Bridging Policies, Platforms and Research: Advancing NLP for Hate Speech Proactive Mitigation

Anonymous ACL submission

Abstract

Despite regulations imposed by nations and social media platforms (Government of India, 2021; European Parliament and Council of the European Union, 2022), hateful content persists as a significant challenge. Existing approaches primarily rely on reactive measures such as blocking or suspending offensive messages, with emerging strategies focusing on **proactive measurements** like detoxification and counterspeech. In this work, we conduct a comprehensive examination of hate speech regulations and strategies from multiple perspectives: *country regulations*, *social platform policies*, and *NLP research datasets*. Our findings reveal significant inconsistencies in hate speech definitions and moderation practices across jurisdictions and platforms, alongside a lack of alignment with research efforts. Based on these insights, we suggest ideas and research direction for further exploration of a unified framework for automated hate speech moderation incorporating diverse strategies.

1 Introduction

AI continues to advance rapidly across various domains, offering diverse applications. Among these, leveraging AI for societal positive impact (Shi et al., 2020) is becoming an important direction to explore. Specifically, in the field of NLP (Jin et al., 2021), one of the important societal applications lies in mitigating *digital violence* (Kaye, 2019).

Digital violence persists as a pressing issue in online social environments, posing tangible risks to users (Barbieri et al., 2019; Kara et al., 2022). It involves using information and communication technologies to hurt, humiliate, disturb, frighten, exclude, and victimize individuals. This often results in increased anxiety, sadness, tension, and a loss of motivation at work. It includes harmful online activities such as abusive behavior, hate speech, toxic speech and offensive language, significantly affecting an individual's professional and social effectiveness and efficiency (Özsungur, 2022).

Traditional automated moderation methods typically involve measures such as blocking or suspending accounts that disseminate offensive messages (MacAvaney et al., 2019; Cobbe, 2021). Major technology companies, including Meta and X, have implemented these strategies to manage hate speech. However, such measures have proved insufficient in curbing hateful sentiments over the long term (Parker and Ruths, 2023). Alternatives such as counterspeech have gained traction as promising strategies to mitigate hate speech by engaging in dialogue aimed at challenging harmful narratives (Alsagheer et al., 2022; Kulenović, 2023). Furthermore, text detoxification represents an approach intended to reduce the toxicity of communications while maintaining the original message (Nogueira dos Santos et al., 2018; Logacheva et al., 2022). Despite their potential, these approaches have yet to be widely adopted as part of social media platforms' moderation strategies.

In this work, we conduct a comprehensive examination of the measures currently employed to mitigate digital violence, focusing on insights drawn from government regulations, social media platform policies, and NLP research datasets. While our primary objective is to investigate and document these existing frameworks, we also recognize the critical need for empirical evaluation of their practical effectiveness. Our study highlights the current approaches to handle hate speech, emphasizing the disparities and gaps that persist among them. These insights reveal areas ripe for improvement and suggest the need for further empirical research to assess the real-world impact of these measures. Based on our analysis, we propose exploring the potential development of a more unified and cohesive framework in the future to effectively address these gaps. Our contributions in this study are as follows:

(i) We provide a comprehensive survey of hate speech definitions and mitigation strategies from three main perspectives: (a) government regulations across nations; (b) policies of social media platforms; (c) NLP research datasets.
(ii) We conduct an extensive comparative analysis of documents from these domains to identify inconsistencies and opportunities for improvement in current moderation practices.
(iii) Based on our analysis, we suggest exploring a framework for more formalized methods to combat hate speech in the future.

Key contributions of our paper is highlighted in Figure 1.

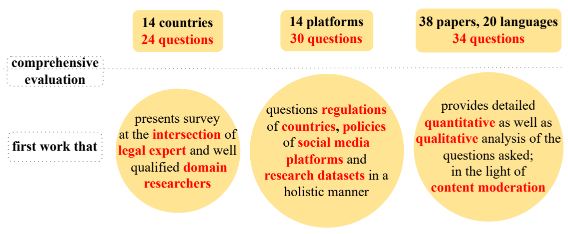

Figure 1: **Key Contributions of this Survey Paper:** Our work is the first of its kind that explores hate speech across country-wise regulations, social media platform policies and dataset research papers.

2 Background

Violence is an umbrella term that refers to words or actions that cause harm to an individual or a community. Digital violence is a special form that anchors digital technologies, with harm typically spread through electronic devices such as computers, smartphones, and IoT sensors. This form of violence can occur publicly on social media platforms or privately on personal devices and in alternative digital environments like the metaverse. Our study focuses on digital violence, more specifically, expressed in a textual form.

Banko et al. (2020) classified harmful content as either **abusive** or **online harm** and offered a corresponding typology. The typology includes four categories: **hate and harassment**, **self-inflicted harm**, **ideological harm**, and **exploitation**. The study by Lewandowska-Tomaszczyk et al. (2023) categorizes harmful content as **offensive speeches**, including 17 sub-categories like **taboo**, **insulting**, **hate speech**, **harassment**, and **toxic**. When it comes to defining **hate speech**, there is no consensus among legislators, platform operators, and researchers (Brown, 2015). One of the most comprehensive definitions widely followed in the computer science literature, as proposed by the United Nations,[1] describes hate speech as any kind of communication in speech, writing, or behavior that attacks or uses pejorative or discriminatory language with reference to a person or a group based on who they are. To address textual digital violence, traditional automated moderation practice often involves **content moderation** measures. Content moderation, both human and algorithmic, involves overseeing user-generated content to align with legal standards, community norms, and platform policies (Banko et al., 2020; Hietanen and Eddebo, 2023). Algorithmic moderation, primarily aimed at **removing** or **banning** non-compliant content, boosts online safety, curbs abuse, and swiftly detects serious infractions, thus reducing the limitations of depending entirely on human moderators.

