

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

Ming Shan Hee^{1*}, Shivam Sharma^{2*}, Rui Cao³, Palash Nandi²,
Tanmoy Chakraborty² and Roy Ka-Wei Lee¹

¹SUTD, ²IIT Delhi, ³SMU

{mingshan.hee@mymail., roy_lee@ }sutd.edu.sg,
{shivam.sharma, palash.nandi., tanchak}@ee.iitd.ac.in,
ruicao.2020@phdcs.smu.edu.sg

Abstract

In the evolving landscape of online communication, moderating hate speech (HS) presents an intricate challenge, compounded by the multimodal nature of digital content. This comprehensive survey delves into the recent strides in HS moderation, spotlighting the burgeoning role of large language models (LLMs) and large multimodal models (LMMs). Our exploration begins with a thorough analysis of current literature, revealing the nuanced interplay between textual, visual, and auditory elements in propagating HS. We uncover a notable trend towards integrating these modalities, primarily due to the complexity and subtlety with which HS is disseminated. A significant emphasis is placed on the advances facilitated by LLMs and LMMs, which have begun to redefine the boundaries of detection and moderation capabilities. We identify existing gaps in research, particularly in the context of underrepresented languages and cultures, and the need for solutions to handle low-resource settings. The survey concludes with a forward-looking perspective, outlining potential avenues for future research, including exploring novel AI methodologies, the ethical governance of AI in moderation, and developing more nuanced, context-aware systems. This comprehensive overview aims to catalyze further research and foster a collaborative effort towards more sophisticated, responsible, and human-centric approaches to HS moderation in the digital era.¹

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. Hate speech, which is any communication demeaning a person or group based on social or ethnic characteristics,

*Both authors contributed equally to this work

¹WARNING: This paper contains offensive examples.

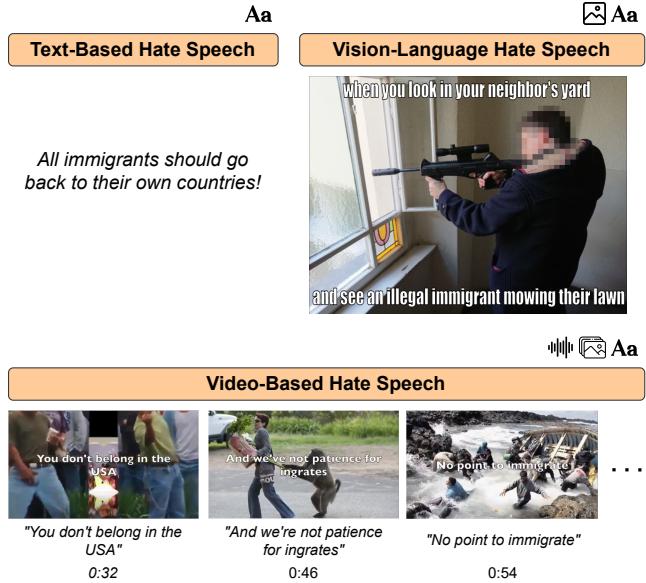


Figure 1: Examples of an anti-migrant HS in different forms, encompassing text, image and/or audio modalities. The text-based, vision-language and video-based HS are taken from the Social Bias Inference Corpus (SBIC) dataset, the Facebook Hateful Memes (FHM) dataset and the Bitchute website, respectively.

undermines social harmony and individual safety, both online and offline [Awan and Zempi, 2016; Lupu *et al.*, 2023; Wiedlitzka *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 social media platform X.² Additionally, 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 impact of HS, transcending geographical boundaries.

²<https://www.nytimes.com/2023/11/15/technology/hate-speech-israel-gaza-internet.html>

³https://www.cair.com/press_releases/cair-reports-sharp-increase-in-complaints-reported-bias-incidents-since-107/

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 [Ana, 1999], or verbalized in posts, broadcasts, and public speeches [Nielsen, 2002]. However, 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 textual content, commonly disseminated through captioned images and memes [Kiela *et al.*, 2020; Fersini *et al.*, 2022; Bhandari *et al.*, 2023]. 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 exemplifies various HS forms targeting immigrants, underscoring animosity towards individuals of diverse nationalities. The text-based approach overtly projects hostile attitudes towards immigrants 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 the digital landscape has seen increasing diversity in HS forms, existing research surveys [Rini *et al.*, 2020; Chhabra and Vishwakarma, 2023; Subramanian *et al.*, 2023] predominantly concentrate on textual HS, often overlooking the complexity introduced by multimodal content. This survey aims to address this critical gap. Unlike prior surveys, we provide a comprehensive analysis of HS across various digital platforms, encompassing not only text but also visual, auditory, and combined multimodal expressions. We delve into the nuanced ways in which HS manifests in these different formats, offering insights into their unique characteristics and the challenges they pose for moderation.

