Recent Advances in Online Hate Speech Moderation: Multimodality and the Role of Large Models

Ming Shan Hee[1]*, **Shivam Sharma**[2,*], **Rui Cao**[3], **Palash Nandi**[2], **Preslav Nakov**[4], **Tanmoy Chakraborty**[2], **Roy Ka-Wei Lee**[1],
[1]SUTD, [2]IIT Delhi, [3]SMU, [4]MBZUAI
[1]{mingshan_hee@mymail., roy_lee@}sutd.edu.sg
[2]{shivam.sharma, palash.nandi, tanchak}@ee.iitd.ac.in
[3]ruicao.2020@phdcs.smu.edu.sg
[4]preslav.nakov@mbzuai.ac.ae

Abstract

Moderating hate speech (HS) in the evolving online landscape is a complex challenge, compounded by the multimodal nature of digital content. This survey examines recent advancements in HS moderation, focusing on the burgeoning role of large language models (LLMs) and large multimodal models (LMMs) in detecting, explaining, debiasing, and countering HS. We begin with a comprehensive analysis of current literature, uncovering how text, images, and audio interact to spread HS. The combination of these modalities adds complexity and subtlety to HS dissemination. We also identified research gaps, particularly in underrepresented languages and cultures, and highlight the need for solutions in low-resource settings. The survey concludes with future research directions, including novel AI methodologies, ethical AI governance, and the development of context-aware systems. This overview aims to inspire further research and foster collaboration towards responsible and human-centric approaches to HS moderation in the digital age.[1]

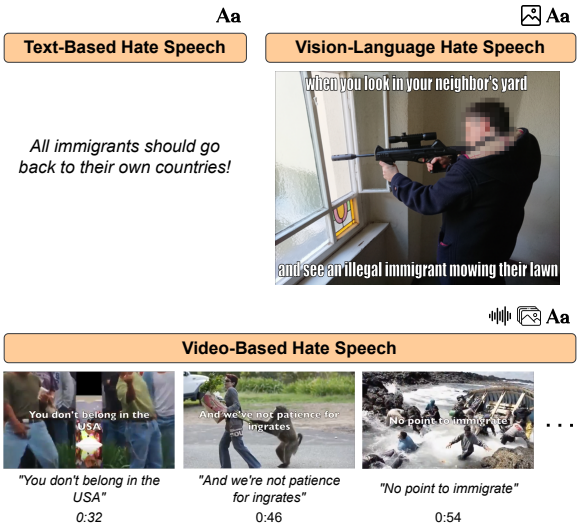

Figure 1: Examples of an anti-migrant HS in different forms, encompassing text, image and/or audio modalities. The text-based, vision-language and video-based HS are taken from the Social Bias Inference Corpus (SBIC) dataset, the Facebook Hateful Memes (FHM) dataset and the Bitchute website, respectively.

1 Introduction

In the era of rapid information exchange and digital connectivity, the rise of hate speech (HS) presents a significant challenge with profound implications for global societies. HS, which is any communication demeaning a person or a group based on social or ethnic characteristics, undermines social harmony and individual safety, both online and offline (Lupu et al., 2023). The recent Israel–Hamas conflict has notably escalated both anti-Muslim and anti-Semitic sentiments worldwide, evidenced by the trending of hashtags such as *#HitlerWasRight* and *#DeathToMuslim* on the social media platform X.[2] Moreover, the Council on American–Islamic Relations reported receiving 774 help requests and bias reports from Muslims in the USA within a 16-day period.[3] While digital interconnectivity facilitates swift information sharing, it simultaneously amplifies the spread and the impact of HS, transcending geographical boundaries.

Technological advancements have transformed the expression of HS, leading to its manifestation in various novel forms. Traditionally, HS was predominantly text-based, found in written materials (Rini et al., 2020), or verbalized in posts, broadcasts, and public speeches (Nielsen, 2002). The digital era has ushered in more complex and subtle variants of HS, engaging multiple sensory modalities. A notable instance is vision-language HS, which fuses visual elements with text, commonly

*These authors contributed equally to this work.
[1]WARNING: This paper contains offensive examples.
[2]https://www.nytimes.com/2023/11/15/technology/hate-speech-israel-gaza-internet.html
[3]https://www.cair.com/press_releases/cair-reports-sharp-increase-in-complaints-reported-bias-incidents-since-107/

Findings of the Association for Computational Linguistics: EACL 2024, pages 4407–4419
November 12-16, 2024

disseminated through captioned images and memes (Uyheng et al., 2020; Kiela et al., 2020). Video-based HS, another emerging form, amalgamates text, visuals, and audio, creating a multi-faceted and potentially more influential mode of communication (Das et al., 2023). Figure 1 shows various HS forms targeting immigrants, underscoring animosity towards individuals of diverse nationalities. The text-based approach overtly projects hostile attitudes towards them in the host country. In vision-language HS, visual (e.g., a person preparing to shoot) and textual elements (e.g., sighting an illegal immigrant mowing the lawn) jointly convey antagonism. The figure also includes a music parody, integrating derogatory visuals with discriminatory audio lyrics, to showcase contempt for immigrants.

While existing research surveys (Rini et al., 2020; Chhabra and Vishwakarma, 2023; Subramanian et al., 2023) have largely focused on text-based HS, they often overlook the complexity of multimodal content. Our survey addresses this gap by offering a comprehensive analysis of HS across various digital platforms, including text, visual, auditory, and combined multimodal expressions. We explore the distinct ways HS manifests in these formats, providing insights into their characteristics and moderation challenges. Additionally, we emphasize the critical role of large language models (LLMs) and large multimodal models (LMMs) in moderating HS, given their ability to process and interpret diverse data types. This survey critically evaluates existing solutions, identifies areas for improvement, and advocates for a shift towards multimodal approaches in HS moderation.

In summary, our paper not only bridges the gap in the existing literature by providing a detailed exploration of multimodal HS but also paves the way for future research in this area. We aim to inspire advancements in HS moderation technology, particularly in the development and refinement of large models, which are imperative for tackling the complex and ever-changing nature of online HS.

Paper Collection. We systematically examined research pertaining to the moderation of various types of hate speech, encompassing text, images, videos, and audio. Our search involved keywords such as '*hate speech*', '*multimodal hate speech*', '*hateful memes*', and similar terms, across scholarly platforms like Google Scholar, DBLP, IEEE Xplore, and ACM Digital Library. Among related research, we further selected state-of-the-art studies, with a particular interest in those using LLMs and LMMs. Due to the need for a manageable scope, we excluded works that did not leverage LLMs or LMMs, or focused narrowly on regional or multilingual aspects without broader relevance. This decision is not a reflection on the quality or importance of these works but rather a necessity to maintain a focused and coherent survey.