Recently, **counterspeech** has gained attention as a strategy to not only mitigate hate speech but also engage users in constructive dialogue that addresses misunderstandings and challenges harmful narratives (Kulenović, 2023). It is important to note that counterspeech is not solely about generating counterarguments automatically; it encompasses a broader strategy for fostering dialogue and potentially influencing the speaker's mindset. Similarly, **detoxification** is concerned with altering the style of text to make communications less toxic while retaining their original intent (Nogueira dos Santos et al., 2018; Logacheva et al., 2022). Detoxification should be viewed as a suggestive tool that recommends less toxic wording, leaving it to the individual user or moderation framework to adopt these changes. Contrary to concerns about infringing on freedom of speech, both counterspeech and detoxification contribute to more civil discourse by offering voluntary and non-coercive means of improving online interactions.

Both approaches serve as valuable alternatives to traditional moderation methods by promoting positive interaction and personal agency in the moderation process.

3 Related works

Automatic platform content policy analysis: Social media platforms take different approaches to content moderation. Comparison of these different approaches can be a time-consuming task because of the variety of formulations and approaches.

[1]un.org/en/what-is-hate-speech

Most often, only single platforms were conducted and analyzed by researchers (Chandrasekharan et al., 2018; Fiesler et al., 2018). The work by Schaffner et al. (2024) proposed an approach for automated collection and the creation of a unified schema to compare platforms. This identified significant structural differences between the platforms in how they deal with these requirements.

Automatic hate speech detection: Moderation is a fundamental element of social media platforms, involving various measures to limit the visibility of hateful content. These measures range from deleting and hiding posts to issuing warnings or blocking users who fail to adhere to regulations (Trujillo et al., 2023). In line with these moderation efforts, researchers have also focused on improving automatic detection systems. Significant research efforts have been directed toward gathering datasets that enable the development of automatic hate speech classification models (Fortuna et al., 2020; Mathew et al., 2021). These datasets support the creation of models capable of detecting hate speech across various contexts, including those in low-resource languages such as Amharic (Ayele et al., 2024), Arabic (Magnossão de Paula et al., 2022; Alzubi et al., 2022), code-mixed Hindi (Bohra et al., 2018; Ousidhoum et al., 2019), etc.

Automatic counterspeech generation: While access restrictions remain a common strategy supported by platforms and government policies to combat harmful content, countering hate speech through engagement is gaining recognition (Mun et al., 2024). This approach, often encapsulated by the phrase *countering rather than censoring*, is seen as preferable to outright censorship, as it tends to respect the principle of free speech (Yu et al., 2023; Bonaldi et al., 2024). Yu et al. (2023) investigate counterspeech with a focus on addressing the author and the hate content directly, where the former is viewed as less robust. Beyond reducing hate, counterspeech efforts are utilized to foster positive transformations within online communities by promoting discussions and cultivating a sense of community (Buerger, 2022, 2021).

Automatic text detoxification: Another promising avenue in combating toxicity involves text detoxification, which targets eliminating offensive content in messages while preserving the intended meaning (Logacheva et al., 2022; Dementieva et al., 2021; Tran et al., 2020). Detoxification enhances the quality of online interactions by facilitating more respectful and less toxic communications (Tran et al., 2020). Various models applied to detoxification aim to generate acceptable and diverse non-toxic outputs.

Examples of automatic mitigation strategies in deployment: Chung et al. (2021) developed a tool for Twitter (now X) designed to continuously monitor and respond to hateful content related to Islamophobia. The tool was used by non-governmental organization (NGO) operators, and the counter-narrative feature has been highly praised for its potential to significantly impact the fight against online Islamophobia.

Arora et al. (2024) in their study examined research on hate speech and related platform moderation policies. The findings reveal a notable discrepancy between the focus of research and the needs of platform policies. This mismatch underscores a gap between the types of content platforms that need to be moderated and the solutions offered by current research on harmful content detection.

In our work, our objective is to identify the gaps between regulatory policies from countries, policies from social media platforms, and approaches used in NLP research. We suggest the potential for a more proactive moderation approach to address these challenges.

4 Methodology

Our approach incorporates a strategic analysis of hate speech regulation & mitigation through three primary perspectives: **country-specific regulations**, **social media platforms' policies** and **NLP research approaches**.

For each of the three perspectives, we developed specific SELECTION CRITERIA to obtain representative samples and crafted a series of QUESTIONS to analyze and gain deeper insights into each area. This dual strategy ensures a comprehensive examination of the regulatory landscape and the effectiveness of various moderation techniques. Furthermore, our analysis also aims to examine three common approaches of content moderation: **blocking/suspending offensive content**, **detoxification of toxic language**, and **counter of hate speech** to engage users constructively. Hence, these moderation techniques were also considered during curation.

QUESTIONS: For each of the three dimensions, we first tunneled down **meta categories** and then brewed relevant questions for each meta category (refer Figure 2). These two steps were specifically

