analysis for information retrieval", in *Proceedings of the 20th annual international ACM SIGIR conference on research and development in information retrieval*, p. 140-147, https://doi.org/10.1145/258525.25855.

Stowe, K., Palmer, M., Anderson, J., Kogan, M., Palen, L., Anderson, K. M., Morss R., Demuth J. Lazrus, H. (2018), "Developing and evaluating annotation procedures for twitter data during hazard events", *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, p. 133-143, https://aclanthology.org/W18-4915.

Sulc, A. de Maiti, K. P. (2020), "No room for hate: What research about hate speech taught us about collaboration?", *TwinTalks@DH/DHN.*

Terkourafi, M., Catedral, L., Haider, I., Karimzad, F., Melgares, J., Mostacero-Pinilla, C., Nelson, J., Weissman, B. (2018), "Uncivil Twitter: A sociopragmatic analysis", *Journal of Language Aggression and Conflict*, 6/1, p. 26-57, https://doi.org/10.1017/9781108954105.024.

Touvron, H. *et al.* (2023), "LLaMA: Open and Efficient Foundation Language Models", arXiv:2302.13971, https://doi.org/10.48550/arXiv.2302.13971.

Unsvåg, E. F., Gambäck, B. (2018), "The effects of user features on Twitter hate speech detection", *Proceedings of the 2nd workshop on abusive language online (ALW2)*, p. 75-85, https://doi.org/10.18653/v1/W18-5110.

Van Aken, B., Risch, J., Krestel, R., Löser, A. (2018), "Challenges for toxic comment classification: An in-depth error analysis", *arXiv preprint arXiv:1809.07572*, https://doi.org/10.18653/v1/W18-5105.

Vehovar, V., Povž, B., Fišer, D., Ljubešić, N., Šulc, A., Jontes, D. (2020), "Družbeno nesprejemljivi diskurz na Facebookovih straneh novičarskih portalov", *Teorija in praksa*, 57/2, p. 622-645.

Vidak, M., Jackiewicz, A. (2016), « Les outils multimodaux de Twitter comme moyens d'expression des émotions et des prises de position », *Cahiers de praxématique*, 66, https://doi.org/10.4000/praxematique.4247.

Vidgen, B., Nguyen, D., Margetts, H. Z., Rossini, P.G., Tromble, R. (2021), "Introducing CAD: the Contextual Abuse Dataset", *North American Chapter of the Association for Computational Linguistics*, https://doi.org/10.18653/V1/2021.NAACL-MAIN.182.

Wang, B., Ding, Y., Liu, S., Zhou, X. (2019), "YNU_Wb at HASOC 2019: Ordered Neurons LSTM with Attention for Identifying Hate Speech and Offensive Language", *Fire*, https://ceur-ws.org/Vol-2517/T3-2.pdf.

Waseem, Z., Hovy, D. (2016), "Hateful symbols or hateful people? predictive features for hate speech detection on twitter", *Proceedings of the NAACL student research workshop*, p. 88-93, https://doi.org/10.18653/v1/N16-2013.

Wright, R. E. (1995), "Logistic regression", in Grimm, L. G., Yarnold, P. R. (eds), *Reading and understanding multivariate statistics*, p. 217-244.

Wulczyn, E., Thain, N., Dixon, L. (2017), "Ex machina: Personal attacks seen at scale", *Proceedings of the 26th international conference on world wide web*, p. 1391-1399, https://doi.org/10.48550/arXiv.1610.08914.

Xenos, A., Pavlopoulos, J., Androutsopoulos, I. (2021), "Context sensitivity estimation in toxicity detection", *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, p. 140-145, https://doi.org/10.18653/v1/2021.woah-1.15.

Xu, Y., Zhou, Y., Sekula, P., Ding, L. (2021), "Machine learning in construction: From shallow to deep learning", *Developments in the built environment*, 6, 100045. https://doi.org/10.1016/j.dibe.2021.100045.

Yang, F., Peng, X., Ghosh, G., Shilon, R., Ma, H., Moore, E., Predović, G. (2019), "Exploring Deep Multimodal Fusion of Text and Photo for Hate Speech Classification", *Proceedings of the Third Workshop on Abusive Language Online*, Florence, p. 11-18, https://doi.org/10.18653/v1/W19-3502.

Yu, X., Zhao, A., Blanco, E., Hong, L. (2023), "A Fine-Grained Taxonomy of Replies to Hate Speech", *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, p. 7275-7289, https://doi.org/10.18653/v1/2023.emnlp-main.450.

Yuan, L. Rizoiu, M.-A. (2022), "Detect hate speech in unseen domains using multi-task learning: A case study of political public figures", arXiv:2208.10598, https://doi.org/10.48550/arXiv.2208.10598.

Yuan, L., Wang, T., Ferraro, G., Suominen, H., Rizoiu, M.-A. (2022), "Transfer learning for hate speech detection in social media", *J Comput Soc Sc*, 6, p. 1081-1101, https://doi.org/10.1007/s42001-023-00224-9.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., Kumar, R. (2019), "Predicting the Type and Target of Offensive Posts in Social Media", *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1 (Long and Short Papers), Minneapolis, p. 1415-1420, https://doi.org/10.18653/v1/N19-1144.

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang S, Yin D., Du, M. (2024), "Explainability for large language models: A survey", *ACM Transactions on Intelligent Systems and Technology*, 15/2, p. 1-38, https://doi.org/10.1145/3639372.

Zhou, X., Zhu, H., Yerukola, A., Davidson, T., Hwang, J. D., Swayamdipta, S., Sap, M. (2023), "Cobra frames: Contextual reasoning about effects and harms of offensive statements", *Findings of the Association for Computational Linguistics*, arXiv preprint arXiv:2306.01985, p. 6294-6315, https://doi.org/10.18653/v1/2023.findings-acl.392.

Ziems, C., Vigfusson, Y., Morstatter, F. (2020), "Aggressive, repetitive, intentional, visible, and imbalanced: Refining representations for cyberbullying classification", *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media* (vol. 14), p. 808-819, https://doi.org/10.1609/icwsm.v14i1.7345.