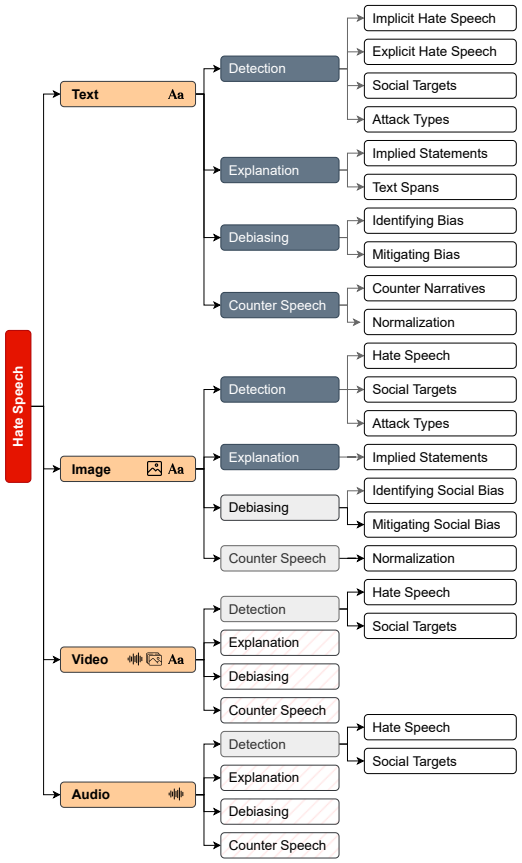

Figure 2: Typology of HS based on modalities and tasks. The dark blue boxes are mature areas with multiple studies; light grey boxes are ongoing research, and hatched boxes are unexplored topics.

2 Hate Speech

HS takes various forms — written text, images, spoken words, and multimedia content — each posing risks of violence, animosity, or prejudice against specific groups. This section reviews existing literature on HS, categorizing it into text-based, image-based, video-based, and audio-based types. For each HS form, we provide a detailed categorization across four tasks: *detection*, *explanation*, *debiasing*, and *counter-speech*. Detection identifies hateful content, forming the basis for further actions. Explanation promotes transparency by clarifying why content is flagged, building trust in automated systems. Debiasing is essential to refine detection systems, ensuring fairness and reduc-

Mod.	Dataset	Task	Labels	Source	# Records
Text	WZ-LS (Waseem and Hovy, 2016)	Det.	**[M.C.]** Sexism, Racism, Neither	Twitter	16,914
	GHC (Kennedy et al., 2018)	Det.	**[M.C.]** VO, HD, CV **[B]** Implicitness **[M.C.]** Hate Targets	Forums	27,665
	Stormfront (de Gibert et al., 2018)	Det.	**[B]** Hateful	StormFront	9,916
	DT (Davidson et al., 2017)	Det.	**[M.C.]** Hateful, Offensive, Neither	Twitter	24,802
	Founta (Founta et al., 2018)	Det.	**[M.C.]** Offensive, Abusive, Hateful Speech, Aggressive, Cyberbullying, Spam, Normal	Twitter	80,000
	DynaHate (Vidgen et al., 2021)	Det.	**[B]** Hateful, **[M.C.]** Hate Targets, **[M.C]** Animosity, Derogation, Dehumanization, Threatening, Support	H-M Adv	41,134
	SBIC (Sap et al., 2020)	Det. Expl.	**[B]** Offensive **[M.C]** Hate Targets **[B]** Intent **[B]** Lewd **[B]** Group **[B]** In-Group **[F.T.]** Implied Statement	Mixed	44,671
	IHC (ElSherief et al., 2021)	Det. Expl.	**[M.C.]** Implicit, Explicit, Non-Hate **[M.C.]** Grievance, Incitement, Inferiority, Irony, Stereotypical, Threatening, Others **[F.T.]** Implied Statement	Twitter	22,584
	HateXplain (Mathew et al., 2021)	Det. Expl.	**[M.C]** Hate, Offensive, Normal **[M.C.]** Hate Targets **[M.L]** Text Rationales/Snippets	Mixed	20,148
	NACL (Masud et al., 2022)	Det. Ctr.	**[M.C.]** Hate Intensity **[M.L.]** Hate Spans **[F.T.]** Hate Speech Normalization	Mixed	4,423
	CONAN (Chung et al., 2019)	Det. Ctr.	**[M.C.]** Hate Types **[M.C.]** Hate Sub-Topic **[F.T.]** CN Generation	Synthetic	14,988
	Multitarget CONAN (Fanton et al., 2021)	Det. Ctr.	**[M.C.]** Hate Targets **[F.T.]** CN Generation	GPT-2	5,003
	Counter Narratives (Das et al., 2023)	Ctr.	**[F.T.]** CN Generation	YouTube	9,119
Img	MMHS150K (Gomez et al., 2020)	Det.	**[B]** Hateful	Twitter	150,000
	FHM (Kiela et al., 2020)	Det.	**[B]** Hateful	Synthetic	10,000
	Finegrained FHM (Mathias et al., 2021)	Det.	**[B]** Hateful **[M.L.M.C]** Protected Category **[M.L.M.C]** Protected Attacks	Synthetic	10,000
	Misogynous Meme (Gasparini et al., 2022)	Det.	**[B]** Misogynistic **[B]** Aggressive **[B]** Ironic	Mixed	800
	MAMI (Fersini et al., 2022)	Det.	**[B]** Misogyny **[M.L.M.C.]** Misogynous, Shaming, Stereotype, Objectification, Violence	Mixed	10,000
	UA-RU Conflict (Thapa et al., 2022)	Det.	**[B]** Hateful	Twitter	5,680
	CrisisHateMM (Bhandari et al., 2023)	Det.	**[B]** Hateful **[B]** Directed **[M.C.]** Hate Targets	Mixed	4,723
	RUHate-MM (Thapa et al., 2024)	Det.	**[B]** Hateful **[M.C]** Hate Targets	Twitter	20,675
	HatReD (Hee et al., 2023)	Expl.	**[F.T]** Explanations	Synthetic	3,228
Video	Bangla Hate Videos (Junaid et al., 2021)	Det.	**[B]** Hateful	YouTube	300
	HateMM (Das et al., 2023)	Det.	**[B]** Hateful **[M.C.]** Hate Targets	Mixed	1,083
	MultiHateClip (Wang et al., 2024)	Det.	**[M.C]** Hateful, Offensive, Normal	YouTube & Bilibili	2,000
Audio	DeToxy (Ghosh et al., 2021)	Det.	**[B]** Hateful	Mixed	2M
	MuTox (Costa-jussà et al., 2024)	Det.	**[B]** Hateful	Mixed	116,000