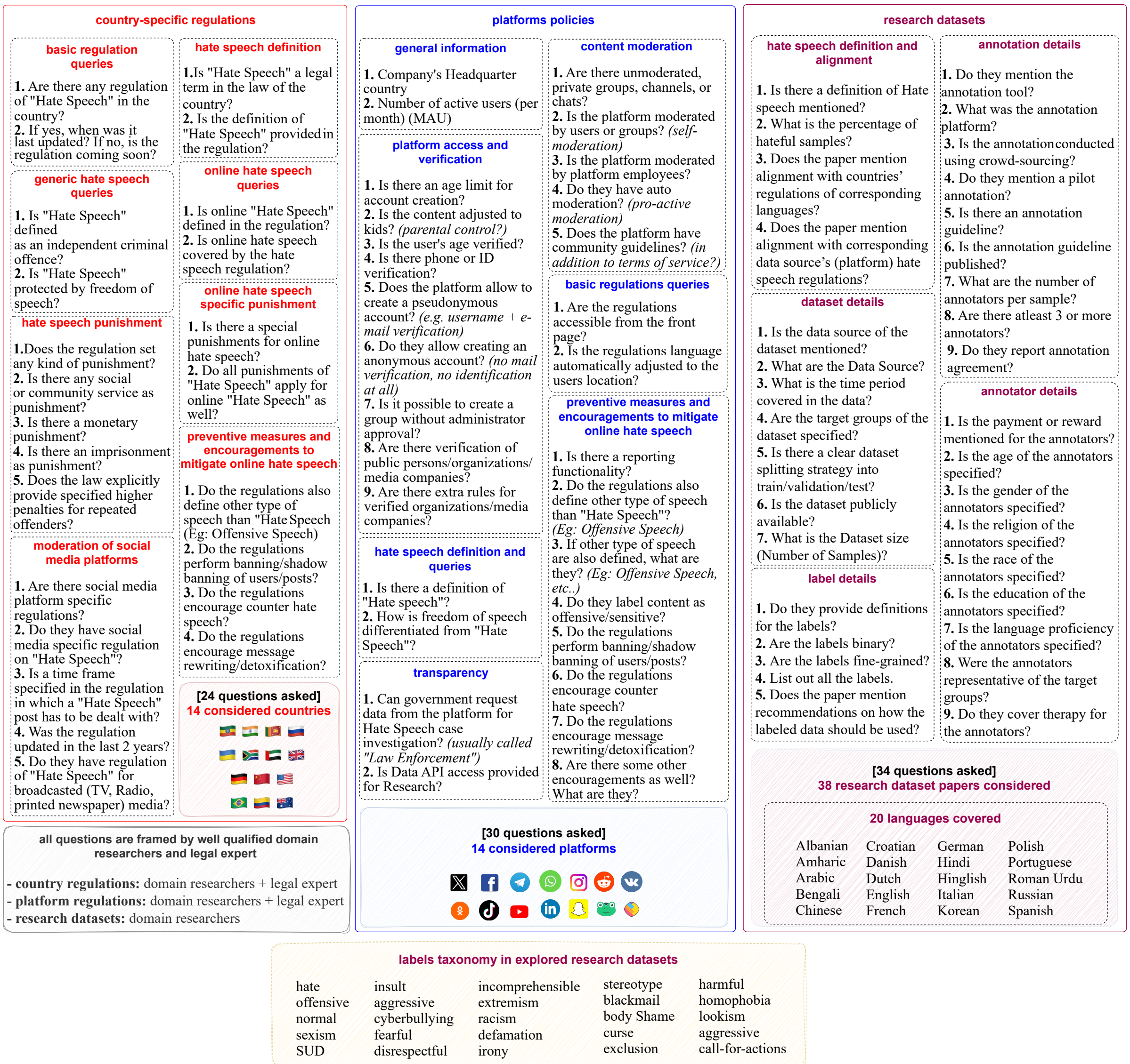

Figure 2: **Selection and Questionnaire:** Selected countries, platforms, meta categories, full list of questionnaire and other statistics spanning *country-specific regulations*, *social media platform policies* and *research datasets*.

performed to audit different perspectives in a relevant and robust manner. Note that the surveys used in this study were carefully designed and answered by a group of qualified researchers, including *PhD students* and *postdoctoral fellows*, who have expertise in social media policies and online hate speech regulation. Information regarding social media platform policies were gathered through a thorough examination of policy documents and guidelines available on the platforms' official websites.

Validity assurance: To ensure the validity and comprehensiveness of our surveys on social media policies and country regulations, we collaborated with LegalTech researchers from a university law school. This collaboration helped us refine our survey questions and ensure that our research methodologies align with the latest legal and regulatory standards.

4.1 Country-specific regulations

In this subsection, we examine the regulations concerning hate speech that have been established by individual countries. Hate speech can manifest itself in various forms and requires different regulatory approaches depending on cultural, legal, and societal contexts and we maximally incorporate these as discussed below.

Selection criteria: To ensure a diverse and representative sample of countries, we selected them based on extensive **familiarity and exper-**

tise of the research team to ensure a detailed and contextually rich analysis. Then we also selected diverse **geographic representation** by selecting atleast one country from each continent based on population to capture a wide range of regulatory approaches. Finally, countries with significant **online presence & engagement** and where incidents of hate speech are prevalent were also considered to strengthen our **focus on hate speech regulation**.

QUESTIONS: First we narrowed down the meta categories to have a solid overview of the regulations. For this purpose, we aimed at following rationales for extracting key insights from each country's approach to hate speech regulation. First, we considered **freedom of speech** and **hate speech definition** as they are very crucial for gaining insights into country's tolerance of expression and for reflecting upon their conceptualization and legal stance on hate. Then we considered different **punishments** like monetary fine or imprisonment which is of immense importance; since it is related to the consequences of violation of hate speech regulations. We also employed **preventive measures** to emphasize censorship or content moderation like counterspeech regulatory support and message detoxification. Finally, we also consider **social media regulations** as it is vital for deep diving into regulations related to online hate. After finalizing meta categories on these key insights, we then added relevant questions into each of them.

In total, we selected 14 countries from around the world,[2] to provide a comprehensive representation of how hate speech and related issues are regulated at the national level. This selection ensures at least one country from each continent is included to capture a diverse set of regulatory approaches and perspectives. Please refer to Figure 2 for a holistic view. The insights derived from the analysis of these regulations are discussed in Section 5.1.

4.2 Platform policies

We analyze the policies developed by social media platforms to regulate hate speech with the goal to understand how these platforms define, detect, and respond to such content. Our analysis provides insights into the accessibility and transparency of platform policies, the use of automated and human moderation, and the preventive measures in place to protect users.

SELECTION CRITERIA: Our selection criteria were designed to ensure a thorough examination of policies across globally popular social media platforms while also accounting for regional variations. **Globally popular platforms** were selected based on their monthly active user count, prioritizing the most widely used platforms worldwide to ensure broad coverage and relevance. For **regionally relevant platforms**, importance was given to the popularity of platforms within the countries mentioned in Section 4.1.

QUESTIONS: We first curated meta categories before concluding the final questionnaire. Social media platform specific rationales targeted at distinct aspect of platform functionalities and their strategies for addressing hate speech were considered. **Hate speech definition** is among the first major rationale we considered for identifying the platform's foundation on content moderation and enforcement actions. Then we pillared on **platform access & verification**, **regulation accessibility** and **content moderation** as the most crucial rationales. These rationales were chosen to understand the mechanisms for user access and verification, including age restrictions and verification processes, inquiry into the accessibility and language of platform regulations aimed to assess the transparency and helped us further delve into the mechanisms and actors involved in content moderation, including user-driven moderation, automated systems, and employee-led moderation teams. In addition, examination of policy alignment with country-specific regulations provided insights into platform compliance and adaptability to legal frameworks. Similar to rationales in country-specific regulations, here also we include **preventive measures** as they focused on the platform's efforts to empower users in reporting hate speech, as well as initiatives aimed at promoting counterspeech and detoxification of harmful content. Additionally, we also include **data access** as an inquiry as well to assess the platform's transparency and willingness to collaborate with researchers and law enforcement agencies in hate speech investigations. Access to platform data is critical for conducting comprehensive research and ensuring accountability.