Furthermore, our survey underscores the pivotal role of large language models (LLMs) and large multimodal models (LMMs) in moderating HS. These advanced models, which can process and interpret multiple types of data simultaneously, are crucial in understanding and countering the multifaceted nature of modern HS. We critically assess the existing solutions, evaluating their effectiveness and highlighting areas for improvement. Our survey presents a novel perspective on HS moderation, shifting the focus from traditional text-based analyses to multimodal approaches.

For this survey, we methodically examined research on moderating various types of hate speech, including text, images, videos, and audio. We searched with keywords like ‘hate speech’, ‘multimodal hate speech’, ‘harmful memes’, etc., on various scholarly platforms such as Google Scholar, DBLP, IEEE Xplore, and ACM Digital Library.

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.

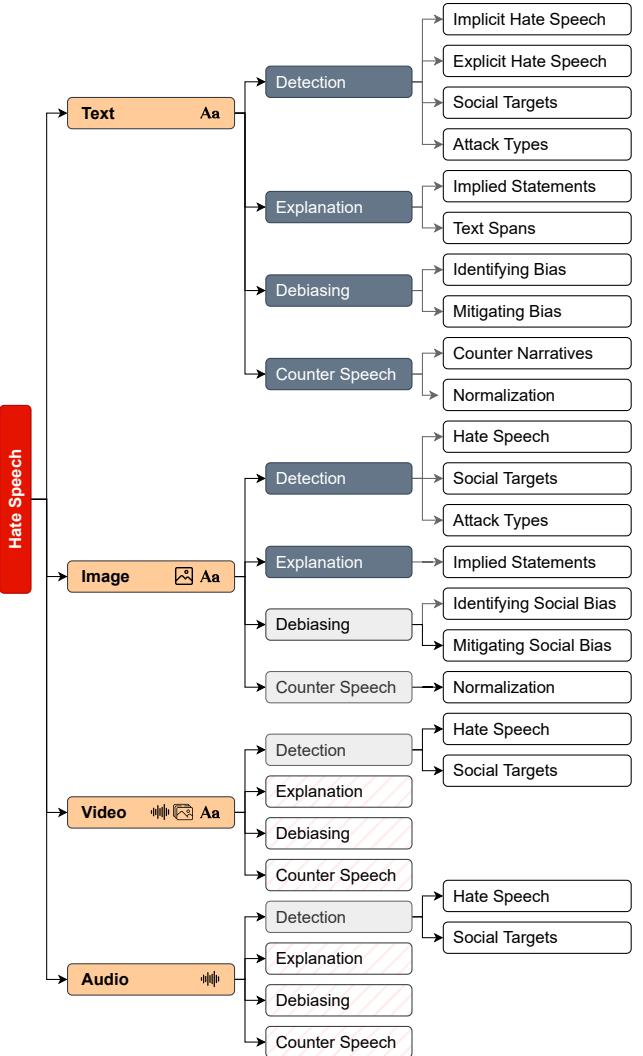


Figure 2: Typology of HS based on modalities and tasks. The dark blue boxes are mature research areas with multiple studies; light grey boxes are ongoing research areas, while the hat can be found, and hatched boxes are unexplored topics.

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. Figure 2 illustrates the range of online HS forms. Additionally, Table 1 lists publicly accessible HS datasets for different modalities, providing researchers with essential resources for data collection and analysis, and supporting comprehensive and consistent research in HS detection and moderation.

2.1 Text-based Hate Speech

Text-based HS encompasses written or typed expressions manifesting across various digital formats, including social media posts [Waseem and Hovy, 2016; Founta *et al.*, 2018]. Researchers addressing online HS have compiled datasets to

detect and analyze its complexity and multifaceted nature, as detailed in existing literature. These studies explore diverse aspects of hateful and derogatory language, focusing on themes such as implicit HS [ElSherief *et al.*, 2021], targeted groups [Kennedy *et al.*, 2018], and types of attacks [ElSherief *et al.*, 2021]. As detection models evolve, it becomes imperative to understand and elucidate their decision-making processes, leading to the development of datasets for implicit statements and methodologies for bias analysis [Sap *et al.*, 2020; ElSherief *et al.*, 2021]. Additionally, some research shifted towards proactive strategies, including countering HS [Masud *et al.*, 2022].

The detection of text-based HS is fraught with significant challenges. A primary obstacle is contextual understanding; HS often subtly or covertly conveys its message through sarcasm, irony, or cultural references. Linguistic variations, including slang, dialects, and unconventional language forms, exacerbate this intricacy, posing difficulties for automated recognition systems. Additionally, the multilingual nature of online content necessitates language-specific tools or models for effective detection across diverse linguistic contexts. A further complication is the rapid evolution of online language, which can quickly render existing detection models outdated.