Table 1: Publicly available datasets for HS detection (Det.), HS explanation (Expl.) and counter HS (Ctr.). Abbreviation: **M.L.**: multi-label, **M.C.**: multi-class, **M.L.M.C.**: multi-label multi-class, **B**: binary, **F.T**: free-text, **H-M Adv**: Human-Machine Adversarial. *Note that multilingual HS is out of the scope for the current review.*

ing bias. Counter-speech involves taking proactive steps to mitigate the impact of hate speech, fostering healthier online dialogue. Although these tasks address different aspects, they collectively form the foundation of an effective content moderation strategy, highlighting both the interconnectedness of HS forms and the research gaps in advancing multimodal HS moderation. Figure 2 illustrates the range of online HS forms. Additionally, Table 1 lists publicly accessible HS datasets in different modalities, providing researchers with essential resources for HS moderation.

2.1 Text-based Hate Speech

Text-based HS encompasses written or typed expressions manifested across online platforms, such as social media posts (Waseem and Hovy, 2016; Founta et al., 2018). Recent studies explored di-

verse aspects of hate and derogatory language, focusing on implicit HS (Sap et al., 2020), targeted groups (Kennedy et al., 2018; Yoder et al., 2022), and types of attacks (ElSherief et al., 2021). As HS detection models improve, it becomes imperative to understand and explain their decision-making processes, mitigating unintended bias (Garg et al., 2023). Additionally, some research shifted towards proactive strategies, including countering HS (Masud et al., 2022).

The detection of text-based HS poses numerous challenges. Detecting hate speech (HS) in a single statement often requires understanding dark humor and cultural nuances (Hee et al., 2024). HS can express underlying intent through sarcasm, irony, or cultural references, which may not be immediately apparent. Linguistic variations, such as slang, dialects, and unconventional language use, further complicate the task. The challenge intensifies when considering the broader context of an utterance (Nagar et al., 2023; Yu et al., 2022a), as statements that seem neutral in isolation may reveal hateful intent when viewed within a conversation. Conversely, what appears offensive might be harmless in context. Therefore, context-aware models are essential for accurately identifying HS by analyzing both individual statements and their surrounding situational context. Expanding the analysis to conversations, such as Reddit threads or WhatsApp chat, adds additional layers of complexity (Naseem et al., 2019). The intent behind a single message can shift based on prior exchanges and the overall tone of the conversation. Furthermore, user-specific features may be important for HS detection (Qian et al., 2018). Data such as a user's posting history, profile, and behavior provide valuable context for identifying hate speech, though using such data raises ethical concerns, particularly regarding privacy.

2.2 Image-based Hate Speech

Image-based HS utilizes visual elements, such as photographs, cartoons, and illustrations, to propagate hate or discrimination against specific groups. A common manifestation of this HS form is memes, which typically consist of images combined with short overlaid text. Although memes often serve humorous or satirical purposes, they are increasingly used to spread hateful content online (Kiela et al., 2020). Recent studies have developed datasets for identifying HS (Gomez et al., 2020), specific targets (Mathias et al., 2021) and types of attacks (Fersini et al., 2022) within these memes. Beyond detection, new approaches analyze and mitigate bias in image-based HS detection models (Hee et al., 2023; Lin et al., 2023). Additionally, new methodologies are emerging to counteract HS transmitted through memes (Van and Wu, 2023).

Image-based HS presents new challenges due to the subtlety of offensive messages concealed within multiple modalities. Images, often embedding symbols, memes, or culturally specific visual cues, require deep cultural and contextual understanding for accurate interpretation. The visual elements and text can subtly imply meanings not immediately evident (Kiela et al., 2020). For example, Figure 1 depicts a man with a gun and text suggesting hostility towards immigrants. Differentiating humour from hate in memes is particularly challenging, influenced by varying cultural, societal, and personal perspectives (Schmid, 2023).

2.3 Video-based Hate Speech

Video-based HS presents a complex challenge, comprising a blend of visuals, audio tracks, and/or textual elements. This form of HS ranges from professionally produced propaganda to amateur videos on social media platforms like YouTube and TikTok (Das et al., 2023; Wang et al., 2024). The engaging nature of video content and its easy dissemination across digital networks significantly heighten its potential for harm. Echoing the concerns of image-based HS, video-based HS also contributes to the normalization of hateful ideologies and can profoundly influence public opinion. Contemporary research primarily focuses on identifying video-based HS and categorizing its various subtypes (Wu and Bhandary, 2020). Nonetheless, the amount of research on video-based HS is less developed than text-based and image-based HS, particularly in areas such as analyzing and mitigating model bias, elucidating decision-making processes, and devising counterstrategies. These gaps, likely stemming from the rapid pace of technological advancements and evolving digital trends, underscore the need for further research to promote a more harmonious online environment.

Detecting hate speech (HS) in videos is challenging and resource-intensive because it requires understanding various elements, including text, images, and audio, both independently and in combination. Each component can independently contain hateful content, further complicating the detection process. The duration of videos further exacerbates this challenge, as longer content necessitates

more extensive review and analysis, with potential shifts in context over time. Moreover, subtle visual cues and sophisticated editing techniques can be employed to discreetly embed hate messages, making their detection by automated tools particularly challenging. Additionally, video content analysis requires considerable computational resources and time, posing a substantial challenge for organizations to detect and address HS in video formats.

2.4 Audio-based Hate Speech

Audio-based HS entails the analysis of sound waves to discern elements such as pitch, intonation, and the contextual meaning of spoken words. This form of HS can originate from a variety of audio channels, including real-time conversations, podcasts, and other forms of audio media. The methodologies for addressing audio-based HS are diverse, targeting different facets of the issue. For instance, Barakat et al. (2012) employed a straightforward keyword-based approach to identify segments of HS, while Wazir et al. (2020) engaged in a detailed classification of offensive categories in audio-based HS, showcasing a nuanced method of understanding and categorizing this form of HS. This research area is still in its developmental stages, partly due to the scarcity of dataset. Nonetheless, recognizing the variety and significance of the approaches and techniques employed in this field is imperative. This recognition not only sheds light on the current state of research but also illuminates potential avenues for future exploration.