In total, 14 social media platforms were selected based on the established selection criteria and analyzed through our detailed questionnaire.[3] Refer

[2]Countries: Ethiopia, India, Sri Lanka, Russia, Ukraine, South Africa, United States, United Arab Emirates, United Kingdom, Germany, China, Brazil, Colombia and Australia.

[3]X, Facebook, Telegram, WhatsApp, Instagram, Reddit,

to Figure 2 for detailed questionnaire with meta categories of social media platforms' regulations. The findings, which elucidate the platforms' approaches to hate speech, are presented in Section 5.2.

4.3 Research datasets

In our third pillar, we bridge the gap with NLP research by examining the current state of automatic **hate speech detection in texts**. Our focus centers on datasets designed for fine-tuning machine learning models, allowing us to gain a comprehensive understanding of the landscape across diverse languages. This exploration will highlight the methodologies used in dataset creation, their definitions of hate speech, and their relevance in addressing the challenges posed by hateful content in digital environments.

SELECTION CRITERIA: Our selection criteria were crafted to ensure the inclusion of diverse perspectives while maintaining a high standard of relevance and credibility. These criteria included the following points. **Language inclusivity** was chosen as one of the most crucial criterion as it encompasses a wide array of languages prevalent in the countries considered in Section 4.1. **Citations** and **publication venue** are one of the most important parameters of success of a research work. We therefore prioritized dataset papers that have significantly influenced the academic community, as indicated by their citation metrics. For low-resource languages, we included the majority or all of the available datasets to ensure comprehensive representation in our analysis. Preference was given to papers published in esteemed NLP venues such as *ACL Anthology*, *AAAI*, *LREC*, *COLING*, or *WOAH*, ensuring a standard of quality and rigor in the selected dataset papers. Finally, we further **cross-verified** to bolster the credibility of our selection and cross-checked our choices with established repositories such as `hatespeechdatasets.com`, thus validating the inclusion of well-established datasets.

QUESTIONS: For the formulation of meta categories we designed to extract key insights essential for a comprehensive understanding of hate speech datasets. **Hate speech definition** is the crucial rationale here as well as it provides with the complex nature of hate speech, and helps in exploring how researchers conceptualize and define it. Next, we anchor on **annotation process** and diverse set of **labels** as they help in investigating that how the annotation process sheds light on the methodologies employed, including the existence of guidelines, pilot annotations, and quality control measures, which are crucial for evaluating the quality and reliability of the dataset. The labels used for annotation and their descriptions provide insights into the granularity and depth of the dataset's understanding of hate speech nuances. **Annotator demographics** are also very crucial as they help in exploring the demographics of annotators, encompassing factors such as age, gender, religion, and race, facilitated an assessment of dataset inclusivity and annotator suitability. Finally, **dataset material** which queries aspects such as data source, modality, size, and availability is vital for understanding the dataset's scope and applicability in hate speech research.

We selected 38 dataset papers spanning 20 languages based on our criteria and analyzed them using our comprehensive questionnaire. The complete questionnaire is available in Figure 2, cited datasets are present in Appendix 1 and the results from this analysis are presented in Section 5.3.

5 Results and analysis

In this section, we will discuss the outcomes of our investigation across three key areas aimed at mitigating hate speech: *country regulations*, *platform policies* and *research datasets*. We have summarized our analysis quantitatively in Figure 3 and have also uploaded full list of questionnaire as a supplementary material.

5.1 Regulation results

As stated earlier, we selected 14 countries from all over the world in order to have a comprehensive picture of how hate speech and related issues are regulated on a governmental level. The quantitative results of our investigation are summarized in ***Figure 3(a)*** and below we perform qualitative analysis.

First of all, we note that all the countries considered regulate hate speech except the USA and the majority of the regulations have been updated no earlier than ***four years*** ago, keeping the nations up-to-date with the current hate speech challenges. The ***definition of hate speech***, inspite of the widespread recognition of the need to address hate speech at the governmental level, lacks single universally accepted definition of what constitutes

VK, Odnoklassniki, TikTok, YouTube, LinkedIn, Snapchat, GAB, ShareChat

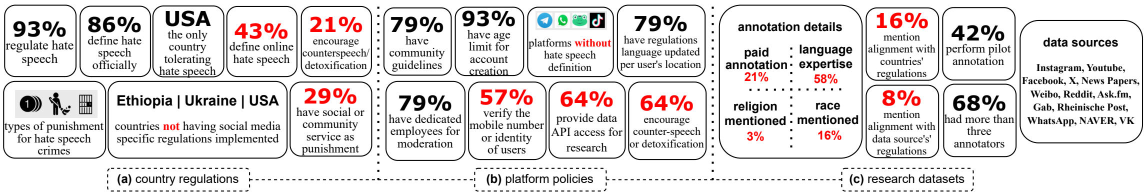

Figure 3: **EXPLORATORY RESULTS** on regulations by nations and social media platforms.

hate speech. Different countries have developed their own definitions, reflecting their unique cultural, legal, and social contexts. Understanding these context-specific definitions is crucial for developing targeted interventions that respect local norms while safeguarding individuals from harmful speech.

Although most countries have laws regulating hate speech, only **43%** have specific definitions related to ***online hate speech***. Countries such as the USA, Russia, and Ukraine do not independently address online hate speech at the legislative level, whereas hate speech is protected under freedom of speech in the USA.

Coming to the ***punishments***, most countries adopt various approaches to punish hate speech offenders, with penalties ranging from fines and community service to imprisonment. While imprisonment is a potential consequence, the duration of sentences is typically relatively short, and varies from one country to the other. Finally, ***proactively mitigation of hate speech*** is being used in a limited manner. At both national and regional levels, specific laws addressing counterspeech and detoxification are lacking. However, many countries have emphasized the creation of a safe environment through proactive methods, which appears to be a positive initial step in this direction.