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 type 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]. The viral nature of memes on digital platforms further amplifies their impact, allowing them to reach vast audiences rapidly. To address the proliferation of hateful memes, recent studies have focused on developing datasets for their detection [Kiela *et al.*, 2020; Gomez *et al.*, 2020] and for identifying specific targets and types of attacks within these memes [Fersini *et al.*, 2022]. Beyond detection, new approaches are being pursued to 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].

Detecting image-based HS presents challenges due to the nuanced expression of offensive messages in visual content. Images, often embedding symbols, memes, or culturally specific visual cues, require deep cultural and contextual understanding for accurate interpretation. The combination of 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 humor 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. 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 in this domain primarily focuses on identifying video-based HS and categorizing its various subtypes [Wu and Bhandary, 2020]. Recent studies have explored detecting hateful segments within videos alongside multi-dimensional analysis. Nonetheless, the body 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.

Identifying HS within videos is a demanding and resource-intensive endeavor. This complexity stems from the amalgamation of various elements, such as text, images, and audio, within a single medium. Each component can independently harbor HS, thereby compounding 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, the analysis of video content requires considerable computational resources and time, posing a substantial challenge for organizations tasked with detecting and addressing 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, broadcasts, 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, back-

ground 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

In this section, we review state-of-the-art methodologies that have made significant contributions to key areas of HS research: *HS detection*, *HS explanation*, *HS debiasing*, and *counter speech*. We emphasise research studies incorporating LLMs or LMMs, aiming to garner comprehensive insights into emerging trends within this domain. This allows us to understand how these advanced models enhance our understanding and management of HS in 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.

In this survey, we regard both 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 instance, AngryBERT [Awal *et al.*, 2021] employs a multi-task learning approach to fine-tune BERT specifically for binary text-based HS detection. PromptHate [Cao *et al.*, 2022] integrates in-context learning with a demonstration sampling strategy for mask prompt fine-tuning on RoBERTa, targeting hateful memes. In audio-based HS classification, studies [Boishakhi *et al.*, 2021; Ibañez *et al.*, 2021] have leveraged ensemble techniques involving AdaBoost, Naive Bayes, and random forest. Additionally, CNNs have been utilized to convert audio into spectrograms [Medina *et al.*, 2022], with self-attentive CNNs [Yousefi and Emmanouilidou, 2021] employed to extract audio features. For video-based HS classification, a combination of BERT, ViT, and MFCC has been used for text, image, and audio modality analysis, respectively [Das *et al.*, 2023]. Note that

video-based and audio-based HS detection represent emerging fields, offering substantial scope for further development and innovation.

Transformer-based models have significantly advanced the detection of text-based and image-based HS; yet they encounter specific 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 difficulty 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. ProCap [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 [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; ElSherief *et al.*, 2021; Hee *et al.*, 2023]. 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, the process of gathering 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 innovative methods for generating rationales in HS detection, which have