Detecting HS in audio recordings presents unique challenges, primarily related to the transcription and interpretation of spoken words. The accuracy of speech recognition is crucial, especially when dealing with diverse accents, background noise, or poor audio quality. Additionally, the tone and intonation of spoken language play a significant role in conveying intent, which can substantially alter the meaning of words. This aspect poses a challenge for detection based solely on text transcripts, as subtle nuances in vocal expression may be lost during transcription. Moreover, non-verbal audio elements, such as sound cues or background noises, are pivotal in contextualizing speech. However, these elements are often difficult to interpret using automated methods.

3 Methodology

This section reviews the state-of-the-art methodologies that have significantly contributed to primary areas of HS research, particularly those involving large models. First, we discuss the recent capabilities of large models (Section 3.1). Subsequently, we explore studies in four important HS areas: *detection* (Section 3.2), *explanation* (Section 3.3), *debiasing* (Section 3.4), and *counter-speech* (Section 3.5), focusing on works using large models. This review highlights the emerging trends, providing insights into how large models can be used to understand and address HS in its various forms.

3.1 Large Models

The emergence of large foundation models, such as LLMs and LMMs, marks a significant milestone in artificial intelligence research, showcasing unprecedented capabilities in understanding and generating data across different formats (Zhao et al., 2023). LLMs are designed to excel in language understanding and text generation (Touvron et al., 2023). In contrast, LMMs are adept at processing and interpreting various data types, including visual, textual, and auditory inputs, enabling a broader spectrum of applications (Yang et al., 2023b). These foundation models have opened new avenues for identifying and mitigating hateful content, which requires nuanced understanding of language and context.

Here, we regard LLMs and LMMs as models with several billion parameters, aligning with the definition widely accepted and analyzed in numerous studies of large-scale models (Luo et al., 2023).

3.2 Hate Speech Detection

The leading detection techniques for HS vary according to the modality of the content, encompassing approaches from transformer-based models to spectrogram-based classification models. For text-based HS detection, approaches range from embedding-based methods to advanced neural models (Cao et al., 2020; Davidson et al., 2017; Badjatiya et al., 2017; Fortuna and Nunes, 2018). AngryBERT (Awal et al., 2021) fine-tunes BERT using a multi-task learning strategy for binary text HS detection. PromptHate (Cao et al., 2022) combines demonstration sampling and in-context learning to fine-tune RoBERTa for hateful meme detection. In audio-based HS detection, ensemble techniques such as AdaBoost, Naive Bayes, and Random Forest have been employed. (Boishakhi et al., 2021;

Ibañez et al., 2021). CNNs are also used to convert audio into spectrograms (Medina et al., 2022), with self-attentive CNNs extracting audio features (Yousefi and Emmanouilidou, 2021). For video-based HS detection, a combination of BERT, ViT, and MFCC has been used for text, image, and audio modality analysis, respectively (Das et al., 2023). Note that audio-based and video-based HS detection are emerging areas with significant potential for future advancements.

Transformer-based models have significantly advanced the detection of text-based and image-based HS; yet they encounter challenges. For text-based models, a major hurdle is generalizing to out-of-distribution datasets, often hindered by limited vocabulary and the rarity of implicit HS in many datasets (Ocampo et al., 2023b). To overcome this, recent initiatives include adversarial HS generation and in-context learning with LLMs. Ocampo et al. (2023a) introduced a method using GPT-3 to generate implicit HS, aiming to both challenge and improve HS classifiers. Concurrently, Wang et al. (2023b) developed a technique for optimizing example selection for in-context learning in LLMs.

In image-based HS, the primary challenge lies in deciphering implicit hate messages within memes. This often stems from the loss of information during the extraction of text-based features from images, a common step in many methodologies (Lee et al., 2021; Pramanick et al., 2021; Cao et al., 2022). Furthermore, the implicit HS in memes can be concealed by seemingly unrelated text and images, as illustrated in Figure 1. To address these challenges, recent strategies include employing LMMs with prompting techniques and/or knowledge distillation. Pro-Cap (Cao et al., 2023) addresses the issue of information loss in image-to-text conversion by prompting an LMM in a QA format, enhancing the generated caption's quality and informativeness. To tackle the problem of disconnected text and images, MR.HARM (Lin et al., 2023) utilizes an LMM to generate potential rationales. These rationales are subsequently employed to fine-tune supervised HS classification systems through knowledge distillation, improving the detection of hateful memes.

3.3 Hate Speech Explanation

A major challenge in contemporary HS detection methods is their lack of explainability in decision-making processes. Explainability is crucial for fostering user trust and facilitating systems that require human interaction (Balkir et al., 2022). One proposed solution involves training supervised models that not only categorize HS but also provide rationales for these classifications. Sap et al. (2020) and ElSherief et al. (2021) developed text-based HS datasets with human-annotated explanations, setting benchmarks for identifying underlying hate. Similarly, Hee et al. (2023) compiled a dataset for hateful memes, complete with human-annotated explanations and benchmarks. However, collecting human-written explanations is not only time-consuming but also susceptible to individual biases. Moreover, it involves the risk of subjecting human annotators to prolonged exposure to HS, which can have adverse psychological effects.

Recent studies have delved into employing LLMs to generate plausible and meaningful explanations for HS. For instance, Wang et al. (2023a) demonstrated that GPT-3 can craft convincing and effective explanations for HS, a finding substantiated by extensive human evaluations. Additionally, HARE (Yang et al., 2023a) introduces two prompting methods that generate rationales for HS, enhancing the training of HS detection models and improving its performance. This approach presents an alternative means of developing insightful explanations, while simultaneously mitigating the risks associated with prolonged human exposure to HS. Nevertheless, this area of research is still nascent, thus presenting numerous opportunities for further investigation and development.

3.4 Hate Speech Debiasing

Bias in HS detection models poses a significant risk to their effectiveness and fairness, leading to potential adverse impacts on individuals and society. Addressing this, numerous studies have focused on identifying and mitigating bias in these models. Sap et al. (2019) found that two widely-used corpora exhibit bias against African American English, which increases the likelihood of classifying tweets in this dialect as hateful. Hee et al. (2022) conducted a quantitative analysis of modality bias in hateful meme detection, observing that the image modality significantly influences model predictions. Their study also highlighted the tendency of these models to generate false positives when encountering specific group identifier terms.