5.2 Platform results

In this subsection we analyze the outcome of our survey on social media platform's policies to robustly corroborate the community guidelines provided by the respective platforms in terms of hateful content and their mitigation strategies. The overall quantitative results from our investigation are summarized in ***Figure 3(b)*** and below we perform qualitative analysis.

The majority of platforms have an age limit for account creation and some sort of parental control. Only *three* out of 14 platforms we studied—Facebook, Instagram, and YouTube—apply age verification methods. Phone number or any other sort of ID verification is present in only **57%** of the platforms that we studied. None of the platforms allow for the creation of completely ***anonymous accounts***, but nine platforms allow for the creation of ***pseudonymous accounts***, i.e., an account that uses a fictitious name or alias to protect the user's digital identity.

All platforms except *GAB* have made their regulations or ***community guidelines*** accessible from their home pages. *X*, *Telegram* and *GAB* ***do not adjust the language*** of the regulations automatically according to the user's geographical location. Platforms like *Telegram*, *WhatsApp*, *TikTok*, and *GAB* do not even have a strict definition of hate speech in their regulations.

Platforms play an important role in ***content moderation***, where administrators or moderators can moderate respective groups or communities. It is highly subjective and dependent on the social and cultural context of the individual and their demographics. Only a small minority of platforms—Facebook, Instagram, TikTok, ShareChat and YouTube —have moderators with ***demographic diversity***. A common solution to this challenge is employing ***auto-moderation***, which is adopted by almost all platforms except *Telegram*, *WhatsApp*, and *GAB*.

As the primary ***preventive measure***, all platforms have a reporting functionality where users can report content they find inappropriate. The users generally flag the reported content according to the category labels provided by the platform. Platforms like *WhatsApp*, *VK*, *Odnoklassniki*, *TikTok*, and *ShareChat* do not provide a label for offensive or sensitive content when reporting.

At last, we analyze the acceptance of counterspeech and message detoxification as a proactive moderation strategy. Surprisingly, we found very few platforms like *Facebook*, *VK*, and *Odnoklassniki* that encourage the promotion of these new moderation paradigms.

5.3 Results based on research datasets

Our analysis of various hate speech dataset papers has yielded several key findings that provide insights into the landscape of hate speech research and dataset construction. Quantitative results are provided in ***Figure 3(c)*** and below we share qualitative analysis.

Interestingly, **66%** of the surveyed papers present a clear ***definition of hate speech*** within their work. We believe, especially for annotation tasks and dataset papers, conceptual clarity in understanding hate speech is highly important. Consequently, our expectation was that almost all papers would have a definition of hate speech, which is unfortunately not true. Further, our analysis reveals that only **16%** of the papers have ***cross-checked*** their definition with *hate speech regulations at the national level*, and only *three papers* referenced *platform-specific regulations*. This lack of alignment with regulatory frameworks highlights potential discrepancies between academic definitions and legal or platform-specific interpretations of hate speech. To our surprise, only one of the 38 surveyed papers formulate recommendations on ***leveraging their work***, datasets, or annotations. This highlights a missed opportunity for academic research to inform practical interventions and policy-making efforts in the fight against hate speech.

Finally, we observe considerable imbalance in investigated ***data sources***. **X** account for over **50%** of the studies, while other platforms such as *YouTube, Instagram, Reddit* and *WhatsApp* were explored in less than **10%** of the papers. Facebook, with its 3 billion users, far exceeds X, which has only 611 million users, indicating that the over-representation of certain platforms does not correlate with actual usage.

Due to paucity of space, we present few more results which we obtained as observation from our analysis in Appendix A.

6 Conclusion and Future Directions

Our three-tiered study highlighted key challenges in addressing hate speech from governmental, platform, and research angles.

Firstly, the lack of a universally accepted definition of hate speech complicated the development of consistent regulations across countries. While most nations have a definition of hate speech, only a third defined it specifically for the online environment. This underscored the need to raise awareness about online hate and its mitigation. However, some of the countries are interested in proactive hate speech moderation methods development.

Secondly, social media platforms showed policy inconsistencies, which hindered effective content moderation. A fifth of the platforms failed to adapt hate definitions to local languages and cultures, and moderation typically focused on banning rather than proactive strategies.

Thirdly, most NLP research did not align with platform or regulatory guidelines, often reusing outdated definitions from previous computer science studies. Moreover, many studies did not explore proactive measures such as counterspeech or detoxification in operational settings. Such data labeling could potentially improve automatic online hate mitigation.

Ultimately, collaboration between platforms, governments, and researchers is essential to create dynamic moderation frameworks. Aligning definitions and promoting proactive strategies will lead to more effective solutions for combating online hate. The further exploration of proactive moderation pipeline which consists of thoughtful combination of text detoxification, counter speech generation, other preventing measures, and preparing such datasets for automatic methods development should be a frontier for future research.

Limitations

While we made diligent efforts to meticulously document our research process, findings and recommendations, it is important to acknowledge that our study has certain limitations:

1) Only text-based content: We only took into consideration textual expression of digital violence in NLP research. We acknowledge that hate can also be extremely taxing in other modalities like images, voice recordings and videos. Our study on hate mitigation do not encompass such cases.

2) Only human-written content: Our mitigation pipeline was initially tailored to address only human-authored messages and comments. However, as text generation systems become more prevalent, there is a growing influx of machine-generated content on social media platforms. It is imperative to incorporate additional measures to detect and address bots and other machine-generated texts that may pose greater risks in inciting hatred.

3) Only digital content: Finally, we performed our studies only in the realm of digital violence. Nevertheless, digital hater can transcend virtual platforms and manifest in real-world scenarios through various means. For this reason, we include an 'authorities' intervention' step in our demarcation pipeline.

Ethics statement

We are committed to upholding freedom of speech and respect the autonomy of stakeholders in deploying moderation technologies tailored to their specific domain, context, and requirements. Our aim is to offer a broader perspective on potential automatic proactive moderation strategies, providing novel insights and recommendations.