Mod.	Dataset	Task	Labels	Source	# Records
	WZ-LS [Waseem and Hovy, 2016]	Det.	[M.C.] Sexism, Racism, Neither	Twitter	16,914
	GHC [Kennedy <i>et al.</i> , 2018]	Det.	[M.C.] VO, HD, CV [B] Implicit, Explicit [M.C.] Hate Targets	Forums	27,665
	Stormfront [de Gibert <i>et al.</i> , 2018]	Det.	[B] Hateful	StormFront	9,916
	DT [Davidson <i>et al.</i> , 2017]	Det.	[M.C.] Hateful, Offensive, Neither	Twitter	24,802
	HatEval [Basile <i>et al.</i> , 2019]	Det.	[B] Hateful [B] Target [B] Aggressive	Twitter	19,600
	Founta [Founta <i>et al.</i> , 2018]	Det.	[M.C.] Offensive, Abusive, Hateful Speech, Aggressive, Cyberbullying, Spam, Normal	Twitter	80,000
	SBIC [Sap <i>et al.</i> , 2020]	Det.	[B] Offensive [B] Intent [B] Lewd [B] Group [B] Hate Targets [B] In-Group	Mixed	44,671
		Expl.	[F.T.] Implied Statement		
Text	IHC [ElSherief <i>et al.</i> , 2021]	Det.	[M.C.] Implicit, Explicit, Non-Hate [M.C.] Grievance, Incitement, Inferiority, Irony, Stereotypical, Threatening, Others	Twitter	22,584
		Expl.	[F.T.] Implied Statement		
	HateXplain [Mathew <i>et al.</i> , 2021]	Det.	[M.C] Hate, Offensive, Normal [M.C.] Hate Targets	Mixed	20,148
		Expl.	[M.L] Text Rationales/Snippets		
	DynaHate [Vidgen <i>et al.</i> , 2021]	Det.	[B] Hateful, [M.C] Animosity, Derogation, Dehumanization, Threatening, Support, [M.C.] Hate Targets	H-M Adv	41,134
	ToxiGen [Mathew <i>et al.</i> , 2021]	Det.	[B] Toxic, Benign	GPT-3	274,186
	NACL [Masud <i>et al.</i> , 2022]	Det.	[M.C.] Hate Intensity [M.L.] Hate Spans	Mixed	4,423
		Ctr.	[F.T.] Hate Speech Normalization		
	CONAN [Chung <i>et al.</i> , 2019]	Det.	[M.C.] Hate Types [M.C.] Hate Sub-Topic	Synthetic	14,988
		Ctr.	[F.T.] CN Generation		
Img	Multitarget CONAN [Fanton <i>et al.</i> , 2021]	Det.	[M.C.] Hate Targets	GPT-2	5,003
		Ctr.	[F.T.] CN Generation		
	Counter Narratives [Das <i>et al.</i> , 2023]	Ctr.	[F.T.] CN Generation	YouTube	9,119
	MMHS150K [Gomez <i>et al.</i> , 2020]	Det.	[B] Hateful	Twitter	150,000
	FHM [Kiela <i>et al.</i> , 2020]	Det.	[B] Hateful	Synthetic	10,000
	Finegrained FHM [Mathias <i>et al.</i> , 2021]	Det.	[B] Hateful [M.L.M.C] Protected Category [M.L.M.C] Protected Attacks	Synthetic	10,000
	Misogynous Meme [Gasparini <i>et al.</i> , 2022]	Det.	[B] Misogynistic [B] Aggressive [B] Ironic	Mixed	800
	MAMI [Fersini <i>et al.</i> , 2022]	Det.	[B] Misogyny [M.L.M.C.] Misogynous, Shaming, Stereotype, Objectification, Violence	Mixed	10,000
	UA-RU Conflict [Thapa <i>et al.</i> , 2022]	Det.	[B] Hateful	Twitter	5,680
Video	CrisisHateMM [Bhandari <i>et al.</i> , 2023]	Det.	[B] Hateful [B] Directed [M.C.] Hate Targets	Mixed	4,723
	HatReD [Hee <i>et al.</i> , 2023]	Expl.	[F.T] Explanations	Synthetic	3,228
Audio	HateMM [Das <i>et al.</i> , 2023]	Det.	[B] Hateful [M.C.] Hate Targets	Mixed	1,083
	Bangla Hate Videos [Junaid <i>et al.</i> , 2021]	Det.	[B] Hateful	YouTube	300
Audio	DeToxy [Ghosh <i>et al.</i> , 2021]	Det.	[B] Hateful / Non-hateful	Mixed	2M
	MuTox [Costa-jussà <i>et al.</i> , 2024]	Det.	[B] Hateful / Non-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.

been shown to improve both the training process and the performance of detection models. 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 [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 actively mitigating bias in HS detection models. This is especially vital as large models increasingly dominate the landscape for generating explanations and enabling transfer learning. Addressing biases within these large models represents a significant opportunity to advance new methodologies and strategies for both the identification and reduction of bias.

3.5 Counter Speech

The approach to counter HS emphasizes creating non-aggressive reactions, aimed at curbing the spread of HS and potentially converting it into respectful and inoffensive language. The primary approach for counter HS involves producing counter-narratives (CN), which rely on reliable evidence, logical arguments, and different perspectives to challenge the hate directed at marginalized groups [Chung *et al.*, 2021]. Multitarget CONAN [Fanton *et al.*, 2021] uses GPT-2 to generate potential CN candidates and then engages human evaluators to identify effective CNs, leading to a dataset of 4,078 CNs. An alternative approach involves either diminishing (i.e., normalization) or eliminating (i.e., correction) the level of hate in HS. NACL [Masud *et al.*, 2022] employs neural networks to paraphrase sections of text containing hate, effectively lessening the intensity of hate in HS. Van and Wu [2023] utilized LMM to identify and correct HS, substituting the text in memes 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 LMMs and LMMs in enhancing approaches and techniques to combat hate speech effectively.