Beyond merely identifying biases, various studies have introduced innovative methods to reduce these biases within models. Kennedy et al. (2020) developed a regularization technique utilizing SOC

post-hoc explanations to address group identifier bias. Similarly, Rizzi et al. (2023) observed that models exhibit biases towards terms linked with stereotypical notions about women, such as *dishwasher* and *broom*. To counteract this, the authors proposed a bias mitigation strategy using Bayesian Optimization, which effectively lessened the bias while preserving overall model performance.

These efforts underscore the critical importance of not only recognizing, but also actively mitigating bias. This is especially vital as large models increasingly dominate the landscape for generating explanations and enabling transfer learning.

3.5 Counter Speech

The approach to countering HS focuses on generating non-aggressive responses that either reduce the spread of HS or transform it into respectful and inoffensive speech. Recent research categorizes counter-speech into various response types and emphasizes the importance of contextual understanding. Yu et al. (2023) developed a taxonomy of responses to HS, showcasing the diversity of counter-speech tactics. Mathew et al. (2019) proposed context-specific strategies such as narrative persuasion and active rebuttal. CONAN (Chung et al., 2021) focused on generating counter-narratives that challenge hate directed at marginalized groups using reliable evidence, logical arguments, and diverse perspectives. These non-aggressive strategies reduce the spread of hate speech and foster positive discourse. Beyond generating non-aggressive responses, other approaches involve diminishing (i.e., normalization) or eliminating (i.e., correction) the level of hate in HS. NACL (Masud et al., 2022) used neural networks to paraphrase hate speech, effectively lowering the intensity of hate. Van and Wu (2023) prompted LMMs to correct HS in memes by replacing hateful text with positive and respectful language.

These studies underscore the critical role of generative models in annotating and developing counter-speech strategies. This further signifies the future opportunities of LLMs and LMMs in enhancing approaches to combat hate speech.

4 Challenges

In the dynamic realm of research, especially in areas related to user-generated content and online harmfulness, numerous challenges persist that shape the trajectory and emphasis of scholarly investigations. These challenges, ranging from technical to ethical, define the landscape in which research on HS moderation and detection operates.

Data Complexity, Quality, and Sourcing. The subtlety of some hate speech, known as implicit hate speech, presents a considerable challenge in identifying and understanding the underlying intent, as these intents can hide within seemingly neutral language or actions (Sap et al., 2020; Ocampo et al., 2023b; Kiela et al., 2020). This difficulty highlights the complexity of human communication and biases, where hateful messages can be conveyed indirectly or through coded language. Furthermore, sourcing data from diverse platforms such as Gab, YouTube, and 4chan introduces difficulties in standardization and interpretation (Mariconti et al., 2019). Additionally, the uneven distribution of hate instances across datasets poses significant obstacles for accurate model training. (Cao and Lee, 2020). These challenges underscore the need for advanced methods capable of navigating the intricate and multifaceted nature of data.

Model Performance and Generalizability. Recent research highlights the importance of enhancing HS detection models for adaptability in various scenarios and contexts. An exemplary example is making HS detection generalizable and effective across domains (Awal et al., 2021), underscoring the need for models to be versatile and not overly reliant on specific content cues such as domain, region, demography, and more. The development of systems like VulnerCheck (Mariconti et al., 2019) exemplifies the demand for models that perform well regardless of the context, and that can adapt to the ever-evolving nature of online material. Such adaptability is crucial for identifying and managing new hateful content, especially such designed to bypass advanced AI technologies. The adoption of technologies, like Few-Shot Learner (FSL), for quick adaptation to this evolving landscape is a promising direction.[4] However, it is imperative that these technologies not only understand the content, but also integrate critical aspects of cultural, behavioral, and conversational contexts.

Expression and Modality Variabilities. Research has highlighted the complexities of interpreting hate speech (HS) across various modalities

[4] https://ai.meta.com/blog/harmful-content-can-evolve-quickly-our-new-ai-system-adapts-to-tackle-it/

(Boishakhi et al., 2021; Kiela et al., 2020). In text-based HS, implicit hate messages often use dark humor or sarcasm to obscure their true intent, making detection particularly challenging (Sap et al., 2020; ElSherief et al., 2021). For image-based HS, models face difficulty identifying subtle cross-modality nuances that convey the underlying message (Hee et al., 2022; Rizzi et al., 2023). These challenges become even more pronounced with audio and video content due to factors such as accents, background noise, inconsistent audio quality, and the inherently ambiguous nature of toxic content (Yousefi and Emmanouilidou, 2021). Additionally, the potential for misinterpretation and the models' sensitivity to specific trigger words exacerbate these issues (Sridhar and Yang, 2022). Addressing these variabilities is essential and calls for dedicated efforts in future research.

Contextualization. Effective hate speech moderation requires a nuanced understanding of the HS context. Multi-turn interactions in social media conversations, like Reddit threads or Twitter discussions, play a key role in detecting implicit hate speech (Ghosh et al., 2023) and generating counter speech (Yu et al., 2022b). Additionally, Meng et al. proposed DRAG++, a model that predicts hate intensity by analyzing both the content and the full context of conversation threads (Meng et al., 2023). Furthermore, geographic-specific factors, such as local slang and cultural differences, influence a model's ability to generalize across different regions (Lee et al., 2023). These challenges highlight the need for sophisticated algorithms capable of interpreting language within its contextual usage, thereby enhancing the accuracy and effectiveness of hate speech moderation strategies.

Emerging Domains. Exploring new and evolving fields, such as the metaverse, presents a distinct set of challenges (Medina et al., 2022). The core of these challenges lies the need to adapt current HS detection methods to new contexts and to develop new strategies specifically designed for the unique characteristics of these platforms. The dynamic and immersive nature of these emerging environments necessitates a re-evaluation and potential re-engineering of current HS detection and mitigation strategies. Future research requires a deep understanding of both technological advancements and the social dynamics within these virtual spaces to ensure effectiveness in detecting and mitigating HS within these evolving digital landscapes.

Bias and Ethical Concerns. Addressing bias and upholding ethical considerations in HS detection systems poses a significant challenge. Several studies have highlighted these concerns, introducing functional tests for evaluating HS detection models (Röttger et al., 2022; Ng et al., 2024; Röttger et al., 2020). These challenges are not purely technical but also moral, underscoring the importance of ensuring HS systems operate equitably and do not perpetuate societal biases. Developing responsible HS systems, therefore, require a multidisciplinary approach that combines technical expertise with ethical and societal awareness, ensuring alignment with ethical standards and societal values.