References

Zeleke Abebaw, Andreas Rauber, and Solomon Atnafu. 2022. Multi-channel convolutional neural network for hate speech detection in social media. In *Advances of Science and Technology*, pages 603–618, Cham. Springer International Publishing.

Nuha Albadi, Maram Kurdi, and Shivakant Mishra. 2018. Are they our brothers? analysis and detection of religious hate speech in the arabic twittersphere. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 69–76.

Dana Alsagheer, Hadi Mansourifar, and Weidong Shi. 2022. Counter hate speech in social media: A survey. *CoRR*, abs/2203.03584.

Salaheddin Alzubi, Thiago Castro Ferreira, Lucas Pavanelli, and Mohamed Al-Badrashiny. 2022. aiXplain at Arabic hate speech 2022: An ensemble based approach to detecting offensive tweets. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 214–217, Marseille, France. European Language Resources Association.

Arnav Arora, Preslav Nakov, Momchil Hardalov, Sheikh Muhammad Sarwar, Vibha Nayak, Yoan Dinkov, Dimitrina Zlatkova, Kyle Dent, Ameya Bhatawdekar, Guillaume Bouchard, and Isabelle Augenstein. 2024. Detecting harmful content on online platforms: What platforms need vs. where research efforts go. *ACM Comput. Surv.*, 56(3):72:1–72:17.

Dennis Assenmacher, Marco Niemann, Kilian Müller, Moritz Seiler, Dennis Riehle, Heike Trautmann, and Heike Trautmann. 2021. Rp-mod & rp-crowd: Moderator- and crowd-annotated german news comment datasets. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Abinew Ali Ayele, Skadi Dinter, Tadesse Destaw Belay, Tesfa Tegegne Asfaw, Seid Muhie Yimam, and Chris Biemann. 2022. The 5js in ethiopia: Amharic hate speech data annotation using toloka crowdsourcing platform. In *2022 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, pages 114–120.

Abinew Ali Ayele, Esubalew Alemneh Jalew, Adem Chanie Ali, , Seid Muhie Yimam, and Chris Biemann. 2024. Exploring boundaries and intensities in offensive and hate speech: Unveiling the complex spectrum of social media discourse. In *Proceedings of The Fourth Workshop on Threat, Aggression & Cyberbullying*.

Michele Banko, Brendon MacKeen, and Laurie Ray. 2020. A unified taxonomy of harmful content. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 125–137, Online. Association for Computational Linguistics.

Davide Barbieri, Charlotte Dahin, Brianna Guidorzi, Zuzana Madarova Marre Karu, Blandine Mollard, Jolanta Reingardė, Lina Salanauskaitė, onika Natter, Renate Haupfleisch, Katja Korolkova, Monica Barbovschi, Liza Tsaliki, Brian O'Neill, Clara Faulí, Federica Porcu, Francisco Lupiáñez Villanueva, and Alexandra Theben. 2019. Gender equality and youth: opportunities and risks of digitalisation – main report. Technical report, The European Institute for Gender Equality.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Mohit Bhardwaj, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2020. Hostility detection dataset in hindi. *Preprint*, arXiv:2011.03588.

Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. A dataset of Hindi-English code-mixed social media text for hate speech detection. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 36–41, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Helena Bonaldi, Yi-Ling Chung, Gavin Abercrombie, and Marco Guerini. 2024. NLP for Counterspeech against Hate: A Survey and How-To Guide. *arXiv preprint arXiv:2403.20103*.

Alex Brown. 2015. *Hate speech law: A philosophical examination*. Taylor & Francis.

Catherine Buerger. 2021. #iamhere: Collective Counterspeech and the Quest to Improve Online Discourse. *Social Media + Society*, 7(4).

Catherine Buerger. 2022. Why They Do It: Counter-speech Theories of Change. *SSRN Electronic Journal*.

Ben Burtenshaw and Mike Kestemont. 2021. A Dutch dataset for cross-lingual multilabel toxicity detection. In *Proceedings of the 14th Workshop on Building and Using Comparable Corpora (BUCC 2021)*, pages 75–79, Online (Virtual Mode). INCOMA Ltd.

Miguel Angel Carmona, Estefanía Guzmán-Falcón, Manuel Montes, Hugo Jair Escalante, Luis Villaseñor-Pineda, Veronica Reyes-Meza, and Antonio Rico-Sulayes. 2018. Overview of mex-a3t at ibereval 2018: Authorship and aggressiveness analysis in mexican spanish tweets.

Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The internet's hidden rules: An empirical study of reddit norm violations at micro, meso, and macro scales. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW).

Patricia Chiril, Farah Benamara Zitoune, Véronique Moriceau, Marlène Coulomb-Gully, and Abhishek Kumar. 2019. Multilingual and multitarget hate speech detection in tweets. In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019. Volume II : Articles courts*, pages 351–360, Toulouse, France. ATALA.

Yi-Ling Chung, Serra Sinem Tekiroglu, Sara Tonelli, and Marco Guerini. 2021. Empowering ngos in countering online hate messages. *Online Soc. Networks Media*, 24:100150.

Jennifer Cobbe. 2021. Algorithmic censorship by social platforms: Power and resistance. *Philosophy & Technology*, 34(4):739–766.

Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2019. Cross-Platform Evaluation for Italian Hate Speech Detection. In *CLiC-it 2019 - 6th Annual Conference of the Italian Association for Computational Linguistics*, Bari, Italy.

Mithun Das, Somnath Banerjee, Punyajoy Saha, and Animesh Mukherjee. 2022. Hate speech and offensive language detection in Bengali. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 286–296, Online only. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):512–515.

Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook.

Daryna Dementieva, Daniil Moskovskiy, Varvara Logacheva, David Dale, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. 2021. Methods for Detoxification of Texts for the Russian Language. *Multimodal Technol. Interact.*, 5(9):54.

Christoph Demus, Jonas Pitz, Mina Schütz, Nadine Probol, Melanie Siegel, and Dirk Labudde. 2022. Detox: A comprehensive dataset for German offensive language and conversation analysis. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 143–153, Seattle, Washington (Hybrid). Association for Computational Linguistics.