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. Addressing multimodal HS brings to the fore complex challenges concerning data complexity, quality, and sourcing. In text and image-based contexts, issues range from reliance on pre-trained models to the nuanced interpretation of visual cues [Cao *et al.*, 2020; Awal *et al.*, 2021; Sridhar and Yang, 2022]. For audio and video content, challenges are compounded by factors like varied accents, background noise, and the ambiguous nature of toxic content, coupled with inconsistent audio quality [Yousefi and Emmanouilidou, 2021]. 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 profanity across datasets poses significant obstacles for accurate model training. These challenges underscore the need for advanced methods capable of navigating the intricate and multifaceted nature of data across different modalities.

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 different 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, emphasizing the importance of creating models that can adapt to the ever-evolving nature of online material. Such adaptability is crucial for identifying and managing new hateful content, especially those designed to bypass advanced AI technologies. The adoption of technologies, like Few-Shot Learner (FSL), for quick adaptation to this evolving landscape shows a promising direction⁴. However, it is imperative that these technologies not only understand the content but also integrate critical aspects of cultural, behavioral, and conversational contexts into their analysis.

Modality and Expression Variabilities. Research has highlighted the complexities involved in interpreting expressions of HS across individual and combination of various modalities, particularly in cyberbullying scenarios [Boishakhi *et al.*, 2021]. The main challenges in this area include the limitations of pre-trained models in accurately capturing the subtleties of HS expressed via multiple modalities [Cao and Lee, 2020], the difficulty in balancing sensitivity and specificity in detection algorithms [Cao *et al.*, 2020], and the imperative for models to distinguish between explicit and implicit forms of HS [Kim *et al.*, 2022]. Additionally, the potential for misinterpretation and the sensitivity of models to specific trigger language [Sridhar and Yang, 2022] pose significant challenges. Addressing these issues is crucial and requires concerted efforts in future research endeavors.

Contextualization. Recent research underscores the complexities of identifying and countering online hate speech (HS), requiring an accurate and comprehensive understand-

⁴<https://ai.meta.com/blog/harmful-content-can-evolve-quickly-our-new-ai-syst>

ing of HS, ranging from the targeted victims to its implicit meanings [Sap *et al.*, 2020; Masud *et al.*, 2022]. An effective strategy to counter hate speech depends on this precise interpretation, enabling the creation of context-specific responses. The main obstacles in accurately decoding hate speech involve recognizing the subtle nuances in language use, such as wordplays, dual meanings, or ambiguous expressions [Kennedy *et al.*, 2020]. Additionally, the interpretation of hate speech often depends on its context, including cultural, societal, and situational elements [Rizzi *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 detection and mitigation 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 hate speech (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: Confronting bias and upholding ethical considerations in AI systems, particularly in the context of HS detection, represents a challenge, as detailed in works such as [Sandulescu, 2020]. These issues are not just technical but also moral, encompassing the need to ensure that AI systems operate equitably and do not inadvertently perpetuate or amplify existing societal biases. The development and implementation of responsible AI systems, therefore, require a multidisciplinary approach that integrates technical proficiency with ethical and societal awareness. This necessity underlines the importance of continuous efforts to refine AI algorithms, ensuring they align 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: Role of Large Models

To conclude this survey, we explore prospective research opportunities in utilizing large-scale models for moderating hate

content.

Cross-Modality Context Understanding. As hate speech extends beyond mere text to encompass multiple forms of media (multimodality), it becomes crucial for models to possess a proficient understanding of context across these different 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 overall 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 different modalities.

Low-Resource Hate Speech Adaptation. Domain adaptation between related tasks has gained significant attention, with transfer learning emerging as a prominent strategy for addressing this challenge. In the domain of hate speech, an exemplary application is the cross-lingual transfer learning for detecting hate speech across different 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.* [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

This survey has highlighted significant advancements in HS moderation, underscoring the pivotal role of large language and multimodal models. 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 balance innovation with responsibility, striving for a digital landscape that is safer and more inclusive for all.