In summary, the areas of HS detection and moderation are confronted with multifaceted challenges. These arise from the inherent complexities of data, technological limitations, modality variabilities, dataset biases, and the uncharted territories of emerging domains like the metaverse. To effectively navigate these obstacles, a concerted and multidisciplinary effort is essential. It calls for the development of methodologies that are not only sophisticated and robust but also highly adaptable. Such methodologies must be capable of contending with the dynamic and often unpredictable nature of user-generated content and online interactions. The future of this field hinges on our ability to continuously evolve and innovate, ensuring that our approaches remain relevant, effective, and ethically sound in an ever-changing digital landscape.

5 Future Directions

Cross-Modality Context Understanding. As hate speech extends beyond mere text to encompass multiple forms of media (multimodality), it becomes crucial for models to have a proficient understanding of context across modalities. Hence, it is imperative that models not only identify hateful content within text or images separately, but also grasp how the combination of text and images can alter the message (Kiela et al., 2020). For instance, an image that is benign on its own might become hateful when paired with specific text. Research could focus on developing models that more effectively understand context across modalities.

Low-Resource Hate Speech Adaptation. Domain adaptation between related tasks has gained significant attention. In the domain of hate speech, an exemplary application is the cross-lingual transfer learning for detecting hate speech across differ-

ent languages. Winata et al. (Winata et al., 2022) use few-shot in-context learning and fine-tuning techniques to adapt insights from languages with abundant resources to those with fewer resources. Given the widespread presence of hate speech and its relatively consistent definitions across different forms, there is potential to extend knowledge from text-based hate speech with abundant resources to other low-resource forms of hate speech. Future research should aim to develop models capable of pre-training on a broad spectrum of multimodal data, including text, images, audio, and videos, to enhance transfer learning capabilities.

Humour & Sarcasm Understanding. Comprehending humor and sarcasm involves recognizing subtle linguistic signals and understanding the broader context, which includes cultural, social, and environmental factors. LLMs are adept at processing language but might not entirely capture these intricate details or fully understand the specific circumstances surrounding a statement. Additionally, humor and sarcasm often hinge on wordplay, double meanings, or ambiguous interpretations. Although LLMs can identify language patterns, they might struggle to differentiate between straightforward and figurative speech. Research efforts can focus on enhancing the capability of LLMs and LMMs to interpret sarcasm and humor, particularly dark humor, which conceals itself within the context of a sentence.

Multicultural Moderation. A challenge in hate speech detection lies in the varying cultural and contextual cues across different countries and regions. These subtle cues often require a nuanced understanding of local languages, dialects, slang, and social norms. This complexity makes it difficult for automated systems to identify and differentiate hate speech from non-offensive content. Nguyen et al. (2023) demonstrated how providing cultural common-sense knowledge can alter GPT-3's behaviour, leading it to produce more accurate and culturally sensitive questions. Similarly, future research could aim to curate HS dataset with regional culture information and build culturally-aware LLMs and LMMs by injecting and fine-tuning models with cultural knowledge.

Real-Time Monitoring. The vocabulary of hate speech is constantly changing and evolving, particularly in online spaces. Although adapting to different domains can enhance a model's capacity to apply its knowledge across a range of current datasets, the ongoing development of new slurs, coded terms, and symbolic expressions presents a considerable obstacle to the successful detection of hate speech. Research efforts can focus on continual learning methods that enable these models to be updated regularly while minimising the adjustments to their parameters.

Factual Grounding. Although current methods in generating HS explanation using large-scale models (refer to Section 3.3) have shown promise, they still face significant challenges. These large models are prone to "hallucinations" producing responses that can be factually incorrect, illogical, or unrelated to the initial prompt (Ji et al., 2023). Consequently, while these recent advancements are promising, the explanations generated by these models are susceptible to misinformation and require verification. Future research should aim to improve the accuracy and relevance of these explanations, which might involve anchoring the explanations in verifiable facts and developing techniques to identify and rectify any discrepancies.

6 Conclusion

We highlighted the advancements in HS moderation, underscoring the pivotal role of LLMs and LMMs. Despite these strides, challenges remain, particularly in inclusivity and nuanced detection. Future research should focus on developing AI methodologies that are more context-aware and ethically governed. This endeavor is not only a technological challenge, but also a moral imperative, necessitating interdisciplinary collaboration. As we advance, it is crucial to ensure that technological advancements are matched with a commitment to responsibility, striving for a digital environment that is secure and welcoming for everyone.

Acknowledgement

This research is supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 2 (Award ID: MOE-T2EP20222-0010). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of the Ministry of Education, Singapore. Tanmoy Chakraborty acknowledges Anusandhan National Research Foundation (CRG/2023/001351) for the financial support.

Limitations

There are two limitations to the scope and coverage of our survey paper.

Scope. In this survey paper, we specifically focus on the role of large models in multimodal "hateful speech moderation". We recognize that there is extensive research on toxic and harmful content that is closely related to hatefulness. However, hate speech has a distinct definition that attacks need to discriminate against a group of people based on specific traits such as race, gender and sexual orientation. Hence, while these closely related areas are significant, they fall outside the scope of this survey paper. We also recognize different forms of hate speech can exist in multiple languages, resulting in exciting research on multilingual hate speech. However, the primary goal of this paper is to highlight the evolution and adaptation of hate speech in various forms of digital content. Therefore, the topic of multilingualism is beyond the scope of this paper.

Research Paper Coverage. Although numerous research studies on hate speech have been conducted over the past decades, we have focused on state-of-the-art works that either employed large models in their studies or pioneered specific hate speech tasks. This selection enables us to maintain the brevity of this survey paper while focusing our discussion on the promising areas of using large models for hate speech moderation.

References

Md Rabiul Awal, Rui Cao, Roy Ka-Wei Lee, and Sandra Mitrović. 2021. Angrybert: Joint learning target and emotion for hate speech detection. In *PAKDD*.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, pages 759–760.

Esma Balkir, Isar Nejadgholi, Kathleen C Fraser, and Svetlana Kiritchenko. 2022. Necessity and sufficiency for explaining text classifiers: A case study in hate speech detection. In *NAACL*.

M. S. Barakat, C. H. Ritz, and D. A. Stirling. 2012. Detecting offensive user video blogs: An adaptive keyword spotting approach. In *ICALIP*.