European Parliament and Council of the European Union. 2022. Regulation (eu) 2022/2065 of the european parliament and of the council of 19 october 2022 on a single market for digital services (digital services act) and amending directive 2000/31/ec. Official Journal of the European Union, L 277, 27.10.2022, p. 1–102.

Casey Fiesler, Jialun Jiang, Joshua McCann, Kyle Frye, and Jed Brubaker. 2018. Reddit rules! characterizing an ecosystem of governance. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Paula Fortuna, João Rocha da Silva, Juan Soler-Company, Leo Wanner, and Sérgio Nunes. 2019. A hierarchically-labeled Portuguese hate speech dataset. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 94–104, Florence, Italy. Association for Computational Linguistics.

Paula Fortuna, Juan Soler, and Leo Wanner. 2020. Toxic, Hateful, Offensive or Abusive? What Are We Really Classifying? An Empirical Analysis of Hate Speech Datasets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6786–6794, Marseille, France. European Language Resources Association.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

Government of India. 2021. Information technology (intermediary guidelines and digital media ethics code) rules, 2021. Ministry of Electronics and Information Technology, Government of India.

Hatem Haddad, Hala Mulki, and Asma Oueslati. 2019. T-hsab: A tunisian hate speech and abusive dataset. In *Arabic Language Processing: From Theory to Practice*, pages 251–263, Cham. Springer International Publishing.

Mika Hietanen and Johan Eddebo. 2023. Towards a definition of hate speech—with a focus on online contexts. *Journal of communication Inquiry*, 47(4):440–458.

Younghoon Jeong, Juhyun Oh, Jaimeen Ahn, Jongwon Lee, Jihyung Moon, Sungjoon Park, and Alice Oh. 2022. Kold: Korean offensive language dataset. *Preprint*, arXiv:2205.11315.

Aiqi Jiang, Xiaohan Yang, Yang Liu, and Arkaitz Zubiaga. 2022. Swsr: A chinese dataset and lexicon for online sexism detection. *Online Social Networks and Media*, 27:100182.

Zhijing Jin, Geeticka Chauhan, Brian Tse, Mrinmaya Sachan, and Rada Mihalcea. 2021. How good is NLP? a sober look at NLP tasks through the lens of social impact. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3099–3113, Online. Association for Computational Linguistics.

Ergün Kara, Gülşen Kirpik, and Attila Kaya. 2022. A research on digital violence in social media. In *Handbook of research on digital violence and discrimination studies*, pages 270–290. IGI Global.

Md. Rezaul Karim, Sumon Kanti Dey, Tanhim Islam, Sagor Sarker, Mehadi Hasan Menon, Kabir Hossain, Md. Azam Hossain, and Stefan Decker. 2021. Deephateexplainer: Explainable hate speech detection in under-resourced bengali language. In *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10.

David Kaye. 2019. *Speech Police: The Global Struggle to Govern the Internet*. Columbia Global Reports.

Enes Kulenović. 2023. Should democracies ban hate speech? hate speech laws and counterspeech. *Ethical Theory and Moral Practice*, 26(4):511–532.

Barbara Lewandowska-Tomaszczyk, Slavko Žitnik, Chaya Liebeskind, Giedre Valunaite Oleskevicienė, Anna Bączkowska, Paul A Wilson, Marcin Trojszczak, Ivana Brač, Lobel Filipić, Ana Ostroški Anić, et al. 2023. Annotation scheme and evaluation: The case of offensive language. *Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje*, 49(1).

Nikola Ljubešić, Darja Fišer, Tomaž Erjavec, and Ajda Šulc. 2021. Offensive language dataset of croatian, english and slovenian comments FRENK 1.1. Slovenian language resource repository CLARIN.SI.

Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 2022. ParaDetox: Detoxification with parallel data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6804–6818, Dublin, Ireland. Association for Computational Linguistics.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152.

Angel Felipe Magnossão de Paula, Paolo Rosso, Imene Bensalem, and Wajdi Zaghouani. 2022. UPV at the Arabic hate speech 2022 shared task: Offensive language and hate speech detection using transformers and ensemble models. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 181–185, Marseille, France. European Language Resources Association.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '19, page 14–17, New York, NY, USA. Association for Computing Machinery.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875.

Puneet Mathur, Rajiv Shah, Ramit Sawhney, and Debanjan Mahata. 2018. Detecting offensive tweets in Hindi-English code-switched language. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 18–26, Melbourne, Australia. Association for Computational Linguistics.

Zewdie Mossie and Jenq-Haur Wang. 2020. Vulnerable community identification using hate speech detection on social media. *Information Processing & Management*, 57(3):102087.

Hala Mulki, Hatem Haddad, Chedi Bechikh Ali, and Halima Alshabani. 2019. L-HSAB: A Levantine Twitter dataset for hate speech and abusive language. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 111–118, Florence, Italy. Association for Computational Linguistics.

Jimin Mun, Cathy Buerger, Jenny T. Liang, Joshua Garland, and Maarten Sap. 2024. Counterspeakers' perspectives: Unveiling barriers and AI needs in the fight against online hate. *CoRR*, abs/2403.00179.

Cicero Nogueira dos Santos, Igor Melnyk, and Inkit Padhi. 2018. Fighting offensive language on social media with unsupervised text style transfer. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 189–194, Melbourne, Australia. Association for Computational Linguistics.

Erida Nurce, Jorgel Keci, and Leon Derczynski. 2022. Detecting abusive albanian. *Preprint*, arXiv:2107.13592.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. 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)*, pages 4675–4684, Hong Kong, China. Association for Computational Linguistics.

Sara Parker and Derek Ruths. 2023. Is hate speech detection the solution the world wants? *Proceedings of the National Academy of Sciences*, 120(10):e2209384120.

Michal Ptaszynski, Agata Pieciukiewicz, and Paweł Dybała. 2019. *Results of the PolEval 2019 Shared Task 6: first dataset and Open Shared Task for automatic cyberbullying detection in Polish Twitter*, page 89–110.

Juan Manuel Pérez, Franco M. Luque, Demian Zayat, Martín Kondratzky, Agustín Moro, Pablo Santiago Serrati, Joaquín Zajac, Paula Miguel, Natalia Debandi, Agustín Gravano, and Viviana Cotik. 2023. Assessing the impact of contextual information in hate speech detection. *IEEE Access*, 11:30575–30590.