References

- [Ana, 1999] Otto Santa Ana. ‘like an animal i was treated’: Anti-immigrant metaphor in us public discourse. *Discourse & Society*, 1999.
- [Awal *et al.*, 2021] Md Rabiul Awal, Rui Cao, Roy Ka-Wei Lee, and Sandra Mitrović. Angrybert: Joint learning target and emotion for hate speech detection. In *PAKDD*, 2021.
- [Awan and Zempi, 2016] Imran Awan and Irene Zempi. The affinity between online and offline anti-muslim hate crime: Dynamics and impacts. *Aggression and violent behavior*, 2016.
- [Balkir *et al.*, 2022] Esma Balkir, Isar Nejadgholi, Kathleen C Fraser, and Svetlana Kiritchenko. Necessity and sufficiency for explaining text classifiers: A case study in hate speech detection. In *NAACL*, 2022.
- [Barakat *et al.*, 2012] M. S. Barakat, C. H. Ritz, and D. A. Stirling. Detecting offensive user video blogs: An adaptive keyword spotting approach. In *ICALIP*, 2012.
- [Basile *et al.*, 2019] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *SemEval*, 2019.
- [Bhandari *et al.*, 2023] Aashish Bhandari, Siddhant Bikram Shah, Surendrabikram Thapa, Usman Naseem, and Mehwish Nasim. Crisishatemm: Multimodal analysis of directed and undirected hate speech in text-embedded images from russia-ukraine conflict. In *CVPR Workshops*. IEEE, 2023.
- [Boishakhi *et al.*, 2021] Fariha Tahsin Boishakhi, Ponkoj Chandra Shill, and Md Golam Rabiul Alam. Multi-modal hate speech detection using machine learning. In *Big Data*. IEEE, 2021.
- [Cao and Lee, 2020] Rui Cao and Roy Ka-Wei Lee. Hategan: Adversarial generative-based data augmentation for hate speech detection. In *COLING*, 2020.
- [Cao *et al.*, 2020] Rui Cao, Roy Ka-Wei Lee, and Tuan-Anh Hoang. Deepbate: Hate speech detection via multi-faceted text representations. In *WebSci*, 2020.
- [Cao *et al.*, 2022] Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. Prompting for multimodal hateful meme classification. In *EMNLP*, 2022.
- [Cao *et al.*, 2023] Rui Cao, Ming Shan Hee, Adriel Kuek, Wen-Haw Chong, Roy Ka-Wei Lee, and Jing Jiang. Pro-cap: Leveraging a frozen vision-language model for hateful meme detection. In *ACMMM*, 2023.
- [Chhabra and Vishwakarma, 2023] Anusha Chhabra and Dinesh Kumar Vishwakarma. A literature survey on multimodal and multilingual automatic hate speech identification. *Multimedia Systems*, 2023.
- [Chung *et al.*, 2019] Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. CONAN - COunter Narratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *ACL*, 2019.
- [Chung *et al.*, 2021] Yi-Ling Chung, Serra Sinem Tekiroğlu, and Marco Guerini. Towards knowledge-grounded counter narrative generation for hate speech. In *ACL (Findings)*, 2021.
- [Costa-jussà *et al.*, 2024] Marta R Costa-jussà, Mariano Coria Meglioli, Pierre Andrews, David Dale, Prangthip Hansanti, Elahe Kalbassi, Alex Mourachko, Christophe Ropers, and Carleigh Wood. Mutox: Universal multilingual audio-based toxicity dataset and zero-shot detector. *arXiv preprint arXiv:2401.05060*, 2024.
- [Das *et al.*, 2023] Mithun Das, Rohit Raj, Punyajoy Saha, Binny Mathew, Manish Gupta, and Animesh Mukherjee. Hatemm: A multi-modal dataset for hate video classification. In *ICWSM*, 2023.
- [Davidson *et al.*, 2017] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *ICWSM*, 2017.
- [de Gibert *et al.*, 2018] Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. Hate speech dataset from a white supremacy forum. In Darja Fišer, Ruihong Huang, Vinodkumar Prabhakaran, Rob Voigt, Zeerak Waseem, and Jacqueline Wernimont, editors, *Proc. of the 2st workshop on ab. lang. online*, 2018.
- [ElSherief *et al.*, 2021] Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. Latent hatred: A benchmark for understanding implicit hate speech. In *EMNLP*, 2021.