Aashish Bhandari, Siddhant Bikram Shah, Surendrabikram Thapa, Usman Naseem, and Mehwish Nasim. 2023. Crisishatemm: Multimodal analysis of directed and undirected hate speech in text-embedded images from russia-ukraine conflict. In *CVPR Workshops*. IEEE.

Fariha Tahosin Boishakhi, Ponkoj Chandra Shill, and Md Golam Rabiul Alam. 2021. Multi-modal hate speech detection using machine learning. In *Big Data*. IEEE.

Rui Cao, Ming Shan Hee, Adriel Kuek, Wen-Haw Chong, Roy Ka-Wei Lee, and Jing Jiang. 2023. Procap: Leveraging a frozen vision-language model for hateful meme detection. In *ACMMM*.

Rui Cao and Roy Ka-Wei Lee. 2020. Hategan: Adversarial generative-based data augmentation for hate speech detection. In *COLING*.

Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. 2022. Prompting for multimodal hateful meme classification. In *EMNLP*.

Rui Cao, Roy Ka-Wei Lee, and Tuan-Anh Hoang. 2020. Deephate: Hate speech detection via multi-faceted text representations. In *WebSci*.

Anusha Chhabra and Dinesh Kumar Vishwakarma. 2023. A literature survey on multimodal and multilingual automatic hate speech identification. *Multimedia Systems*.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *ACL*.

Yi-Ling Chung, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Towards knowledge-grounded counter narrative generation for hate speech. In *ACL (Findings)*.

Marta R Costa-jussà, Mariano Coria Meglioli, Pierre Andrews, David Dale, Prangthip Hansanti, Elahe Kalbassi, Alex Mourachko, Christophe Ropers, and Carleigh Wood. 2024. Mutox: Universal multilingual audio-based toxicity dataset and zero-shot detector. *arXiv preprint arXiv:2401.05060*.

Mithun Das, Rohit Raj, Punyajoy Saha, Binny Mathew, Manish Gupta, and Animesh Mukherjee. 2023. Hatemm: A multi-modal dataset for hate video classification. In *ICWSM*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *ICWSM*.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Proc. of the 2st workshop on ab. lang. online*.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *EMNLP*.

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroglu, and Marco Guerini. 2021. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. In *ACL*.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. Semeval-2022 task 5: Multimedia automatic misogyny identification. In *SemEval@NAACL*.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

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. In *ICWSM*.

Tanmay Garg, Sarah Masud, Tharun Suresh, and Tanmoy Chakraborty. 2023. Handling bias in toxic speech detection: A survey. *ACM Computing Surveys*, 55(13s):1–32.

Francesca Gasparini, Giulia Rizzi, Aurora Saibene, and Elisabetta Fersini. 2022. Benchmark dataset of memes with text transcriptions for automatic detection of multi-modal misogynistic content. *Data in brief*.

Sreyan Ghosh, Samden Lepcha, S Sakshi, Rajiv Ratn Shah, and Srinivasan Umesh. 2021. Detoxy: A large-scale multimodal dataset for toxicity classification in spoken utterances. *arXiv preprint arXiv:2110.07592*.

Sreyan Ghosh, Manan Suri, Purva Chiniya, Utkarsh Tyagi, Sonal Kumar, and Dinesh Manocha. 2023. CoSyn: Detecting implicit hate speech in online conversations using a context synergized hyperbolic network. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6159–6173.

Raul Gomez, Jaume Gibert, Lluís Gómez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *IEEE/WACV*.

Ming Shan Hee, Rui Cao, Tanmoy Chakraborty, and Roy Ka-Wei Lee. 2024. Understanding (dark) humour with internet meme analysis. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 1276–1279.

Ming Shan Hee, Wen-Haw Chong, and Roy Ka-Wei Lee. 2023. Decoding the underlying meaning of multimodal hateful memes. In *IJCAI*.

Ming Shan Hee, Roy Ka-Wei Lee, and Wen-Haw Chong. 2022. On explaining multimodal hateful meme detection models. In *WWW*.

Michael Ibañez, Ranz Sapinit, Lloyd Antonie Reyes, Mohammed Hussien, Joseph Marvin Imperial, and Ramon Rodriguez. 2021. Audio-based hate speech classification from online short-form videos. In *IALP*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*.

Mohd Istiaq Hossain Junaid, Faisal Hossain, and Rashedur M Rahman. 2021. Bangla hate speech detection in videos using machine learning. In *UEMCON*, pages 0347–0351. IEEE.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, et al. 2018. The gab hate corpus: A collection of 27k posts annotated for hate speech. *PsyArXiv. July*.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing hate speech classifiers with post-hoc explanation. In *ACL*.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *NeurIPS*.

Nayeon Lee, Chani Jung, and Alice Oh. 2023. Hate speech classifiers are culturally insensitive. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 35–46.

Roy Ka-Wei Lee, Rui Cao, Ziqing Fan, Jing Jiang, and Wen-Haw Chong. 2021. Disentangling hate in online memes. In *ACMMM*.

Hongzhan Lin, Ziyang Luo, Jing Ma, and Long Chen. 2023. Beneath the surface: Unveiling harmful memes with multimodal reasoning distilled from large language models. In *EMNLP (Findings)*.

Junyu Luo, Cao Xiao, and Fenglong Ma. 2023. Zero-resource hallucination prevention for large language models. *arXiv preprint arXiv:2309.02654*.

Yonatan Lupu, Richard Sear, Nicolas Velásquez, Rhys Leahy, Nicholas Johnson Restrepo, Beth Goldberg, and Neil F Johnson. 2023. Offline events and online hate. *PLoS one*.

Enrico Mariconti, Guillermo Suarez-Tangil, Jeremy Blackburn, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Jordi Luque Serrano, and Gianluca Stringhini. 2019. You know what to do proactive detection of youtube videos targeted by coordinated hate attacks. *Proc. of the ACM on HCI*.

Sarah Masud, Manjot Bedi, Mohammad Aflah Khan, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Proactively reducing the hate intensity of online posts via hate speech normalization. In *ACM-SIGKDD*.

Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2019. Thou shalt not hate: Countering online hate speech. In *ICWSM*.

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 *AAAI*.

Lambert Mathias, Shaoliang Nie, Aida Mostafazadeh Davani, Douwe Kiela, Vinodkumar Prabhakaran, Bertie Vidgen, and Zeerak Waseem. 2021. Findings of the WOAH 5 shared task on fine grained hateful memes detection. In *WOAH*.