Hammad Rizwan, Muhammad Haroon Shakeel, and Asim Karim. 2020. Hate-speech and offensive language detection in Roman Urdu. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2512–2522, Online. Association for Computational Linguistics.

Nauros Romim, Mosahed Ahmed, Hriteshwar Talukder, and Md. Saiful Islam. 2021. Hate speech detection in the bengali language: A dataset and its baseline evaluation. In *Proceedings of International Joint Conference on Advances in Computational Intelligence*, pages 457–468, Singapore. Springer Singapore.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An Italian Twitter corpus of hate speech against immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Brennan Schaffner, Arjun Nitin Bhagoji, Siyuan Cheng, Jacqueline Mei, Jay L Shen, Grace Wang, Marshini Chetty, Nick Feamster, Genevieve Lakier, and Chenhao Tan. 2024. "community guidelines make this the best party on the internet": An in-depth study of online platforms' content moderation policies. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Ravi Shekhar, Mladen Karan, and Matthew Purver. 2022. CoRAL: a context-aware Croatian abusive language dataset. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 217–225, Online only. Association for Computational Linguistics.

Zheyuan Ryan Shi, Claire Wang, and Fei Fang. 2020. Artificial intelligence for social good: A survey. *CoRR*, abs/2001.01818.

Gudbjartur Ingi Sigurbergsson and Leon Derczynski. 2020. Offensive language and hate speech detection for Danish. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3498–3508, Marseille, France. European Language Resources Association.

Rachele Sprugnoli, Stefano Menini, Sara Tonelli, Filippo Oncini, and Enrico Piras. 2018. Creating a WhatsApp dataset to study pre-teen cyberbullying. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 51–59, Brussels, Belgium. Association for Computational Linguistics.

Minh Tran, Yipeng Zhang, and Mohammad Soleymani. 2020. Towards a friendly online community: An unsupervised style transfer framework for profanity redaction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2107–2114, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Amaury Trujillo, Tiziano Fagni, and Stefano Cresci. 2023. The DSA transparency database: Auditing self-reported moderation actions by social media. *CoRR*, abs/2312.10269.

Xinchen Yu, Ashley Zhao, Eduardo Blanco, and Lingzi Hong. 2023. A fine-grained taxonomy of replies to hate speech. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 7275–7289. Association for Computational Linguistics.

Nadezhda Zueva, Madina Kabirova, and Pavel Kalaidin. 2020. Reducing unintended identity bias in Russian hate speech detection. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 65–69, Online. Association for Computational Linguistics.

Fahri Özsungur. 2022. *Handbook of Research on Digital Violence and Discrimination Studies: A volume in the Advances in Human and Social Aspects of Technology (AHSAT) Book Series.*

A Analysis continued

Further investigation into hate speech dataset papers revealed a nuanced understanding of hate speech as a multi-faceted phenomenon. Through the analysis of hate speech definitions and descriptions, several key aspects emerged that can be considered for the classification of hate speech. We outline these aspects below:

(i) Target: Understanding the target of hate speech is essential in contextualizing its impact. Inflammatory messages directed at individuals or groups are often considered hate speech, while undirected messages are not.

(ii) Discrimination: Hate speech often manifests through discriminatory language targeting various characteristics such as race, sex, gender, nationality, religion, and more.

(iii) Intent of the perpetrator: Malicious intent, ranging from mocking and causing emotional harm to issuing threats or inciting violence, is typical for hate speech. However, humorous, sarcastic, or troll messages are often not considered hate speech.

(iv) Language usage: Hate speech can manifest in diverse linguistic forms, from threatening, dehumanizing, or fear-inducing speech to overtly violent or obscene language. Again, sarcastic or humorous language is often not considered hate speech.

(v) Emotions of the victim/target: Understanding the emotional impact on hate speech victims is crucial for assessing its harm, as it often induces sadness, anger, fear, and out-group prejudice.

(vi) Frequency: Hate speech can manifest as isolated incidents or persistent harassment, such as mobbing or bullying. Analyzing attack frequency helps gauge the severity of hate speech.

(vii) Time: Hate speech may reference past events, current circumstances, or future actions. Especially, messages that incite violent actions in the near future are dangerous. The temporal dimension should not be neglected.

(viii) Fact-checking: Hate speech often relies on misinformation or distorted facts to perpetuate harmful narratives. Identifying disinformation can aid hate speech detection and inform the severity.

(ix) Topic and context: Hate speech targets various topics, from political ideologies to social identities, and contextual factors must be considered in its assessment. Our analysis underscores the complexity of hate speech, highlighting the need for nuanced approaches to effectively identify, classify, and mitigate its harmful effects.

Year of publication	**Dataset research papers**
2017 *(2)*	(Davidson et al., 2017; Del Vigna et al., 2017)
2018 *(7)*	(Albadi et al., 2018; Founta et al., 2018; Bohra et al., 2018; Mathur et al., 2018) (Sanguinetti et al., 2018; Sprugnoli et al., 2018; Carmona et al., 2018)
2019 *(9)*	(Mulki et al., 2019; Haddad et al., 2019; Chiril et al., 2019) (Ousidhoum et al., 2019; Mandl et al., 2019; Corazza et al., 2019) (Ptaszynski et al., 2019; Fortuna et al., 2019; Basile et al., 2019)
2020 *(5)*	(Mossie and Wang, 2020; Sigurbergsson and Derczynski, 2020) (Bhardwaj et al., 2020; Rizwan et al., 2020; Zueva et al., 2020)
2021 *(6)*	(Karim et al., 2021; Romim et al., 2021; Ljubešić et al., 2021) (Burtenshaw and Kestemont, 2021; Mathew et al., 2021; Assenmacher et al., 2021)
2022 *(8)*	(Nurce et al., 2022; Abebaw et al., 2022; Ayele et al., 2022; Jeong et al., 2022) (Das et al., 2022; Jiang et al., 2022; Shekhar et al., 2022; Demus et al., 2022)
2023 *(1)*	(Pérez et al., 2023)

Table 1: The dataset research papers explored arranged in ascending chronological order. Number in brackets denote the number of explored dataset papers published in the corresponding year.