- [Fanton *et al.*, 2021] Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroglu, and Marco Guerini. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. In *ACL*, 2021.
- [Fersini *et al.*, 2022] Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. Semeval-2022 task 5: Multimedia automatic misogyny identification. In *SemEval@NAACL*, 2022.
- [Founta *et al.*, 2018] Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. Large scale crowdsourcing and characterization of twitter abusive behavior. In *ICWSM*, 2018.
- [Gasparini *et al.*, 2022] Francesca Gasparini, Giulia Rizzi, Aurora Saibene, and Elisabetta Fersini. Benchmark dataset of memes with text transcriptions for automatic detection of multi-modal misogynistic content. *Data in brief*, 2022.
- [Ghosh *et al.*, 2021] Sreyan Ghosh, Samden Lepcha, S Sakshi, Rajiv Ratn Shah, and Srinivasan Umesh. Detox: A large-scale multimodal dataset for toxicity classification in spoken utterances. *arXiv preprint arXiv:2110.07592*, 2021.
- [Gomez *et al.*, 2020] Raul Gomez, Jaume Gibert, Lluís Gómez, and Dimosthenis Karatzas. Exploring hate speech detection in multimodal publications. In *IEEE/WACV*, 2020.
- [Hee *et al.*, 2022] Ming Shan Hee, Roy Ka-Wei Lee, and Wen-Haw Chong. On explaining multimodal hateful meme detection models. In *WWW*, 2022.
- [Hee *et al.*, 2023] Ming Shan Hee, Wen-Haw Chong, and Roy Ka-Wei Lee. Decoding the underlying meaning of multimodal hateful memes. In *IJCAI*, 2023.
- [Ibañez *et al.*, 2021] Michael Ibañez, Ranz Sapinit, Lloyd Antonie Reyes, Mohammed Hussien, Joseph Marvin Imperial, and Ramon Rodriguez. Audio-based hate speech classification from online short-form videos. In *IALP*, 2021.
- [Ji *et al.*, 2023] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 2023.
- [Junaid *et al.*, 2021] Mohd Istiaq Hossain Junaid, Faisal Hossain, and Rashedur M Rahman. Bangla hate speech detection in videos using machine learning. In *UEMCON*, pages 0347–0351. IEEE, 2021.
- [Kennedy *et al.*, 2018] Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, et al. The gab hate corpus: A collection of 27k posts annotated for hate speech. *PsyArXiv*, July, 2018.
- [Kennedy *et al.*, 2020] Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. Contextualizing hate speech classifiers with post-hoc explanation. In *ACL*, 2020.
- [Kiela *et al.*, 2020] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. In *NeurIPS*, 2020.
- [Kim *et al.*, 2022] Youngwook Kim, Shinwoo Park, and Yo-Sub Han. Generalizable implicit hate speech detection using contrastive learning. In *COLING*, 2022.
- [Lee *et al.*, 2021] Roy Ka-Wei Lee, Rui Cao, Ziqing Fan, Jing Jiang, and Wen-Haw Chong. Disentangling hate in online memes. In *ACMMM*, 2021.
- [Lin *et al.*, 2023] Hongzhan Lin, Ziyang Luo, Jing Ma, and Long Chen. Beneath the surface: Unveiling harmful memes with multimodal reasoning distilled from large language models. In *EMNLP (Findings)*, 2023.
- [Luo *et al.*, 2023] Junyu Luo, Cao Xiao, and Fenglong Ma. Zero-resource hallucination prevention for large language models. *arXiv preprint arXiv:2309.02654*, 2023.
- [Lupu *et al.*, 2023] Yonatan Lupu, Richard Sear, Nicolas Velásquez, Rhys Leahy, Nicholas Johnson Restrepo, Beth Goldberg, and Neil F Johnson. Offline events and online hate. *PLoS one*, 2023.
- [Mariconti *et al.*, 2019] Enrico Mariconti, Guillermo Suarez-Tangil, Jeremy Blackburn, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Jordi Luque Serrano, and Gianluca Stringhini. You know what to do proactive detection of youtube videos targeted by coordinated hate attacks. *Proc. of the ACM on HCI*, 2019.
- [Masud *et al.*, 2022] Sarah Masud, Manjot Bedi, Mohammad Afrah Khan, Md Shad Akhtar, and Tanmoy Chakraborty. Proactively reducing the hate intensity of online posts via hate speech normalization. In *ACM-SIGKDD*, 2022.
- [Mathew *et al.*, 2021] Binny Mathew, Punyajoy Saha, Seid Muhib Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. Hateexplain: A benchmark dataset for explainable hate speech detection. In *AAAI*, 2021.
- [Mathias *et al.*, 2021] Lambert Mathias, Shaoliang Nie, Aida Mostafazadeh Davani, Douwe Kiela, Vinodkumar Prabhakaran, Bertie Vidgen, and Zeerak Waseem. Findings of the WOAH 5 shared task on fine grained hateful memes detection. In *WOAH*, 2021.
- [Medina *et al.*, 2022] Robin Medina, Judith Njoku, Jae Min Lee, and Dong-Seong Kim. Audio-based hate speech detection for the metaverse using cnn. In *KICS*, 2022.
- [Nguyen *et al.*, 2023] Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. Extracting cultural commonsense knowledge at scale. In *WWW*, pages 1907–1917, 2023.
- [Nielsen, 2002] Laura Beth Nielsen. Subtle, pervasive, harmful: Racist and sexist remarks in public as hate speech. *Journal of Social issues*, 2002.
- [Ocampo *et al.*, 2023a] Nicolas Ocampo, Elena Cabrio, and Serena Villata. Playing the part of the sharp bully: Generating adversarial examples for implicit hate speech detection. In *ACL (Findings)*, 2023.
- [Ocampo *et al.*, 2023b] Nicolas Benjamin Ocampo, Ekaterina Sviridova, Elena Cabrio, and Serena Villata. An in-depth analysis of implicit and subtle hate speech messages. In *ACL*, 2023.
- [Pramanick *et al.*, 2021] Shravan Pramanick, Shivam Sharma, Dimitar Dimitrov, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. Momenta: A multimodal framework for detecting harmful memes and their targets. pages 4439–4455. ACL, 2021.
- [Rini *et al.*, 2020] Rini Rini, Ema Utami, and Anggit Dwi Hartanto. Systematic literature review of hate speech detection with text mining. In *ICORIS*. IEEE, 2020.
- [Rizzi *et al.*, 2023] Giulia Rizzi, Francesca Gasparini, Aurora Saibene, Paolo Rosso, and Elisabetta Fersini. Recognizing misogynous memes: Biased models and tricky archetypes. *Inf. Proc. Mgmt.*, 2023.