Robin Medina, Judith Njoku, Jae Min Lee, and Dong-Seong Kim. 2022. Audio-based hate speech detection for the metaverse using cnn. In *KICS*.

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

Seema Nagar, Ferdous Ahmed Barbhuiya, and Kuntal Dey. 2023. Towards more robust hate speech detection: using social context and user data. *Social Network Analysis and Mining*, 13(1):47.

Usman Naseem, Imran Razzak, and Ibrahim A Hameed. 2019. Deep context-aware embedding for abusive and hate speech detection on twitter. *Aust. J. Intell. Inf. Process. Syst.*, 15(3):69–76.

Ri Chi Ng, Nirmalendu Prakash, Ming Shan Hee, Kenny Tsu Wei Choo, and Roy Ka-Wei Lee. 2024. Sghatecheck: Functional tests for detecting hate speech in low-resource languages of singapore. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 312–327.

Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. 2023. Extracting cultural commonsense knowledge at scale. In *WWW*, pages 1907–1917.

Laura Beth Nielsen. 2002. Subtle, pervasive, harmful: Racist and sexist remarks in public as hate speech. *Journal of Social issues*.

Nicolas Ocampo, Elena Cabrio, and Serena Villata. 2023a. Playing the part of the sharp bully: Generating adversarial examples for implicit hate speech detection. In *ACL (Findings)*.

Nicolas Benjamin Ocampo, Ekaterina Sviridova, Elena Cabrio, and Serena Villata. 2023b. An in-depth analysis of implicit and subtle hate speech messages. In *ACL*.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. Momenta: A multimodal framework for detecting harmful memes and their targets. pages 4439–4455. ACL.

Jing Qian, Mai ElSherief, Elizabeth Belding, and William Yang Wang. 2018. Leveraging intra-user and inter-user representation learning for automated hate speech detection. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 118–123.

Rini Rini, Ema Utami, and Anggit Dwi Hartanto. 2020. Systematic literature review of hate speech detection with text mining. In *ICORIS*. IEEE.

Giulia Rizzi, Francesca Gasparini, Aurora Saibene, Paolo Rosso, and Elisabetta Fersini. 2023. Recognizing misogynous memes: Biased models and tricky archetypes. *Inf. Proc. Mgmt.*

Paul Röttger, Haitham Seelawi, Debora Nozza, Zeerak Talat, and Bertie Vidgen. 2022. Multilingual hatecheck: Functional tests for multilingual hate speech detection models. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 154–169.

Paul Röttger, Bertram Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet B Pierrehumbert. 2020. Hatecheck: Functional tests for hate speech detection models. *arXiv preprint arXiv:2012.15606*.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *ACL*.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *ACL*.

Ursula Kristin Schmid. 2023. Humorous hate speech on social media: A mixed-methods investigation of users' perceptions and processing of hateful memes. *New Media & Society*.

Rohit Sridhar and Diyi Yang. 2022. Explaining toxic text via knowledge enhanced text generation. In *NAACL*.

Malliga Subramanian, Veerappampalayam Easwaramoorthy Sathiskumar, G Deepalakshmi, Jaehyuk Cho, and G Manikandan. 2023. A survey on hate speech detection and sentiment analysis using machine learning and deep learning models. *Alexandria Engineering Journal*.

Surendrabikram Thapa, Farhan Ahmad Jafri, Kritesh Rauniyar, Mehwish Nasim, and Usman Naseem. 2024. Ruhate-mm: Identification of hate speech and targets using multimodal data from russia-ukraine crisis. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 1854–1863.

Surendrabikram Thapa, Aditya Shah, Farhan Jafri, Usman Naseem, and Imran Razzak. 2022. A multimodal dataset for hate speech detection on social media: Case-study of russia-ukraine conflict. In *CASE@EMNLP*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Joshua Uyheng, Lynnette Hui Xian Ng, and Kathleen M Carley. 2020. Visualizing vitriol: Hate speech and image sharing in the 2020 singaporean elections. *discourse*, 7:17.

Minh-Hao Van and Xintao Wu. 2023. Detecting and correcting hate speech in multimodal memes with large visual language model. *CoRR*.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *ACL-IJCNLP*.

Han Wang, Ming Shan Hee, Md. Rabiul Awal, Kenny Tsu Wei Choo, and Roy Ka-Wei Lee. 2023a. Evaluating GPT-3 generated explanations for hateful content moderation. In *IJCAI*.

Han Wang, Tan Rui Yang, Usman Naseem, and Roy Ka-Wei Lee. 2024. Multihateclip: A multilingual benchmark dataset for hateful video detection on youtube and bilibili. *arXiv preprint arXiv:2408.03468*.

Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang. 2023b. Large language models are latent variable models: Explaining and finding good demonstrations for in-context learning. In *NeurIPS*.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *NAACL SRW*.

Abdulaziz Saleh Ba Wazir, Hezerul Abdul Karim, Mohd Haris Lye Abdullah, Sarina Mansor, Nouar AlDahoul, Mohammad Faizal Ahmad Fauzi, and John See. 2020. Spectrogram-based classification of spoken foul language using deep cnn. In *MMSP*.

Genta Winata, Shijie Wu, Mayank Kulkarni, Thamar Solorio, and Daniel Preoţiuc-Pietro. 2022. Cross-lingual few-shot learning on unseen languages. In *AACL*.

Ching Seh Wu and Unnathi Bhandary. 2020. Detection of hate speech in videos using machine learning. In *Int. Conf. on Comp. Science and Comp. Int.* IEEE.

Yongjin Yang, Joonkee Kim, Yujin Kim, Namgyu Ho, James Thorne, and Se-Young Yun. 2023a. HARE: Explainable hate speech detection with step-by-step reasoning. In *EMNLP (Findings)*, pages 5490–5505.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023b. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*.

Michael Miller Yoder, Lynnette Hui Xian Ng, David West Brown, and Kathleen M Carley. 2022. How hate speech varies by target identity: a computational analysis. *arXiv preprint arXiv:2210.10839*.

Midia Yousefi and Dimitra Emmanouilidou. 2021. Audio-based toxic language classification using self-attentive convolutional neural network. In *EUSIPCO*.

Xinchen Yu, Eduardo Blanco, and Lingzi Hong. 2022a. Hate speech and counter speech detection: Conversational context does matter. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5918–5930.

Xinchen Yu, Eduardo Blanco, and Lingzi Hong. 2022b. Hate speech and counter speech detection: Conversational context does matter. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

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*, pages 7275–7289.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.