- [Sandulescu, 2020] Vlad Sandulescu. Detecting hateful memes using a multimodal deep ensemble. *CoRR*, 2020.
- [Sap *et al.*, 2019] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. The risk of racial bias in hate speech detection. In *ACL*, 2019.
- [Sap *et al.*, 2020] Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. Social bias frames: Reasoning about social and power implications of language. In *ACL*, 2020.
- [Schmid, 2023] Ursula Kristin Schmid. Humorous hate speech on social media: A mixed-methods investigation of users' perceptions and processing of hateful memes. *New Media & Society*, 2023.
- [Sridhar and Yang, 2022] Rohit Sridhar and Diyi Yang. Explaining toxic text via knowledge enhanced text generation. In *NAACL*, 2022.
- [Subramanian *et al.*, 2023] Malliga Subramanian, Veerappalayam Easwaramoorthy Sathiskumar, G Deepalakshmi, Jaehyuk Cho, and G Manikandan. A survey on hate speech detection and sentiment analysis using machine learning and deep learning models. *Alexandria Engineering Journal*, 2023.
- [Thapa *et al.*, 2022] Surendrabikram Thapa, Aditya Shah, Farhan Jafri, Usman Naseem, and Imran Razzak. A multi-modal dataset for hate speech detection on social media: Case-study of russia-ukraine conflict. In *CASE@EMNLP*, 2022.
- [Touvron *et al.*, 2023] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- [Van and Wu, 2023] Minh-Hao Van and Xintao Wu. Detecting and correcting hate speech in multimodal memes with large visual language model. *CoRR*, 2023.
- [Vidgen *et al.*, 2021] Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *ACL-IJCNLP*, 2021.
- [Wang *et al.*, 2023a] Han Wang, Ming Shan Hee, Md. Rabiul Awal, Kenny Tsu Wei Choo, and Roy Ka-Wei Lee. Evaluating GPT-3 generated explanations for hateful content moderation. In *IJCAI*, 2023.
- [Wang *et al.*, 2023b] Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang. Large language models are latent variable models: Explaining and finding good demonstrations for in-context learning. In *NeurIPS*, 2023.
- [Waseem and Hovy, 2016] Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *NAACL SRW*, 2016.
- [Wazir *et al.*, 2020] Abdulaziz Saleh Ba Wazir, Hezerul Abdul Karim, Mohd Haris Lye Abdullah, Sarina Mansor, Nouar AlDahoul, Mohammad Faizal Ahmad Fauzi, and John See. Spectrogram-based classification of spoken foul language using deep cnn. In *MMSP*, 2020.
- [Wiedlitzka *et al.*, 2023] Susann Wiedlitzka, Gabriele Prati, Rupert Brown, Josh Smith, and Mark A Walters. Hate in word and deed: the temporal ass. between online and offline islamophobia. *Journal of quantitative criminology*, 2023.
- [Winata *et al.*, 2022] Genta Winata, Shijie Wu, Mayank Kulkarni, Thamar Solorio, and Daniel Preoțiu-Pietro. Cross-lingual few-shot learning on unseen languages. In *AACL*, 2022.
- [Wu and Bhandary, 2020] Ching Seh Wu and Unnathi Bhandary. Detection of hate speech in videos using machine learning. In *Int. Conf. on Comp. Science and Comp. Int.* IEEE, 2020.
- [Yang *et al.*, 2023a] Yongjin Yang, Joonkee Kim, Yujin Kim, Namgyu Ho, James Thorne, and Se-Young Yun. HARE: Explainable hate speech detection with step-by-step reasoning. In *EMNLP (Findings)*, pages 5490–5505, 2023.
- [Yang *et al.*, 2023b] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 2023.
- [Yousefi and Emmanouilidou, 2021] Midia Yousefi and Dimitra Emmanouilidou. Audio-based toxic language classification using self-attentive convolutional neural network. In *EUSIPCO*, 2021.
- [Zhao *et al.*, 2023] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.